AEA Continuing Education Program

Cross-section Econometrics

Jeffrey Wooldridge, Michigan State University

January 8-10, 2012

AEA Lectures **Chicago**, **January 2012**

Lecture 2, **Monday**, **January 9**, **8**:**30 am**-**10**:**00 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 discusses methods for estimating dynamic panel data models without strictly exogenous regressors. Recent methods for estimating production functions using firm-level panel data are summarized in Section 5, and Section 6 provides a unified treatment of estimation with pseudo-panel data.

1. 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

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

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 now know that, in

the small T case, viewing the c_i as random draws along with the observed variables is the appropriate posture. 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 (1.3). Of course, if we add $Cov(\mathbf{x}_{it}, c_i) = \mathbf{0}$ for any t, then $\boldsymbol{\beta}$ is identified and can be consistently estimated by a cross section regression using a single time period t, or by pooling across t. But usually the whole point in having panel data is to allow the unobserved effect to be correlated with time-varying $\mathbf{x}_{it}$.

We can allow general correlation between c_i and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ 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. The fixed effects, or within estimator, is the pooled OLS estimator in the equation

$$\ddot{y}_{it} = \ddot{\eta}_t + \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it},\ t = 1,\ldots,T,$$

where $\ddot{y}_{it} = y_{it} - T^{-1}\sum_{r=1}^{T} y_{ir}$ is the deviation of y_{it} from the time average, $\bar{y}_i$ and similarly for $\ddot{\mathbf{x}}_{it}$. Consistency of pooled OLS (for fixed T and $N \to \infty$) essentially requires rests on $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = \sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'u_{it}) = \mathbf{0}$, which means the error u_{it} should be uncorrelated with $\mathbf{x}_{ir}$ for all r and t. The FD estimator is pooled OLS on

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

where $\delta_t = \eta_t - \eta_{t-1}$. Sufficient for consistency is $E(\Delta\mathbf{x}_{it}'\Delta u_{it}) = \mathbf{0}$. See Wooldridge (2010, Chapter 10) for further discussion.

If FE or FD are used, standard inference can and should be made fully robust to heteroskedasticity and serial dependence that could depend on the regressors (or not). These are the now well-known "cluster" standard errors (which we discuss in detail in the notes on cluster sampling). With large N and small T, there is little excuse not to compute them. Even if GLS is used with an unrestricted variance matrix for the $T - 1$ vector $\Delta\mathbf{u}_i$ (in the FD case) or the $T - 1$ vector $\ddot{\mathbf{u}}_i$ (where we drop one time period), the system homoskedasticity assumption, for example, in the FE case, $E(\ddot{\mathbf{u}}_i\ddot{\mathbf{u}}_i'|\ddot{\mathbf{x}}_i) = E(\ddot{\mathbf{u}}_i\ddot{\mathbf{u}}_i')$, need not hold, and so a case can be made for robust inference.

(As an aside, 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, we obtain the same algebraic equivalence using Chamberlain's more flexible devise. That is, if we apply pooled OLS to the equation $y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_{i1}\boldsymbol{\xi}_1 + \ldots + \mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i + u_{it}$, the estimate of $\boldsymbol{\beta}$ is still the FE estimator, even though the $\boldsymbol{\xi}_t$ might change substantially across t. Of course, this estimator is not generally efficient, and Chamberlain shows how to obtain the efficient

minimum distance estimator. See also Wooldridge (2010, Chapter 11).)

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. That is, if a test has no asymptotic power for detecting violation of certain assumptions, the test should be modified to have proper asymptotic size if those assumptions are violated. Much progress has been made in the theoretical literature, but one still sees routine use of Hausman (1978) statistics that maintain a full set of assumptions under the null hypothesis. (Ironically, this often happens in studies where traditional inference about parameters is made fully robust.) Take a leading case, comparing random effects to fixed effects. Once we maintain (1.4), which is used by FE and RE, the key assumption is (1.8), that is, we are interested in finding evidence of whether c_i is correlated with $\mathbf{x}_i$. Of course, the FE estimator is consistent (for the coefficients on time-varying covariates) whether or not c_i is correlated with $\mathbf{x}_i$. And, of course, we need make no assumptions about $Var(\mathbf{u}_i|\mathbf{x}_i, c_i)$ for consistency of FE. Further, RE is consistent under (1.8), whether or not $Var(\mathbf{v}_i|\mathbf{x}_i)$ has the random effects structure, where $v_{it} = c_i + u_{it}$. (In addition to (1.4) and (1.8), sufficient are $Var(\mathbf{u}_i|\mathbf{x}_i, c_i) = \sigma_u^2\mathbf{I}_T$ and $Var(c_i|\mathbf{x}_i) = Var(c_i)$.) In fact, we might be perfectly happy using RE under (1.8) even though it might not be the asymptotically efficient estimator. Therefore, for testing the key assumption (1.8), we should not add the auxiliary assumptions that imply RE is asymptotically efficient. Moreover, as should be clear from the structure of the statistic (and can be shown formally), the usual form of the Hausman statistic has no systematic power for detecting violations of the second moment assumptions on $Var(\mathbf{v}_i|\mathbf{x}_i)$. In particular, if (1.4) and (1.8) hold, the usual statistic converges in distribution to some random variable (not chi-square in general), regardless of the structure of $Var(\mathbf{v}_i|\mathbf{x}_i)$.

To summarize, it makes no sense to report fully robust variance matrices for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions. The regression-based Hausman test from (1.11) is very handy for obtaining a fully robust test, as well as for using the proper degrees of freedom in the limiting distribution. Specifically, suppose the model contains a full set of year intercepts as well as time-constant and time-varying explanatory variables:

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it},\ t = 1,\ldots,T.$$

Now, it is clear that, because we cannot estimate $\boldsymbol{\gamma}$ by FE, it is not part of the Hausman test comparing the RE and FE estimates. What is less clear, but also true, is that the coefficients on the aggregate time variables, $\boldsymbol{\eta}$, cannot be included, either. (RE and FE estimation only with variables that change across *t* 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 also makes it clear that the are only *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 make it easy to obtain a nonnegative test statistic, and it is easy to get confused about the appropriate degrees of freedom in the chi-square distribution. For example, the "Hausman" command in Stata includes year dummies in the comparison between RE and FE; in addition, the test maintains the full set of RE assumptions under the null. 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 (2005) 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 (2.2) 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 $E(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = E(\mathbf{b}_i)$. 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):

$E(\mathbf{b}_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}) = E(\mathbf{b}_i|\bar{\mathbf{x}}_i)$

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 $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, 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 (2005) 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 unobserved "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 $H_0 : \eta_2 = \eta_3 = \ldots = \eta_T$ 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 relationship 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 parameters, 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 (2.9) 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 quasi-differencing method to eliminate 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 (2001) 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, $\mathrm{E}(u_{it}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i) = 0$ 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 individual-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 (2003) 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 simulation results on the performance 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, $\ddot{\hat{\mathbf{v}}}_{it3} = \ddot{\mathbf{y}}_{it3} - \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Pi}}_3$, 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 without 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

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

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 $\text{E}(\mathbf{e}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of $\mathbf{x}_{it}$ are random walks with drift – then $\text{E}(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails.

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

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

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) unobserved 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 the 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 other 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 combined 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 interpreted 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 $|\rho| < 1$ by recursive substitution 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 in the parameters 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 variables, 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 for a random draw i is

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

Of course, if the log likelihood represents the correct density of $(y_{i1},\ldots,y_{iT})$ given y_{i0}, 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. See Blundell and Bond (1998) for further discussion.

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

as

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

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

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

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. The data set can be obtained from the web site for Wooldridge (2010), and is called AIRFARE.RAW.

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 reasonably large *T*. Alvarez and Arellano (2003) 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.

References

Ahn, S.C. Y.H. Lee, and P. Schmidt (2001), "GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects," *Journal of Econometrics* 101, 219-255.

Ahn, S.C. and P. Schmidt (1995), "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics* 68, 5-27.

Alvarez, J. and M. Arellano (2003), "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica* 71, 1121-1159

Anderson, T.W. and C. Hsiao (1982), "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics* 18, 47-82.

Arellano, M. (1993), "On the Testing of Correlated Effects with Panel Data," *Journal of Econometrics* 59, 87-97.

Arellano, M. (2003), *Panel Data Econometrics*. Oxford University Press: Oxford.

Arellano, M. and S.R. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58, 277-297.

Arellano, M. and O. Bover (1995), "Another Look at the Instrumental Variable Estimation of Error Components Models," *Journal of Econometrics* 68, 29-51.

Arellano, M. and B. Honoré (2001), "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, Volume 5, ed. J.J. Heckman and E. Leamer. Amsterdam: North Holland, 3229-3296.

Blundell, R. and S.R. Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics* 87, 115-143.

Blundell, R. and S.R. Bond (2000). "GMM Estimation with Persistent Panel Data: An

Application to Production Functions," *Econometric Reviews* 19, 321-340.

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

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

Engle, R.F., D.F. Hendry, and J.-F. Richard (1983), "Exogeneity," *Econometrica* 51, 277-304.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251-1271.

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

Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988), "Estimating Vector Autoregressions with Panel Data," *Econometrica* 56, 1371-1395.

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

Murtazashvili, I. and J.M. Wooldridge (2007), "Fixed Effects Instrumental Variables Estimation in Correlated Random Coefficient Panel Data Models," *Journal of Econometrics* 142, 539-552.

Robinson, P.M. (1988), "Root-*n* Consistent Semiparametric Regression," *Econometrica* 55, 931-954.

Semykina, A. (2006), "A Semiparametric Approach to Estimating Panel Data Models with Endogenous Explanatory Variables and Selection," mimeo, Florida State University

Department of Economics.

Verbeek, M. and T.E. Nijman (1993), "Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections," *Journal of Econometrics* 59, 125-136.

Wooldridge, J.M. (2003), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters* 79, 185-191.

Wooldridge, J.M. (2005), "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics* 87, 385-390.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. MIT Press: Cambridge, MA.

Cross-Section Econometrics
Lecture 2: Linear Panel Data Models

Jeff Wooldridge
Michigan State University

AEA Lectures, Chicago, January 2012

1. Overview of the Basic Model
2. Recent Insights Into Old Estimators
3. Behavior of Estimators without Strict Exogeneity
4. IV Estimation under Sequential Exogeneity

1. Overview of the Basic Model

- Unless stated otherwise, cover assume a large cross section and small time series, although some approximations are based on T increasing.
- For a generic unit i in the population,

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

- η_t is a separate time period intercept, $\mathbf{x}_{it}$ 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.

- The η_t are parameters (T of them) that can be estimated. c_i is a random draw along with $(\mathbf{x}_{it}, y_{it})$.
- An attractive assumption is *contemporaneous exogeneity conditional on* c_i:

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0, \; t = 1, \ldots, T, \tag{2}$$

which implies

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{3}$$

The β_j are partial effects holding c_i fixed.

- $\boldsymbol{\beta}$ is not identified only under (2). If we add $Cov(\mathbf{x}_{it}, c_i) = \mathbf{0}$, then $\boldsymbol{\beta}$ is identified.

- To allow any correlation between $\mathbf{x}_{it}$ and c_i, assume *strict exogeneity conditional on* c_i:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0, \; t = 1, \ldots, T, \tag{4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_i, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{5}$$

- If $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ has suitable time variation, $\boldsymbol{\beta}$ can be consistently estimated by fixed effects (FE) or first differencing (FD), as well as GLS and GMM versions of them.

• The FE estimator uses the deviations from time averages to remove c_i (absorb time dummies in $\mathbf{x}_{it}$):

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, \quad t = 1, \ldots, T, \tag{6}$$

where $\ddot{y}_{it} = y_{it} - T^{-1}\sum_{r=1}^{T} y_{ir}$, and so on. FE is pooled OLS on (6).

• Canned packages provide proper standard errors and inference (with the proper "degrees-of-freedom" adjustment). But the "usual" (nonrobust) inference assumes homoskedasticity and serial independence in $\{u_{it}\}$.

• Make inference fully robust to heteroksedasticity and serial dependence. With large N and small T, there is little excuse not to compute "cluster robust" standard errors.

• Treating the c_i as parameters to estimate can lead to trouble even in the linear model: an attempt to "cluster" the standard errors to allow arbitrary serial correlation leads to meaningless standard errors for the $\hat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}$.

• An alternative way to remove c_i is to first difference:

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, t = 2, \ldots, T. \tag{7}$$

The FD estimator is pooled OLS on the first differences.

• FD also requires a kind of strict exogeneity, namely, that u_{it} is uncorrelated with $\mathbf{x}_{i,t-1}, \mathbf{x}_{it}$, and $\mathbf{x}_{i,t+1}$.

• Failure of strict exogeneity will cause different inconsistencies in FE and FD when $T > 2$.

• Should make inference robust to serial correlation and heteroskedasticity in the differenced errors, $e_{it} \equiv u_{it} - u_{i,t-1}$. For example, if $\{u_{it}\}$ is serially uncorrelated, $Corr(e_{it}, e_{i,t+1}) = -.5$.

• In unbalanced cases, FD requires that data exists in adjacent time periods. FE does not.

• Even with FE and FD, have to worry about violations of strict exogeneity: strict exogeneity is always violated if $\mathbf{x}_{it}$ contains lagged dependent variables, but also if changes in u_{it} cause changes in $\mathbf{x}_{i,t+1}$ ("feedback effect").

• *Sequential exogeneity condition on c_i*:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}, c_i) = 0, t = 1, \ldots, T \tag{8}$$

or, maintaining the linear model,

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i). \tag{9}$$

Allows for lagged dependent variables and other feedback.

• Sequential exogeneity imposes a certain form of correct dynamics, but does not rule out feedback from shocks to y_{it} to $\mathbf{x}_{i,t+1}$.

• If $\mathbf{x}_{it}$ contains $y_{i,t-1}$ (and perhaps other variables $\mathbf{z}_{it}$ and lags), sequential exogeneity rules out serial correlation in $\{u_{it}\}$.

• If, say, $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i,t-Q})$ then sequential exogeneity implies correct distributed lag dynamics, but allows shocks u_{it} to be correlated with $\mathbf{z}_{i,t+1}$. $\{u_{it}\}$ can be serially correlated.

• Generally, $\boldsymbol{\beta}$ is identified under sequential exogeneity.

• The key "random effects" assumption is

$$E(c_i|\mathbf{x}_i) = E(c_i). \tag{10}$$

• RE leaves c_i in the error term and accounts for the serial correlation in the composite error, $c_i + u_{it}$, via generalized least squares. The nominal assumption is homoskedasticity and serial independence in $\{u_{it}\}$. But RE inference can also be made fully robust to violations of this assumption.

• Can show RE can be computed as a pooled OLS estimator on quasi-time-demeaned data:

$$y_{it} - \theta\bar{y}_i = (\mathbf{x}_{it} - \theta\bar{\mathbf{x}}_i)\boldsymbol{\beta} + v_{it} - \theta\bar{v}_i \tag{11}$$

where $v_{it} = c_i + u_{it}$ and

$$\theta = 1 - \left[\frac{1}{1 + T(\sigma_c^2/\sigma_u^2)}\right]^{1/2}, \tag{12}$$

• RE can be close to FE with large T or when σ_c^2/σ_u^2 is large, or when $E(c_i|\mathbf{x}_i) = E(c_i)$.

• Advantages of RE: (a) RE allows inclusion of time-constant variables; (b) Can be substantially more efficient than FE.

• Under the full set of RE assumptions,

$$Avar(\hat{\boldsymbol{\beta}}_{FE}) = \sigma_u^2[E(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i)]^{-1}/N$$
$$Avar(\hat{\boldsymbol{\beta}}_{RE}) = \sigma_u^2[E(\check{\mathbf{X}}_i'\check{\mathbf{X}}_i)]^{-1}/N,$$

where $\check{\mathbf{x}}_{it} = \mathbf{x}_{it} - \theta\bar{\mathbf{x}}_i$ are the quasi-time demeaned time-varying covariates.

• But RE is inconsistent without $E(c_i|\mathbf{x}_i) = E(c_i)$ (or at least zero correlation).

• Some important algebraic equivalances: If

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

then $\hat{\boldsymbol{\eta}}_{RE} = \hat{\boldsymbol{\eta}}_{FE} = \hat{\boldsymbol{\eta}}_{POLS}$ and $\hat{\boldsymbol{\gamma}}_{RE} = \hat{\boldsymbol{\gamma}}_{POLS}$ (where POLS = Pooled OLS).

• $\hat{\boldsymbol{\eta}}_{RE} = \hat{\boldsymbol{\eta}}_{FE}$ has implications for Hausman test comparing RE and FE.

• Define two *correlated random effects* (CRE) assumptions. The first just uses the definition of a linear projection:

$$L(c_i|\mathbf{x}_i) = \psi + \mathbf{x}_i\boldsymbol{\xi} \tag{13}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$. This is often called the *Chamberlain device*, after Chamberlain (1982).

• Mundlak (1978) used a restricted version

$$E(c_i|\mathbf{x}_i) = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \tag{14}$$

where $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$.

• In the equation

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}. \tag{15}$$

we write $c_i = \psi + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i$ and then get the Mundlak equation

$$y_{it} = \psi + \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + u_{it}, \tag{16}$$

and we can apple pooled OLS or RE because $E(a_i + u_{it}|\mathbf{x}_i) = 0$. Both equal the FE estimator of $\boldsymbol{\delta}$.

• Assumptions such as $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$ are very convenient for nonlinear models.

• The Mundlak equation makes it easy to compute a fully robust Hausman test comparing RE and FE. In the equation

$$y_{it} = \psi + \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + u_{it} \qquad (17)$$

test $H_0 : \boldsymbol{\xi} = \mathbf{0}$ using a fully robust Wald statistic after RE estimation.

• We can only compare $\hat{\boldsymbol{\delta}}_{RE}$ and $\hat{\boldsymbol{\delta}}_{FE}$ (M parameters), not $\hat{\boldsymbol{\eta}}_{RE}$ and $\hat{\boldsymbol{\eta}}_{FE}$.

• Equation (17) allows us to estimate coefficients on $\mathbf{z}_i$ while allowing correlation between c_i and $\bar{\mathbf{w}}_i$. (Should use caution in interpreting the coefficients on $\mathbf{z}_i$).

• Be wary of canned Hausman test procedures that directly compare estimates: the *df* are often wrong (the aggregate time variables are counted) and the tests are nonrobust. Can get negative statistics, too.

EXAMPLE: For $N = 1,149$ U.S. air routes and the years 1997 through 2000, y_{it} is $lfare_{it} = \log(fare_{it})$ and the key explanatory variable is $concen_{it}$, the concentration ratio for route i. Other covariates are year dummies and the time-constant variables $ldist_i = \log(dist_i)$ and $ldist_i^2$. Called AIRFARE.DTA.

$$lfare_{it} = \alpha_t + \beta_1 concen_{it} + \beta_2 ldist_i + \beta_3 ldist_i^2 + c_i + u_{it}$$


```
. des fare lfare concen dist

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
fare            int    %9.0g                  avg. one-way fare, $
lfare           float  %9.0g                  log(fare)
concen          float  %9.0g                  = bmktshr
dist            int    %9.0g                  distance, in miles

. sum fare concen dist

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fare |      4596    178.7968    74.88151         37        522
      concen |      4596    .6101149     .196435      .1605          1
        dist |      4596     989.745    611.8315         95       2724

. tab year

1997, 1998, |
 1999, 2000 |      Freq.     Percent        Cum.
------------+-----------------------------------
       1997 |      1,149       25.00       25.00
       1998 |      1,149       25.00       50.00
       1999 |      1,149       25.00       75.00
       2000 |      1,149       25.00      100.00
------------+-----------------------------------
      Total |      4,596      100.00
```

```
. reg lfare concen ldist ldistsq y98 y99 y00

      Source |       SS       df       MS              Number of obs =    4596
-------------+------------------------------           F(  6,  4589) =  523.18
       Model |  355.453858     6  59.2423096           Prob > F      =  0.0000
    Residual |  519.640516  4589  .113236112           R-squared     =  0.4062
-------------+------------------------------           Adj R-squared =  0.4054
       Total |  875.094374  4595  .190444913           Root MSE      =  .33651

------------------------------------------------------------------------------
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .3601203   .0300691    11.98   0.000     .3011705    .4190702
       ldist |  -.9016004    .128273    -7.03   0.000    -1.153077   -.6501235
     ldistsq |   .1030196   .0097255    10.59   0.000     .0839529    .1220863
         y98 |   .0211244   .0140419     1.50   0.133    -.0064046    .0486533
         y99 |   .0378496   .0140413     2.70   0.007      .010322    .0653772
         y00 |     .09987   .0140432     7.11   0.000     .0723385    .1274015
       _cons |   6.209258   .4206247    14.76   0.000     5.384631    7.033884
------------------------------------------------------------------------------
```



```
. reg lfare concen ldist ldistsq y98 y99 y00, cluster(id)

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .3601203    .058556     6.15   0.000     .2452315    .4750092
       ldist |  -.9016004   .2719464    -3.32   0.001    -1.435168   -.3680328
     ldistsq |   .1030196   .0201602     5.11   0.000     .0634647    .1425745
         y98 |   .0211244   .0041474     5.09   0.000     .0129871    .0292617
         y99 |   .0378496   .0051795     7.31   0.000     .0276872     .048012
         y00 |     .09987   .0056469    17.69   0.000     .0887906    .1109493
       _cons |   6.209258   .9117551     6.81   0.000     4.420364    7.998151
------------------------------------------------------------------------------
```



```
. xtreg lfare concen ldist ldistsq y98 y99 y00, re

Random-effects GLS regression                   Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

R-sq:  within  = 0.1348                         Obs per group: min =         4
       between = 0.4176                                        avg =       4.0
       overall = 0.4030                                        max =         4

Random effects u_i ~Gaussian                    Wald chi2(6)       =   1360.42
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .2089935   .0265297     7.88   0.000     .1569962    .2609907
       ldist |  -.8520921   .2464836    -3.46   0.001    -1.335191   -.3689931
     ldistsq |   .0974604   .0186358     5.23   0.000     .0609348     .133986
         y98 |   .0224743   .0044544     5.05   0.000     .0137438    .0312047
         y99 |   .0366898   .0044528     8.24   0.000     .0279626    .0454171
         y00 |    .098212   .0044576    22.03   0.000     .0894752    .1069487
       _cons |   6.222005   .8099666     7.68   0.000       4.6345     7.80951
-------------+----------------------------------------------------------------
     sigma_u |  .31933841
     sigma_e |  .10651186
         rho |  .89988885   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```



```
. xtreg lfare concen ldist ldistsq y98 y99 y00, re cluster(id)

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .2089935   .0422459     4.95   0.000      .126193    .2917939
       ldist |  -.8520921   .2720902    -3.13   0.002    -1.385379   -.3188051
     ldistsq |   .0974604   .0201417     4.84   0.000     .0579833    .1369375
         y98 |   .0224743   .0041461     5.42   0.000      .014348    .0306005
         y99 |   .0366898   .0051318     7.15   0.000     .0266317     .046748
         y00 |    .098212   .0055241    17.78   0.000     .0873849     .109039
       _cons |   6.222005   .9144067     6.80   0.000     4.429801    8.014209
-------------+----------------------------------------------------------------
     sigma_u |  .31933841
     sigma_e |  .10651186
         rho |  .89988885   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. xtreg lfare concen ldist ldistsq y98 y99 y00, fe

Fixed-effects (within) regression               Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

R-sq:  within  = 0.1352                         Obs per group: min =         4
       between = 0.0576                                        avg =       4.0
       overall = 0.0083                                        max =         4

                                                F(4,3443)          =    134.61
corr(u_i, Xb)  = -0.2033                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |    .168859   .0294101     5.74   0.000     .1111959     .226522
       ldist |  (dropped)
     ldistsq |  (dropped)
         y98 |   .0228328   .0044515     5.13   0.000     .0141048    .0315607
         y99 |   .0363819   .0044495     8.18   0.000     .0276579    .0451058
         y00 |   .0977717   .0044555    21.94   0.000      .089036    .1065073
       _cons |   4.953331   .0182869   270.87   0.000     4.917476    4.989185
-------------+----------------------------------------------------------------
     sigma_u |  .43389176
     sigma_e |  .10651186
         rho |  .94316439   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(1148, 3443) =    36.90          Prob > F = 0.0000
```



```
. xtreg lfare concen ldist ldistsq y98 y99 y00, fe cluster(id)

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |    .168859   .0494587     3.41   0.001     .0718194    .2658985
       ldist |  (dropped)
     ldistsq |  (dropped)
         y98 |   .0228328    .004163     5.48   0.000     .0146649    .0310007
         y99 |   .0363819   .0051275     7.10   0.000     .0263215    .0464422
         y00 |   .0977717   .0055054    17.76   0.000     .0869698    .1085735
       _cons |   4.953331   .0296765   166.91   0.000     4.895104    5.011557
-------------+----------------------------------------------------------------
     sigma_u |  .43389176
     sigma_e |  .10651186
         rho |  .94316439   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```



```
. * First use the Hausman test that maintains all of the RE assumptions under
. * the null:

. qui xtreg lfare concen ldist ldistsq y98 y99 y00, fe

. estimates store b_fe

. qui xtreg lfare concen ldist ldistsq y98 y99 y00, re

. estimates store b_re
```



```
. hausman b_fe b_re

                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      b_fe         b_re        Difference          S.E.
-------------+----------------------------------------------------------------
      concen |     .168859     .2089935       -.0401345        .0126937
         y98 |    .0228328     .0224743        .0003585               .
         y99 |    .0363819     .0366898        -.000308               .
         y00 |    .0977717      .098212       -.0004403               .
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       10.00
                Prob>chi2 =      0.0405
                (V_b-V_B is not positive definite)

.
. di -.0401/.0127
-3.1574803

. * This is the nonrobust H t test based just on the concen variable. There is
. * only one restriction to test, not four. The p-value reported for the
. * chi-square statistic is incorrect.
```

```
. * Using the same variance matrix estimator solves the problem of wrong df.
. * The next command uses the matrix of the relatively efficient estimator.

. hausman b_fe b_re, sigmamore

Note: the rank of the differenced variance matrix (1) does not equal the
      number of coefficients being tested (4); be sure this is what you expect,
      or there may be problems computing the test.  Examine the output of your
      estimators for anything unexpected and possibly consider scaling your
      variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       b_fe         b_re        Difference          S.E.
-------------+------------------------------------------------------------
      concen |     .168859     .2089935       -.0401345        .0127597
         y98 |    .0228328     .0224743        .0003585         .000114
         y99 |    .0363819     .0366898        -.000308        .0000979
         y00 |    .0977717      .098212       -.0004403          .00014
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =          9.89
                Prob>chi2 =        0.0017
```



```
. * The regression-based test is better: it gets the df right and is fully
. * robust to violations of the RE variance-covariance matrix:

. egen concenbar = mean(concen), by(id)

. xtreg lfare concen concenbar ldist ldistsq y98 y99 y00, re cluster(id)

                                     (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |    .168859   .0494749     3.41   0.001       .07189     .2658279
   concenbar |   .2136346   .0816403     2.62   0.009      .0536227     .3736466
       ldist |  -.9089297   .2721637    -3.34   0.001    -1.442361    -.3754987
     ldistsq |   .1038426   .0201911     5.14   0.000      .0642688     .1434164
         y98 |   .0228328   .0041643     5.48   0.000      .0146708     .0309947
         y99 |   .0363819   .0051292     7.09   0.000      .0263289     .0464349
         y00 |   .0977717   .0055072    17.75   0.000      .0869777     .1085656
       _cons |   6.207889   .9118109     6.81   0.000      4.420773     7.995006
-------------+----------------------------------------------------------------
     sigma_u |  .31933841
     sigma_e |  .10651186
         rho |  .89988885   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. * So the robust t statistic is 2.62 --- still a rejection, but not as strong.
```



```
. * What if we do not control for distance in RE?

. xtreg lfare concen y98 y99 y00, re cluster(id)

Random-effects GLS regression                   Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

                                     (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .0468181   .0427562     1.09   0.274    -.0369826     .1306188
         y98 |   .0239229   .0041907     5.71   0.000     .0157093     .0321364
         y99 |   .0354453   .0051678     6.86   0.000     .0253167      .045574
         y00 |   .0964328   .0055197    17.47   0.000     .0856144     .1072511
       _cons |   5.028086   .0285248   176.27   0.000     4.972178     5.083993
-------------+----------------------------------------------------------------
     sigma_u |  .40942871
     sigma_e |  .10651186
         rho |  .93661309   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. * The RE estimate is now way below the FE estimate, .169. Thus, it can be
. * very harmful to omit time-constant variables in RE estimation.
```


Instrumental Variables and Fixed Effects

- Can combine IV methods with unobserved effects. Allow contemporaneous endogeneity: correlation between $\mathbf{x}_{it}$ and u_{it}, in addition to correlation between $\mathbf{x}_{it}$ and c_i.
- Let $\ddot{\mathbf{z}}_{it} = \mathbf{z}_{it} - \bar{\mathbf{z}}_i$ be time-demeaned instruments. Apply IV methods to

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, \; t = 1, \ldots, T, \tag{18}$$

such as pooled 2SLS. This gives the FEIV estimator. (Fully robust inference available in `xtivreg2` in Stata.)

• FEIV allows arbitrary correlation between c_i and $\mathbf{z}_{it}$, but maintains a strict exogeneity assumption with respect to $\{u_{it}\}$:

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

• The FEIV estimator is equivalent to applying REIV to

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{z}}_i\boldsymbol{\xi} + c_i + u_{it}$$

using IVs $(\mathbf{z}_{it}, 1, \bar{\mathbf{z}}_i)$.

• A fully robust test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$ is a fully robust test comparing FEIV and REIV.

2. Recent Insights Into Old Estimators

• Consider an extension of the usual model to allow for unit-specific slopes,

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{19}$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0, t = 1, \ldots, T, \tag{20}$$

where $\mathbf{b}_i$ is $K \times 1$. We act as if $\mathbf{b}_i$ is constant for all i but think c_i might be correlated with $\mathbf{x}_{it}$; we apply usual FE estimator.

• When does the usual FE estimator consistently estimate the population average effect, $\boldsymbol{\beta} = E(\mathbf{b}_i)$?

• A sufficient condition for consistency of the FE estimator, along with $E(u_{it}|\mathbf{x}_i, c_i) = 0$ and the usual FE rank condition, is

$$E(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = E(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \ldots, T \tag{21}$$

where $\ddot{\mathbf{x}}_{it}$ are the time-demeaned covariates. Allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$. Then (21) holds if $E(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = E(\mathbf{b}_i)$.

• Extends to a more general class of FE estimators. Write

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it}, \quad t = 1, \ldots, T \tag{22}$$

where $\mathbf{w}_t$ is a set of deterministic functions of time. Now FE refers to sweeping away $\mathbf{a}_i$ by netting out $\mathbf{w}_t$ from $\mathbf{x}_{it}$.

• In the random trend model, $\mathbf{w}_t = (1, t)$, and now the elements of $\ddot{\mathbf{x}}_{it}$ have had unit-specific linear trends removed. If $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$.

• Generally, need $\dim(\mathbf{w}_t) < T$.

• Can apply to models with time-varying factor loads, η_t :

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i + u_{it}, t = 1, \ldots, T. \tag{23}$$

Sufficient for consistency of FE estimator that ignores the η_t is

$$Cov(\ddot{\mathbf{x}}_{it}, c_i) = \mathbf{0}, t = 1, \ldots, T. \tag{24}$$

• Can use nonlinear GMM to estimate $\boldsymbol{\beta}$ along with the η_t; estimates are often similar to FE.

Robustness of FEIV

• FEIV methods also have some robustness in the general random slopes model

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

The slopes $\mathbf{b}_i$ can be correlated with $\mathbf{x}_{it}$.

• Assume the instruments are strictly exogenous conditioanal on $(\mathbf{a}_i, \mathbf{b}_i)$:

$$E(u_{it}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i) = 0. \tag{25}$$

• Along with

$$E(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = E(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \ldots, T. \tag{26}$$

also assume

$$Cov(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = Cov(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i), t = 1, \ldots, T. \tag{27}$$

The $K \times K$ matrix unconditional covariance matrix, $Cov(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i)$, is unrestricted. The *conditional* covariance cannot depend on the $\ddot{\mathbf{z}}_{it}$. Then, FEIV is consistent for $\boldsymbol{\beta} = E(\mathbf{b}_i)$ provided a full set of time dummies is included.

• Assumption (27) can hold for continuous $\mathbf{x}_{it}$ but is unrealistic when endogenous elements of $\mathbf{x}_{it}$ are discrete.

Testing for Correlated Random Slopes

• Simple test to see whether the slopes $\mathbf{b}_i$ depend on observed factors, say, $\mathbf{w}_i$ (which do not change over time) and $\bar{\mathbf{x}}_i$ (the time averages of time-varying covariates):

$$y_{it} = \eta_t + \mathbf{x}_{it}\mathbf{b}_i + c_i + u_{it} \tag{28}$$

$$c_i = \alpha + (\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h})\boldsymbol{\gamma} + a_i \tag{29}$$

$$\mathbf{b}_i = \boldsymbol{\beta} + \boldsymbol{\Pi}(\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h})' + \mathbf{d}_i \tag{30}$$

where $\mathbf{h}_i = (\bar{\mathbf{x}}_i, \mathbf{w}_i)$ (a row vector) and $E(a_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{w}_i) = 0$, $E(\mathbf{d}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{w}_i) = \mathbf{0}$.

• After some algebra,

$$y_{it} = \alpha_t + (\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h})\boldsymbol{\gamma} + \mathbf{x}_{it}\boldsymbol{\beta} + [(\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h}) \otimes \mathbf{x}_{it}]\boldsymbol{\pi} + a_i + \mathbf{x}_{it}\mathbf{d}_i + u_{it}, \quad (31)$$

which just means interact elements of $\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h}$ with elements of $\mathbf{x}_{it}$.

• In practice, replace $\boldsymbol{\mu}_\mathbf{h}$ with $\bar{\mathbf{h}}$ (sample average) and use pooled OLS.

• Can even estimate (31) by FE, which removes $(\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h})$ but not the interactions. The interactions might be significant even though the estimates of $\boldsymbol{\beta}$ (population averaged effect) might be similar to FE on the equation without the interactions, $[(\mathbf{h}_i - \boldsymbol{\mu}_\mathbf{h}) \otimes \mathbf{x}_{it}]$.

• Can use random effects, too, although it would ignore $\mathbf{x}_{it}\mathbf{d}_i$ (so, at a minimum, inference should be made fully robust).

• A GLS estimator that accounts for $\mathbf{x}_{it}\mathbf{d}_i$ is possible (but may not be worth it if want mean effects).

EXAMPLE: Airfare Application.

```
. egen concenb = mean(concen), by(id)

. sum concenb ldist ldistsq

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     concenb |      4596    .6101149    .1888741       .1862       .9997
       ldist |      4596    6.696482    .6593177   4.553877   7.909857
     ldistsq |      4596    45.27747    8.726898   20.73779   62.56583

. gen cbconcen = (concenb - .61)*concen
. gen ldconcen = (ldist - 6.696)*concen
. gen ldsqconcen = (ldistsq - 45.277)*concen
```



```
. xtreg lfare concen concenb cbconcen ldconcen ldsqconcen ldist ldistsq
      y98 y99 y00, re cluster(id)

                                    (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .1682492   .0496695     3.39   0.001     .0708988    .2655996
     concenb |    .157291   .2085049     0.75   0.451    -.2513711     .565953
    cbconcen |   .0635453   .3033809     0.21   0.834    -.5310704    .6581609
    ldconcen |  -.2994869   .9930725    -0.30   0.763    -2.245873    1.646899
  ldsqconcen |   .0112477   .0746874     0.15   0.880     -.135137    .1576324
       ldist |  -.4394368   .6713288    -0.65   0.513    -1.755217    .8763435
     ldistsq |   .0752147   .0494201     1.52   0.128    -.0216469    .1720764
         y98 |   .0229684   .0041542     5.53   0.000     .0148262    .0311105
         y99 |   .0358549   .0051298     6.99   0.000     .0258007    .0459091
         y00 |   .0976256    .005461    17.88   0.000     .0869221     .108329
       _cons |   4.382552   2.272566     1.93   0.054    -.0715953    8.836699
-------------+----------------------------------------------------------------
. * Nonrobust se for concen is .0295.
. test cbconcen ldconcen ldsqconcen

 ( 1)  cbconcen = 0
 ( 2)  ldconcen = 0
 ( 3)  ldsqconcen = 0

           chi2(  3) =    5.47
         Prob > chi2 =    0.1407
```

3. Behavior of Estimators without Strict Exogeneity

• Both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the strict exogeneity assumption. But inconsistencies (as function of T) can differ.

• If we maintain $E(u_{it}|\mathbf{x}_{it}, c_i) = 0$ and assume $\{(\mathbf{x}_{it}, u_{it}) : t = 1, \ldots, T\}$ is "weakly dependent", can show

$$\text{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{FE} = \boldsymbol{\beta} + O(T^{-1}) \quad (32)$$

$$\text{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{FD} = \boldsymbol{\beta} + O(1). \quad (33)$$

• Result (32) still holds if $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ has unit roots as long as $\{u_{it}\}$ is I(0) and contemporaneous exogeneity holds.

($y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$ is a "cointegrating relationship.")

• Important caveat: if $\{u_{it}\}$ is I(1) – so that the time series "model" is a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not *cointegrated*), then (32) is no longer true. FD is attractive because it eliminates any unit roots.

• Same conclusions hold for instrumental variables versions: FE has bias of order T^{-1} if instruments are contemporaneously exogenous and $\{u_{it}\}$ is weakly dependent.

• Simple test for lack of strict exogeneity (or feedback) in covariates:

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it} \quad (34)$$

where $\mathbf{w}_{it}$ is a subset of $\mathbf{x}_{it}$. Estimate the equation by fixed effects (losing the last time period) and test $H_0 : \boldsymbol{\delta} = \mathbf{0}$.


```
. sort id year
. gen concenp1 = concen[_n+1] if year < 2000
. xtreg lfare concen concenp1 y98 y99 y00, fe cluster(id)

Fixed-effects (within) regression               Number of obs      =      3447
Group variable: id                              Number of groups   =      1149

R-sq:  within  = 0.0558                         Obs per group: min =         3
       between = 0.0535                                        avg =       3.0
       overall = 0.0347                                        max =         3

                                                F(4,1148)          =     25.63
corr(u_i, Xb)  = -0.2949                        Prob > F           =    0.0000

                                    (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      concen |   .2983988    .054797     5.45   0.000     .1908854    .4059122
    concenp1 |  -.0659259   .0467578    -1.41   0.159    -.1576663    .0258145
        y98 |   .0205809   .0042341     4.86   0.000     .0122735    .0288883
        y99 |   .0360638   .0050754     7.11   0.000     .0261058    .0460218
        y00 |  (dropped)
      _cons |   4.914953   .0478488   102.72   0.000     4.821072    5.008834
------------------------------------------------------------------------------
```

4. IV Estimation under Sequential Exogeneity

• We now consider IV estimation of the model

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

under the sequential exogeneity assumption

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

• Actually, for consistency, we can get by with the weaker form

$$Cov(\mathbf{x}_{is}, u_{it}) = 0, \text{all } s \leq t.$$

• This leads to simple moment conditions after first differencing:

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

Therefore, at time t, the available instruments in the FD equation are in the vector $\mathbf{x}_{it}^o \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it})$. The matrix of instruments is

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

which has $T-1$ rows. Routine to apply GMM estimation.

• Simpler strategy (at least to see what can be expected from GMM):

1. 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, $\widehat{\Delta\mathbf{x}}_{it}$. This is a good time to make sure $\Delta\mathbf{x}_{it}$ is sufficiently predictable by $\mathbf{x}_{i,t-1}^o$.

• Estimate the FD equation

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

by pooled IV using instruments (not regressors) $\widehat{\Delta\mathbf{x}}_{it}$.

• Any approach to estimating the FD equation by IV can suffer from a weak instrument problem when $\Delta\mathbf{x}_{it}$ has little correlation with $\mathbf{x}_{i,t-1}^o$. In particular, if

$$\mathbf{x}_{it} = \boldsymbol{\omega}_t + \mathbf{x}_{i,t-1} + \mathbf{r}_{it}$$
$$E(\mathbf{r}_{it}|\mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, \ldots, \mathbf{x}_{i0}) = \mathbf{0}$$

then $E(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1}^o) = E(\Delta\mathbf{x}_{it}) = \boldsymbol{\omega}_t$, and IV fails when a full set of year intercepts is included in the equation.

• More moment restrictions are obtained by assuming dynamic completeness in the mean:

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

• Rules out serial correlation in $\{u_{it}\}$, which is often too restrictive when $\mathbf{x}_{it}$ does not included $y_{i,t-1}$.

• Under the dynamic completeness assumption, many more moment conditions are available. Using linear functions of the data, for $t = 3, \ldots, T$,

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

• Drawbacks: (i) We often do not want to assume dynamic completeness; (ii) Extra moment conditions are nonlinear in parameters.

• Arellano and Bover (1995) suggested instead the restrictions

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

• Suppose $\{\mathbf{x}_{it}\}$ follows

$$\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$$

for a weakly dependent process, $\{\mathbf{r}_{it}\}$. Then $\Delta\mathbf{x}_{it} = \Delta\mathbf{r}_{it}$, and so

$$Cov(\mathbf{r}_{it}, c_i) = \mathbf{0}, t = 1, 2, \ldots$$

is sufficient.

• If $\{\mathbf{x}_{it}\}$ is a random walk with drift, say $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$ then

$$\Delta\mathbf{x}_{it} = \mathbf{h}_i + \Delta\mathbf{r}_{it}$$

and so now we would have to assume

$$Cov(\mathbf{h}_i, c_i) = \mathbf{0},$$

which is less appealing.

• How can the new moment conditions be used? Let $\alpha = E(c_i)$. Then sequential exogeniety plus Arellano-Bover conditions gives

$$E\{\Delta \mathbf{x}_{it}'[(c_i - \alpha) + u_{it}]\} = \mathbf{0}, t = 2, \ldots, T.$$

• Replacing $u_{it} = y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - c_i$ shows that $v_{it} \equiv (c_i - \alpha) + u_{it} = y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta}$, and so we have moment conditions, written in terms of the parameters, in the levels equation $y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}$:

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

• We can use these new moment conditions along with the moment conditions in the FD equation. All moment conditions are linear in the parameters $(\alpha, \boldsymbol{\beta}')'$.

• Because we are mixing conditions in FD and levels, if $\mathbf{x}_{it}$ includes year effects (it should) then these must be differenced in $\Delta \mathbf{x}_{it}$.

$$\begin{pmatrix} E[\mathbf{x}_{i1}'(\Delta y_{i2} - \Delta \mathbf{x}_{i2}\boldsymbol{\beta})] \\ \vdots \\ E[\mathbf{x}_{i,T-1}^{o\prime}(\Delta y_{iT} - \Delta \mathbf{x}_{iT}\boldsymbol{\beta})] \\ E[\Delta \mathbf{x}_{i2}'(y_{i2} - \alpha - \mathbf{x}_{i2}\boldsymbol{\beta})] \\ \vdots \\ E[\Delta \mathbf{x}_{iT}'(y_{iT} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})] \end{pmatrix} = \mathbf{0}.$$

• Use GMM with a general weighting matrix to allow arbitrary correlation across all time periods/equations. (Now called "system GMM.")

• Simple AR(1) model:

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

• Typically, the minimal assumptions imposed are

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

so for $t = 2, \ldots, T$,

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

• Again, can suffer from weak instruments when ρ is close to unity.

Blundell and Bond (1998) showed that if the condition

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

is added to $E(u_{it}|y_{i,t-1}, \ldots, y_{i0}, c_i) = 0$ then

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

which can be added to the usual moment conditions.

• Can be intepreted as a restriction on the initial condition, y_{i0}. Write y_{i0} as a deviation from its steady state, $c_i/(1-\rho)$ (obtained for $|\rho| < 1$ by recursive subsitution and then taking the limit), as $y_{i0} = c_i/(1-\rho) + r_{i0}$. Then the extra condition is

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

The deviation of y_{i0} from its steady state is uncorrelated with the SS.

• Potential problem: As ρ approaches one, how realistic is it to assume there is a steady state?

• Extensions of the AR(1) model, such as

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

and use FD:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \mathbf{z}_{it}\boldsymbol{\gamma} + \Delta u_{it}, \quad t = 2, \ldots, T.$$

• Can use $\Delta \mathbf{z}_{it}$ as own IVs if they are strictly exogenous, $y_{i,t-h}$, $h \geq 2$, and can still add moment conditions in levels.

• If $\{\mathbf{z}_{it}\}$ is not strictly exogenous, can use $\{\mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i1}\}$ as IVs, along with $\{y_{i,t-2}, \ldots, y_{i0}\}$ in the FD equation at time t.

• And, we still might use, for $t = 2, \ldots, T$, moment conditions on the levels:

$$E[\Delta y_{i,t-1}(y_{it} - \alpha - \rho y_{i,t-1} - \mathbf{z}_{it}\boldsymbol{\gamma})] = 0$$
$$E[\Delta \mathbf{z}_{it}'(y_{it} - \alpha - \rho y_{i,t-1} - \mathbf{z}_{it}\boldsymbol{\gamma})] = \mathbf{0}$$

• As usual, time dummies act as their own IVs.

• AIRFAIRE.DTA Example: Dynamic model with a route-specific effect:

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

Dep. Var.	*lfare*			
	(1)	(2)	(3)	(4)
Expl. Var.	Pooled OLS (FD)	FE	Pooled IV (FD)	Arellano-Bond
$lfare_{-1}$	−.126 (.027)	.077 (.032)	.219 (.062)	.333 (.055)
concen	.076 (.053)	.058 (.053)	.126 (.056)	.152 (.040)
N	1,149	1,149	1,149	1,149


```
. gen dlfare = d.lfare
(1149 missing values generated)

. gen dlfare_1 = l.dlfare
(2298 missing values generated)

. gen dconcen = d.concen
(1149 missing values generated)
```



```
. reg dlfare dlfare_1 dconcen y99 y00, cluster(id)

Linear regression                               Number of obs =    2298
                                                F(  3,  1148) =   36.38
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0651
                                                Root MSE      =   .1168

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      dlfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlfare_1 |  -.1264673   .0267104    -4.73   0.000    -.1788739   -.0740606
     dconcen |   .0762671   .0527226     1.45   0.148    -.0271763    .1797106
         y99 |  -.0473536   .0050308    -9.41   0.000    -.0572241    -.037483
         y00 |  (dropped)
       _cons |   .0624434   .0032977    18.94   0.000     .0559732    .0689136
------------------------------------------------------------------------------

. * Pooled OLS on the differenced equation is very misleading. FE on the levels
. * does better, but is still downward biased.
```



```
. xtreg lfare lfare_1 concen y99 y00, fe cluster(id)

Fixed-effects (within) regression               Number of obs      =      3447
Group variable: id                              Number of groups   =      1149

R-sq:  within  = 0.1605                         Obs per group: min =         3
       between = 0.8863                                        avg =       3.0
       overall = 0.5014                                        max =         3

                                                F(4,1148)          =     98.81
corr(u_i, Xb)  = 0.6597                         Prob > F           =    0.0000

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lfare_1 |   .0773594   .0318913     2.43   0.015     .0147877    .1399311
      concen |   .0579086   .0533893     1.08   0.278    -.0468428    .1626601
         y99 |   .0098236   .0037176     2.64   0.008     .0025296    .0171177
         y00 |   .0700164   .0043967    15.92   0.000       .06139    .0786428
       _cons |   4.653928   .1628858    28.57   0.000     4.334341    4.973515
-------------+----------------------------------------------------------------
     sigma_u |  .38891209
     sigma_e |  .09055856
         rho |  .94856899   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. * Try FDIV, generating instruments using first-stage regressions.

. gen lfare_2 = l2.lfare
(2298 missing values generated)

. gen lfare_3 = l3.lfare
(3447 missing values generated)

. reg dlfare_1 lfare_2 dconcen if y99

      Source |       SS       df       MS              Number of obs =    1149
-------------+------------------------------           F(  2,  1146) =  110.84
       Model |  3.63569369     2  1.81784684           Prob > F      =  0.0000
    Residual |  18.7948202  1146  .016400367           R-squared     =  0.1621
-------------+------------------------------           Adj R-squared =  0.1606
       Total |  22.4305139  1148  .019538775           Root MSE      =  .12806

------------------------------------------------------------------------------
    dlfare_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lfare_2 |  -.1221207   .0082417   -14.82   0.000    -.1382913   -.1059502
     dconcen |  -.1754244   .0544243    -3.22   0.001    -.2822069    -.068642
       _cons |   .6389637   .0417491    15.30   0.000     .5570504    .7208769
------------------------------------------------------------------------------

. predict dlfare_1h99
(option xb assumed; fitted values)
(2298 missing values generated)
```



```
. reg dlfare_1 lfare_2 lfare_3 dconcen if y00

      Source |       SS       df       MS              Number of obs =    1149
-------------+------------------------------           F(  3,  1145) =   11.93
       Model |  .524236952     3  .174745651           Prob > F      =  0.0000
    Residual |  16.7684066  1145  .014644897           R-squared     =  0.0303
-------------+------------------------------           Adj R-squared =  0.0278
       Total |  17.2926436  1148  .015063278           Root MSE      =  .12102

------------------------------------------------------------------------------
    dlfare_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lfare_2 |  -.1027683   .0278186    -3.69   0.000    -.1573495   -.0481871
     lfare_3 |   .0744738    .025707     2.90   0.004     .0240356     .124912
     dconcen |  -.1971475   .0483136    -4.08   0.000    -.2919407   -.1023543
       _cons |    .155675   .0429415     3.63   0.000     .0714222    .2399278
------------------------------------------------------------------------------

. * No evidence of weak instruments in either time period.

. predict dlfare_1h00
(option xb assumed; fitted values)
(3447 missing values generated)

. gen dlfare_1h = dlfare_1h99 if y99
(3447 missing values generated)

. replace dlfare_1h = dlfare_1h00 if y00
(1149 real changes made)
```



```
. ivreg dlfare dconcen y00 (dlfare_1 = dlfare_1h), cluster(id)

Instrumental variables (2SLS) regression              Number of obs =    2298
                                                      F(  3,  1148) =   24.03
                                                      Prob > F      =  0.0000
                                                      R-squared     =       .
                                                      Root MSE      =  .12529

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      dlfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlfare_1 |   .2190128   .0619844     3.53   0.000     .0973973    .3406283
     dconcen |   .1262854    .056415     2.24   0.025     .0155974    .2369735
         y00 |    .051385    .006324     8.13   0.000     .0389771    .0637929
       _cons |   .0075111   .0042639     1.76   0.078    -.0008549    .0158771
------------------------------------------------------------------------------
Instrumented:  dlfare_1
Instruments:   dconcen y00 dlfare_1h
------------------------------------------------------------------------------

. * With FDIV, both the lag and the concen variable are positive and
. * statistically significant.
```



```
. * Now use the Arellano and Bond GMM approach.

. xtabond lfare concen y99 y00

Arellano-Bond dynamic panel-data estimation  Number of obs         =      2298
Group variable: id                           Number of groups      =      1149
Time variable: year
                                             Obs per group:    min =         2
                                                               avg =         2
                                                               max =         2

Number of instruments =      7               Wald chi2(4)          =    441.62
                                             Prob > chi2           =    0.0000
One-step results
------------------------------------------------------------------------------
       lfare |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |
         L1. |   .3326355   .0548124     6.07   0.000     .2252051    .4400659
      concen |   .1519406   .0399507     3.80   0.000     .0736386    .2302425
         y99 |   .0051715   .0041216     1.25   0.210    -.0029066    .0132496
         y00 |   .0629313   .0043475    14.48   0.000     .0544103    .0714523
       _cons |   3.304619   .2820506    11.72   0.000      2.75181    3.857428
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).lfare
        Standard: D.concen D.y99 D.y00
Instruments for level equation
        Standard: _cons
```

AEA Lectures **Chicago**, **January 2012**

Lecture 4, **Monday**, **January 9**, **1**:**00 pm-2**:**30 pm**

Cluster Sampling and Difference-in-Differences

These notes consider estimation and inference with cluster samples and samples obtained by stratifying the population. The main focus is on true cluster samples, although the case of applying cluster-sample methods to panel data is treated, including recent work where the sizes of the cross section and time series are similar. Wooldridge (2003, extended version 2006) contains a survey, but more recent work is discussed here.

1. The Linear Model with Cluster Effects

This section considers linear models estimated using cluster samples (of which a panel data set is a special case). For each group or cluster g, let $\{(y_{gm}, x_g, z_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster g, y_{gm} is a scalar response, x_g is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and z_{gm} is a $1 \times L$ vector of covariates that vary within (as well as across) groups.

1.1 Specification of the Model

The linear model with an additive error is

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \quad (1.1)$$

Our approach to estimation and inference in equation (1.1) depends on several factors, including whether we are interested in the effects of aggregate variables (β) or individual-specific variables (γ). Plus, we need to make assumptions about the error terms. In the context of pure cluster sampling, an important issue is whether the v_{gm} contain a common

group effect that can be separated in an additive fashion, as in

$$v_{gm} = c_g + u_{gm}, m = 1, \ldots, M_g, \tag{1.2}$$

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error. (In the statistics literature, (1.1) and (1.2) are referred to as a "hierarchical linear model.") One important issue is whether the explanatory variables in (1.1) can be taken to be appropriately exogenous. Under (1.2), exogeneity issues are usefully broken down by separately considering c_g and u_{gm}.

Throughout we assume that the sampling scheme generates observations that are independent across g. This assumption can be restrictive, particularly when the clusters are large geographical units. We do not consider problems of "spatial correlation" across clusters, although, as we will see, fixed effects estimators have advantages in such settings.

We treat two kinds of sampling schemes. The simplest case also allows the most flexibility for robust inference: from a large population of relatively small clusters, we draw a large number of clusters (G), where cluster g has M_g members. This setup is appropriate, for example, in randomly sampling a large number of families, classrooms, or firms from a large population. The key feature is that the number of groups is large enough relative to the group sizes so that we can allow essentially unrestricted within-cluster correlation. Randomly sampling a large number of clusters also applies to many panel data sets, where the cross-sectional population size is large (say, individuals, firms, even cities or counties) and the number of time periods is relatively small. In the panel data setting, G is the number of cross-sectional units and M_g is the number of time periods for unit g.

A different sampling scheme results in data sets that also can be arranged by group, but is better interpreted in the context of sampling from different populations are different strata within a population. We stratify the population into into $G \geq 2$ nonoverlapping groups. Then,

we obtain a random sample of size M_g from each group. Ideally, the group sizes are large in the population, hopefully resulting in large M_g. This is the perspective for the "small G" case in Section 1.3.

1.2. Large Group Asymptotics

In this section I review methods and estimators justified when the asymptotic approximations theory is with The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed; see, for example, White (1984), Arellano (1987), and Wooldridge (2010, Chapters 10, 11). Here, the emphasis is on how one might wish to use methods robust to cluster sampling even when it is not so obvious.

First suppose that the covariates satisfy

$$E(v_{gm}|x_g, z_{gm}) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.3}$$

For consistency, we can, of course, get by with zero correlation assumptions, but we use (1.3) for convenience because it meshes well with assumptions concerning conditional second moments. Importantly, the exogeneity in (1.3) only requires that z_{gm} and v_{gm} are uncorrelated. In particular, it does not specify assumptions concerning v_{gm} and z_{gp} for $m \neq p$. As we saw in the linear panel data notes, (1.3) is called the "contemporaneous exogeneity" assumption when m represents time. Allowing for correlation between v_{gm} and $z_{gp}, m \neq p$ is useful for some panel data applications and possibly even cluster samples (if the covariates of one unit can affect another unit's response). Under (1.3) and a standard rank condition on the covariates, the pooled OLS estimator, where we regress y_{gm} on $1, x_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G,$ is consistent for $\lambda \equiv (\alpha, \beta', \gamma')'$ (as $G \to \infty$ with M_g fixed) and $\sqrt{G}$-asymptotically normal.

Without more assumptions, a robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|x_g, z_{gm})$, or both. When v_{gm} has the form in

(1.2), the amount of within-cluster correlation can be substantial, which means the usual OLS standard errors can be very misleading (and, in most cases, systematically too small). Write W_g as the $M_g \times (1 + K + L)$ matrix of all regressors for group g. Then the $(1 + K + L) \times (1 + K + L)$ variance matrix estimator is

$$\widehat{Avar}(\hat{\lambda}_{POLS}) = \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \left(\sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g\right) \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \tag{1.4}$$

where $\hat{v}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group g. This asymptotic variance is now computed routinely using "cluster" options.

Pooled OLS estimation of the parameters in (1.1) ignores the within-cluster correlation of the v_{gm}; even if the procedure is consistent (again, with $G \to \infty$ and the M_g fixed), the POLS estimators can be very inefficient. If we strengthen the exogeneity assumption to

$$E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.5}$$

where Z_g is the $M_g \times L$ matrix of unit-specific covariates, then we can exploit the presence of c_g in (1.2) in a generalized least squares (GLS) analysis. With true cluster samples, (1.5) rules out the covariates from one member of the cluster affecting the outcomes on another, holding own covariates fixed. In the panel data case, (1.5) is the strict exogeneity assumption on $\{z_{gm} : m = 1, \ldots, M_g\}$ that we discussed in the linear panel data notes The standard random effects approach makes enough assumptions so that the $M_g \times M_g$ variance-covariance matrix of $v_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the so-called "random effects" form,

$$Var(v_g) = \sigma_c^2 j_{M_g}' j_{M_g} + \sigma_u^2 I_{M_g}, \tag{1.6}$$

where j_{M_g} is the $M_g \times 1$ vector of ones and I_{M_g} is the $M_g \times M_g$ identity matrix. In the standard setup, we also make the "system homoskedasticity" assumption,

$$Var(v_g|x_g, Z_g) = Var(v_g). \tag{1.7}$$

It is important to understand the role of assumption (1,7): it implies that the conditional variance-covariance matrix is the same as the unconditional variance-covariance matrix, but it does not restrict $Var(v_g)$; it can be any $M_g \times M_g$ matrix under (1.7). The particular random effects structure on $Var(v_g)$ is given by (1.6). Under (1.6) and (1.7), the resulting GLS estimator is the well-known random effects (RE) estimator.

The random effects estimator $\hat{\lambda}_{RE}$ is asymptotically more efficient than pooled OLS under (1.5), (1.6), and (1.7) as $G \to \infty$ with the M_g fixed. The RE estimates and test statistics are computed routinely by popular software packages. Nevertheless, an important point is often overlooked in applications of RE: one can, and in many cases should, make inference completely robust to an unknown form of $Var(v_g|x_g, Z_g)$.

The idea in obtaining a fully robust variance matrix of RE is straightforward and we essentially discussed it in the notes on nonlinear panel data models. Even if $Var(v_g|x_g, Z_g)$ does not have the RE form, the RE estimator is still consistent and $\sqrt{G}$-asymptotically normal under (1.5), and it is likely to be more efficient than pooled OLS. Yet we should recognize that the RE second moment assumptions can be violated without causing inconsistency in the RE estimator. For panel data applications, making inference robust to serial correlation in the idiosyncratic errors, especially with more than a few time periods, can be very important. Further, within-group correlation in the idiosyncratic errors can arise for cluster samples, too, especially if underlying (1.1) is a random coefficient model,

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma_g + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.8}$$

By estimating a standard random effects model that assumes common slopes γ, we effectively

include $z_{gm}(\gamma_g - \gamma)$ in the idiosyncratic error; this generally creates within-group correlation because $z_{gm}(\gamma_g - \gamma)$ and $z_{gp}(\gamma_g - \gamma)$ will be correlated for $m \neq p$, conditional on Z_g. Also, the idiosyncratic error will have heteroskedasticity that is a function of z_{gm}. Nevertheless, if we assume $E(\gamma_g|X_g, Z_g) = E(\gamma_g) \equiv \gamma$ along with (1.5), the random effects estimator still consistently estimates the average slopes, γ. Therefore, in applying random effects to panel data or cluster samples, it is sensible (with large G) to make the variance estimator of random effects robust to arbitrary heteroskedasticity and within-group correlation.

One way to see what the robust variance matrix looks like for $\hat{\lambda}_{RE}$ is to use the pooled OLS characterization of RE on a transformed set of data. For each g, define $\hat{\theta}_g = 1 - \{1/[1 + M_g(\hat{\sigma}_c^2/\hat{\sigma}_u^2)]\}^{1/2}$, where $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ are estimators of the variances of c_g and u_{gm}, respectively. Then the RE estimator is identical to the pooled OLS estimator of

$$y_{gm} - \hat{\theta}_g \bar{y}_g \text{ on } (1 - \hat{\theta}_g), (1 - \hat{\theta}_g)x_g, z_{gm} - \hat{\theta}_g \bar{z}_g, m = 1, \ldots, M_g; g = 1, \ldots, G; \tag{1.9}$$

see, for example, Hsiao (2003). For fully robust inference, we can just apply the fully robust variance matrix estimator in (1.4) but on the transformed data.

With panel data, it may make sense to estimate an unrestricted version of $Var(v_g)$, especially if G is large. Even in that case, it makes sense to obtain a variance matrix robust to $Var(v_{gm}|x_g, Z_g) \neq Var(v_g)$, as the GEE literature does. One can also specify a particular structure, such as an AR(1) model for the idiosyncratic errors. In any case, fully robust inference is still a good idea.

In summary, with large G and relatively small M_g, it makes sense to compute fully robust variance estimators even if we apply a GLS procedure that allows $Var(v_g)$ to be unrestricted. Nothing ever guarantees $Var(v_{gm}|x_g, Z_g) = Var(v_g)$. Because RE imposes a specific structure

on $Var(v_g)$, there is a strong case for making RE inference fully robust. When c_g is in the error term, it is even more critical to use robust inference when using pooled OLS because the usual standard errors ignore within-cluster correlation entirely.

If we are only interested in estimating γ, the "fixed effects" (FE) or "within" estimator is attractive. The within transformation subtracts off group averages from the dependent variable and explanatory variables:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.10}$$

and this equation is estimated by pooled OLS. (Of course, the x_g get swept away by the within-group demeaning.) Under a full set of "fixed effects" assumptions – which, unlike pooled OLS and random effects, allows arbitrary correlation between c_g and the z_{gm} – inference is straightforward using standard software. Nevertheless, analogous to the random effects case, it is often important to allow $Var(u_g|Z_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. For panel data, the idiosyncratic errors can always have serial correlation or heteroskedasticity, and it is easy to guard against these problems in inference. Reasons for wanting a fully robust variance matrix estimator for FE applied to cluster samples are similar to the RE case. For example, if we start with the model (1.8) then $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term. As we discussed in the linear panel data notes, the FE estimator is still consistent if $E(\gamma_g|z_{g1} - \bar{z}_g, \ldots, z_{g,M_g} - \bar{z}_g) = E(\gamma_g) = \gamma$, an assumption that allows γ_g to be correlated with $\bar{z}_g$. Nevertheless, u_{gm}, u_{gp} will be correlated for $m \neq p$. A fully robust variance matrix estimator is

$$\widehat{Avar}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1} \left(\sum_{g=1}^{G} \ddot{Z}_g' \hat{u}_g \hat{u}_g' \ddot{Z}_g\right) \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1}, \tag{1.11}$$

where $\ddot{Z}_g$ is the matrix of within-group deviations from means and $\hat{u}_g$ is the $M_g \times 1$ vector of fixed effects residuals. This estimator is justified with large-G asymptotics.

One benefit of a fixed effects approach, especially in the standard model with constant slopes but c_g in the composite error term, is that no adjustments are necessary if the c_g are correlated across groups. When the groups represent different geographical units, we might expect correlation across groups close to each other. If we think such correlation is largely captured through the unobserved effect c_g, then its elimination via the within transformation effectively solves the problem. If we use pooled OLS or a random effects approach, we would have to deal with spatial correlation across g, in addition to within-group correlation, and this is a difficult problem.

The previous discussion extends immediately to instrumental variables versions of all estimators. With large G, one can afford to make pooled two stage least squares (2SLS), random effects 2SLS, and fixed effects 2SLS robust to arbitrary within-cluster correlation and heteroskedasticity. Also, more efficient estimation is possible by applying generalized method of moments (GMM); again, GMM is justified with large G.

1.3. Should we Use the "Large" G Formulas with "Large" M_g?

Until recently, the standard errors and test statistics obtained from pooled OLS, random effects, and fixed effects were known to be valid only as $G \to \infty$ with each M_g fixed. As a practical matter, that means one should have lots of small groups. Consider again formula (1.4), for pooled OLS, when the cluster effect, c_g, is left in the error term. With a large number of groups and small group sizes, we can get good estimates of the within-cluster correlations – technically, of the cluster correlations of the cross products of the regressors and errors – even if they are unrestricted, and that is why the robust variance matrix is consistent as $G \to \infty$ with

M_g fixed. In fact, in this scenario, one loses nothing in terms of asymptotic local power (with local alternatives shrinking to zero at the rate $G^{-1/2}$) if c_g is not present. In other words, based on first-order asymptotic analysis, there is no cost to being fully robust to any kind of within-group correlation or heteroskedasticity. These arguments apply equally to panel data sets with a large number of cross sections and relatively few time periods, whether or not the idiosyncratic errors are serially correlated.

What if one applies robust inference in scenarios where the fixed M_g, $G \to \infty$ asymptotic analysis not realistic? Hansen (2007) has recently derived properties of the cluster-robust variance matrix and related test statistics under various scenarios that help us more fully understand the properties of cluster robust inference across different data configurations. First consider how his results apply to true cluster samples. Hansen (2007, Theorem 2) shows that, with G and M_g both getting large, the usual inference based on (1.4) is valid with arbitrary correlation among the errors, v_{gm}, within each group. Because we usually think of v_{gm} as including the group effect c_g, this means that, with large group sizes, we can obtain valid inference using the cluster-robust variance matrix, provided that G is also large. So, for example, if we have a sample of $G = 100$ schools and roughly $M_g = 100$ students per school, and we use pooled OLS leaving the school effects in the error term, we should expect the inference to have roughly the correct size. Probably we leave the school effects in the error term because we are interested in a school-specific explanatory variable, perhaps indicating a policy change.

Unfortunately, pooled OLS with cluster effects when G is small and group sizes are large fall outside Hansen's theoretical findings: the proper asymptotic analysis would be with G fixed, $M_g \to \infty$, and persistent within-cluster correlation (because of the presence of c_g in the

error). Hansen (2007, Theorem 4) is aimed at panel data where the time series dependence satisfies strong mixing assumptions, that is, where the correlation within each group g is weakly dependent. Even in this case, the variance matrix in (1.4) must be multiplied by $G/(G-1)$ and inference based on the t_{G-1} distribution. (Conveniently, this adjustment is standard in Stata's calculation of cluster-robust variance matrices.) Interestingly, Hansen finds, in simulations, that with $G = 10$ and $M_g = 50$ for all g, using the adjusted robust variance matrix estimator with critical values from the t_{G-1} distribution produces fairly small size distortions. But the simulation study is special (one covariate whose variance is as large as the variance of the composite error).

We probably should not expect good properties of the cluster-robust inference with small groups and very large group sizes when cluster effects are left in the error term. As an example, suppose that $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest is exogenous and varies only at the hospital level, it is tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and reasons to expect it will not work well. In the next section we discuss other approaches available with small G and large M_g.

If the explanatory variables of interest vary within group, FE is attractive for a couple of reasons. The first advantage is the usal one about allowing c_g to be arbitrarily correlated with the z_{gm}. The second advantage is that, with large M_g, we can treat the c_g as parameters to estimate – because we can estimate them precisely – and then assume that the observations are independent across m (as well as g). This means that the usual inference is valid, perhaps with adjustment for heteroskedasticity. Interestingly, the fixed G, large M_g asymptotic results in Theorem 4 of Hansen (2007) for cluster-robust inference apply in this case. But using

cluster-robust inference is likely to be very costly in this situation: the cluster-robust variance matrix actually converges to a random variable, and t statistics based on the adjusted version of (1.11) – that is, multiplied by $G/(G-1)$ – have an asymptotic t_{G-1} distribution. Therefore, while the usual or heteroskedasticity-robust inference can be based on the standard normal distribution, the cluster-robust inference is based on the t_{G-1} distribution (and the cluster-robust standard errors may be larger than the usual standard errors). With small G, inference based on cluster-robust statistics could be very conservative when it need not be. (Also, Hansen's Theorem 4 is not completely general, and may not apply with heterogeneous sampling across groups.)

In summary, for true cluster sample applications, cluster-robust inference using pooled OLS delivers statistics with proper size when G and M_g are both moderately large, but they should probably be avoided with large M_g and small G. When cluster fixed effects are included, the usual inference is often valid, perhaps made robust to heteroskedasticity, and is likely to be much more powerful than cluster-robust inference.

For panel data applications, Hansen's (2007) results, particularly Theorem 3, imply that cluster-robust inference for the fixed effects estimator should work well when the cross section (N) and time series (T) dimensions are similar and not too small. If full time effects are allowed in addition to unit-specific fixed effects – as they often should – then the asymptotics must be with N and T both getting large. In this case, any serial dependence in the idiosyncratic errors is assumed to be weakly dependent. The similulations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007) verify that the fully robust cluster-robust variance matrix works well.

There is some scope for applying the fully robust variance matrix estimator when N is

small relative to T when unit-specific fixed effects are included. Unlike in the true cluster sampling case, it makes sense to treat the idiosyncratic errors as correlated with only weakly dependent. But Hansen's (2007, Theorem 4) does not allow time fixed effects (because the asymptotics is with fixed N and $T \to \infty$, and so the inclusion of time fixed effects means adding more and more parameters without more cross section data to estimate them). As a practical matter, it seems dangerous to rely on omitting time effects or unit effects with panel data. Hansen's result that applies in this case requires N and T both getting large.

2. Estimation with a Small Number of Groups and Large Group Sizes

We can summarize the findings of the previous section as follows: fully robust inference justified for large G (N) and small M_g (T) can also be relied on when M_g (T) is also large, provided G (N) is also reasonably large. However, whether or not we leave cluster (unobserved) effects in the error term, there are good reasons not to rely on cluster-robust inference when G (N) is small and M_g (T) is large.

In this section, we describe approaches to inference when G is small and the M_g are large. These results apply primarily to the true cluster sample case, although we will draw on them for difference-in-differences frameworks using pooled cross sections in a later set of notes.

In the large G, small M_g case, it often makes sense to think of sampling a large number of groups from a large population of clusters, where each cluster is relatively small. When G is small while each M_g is large, this thought experiment needs to be modified. For most data sets with small G, a stratified sampling scheme makes more sense: we have defined G groups in the population, and we obtain our data by randomly sampling from each group. As before, M_g is the sample size for group g. Except for the relative dimensions of G and M_g, the resulting data

set is essentially indistinguishable from that described in Section 1.2.

The problem of proper inference when M_g is large relative to G was brought to light by Moulton (1990), and has been recently studied by Donald and Lang (2007). DL focus on applications that seem well described by the stratified sampling scheme, but their approach seems to imply a different sampling experiment. In particular, they treat the parameters associated with the different groups as outcomes of random draws. One way to think about the sampling in this case is that a small number of groups is drawn from a (large) population of potential groups; therefore, the parameters common to all members of the group can be viewed as random. Given the groups, samples are then obtained via random sampling within each group.

To illustrate the issues considered by Donald and Lang, consider the simplest case, with a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \tag{2.1}$$

$$= \delta_g + \beta x_g + u_{gm}, \; m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{2.2}$$

Notice how (2.2) is written as a model with common slope, β, but intercept, δ_g, that varies across g. Donald and Lang focus on (2.1), where c_g is assumed to be independent of x_g with zero mean. They use this formulation to highlight the problems of applying standard inference to (2.1), leaving c_g as part of the composite error term, $v_{gm} = c_g + u_{gm}$. We know this is a bad idea even in the large G, small M_g case, as it ignores the persistent correlation in the errors within each group. Further, from the discussion of Hansen's (2007) results, using cluster-robust inference when G is small is likely to produce poor inference.

One way to see the problem with the usual inference in applying standard inference is to note that when $M_g = M$ for all $g = 1, \ldots, G$, the pooled OLS estimator, $\hat{\beta}$, is identical to the

"between" estimator obtained from the regression

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.3}$$

Conditional on the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g : g = 1, \ldots, G\}$, the within-group averages of the composite errors $v_{gm} \equiv c_g + u_{gm}$. The presence of c_g means new observations within group do not provide additional information for estimating β beyond how they affect the group average, $\bar{y}_g$. In effect, we only have G useful pieces of information.

If we add some strong assumptions, there is a solution to the inference problem. In addition to assuming $M_g = M$ for all g, assume $c_g | x_g \sim \text{Normal}(0, \sigma_c^2)$ and assume $u_{gm} | x_g, c_g \sim \text{Normal}(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim \text{Normal}(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G \tag{2.4}$$

satisfies the classical linear model assumptions. Therefore, we can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$. When G is very small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent then if we use the $t_{M_1+M_2+\ldots+M_G-2}$ distribution – which is what we would be doing if we use the usual pooled OLS statistics. (Interestingly, if we use cluster-robust inference and apply Hansen's results – even though they do not apply – we would use the t_{G-1} distribution.)

When x_g is a $1 \times K$ vector, we need $G > K + 1$ to use the t_{G-K-1} distribution for inference. [In Moulton (1990), $G = 50$ states and x_g contains 17 elements]

As pointed out by DL, performing the correct inference in the presence of c_g is *not* just a matter of correcting the pooled OLS standard errors for cluster correlation – something that does not appear to be valid for small G, anyway – or using the RE estimator. In the case of

common group sizes, there is only estimator: pooled OLS, random effects, and the between regression in (2.4) all lead to the *same* $\hat{\beta}$. The regression in (2.4), by using the t_{G-K-1} distribution, yields inference with appropriate size.

We can apply the DL method without normality of the u_{gm} if the common group size M is large: by the central limit theorem, $\bar{u}_g$ will be approximately normally distributed very generally. Then, because c_g is normally distributed, we can treat $\bar{v}_g$ as approximately normal with constant variance. Further, even if the group sizes differ across g, for very large group sizes $\bar{u}_g$ will be a negligible part of $\bar{v}_g$: $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$. Provided c_g is normally distributed and it dominates $\bar{v}_g$, a classical linear model analysis on (2.4) should be roughly valid.

The broadest applicability of DL's setup is when the average of the idiosyncratic errors, $\bar{u}_g$, can be ignored – either because σ_u^2 is small relative to σ_c^2, M_g is large, or both. In fact, applying DL with different group sizes or nonnormality of the u_{gm} is identical to ignoring the estimation error in the sample averages, $\bar{y}_g$. In other words, it is as if we are analyzing the simple regression $\mu_g = \alpha + \beta x_g + c_g$ using the classical linear model assumptions (where $\bar{y}_g$ is used in place of the unknown group mean, μ_g). With small G, we need to further assume c_g is normally distributed.

If z_{gm} appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + x_g\beta + \bar{z}_g\gamma + \bar{v}_g, g = 1, \ldots, G, \tag{2.5}$$

provided $G > K + L + 1$. If c_g is independent of $(x_g, \bar{z}_g)$ with a homoskedastic normal distribution and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution.

The DL solution to the inference problem with small G is pretty common as a strategy to check robustness of results obtained from cluster samples, but often it is implemented with somewhat large G (say, $G = 50$). Often with cluster samples one estimates the parameters using the disaggregated data and also the averaged data. When some covariates that vary within cluster, using averaged data is generally inefficient. But it does mean that standard errors need not be made robust to within-cluster correlation. We now know that if G is reasonably large and the group sizes not too large, the cluster-robust inference can be acceptable. DL point out that with small G one should think about simply using the group averages in a classical linear model analysis.

For small G and large M_g, inference obtained from analyzing (2.5) as a classical linear model will be very conservative in the absense of a cluster effect. Perhaps in some cases it is desirable to inject this kind of uncertainty, but it rules out some widely-used staples of policy analysis. For example, suppose we have two populations (maybe men and women, two different cities, or a treatment and a control group) with means $\mu_g, g = 1, 2$, and we would like to obtain a confidence interval for their difference. In almost all cases, it makes sense to view the data as being two random samples, one from each subgroup of the population. Under random sampling from each group, and assuming normality and equal population variances, the usual comparison-of-means statistic is distributed exactly as $t_{M_1+M_2-2}$ under the null hypothesis of equal population means. (Or, we can construct an exact 95% confidence interval of the difference in population means.) With even moderate sizes for M_1 and M_2, the $t_{M_1+M_2-2}$ distribution is close to the standard normal distribution. Plus, we can relax normality to obtain approximately valid inference, and it is easy to adjust the t statistic to allow for different population variances. With a controlled experiment the standard difference-in-means analysis

is often quite convincing. Yet we cannot even study this estimator in the DL setup because $G = 2$. The problem can be seen from (2.2): in effect, we have three parameters, δ_1, δ_2, and β, but only two observations.

DL criticize Card and Krueger (1994) for comparing mean wage changes of fast-food workers across two states because Card and Krueger fail to account for the state effect (New Jersery or Pennsylvania), c_g, in the composite error, v_{gm}. But the DL criticism in the $G = 2$ case is no different from a common question raised for any difference-in-differences analyses: How can we be sure that any observed difference in means is due entirely to the policy change? To characterize the problem as failing to account for an unobserved group effect is not necessarily helpful.

To further study the $G = 2$ case, recall that c_g is independent of x_g with mean zero. In other words, the expected deviation in using the simple comparison-of-means estimator is zero. In effect, it estimates

$$\mu_2 - \mu_1 = (\delta_2 + \beta) - \delta_1 = (\alpha + c_2 + \beta) - (\alpha + c_1) = \beta + (c_2 - c_1). \tag{2.6}$$

Under the DL assumptions, $c_2 - c_1$ has mean zero, and so estimating it should not bias the analysis. DL work under the assumption that β is the parameter of interest, but, if the experiment is properly randomized – as is maintained by DL – it is harmless to include the c_g in the estimated effect.

Consider now a case where the DL approach can be applied. Assume there are $G = 4$ groups with groups one and two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). The DL approach would involve computing the averages for each group, $\bar{y}_g$, and running the regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$. Inference is based on the t_2 distribution. The estimator $\hat{\beta}$ in this case can be written as

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{2.7}$$

(The pooled OLS regression using the disaggregated data results in the weighted average $(p_3\bar{y}_3 + p_4\bar{y}_4) - (p_1\bar{y}_1 + p_2\bar{y}_2)$, where $p_1 = M_1/(M_1 + M_2)$, $p_2 = M_2/(M_1 + M_2)$, $p_3 = M_3/(M_3 + M_4)$, and $p_4 = M_4/(M_3 + M_4)$ are the relative proportions within the control and treatment groups, respectively.) With $\hat{\beta}$ written as in (2.7), we are left to wonder why we need to use the t_2 distribution for inference. Each $\bar{y}_g$ is usually obtained from a large sample – $M_g = 30$ or so is usually sufficient for approximate normality of the standardized mean – and so $\hat{\beta}$, when properly standardized, has an approximate standard normal distribution quite generally.

In effect, the DL approach rejects the usual inference based on group means from large sample sizes because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. In other words, the control group may be heterogeneous as might be the treatment group. But this in itself does not invalidate standard inference applied to (2.7). In fact, if we *define* the object of inference as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2, \tag{2.8}$$

which is an average treatment effect of sorts, then $\hat{\beta}$ is consistent for τ and (when properly scaled) asymptotically normal as the M_g get large.

Equation (2.7) hints at a different way to view the small G, large M_g setup. In this particular application, we estimate two parameters, α and β, given four moments that we can estimate with the data. The OLS estimates from (2.4) in this case are minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. If we use the 4×4 identity matrix as the weight matrix, we get $\hat{\beta}$ as in (2.7) and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$. Using the MD approach, we see there are two overidentifying restrictions, which are easily tested. But even if

we reject them, it simply implies at least one pair of means within each of the control and treatment groups is different.

With large group sizes, and whether or not G is especially large, we can put the general problem into an MD framework, as done, for example, by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group. For each group g, write

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}, m = 1, \ldots, M_g, \tag{2.9}$$

where we assume random sampling within group and independent sampling across groups. We make the standard assumptions for OLS to be consistent (as $M_g \to \infty$) and $\sqrt{M_g}$-asymptotically normal; see, for example, Wooldridge (2010, Chapter 4). The presence of group-level variables x_g in a "structural" model can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in (2.9). In particular,

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G, \tag{2.10}$$

where we think of x_g as fixed, observed attributes of heterogeneous groups. With K attributes we must have $G \geq K + 1$ to determine α and β. If M_g is large enough to estimate the δ_g precisely, a simple two-step estimation strategy suggests itself. First, obtain the $\hat{\delta}_g$, along with $\hat{\gamma}_g$, from an OLS regression within each group. If $G = K + 1$ then, typically, we can solve for $\hat{\theta} \equiv (\hat{\alpha}, \hat{\beta}')'$ uniquely in terms of the $G \times 1$ vector $\hat{\delta}$:. $\hat{\theta} = X^{-1}\hat{\delta}$, where X is the $(K + 1) \times (K + 1)$ matrix with g^{th} row $(1, x_g)$. If $G > K + 1$ then, in a second step, we can use a minimum distance approach, as described in Wooldridge (2010, Section 14.5). If we use as the weighting matrix I_G, the $G \times G$ identity matrix, then the minimum distance estimator can be computed from the OLS regression

$$\hat{\delta}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.10}$$

Under asymptotics such that $M_g = \rho_g M$ where $0 < \rho_g \leq 1$ and $M \to \infty$, the minimum distance estimator $\hat{\theta}$ is consistent and $\sqrt{M}$-asymptotically normal. Still, this particular minimum distance estimator is asymptotically inefficient except under strong assumptions. Because the samples are assumed to be independent, it is not appreciably more difficult to obtain the efficient minimum distance (MD) estimator, also called the "minimum chi-square" estimator.

First consider the case where z_{gm} does not appear in the first stage estimation, so that the $\hat{\delta}_g$ is just $\bar{y}_g$, the sample mean for group g. Let $\hat{\sigma}_g^2$ denote the usual sample variance for group g. Because the $\bar{y}_g$ are independent across g, the efficient MD estimator uses a diagonal weighting matrix. As a computational device, the minimum chi-square estimator can be computed by using the weighted least squares (WLS) version of (2.11), where group g is weighted by $M_g/\hat{\sigma}_g^2$ (groups that have more data and smaller variance receive greater weight). Conveniently, the reported t statistics from the WLS regression are asymptotically standard normal as the group sizes M_g get large. (With fixed G, the WLS nature of the estimation is just a computational device; the standard asymptotic analysis of the WLS estimator has $G \to \infty$.). The minimum distance approach works with small G provided $G \geq K + 1$ and each M_g is large enough so that normality is a good approximation to the distribution of the (properly scaled) sample average within each group.

If z_{gm} is present in the first-stage estimation, we use as the minimum chi-square weights the inverses of the asymptotic variances for the g intercepts in the separate G regressions. With large M_g, we might make these fully robust to heteroskedasticity in $E(u_{gm}^2|z_{gm})$ using the White (1980) sandwich variance estimator. At a minimum we would want to allow different σ_g^2 even if we assume homoskedasticity within groups. Once we have the $\widehat{Avar}(\hat{\delta}_g)$ – which are just the

squared reported standard errors for the $\hat{\delta}_g$ – we use as weights $1/\widehat{Avar}(\hat{\delta}_g)$ in the computationally simple WLS procedure. We are still using independence across g in obtaining a diagonal weighting matrix in the MD estimation.

An important by-product of the WLS regression is a minimum chi-square statistic that can be used to test the $G - K - 1$ overidentifying restrictions. The statistic is easily obtained as the weighted sum of squared residuals, say SSR_w. Under the null hypothesis in (2.10), $SSR_w \overset{a}{\sim} \chi^2_{G-K-1}$ as the group sizes, M_g, get large. If we reject H_0 at a reasonably small significance level, the x_g are not sufficient for characterizing the changing intercepts across groups. If we fail to reject H_0, we can have some confidence in our specification, and perform inference using the standard normal distribution for t statistics for testing linear combinations of the population averages.

We might also be interested in how one of the slopes in γ_g depends on the group features, x_g. Then, we simple replace $\hat{\delta}_g$ with, say $\hat{\gamma}_{g1}$, the slope on the first element of z_{gm}. Naturally, we would use $1/\widehat{Avar}(\hat{\gamma}_{g1})$ as the weights in the MD estimation.

The minimum distance approach can also be applied if we impose $\gamma_g = \gamma$ for all g, as in the original model (1). Obtaining the $\hat{\delta}_g$ themselves is easy: run the pooled regression

$$y_{gm} \text{ on } d1_g, d2_g, \ldots, dG_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G \tag{2.11}$$

where $d1_g, d2_g, \ldots, dG_g$ are group dummy variables. Using the $\hat{\delta}_g$ from the pooled regression (2.12) in MD estimation is complicated by the fact that the $\hat{\delta}_g$ are no longer asymptotically independent; in fact, $\hat{\delta}_g = \bar{y}_g - \bar{z}_g\hat{\gamma}$, where $\hat{\gamma}$ is the vector of common slopes, and the presence of $\hat{\gamma}$ induces correlation among the intercept estimators. Let $\hat{V}$ be the $G \times G$ estimated (asymptotic) variance matrix of the $G \times 1$ vector $\hat{\delta}$. Then the MD estimator is

$\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{\delta}$ and its estimated asymptotic variance is $(X'\hat{V}^{-1}X)^{-1}$. If the OLS regression (2.11) is used, or the WLS version, the resulting standard errors will be incorrect because they ignore the across group correlation in the estimators. (With large group sizes the errors might be small; see the next section.)

Intermediate approaches are available, too. Loeb and Bound (1996) (LB for short) allow different group intercepts and group-specific slopes on education, but impose common slopes on demographic and family background variable. The main group-level covariate is the student-teacher ratio. Thus, LB are interested in seeing how the student-teach ratio affects the relationship between test scores and education levels. LB use both the unweighted estimator and the weighted estimator and find that the results differ in unimportant ways. Because they impose common slopes on a set of regressors, the estimated slopes on education (say $\hat{\gamma}_{g1}$) are not asymptotically independent, and perhaps using a nondiagonal estimated variance matrix $\hat{V}$ (which would be 36×36 in this case) is more appropriate; but see Section 3.

If we reject the overidentifying restrictions, we are essentially concluding that $\delta_g = \alpha + x_g\beta + c_g,$ where c_g can be interpreted as the deviation from the restrictions in (2.10) for group *g*. As *G* increases relative to *K*, the likelihood of rejecting the restrictions increases. One possibility is to apply the Donald and Lang approach, which is to analyze the OLS regression in (2.11) in the context of the classical linear model (CLM), where inference is based on the t_{G-K-1} distribution. Why is a CLM analysis justified? Since $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, we can ingore the estimation error in $\hat{\delta}_g$ for large M_g (Recall that the same "large M_g" assumption underlies the minimum distance approach.) Then, it is as if we are estimating the equation $\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G$ by OLS. If the c_g are drawn from a normal distribution, classical analysis is applicable because c_g is assumed to be independent of

x_g. This approach is desirable when one cannot, or does not want to, find group-level observables that completely determine the δ_g. It is predicated on the assumption that the other factors in c_g are not systematically related to x_g, a reasonable assumption if, say, x_g is a randomly assigned treatment at the group level, a case considered by Angrist and Lavy (2002).

Beyond the treatment effect case, the issue of how to define parameters of interest appears complicated, and deserves further study. In the example with $G = 4$ and two control and two treatment groups, it can be shown that defining the treatment effect as (2.8) is the same as defining the parameters of interest as $\theta = (X'X)^{-1}X'\delta$, where X is the 4×2 matrix

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.12}$$

and $\beta = \tau$ is the second element of θ. Generally, if it makes sense to define the object of interest as $\theta = (X'X)^{-1}X'\delta$, and if we estimate θ as $\hat{\theta} = (X'X)^{-1}X'\hat{\delta}$, then $\sqrt{M}(\hat{\theta} - \theta)$ inherits its asymptotic distribution from that of $\sqrt{M}(\hat{\delta} - \delta)$, where we assume, as before, that $M_g = \rho_g M$ with $0 < \rho_g \leq 1$ and $M \to \infty$. Such a setting implies

$$\widehat{Avar(\hat{\theta})} = (X'X)^{-1}X'[\widehat{Avar(\hat{\delta})}]X(X'X)^{-1}. \tag{2.13}$$

3. What if G and M_g are Both "Large"?

Section 1 reviewed methods appropriate for a large number of groups and relatively small group sizes. Section 2 considered two approaches appropriate for large group sizes and a small number of groups. The DL and minimum distance approaches use the large group sizes

assumption differently: in its most applicable setting, DL use the large M_g assumption to ignore the first-stage estimation error entirely, with all uncertainty coming from unobserved group effects, while the asymptotics underlying the MD approach is based on applying the central limit theorem within each group. Not surprisingly, more flexibility is afforded if G and M_g are both "large."

For example, suppose we adopt the DL specification (with an unobserved cluster effect),

$$\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G, \tag{3.1}$$

and $G = 50$ (say, states in the U.S.). Further, assume first that the group sizes are large enough (or the cluster effects are so strong) that the first-stage estimation error can be ignored. Then, it matters not whether we impose some common slopes or run separate regressions for each group (state) in the first stage estimation. In the second step, we can treat the group-specific intercepts, $\hat{\delta}_g, g = 1, \ldots, G$, as independent "observations" to be used in the second stage. This means we apply regression (2.10) and apply the usual inference procedures. The difference now is that with $G = 50$, the usual t statistics have some robustness to nonnormality of the c_g, assuming the CLT approximation works well With small G, the exact inference was based on normality of the c_g.

Loeb and Bound (1996), with $G = 38$, essentially use regression (2.10), but with estimated slopes as the dependent variable in place of estimated intercepts. As mentioned in Section 2, LB impose some common slopes across groups, which means the estimated parameters are dependent across group. The minimum distance approach without cluster effects is one way to account for the dependence. Alternatively, one can simply adopt the DL perspective and just assume the estimation error is swamped by c_g; then standard OLS analysis is approximately

justfied.

4. The Traditional Difference-in-Differences Methodology

Since the work by Ashenfelter and Card (1985), the use of difference-in-differences methods has become very widespread. The simplest set up is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. In the case where the same units within a group are observed in each time period, the average gain in the second (control) group is substracted from the average gain in the first (treatment) group. This removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of trends. We will treat the panel data case in Section 4.

With repeated cross sections, we can write the model for a generic member of any of groups as

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \tag{4.1}$$

where y is the outcome of interest, $d2$ is a dummy variable for the second time period. The dummy variable dB captures possible differences between the treatment and control groups prior to the policy change. The time period dummy, $d2$, captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest, δ_1, multiplies the interaction term, $d2 \cdot dB$, which is the same as a dummy variable equal to one

for those observations in the treatment group in the second period. The difference-in-differences estimate is

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}). \tag{4.2}$$

Inference based on even moderate sample sizes in each of the four groups is straightforward, and is easily made robust to different group/time period variances in the regression framework.

In some cases a more convincing analysis of a policy change is available by further refining the definition of treatment and control groups. For example, suppose a state implements a change in health care policy aimed at the elderly, say people 65 and older, and the response variable, *y*, is a health outcome. One possibility is to use data only on people in the state with the policy change, both before and after the change, with the control group being people under 65 (or, say, 55 to 64), and the treatment group being people 65 and older. The potential problem with this DD analysis is that other factors unrelated to the state's new policy might affect the health of the elderly relative to the younger population, for example, changes in health care emphasis at the federal level. A different DD analysis would be to use another state as the control group and use the elderly from the non-policy state as the control group. Here, the problem is that *changes* in the health of the elderly might be systematically different across states due to, say, income and wealth differences, rather than the policy change.

A more robust analysis than either of the DD analyses described above can be obtained by comparing the DD estimate for the state where the policy was implemented with the same estimate from a control state. If we again label the two time periods as one and two, let *B* represent the state implementing the policy, and let *E* denote the group of elderly, then an expanded verson of (4.1) is

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (4.3)$$

The coefficient of interest is now δ_3, the coefficient on the triple interaction term, $d2 \cdot dB \cdot dE$. The OLS estimate $\hat{\delta}_3$ can be expressed as

$$\hat{\delta}_3 = [(\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1})] - [(\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{A,N,2} - \bar{y}_{A,N,1})], \quad (4.4)$$

where the *A* subscript means the state not implementing the policy and the *N* subscript represents the non-elderly. The estimate in (4.4) is usually called the *difference-in-difference-in-differences (DDD)* estimate. The first term in [•] is the DD estimate obtained by using the non-elderly as the control group and the time periods before and after the policy change. To ensure that this DD estimate is not simply picking up different trends in health outcomes between the old and young, the DDD estimate subtracts off the same estimated difference in trends for the control state (the second term in [•]).

When implemented as a regression, a standard error for $\hat{\delta}_3$ is easily obtained, including a heteroskedasticity-robust standard error. As in the DD case, it is straightforward to add additional covariates to (4.3) and inference robust to heteroskedasticity.

5. How Should We View Uncertainty in DD Settings?

The standard approach just described assumes that all uncertainty in inference enters through sampling error in estimating the means of each group/time period combination. This approach has a long history in statistics, as it is equivalent to analysis of variance. Recently, different approaches have been suggest that focus on different kinds of uncertainty – perhaps in addition to sampling error in estimating means. Recent work by Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a,b), and Abadie, Diamond, and Hainmueller (2007) argues for additional sources of uncertainty. In fact, in most cases the additional uncertainty is assumed to swamp the sampling error in estimating group/time period means. We already discussed the DL approach in the cluster sample notes, although we did not explicitly introduce a time dimension. One way to view the uncertainty introduced in the DL framework – and a perspective explicitly taken by ADH – is that our analysis should better reflect the uncertainty in the quality of the control groups.

Before we turn to a general setting, it is useful to ask whether introducing more than sampling error into DD analyses is necessary, or desirable. As we discussed in the cluster sample notes, the DL approach does not allow inference in the basic comparison-of-mean case for two groups. While the DL estimate is the usual difference in means, the error variance of the cluster effect cannot be estimated, and the *t* distribution is degenerate. It is also the case that the DL approach cannot be applied to the standard DD or DDD cases covered in Section 1. We either have four different means to estimate or six, and the DL regression in these cases produces a perfect fit with no residual variance. Should we conclude nothing can be learned in such settings?

Consider the example from Meyer, Viscusi, and Durbin (1995) on estimating the effects of

benefit generosity on length of time a worker spends on workers' compensation. MVD have a before and after period, where the policy change was to raise the cap on covered earnings. The treatment group is high earners, and the control group is low earners – who should not have been affected by the change in the cap. Using the state of Kentucky and a total sample size of 5,626, MVD find the DD estimate of the policy change is about 19.2% (longer time on workers' compensation). The *t* statistic is about 2.76, and the estimate changes little when some controls are added. MVD also use a data set for Michigan. Using the same DD approach, they estimate an almost identical effect: 19.1%. But, with "only" 1,524 observations, the *t* statistic is 1.22. It seems that, in this example, there is plenty of uncertainty in estimation, and one cannot obtain a tight estimate without a fairly large sample size. It is unclear what we gain by concluding that, because we are just identifying the parameters, we cannot perform inference in such cases. In this example, it is hard to argue that the uncertainty associated with choosing low earners within the same state and time period as the control group somehow swamps the sampling error in the sample means.

6. General Settings for DD Analysis: Multiple Groups and Time Periods

The DD and DDD methodologies can be applied to more than two time periods. In the first case, a full set of time-period dummies is added to (4.1), and a policy dummy replaces $d2 \cdot dB$; the policy dummy is simply defined to be unity for groups and time periods subject to the policy. This imposes the restriction that the policy has the same effect in every year, and assumption that is easily relaxed. In a DDD analysis, a full set of dummies is included for each of the two kinds of groups and all time periods, as well as all pairwise interactions. Then, a

policy dummy (or sometimes a continuous policy variable) measures the effect of the policy. See Gruber (1994) for an application to mandated maternity benefits.

With many time periods and groups, a general framework considered by BDM (2004) and Hansen (2007b) is useful. The equation at the individual level is

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{6.1}$$

where i indexes individual, g indexes group, and t indexes time. This model has a full set of time effects, λ_t, a full set of group effects, α_g, group/time period covariates, x_{gt} (these are the policy variables), individual-specific covariates, $\mathbf{z}_{igt}$, unobserved group/time effects, v_{gt}, and individual-specific errors, u_{igt}. We are interested in estimating $\boldsymbol{\beta}$. Equation (6.1) is an example of a *multilevel model*.

One way to write (6.1) that is useful is

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{6.2}$$

which shows a model at the individual level where both the intercepts and slopes are allowed to differ across all (g,t) pairs. Then, we think of δ_{gt} as

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{6.3}$$

Equation (6.3) is very useful, as we can think of it as a regression model at the group/time period level.

As discussed by BDM, a common way to estimate and perform inference in (6.1) is to ignore v_{gt}, in which case the observations at the individual level are treated as independent. When v_{gt} is present, the resulting inference can be very misleading. BDM and Hansen (2007b) allow serial correlation in $\{v_{gt} : t = 1, 2, \ldots, T\}$ and assume independence across groups, g.

A simple way to proceed is to view (6.3) as ultimately of interest. We observe $\mathbf{x}_{gt}$, λ_t is

handled with year dummies,and α_g just represents group dummies. The problem, then, is that we do not observe δ_{gt}. But we can use the individual-level data to estimate the δ_{gt}, provided the group/time period sizes, M_{gt}, are reasonably large. With random sampling within each (g,t), the natural estimate of δ_{gt} is obtained from OLS on (6.2) for each (g,t) pair, assuming that $E(\mathbf{z}_{igt}'u_{igt}) = \mathbf{0}$. (In most DD applications, this assumption almost holds by definition, as the individual-specific controls are included to improve estimation of δ_{gt}.) If a particular model of heteroskedasticity suggests itself, and $E(u_{it}|\mathbf{z}_{igt}) = 0$ is assumed, then a weighted least squares procedure can be used. Sometimes one wishes to impose some homogeneity in the slopes – say, $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}_g$ or even $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}$ – in which case pooling can be used to impose such restrictions. In any case, we proceed as if the M_{gt} are large enough to ignore the estimation error in the $\hat{\delta}_{gt}$; instead, the uncertainty comes through v_{gt} in (6.3). Hansen (2007b) considers adjustments to inference that accounts for sampling error in the $\hat{\delta}_{gt}$, but the methods are more complicated. The minimum distance approach we discussed in the cluster sampling notes, applied in the current context, effectively drops v_{gt} from (6.3) and views $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta}$ as a set of deterministic restrictions to be imposed on δ_{gt}. Inference using the efficient minimum distance estimator uses only sampling variation in the $\hat{\delta}_{gt}$, which will be independent across all (g,t) if they are separately estimated, or which will be correlated if pooled methods are used.

Because we are ignoring the estimation error in $\hat{\delta}_{gt}$, we proceed simply by analyzing the panel data equation

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}, \quad t = 1,\ldots,T, g = 1,\ldots,G, \tag{6.4}$$

where we keep the error as v_{gt} because we are treating $\hat{\delta}_{gt}$ and δ_{gt} interchangeably. If we

assume that We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (6.4) by OLS – which means full year and group effects, along with x_{gt} – then the OLS estimator has satisfying properties as G and T both increase, provided $\{v_{gt} : t = 1, 2, \ldots, T\}$ is a weakly dependent (mixing) time series for all g. The simulations in BDM and Hansen (2007a) indicate that cluster-robust inference, where each cluster is a set of time periods, work reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and G is moderately large.

Hansen (2007b), noting that the OLS estimator (the fixed effects estimator) applied to (6.4) is inefficient when v_{gt} is serially uncorrelated (and possibly heteroskedastic), proposes feasible GLS. As is well known, if T is not large, estimating parameters for the variance matrix $\Omega_g = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. In other words, using the FE residuals, $\hat{v}_{gt}$, to estimate Ω_g can result in severe bias for small T. Solon (1984) highlighted this problem for the homoskedastic AR(1) model. Of course, the bias disappears as $T \to \infty$, and regression packages such as Stata, that have a built-in command to do fixed effects with AR(1) errors, use the usual AR(1) coefficient $\hat{\rho}$, obtained from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1}, \; t = 2, \ldots, T, g = 1, \ldots, G. \tag{6.5}$$

As discussed in Wooldridge (2003) and Hansen (2007b), one way to account for the bias in $\hat{\rho}$ is to still use a fully robust variance matrix estimator. But Hansen's simulations show that this approach is quite inefficient relative to his suggestion, which is to bias-adjust the estimator $\hat{\rho}$ and then use the bias-adjusted estimator in feasible GLS. (In fact, Hansen covers the general $AR(p)$ model.) Hansen derives many attractive theoretical properties of his the estimator.An iterative bias-adjusted procedure has the same asymptotic distribution as $\hat{\rho}$ in the case $\hat{\rho}$ should

work well: G and T both tending to infinity. Most importantly for the application to DD problems, the feasible GLS estimator based on the iterative procedure has the same asymptotic distribution as the GLS etsimator when $G \to \infty$ and T is fixed. When G and T are both large, there is no need to iterated to achieve efficiency.

Hansen further shows that, even when G and T are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-adusted estimates deliver higher-order improvements in the asymptotic distribution. One limitation of Hansen's results is that they assume $\{\mathbf{x}_{gt} : t = 1, \ldots, T\}$ are strictly exogenous. We know that if we just use OLS – that is, the usual fixed effects estimate – strict exogeneity is not required for consistency as $T \to \infty$. GLS, in exploiting correlations across different time periods, tends to exacerbate bias that results from a lack of strict exogeneity. In policy analysis cases, this is a concern if the policies can switch on and off over time, because one must decide whether the decision to implement or remove a program is related to past outcomes on the response.

With large G and small T, one can estimate an unstricted variance matrix Ω_g and proceed with GLS – this is the approach suggested by Kiefer (1980) and studied more recently by Hausman and Kuersteiner (2005). It is equivalent to dropping a time period in the time-demeaned equation and proceeding with full GLS (and this avoids the degeneracy in the variance matrix of the time-demeaned errors). Hausman and Kuersteiner show that the Kiefer approach works pretty well when $G = 50$ and $T = 10$, although substantial size distortions exist for $G = 50$ and $T = 20$.

Especially if the M_{gt} are not especially large, we might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. One simple way to avoid ignoring the estimation error in $\hat{\delta}_{gt}$ is to aggregate equation (6.1) over individuals, giving

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt},\ t = 1,..,T, g = 1,\ldots,G. \tag{6.6}$$

Of course, this equation can be estimated by fixed effects, too, and fully robust inference is available using Hansen (2007a) because the composite error, $\{r_{gt} \equiv v_{gt} + \bar{u}_{gt}\}$, is weakly dependent. Fixed Effects GLS using an unrestricted variance matrix can be used with large G and small T. The complication with using specific time series model for the error is the presence of $\bar{u}_{gt}$. With different M_{gt}, $Var(\bar{u}_{gt})$ is almost certainly heteroskedastic (and might be with the same M_{gt}, of course). So, even if we specify, say, an AR(1) model $v_{gt} = \rho v_{g,t-1} + e_{gt}$, the variance matrix of $\mathbf{r}_g$ is more complicated. One possibility is to just assume the composite error, r_{gt}, follows a simple model, implement Hansen's methods, but then use fully robust inference.

The Donald and Land (2007) approach applies in the current setting by using finite sample analysis applied to the pooled regression (6.4). However, DL assume that the errors $\{v_{gt}\}$ are uncorrelated across time, and so, even though for small G and T it uses small degrees-of-freedom in a t distribution, it does not account for uncertainty due to serial correlation in $\{v_{gt} : t = 1,\ldots,T\}$.

References

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

Angrist, J.D. and V. Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER Working Paper 9389.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics* 49, 431-434.

Baker, M. and N.M. Fortin (2001), "Occupational Gender Composition and Wages in Canada, 1987-1988," *Canadian Journal of Economics* 34, 345-376.

Bhattacharya, D. (2005), "Asymptotic Inference from Multi-stage Samples," *Journal of Econometrics* 126, 145-171.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119, 249-275.

Cameron, A.C., J.B. Gelbach, and D.L. Miller (2006), "Robust Inference with Multi-way Clustering," NBER Technical Working Paper Number 327.

Campolieti, M. (2004), "Disability Insurance Benefits and Labor Supply: Some Additional Evidence," *Journal of Labor Economics* 22, 863-889.

Card, D., and A.B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84, 772-793.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of*

Economic Studies 47, 225-238.

Cosslett, S.R. (1993), "Estimation from Endogenously Stratified Samples," in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, Volume 11, 1-43. North-Holland: Amsterdam.

Donald, S.G. and K. Lang (2001), "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics* 89, 221-233.

Hansen, C.B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics* 141, 597-620.

Hausman, J.A., B.H. Hall, and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52, 909-938.

Hsiao, C. (2003), *Analysis of Panel Data*. Cambridge: Cambridge University Press, second edition.

Imbens, G.W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica* 60, 1187-1214.

Imbens, G.W. and T. Lancaster (1996), "Efficient Estimation and Stratified Sampling," *Journal of Econometrics* 74, 289-318.

Kézdi, G. (2001), "Robust Standard Error Estimation in Fixed-Effects Panel Models," mimeo, University of Michigan Department of Economics.

Liang, K.-Y., and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika* 73, 13-22.

Loeb, S. and J. Bound (1996), "The Effect of Measured School Inputs on Academic Achievement: Evidence form the 1920s, 1930s and 1940s Birth Cohorts," *Review of Economics and Statistics* 78, 653-664

Moulton, B.R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics* 72, 334-338.

Pepper, J.V. (2002), "Robust Inferences from Random Clustered Samples: An Application Using Data from the Panel Study of Income Dynamics," Economics Letters 75, 341-345.

Petrin, A. and K. Train (2003), "Omitted Product Attributes in Discrete Choice Models," NBER Working Paper Number 9452.

White, H. (1980), "A Heterosekdasticity-Consistent Covariance Matrix Estimator and A Direct Test for Heteroskeasticity," *Econometrica* 48, 817-838.

White, H. (1982), "Maximum Likelihood Estimation with Misspecified Models," *Econometrica* 50, 1-26.

White, H. (1984), *Asymptotic Theory for Econometricians*. Academic Press: Orlando, FL.

Wooldridge, J.M. (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica* 67, 1385-1406.

Wooldridge, J.M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory* 17, 451-470.

Wooldridge, J.M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93, 133-138.

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

Wooldridge, J.M. (2006), "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," manuscript, Michigan State University Department of Economics.

Wooldridge, J.M. (2007), "Inverse Probability Weighted M-Estimation for General Missing Data Problems," *Journal of Econometrics* 141, 1281-1301.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. MIT Press: Cambridge, MA.

Cross-Section Econometrics
Lecture 4: Cluster Sampling and Difference-in-Differences

Jeff Wooldridge
Michigan State University

AEA Lectures, Chicago, January 2012

1. The Linear Model with Cluster Effects

• For each group or cluster g, let $\{(y_{gm}, \mathbf{x}_g, \mathbf{z}_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster g, y_{gm} is a scalar response, $\mathbf{x}_g$ is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and $\mathbf{z}_{gm}$ is a $1 \times L$ vector of covariates that vary within (as well as across) groups.

• The linear model with an additive error is

$$y_{gm} = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + v_{gm} \tag{1}$$

for $m = 1, \ldots, M_g$, $g = 1, \ldots, G$.

• Key questions:

1. What is the sampling scheme, and how large are the group sizes (M_g) and number of groups (G)?

• Easiest sampling scheme: From a large population of relatively small clusters, we draw a large number of clusters (G), where cluster g has M_g members. For example, sampling a large number of families, classrooms, or firms from a large population.

• In the panel data setting, G is the number of cross-sectional units and M_g is the number of time periods for unit g.

2. Does v_{gm} contain a common group effect, as in

$$v_{gm} = c_g + u_{gm}, m = 1, \ldots, M_g, \tag{2}$$

where c_g is an unobserved group (cluster) effect and u_{gm} is the idiosyncratic component? Usually this is assumed.

3. Are the regressors $(\mathbf{x}_g, \mathbf{z}_{gm})$ appropriately exogenous?

Large Group Asymptotics

• The case of independently sampling groups with $G \to \infty$, fixed group sizes, M_g, is well developed [White (1984), Arellano (1987)]. How should one use these methods? If

$$E(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm}) = 0 \tag{3}$$

then pooled OLS estimator of y_{gm} on $1, \mathbf{x}_g, \mathbf{z}_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G,$ is consistent for $\boldsymbol{\lambda} \equiv (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$ (as $G \to \infty$ with M_g fixed) and $\sqrt{G}$-asymptotically normal.

• Robust variance matrix to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|\mathbf{x}_g, \mathbf{z}_{gm})$, or both. Write $\mathbf{W}_g$ as the $M_g \times (1 + K + L)$ matrix of all regressors for group g. Then the $(1 + K + L) \times (1 + K + L)$ variance matrix estimator is

$$\left(\sum_{g=1}^{G} \mathbf{W}_g'\mathbf{W}_g\right)^{-1}\left(\sum_{g=1}^{G} \mathbf{W}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbf{W}_g\right)\left(\sum_{g=1}^{G} \mathbf{W}_g'\mathbf{W}_g\right)^{-1} \tag{4}$$

where $\hat{\mathbf{v}}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group g.

• This "sandwich" estimator is now computed routinely using "cluster" options.

• Generalized Least Squares: Strengthen the exogeneity assumption to

$$E(v_{gm}|\mathbf{x}_g, \mathbf{Z}_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{5}$$

where $\mathbf{Z}_g$ is the $M_g \times L$ matrix of unit-specific covariates. May have to include "peer effects."

• Full RE approach: the $M_g \times M_g$ variance-covariance matrix of $\mathbf{v}_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the "random effects" form,

$$Var(\mathbf{v}_g) = \sigma_c^2 \mathbf{j}_{M_g}'\mathbf{j}_{M_g} + \sigma_u^2 \mathbf{I}_{M_g}, \tag{6}$$

where $\mathbf{j}_{M_g}$ is the $M_g \times 1$ vector of ones and $\mathbf{I}_{M_g}$ is the $M_g \times M_g$ identity matrix.

• The usual assumptions include the "system homoskedasticity" assumption,

$$Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g) = Var(\mathbf{v}_g). \tag{7}$$

• The random effects estimator $\hat{\boldsymbol{\lambda}}_{RE}$ is asymptotically more efficient than pooled OLS under (5), (6), and (7) as $G \to \infty$ with the M_g fixed. The RE estimates and test statistics are computed routinely by popular software packages.

• Important point: In many cases one should, make RE inference robust to an unknown form of $Var(\mathbf{v}_g|\mathbf{x}_g, \mathbf{Z}_g)$, whether we have a true cluster sample or panel data.

• Why? Random coefficient model:

$$\begin{aligned} y_{gm} &= \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + c_g + u_{gm} \\ &= \alpha + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{z}_{gm}\boldsymbol{\gamma} + \mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma}) + c_g + u_{gm} \end{aligned} \tag{8}$$

By using a standard random effects estimator that assumes common slopes $\boldsymbol{\gamma}$, we effectively include $\mathbf{z}_{gm}(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ in the idiosyncratic error.

• If only $\boldsymbol{\gamma}$ is of interest, fixed effects is attractive. Apply pooled OLS to the equation with group means removed:

$$y_{gm} - \bar{y}_g = (\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)\boldsymbol{\gamma} + u_{gm} - \bar{u}_g. \tag{9}$$

• Can be important to allow $Var(\mathbf{u}_g|\mathbf{Z}_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. Certainly for panel data (serial correlation), but also for cluster sampling.

• From linear panel data notes, FE can consistently estimate the average effect in the random coefficient case. But $(\mathbf{z}_{gm} - \bar{\mathbf{z}}_g)(\boldsymbol{\gamma}_g - \boldsymbol{\gamma})$ appears in the error term.

• A fully robust variance matrix estimator of $\hat{\boldsymbol{\gamma}}_{FE}$ is

$$\left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}\left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\hat{\ddot{\mathbf{u}}}_g\hat{\ddot{\mathbf{u}}}_g'\ddot{\mathbf{Z}}_g\right)\left(\sum_{g=1}^{G} \ddot{\mathbf{Z}}_g'\ddot{\mathbf{Z}}_g\right)^{-1}, \tag{10}$$

where $\ddot{\mathbf{Z}}_g$ is the matrix of within-group deviations from means and $\hat{\ddot{\mathbf{u}}}_g$ is the $M_g \times 1$ vector of fixed effects residuals. This estimator is justified with large-G asymptotics.

• Above results are for "one-way clustering." Cameron, Gelbach, and Miller (2006) have shown how to extend the formulas to multi-way clustering. For example, we have individual-level data with industry and occupation representing different clusters. So we have y_{ghm} for $g = 1, \ldots, G$, $h = 1, \ldots, H$, $m = 1, \ldots, M_{gh}$. An individual belongs to two clusters, implying some correlation across groups.

• Two-way clustering seems to be overused, especially in cases where one dimension is short.

• If explanatory variables vary by individual, two-way fixed effects is attractive and often eliminates the need for cluster-robust inference – at least in one direction.

Should we Use the "Large" G Formulas with "Large" M_g?

- What if one applies robust inference in scenarios where the fixed M_g, $G \to \infty$ asymptotic analysis not realistic? Can apply recent results of Hansen (2007a) to various scenarios.
- Hansen (2007a, Theorem 2): Usual inference based on the robust "sandwich" estimator is valid with arbitrary correlation among the errors, v_{gm}, within each group (but still independence across groups) if G and M_g both grow.

- Example: We sample $G = 100$ schools and roughly $M_g = 100$ students per school, and we use pooled OLS leaving the school effects in the error term. Hansen's results (and simulations) show we should expect the inference to have roughly the correct size.

- Unfortunately, in the presence of cluster effects with a small number of groups (G) and large group sizes (M_g), cluster-robust inference with pooled OLS falls outside Hansen's theoretical findings. We should not expect good properties of the cluster-robust inference with small groups and large group sizes.
- Example: Suppose $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest varies only at the hospital level, tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and reasons to expect it will not work well.

- If the explanatory variables of interest vary within group, FE is attractive. First, allows c_g to be arbitrarily correlated with the $\mathbf{z}_{gm}$. Second, with large M_g, can treat the c_g as parameters to estimate – because we can estimate them precisely – and then assume that the observations are independent across m (as well as g). This means that the usual inference is valid, perhaps with adjustment for heteroskedasticity. The fixed G, large M_g results in Hansen (2007a, Theorem 4) for cluster-robust inference apply, but are likely to be very costly: the usual variance matrix is multiplied by $G/(G-1)$ and the t statistics are approximately distributed as t_{G-1} (not standard normal).

• For panel data applications, Hansen's (2007a) results, particularly Theorem 3, imply that cluster-robust inference for the fixed effects estimator should work well when the cross section (N) and time series (T) dimensions are similar and not too small.

• If full time effects are allowed in addition to unit-specific fixed effects, any serial dependence in the idiosyncratic errors is assumed to be weakly dependent (no unit roots).

• Simulations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007a) verify that the fully robust cluster-robust variance matrix works well when N and T are about 50 and the idiosyncratic errors follow a stable AR(1) model.

2. Estimation with Few Groups and Large Group Sizes

• Recent interest when we have few groups (small G) and large group sizes (each M_g is large).

• It is important to know the sampling scheme. With random sampling from a large population, no clustering is needed.

• Sometimes we have random sampling within each segment (group) of the population. Except for the relative dimensions of G and M_g, the resulting data set is essentially indistinguishable from a data set obtained by sampling entire clusters.

• The problem of proper inference when M_g is large relative to G – the "Moulton (1990) problem" – has been recently studied by Donald and Lang (2007).

• DL treat the parameters associated with the different groups as outcomes of random draws.

• Simplest case: a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \quad (11)$$
$$= \delta_g + \beta x_g + u_{gm}. \quad (12)$$

(12) has a common slope, β but intercept, δ_g, that varies across g.

• Donald and Lang focus on (11), where c_g is assumed to be independent of x_g with zero mean. Define the composite error $v_{gm} = c_g + u_{gm}$.

• Standard pooled OLS inference applied to (11) can be badly biased because it ignores the cluster correlation. Hansen's results do not apply. (And we cannot use fixed effects.)

• DL propose studying the regression in averages:

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G. \quad (13)$$

If we add some strong assumptions, we can perform inference on (13) using standard methods. In particular, assume that $M_g = M$ for all g, $c_g | x_g \sim Normal(0, \sigma_c^2)$ and $u_{gm} | x_g, c_g \sim Normal(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim Normal(0, \sigma_c^2 + \sigma_u^2/M)$. Because we assume independence across g, (13) satisfies the classical linear model assumptions.

• So, we can just use the "between" regression,

$$\bar{y}_g \text{ on } 1, \; x_g, \; g = 1, \ldots, G. \quad (14)$$

With same group sizes, identical to pooled OLS across g and m.

• Conditional on the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g : g = 1, \ldots, G\}$, the within-group averages of the composite errors.

• We can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$.

• If G is small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent then if we use the $t_{M_1+M_2+\ldots+M_G-2}$ distribution (traditional approach).

• Using the between regression is *not* the same as using cluster-robust standard errors for pooled OLS. Those are not justified and, anyway, we would use the wrong df in the t distribution.

• So the DL method uses a standard error from the aggregated regression and degrees of freedom $G - 2$.

• We can apply the DL method without normality of the u_{gm} if the group sizes are large because $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$ so that $\bar{u}_g$ is a negligible part of $\bar{v}_g$. But we still need to assume c_g is normally distributed.

• If $\mathbf{z}_{gm}$ appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, g = 1, \ldots, G, \tag{15}$$

provided $G > K + L + 1$.

• If c_g is independent of $(\mathbf{x}_g, \bar{\mathbf{z}}_g)$ with a homoskedastic normal distribution, and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution. Regressions like (15) are reasonably common, at least as a check on results using disaggregated data, but usually with larger G then just a handful.

• If $G = 2$, should we give up? Suppose x_g is binary, indicating treatment and control ($g = 2$ is the treatment, $g = 1$ is the control). The DL estimate of β is the usual one: $\hat{\beta} = \bar{y}_2 - \bar{y}_1$. But in the DL setting, we cannot do inference (there are zero df). So, the DL setting rules out the standard comparison of means.

• Can we still obtain inference on estimated policy effects using randomized or quasi-randomized interventions when the policy effects are just identified? Not according the DL approach.

• If $y_{gm} = \Delta w_{gm}$ – the change of some variable over time – and x_g is binary, then application of the DL approach to

$$\Delta w_{gm} = \alpha + \beta x_g + c_g + u_{gm},$$

leads to a difference in mean changes, $\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$. But inference is not available no matter the sizes of M_1 and M_2.

• $\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$ has been a workhorse in the quasi-experiemental literature, and obtaining inference in the traditional setting is straightforward [Card and Krueger (1994), for example.]

• Even when DL approach can be applied, should we? Suppose $G = 4$ with two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). DL involves the OLS regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$; inference is based on the t_2 distribution.

• Can show the DL estimate is

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{16}$$

• With random sampling from each group, $\hat{\beta}$ is approximately normal even with moderate group sizes M_g. In effect, the DL approach rejects usual inference based on means from large samples because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$.

• Why not tackle mean heterogeneity directly? Could just define the treatment effect as

$$\tau = (\mu_3 + \mu_4)/2 - (\mu_1 + \mu_2)/2,$$

or weight by population frequencies.

• The expression $\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2$ hints at a different way to view the small G, large M_g setup. DL estimates two parameters, α and β, but there are four population means.

• The DL estimates of α and β can be interpreted as minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. If we use the 4×4 identity matrix as the weight matrix, we get $\hat{\beta}$ and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$.

• With large group sizes, and whether or not G is especially large, we can put the problem into an MD framework, as done by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group.

• For each group g, write

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm}. \tag{17}$$

Assume random sampling within group and independence across groups. OLS estimates within group are $\sqrt{M_g}$-asymptotically normal.

• The presence of $\mathbf{x}_g$ can be viewed as putting restrictions on the intercepts:

$$\delta_g = \alpha + \mathbf{x}_g\boldsymbol{\beta}, g = 1, \ldots, G, \tag{18}$$

where we think of x_g as fixed, observed attributes of heterogeneous groups. With K attributes we must have $G \geq K + 1$ to determine α and $\boldsymbol{\beta}$. In the first stage, obtain $\hat{\delta}_g$, either by group-specific regressions or pooling to impose some common slope elements in $\boldsymbol{\gamma}_g$.

• Let $\hat{\mathbf{V}}$ be the $G \times G$ estimated (asymptotic) variance of $\hat{\boldsymbol{\delta}}$. Let $\mathbf{X}$ be the $G \times (K + 1)$ matrix with rows $(1, \mathbf{x}_g)$. The MD estimator is

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}} \tag{19}$$

• Asymptotics are as the M_g get large, and $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution; its estimated asymptotic variance is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.

• When separate group regressions are used, the $\hat{\delta}_g$ are independent and $\hat{\mathbf{V}}$ is diagonal.

• Estimator looks like "GLS," but inference is with G (number of rows in $\mathbf{X}$) fixed with M_g growing.

• Can test the overidentification restrictions. If reject, can go back to the DL approach, applied to the $\hat{\delta}_g$. With large group sizes, can analyze

$$\hat{\delta}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + c_g, g = 1, \ldots, G \tag{20}$$

as a classical linear model because $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, provided c_g is homoskedastic, normally distributed, and independent of $\mathbf{x}_g$.

• Alternatively, can define the parameters of interest in terms of the δ_g, as in the treatment effects case.

3. The Traditional Difference-in-Differences Methodology

- Standard case: Outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. Structure can apply to repeated cross sections or panel data.
- With repeated cross sections, let A be the control group and B the treatment group. Write

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \tag{21}$$

where y is the outcome of interest.

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u$$

- dB captures possible differences between the treatment and control groups prior to the policy change.
- $d2$ captures aggregate factors that would cause changes in y over time even in the absense of a policy change.
- Coefficient of interest is δ_1.
- The difference-in-differences (DD) estimate is

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}). \tag{22}$$

- Given random samples from each of the four groups, inference is straightforward, and is easily made robust to different group/time period variances in the regression framework.
- Often moderate sample sizes are enough for asymptotics to work well.

- Can refine the definition of treatment and control groups.
- Example: Change in state health care policy aimed at elderly.
- Could use data only on people in the state with the policy change, both before and after the change, with the control group being people 55 to 65 (say) and and the treatment group being people over 65. This DD analysis assumes that the paths of health outcomes for the younger and older groups would not be systematically different in the absense of intervention.

• Instead, use a similar state as an additional control.

• Let A be the control state, B the state where the intervention took place, E the group of elderly, and N the non-elderly. $d2$ is the post-intervention time period.

• Two states, two groups, and two time periods: eight total groups. Estimate the means for each of the eight groups.

• Or, use a regression approach:

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (23)$$

• The OLS estimate $\hat{\delta}_3$ is

$$\hat{\delta}_3 = [(\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1})] - [(\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{A,N,2} - \bar{y}_{A,N,1})] \quad (24)$$

where the A subscript means the state not implementing the policy and the N subscript represents the non-elderly. This is the *difference-in-difference-in-differences (DDD)* estimate.

• Can add covariates to either the DD or DDD analysis to (hopefully) control for compositional changes.

4. How Should We View Uncertainty in DD Settings?

• Traditional approach: All uncertainty in inference enters through sampling error in estimating the means of each group/time period combination. Long history in analysis of variance.

• Recently, a focus on different kinds of uncertainty – perhaps in addition to sampling error in estimating means. Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a,b), and Abadie, Diamond, and Hainmueller (2011) argue for additional sources of uncertainty.

• In the "new" view, the additional uncertainty is often assumed to swamp the sampling error in estimating group/time period means.

• One way to view the uncertainty introduced in the DL framework – and a perspective explicitly taken by ADH – is that our analysis should better reflect the uncertainty in the quality of the control groups.

• Issue: In the standard DD and DDD cases, the policy effect is just identified in the sense that we do not have multiple treatment or control groups assumed to have the same mean responses. So, for example, the DL approach does not allow inference in such cases.

• Example from Meyer, Viscusi, and Durbin (1995) on estimating the effects of raising the salary cap on length of time a worker spends on workers' compensation. MVD have the standard DD before-after setting.

```
. reg ldurat afchnge highearn afhigh if ky, robust

Linear regression                               Number of obs =    5626
                                                F(  3,  5622) =   38.97
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0207
                                                Root MSE      =  1.2692

------------------------------------------------------------------------------
             |               Robust
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    afchnge |   .0076573   .0440344     0.17   0.862    -.078667    .0939817
   highearn |   .2564785   .0473887     5.41   0.000     .1635785    .3493786
     afhigh |   .1906012    .068982     2.76   0.006     .0553699    .3258325
      _cons |   1.125615   .0296226    38.00   0.000     1.067544    1.183687
------------------------------------------------------------------------------
```



```
. reg ldurat afchnge highearn afhigh if mi, robust

Linear regression                               Number of obs =    1524
                                                F(  3,  1520) =    5.65
                                                Prob > F      =  0.0008
                                                R-squared     =  0.0118
                                                Root MSE      =  1.3765

------------------------------------------------------------------------------
             |               Robust
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    afchnge |   .0973808   .0832583     1.17   0.242    -.0659325    .2606941
   highearn |   .1691388   .1070975     1.58   0.114    -.0409358    .3792133
     afhigh |   .1919906   .1579768     1.22   0.224     -.117885    .5018662
      _cons |   1.412737   .0556012    25.41   0.000     1.303674      1.5218
------------------------------------------------------------------------------
```


5. Multiple Groups and Time Periods

• The Donald-Lang approach requires even stronger assumptions when the data stretches across different time periods (except in the simple case where differences of the same unit can be used to produce a single cross section). Because the classical linear model is relied on, difficult to allow serial correlation.

• With many time periods and groups, the approach in BDM (2004) and Hansen (2007a) is useful. At the individual level,

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \qquad (25)$$
$$i = 1, \ldots, M_{gt}.$$

• i indexes individual, g indexes group, and t indexes time. Full set of time effects, λ_t, full set of group effects, α_g, group/time period covariates (policy variabels), $\mathbf{x}_{gt}$, individual-specific covariates, $\mathbf{z}_{igt}$, unobserved group/time effects, v_{gt}, and individual-specific errors, u_{igt}. Interested in $\boldsymbol{\beta}$.

• Write

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{26}$$

a model at the individual level where intercepts and slopes are allowed to differ across all (g, t) pairs.

• Then, we think of δ_{gt} as

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{27}$$

Think of (27) as a model at the group/time period level.

• As discussed by BDM, a common way to estimate and perform inference in (27) is to ignore v_{gt}, so the individual-level observations are treated as independent. When v_{gt} is present, the resulting inference can be very misleading.

• BDM and Hansen (2007a) allow serial correlation in $\{v_{gt} : t = 1, 2, \ldots, T\}$ but assume independence across g.

• If we view (27) as ultimately of interest, there are simple ways to proceed. We observe $\mathbf{x}_{gt}$, λ_t is handled with year dummies,and α_g just represents group dummies. The problem, then, is that we do not observe δ_{gt}. Use OLS on the individual-level data to estimate the δ_{gt}, assuming $E(\mathbf{z}_{igt}' u_{igt}) = \mathbf{0}$ and the group/time period sizes, M_{gt}, are reasonably large.

• Sometimes one wishes to impose some homogeneity in the slopes – say, $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}_g$ or even $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}$ – in which case pooling can be used to impose such restrictions.

• In any case, proceed as if M_{gt} are large enough to ignore the estimation error in the $\hat{\delta}_{gt}$; instead, the uncertainty comes through v_{gt} in (27). In other words, ignore the estimation error in the $\hat{\delta}_{gt}$, and proceed as if, for $t = 1, \ldots, T, g = 1, \ldots, G$,

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{28}$$

• We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (28) by OLS – which means full year and group effects, along with $\mathbf{x}_{gt}$ – then the OLS estimator has satisfying properties as G and T both increase, provided $\{v_{gt} : t = 1,2,\ldots,T\}$ is a weakly dependent time series for all g. The simulations in BDM and Hansen (2007a) indicate that cluster-robust inference, where each cluster is a set of time periods, work reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and G is moderately large.

• Hansen (2007b), noting that the OLS estimator (the fixed effects estimator) applied to (28) is inefficient when v_{gt} is serially uncorrelated, proposes feasible GLS. When T is small, estimating the parameters in $\Omega = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. Bias in estimates based on the FE residuals, $\hat{v}_{gt}$, disappears as $T \to \infty$, but can be substantial even for moderate T. In AR(1) case, $\hat{\rho}$ comes from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1},\ t = 2,\ldots,T, g = 1,\ldots,G. \tag{29}$$

• Hansen shows that an iterative bias-adjusted procedure has the same asymptotic distribution as $\hat{\rho}$ in the case $\hat{\rho}$ should work well: G and T both tending to infinity. Most importantly for the application to DD problems, the feasible GLS estimator based on the iterative procedure has the same asymptotic distribution as the infeasible GLS etsimator when $G \to \infty$ and T is fixed.

• Even when G and T are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-adusted estimates deliver higher-order improvements in the asymptotic distribution.

• One limitation of Hansen's results: they assume $\{\mathbf{x}_{gt} : t = 1,\ldots,T\}$ are strictly exogenous. If we just use OLS, that is, the usual fixed effects estimate – strict exogeneity is not required for consistency as $T \to \infty$. Of course, GLS approaches to serial correlation generally rely on strict exogeneity. In intervention analyis, might be concerned if the policies can switch on and off over time.

• If the M_{gt} are not large, might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. Instead, aggregate over individuals:

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \quad (30)$$
$$t = 1,..,T, g = 1,\ldots,G.$$

• Adding $\bar{\mathbf{z}}_{gt}$ reduces the degrees-of-freedom. Can estimate (22) by FE and use fully robust inference (to account for time series dependence) because the composite error, $\{r_{gt} \equiv v_{gt} + \bar{u}_{gt}\}$, is weakly dependent.

• The Donald and Lang (2007) approach applies in the current setting by using finite sample analysis applied to the pooled regression (30). However, DL assume that the errors $\{v_{gt}\}$ are uncorrelated across time, and so, even though for small G and T it uses small degrees-of-freedom in a t distribution, it does not account for uncertainty due to serial correlation in v_{gt}.

• A minimum distance approach can be used.

1. Obtain the intercepts, $\hat{\delta}_{gt}$, from GT regressions using the individual-level data (same as Hansen) and also the estimates of the sampling variances (the squared standard errors of $\hat{\delta}_{gt}$ made robust to heteroskedasticity); call these s_{gt}^2. (These depend on the M_{gt}.)

2. The efficient minimum distance estimator is the "weighted least squares" estimator from estimating

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \quad (31)$$

using weights $1/s_{gt}^2$. Inference from this WLS regression is valid (even though the sampling theory is different from usual WLS).

• The MD approach puts $GT - G - T - (K+1)$ extra restrictions on the δ_{gt}. These are easily tested after the WLS estimation. The statistic is the weighted sum of squared residuals, say SSR_w. Under the null hypothesis that the restrictions are valid, $SSR_w \overset{a}{\sim} \chi^2_{GT-G-T-(K+1)}$ as the group sizes, M_{gt}, get large.

• With separate slopes $\hat{\boldsymbol{\gamma}}_{gt}$ obtained from each regression, the $\hat{\delta}_{gt}$ are independent across g and t, and there is no serial correlation as in the BDM/Hansen framework. If impose common slopes, MD should account for correlation in the $\hat{\delta}_{gt}$ across g and t.

AEA Lectures **Chicago, January 2012**

Lecture 6, Monday, January 9, 4:30 pm to 5:30 pm

Nonlinear Panel Data Models

These notes summarize some recent, and perhaps not-so-recent, advances in the estimation of nonlinear panel data models. Research in the last 10 to 15 years has branched off in two directions. In one, the focus has been on parameter estimation, possibly only up to a common scale factor, in semiparametric models with unobserved effects that can be arbitrarily correlated with covariates. Another branch has focused on estimating partial effects when restrictions are made on the distribution of heterogeneity conditional on the history of the covariates. These notes attempt to lay out the pros and cons of each approach, paying particular attention to the tradeoff in assumptions and the quantities that can be estimated.

1. Basic Issues and Quantities of Interest

Most microeconomic panel data sets are best characterized as having few time periods and (relatively) many cross section observations. Therefore, most of the discussion in these notes assumes T is fixed in the asymptotic analysis while N is increasing. We assume random sampling in the cross section, that is, $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$, is a random draw of T time periods for observation i. We take the response y_{it} to be a scalar for simplicity.

If we are not concerned about traditional (contemporaneous) endogeneity, then we are typically interested in the conditional distribution

$$D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \tag{1.1}$$

where $\mathbf{c}_i$ is the onobserved heterogeneity for observation i drawn along with the observables. Often we are interested in a particular feature of this distribution, such as $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or a

conditional median. Generally, with nonlinear models, we must deal with the issue of summarizing the effects of the observed covariates while accounting for the presense of $\mathbf{c}_i$. For example, in the case of a mean, how do we summarize the partial effects when they depend on the unobserved heterogeneity? Let $E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t, \mathbf{c})$ be the mean function. If x_{tj} is continuous, then the partial effect can be defined as

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}. \tag{1.2}$$

For discrete (or continuous) variables, we can instead look at discrete changes in the mean function. Either way, a key question is: How do we account for unobserved $\mathbf{c}$? If we want to estimate magnitudes of effects, we need to know enough about the distribution of $\mathbf{c}_i$ so that we can either insert meaningful values for $\mathbf{c}$, or we can average the partial effects across the distribution of $\mathbf{c}_i$. As an example of the former, suppose we can identify $\boldsymbol{\mu}_\mathbf{c} = E(\mathbf{c}_i)$. Then we can compute the *partial effect at the average (PEA)*,

$$\theta_j(\mathbf{x}_t, \boldsymbol{\mu}_\mathbf{c}). \tag{1.3}$$

Of course, to estimate (1.3), we need to estimate the function m_t and the mean of $\mathbf{c}_i$. If we know more about the distribution of $\mathbf{c}_i$, we can insert different quantiles, for example, or a certain number of standard deviations from the mean.

As an alternative to plugging in specific values for $\mathbf{c}$, we can average the partial effects across the distribution of $\mathbf{c}_i$:

$$\text{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)], \tag{1.4}$$

the so-called *average partial effect (APE)*. The difference between (1.3) and (1.4) can be nontrivial for nonlinear mean functions. The definition in (1.4) dates back at least to

Chamberlain (1984), and is closely related to the notion of the *average structural function (ASF)* [Blundell and Powell (2003)]. The ASF is defined as

$$\text{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \tag{1.5}$$

Assuming the derivative passes through the expectation results in (1.5), the average partial effect. Of course, computing a discrete change in the ASF always gives the corresponding APE. A useful feature of APEs is that they can be compared across models, where the functional form of the mean or the distribution of the heterogeneity can be different. In particular, APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Semiparametric methods that are totally silent about the distribution of $\mathbf{c}_i$, unconditionally or conditional on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, cannot generally deliver estimates of PAEs or APEs essentially by design. Instead, an index structure is usually imposed so that parameters can be consistently estimated. A common setup with scalar heterogeneity is

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.6}$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable (and, in some cases, is known, and in others, is not). The partial effects are proportional to the parameters:

$$\theta_j(\mathbf{x}_t, \mathbf{c}) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.7}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Therefore, if we can estimate β_j then we can estimate the sign of the partial effect, and even the relative effects of any two continuous variables. But, even if $G(\cdot)$ is specified (the more common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i: the size of the scale factor for a random draw i, $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$, depends on c_i. Without knowing something about the

distribution of c_i we cannot generally locate $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$ or average out the heterogeneity.

Returning to the general case, Altonji and Matzkin (2005) focus on what they call the *local average response (LAR)* as opposed to the APE or PAE. The LAR at $\mathbf{x}_t$ for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t), \tag{1.8}$$

where $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a "local" partial effect because it averages out the heterogeneity for the slice of the population described by the vector of observed covariates, $\mathbf{x}_t$. The APE, which, by comparison, could be called a "global average response," averages out over the entire distribution of $\mathbf{c}_i$. See also Florens, Heckman, Meghir, and Vytlacil (2007).

It is important to see that the previous definitions of partial effects does not depend on the nature of the variables in $\mathbf{x}_t$ (except for whether it makes sense to use the calculus approximation or use changes). In particular, $\mathbf{x}_t$ can include lagged dependent variables and lags of other variables, which may or may not be strictly exogenous. Unfortunately, we cannot identify the APEs, or even relative effects in index models, without some assumptions.

2. Exogeneity Assumptions on the Covariates

Ideally, we would only have to specify a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or some feature, to estimate parameters and partial effects. Unfortunately, it is well known that specifying a full parametric model is not sufficient for identifying either the parameters of the model or the partial effects defined in Section 1. In this section, we consider two useful exogeneity assumptions on the covariates.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i), \tag{2.1}$$

which means that $\mathbf{x}_{ir}$, $r \neq t$, does not appear in the conditional distribution of $\mathbf{y}_{it}$ once $\mathbf{x}_{it}$ and $\mathbf{c}_i$ have been counted for. Chamberlain (1984) labeled (2.1) *strict exogeneity conditional on the unobserved (or latent) effects* $\mathbf{c}_i$; as discussed by Chamberlain, (2.1) is much more plausible than if we did not condition on $\mathbf{c}_i$. Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i), \tag{2.2}$$

which we already saw for linear models. (In other cases a condition stated in terms of conditional medians is more convenient.) Assumption (2.1), or its conditional mean version, are less restrictive than if we do not condition on $\mathbf{c}_i$. Indeed, it is easy to see that, if (2.1) holds and $D(\mathbf{c}_i|\mathbf{x}_i)$ depends on $\mathbf{x}_i$, then strict exogeneity without conditioning on $\mathbf{c}_i$, $D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$, cannot hold. Unfortunately, both (2.1) and (2.2) rule out lagged dependent variables, as well as other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in $\mathbf{x}_{ir}$, $r > t$. (Essentially the same problem shows up in linear models, but there it is more easily dealt with.) Neverthless, the conditional strict exogeneity assumption underlies the most common estimation methods for nonlinear models.

More natural is *sequential exogeneity conditional on the unobserved effects*, which we can state generally as

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{it},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i), \tag{2.3}$$

or, sometimes, in terms of specific features of the distribution. Assumption (2.3) allows for

lagged dependent variables and does not restrict feedback. Unfortunately, (2.3) is more difficult to work with than (2.1) for general nonlinear models.

Because we condition on $\mathbf{x}_{it}$, neither (2.1) nor (2.3) allows for contemporaneous endogeneity of one or more elements of $\mathbf{x}_{it}$, where, say, x_{itj} is correlated with unobserved, time-varying unobservables that affect y_{it}, or where x_{itj} is simultaneously determined along with y_{it}. Such cases will be covered in later notes on control function methods.

3. Conditional Independence Assumption

The exogeneity conditions stated in Section 2 generally do not restrict the dependence in the responses, $\{y_{it} : t = 1, \ldots, T\}$, although in special cases (2.3) does. Often, a *conditional independence* assumption is explicitly imposed. We can write the condition generally as

$$D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it} | \mathbf{x}_i, \mathbf{c}_i). \tag{3.1}$$

Equation (3.1) simply means that, conditional on the entire history $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ and the unobserved heterogeneity $\mathbf{c}_i$, the responses are independent across time. One way to think about (3.1) is that time-varying unobservables are independent over time. Because (3.1) conditions on $\mathbf{x}_i$, it is useful only in the context of the strict exogeneity assumption (2.1). Then, conditional independence can be written as

$$D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i). \tag{3.2}$$

In a parametric context, the conditional independence assumption reduces our task to specifying a model for $D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i)$, and then determining how to treat the unobserved

heterogeneity, $\mathbf{c}_i$. In random effects and correlated RE frameworks, conditional independence can play a critical role in being able to estimate the parameters and the distribution of $\mathbf{c}_i$. We could get by with less restrictive assumptions by parameterizing the dependence in the joint distribution $D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i)$ – something that is sometimes done – but that increases computational burden. As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. [That is, we do not need to place restrictions on $D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i)$.] Before we can study estimation, we must discuss the critical issue of the dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$.

4. Assumptions about the Unobserved Heterogeneity

The modern approach to panel data analysis with micro data treats the unobserved heterogeneity as random draws along with the observed data, and that is the view taken here. Nevertheless, in order to avoid making distributional assumptions about $\mathbf{c}_i$, one sometimes treats the $\mathbf{c}_i$ as parameters to estimate, and so we allow for that possibility in our discussion.

Random Effects

For general nonlinear models, what we call the *random effects assumption* is independence between $\mathbf{c}_i$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$:

$$D(\mathbf{c}_i | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{4.1}$$

If we combine this assumption with a model for the conditional mean, $m_t(\mathbf{x}_t, \mathbf{c})$, then the APEs are actually nonparametrically identified. (And, in fact, we do not need to assume strict or sequential exogeneity to use a pooled estimation method, or to use just a single time period.) In

fact, if $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ and $D(\mathbf{c}_i|\mathbf{x}_{it}) = D(\mathbf{c}_i)$, then the APEs are obtained from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{4.2}$$

(The argument is a simple application of the law of interated expectations; it is discussed in detail in Wooldridge (2005a).) In principle, $E(y_{it}|\mathbf{x}_{it})$ can be estimated nonparametrically, and we only need a single time period to identify the partial effects for that time period.

In some leading cases (for example random effects probit and Tobit models with heterogeneity normally distributed and homoskedastic), if we want to obtain partial effects for different values of $\mathbf{c}$, we must assume more: the strict exogeneity assumption (2.1), the conditional independence assumption (3.1), and the random effects assumption (4.1) – with a parametric distribution for $D(\mathbf{c}_i)$ – are typically sufficient. We postpone this discussion because it takes us into the realm of specifying parametric models.

Correlated Random Effects

A *correlated random effects framework* allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but the dependence in restricted in some way. In a parametric setting, we specify a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as in Mundlak (1978), Chamberlain (1982), and many subsequent authors; see Wooldridge (2010). For many models, including probit and Tobit, one can allow $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend in a "nonexchangeable" manner on the time series of the covariates; Chamberlain's correlated random effects probit model does this. But the distributional assumptions that lead to simple estimation – namely, homoskedastic normal with a linear conditional mean — are restrictive. But it is aslo possible to assume

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i) \tag{4.3}$$

without specifying $\mathrm{D}(c_i|\bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. We will see in the

next section that (4.3) is very powerful.

We can go further. For example, suppose that we think the heterogeneity $\mathbf{c}_i$ is correlated with features of the covariate history other than just the time average. Altonji and Matzkin (2005) allow for $\bar{\mathbf{x}}_i$ in equation (4.3) to be replaced by other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as sample variances and covariance. These are examples of "exchangeable" functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, say, $\mathbf{w}_i$ – that is, statistics whose value is the same regardless of the ordering of the $\mathbf{x}_{it}$. Non-exchangeable functions can be used, too. For example, we might think that $\mathbf{c}_i$ is correlated with individual-specific trends, and so we define $\mathbf{w}_i$ to include the intercept and slope from the unit-specific regressions $\mathbf{x}_{it}$ on 1, t, $t = 1, \ldots, T$ (for $T \geq 3$); we can also add the error variance from this individual specific regression if we have a sufficient number of time periods. Regardless of how we choose $\mathbf{w}_i$, the key condition is

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\mathbf{w}_i). \tag{4.4}$$

Practically, we need to specify $\mathbf{w}_i$ and then establish that there is enough variation in $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ separate from $\mathbf{w}_i$ in order to identify either parameters or, more like, average partial effects; this will be clear in the next section.

Fixed Effects

Unfortunately, the label "fixed effects" is used in different ways by different researchers (and, sometimes, by the same researcher). The traditional view is that a fixed effects framework meant $\mathbf{c}_i$, $i = 1, \ldots, N$ were treated as parameters to estimate. This view is still around, and, when researchers say they estimated a nonlinear panel data model by "fixed effects," they sometimes mean the $\mathbf{c}_i$ were treated as parameters to estimate along with other parameters (whose dimension does not change with N). As is well known, except in special

cases, estimation of the $\mathbf{c}_i$ generally introduces an "incidental parameters" problem. (More on this later when we discuss estimation methods.) Subject to computational feasibility, the approach that treats the $\mathbf{c}_i$ as parameters is widely applicable. The practical question is whether the stance of treating the $\mathbf{c}_i$ as parameters delivers "good" estimates of the population parameters and the partial effects.

Rather than meaning the $\mathbf{c}_i$ are parameters to estimate, the "fixed effects" label can mean that $\mathbf{c}_i$ is random but $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted. Even in that case, there are different approaches to estimation of parameters. One is to specify a joint distribution $D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i)$ such that a sufficient statistic, say $\mathbf{s}_i$, can be found such that

$$D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{s}_i), \tag{4.5}$$

and where the latter distribution still depends on the parameters of interest in a way that identifies them. In most cases, the conditional independence assumption (3.1) is maintained, although one conditional MLE is known to have robustness properties: the so-called "fixed effects" Poisson estimator. We cover that in Section 7.

5. Nonparametric Identification of Average Partial and Local Average Effects

Before considering identification and estimation of parameters in parametric models, it is useful to ask which quantities, if any, are identified without imposing parametric assumptions. Not surprisingly, there are no known results on nonparametric identificiation of partial effects evaluated at specific values of $\mathbf{c}$, such as $\boldsymbol{\mu}_\mathbf{c}$ – except, of course, when the partial effects do not depend on $\mathbf{c}$. Interestingly, identification can fail even under a full set of strong parametric

assumptions. For example, in the probit model with unobserved heterogeneity,

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\boldsymbol{\beta} + c), \tag{5.1}$$

where $\mathbf{x}$ is $1 \times K$ and includes unity, the partial effect for a continuous variable x_j is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta} + c)$. Assuming $E(c) = 0$, which is without loss of generality when $x_1 = 1$, the partial effect at the mean of c is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$. Suppose we make the stronger assumption that $c|\mathbf{x}$ ~Normal$(0, \sigma_c^2)$. Then it is easy to show (see Wooldridge (2010, Chapter 15)) that

$$P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}), \tag{5.2}$$

which means that only the scaled parameter vector $\boldsymbol{\beta}_c \equiv \boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}$ is identified. Therefore, $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$ is evidently unidentified. (The fact that probit of y on $\mathbf{x}$ estimates $\boldsymbol{\beta}_c$ rather than $\boldsymbol{\beta}$ has been called the "attenuation bias" that results from omitted variables in the context of probit, even when the omitted variable is independent of the covariates and normally distributed. As mentioned earlier more generally, the average partial effects are obtained directly from $P(y = 1|\mathbf{x})$, and, in fact, are given by $\beta_{cj} \phi(\mathbf{x}\boldsymbol{\beta}_c)$. As discussed in Wooldridge (2010, Chapter 15), $\beta_{cj} \phi(\mathbf{x}\boldsymbol{\beta}_c)$ can be larger or smaller in magnitude than the PEA $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$: $|\beta_{cj}| \leq |\beta_j|$ but $\phi(\mathbf{x}\boldsymbol{\beta}_c) \geq \phi(\mathbf{x}\boldsymbol{\beta})$.)

A related example is due to Hahn (2001), and is related to the nonidentification results of Chamberlain (1993). Suppose that x_{it} is a binary indicator (for example, a policy variable). Consider the unobserved effects probit model for two time periods,

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = \Phi(\beta x_{it} + c_i), \; t = 1, 2. \tag{5.3}$$

As discussed by Hahn, β is not known to be identified in this model, even under the conditional independence assumption (2.1) *and* the random effects assumption $D(c_i|\mathbf{x}_i) = D(c_i)$. But the average partial effect, which in this case is an average treatment

effect, is simply $\tau \equiv E[\Phi(\beta + c_i)] - E[\Phi(c_i)]$. By the general result cited earlier, τ is consistently estimated (in fact, unbiasedly estimated) by using a difference of means for the treated and untreated groups, for either time period. (If treatment is only in the second time period, as in Hahn (2001), then the difference must be for the second time period.) In fact, as discussed in Wooldridge (2005a), identification of the APE holds if we replace Φ with an unknown function G and allow $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. But the parameters are still not identified.

The previous examples raise the following question: Are we focusing too much on parameters in nonlinear models with unobserved heterogeneity? The answer seems to be yes, but with qualifications. Consider a third example, due to Wooldridge (2005c). The binary variable y is determined by the index model $y = 1[\mathbf{x}\boldsymbol{\beta} + u > 0]$, where $u|\mathbf{x}$ ~Normal$(0, \exp(2\mathbf{x}_1\boldsymbol{\delta}))$, where $\mathbf{x}_1$ is a subset of $\mathbf{x}$ that does not contain an intercept. This model is often called a *heteroskedastic probit* model. Of course, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are estimable by MLE because $P(y = 1|\mathbf{x}) = \Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$. However, the APE for, say, the continuous variable x_j is *not* obtained by differentiating $P(y = 1|\mathbf{x})$ with respect to x_j; in fact, as is well known, this derivative can have a sign different from the sign of β_j. Instead, the average structural function is consistently estimated by

$$\widehat{ASF}(\mathbf{x}) = \left\{N^{-1}\sum_{i=1}^{N}\Phi[\exp(-\mathbf{x}_{i1}\hat{\boldsymbol{\delta}})\mathbf{x}\hat{\boldsymbol{\beta}}]\right\},$$

and the partial derivative with respect to x_j always has the same sign as $\hat{\beta}_j$. Notice how the ASF averages across the argument $\mathbf{x}_{i1}$ in the heteroskedasticity function. That comes about because we can write $ASF(\mathbf{x}) = E_{\mathbf{x}_{i1}}\{E(1[\mathbf{x}\boldsymbol{\beta} + u_i > 0]|\mathbf{x}_{i1})\} = E\{\Phi[\exp(-\mathbf{x}_{i1}\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]\}$. The point of this example is that in this case the parameters actually give us the APEs up to the same, positive factor (which depends on the parameters and $\mathbf{x}$), and so the sign of the β_j gives

us the direction of the effect on the APE, and ratios of parameters on continuous variables provide the relative APEs. By contrast, if we blindly differentiate $\Phi[\exp(-\mathbf{x}_1\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$ with respect to x_j and x_j appears in $\mathbf{x}_1$, the resulting expression is not the APE. In other words, parameters tell us more than derivatives in this case. Of course, we will prefer to take derivatives of the appropriate function in (5.4), thereby getting consistent estimates of the APEs. See Wooldridge (2005c) for further discussion of this kind of example, including the negative finding that the is no way to distinguish between the heteroskedastic probit model and a model with random slope coefficients. (And, in the latter case, we *do* obtain the APEs by differentiating $P(y = 1|\mathbf{x})$ with respect to x_j.)

Returning to the panel data case, we can establish identification of average partial effects much more generally. Assume only that the strict exogeneity assumption (2.1) holds along with $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. These two assumptions are sufficient to identify the APEs. To see why, note that the average structural function at time t can be written as

$$\mathrm{ASF}_t(\mathbf{x}_t) = \mathrm{E}_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] = \mathrm{E}_{\bar{\mathbf{x}}_i}\{\mathrm{E}[m_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]\} \equiv \mathrm{E}_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)], \tag{5.4}$$

where $r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i) \equiv \mathrm{E}[m_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]$. It follows that, given an estimator $\hat{r}_t(\bullet, \bullet)$ of the function $r_t(\bullet, \bullet)$, the ASF can be estimated as

$$\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1}\sum_{i=1}^{N}\hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i), \tag{5.5}$$

and then we can take derivatives or changes with respect to the entries in $\mathbf{x}_t$. Notice that (5.4) holds without the strict exogeneity assumption (2.1) or the assumption $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. However, these assumptions come into play in our ability to estimate $r_t(\bullet, \bullet)$. If we combine (2.1) and (4.3) we have

$$\mathrm{E}(y_{it}|\mathbf{x}_i) = \mathrm{E}[\mathrm{E}(y_{it}|\mathbf{x}_i, \mathbf{c}_i)|\mathbf{x}_i] = \mathrm{E}[m_t(\mathbf{x}_{it}, \mathbf{c}_i)|\mathbf{x}_i] = \int m_t(\mathbf{x}_{it}, \mathbf{c})\mathrm{d}F(\mathbf{c}|\mathbf{x}_i)$$
$$= \int m_t(\mathbf{x}_{it}, \mathbf{c})\mathrm{d}F(\mathbf{c}|\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i), \tag{5.6}$$

where $F(\mathbf{c}|\mathbf{x}_i)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_i)$ (which can be a discrete, continuous, or mixed distribution), the second equality follows from (2.1), the fourth equality follows from assumption (4.3), and the last equality folllows from the definition of $r_t(\cdot, \cdot)$ Of course, because $\mathrm{E}(y_{it}|\mathbf{x}_i)$ depends only on $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$, we must have

$$\mathrm{E}(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i). \tag{5.7}$$

Further, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is assumed to have time variation, and so $\mathbf{x}_{it}$ and $\bar{\mathbf{x}}_i$ can be used as separate regressors even in a fully nonparametric setting.

Altonji and Matskin (2005).use this idea more generally, and focus on estimating the local average response. Wooldridge (2005a) used $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$ generally in the case $\mathbf{x}_{it}$ is discrete, in which case a full nonparametric analysis is easy. When

$$\mathrm{D}(\mathbf{c}_i|\mathbf{x}_i) = \mathrm{D}(\mathbf{c}_i|\mathbf{w}_i) \tag{5.8}$$

for $\mathbf{w}_i$ a function of $\mathbf{x}_i$, Altonji and Matzkin (2005) show that the LAR can be obtained as

$$\int \frac{\partial r_t(\mathbf{x}_t, \mathbf{w})}{\partial x_{tj}} \mathrm{d}K_t(\mathbf{w}|\mathbf{x}_t), \tag{5.9}$$

where $r(\mathbf{x}_t, \mathbf{w}) = \mathrm{E}(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{w}_i = \mathbf{w})$ and $K_t(\mathbf{w}|\mathbf{x}_t)$ is the cdf of $\mathrm{D}(\mathbf{w}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. Altonji and Matskin demonstrate how to estimate the LAR based on nonparametric estimation of $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$ followed by "local" averaging, that is, averaging $\partial r(y_{it}|\mathbf{x}_t, \mathbf{w}_i)/\partial x_{tj}$ over observations i with $\mathbf{x}_{it}$ "close" to $\mathbf{x}_t$.

This analysis demonstrates that APEs are nonparametrically identified under the conditional mean version of strict exogeneity, $E(y_{it}|\mathbf{x}_i, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, and (5.8), at least for

time-varying covariates if $\mathbf{w}_i$ is restricted in some way. In fact, we can identify the APEs for a single time period with just one year of data on y. We only need to obtain $\mathbf{w}_i$ (with $\mathbf{w}_i = \bar{\mathbf{x}}_i$) the leading case) and, in effect, include it as a control. Of course, efficiency would be gained by assuming some stationarity across t and using a pooled method.

6. Dynamic Models

General models with only sequentially exogenous variables are difficult to deal with. Arellano and Carrasco (2003) consider probit models. Wooldridge (2000) suggests a strategy the requires modeling the dynamic distribution of the variables that are not strictly exogenous.

Much more is known about models with lagged dependent variables and otherwise strictly exogenous variables. So, we start with a model for

$$D(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \ldots, \mathbf{z}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i), \; t = 1, \ldots, T, \tag{6.1}$$

which we assume also is $D(\mathbf{y}_{it}|\mathbf{z}_i, \mathbf{y}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i)$ where $\mathbf{z}_i$ is the entire history $\{\mathbf{z}_{it} : t = 1, \ldots, T\}$. This is the sense in which the $\mathbf{z}_{it}$ are strictly exogenous.

Suppose this model depends only on $(\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$, so $f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta})$. The joint density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{y}_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}). \tag{6.2}$$

The problem with using this for estimation is the presence of $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$. Several approaches have bee suggestd: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (incidental parameters problem, although recent research has attempted to reduce the asymptotic bias in the partial effects). (ii) Try to estimate the parameters without specifying conditional or

unconditional distributions for c_i. (Available in some special cases covered below, but other restrictions are needed. And, generally, cannot estimate partial effects.). (iii) Find or, more practically, approximate $D(\mathbf{y}_{i0}|\mathbf{c}_i,z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. After integrating out c_i we obtain the density for $D(\mathbf{y}_{i0},\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{z}_i)$ and we can use MLE (conditional on z_i), (iv) Model $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$. After integrating out c_i we obtain the density for $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{y}_{i0},\mathbf{z}_i)$, and we can use MLE (conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$). As shown by Wooldridge (2005b), in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ that mixes with the density $f(\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{y}_0,\mathbf{z},\mathbf{c})$ to produce a log-likelihood that is in a common family and carried out by standard software.

If $m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})$ is the mean function $E(y_t|\mathbf{x}_t,\mathbf{c})$ for a scalar y_t, then average partial effects are easy to obtain. The average structural function is

$$ASF(\mathbf{x}_t) = E_{c_i}[m_t(\mathbf{x}_t,\mathbf{c}_i,\theta)] = E\left\{\left[\int m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})h(\mathbf{c}|y_{i0},\mathbf{z}_i,\boldsymbol{\gamma})d\mathbf{c}\right]|y_{i0},\mathbf{z}_i\right\}. \quad (6.3)$$

The term inside the brackets, say $r_t(\mathbf{x}_t,y_{i0},\mathbf{z}_i,\boldsymbol{\theta},\boldsymbol{\gamma})$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, in fact. Generally, it can be simulated. In any case, $ASF(\mathbf{x}_t,\boldsymbol{\theta})$ is consistently estimated by

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{t=1}^{T} r_t(\mathbf{x}_t,y_{i0},\mathbf{z}_i,\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}).$$

Partial derivatives and differences with respect to elements of x_t (which, remember, can include y_{t-1}) can be computed. With large N and small T, the panel data bootstrap can be used for standard errors and inference.

7. Applications to Specific Models

We now turn to some common parametric models and highlight the difference between estimation partial effects at different values of the heterogeneity and estimating average partial effects. An analysis of Tobit models follows in a very similar way to those in the following two sections. See Wooldridge (2010, Chapter 17) and Honoré and Hu (2004).

7.1 Binary and "Fractional" Response Models

We start with the standard specification for the unobserved effects (UE) probit model, which is

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1, \ldots, T, \tag{7.1}$$

where $\mathbf{x}_{it}$ does not contain an overall intercept but would usually include time dummies. We cannot identify $\boldsymbol{\beta}$ or the APEs without further assumptions. The traditional RE probit models imposes a strong set of assumptions: strict exogeneity, conditional serial independence, and independence between c_i and $\mathbf{x}_i$ with c_i ~Normal(μ_c, σ_c^2). Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE (conditional on $\mathbf{x}_i$).

We can relax independence between c_i and $\mathbf{x}_i$ using the Chamberlain-Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2), \tag{7.2}$$

where the time average is often used to save on degrees of freedom. We can relax (7.2) and allow Chamberlain's (1980) more flexible device:

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1}\boldsymbol{\xi}_1 + \ldots + \mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i \tag{7.3}$$

Even when the $\boldsymbol{\xi}_r$ seem to be very different, the Mundlak restriction can deliver similar

estimates of the other parameters and the APEs. (In the linear case, they both produce the usual FE estimator of $\boldsymbol{\beta}$.)

If we still assume conditional serial independence then all parameters are identified. We can estimate the mean of c_i as $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Of course, c_i is not generally normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. The approximation might get better as T gets large. In any case, we can plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated as

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{7.4}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\text{ASF}}(\mathbf{x}_t)$ with respect to elements of $\mathbf{x}_t$ can be compared with fixed effects estimates from a linear model. Often, if we also average out across $\mathbf{x}_{it}$, the linear FE estimates are similar to the averaged effects.

As we discussed generally in Section 5, the APEs are defined without the conditional serial independence assumption. Without $D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,c_i) = \prod_{t=1}^{T}D(y_{it}|\mathbf{x}_{it},c_i)$, we can still estimate the scaled parameters because

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a), \tag{7.5}$$

and so pooled probit consistently estimates the scaled parametes. (Time dummies have been supressed for simplicity.) Now we have direct estimates of $\boldsymbol{\beta}_a$, ψ_a, and $\boldsymbol{\xi}_a$, and we insert those directly into (7.4).

Using pooled probit can be inefficient for estimating the scaled parameters, whereas the

full MLE is efficient but not (evidently) robust to violation of the conditional serial independence assumption. It is possible to estimate the parameters more efficiently than pooled probit that is still consistent under the same set of assumptions. One possibility is minimum distance estimation. That is, estimate a separate models for each t, and then impose the restrictions using minimum distance methods. (This can be done with or without the Mundlak device.)

A different approach is to apply the so called "generalized estimating equations" (GEE) approach. Briefly, GEE applied to panel data is essentially weighted multivariate nonlinear least squares (WMNLS) with explicit recognition that the weighting matrix might not be the inverse of the conditional variance matrix. In most nonlinear panel data models, obtaining the actual matrix $Var(\mathbf{y}_i|\mathbf{x}_i)$ is difficult, if not impossible, because integrating out the heterogeneity does not deliver a closed form. The GEE approach uses $Var(y_{it}|\mathbf{x}_i)$ implied by the specific distribution – in the probit case, we have the correct conditional variances,

$$Var(y_{it}|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)[1 - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)] \equiv v_{it}. \tag{7.6}$$

The "working" correlation matrix oftenusually specified as "exchangeable,"

$$Corr(e_{it}, e_{is}|\mathbf{x}_i) \text{ " = "} \rho, \tag{7.7}$$

where $e_{it} = [y_{it} - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)]v_{it}^{1/2}$ is the standardized error. Or, each pair (t,s) is allowed to have its own correlation but which is assumed to be independent of $\mathbf{x}_i$ ("unstructured"). The conditional correlation $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ is not constant, but that is the working assumption. The hope is to improve efficiency over the pooled probit estimator while maintaining the robustness of the pooled estimator. (The full RE probit estimator is not robust to serial dependence.) A robust sandwich matrix is easily computed provided the conditional

mean function (in this case, response probability) is correctly specified.

Because the Bernoulli log-likelihood is in the linear exponential family (LEF), exactly the same methods can be applied if $0 \leq y_{it} \leq 1$ – that is, y_{it} is a "fractional" response – but where the model is for the conditional mean: $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$. Pooled "probit" or minimum distance estimation or GEE can be used. Now, however, we must make inference robust to $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ not having the probit form. (There are cases where $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ is proportional to (7.6), and so it still makes sense to use the probit quasi-log-likelihood. Pooled nonlinear regression is another possibility or weighted multivariate nonlinear regression are also possible and a special case of GEE.)

A more radical suggestion, but in the spirit of Altonji and Matzkin (2005) and Wooldridge (2005a), is to just use a flexible model for $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly. For example, if y_{it} is binary, or a fractional response, $0 \leq y_{it} \leq 1$, we might just specify a flexible parametric model, such as

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.8}$$

or the "heteroskedastic probit" model

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[(\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})\exp(-\bar{\mathbf{x}}_i\boldsymbol{\eta})]. \tag{7.9}$$

If we write either of these functions as $r_t(\mathbf{x}_t, \bar{\mathbf{x}})$ then the average structural function is estimated as $\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1}\sum_{i=1}^{N} \hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)$, where the "^" indicates that we have substituted in the parameter estimates. We can let all parameters depend on *t*, or we can estimate the parameters separately for each *t* and then use minimum distance estimation to impose the parameter restrictions. The justification for using, say, (7.8) is that we are interested in the average partial effects, and how parameters appear is really not the issue. Even though (7.8) cannot be derived from $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$ or any other standard model, there is nothing sacred about this

formulation. In fact, it is fairly simplistic. We can view (7.8) as the approximation to the true $E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ obtained after integrating c_i out of the unknown function $m(\mathbf{x}_t, c_i)$. (We could formalize this process by using series estimation, as in Newey (1988), where the number of terms is allowed to grow with N.) This is the same argument used by, say, Angrist (2001) in justifying linear models for limited dependent variables when the focus on primarily on average effects.

The argument is essentially unchanged if we replace $\bar{\mathbf{x}}_i$ with other statistics $\mathbf{w}_i$. For example, we might run, for each i, the regression $\mathbf{x}_{it}$ on $1, t$, $t = 1, \ldots, T$ and use the intercept and slope (on the time trend) as the elements of $\mathbf{w}_i$. Or, we can use sample variances and covariances for each i, along with the sample mean. Or, we can use initial values and average growth rates. The key condition is $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\mathbf{w}_i)$, and then we need sufficient variation in $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ not explained by $\mathbf{w}_i$ for identification. (Naturally, as we expand $\mathbf{w}_i$, the number of time periods required generally increases.)

Of course, once we just view (7.8) as an approximation, we can are justified in using the logistic function, say

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Lambda[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.10}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$, which, again, can be applied to binary or fractional responses. The focus on partial effects that average out the heterogeneity can be liberating in that it means the step of specifying $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ is largely superfluous, and, in fact, can get in the way of pursuing a suitably flexible analysis. On the other hand, if we start with, say, a "structural" model such as $P(y_{i1} = 1|\mathbf{x}_i, \mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$, which is a heterogeneous index model, then we cannot derive equations such as (7.8) or (7.9), even under the strong assumption that $\mathbf{c}_i$ is independent of $\mathbf{x}_i$ and multivariate normal. If we imposed the

Chamberlain device for the elements of $\mathbf{c}_i$ we can get expressions "close" to a combination of (7.8) and (7.9). Whether one is willing to simply estimate relative simple models such as (7.8) in order to estimate APEs depends on one's taste for bypassing more traditional formulations.

If we start with the logit formulation

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{7.11}$$

then we can estimate the parameters, $\boldsymbol{\beta}$ without restricting $D(c_i|\mathbf{x}_i)$ in any way, but we must add the conditional independence assumption. (No one has been able to show that, unlike in the linear model, or the Poisson model covered below, that the MLE that conditions on the number of successes $n_i = \sum_{t=1}^{T} y_{it}$ is robust to serial dependence. It appears not to be. Plus, the binary nature of y_{it} appears to be critical, so the conditional MLE cannot be applied to fractional responses even under serial independence.) Because we have not restricted $D(c_i|\mathbf{x}_i)$ in any way, it appears that we cannot estimate average partial effects. As commonly happens in nonlinear models, if we relax assumptions about the distribution of heterogeneity, we lose the ability to estimate partial effects. We can estimate the effects of the covariates on the log-odds ratio, and relative partial effects of continuous variables. But for partial effects themselves, we do not have sensible values to plug in for c, and we cannot average across its distribution.

The following table summarizes the features of various approaches to estimating binary response unobserved effects models.

Model, Estimation Method	$P(y_{it}= 1\|x_{it}, c_i)$	Restricts $D(c_i\|x_i)$?	Idiosyncratic Serial	PEs	APEs?
	Bounded in (0,1)?		Dependence?	at $E(c_i)$?	
RE Probit, MLE	Yes	Yes (indep, normal)	No	Yes	Yes
RE Probit, Pooled MLE	Yes	Yes (indep, normal)	Yes	No	Yes
RE Probit, GEE	Yes	Yes (indep, normal)	Yes	No	Yes
CRE Probit, MLE	Yes	Yes (lin. mean, normal)	No	Yes	Yes
CRE Probit, Pooled MLE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
CRE Probit, GEE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
LPM, Within	No	No	Yes	Yes	Yes
FE Logit, MLE	Yes	No	No	No	No

As an example, we apply several of the methods to women's labor force participation data, used by Chay and Hyslop (2001), where the data are for five time periods spaced four months apart. The results are summarized in the following table. The standard errors for the APEs were obtained with 500 bootstrap replications. The time-varying explanatory variables are log of husband's income and number of children, along with a full set of time period dummies. (The time-constant variables race, education, and age are also included in columns (2), (3), and (4).)

	(1)	(2)		(3)		(4)		(5)
Model	Linear	Probit		CRE Probit		CRE Probit		FE Logit
Estimation Method	Fixed Effects	Pooled MLE		Pooled MLE		MLE		MLE
	Coefficient	Coefficient	APE	Coefficient	APE	Coefficient	APE	Coefficient
kids	−.0389	−.199	−.0660	−.117	−.0389	−.317	−.0403	−.644
	(.0092)	(.015)	(.0048)	(.027)	(.0085)	(.062)	(.0104)	(.125)
lhinc	−.0089	−.211	−.0701	−.029	−.0095	−.078	−.0099	−.184
	(.0046)	(.024)	(.0079)	(.014)	(.0048)	(.041)	(.0055)	(.083)
$\overline{kids}$	—	—	—	−.086	—	−.210	—	—
	—	—	—	(.031)	—	(.071)	—	—
$\overline{lhinc}$	—	—	—	−.250	—	−.646	—	—
	—	—	—	(.035)	—	(.079)	—	—
$(1+\hat{\sigma}_a^2)^{-1/2}$	—	—		—		.387		—
Log Likelihood	—	−16,556.67		−16,516.44		−8,990.09		−2,003.42
Number of Women	5,663	5,663		5,663		5,663		1,055

In the three methods that allow for unobserved heterogeneity correlated with the covariates and where we can estimate APEs – columns (1), (3), and (4) – the estimated APEs are pretty similar. Column (2) contains the pooled probit estimates without allowing the Chamberlain-Mundlak device, and the APEs are much larger, especially on *lhinc*. Comparing columns (2) and (3) stronly suggest the presence of unobserved heterogeneity correlated with the covariates. To compare the estimates in (1), (3), and (4) to FE logit, we can look only at the ratio of the coefficients on *kids* and *lhinc*, which is 3.50 in column (5). In columns (1), (3), and (4) the ratios are 4.37, 4.03, and 4.06. Even if we think these differ substantially from the ratio in column (5), we cannot be sure if this is due to the parametric assumptions on $D(c_i|\mathbf{x}_i)$ used in the probit models or the conditional independence used by FE logit. Of course, both could be misspecified.

Generally, CMLE approaches are fragile to changes in the specification. For example, a natural extension is

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(a_i + \mathbf{x}_{it}\mathbf{b}_i), \tag{7.12}$$

where $\mathbf{b}_i$ is a vector of heterogeneous slopes with $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$; let $\alpha \equiv E(a_i)$. This extension of the standard unobserved effects logit model raises several issues. First, what do we want to estimate? Perhaps the partial effects at the mean values of the heterogeneity. But the APEs, or local average effects, are probably of more interest.

Nothing seems to be known about what the logit CMLE would estimate if applied to (7.12), where we assume $\boldsymbol{\beta} = \mathbf{b}_i$. On the other hand, if, say, $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$, a flexible binary response model with covariates $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ (and allowing sufficiently for changes over time) identifies the APEs – without the conditional serial independence assumption. The same is true of the extension to time-varying factor loads, $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i)$.

There are methods that allow estimation, up to scale, of the coefficients without even specifying the distribution of u_{it} in

$$y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]. \tag{7.13}$$

under strict exogeneity.conditional on c_i. Arellano and Honoré (2001) survey methods, including variations on Manski's maximum score estimator.

Estimation of parameters and APEs is much more difficult even in simple dynamic probit models. Consider

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, c_i) = P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i), \; t = 1, \ldots, T,$$

which combines correct dynamic specification with strict exogeneity of $\{\mathbf{z}_{it}\}$. For a dynamic probit model

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{7.14}$$

Treating the c_i as parameters to estimate causes inconsistency in β and ρ (although there is recent work by Woutersen and Fernández-Val that shows how to make the asymptotic bias of order $1/T^2$; see the next section). A simple analysis is available if we specify

$$c_i|\mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2) \tag{7.15}$$

Then

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{7.16}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$. Because a_i is independent of $(y_{i0}, \mathbf{z}_i)$, it turns out we can use standard random effects probit software, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. Easily get the average partial effects, too:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{7.17}$$

and take differences or derivatives with respect to elements of $(\mathbf{z}_t, y_{t-1})$. As before, the coefficients are multiplied by $(1 + \hat{\sigma}_a^2)^{-1/2}$. Of course, both the structural model and model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $Var(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic).

We apply this method to the Chay and Hyslop data and estimate a model for $\mathrm{P}(lfp_{it} = 1|kids_{it}, lhinc_{it}, lfp_{i,t-1}, c_i)$, where one lag of labor force participation is assumed to suffice for the dynamics and $\{(kids_{it}, lhinc_{it}) : t = 1, \ldots, T\}$ is assumed to be strictly exogenous conditional on c_i. Also, we include the time-constant variables *educ*, *black*, *age*, and age^2 and a full set of time-period dummies. (We start with five periods and lose one with the lag. Therefore, we estimate the model using four years of data.) We include among the

regressors the initial value, lfp_{i0}, $kids_{i1}$ through $kids_{i4}$, and $lhinc_{i1}$ through $lhinc_{i4}$. Estimating the model by RE probit gives $\hat{\rho} = 1.541$ (se = .067), and so, even after controlling for unobserved heterogeneity, there is strong evidence of state dependence. But to obtain the size of the effect, we compute the APE for lfp_{t-1}. The calculation involves averaging $\Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a) - \Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a)$ across all t and i; we must be sure to scale the original coefficients by $(1 + \hat{\sigma}_a^2)^{-1/2}$, where, in this application, $\hat{\sigma}_a^2 = 1.103$. The APE estimated from this method is about .259. In other words, averaged across all women and all time periods, the probability of being in the labor force at time t is about .26 higher if the women was in the labor force at time $t - 1$ than if she was not. This estimate controls for unobserved heterogeneity, number of young children, husband's income, and the woman's education, race, and age. (This APE estimate can be directly compared to a dynamic linear probability model estimated using, say, the Arellano and Bond (1991) method and its extensions.)

It is instructive to compare the APE with the estimate of a dynamic probit model that ignores c_i. In this case, we just use pooled probit of lfp_{it} on $1, kids_{it}, lhinc_{it}, lfp_{i,t-1} educ_i, black_i, age_i$, and age_i^2 and include a full set of period dummies. The coefficient on $lfp_{i,t-1}$ is 2.876 (se = .027), which is much higher than in the dynamic RE probit model. More importantly, the APE for state dependence is about .837, which is much higher than when heterogeneity is controlled for. Therefore, in this example, much of the persistence in labor force participation of married women is accounted for by the unobserved heterogeneity. There is still some state dependence, but its value is much smaller than a simple dynamic probit indicates.

Arellano and Carrasco (2003) use a different approach to estimate the parameters and

APEs in dynamic binary response models with only sequentially exogenous variables. Thus, their method applies to models with lagged dependent variables, but also other models where there made be feedback from past shocks to future covariates. (Their assumptions essentially impose serial conditional serial independence.) Rather than impose an assumption such as (7.15), they use a different approximation. Let $v_{it} = c_i + u_{it}$ be the composed error in $y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]$. Then, in the context of a probit model, they assume

$$v_{it}|\mathbf{w}_{it} \sim \text{Normal}(E(c_i|\mathbf{w}_{it}), \sigma_t^2) \tag{7.18}$$

where $\mathbf{w}_{it} = (\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, y_{i1}\mathbf{x}_{i1})$. The mean $E(c_i|\mathbf{w}_{it})$ is unrestricted (although, of course, they are linked across time by iterated expectations because $\mathbf{w}_{it} \subset \mathbf{w}_{i,t+1}$), but the shape of the distribution is assumed to be the same across t. Arellano and Carrasco discuss identification and estimation, and extensions to models with time-varying factor loads.

Honoré and Kyriazidou (2000) extend an idea of Chamberlain's (1993) and show how to estimate $\boldsymbol{\delta}$ and ρ in a logit model without distributional assumptions for c_i. They find conditional probabilities that do not depend on c_i but still depend on $\boldsymbol{\delta}$ and ρ. However, in the case with four time periods, $t = 0, 1, 2,$ and 3, the conditioning that removes c_i requires $\mathbf{z}_{i2} = \mathbf{z}_{i3}$. HK show how to use a local version of this condition to consistently estimate the parameters. The estimator is also asymptotically normal, but converges more slowly than the usual $\sqrt{N}$-rate.

The condition that $\mathbf{z}_{i2} - \mathbf{z}_{i3}$ has a distribution with support around zero rules out aggregate year dummies or even linear time trends. Plus, using only observations with $\mathbf{z}_{i2} - \mathbf{z}_{i3}$ in a neighborhood of zero results in much lost data. Finally, estimates of partial effects or average partial effects are not available.

While semiparametric approaches can be valuable to comparing parameter estimates with

more parametric approaches, such comparisons have limitations. For example, the coefficients on y_{t-1} in the dynamic logit model and the dynamic probit model are comparable only in sign; we cannot take the derivative with respect to y_{t-1} because it is discrete. Because we do not know where the evaluate the partial effects – that is, the values of c to plug in, or average out across the distribution of c_i, we cannot compare the magnitudes of the FE logit estimates with CRC approaches. We can compare the relative effects on the continuous elements in $\mathbf{z}_t$ based on partial derivatives. But even here, if we find a difference between semiparametric and parametric methods, is it because aggregate time effects were excluded in the semiparametric estimation or because the model of $D(c_i|y_{i0}, \mathbf{z}_i)$ was misspecified? Currently, we have no good ways of deciding. (Recently, Li and Zheng (2006) use Bayesian methods to estimate a dynamic Tobit model with unobserved heterogeneity, where they distribution of unosberved heterogeneity is an infinite mixture of normals. They find that all of the average partial effects are very similar to those obtained from the much simpler specification in (7.15).)

Honoré and Lewbel (2002) show how to estimate β in the model

$$y_{it} = 1[v_{it} + \mathbf{x}_{it}\beta + c_i + u_{it} \geq 0] \tag{7.19}$$

without distributional assumptions on $c_i + u_{it}$. The special continuous explanatory variable v_{it}, which need not be time varying, is assumed to appear in the equation (and its coefficient is normalized to one). More importantly, v_{it} is assumed to satisfy $D(c_i + u_{it}|v_{it}, \mathbf{x}_{it}, \mathbf{z}_i) = D(c_i + u_{it}|\mathbf{x}_{it}, \mathbf{z}_i)$, which is a conditional independence assumption. The vector $\mathbf{z}_i$ is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, z_i could be functions of variables determined prior to the earliest time period.) The most likely scenario is when v_{it} is randomized and therefore independent of $(\mathbf{x}_{it}, \mathbf{z}_i, e_{it})$, where $e_{it} = c_i + u_{it}$. It seems unlikely to hold if v_{it} is related to past outcomes on y_{it}. The estimator

derived by Honoré and Lewbel is $\sqrt{N}$-asymptotically normal, and fairly easy to compute; it requires estimation of the density of v_{it} given $(\mathbf{x}_{it}, \mathbf{z}_i)$ and then a simple IV estimation.

Honoré and Tamer (2006) have recently shown how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model; these are covered in the notes on partial identification.

Very similar analysis hold for ordered probit models. See Wooldridge (2010, Chapter 15) for the static case and Wooldridge (2005b) for the dynamic case. The dependence of heterogeneity on the initial condition can be made flexible while keeping the likelihood in the class of random effects ordered probit models.

7.2 Count and Other Multiplicative Models

Several options are available for models with conditional means multiplicative in the heterogeneity. The most common is

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.20}$$

where $c_i \geq 0$ is the unobserved effect and x_{it} would incude a full set of year dummies in most cases. First consider estimation under strict exogeneity (conditional on c_i):

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i). \tag{7.21}$$

If we add independence between c_i and x_i – a random effects approach – then, using $E(c_i) = 1$ as a normalization,

$$E(y_{it}|\mathbf{x}_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}), \tag{7.22}$$

and various estimation methods can be used to account for the serial dependence in $\{y_{it}\}$ if only x_i is conditioned on. (Serial correlation is certainly present because of c_i, but it could be

present due to idiosyncratic shocks, too.) Regardless of the actual distribution of y_{it}, or even its nature – other than $y_{it} \geq 0$ – the pooled Poisson quasi-MLE is consistent for β under (7.22) but likely very inefficient; robust inference is straightforward with small T and large N.

Random effects Poisson requires that $D(y_{it}|\mathbf{x}_i, c_i)$ has a Poisson distribution with mean (7.20), and maintains the conditional independence assumption,

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it}, c_i),$$

along with a specific distribution for c_i – usually a Gamma distribution with unit mean. Unfortunately, like RE probit, the full MLE has no known robustness properties. The Poisson distribution needs to hold along with the other assumptions. A generalized estimating approach is available, too. If the Poisson quasi-likelihood is used, the GEE estimator is fully robust provided the mean is correctly specified. One can use an exchangeable, or at least constant, working correlation matrix. See Wooldridge (2010, Chapter 18).

A CRE model can be allowed by writing $c_i = \exp(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi})a_i$ where a_i is independent of x_i with unit mean. Then

$$E(y_{it}|\mathbf{x}_i) = \exp(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}) \tag{7.23}$$

and now the same methods described above can be applied but with $\bar{x}_i$ added as regressors. This approach identifies average partial effects. In fact, we could use Altonji and Matzkin (2005) and specify $E(c_i|x_i) = h(\bar{\mathbf{x}}_i)$ (say), and then estimate the semiparametric model $E(y_{it}|\mathbf{x}_i) = h(\bar{\mathbf{x}}_i)\exp(\mathbf{x}_{it}\boldsymbol{\beta}) = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + g(\bar{\mathbf{x}}_i))$ where $g(\bar{\mathbf{x}}_i) = \log[h(\bar{\mathbf{x}}_i)]$ is also unrestricted. Other features of the series $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as individual-specific trends or sample variances, can be added to $h(\cdot)$.

An important estimator that can be used under just

$$E(y_{it}|\mathbf{x}_i, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.24}$$

is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption. It is often called the fixed effects Poisson estimator, and, in fact, $\hat{\boldsymbol{\beta}}$ turns out to be identical to using pooled Poisson QMLE and treating the c_i as parameters to estimate. (A rare case, like the linear model, where this does not result in an incidental parameters problem.). It is easy to obtain fully robust inference, too (although it is not currently part of standard software, such as Stata). The fact that the quasi-likelihood is derived for a particular, discrete distribution appears to make people queasy about using it, but it is analogous to using the normal log-likelihood in the linear model: the resulting estimator, the usual FE estimator, is fully robust to nonnormality, heteroskedasticity, and serial correlation. See Wooldridge (1999).

Estimation of models under sequential exogeneity has been studied by Chamberlain (1992) and Wooldridge (1997). In particular, they obtain moment conditions for models such as

$$E(y_{it}|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}). \tag{7.25}$$

Under this assumption, it can be shown that

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}] = 0, \tag{7.26}$$

and, because these moment conditions depend only on observed data and the parameter vector $\boldsymbol{\beta}$, GMM can be used to estimate $\boldsymbol{\beta}$, and fully robust inference is straightforward.

Blundell, Griffiths, and Windmeijer (2002) consider a model with additive heterogeneity and a lagged dependent variable that appears linearly, and derive estimating equations.

The moment conditions in (7.26) involve the differences $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$, and we saw for the

linear model that, if elements of $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$ are persistent, IV and GMM estimators can be badly biased and imprecise. If we make more assumptions, models with lagged dependent variables and other regressors that are strictly exogenous can be handled using the conditional MLE approach in Section 6. Wooldridge (2005b) shows how a dynamic Poisson model with conditional Gamma heterogeneity can be easily estimated.

8. Estimating the Fixed Effects

It is well known that, except in special cases (linear and Poisson), treating the c_i as parameters to estimate leads to inconsistent estimates of the common parameters $\boldsymbol{\theta}$. But two questions arise. First, are there ways to adjust the "fixed effects" estimate of $\boldsymbol{\theta}$ to at least partially remove the bias? Second, could it be that estimates of the average partial effects, based generally on

$$N^{-1}\sum_{i=1}^{N}\frac{\partial m_t(\mathbf{x}_t,\hat{\boldsymbol{\theta}},\hat{\mathbf{c}}_i)}{\partial x_{tj}}, \tag{8.1}$$

where $m_t(\mathbf{x}_t,\boldsymbol{\theta},\mathbf{c}) = E(y_t|\mathbf{x}_t,\mathbf{c})$, are better behaved than the parameter estimates, and can their bias be removed? In the unobserved effects probit model, (8.1) becomes

$$N^{-1}\sum_{i=1}^{N}\hat{\beta}_j\phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}+\hat{c}_i), \tag{8.2}$$

which is easy to compute once $\hat{\boldsymbol{\beta}}$ and the $\hat{c}_i$ (N of them) have been obtained.

Hahn and Newey (2004) propose both jackknife and analytical bias corrections for the parameters and show that they work well for estimating the parameters in the probit model. Generally, the jackknife procedure to remove the bias in $\hat{\boldsymbol{\theta}}$ is simple but can be

computationally intensive. The idea is this. The estimator based on T time periods has probability limit that can be written as

$$\boldsymbol{\theta}_T = \boldsymbol{\theta} + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \tag{8.3}$$

for vectors $\mathbf{b}_1$ and $\mathbf{b}_2$. Now, let $\hat{\boldsymbol{\theta}}_{(t)}$ denote the estimator that drops time period t. Then, assuming stability across t, the plim of $\hat{\boldsymbol{\theta}}_{(t)}$ is

$$\boldsymbol{\theta}_{(t)} = \boldsymbol{\theta} + \mathbf{b}_1/(T-1) + \mathbf{b}_2/(T-1)^2 + O(T^{-3}). \tag{8.4}$$

It follows that

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} (T\hat{\boldsymbol{\theta}} - (T-1)\hat{\boldsymbol{\theta}}_{(t)}) &= (T\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/T) - [(T-1)\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/(T-1)] + O(T^{-3}) \\ &= \boldsymbol{\theta} - \mathbf{b}_2/[T(T-1)] + O(T^{-3}) = \boldsymbol{\theta} + O(T^{-2}).\end{aligned} \tag{8.5}$$

If, for given heterogeneity c_i, the data are independent and identically distributed across t, then (8.5) holds for all leave-one-time-period-out estimators, so we use the average of all such estimators in computing the panel jackknife estimator:

$$\tilde{\boldsymbol{\theta}} \equiv T\hat{\boldsymbol{\theta}} - (T-1)T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\theta}}_{(t)}. \tag{8.6}$$

From the argument above, theasymptotic bias of $\tilde{\boldsymbol{\theta}}$ is on the order of T^{-2}.

Unfortunately, there are some practical limitations to the jackknife procedure, as well as to the analytical corrections derived by Hahn and Newey. First, aggregate time effects are not allowed, and they would be very difficult to include because the analysis is with $T \to \infty$. (In other words, they would introduce an incidental parameters problem in the time dimension as well as the cross section dimension.) Generally, heterogeneity in the distributions across t changes the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ when a time period is dropped, and so the simple transformation in (8.5) does not remove the bias terms. Second, Hahn and Newey assume

independence across t conditional on c_i. It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, as noted by Hahn and Keursteiner (2002), applying the "leave-one-out" method to dynamic models is problematical because the $\mathbf{b}_1$ and $\mathbf{b}_2$ in (8.4) would depend on t so, again, the transformation in (8.5) will not eliminate the $\mathbf{b}_1$ term.

Recently, Dhaene, Jochmans, and Thuysbaert (2006) propose a modification of the Hahn-Newey procedure that appears promising for dynamic models. In the simplest case, in addition to the "fixed effects" estimator using all time periods, they obtain estimators for two subperiods: one uses the earlier time periods, one uses later time periods, and they have some overlap (which is small as T gets large). Unfortunately, the procedure still requires stationarity and rules out aggregate time effects.

For the probit model, Fernández-Val (2007) studies the properties of estimators and average partial effects and allows time series dependence in the strictly exogenous regressors. Interestingly, in the probit model with exogenous regressors under the conditional independence assumption, the estimates of the APEs based on the "fixed effects" estimator has bias of order T^{-2} in the case that there is no heterogeneity. Unfortunately, these findings do not carry over to models with lagged dependent variables, and the bias corrections in that case are difficult to implement (and still do not allow for time heterogeneity).

As the resurgent literature on "fixed effects" approaches stands, there is still a tradeoff in the assumptions when compared with the correlated random effecst approach. The FE approach allows $D(\mathbf{c}_i|\mathbf{x}_i)$ to be unrestricted, but, currently, the corrections to the parameter estimates and partial effects impose stationarity across time and restricts the time dependence, often in very restrictive ways (such as serial independence). The CRE approach restricts

$D(\mathbf{c}_i|\mathbf{x}_i)$ but, because it can be applied for small T, does not restrict nonstationarity or serial dependence in the time series dimension. With recent advances such as those in Altonji and Matzkin (2005) that impose weak restrictions on $D(\mathbf{c}_i|\mathbf{x}_i)$, the CRE approach continues to be attractive, particularly because it identifies average partial effects. Generally, the FE and CRE approaches should be viewed as being complementary.

One final comment. The CRE approach has only been fully worked out in the case of balanced panels. When we introduce a set of sample selection indicators for each i, $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iT})$, where $s_{it} = 1$ if $(\mathbf{x}_{it}, \mathbf{y}_{it})$ is osberved, the CRE method requires us to model $D(\mathbf{c}_i|\mathbf{x}_i, \mathbf{s}_i)$. It may still make sense, in some cases, to assume exchangeability, so that, say, $D(\mathbf{c}_i|\mathbf{x}_i, \mathbf{s}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$, where $\bar{\mathbf{x}}_i = T_i^{-1} \sum_{t=1}^{T} s_{it}\mathbf{x}_{it}$ is the average using the selected sample, but this possibility has not been explored. By contrast, provided selection is strictly exogenous conditional on $(\mathbf{x}_i, \mathbf{c}_i)$ – see the notes on missing data – the FE procedure on the unbalanced panel is fundamentally unchanged. (However, the jackknife corrections discussed above would no long be valid with an unbalanced panel.) The properties and merits of FE and CRE approaches using unbalanced panels needs to be explored in future research.

References

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

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

Arellano, M. and R. Carrasco (2003), "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics* 115, 125-157.

Arellano, M. and B. Honoré (2001), "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, Volume 5, ed. J.J. Heckman and E. Leamer. Amsterdam: North Holland, 3229-3296.

Blundell, R., R. Griffith, and F. Windmeijer (2002), "Individual Effects and Dynamics in Count Data Models," *Journal of Econometrics* 108, 113-131.

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

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

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

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

Chamberlain, G. (1992), "Sequential Moment Restrictions in Panel Data: Comment," *Journal of Business and Economic Statistics* 10, 20-26.

Chamberlain, G. (1993), "Feedback in Panel Data Models," Harvard Institute of Economic Research Discussion Paper No. 1656.

Chay, K.Y. and D. Hyslop (2001), ""Identification and Estimation of Dynamic Binary Response Panel Data Models: Empirical Evidence using Alternative Approaches," mimeo, U.C. Berkeley Department of Economics.

Dhaene, G. K. Jochmans, and B. Thuysbaert (2006), "Jackknife Bias Reduction for Nonlinear Dynamic Panel Data Models with Fixed Effects," mimeo, Katholieke Universiteit Leuven Department of Economics.

Fernández-Val, I. (2007), "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," mimeo, Boston University Department of Economics.

Florens, J.P., J.J. Heckman, C. Meghir, and E. Vytlacil (2007), "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous

Treatment and Heterogeneous Effects," mimeo, Columbia University Department of Economics.

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

Hahn, J. and G. Kuersteiner (2002), "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T Are Large," *Econometrica* 70, 1639-1657.

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

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

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

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

Honoré, B.E. and A. Lewbel (2002), "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica* 70, 2053-2063.

Honoré, B.E. and E. Tamer (2006), "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica* 74, 611-629.

Li, T. and X. Zheng (2006), "Semiparametric Bayesian Inference for Dynamic Tobit Panel Data Models with Unobserved Heterogeneity," mimeo, Vanderbilt University Department of Economics.

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

Newey, W.K. (1988), "Adaptive Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics* 38, 301-339.

Wooldridge, J.M. (1997), "Multiplicative Panel Data Models without the Strict Exogeneity Assumption," *Econometric Theory*, 13, 667-678.

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

Wooldridge, J.M. (2000), "A Framework for Estimating Dynamic, Unobserved Effects Panel Data Models with Possible Feedback to Future Explanatory Variables," *Economics Letters* 68, 245-250.

Wooldridge, J.M. (2005a), "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and*

Statistics 87, 385-390.

Wooldridge, J.M. (2005b), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics* 20, 39-54.

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

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. MIT Press: Cambridge, MA.

Cross-Section Econometrics
Lecture 6: Nonlinear Panel Data Models

Jeff Wooldridge
Michigan State University

AEA Lectures, Chicago, January 2012

1. Introduction: Why Nonlinear Models?

- We usually apply nonlinear models when y_{it} is a limited dependent variable, such as a binary variable, a corner solution, and so on. Here we assume the outcome y_{it} is what we want to measure (no data censoring).
- With observed covariates $\mathbf{x}_{it}$ and unobserved heterogeneity $\mathbf{c}_i$, we are interested in the response probability:

$$P(y_{it} = 1|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = p_t(\mathbf{x}_t, \mathbf{c})$$

- We can define quantities of interest – partial effects – without specifying a model.

- Suppose we use a standard linear unobserved effects model for binary y_{it},

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

under strict exogeneity (conditional on c_i),

$$E(u_{it}|\mathbf{x}_i, c_i) = 0$$

- Then

$$P(y_{it} = 1|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\beta} + c_i$$

- The linear probability model (LPM) is simple to interpret: β_j directly measures the partial effect of x_{tj} on $P(y_{it} = 1|\mathbf{x}_{it}, c_i)$.
- If we believe the linearity and strict exogeneity assumptions, we can estimate $\boldsymbol{\beta}$ by FE or FD.

• Recall that FE allows any relationship between c_i and $\mathbf{x}_{it}$ and we can have any kind of serial correlation in $\{u_{it}\}$ (and u_{it} is necessarily heteroskedastic in the LPM).

• Nonlinear methods we use will be more restrictive. For example, we will have to model the conditional distribution $D(c_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, or assume a serial independence assumption, or both.

• So what is wrong with standard FE estimation of a linear model?

• There is only one problem with the LPM: the linear functional form for $P(y_{it} = 1|\mathbf{x}_{it}, c_i)$ usually cannot be literally true.

• If we take the LPM literally, we must have

$$0 \leq \mathbf{x}_{it}\boldsymbol{\beta} + c_i \leq 1, \text{ all } \mathbf{x}_{it}$$

which puts strange restrictions on $D(c_i|\mathbf{x}_i)$.

• Rather than treat the LPM as the true model for $P(y_{it} = 1|\mathbf{x}_{it}, c_i)$, we should treat it as a linear approximation. The β_j estimated by (say) FE can give reasonable estimates of average partial effects (to be defined precisely).

• The main value of nonlinear models is to study how the partial effects change with the covariates and, possibly, heterogeneity.

• For emphasis: Talking about "biased and inconsistent" estimation of "parameters" from using an LPM is off point. A valid criticism is that the linear model may poorly approximate $P(y_{it} = 1|\mathbf{x}_{it} = \mathbf{x}_t, c_i = c)$ over a wide range of values $(\mathbf{x}_t, c)$.

• An alternative to the LPM is the unobserved effects probit model:

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

• If x_{tj} is continuous, its partial effect is

$$\frac{\partial P(y_{it} = 1|\mathbf{x}_{it} = \mathbf{x}_t, c_i = c)}{\partial x_{tj}} = \beta_j \phi(\mathbf{x}_t\boldsymbol{\beta} + c). \tag{2}$$

• If x_{tj} and x_{th} are both continuous:

$$\frac{\beta_j \phi(\mathbf{x}_t\boldsymbol{\beta} + c)}{\beta_h \phi(\mathbf{x}_t\boldsymbol{\beta} + c)} = \frac{\beta_j}{\beta_h}$$

But this still does not tell us the absolute size of each effect.

• Can look at discrete changes, too:

$$\Phi(\mathbf{x}_t^{(1)}\boldsymbol{\beta} + c) - \Phi(\mathbf{x}_t^{(0)}\boldsymbol{\beta} + c)$$

where $\mathbf{x}_t^{(0)}$ and $\mathbf{x}_t^{(1)}$ are set at different values. Again, this depends on c (as well as the values of the covariates).

• Questions: (i) Assuming we can estimate $\boldsymbol{\beta}$, what should we do about the unobservable c? (ii) If we can only estimate $\boldsymbol{\beta}$ up-to-scale, can we still learn something useful about magnitudes of partial effects? (iii) What kinds of assumptions do we need to estimate partial effects?

2. General Setup and Quantities of Interest

• Let $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$ be a random draw from the cross section. Suppose we are interested in

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i). \tag{3}$$

$\mathbf{c}_i$ can be a vector of unobserved heterogeneity.

• Partial effects: if x_{tj} is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}, \tag{4}$$

or discrete changes.

• How do we account for unobserved $\mathbf{c}_i$? If we know enough about the distribution of $\mathbf{c}_i$ we can insert meaningful values for $\mathbf{c}$. For example, if $\boldsymbol{\mu}_\mathbf{c} = E(\mathbf{c}_i)$, then we can compute the *partial effect at the average (PEA)*,

$$PEA_j(\mathbf{x}_t) = \theta_j(\mathbf{x}_t, \boldsymbol{\mu}_\mathbf{c}). \tag{5}$$

Of course, we need to estimate the function m_t and $\boldsymbol{\mu}_\mathbf{c}$. If we can estimate the distribution of $\mathbf{c}_i$, or features in addition to its mean, we can insert different quantiles, or a certain number of standard deviations from the mean.

• Alternatively, we can obtain the *average partial effect* (APE) (or *population average effect*) by averaging across the distribution of $\mathbf{c}_i$:

$$APE(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)]. \quad (6)$$

The difference between (5) and (6) can be nontrivial. In some leading cases, (6) is identified while (5) is not. (6) is closely related to the notion of the *average structural function* (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$ASF(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \quad (7)$$

• Passing the derivative through the expectation in (7) gives the APE.

• How do APEs relate to parameters? Index model:

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \quad (8)$$

where $G(\cdot)$ is differentiable. Then

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \quad (9)$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Even if $G(\cdot)$ is known, magnitude of effects cannot be estimated without making assumptions about the distribution of c_i

• Important: Definitions of partial effects do not depend on whether $\mathbf{x}_t$ is correlated with $\mathbf{c}$. Of course, whether and how we estimate them certainly does.

3. Exogeneity Assumptions

• Cannot get by with only specifying a model for the contemporaneous conditional distribution, $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$.

• The most useful definition of strict exogeneity for nonlinear panel data models is *strict exogeneity conditional on the unobserved effects* $\mathbf{c}_i$ [Chamberlain (1984)]:

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (10)$$

Conditional mean version (for quasi-MLE):

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (11)$$

• The *sequential exogeneity* assumption is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (12)$$

Unfortunately, it is much more difficult to allow sequential exogeneity in in nonlinear models. (Most progress for lagged dependent variables or specific functional forms, such as exponential.)

• Neither strict nor sequential exogeneity allows for contemporaneous endogeneity of one or more elements of $\mathbf{x}_{it}$, where, say, x_{itj} is correlated with unobserved, time-varying unobservables that affect y_{it}.

4. Conditional Independence

- In linear models, serial dependence of idiosyncratic shocks is easily dealt with, either by "cluster robust" inference or Generalized Least Squares extensions of Fixed Effects and First Differencing.
- With strictly exogenous covariates, serial correlation never results in inconsistent estimation, even if improperly modeled. The situation is different with most nonlinear models estimated by MLE.

- *Conditional independence* (CI) (under strict exogeneity):

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \tag{13}$$

- In a parametric context, the CI assumption reduces our task to specifying a model for $D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$.

- In random effects and correlated random frameworks (next section), CI plays a critical role in being able to estimate the "structural" parameters and the parameters in the distribution of $\mathbf{c}_i$ (and therefore, in estimating PEAs). In a broad class of popular models, CI plays no essential role in estimating APEs.

5. Assumptions about the Unobserved Heterogeneity

Random Effects

- Generally stated, the key RE assumption is

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{14}$$

Under (14), the APEs are nonparametrically identified from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{15}$$

- In leading cases (RE probit and RE Tobit with heterogeneity normally distributed), if we want PEs for different values of $\mathbf{c}$, we must assume more: strict exogeneity, conditional independence, and (14) with a parametric distribution for $D(\mathbf{c}_i)$.

Correlated Random Effects

- Allow dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but restricted in some way.
- Can specify a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$, as in Chamberlain (1980,1982), and much work since.
- Distributional assumptions that lead to simple estimation – homoskedastic normal with a linear conditional mean — can be restrictive.

- Possible to drop parametric assumptions and just assume

$$D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i), \tag{16}$$

without restricting $D(c_i|\bar{\mathbf{x}}_i)$. Altonji and Matzkin (2005, Econometrica).
- Other functions of $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$ are possible.

- APEs are identified very generally. Under $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$, a consistent estimate of the average structural function is

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N} m_t(\mathbf{x}_t,\bar{\mathbf{x}}_i), \tag{17}$$

where $m_t(\cdot)$ is the mean function $E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i)$.
- Need a random sample $\{\bar{\mathbf{x}}_i : i = 1,\ldots,N\}$ for the averaging out to work.

Fixed Effects

- The label "fixed effects" is used in different ways by different researchers. One view: $\mathbf{c}_i$, $i = 1,\ldots,N$ are unit-specific parameters to be estimated. Usually leads to an "incidental parameters problem" for the population parameters.
- Ongoing research: How can the bias be adjusted (as a function of T)? Also, behavior of partial effects.

• Second meaning of "fixed effects": $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted and we look for objective functions that do not depend on $\mathbf{c}_i$ but still identify the population parameters. Leads to "conditional MLE" if we can find "sufficient statistics" $\mathbf{s}_i$ such that

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{s}_i). \tag{18}$$

• Conditional Independence is usually maintained for CMLE.

• PEAs and APEs are generally unidentified using CMLE.

6. Dynamic Models

• Model with lagged dependent variables and other strictly exogenous variables:

$$D(\mathbf{y}_{it}|\mathbf{z}_i,\mathbf{y}_{i,t-1},\ldots,\mathbf{y}_{i1},\mathbf{y}_{i0},\mathbf{c}_i) = D(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\ldots,\mathbf{z}_{i1},\mathbf{y}_{i0},\mathbf{c}_i) \tag{19}$$

• Suppose this distribution depends only on $(\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c}_i)$ with density $f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta})$. The joint density of $(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ given $(\mathbf{y}_{i0},\mathbf{z}_i,\mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}). \tag{20}$$

• How do we deal with $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$? Simple approach studied by Wooldridge (2005, Journal of Applied Econometrics): Model $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$ directly.

• Leads to a model for $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{y}_{i0},\mathbf{z}_i)$ and MLE conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$. This can be computationally simple for popular models, and can be made somewhat flexible in $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$.

• The APEs for the conditional means (and therefore conditional probabilities) are easy to obtain.

7. Estimating Popular Models

• Pooled and random effects estimation commands in Stata (for probit, Tobit, Poisson, GLM, GEE) often can be used.

• Stata `egen` command for generating time averages. Need leads and lags of exogenous variables, and the initial condition, for dynamic models.

• For pooled methods, use the "panel bootstrap" feature in Stata to obtain standard errors or confidence intervals.

• Computational time is an issue for dynamic models because it uses full "random effects" with lots of covariates.

Example: **Binary and Fractional Response**

• Unobserved effects (UE) "probit" model for a binary or fractional y_{it}:

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \; t = 1, \ldots, T. \tag{26}$$

Assume strict exogeneity (conditional on c_i) and Chamberlain-Mundlak device:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, \; a_i|\mathbf{x}_i \sim Normal(0, \sigma_a^2). \tag{27}$$

• In binary response case under (conditional) serial independence, all parameters are identified and MLE (Stata: `xtprobit`) can be used. Just add the time averages $\bar{\mathbf{x}}_i$ as an additional set of regressors. Gives $\hat{\boldsymbol{\beta}}$, $\hat{\psi}$, $\hat{\boldsymbol{\xi}}$, and $\hat{\sigma}_a^2$.

• Then

$$\hat{\mu}_c = \hat{\psi} + \bar{\mathbf{x}}\hat{\boldsymbol{\xi}} \xrightarrow{p} \mu_c, \text{ where } \bar{\mathbf{x}} = N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i$$

$$\hat{\sigma}_c^2 = \hat{\boldsymbol{\xi}}'\left[N^{-1}\sum_{i=1}^{N}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right]\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2 \xrightarrow{p} \sigma_c^2$$

• Can evaluate PEs at, say, $\hat{\mu}_c \pm k\hat{\sigma}_c$ for values k.

• The APEs are identified from the ASF, estimated as

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{28}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1 + \hat{\sigma}_a^2)^{1/2}$.

• For binary or fractional response (and in general), APEs are identified without the conditional serial independence assumption. Use pooled Bernoulli quasi-MLE (Stata: `glm`) or generalized estimating equations (Stata: `xtgee`) to estimate scaled coefficients based on

$$E(y_{it}|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a). \tag{29}$$

(Time dummies supressed for simplicity.)

• A more radical suggestion, in the spirit of Altonji and Matzkin (2005): Just use a flexible model for $E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly, say,

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}]. \quad (30)$$

• Average out over $\bar{\mathbf{x}}_i$ to get APEs.

• In the binary response case, we can use a conditional MLE if we start with

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \quad (31)$$

and assume strict exogeneity and conditional independence.

• Can estimate $\boldsymbol{\beta}$ by FE logit without restricting $D(c_i|\mathbf{x}_i)$, but conditional independence is necessary.

• **Example**: Married Women's Labor Force Participation: $N = 5,663$, $T = 5$ (four month intervals). LFP.DTA in Wooldridge (2010, MIT Press).

• Following results include a full set of time period dummies (not reported).

• The APEs are directly comparable across models, and can be compared with the linear model coefficients.

LFP	(1)	(2)		(3)		(4)		(5)
Model	Linear	Probit		CRE Probit		CRE Probit		FE Logit
Est. Method	FE	Pooled MLE		Pooled MLE		MLE		MLE
	Coef.	Coef.	APE	Coef.	APE	Coef.	APE	Coef.
kids	−.0389	−.199	−.0660	−.117	−.0389	−.317	−.0403	−.644
	(.0092)	(.015)	(.0048)	(.027)	(.0085)	(.062)	(.0104)	(.125)
lhinc	−.0089	−.211	−.0701	−.029	−.0095	−.078	−.0099	−.184
	(.0046)	(.024)	(.0079)	(.014)	(.0048)	(.041)	(.0055)	(.083)
$\overline{kids}$	—	—	—	−.086	—	−.210	—	—
	—	—	—	(.031)	—	(.071)	—	—
$\overline{lhinc}$	—	—	—	−.250	—	−.646	—	—
	—	—	—	(.035)	—	(.079)	—	—

```
. des lfp kids hinc

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
lfp             byte   %9.0g                  =1 if in labor force
kids            byte   %9.0g                  number children < 18
hinc            float  %9.0g                  husband's monthly income, $

. tab period

  1 through |
  5, each 4 |
months long |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      5,663       20.00       20.00
          2 |      5,663       20.00       40.00
          3 |      5,663       20.00       60.00
          4 |      5,663       20.00       80.00
          5 |      5,663       20.00      100.00
------------+-----------------------------------
      Total |     28,315      100.00

. egen kidsbar = mean(kids), by(id)

. egen lhincbar = mean(lhinc), by(id)
```



```
. * Linear model by FE:

. xtreg lfp kids lhinc per2-per5, fe cluster(id)

Fixed-effects (within) regression               Number of obs      =     28315
Group variable (i): id                          Number of groups   =      5663


                                    (Std. Err. adjusted for 5663 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.0388976   .0091682    -4.24   0.000    -.0568708   -.0209244
       lhinc |  -.0089439   .0045947    -1.95   0.052    -.0179513    .0000635
        per2 |  -.0042799    .003401    -1.26   0.208    -.0109472    .0023875
        per3 |  -.0108953   .0041859    -2.60   0.009    -.0191012   -.0026894
        per4 |  -.0123002   .0044918    -2.74   0.006    -.0211058   -.0034945
        per5 |  -.0176797   .0048541    -3.64   0.000    -.0271957   -.0081637
       _cons |   .8090216   .0375234    21.56   0.000     .7354614    .8825818
-------------+----------------------------------------------------------------
     sigma_u |  .42247488
     sigma_e |  .21363541
         rho |  .79636335   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```



```
. * Fixed Effects Logit:

. xtlogit lfp kids lhinc per2-per5, fe

Conditional fixed-effects logistic regression   Number of obs      =      5275
Group variable: id                              Number of groups   =      1055

                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

                                                LR chi2(6)         =     57.27
Log likelihood  = -2003.4184                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.6438386   .1247828    -5.16   0.000    -.8884084   -.3992688
       lhinc |  -.1842911   .0826019    -2.23   0.026    -.3461878   -.0223943
        per2 |  -.0928039   .0889937    -1.04   0.297    -.2672283    .0816205
        per3 |  -.2247989   .0887976    -2.53   0.011     -.398839   -.0507587
        per4 |  -.2479323   .0888953    -2.79   0.005     -.422164   -.0737006
        per5 |  -.3563745   .0888354    -4.01   0.000    -.5304886   -.1822604
------------------------------------------------------------------------------

. di 644/184
3.5

. di 389/89
4.3707865
```



```
. * CRE probit:

. xtprobit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5, re

Random-effects probit regression                Number of obs      =     28315
Group variable (i): id                          Number of groups   =      5663

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.3174051     .06203    -5.12   0.000    -.4389816   -.1958287
       lhinc |  -.0777949   .0414033    -1.88   0.060    -.1589439    .0033541
     kidsbar |  -.2098409   .0708676    -2.96   0.003    -.3487389   -.0709429
    lhincbar |  -.6463674   .0792719    -8.15   0.000    -.8017374   -.4909974
        educ |    .221596   .0147891    14.98   0.000     .1926099    .2505821
       black |   .5226558   .1502331     3.48   0.001     .2282042    .8171073
         age |   .4036543   .0287538    14.04   0.000     .3472979    .4600107
       agesq |  -.0054898   .0003536   -15.52   0.000    -.0061829   -.0047966
        per2 |   -.034359   .0438562    -0.78   0.433    -.1203156    .0515976
        per3 |  -.0954482   .0439688    -2.17   0.030    -.1816253    -.009271
        per4 |  -.1046944   .0439108    -2.38   0.017    -.1907581   -.0186308
        per5 |  -.1559446   .0435241    -3.58   0.000    -.2412502   -.0706389
       _cons |  -2.080352   .6567295    -3.17   0.002    -3.367518   -.7931854
-------------+----------------------------------------------------------------
    /lnsig2u |    1.73677   .0266277                     1.684581     1.78896
-------------+----------------------------------------------------------------
     sigma_u |   2.383059   .0317277                     2.321679    2.446063
         rho |   .8502764   .0033899                     .8435102    .8567997
------------------------------------------------------------------------------
```

```
. predict xdhat, xb

. gen xdhata = xdhat/sqrt(1 + 2.383059^2)

. di 1/sqrt(1 + 2.383059^2)
.38694144

. * Scaled coefficients to compare with pooled probit:

. di (1/sqrt(1 + 2.383059^2))*_b[kids]
-.1228172

. di (1/sqrt(1 + 2.383059^2))*_b[lhinc]
-.03010209
```



```
. probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5,
        cluster(id)

Probit regression                                 Number of obs   =      28315
                                                  Wald chi2(12)   =     538.09
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -16516.436                 Pseudo R2       =     0.0673

                                   (Std. Err. adjusted for 5663 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1173749   .0269743    -4.35   0.000    -.1702435   -.0645064
       lhinc |  -.0288098    .014344    -2.01   0.045    -.0569234   -.0006961
     kidsbar |  -.0856913   .0311857    -2.75   0.006     -.146814   -.0245685
    lhincbar |  -.2501781   .0352907    -7.09   0.000    -.3193466   -.1810097
        educ |   .0841338   .0067302    12.50   0.000     .0709428    .0973248
       black |   .2030668   .0663945     3.06   0.002     .0729359    .3331976
         age |   .1516424   .0124831    12.15   0.000      .127176    .1761089
       agesq |  -.0020672   .0001553   -13.31   0.000    -.0023717   -.0017628
        per2 |  -.0135701   .0103752    -1.31   0.191    -.0339051    .0067648
        per3 |  -.0331991   .0127197    -2.61   0.009    -.0581293    -.008269
        per4 |  -.0390317   .0136244    -2.86   0.004    -.0657351   -.0123284
        per5 |  -.0552425   .0146067    -3.78   0.000    -.0838711   -.0266139
       _cons |  -.7260562   .2836985    -2.56   0.010    -1.282095   -.1700173
------------------------------------------------------------------------------
```



```
. drop xdhat xdhata

. predict xdhat, xb

. gen scale = normden(xdhat)

. sum scale

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       scale |     28315    .3310079     .057301   .0694435   .3989423

. di .331*(-.117375)
-.03885113

. di .331*(-.02881)
-.00953611
```



```
. margeff

Average marginal effects on Prob(lfp==1) after probit

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |   -.038852   .0089243    -4.35   0.000    -.0563433   -.0213608
       lhinc |  -.0095363   .0047482    -2.01   0.045    -.0188426     -.00023
     kidsbar |  -.0283645   .0102895    -2.76   0.006    -.0485315   -.0081974
    lhincbar |  -.0828109   .0115471    -7.17   0.000    -.1054428    -.060179
        educ |    .027849   .0021588    12.90   0.000     .0236178    .0320801
       black |   .0643443   .0200207     3.21   0.001     .0251043    .1035842
         age |   .0501948   .0039822    12.60   0.000     .0423898    .0579998
       agesq |  -.0006843   .0000493   -13.88   0.000    -.0007809   -.0005876
        per2 |  -.0044999   .0034482    -1.30   0.192    -.0112583    .0022585
        per3 |  -.0110375   .0042512    -2.60   0.009    -.0193698   -.0027052
        per4 |  -.0129865   .0045606    -2.85   0.004    -.0219252   -.0040479
        per5 |  -.0184197   .0049076    -3.75   0.000    -.0280385    -.008801
------------------------------------------------------------------------------
```

```
. probit lfp kids lhinc educ black age agesq per2-per5, cluster(id)

Probit regression                                 Number of obs   =      28315
                                                  Wald chi2(10)   =     537.36
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -16556.671                 Pseudo R2       =     0.0651

                                  (Std. Err. adjusted for 5663 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1989144   .0153153   -12.99   0.000    -.2289319   -.1688969
       lhinc |  -.2110739   .0242901    -8.69   0.000    -.2586816   -.1634661
        educ |   .0796863   .0065453    12.17   0.000     .0668577    .0925149
       black |   .2209396   .0659041     3.35   0.001        .09177   .3501093
         age |   .1449159   .0122179    11.86   0.000     .1209693    .1688624
       agesq |  -.0019912   .0001522   -13.08   0.000    -.0022895   -.0016928
        per2 |  -.0124245   .0104551    -1.19   0.235    -.0329162    .0080672
        per3 |  -.0325178   .0127431    -2.55   0.011    -.0574938   -.0075418
        per4 |   -.046097   .0136286    -3.38   0.001    -.0728087   -.0193853
        per5 |  -.0577767    .014632    -3.95   0.000    -.0864548   -.0290985
       _cons |  -1.064449    .261872    -4.06   0.000    -1.577709   -.5511895
------------------------------------------------------------------------------
```



```
. margeff

Average marginal effects on Prob(lfp==1) after probit

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.0660184   .0049233   -13.41   0.000    -.0756678    -.056369
       lhinc |   -.070054   .0079819    -8.78   0.000    -.0856981   -.0544099
        educ |   .0264473   .0021119    12.52   0.000     .0223082    .0305865
       black |   .0698835   .0197251     3.54   0.000      .031223     .108544
         age |   .0480966   .0039216    12.26   0.000     .0404105    .0557828
       agesq |  -.0006609   .0000486   -13.60   0.000    -.0007561   -.0005656
        per2 |  -.0041304   .0034828    -1.19   0.236    -.0109565    .0026957
        per3 |   -.010839   .0042694    -2.54   0.011    -.0192069   -.0024712
        per4 |  -.0153921   .0045809    -3.36   0.001    -.0243705   -.0064137
        per5 |  -.0193224   .0049309    -3.92   0.000    -.0289867   -.0096581
------------------------------------------------------------------------------

. * So, without accounting for heterogeneity through the time averages,
. * the effects are much larger.
```


• Simple dynamic model (for binary response only, not fractional responses):

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{32}$$

A simple analysis is available if we specify

$$c_i|\mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2) \tag{33}$$

Then

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{34}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$.

• Turns out we can use standard random effects probit software (Stata: xtprobit), with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. Easily get the average partial effects, too:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{35}$$

with coefficients scaled by $(1 + \hat{\sigma}_a^2)^{-1/2}$.

EXAMPLE: (Dynamic Married Women's Labor Force Participation)

$$\begin{aligned} P(lfp_{it} &= 1|kids_{it}, lhinc_{it}, lfp_{i,t-1}, c_i) \\ &= \Phi(\alpha_t + \delta_1 kids_{it} + \delta_2 lhinc_{it} + \rho lfp_{i,t-1} + c_i) \\ c_i|\mathbf{z}_i, lfp_{i0} &\sim Normal(\psi + \xi_0 lfp_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2) \end{aligned}$$

- To get a measure of the magnitude of state dependence, estimate

$$E_{c_i}[\Phi(\alpha_t + \delta_1 kids_t + \delta_2 lhinc_t + \rho + c_i) - \Phi(\alpha_t + \delta_1 kids_t + \delta_2 lhinc_t + c_i)]$$

and put in interesting values for $kids_t$ and $lhinc_t$, or average those out in the sample.

- Data from LFP.DTA. The APE from dynamic probit with heterogeneity is about .260 (.026). If we ignore the heterogeneity, estimated APE is .837 (.005); standard errors from 500 panel bootstrap replications.
- Linear model estimates: .382 (.020) with heterogeneity, .851 (.004) without.


```
. * Start with a linear model estimated by Arellano and Bond:

. xtabond lfp kids lhinc per3 per4 per5

Arellano-Bond dynamic panel-data estimation  Number of obs         =     16989
Group variable: id                           Number of groups      =      5663
Time variable: period
                                             Obs per group:    min =         3
                                                               avg =         3
                                                               max =         3

Number of instruments =     12               Wald chi2(6)          =    378.77
                                             Prob > chi2           =    0.0000
One-step results
------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lfp_L1. |   .3818295   .0201399    18.96   0.000     .3423559    .4213031
        kids |  -.0130903   .0091827    -1.43   0.154     -.031088    .0049075
       lhinc |  -.0058375   .0053704    -1.09   0.277    -.0163633    .0046882
        per3 |  -.0053284   .0039777    -1.34   0.180    -.0131245    .0024677
        per4 |  -.0038833   .0039916    -0.97   0.331    -.0117067     .00394
        per5 |  -.0090286   .0039853    -2.27   0.023    -.0168396   -.0012176
       _cons |   .4848731   .0458581    10.57   0.000      .394993    .5747533
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).lfp
        Standard: D.kids D.lhinc D.per3 D.per4 D.per5
Instruments for level equation
        Standard: _cons
```



```
. * Accounting for heterogeneity is important, even in the linear
. * approximation. Without heterogeneity, the estimated state dependence is
. * much higher:

. reg lfp l.lfp kids lhinc per3 per4 per5, robust

Linear regression                                      Number of obs =   22652
                                                       F(  6, 22645) = 7938.78
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7207
                                                       Root MSE      =  .24664

------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lfp |
          L1. |   .8510015   .0039478   215.57   0.000     .8432637    .8587394
              |
        kids |  -.0021431   .0014379    -1.49   0.136    -.0049615    .0006754
       lhinc |  -.0071892   .0025648    -2.80   0.005    -.0122164   -.0021619
        per3 |  -.0036044   .0047215    -0.76   0.445    -.0128588      .00565
        per4 |   .0010464   .0046287     0.23   0.821    -.0080262     .010119
        per5 |  -.0036555   .0045471    -0.80   0.421    -.0125681    .0052571
       _cons |    .157911   .0210127     7.52   0.000     .1167247    .1990972
------------------------------------------------------------------------------
```

```
. * Generate variables needed for dynamic probit.

. sort id period

. gen lfp_1 = lfp[_n-1] if period > 1
(5663 missing values generated)

. * Put initial condition in periods 2-5:
. gen lfp1 = lfp[_n-1] if per2
(22652 missing values generated)

. replace lfp1 = lfp[_n-2] if per3
(5663 real changes made)

. replace lfp1 = lfp[_n-3] if per4
(5663 real changes made)

. replace lfp1 = lfp[_n-4] if per5
(5663 real changes made)

. * Put all kids variables in periods 2-5:
. gen kids2 = kids if per2
(22652 missing values generated)

. replace kids2 = kids[_n-1] if per3
(5663 real changes made)

. replace kids2 = kids[_n-2] if per4
(5663 real changes made)

. replace kids2 = kids[_n-3] if per5
(5663 real changes made)

. gen kids3 = kids[_n+1] if per2
(22652 missing values generated)
```



```
. replace kids3 = kids if per3
(5663 real changes made)

. replace kids3 = kids[_n-1] if per4
(5663 real changes made)

. replace kids3 = kids[_n-2] if per5
(5663 real changes made)

. gen kids4 = kids[_n+2] if per2
(22652 missing values generated)

. replace kids4 = kids[_n+1] if per3
(5663 real changes made)

. replace kids4 = kids if per4
(5663 real changes made)

. replace kids4 = kids[_n-1] if per5
(5663 real changes made)

. gen kids5 = kids[_n+3] if per2
(22652 missing values generated)

. replace kids5 = kids[_n+2] if per3
(5663 real changes made)

. replace kids5 = kids[_n+1] if per4
(5663 real changes made)

. replace kids5 = kids if per5
(5663 real changes made)

. * Put all lhinc variables in periods 2-5:
```



```
. gen lhinc2 = lhinc if per2
(22652 missing values generated)

. replace lhinc2 = lhinc[_n-1] if per3
(5663 real changes made)

. replace lhinc2 = lhinc[_n-2] if per4
(5663 real changes made)

. replace lhinc2 = lhinc[_n-3] if per5
(5663 real changes made)

. gen lhinc3 = lhinc[_n+1] if per2
(22652 missing values generated)

. replace lhinc3 = lhinc if per3
(5663 real changes made)

. replace lhinc3 = lhinc[_n-1] if per4
(5663 real changes made)

. replace lhinc3 = lhinc[_n-2] if per5
(5663 real changes made)

. gen lhinc4 = lhinc[_n+2] if per2
(22652 missing values generated)

. replace lhinc4 = lhinc[_n+1] if per3
(5663 real changes made)

. replace lhinc4 = lhinc if per4
(5663 real changes made)

. replace lhinc4 = lhinc[_n-1] if per5
(5663 real changes made)
```



```
. gen lhinc5 = lhinc[_n+3] if per2
(22652 missing values generated)

. replace lhinc5 = lhinc[_n+2] if per3
(5663 real changes made)

. replace lhinc5 = lhinc[_n+1] if per4
(5663 real changes made)

. replace lhinc5 = lhinc if per5
(5663 real changes made)
```

```
. * Now include initial condition, leads and lags, and other
. * time-constant variables in RE probit
.
. xtprobit lfp lfp_1 lfp1 kids kids2-kids5 lhinc lhinc2-lhinc5 educ
            black age agesq per3-per5, re


Random-effects probit regression                Number of obs      =     22652
Group variable (i): id                          Number of groups   =      5663

Random effects u_i ~Gaussian                    Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4

                                                Wald chi2(19)      =   4091.17
Log likelihood  = -5028.9785                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfp_1 |   1.541288    .066803    23.07   0.000     1.410357     1.67222
        lfp1 |   2.530053   .1565322    16.16   0.000     2.223256    2.836851
        kids |  -.1455379   .0787386    -1.85   0.065    -.2998626    .0087868
       kids2 |   .3236282   .0968499     3.34   0.001      .133806    .5134504
       kids3 |   .1072842   .1235197     0.87   0.385    -.1348099    .3493784
       kids4 |     .01792   .1275595     0.14   0.888    -.2320921    .2679322
       kids5 |  -.3912412   .1058482    -3.70   0.000    -.5986998   -.1837825
       lhinc |  -.0748846   .0508406    -1.47   0.141    -.1745304    .0247612
      lhinc2 |  -.0232267   .0590167    -0.39   0.694    -.1388973    .0924438
      lhinc3 |   -.083386   .0626056    -1.33   0.183    -.2060908    .0393188
      lhinc4 |  -.0862979    .060961    -1.42   0.157    -.2057793    .0331835
      lhinc5 |   .0627793   .0592742     1.06   0.290     -.053396    .1789547
        educ |    .049906   .0100314     4.97   0.000     .0302447    .0695672
       black |   .1316009   .0982941     1.34   0.181     -.061052    .3242539
         age |   .1278946   .0193999     6.59   0.000     .0898715    .1659177
```



```
       agesq |  -.0016882     .00024    -7.03   0.000    -.0021586   -.0012177
        per3 |  -.0560723   .0458349    -1.22   0.221    -.1459071    .0337625
        per4 |   -.029532   .0463746    -0.64   0.524    -.1204245    .0613605
        per5 |  -.0784793   .0464923    -1.69   0.091    -.1696025     .012644
       _cons |  -2.946082   .4367068    -6.75   0.000    -3.802011   -2.090152
-------------+----------------------------------------------------------------
    /lnsig2u |   .0982792   .1225532                     -.1419206     .338479
-------------+----------------------------------------------------------------
     sigma_u |   1.050367   .0643629                      .9314989    1.184404
         rho |     .52455   .0305644                      .4645793     .583821
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   160.73 Prob >= chibar2 = 0.000

. predict xdh, xb
(5663 missing values generated)

. gen xd0 = xdh - _b[lfp_1]*lfp_1
(5663 missing values generated)

. gen xd1 = xd0 + _b[lfp_1]
(5663 missing values generated)

. gen xd0a = xd0/sqrt(1 + (1.050367)^2)
(5663 missing values generated)

. gen xd1a = xd1/sqrt(1 + (1.050367)^2)
(5663 missing values generated)
```



```
. gen PHI0 = norm(xd0a)
(5663 missing values generated)

. gen PHI1 = norm(xd1a)
(5663 missing values generated)

. gen pelfp_1 = PHI1 - PHI0
(5663 missing values generated)

. sum pelfp_1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     pelfp_1 |     22652    .2591284    .0551711   .0675151   .4047995

. * .259 is the average probability of being in the labor force in
. * period t, given participation in t-1. This is somewhat lower than
. * the linear model estimate, .382.
```



```
. * A nonlinear model without heterogeneity gives a much larger
. * estimate:

. probit lfp lfp_1 kids lhinc educ black age agesq per3-per5

Probit regression                               Number of obs      =     22652
                                                LR chi2(10)        =  17744.22
                                                Prob > chi2        =    0.0000
Log likelihood = -5332.5289                     Pseudo R2          =    0.6246

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfp_1 |   2.875679   .0269811   106.58   0.000     2.822797    2.928561
        kids |   -.060792    .012217    -4.98   0.000    -.0847368   -.0368472
       lhinc |  -.1143176   .0211668    -5.40   0.000    -.1558037   -.0728315
        educ |   .0291868   .0052362     5.57   0.000     .0189241    .0394495
       black |   .0792495   .0536694     1.48   0.140    -.0259406    .1844395
         age |    .084403   .0099983     8.44   0.000     .0648067    .1039993
       agesq |  -.0010991   .0001236    -8.90   0.000    -.0013413    -.000857
        per3 |  -.0340795   .0369385    -0.92   0.356    -.1064777    .0383187
        per4 |   .0022816   .0371729     0.06   0.951    -.0705759    .0751391
        per5 |  -.0304156   .0371518    -0.82   0.413    -.1032318    .0424006
       _cons |  -2.170796   .2219074    -9.78   0.000    -2.605727   -1.735866
------------------------------------------------------------------------------
```

```
. predict xdph, xb
(5663 missing values generated)

. gen xdp0 = xdph - _b[lfp_1]*lfp_1
(5663 missing values generated)

. gen xdp1 = xdp0 + _b[lfp_1]
(5663 missing values generated)

. gen PHI0p = norm(xdp0)
(5663 missing values generated)

. gen PHI1p = norm(xdp1)
(5663 missing values generated)

. gen pelfp_1p = PHI1p - PHI0p
(5663 missing values generated)

. sum pelfp_1p

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pelfp_1p |     22652    .8373056     .012207   .6019558   .8495204

. * Without accounting for heterogeneity, the average state dependence
. * is much larger: .837 versus .259.

. * The .837 estimate is pretty close to the dynamic linear model without
. * heterogeneity, .851.
```

AEA Lectures **Chicago, January 2012**

Lecture 8, Tuesday, January 10, 9:45 am to 11 am

Control Function and Related Methods

These notes review the control function approach to handling endogeneity in models linear in parameters, and draws comparisons with standard methods such as 2SLS and maximum likelihood methods. Certain nonlinear models with endogenous explanatory variables are most easily estimated using the CF method, and the recent focus on average marginal effects suggests some simple, flexible strategies. Recent advances in semiparametric and nonparametric control function method are covered, and an example for how one can apply CF methods to nonlinear panel data models is provided.

1. Linear-in-Parameters Models: IV versus Control Functions

Most models that are linear in parameters are estimated using standard instrumental variables methods – either two stage least squares (2SLS) or generalized method of moments (GMM). An alternative, the control function (CF) approach, relies on the same kinds of identification conditions. In the standard case where a endogenous explanatory variables appear linearly, the CF approach leads to the usual 2SLS estimator. But there are differences for models nonlinear in endogenous variables even if they are linear in parameters. And, for models nonlinear in parameters, the CF approach offers some distinct advantages.

To illustrate the CF approach, let y_1 denote the response variable, y_2 the endogenous explanatory variable (a scalar for simplicity), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (which includes unity as its first element). Consider the model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{1.1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$ that also includes a constant. The sense in which $\mathbf{z}$ is exogenous is given by the L orthogonality (zero covariance) conditions

$$\mathrm{E}(\mathbf{z}'u_1) = \mathbf{0}. \tag{1.2}$$

Of course, this is the same exogeneity condition that we use for consistency of the 2SLS estimator, and we can consistently estimate $\boldsymbol{\delta}_1$ and α_1 by 2SLS under (1.2) and the rank condition, which reduces to *rank* $E(\mathbf{z}'\mathbf{x}_1) = K_1$, where $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$ is a $1 \times K_1$ vector. (We also need to assume $E(\mathbf{z}'\mathbf{z})$ is nonsingular, but this assumption is rarely a concern.)

Just as with 2SLS, the reduced form of y_2 – that is, the linear projection of y_2 onto the exogenous variables – plays a critical role. Write the reduced form with an error term as

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{1.3}$$

$$\mathrm{E}(\mathbf{z}'v_2) = \mathbf{0} \tag{1.4}$$

where $\boldsymbol{\pi}_2$ is $L \times 1$. Endogeneity of y_2 arises if and only if u_1 is correlated with v_2. Write the linear projection of u_1 on v_2, in error form, as

$$u_1 = \rho_1 v_2 + e_1, \tag{1.5}$$

where $\rho_1 = \mathrm{E}(v_2u_1)/\mathrm{E}(v_2^2)$ is the population regression coefficient. By definition, $\mathrm{E}(v_2e_1) = 0$, and $\mathrm{E}(\mathbf{z}'e_1) = \mathbf{0}$ because u_1 and v_2 are both uncorrelated with $\mathbf{z}$.

Plugging (1.5) into equation (1.1) gives

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{1.6}$$

where we now view v_2 as an explanatory variable in the equation. As just noted, e_1, is uncorrelated with v_2 and $\mathbf{z}$. Plus, y_2 is a linear function of $\mathbf{z}$ and v_2, and so e_1 is also uncorrelated with y_2.

Because e_1 is uncorrelated with $\mathbf{z}_1$, y_2, and v_2, (1.6) suggests a simple procedure for

consistently estimating δ_1 and α_1 (as well as ρ_1): run the OLS regression of y_1 on $\mathbf{z}_1, y_2$, and v_2 using a random sample. (Remember, OLS consistently estimates the parameters in any equation where the error term is uncorrelated with the right hand side variables.) The only problem with this suggestion is that we do not observe v_2; it is the error in the reduced form equation for y_2. Nevertheless, we can write $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and, because we collect data on y_2 and $\mathbf{z}$, we can consistently estimate $\boldsymbol{\pi}_2$ by OLS. Therefore, we can replace v_2 with $\hat{v}_2$, the OLS residuals from the first-stage regression of y_2 on $\mathbf{z}$. Simple substitution gives

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error, \tag{1.7}$$

where, for each i, $error_i = e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, which depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$. Standard results on two-step estimation imply the OLS estimators from (1.7) will be consistent for δ_1, α_1, and ρ_1.

The OLS estimates from (1.7) are control function estimates. The inclusion of the residuals $\hat{v}_2$ "controls" for the endogeneity of y_2 in the original equation (although it does so with sampling error because $\hat{\boldsymbol{\pi}}_2 \neq \boldsymbol{\pi}_2$).

It is a simple exercise in the algebra of least squares to show that the OLS estimates of $\boldsymbol{\delta}_1$ and α_1 from (1.7) are *identical* to the 2SLS estimates starting from (1.1) and using $\mathbf{z}$ as the vector of instruments. [Standard errors from (1.7) must adjust for the generated regressor.]

It is trivial to use (1.7) to test $H_0 : \rho_1 = 0$, as the usual t statistic is asymptotically valid under homoskedasticity ($Var(u_1|\mathbf{z}, y_2) = \sigma_1^2$ under H_0); or use the heteroskedasticity-robust version (which does *not* account for the first-stage estimation of $\boldsymbol{\pi}_2$).

An estimator that can be different from the CF and 2SLS estimators is the limited information (quasi-) maximum likelihood (LIML) estimator. The LIML estimator is obtained from equations (1.1) and (1.3) under the assumption that (u_1, v_2) is independent of $\mathbf{z}$ with a

mean-zero bivariate normal distribution. In fact, we can work off of (1.3) and (1.6) and use the relationship $f(y_1, y_2|\mathbf{z}) = f(y_1|y_2, \mathbf{z})f(y_2|\mathbf{z})$. If $\eta_1^2 = Var(e_1)$ and $\tau_2^2 = Var(v_2)$, the quasi-log-likelihood for observation i is

$$\begin{aligned} &-\log(\eta_1^2)/2 - [(y_{i1} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \alpha_1 y_{i2} - \rho_1(y_{i2} - \mathbf{z}_i\boldsymbol{\pi}_2)]^2/(2\eta_1^2) \\ &\quad -\log(\tau_2^2)/2 - (y_{i2} - \mathbf{z}_i\boldsymbol{\pi}_2)^2/(2\tau_2^2), \end{aligned} \tag{1.8}$$

and all parameters are estimated simultaneously. When (1.1) is overidentified, LIML is generally different from CF (2SLS). And, as the weak instruments notes document, LIML typically has better statistical properties than 2SLS in situations with overidentification. The CF approach can be seen to be a two-step version of LIML, where $\boldsymbol{\pi}_2$ is obtained in a first step and then $\boldsymbol{\delta}_1, \alpha_1$, and ρ_1 are estimated in a second step. (The variance parameters can be estimated in the two-step procedure, too.) Fortunately, while LIML is derived under joint normality, it is just as robust as the CF estimator: independence between the errors and $\mathbf{z}$ and normality are not needed.

[Incidentally, full information maximum likelihood (FIML) arises in systems with true simultaneity when interest lies in estimating all structural equations. In these notes, we assume that one equation is of particular interest. This could be because it is the main equation in a truly simultaneous system or because the endogeneity we are worried about is due to omitted variables.]

Now extend the model to include a quadratic:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{1.9}$$

$$\mathrm{E}(u_1|\mathbf{z}) = 0. \tag{1.10}$$

For simplicity, assume that we have a scalar, z_2, that is not also in $\mathbf{z}_1$. Then, under (1.10) – which is stronger than (1.2), and is essentially needed to identify nonlinear models – we can

use, say, z_2^2 (if z_2 is not binary) as an instrument for y_2^2 because any function of z_2 is uncorrelated with u_1. In other words, we can apply the standard IV estimator with explanatory variables $(\mathbf{z}_1, y_2, y_2^2)$ and instruments $(\mathbf{z}_1, z_2, z_2^2)$; note that we have two endogenous explanatory variables, y_2 and y_2^2.

What would the CF approach entail in this case? To implement the CF approach in (1.9), we obtain the conditional expectation $\mathrm{E}(y_1|\mathbf{z}, y_2)$ – a linear projection argument no longer works because of the nonlinearity – and that requires an assumption about $\mathrm{E}(u_1|\mathbf{z}, y_2)$. A standard assumption is

$$\mathrm{E}(u_1|\mathbf{z}, y_2) = \mathrm{E}(u_1|\mathbf{z}, v_2) = \mathrm{E}(u_1|v_2) = \rho_1 v_2, \tag{1.11}$$

where the first equality follows because y_2 and v_2 are one-to-one functions of each other (given $\mathbf{z}$) and the second would hold if (u_1, v_2) is independent of $\mathbf{z}$ – a nontrivial restriction on the reduced form error in (1.3), not to mention the structural error u_1.. The final assumption is linearity of the conditional expectation $\mathrm{E}(u_1|v_2)$, which is more restrictive than simply defining a linear projection. Under (1.11),

$$\begin{aligned} \mathrm{E}(y_1|\mathbf{z}, y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1(y_2 - \mathbf{z}\boldsymbol{\pi}_2) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2. \end{aligned} \tag{1.12}$$

Implementing the CF approach means running the OLS regression y_1 on $\mathbf{z}_1, y_2, y_2^2, \hat{v}_2$,where $\hat{v}_2$ still represents the reduced form residuals. The CF estimates are *not* the same as the 2SLS estimates using any choice of instruments for (y_2, y_2^2).

The CF approach, while likely more efficient than a direct IV approach, is less robust. For example, it is easily seen that (1.10) and (1.11) imply that $\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$. A linear conditional expectation for y_2 is a substantive restriction on the conditional distribution of y_2. Therefore, the CF estimator will be inconsistent in cases where the 2SLS estimator will be consistent. On

the other hand, because the CF estimator solves the endogeneity of y_2 and y_2^2 by adding the scalar $\hat{v}_2$ to the regression, it will generally be more precise – perhaps much more precise – than the IV estimator. [I do not know of a systematic analysis comparing the two approaches in models such as (1.9).]

The equivalence between CF approaches and IV methods is broken even in the simple model (1.1) if we allow y_2 to have discreteness in its distribution and we use a distributional assumption to exploit that discreteness. For example, suppose y_2 is a binary response. The standard CF approach involves estimating

$$E(y_1|\mathbf{z},y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z},y_2), \tag{1.13}$$

and so we must be able to estimate $E(u_1|\mathbf{z},y_2)$. If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$, (u_1, e_2) is independent of $\mathbf{z}$, $E(u_1|e_2) = \rho_1 e_2$, and e_2 ~Normal$(0,1)$, then

$$\begin{aligned} E(u_1|\mathbf{z},y_2) &= E[E(u_1|\mathbf{z},e_2)|\mathbf{z},y_2] = \rho_1 E(v_2|\mathbf{z},y_2) \\ &= \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \end{aligned} \tag{1.14}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio (IMR). A simple two-step estimator is to obtain the probit estimate $\hat{\boldsymbol{\delta}}_2$ and then to add the "generalized residual," $\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as a regressor:

$$y_{i1} \text{ on } \mathbf{z}_{i1},\ y_{i2},\ \widehat{gr}_{i2},\ i = 1,\ldots,N. \tag{1.15}$$

The estimators from this regression are consistent and $\sqrt{N}$-asymptotically normal provided $D(y_2|\mathbf{z})$ follows a probit, $E(u_1|v_2)$ is linear, and $E(u_1|\mathbf{z},v_2) = E(u_1|v_2)$. (Standard errors need to be adjusted for the two-step estimation, except when $\rho_1 = 0$. A simple t test on $\widehat{gr}_{i2}$ is valid as a test of $H_0 : \rho_1 = 0$.)

Of course, if we just apply 2SLS directly to $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$, we make no distinction

among discrete, continuous, or some mixture for y_2. 2SLS is consistent if $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ actually depends on $\mathbf{z}_2$ and (1.2) holds. So, while estimating (1.1) using CF methods when y_2 is binary is somewhat popular (Stata's "treatreg" even has the option of full MLE, where (u_1, e_2) is bivariate normal), one should remember that it is less robust than standard IV approaches. In principal, it is much less robust, but whether estimates obtained from (1.15) differ substantially from 2SLS estimates is an empirical issue.

Often researchers look to exploit the binary nature of the endogenous explanatory variable, and there may even be some confusion about the properties of 2SLS in such contexts. Again, it is important to understand that 2SLS is consistent, $\sqrt{N}$-asymptotically normal, and inference is standard. But it could be asymptotically inefficient. Therefore, a natural question is: How might one use the binary nature of y_2 in IV estimation [as opposed to the CF approach in (1.15)]? We need to assume $E(u_1|\mathbf{z}) = 0$ to exploit nonlinear functions $\mathbf{z}$ as IVs. Nominally, the same probit model for $D(y_2|\mathbf{z})$ that is used in the CF approach. Then, after estimating the probit model, obtain the fitted probabilities, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$. These fitted probabilities are then used as IVs for y_{i2} in estimating (1.1). This method has several attractive features: it is fully robust to misspecification of the probit model, provided one uses $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as an IV for y_{i2}, not as a regressor in place of y_{i2}; the standard errors need not be adjusted for the first-stage probit (asymptotically); and it is the efficient IV estimator if $P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$. Probably it is less efficient than the CF estimator if the additional assumptions needed for CF consistency hold; a careful study could shed light on the tradeoffs. See Wooldridge (2010, Chapter 21) for further discussion.

We can briefly summarize the main points of this section. In the model (1.1), CF methods based on $E(y_1|\mathbf{z},y_2)$ impose additional assumptions compared with standard IV methods. When

y_2 has special features (such as being binary, or even a corner solution), models for $E(y_2|\mathbf{z})$ can be used to generate instruments (not regressors) for y_2. The resulting IV estimates are robust to misspecification of the model for $E(y_2|\mathbf{z})$ and the first-step estimation can be ignored asymptotically.

2. Correlated Random Coefficient Models

Control function methods can be used for random coefficient models – that is, models where unobserved heterogeneity interacts with endogenous explanatory variables. In some cases, CF methods are indispensable; in other cases, standard IV methods are more robust. To illustrate, we modify equation (1.1) as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + u_1, \tag{2.1}$$

where $\mathbf{z}_1$ is $1 \times L_1$, y_2 is the endogenous explanatory variable, and a_1, the "coefficient" on y_2 – an unobserved random variable. [It is now convenient to set apart the intercept.] We could replace $\boldsymbol{\delta}_1$ with a random vector, say $\mathbf{d}_1$, and this would not affect our analysis of the IV estimator (but, as we will see, does change the control function estimator). Following Heckman and Vytlacil (1998), we refer to (2.1) as a **correlated random coefficient** (**CRC**) **model**.

It is convenient to write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = \mathrm{E}(a_1)$ is the object of interest. We can rewrite the equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1, \tag{2.2}$$

where $e_1 = v_1 y_2 + u_1$. Equation (2.2) shows explicitly a constant coefficient on y_2 (which we hope to estimate) but also an interaction between the observed heterogeneity, v_1, and y_2. Remember, (2.2) is a population model. For a random draw, we would write $y_{i1} = \eta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + v_{i1} y_{i2} + u_{i1}$, which makes it clear that $\boldsymbol{\delta}_1$ and α_1 are parameters to estimate and v_{i1} is specific to observation i.

As discussed in Wooldridge (1997, 2003), the potential problem with applying instrumental variables (2SLS) to (2.2) is that the error term $v_1 y_2 + u_1$ is not necessarily uncorrelated with

the instruments **z**, even if we make the assumptions

$$\mathrm{E}(u_1|\mathbf{z}) = \mathrm{E}(v_1|\mathbf{z}) = 0, \tag{2.3}$$

which we maintain from here on. Generally, the term $v_1 y_2$ can cause problems for IV estimation, but it is important to be clear about the nature of the problem. If we are allowing y_2 to be correlated with u_1 then we also want to allow y_2 and v_1 to be correlated. In other words, $\mathrm{E}(v_1 y_2) = \mathrm{Cov}(v_1, y_2) \equiv \tau_1 \neq 0$. But a nonzero unconditional covariance is *not* a problem with applying IV to (2.2): it simply implies that the composite error term, e_1, has (unconditional) mean τ_1 rather than a zero. As we know, a nonzero mean for e_1 means that the orginal intercept, η_1, would be inconsistenly estimated, but this is rarely a concern.

Therefore, we can allow $\mathrm{Cov}(v_1, y_2)$, the unconditional covariance, to be unrestricted. But the usual IV estimator is generally inconsistent if $\mathrm{E}(v_1 y_2|\mathbf{z})$ depends on **z**. Note that, because $\mathrm{E}(v_1|\mathbf{z}) = 0$, $\mathrm{E}(v_1 y_2|\mathbf{z}) = \mathrm{Cov}(v_1, y_2|\mathbf{z})$. Therefore, as shown in Wooldridge (2003), a sufficient condition for the IV estimator applied to (2.2) to be consistent for $\boldsymbol{\delta}_1$ and α_1 is

$$\mathrm{Cov}(v_1, y_2|\mathbf{z}) = \mathrm{Cov}(v_1, y_2). \tag{2.4}$$

The 2SLS intercept estimator is consistent for $\eta_1 + \tau_1$. Condition (2.4) means that the conditional covariance between v_1 and y_2 is not a function of **z**, but the unconditional covariance is unrestricted.

Because v_1 is unobserved, we cannot generally verify (2.4). But it is easy to find situations where it holds. For example, if we write

$$y_2 = m_2(\mathbf{z}) + v_2 \tag{2.5}$$

and assume (v_1, v_2) is independent of **z** (with zero mean), then (2.4) is easily seen to hold because $\mathrm{Cov}(v_1, y_2|\mathbf{z}) = \mathrm{Cov}(v_1, v_2|\mathbf{z})$, and the latter cannot be a function of **z** under

independence. Of course, assuming v_2 in (2.5) is independent of $\mathbf{z}$ is a strong assumption even if we do not need to specify the mean function, $m_2(\mathbf{z})$. It is much stronger than just writing down a linear projection of y_2 on $\mathbf{z}$ (which is no real assumption at all). As we will see in various models in Part IV, the representation (2.5) with v_2 independent of $\mathbf{z}$ is not suitable for discrete y_2, and generally (2.4) is not a good assumption when y_2 has discrete characteristics. Further, as discussed in Card (2001), (2.4) can be violated even if y_2 is (roughly) continuous. Wooldridge (2005) makes some headway in relaxing (2.44) by allowing for parametric heteroskedasticity in u_1 and v_2.

A useful extension of (1.1) is to allow observed exogenous variables to interact with y_2. The most convenient formulation is

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + (\mathbf{z}_1 - \boldsymbol{\psi}_1)y_2\boldsymbol{\gamma}_1 + v_1 y_2 + u_1 \tag{2.6}$$

where $\boldsymbol{\psi}_1 \equiv \mathrm{E}(\mathbf{z}_1)$ is the $1 \times L_1$ vector of population means of the exogenous variables and $\boldsymbol{\gamma}_1$ is an $L_1 \times 1$ parameter vector. As we saw in Chapter 4, subtracting the mean from $\mathbf{z}_1$ before forming the interaction with y_2 ensures that α_1 is the average partial effect.

Estimation of (2.6) is simple if we maintain (2.4) [along with (2.3) and the appropriate rank condition]. Typically, we would replace the unknown $\boldsymbol{\psi}_1$ with the sample averages, $\bar{\mathbf{z}}_1$, and then estimate

$$y_{i1} = \theta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}\boldsymbol{\gamma}_1 + error_i \tag{2.7}$$

by instrumental variables, ignoring the estimation error in the population mean. The only issue is choice of instruments, which is complicated by the interaction term. One possibility is to use interactions between $\mathbf{z}_{i1}$ and all elements of $\mathbf{z}_i$ (including $\mathbf{z}_{i1}$). This results in many overidentifying restrictions, even if we just have one instrument z_{i2} for y_{i2}. Alternatively, we

could obtain fitted values from a first stage linear regression y_{i2} on $\mathbf{z}_i$, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, and then use IVs $[1, \mathbf{z}_i, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$, which results in as many overidentifying restrictions as for the model without the interaction. Importantly, the use of $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}$ as IVs for $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}$ is asymptotically the same as using instruments $(\mathbf{z}_{i1} - \boldsymbol{\psi}_1) \cdot (\mathbf{z}_i\boldsymbol{\pi}_2)$, where $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ is the linear projection. In other words, consistency of this IV procedure does not in any way restrict the nature of the distribution of y_2 given $\mathbf{z}$. Plus, although we have generated instruments, the assumptions sufficient for ignoring estimation of the instruments hold, and so inference is standard (perhaps made robust to heteroskedasticity, as usual).

We can just identify the parameters in (2.6) by using a further restricted set of instruments, $[1, \mathbf{z}_{i1}, \hat{y}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$. If so, it is important to use these as instruments and not as regressors. If we add the assumption. The latter procedure essentially requires a new assumption:

$$\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2 \tag{2.8}$$

(where $\mathbf{z}$ includes a constant). Under (2.3), (2.4), and (2.8), it is easy to show

$$\mathrm{E}(y_1|\mathbf{z}) = (\eta_1 + \tau_1) + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\pi}_2) + (\mathbf{z}_1 - \boldsymbol{\psi}_1) \cdot (\mathbf{z}\boldsymbol{\pi}_2)\boldsymbol{\gamma}_1, \tag{2.9}$$

which is the basis for the Heckman and Vytlacil (1998) plug-in estimator. The usual IV approach applied to (2.7) simply relaxes (2.8) and does not require adjustments to the standard errors (because it uses generated instruments, not generated regressors).

We can also use a control function approach if we assume

$$\mathrm{E}(u_1|\mathbf{z}, v_2) = \rho_1 v_2, \mathrm{E}(v_1|\mathbf{z}, v_2) = \xi_1 v_2. \tag{2.10}$$

Then

$$\mathrm{E}(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \xi_1 v_2 y_2 + \rho_1 v_2, \tag{2.11}$$

and this equation is estimable once we estimate $\boldsymbol{\pi}_2$. Garen's (1984) control function procedure

is to first regress y_2 on $\mathbf{z}$ and obtain the reduced form residuals, $\hat{v}_2$, and then to run the OLS regression y_1 on $1, \mathbf{z}_1, y_2, \hat{v}_2 y_2, \hat{v}_2$. Under the maintained assumptions, Garen's method consistently estimates $\boldsymbol{\delta}_1$ and α_1. Because the second step uses generated regressors, the standard errors should be adjusted for the estimation of $\boldsymbol{\pi}_2$ in the first stage. Nevertheless, a test that y_2 is exogenous is easily obtained from the usual F test of $\mathrm{H}_0 : \xi_1 = 0, \rho_1 = 0$ (or a heteroskedasticity-robust version). Under the null, no adjustment is needed for the generated standard errors.

Garen's assumptions are more restrictive than those needed for the standard IV estimator to be consistent. For one, it would be a fluke if (2.10) held without the conditional covariance $\mathrm{Cov}(v_1, y_2|\mathbf{z})$ being independent of $\mathbf{z}$. Plus, like HV (1998), Garen relies on a linear model for $\mathrm{E}(y_2|\mathbf{z})$. Further, Garen adds the assumptions that $\mathrm{E}(u_1|v_2)$ and $\mathrm{E}(v_1|v_2)$ are linear functions, something not needed by the IV approach.

Of course, one can make Garen's approach less parametric by replacing the linear functions in (2.10) with unknown functions. But independence of (u_1, v_1, v_2) and $\mathbf{z}$ – or something very close to independence – is needed. And this assumption is not needed for the usual IV estimator,

If the assumptions needed for Garen's CF estimator to be consistent hold, it is likely more efficient than the IV estimator, although a comparison of the correct asymptotic variances is complicated. Again, there is a tradeoff between efficiency and robustness.

In the case of binary y_2, we have what is often called the "switching regression" model. Now, the right hand side of equation (2.11) represents $E(y_1|\mathbf{z}, v_2)$ where $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]$. If we assume (2.10) and that $v_2|\mathbf{z}$ is Normal$(0, 1)$, then

$$E(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) y_2, \tag{2.12}$$

where

$$h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) = y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2) \tag{2.13}$$

is the generalized residual function. The two-step estimation method is the one due to Heckman (1976).

There are two ways to embellish the model. The first is common: interact $(\mathbf{z}_1 - \boldsymbol{\mu}_1)$ with y_2 to allow different slopes for the "treated" and non-treated groups (keeping α_1 as the average treatment effect). With this extension, the CF regression is

$$y_{i1} \text{ on } 1,\ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}, \tag{2.14}$$

and is identical to running two separate regressions, including the IMRs for $y_2 = 0$ and $y_2 = 1$. The estimate of α_1 is the difference in the two intercepts.

An extension that is not so common – in fact, it seems not to appear in the literature – comes from allowing $\mathbf{z}_1$ to also interact with heterogeneity, as in

$$y_1 = \mathbf{z}_1\mathbf{d}_1 + a_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\mathbf{g}_1 + u_1. \tag{2.15}$$

Now all coefficients are heterogeneous. If we assume that $E(a_1|v_2)$, $E(\mathbf{d}_1|v_2)$, and $E(\mathbf{g}_1|v_2)$ are linear in v_2, then

$$\begin{aligned} E(y_1|\mathbf{z}, y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\xi}_1 + \rho_1 E(v_2|\mathbf{z}, y_2) + \xi_1 E(v_2|\mathbf{z}, y_2)y_2 \\ &\quad + \mathbf{z}_1 E(v_2|\mathbf{z}, y_2)\boldsymbol{\psi}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)E(v_2|\mathbf{z}, y_2)\boldsymbol{\omega}_1 \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2)y_2 \\ &\quad + h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2)\mathbf{z}_1\boldsymbol{\psi}_1 + h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2)y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\omega}_1 \end{aligned} \tag{2.16}$$

and the second-step estimation after the first stage probit is a regression

$$\begin{aligned} y_{i1} \text{ on } 1,\ &\mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}, \\ &h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1},\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1). \end{aligned} \tag{2.17}$$

across all observations i. Bootstrapping can be used to obtain valid standard errors because the first-stage estimation is just probit and the second stage is just linear regression.

If not for the term $v_1 y_2$ in (2.6), we could, in a much more robust manner, apply IV directly to (2.7) (and the standard errors are easier to obtain, too). The IVs would be, say, $[1, \mathbf{z}_{i1}, \hat{\Phi}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1) \cdot \hat{\Phi}_{i2}]$, and the same procedure consistently estimates the average effects whether or not there are random coefficients on $\mathbf{z}_{i1}$.

Interestingly, the addition of the terms $h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1}$ and $h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)$ has similarities with methods that allow $E(v_1|v_2)$ and so on to be more flexible. For example, as shown in Heckman and MaCurdy (1986), if $E(u_1|v_2) = \rho_1 v_2 + \kappa_1(v_2^2 - 1)$, then the extra term in the expected value when $y_2 = 1$ is $-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, and there is a similar expression for $y_{i2} = 0$.

Newey (1988), in the standard switching regression framework, proposed a flexible two-step procedure that estimates $\boldsymbol{\delta}_2$ semiparametrically in the first stage – see Powell (1994) for a survey of such methods – and then uses series in $\mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ in place of the usual IMR terms. He obtains valid standard errors and, in most cases, bootstrapping is valid, too.

Finally, we should not forget that maximum likelihood estimation is possible, too. If $D(y_2|\mathbf{z})$ is specified as a probit and all unobservables are assumed to be jointly normal and independent of $\mathbf{z}$, $D(y_1|y_2, \mathbf{z})$ can be obtained and all parameters can be estimated jointly.

3. Some Common Nonlinear Models and Limitations of the CF Approach

Like standard IV methods, control function approaches are more difficult to apply to nonlinear models, even relatively simple ones. Methods are available when the endogenous explanatory variables are continuous, but few if any results apply to cases with discrete y_2. Therefore, maximum likelihood approaches continue to be popular for nonlinear models.

3.1. Binary and Fractional Responses

The probit model provides a good illustration of the general approach. With a single endogenous explanatory variable, the simplest specification is

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \geq 0], \tag{3.1}$$

where $u_1|z \sim \text{Normal}(0,1)$. But the analysis goes through if we replace (z_1, y_2) with any known function $g_1(z_1, y_2)$, provided we have sufficient identifying assumptions. An example is $y_1 = [\mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_1\boldsymbol{\alpha}_1 + \gamma_1 y_2^2 + u_1 > 0]$. The nonlinearity in y_2 is not itself a problem (unless we inappropriately try to mimic 2SLS – more on this later).

The Smith-Blundell (1986) and Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for y_2,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, \;\; v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2). \tag{3.2}$$

A key point is that the RV approach essentially requires

$$(u_1, v_2) \text{ independent of } \mathbf{z}; \tag{3.3}$$

as we will see in the next section, semiparametric and nonparametric CF methods also rely on (3.3), or at least something close to it..

If we assume

$$(u_1, v_2) \sim \text{Bivariate Normal} \tag{3.4}$$

with $\rho_1 = Corr(u_1, v_2)$, then we can proceed with MLE based on $f(y_1, y_2|\mathbf{z})$. A simpler two-step approach, which is convenient for testing $H_0 : \rho_1 = 0$ (y_2 is exogenous), is also available, and it works if we replace the normality assumption in (3.2), the independence assumption in (3.3), and joint normality in (3.4) with

$$D(u_1|v_2, \mathbf{z}) = \text{Normal}(\theta_1 v_2, 1 - \rho_1^2), \tag{3.5}$$

where $\theta_1 = \rho_1/\tau_2$ is the regression coefficient. That we can relax the assumptions to some degree using a two-step CF approach has implications for less parametric approaches. Certainly we can relax the homoskedasticity and linear expectation in (3.3) without much additional work, as discussed in Wooldridge (2005).

Under the weaker assumption (3.5) we can write

$$P(y_1 = 1|\mathbf{z}, y_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\rho 1} + \alpha_{\rho 1} y_2 + \theta_{\rho 1} v_2) \tag{3.6}$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$.

The RV two-step approach is

(1) OLS of y_2 on $\mathbf{z}$, to obtain the residuals, $\hat{v}_2$.

(2) Probit of y_1 on $\mathbf{z}_1, y_2, \hat{v}_2$ to estimate the scaled coefficients.

The original coefficients, which appear in the partial effects, are easily obtained from the set of two-step estimates:

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\rho 1}/(1 + \hat{\theta}_{\rho 1}^2 \hat{\tau}_2^2)^{1/2}, \tag{3.7}$$

where $\hat{\theta}_{\rho 1}$ is the coeffcient on $\hat{v}_2$ and $\hat{\tau}_2^2$ is the usual error variance estimator from the first step

OLS, and $\hat{\boldsymbol{\beta}}_{\rho 1}$ includes $\hat{\boldsymbol{\delta}}_{\rho 1}$ and $\hat{\alpha}_{\rho 1}$. Standard errors can be obtained from the delta method of bootstrapping. Of course, they are computed directly from MLE. Partial effects are based on $\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1)$ where $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$. It should be clear that nothing changes for estimation if $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)$; of course, we would change how partial effects are computed to account for the specific function $\mathbf{g}_1(\cdot, \cdot)$.

Testing the null hypothesis that y_2 is exogenous is simple using the two-step control function approache. Asymptotically, a simple t test on $\hat{v}_2$ is valid to test $\mathrm{H}_0 : \rho_1 = 0$.

Under (3.3), we can also apply maximum likelihood by combining (3.2) and (3.6), recognizing that $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and estimating all parameters jointly. For details, see Wooldridge (2010, Section 15.7.2).

A different way to obtain partial effects is to use the average structural function approach, which leads to estimation of $\mathrm{E}_{v_2}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_{\rho 1} + \theta_{\rho 1} v_2)]$. Whether or not v_2 is normally distributed, a consistent, $\sqrt{N}$-asymptotically normal estimator of the average structural function (evaluated at a given vector $\mathbf{x}_1$) is

$$\widehat{\mathrm{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1}\hat{v}_{i2}); \tag{3.8}$$

that is, we average out the reduced form residuals, $\hat{v}_{i2}$. This formulation is also useful for more complicated models.

Given that the probit structural model is essentially arbitrary, one might be so bold as to specify models for $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$ directly. For example, we can add polynomials in v_2 or even interact v_2 with elements of $\mathbf{x}_1$ side a probit or logit function. We return to such possibilities in the next section.

The two-step CF approach easily extends to fractional responses. Now, we start with an omitted variables formulation in the conditional mean:

$$E(y_1|\mathbf{z},y_2,q_1) = E(y_1|\mathbf{z}_1,y_2,q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \tag{3.9}$$

where $\mathbf{x}_1$ is a function of $(\mathbf{z}_1,y_2)$ and q_1 contains unobservables. As usual, we need some exclusion restrictions, embodied by omitting $\mathbf{z}_2$ from $\mathbf{x}_1$. The specification in equation (3.9) allows for responses at the corners, zero and one, and y_1 may take on any values in between. Under the assumption that

$$D(q_1|v_2,\mathbf{z}) \sim \text{Normal}(\theta_1 v_2, \eta_1^2) \tag{3.10}$$

Given (3.9) and (3.10), it can be shown, using the mixing property of the normal distribution, that

$$E(y_1|\mathbf{z},y_2,v_2) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1} + \theta_{\eta 1} v_2), \tag{3.11}$$

where the index "η" denotes coefficients multiplied by $(1 + \eta_1^2)^{-1/2}$. Because the Bernoulli log likelihood is in the linear exponential family, maximizing it consistently estimates the parameters of a correctly specified mean; naturally, the same is true for two-step estimation. That is, the *same* two-step method can be used in the binary and fractional cases. Of course, the variance associated with the Bernoulli distribution is generally incorrect. In addition to correcting for the first-stage estimates, a robust sandwich estimator should be computed to account for the fact that $D(y_1|\mathbf{z},y_2)$ is not Bernoulli. The best way to compute partial effects is to use (3.8), with the slight notational change that the implicit scaling in the coefficients is different. By using (3.8), we can directly use the scaled coefficients estimated in the second stage – a feature common across CF methods for nonlinear models. The bootstrap that reestimates the first and second stages for each iteration is an easy way to obtain standard

errors. Of course, having estimates of the parameters up to a common scale allows us to determine signs of the partial effects in (3.9) as well as relative partial effects on the continuous explanatory variables.

Wooldridge (2005) describes some simple ways to make the analysis starting from (3.9) more flexible, including allowing $Var(q_1|v_2)$ to be heteroskedastic. We can also use strictly monotonic transformations of y_2 in the reduced form, say $h_2(y_2)$, regardless of how y_2 appears in the structural model: the key is that y_2 can be written as a function of $(\mathbf{z}, v_2)$. The extension to multivariate $\mathbf{y}_2$ is straightforward with sufficient instruments provide the elements of $\mathbf{y}_2$, or strictly monotonic functions of them, have reduced forms with additive errors that are effectively indendent of $\mathbf{z}$. (This assumption rules out applications to y_2 that are discrete (binary, multinomial, or count)or have a discrete component (corner solution).

The control function approach has some decided advantages over another two-step approach – one that appears to mimic the 2SLS estimation of the linear model. Rather than conditioning on v_2 along with $\mathbf{z}$ (and therefore y_2) to obtain $\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{P}(y_1 = 1|\mathbf{z}, y_2, v_2)$, we can obtain $\mathrm{P}(y_1 = 1|\mathbf{z})$. To find the latter probability, we plug in the reduced form for y_2 to get $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2) + \alpha_1 v_2 + u_1 > 0]$. Because $\alpha_1 v_2 + u_1$ is independent of $\mathbf{z}$ and (u_1, v_2) has a bivariate normal distribution, $\mathrm{P}(y_1 = 1|\mathbf{z}) = \Phi\{[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2)]/\omega_1\}$ where $\omega_1^2 \equiv \mathrm{Var}(\alpha_1 v_2 + u_1) = \alpha_1^2\tau_2^2 + 1 + 2\alpha_1\mathrm{Cov}(v_2, u_1)$. (A two-step procedure now proceeds by using the same first-step OLS regression – in this case, to get the fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ – now followed by a probit of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. It is easily seen that this method estimates the coefficients up to the common scale factor $1/\omega_1$, which can be any positive value (unlike in the CF case, where we know the scale factor is greater than unity).

One danger with plugging in fitted values for y_2 is that one might be tempted to plug $\hat{y}_2$ into nonlinear functions, say y_2^2 or $y_2\mathbf{z}_1$. This does not result in consistent estimation of the scaled parameters or the partial effects. If we believe y_2 has a linear RF with additive normal error independent of $\mathbf{z}$, the addition of $\hat{v}_2$ solves the endogeneity problem regardless of how y_2 appears. Plugging in fitted values for y_2 only works in the case where the model is linear in y_2. Plus, the CF approach makes it much easier to test the null that for endogeneity of y_2 as well as compute APEs.

In standard index models such as (3.9), or, if you prefer, (3.1), the use of control functions to estimate the (scaled) parameters and the APEs produces no surprises. However, one must take care when, say, we allow for random slopes in nonlinear models. For example, suppose we propose a random coefficient model

$$E(y_1|\mathbf{z},y_2,\mathbf{c}_1) = E(y_1|\mathbf{z}_1,y_2,\mathbf{c}_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1y_2 + q_1), \tag{3.12}$$

where a_1 is random with mean α_1 and q_1 again has mean of zero. If we want the partial effect of y_2, evaluated at the mean of heterogeneity, we have

$$\alpha_1\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2), \tag{3.13}$$

where $\phi(\cdot)$ is the standard normal pdf, and this equation is obtained by differentiating (3.12) with respect to y_2 and then plugging in $a_1 = \alpha_1$ and $q_1 = 0$. Suppose we write $a_1 = \alpha_1 + d_1$ and assume that (d_1, q_1) is bivariate normal with mean zero. Then, for given $(\mathbf{z}_1, y_2)$, the average structural function can be shown to be

$$E_{(d_1,q_1)}[\Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2 + d_1y_2 + q_1)] = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2)/(1 + \sigma_q^2 + 2\sigma_{dq}y_2 + \sigma_d^2y_2^2)^{1/2}], \tag{3.14}$$

where $\sigma_q^2 = Var(q_1)$, $\sigma_d^2 = Var(d_1)$, and $\sigma_{dq} = Cov(d_1, q_1)$. The average partial effect with respect to, say, y_2, is the derivative of this function with respect to y_2. While this partial effect

depends on α_1, it is messier than (3.13) and need not even have the same sign as α_1. Wooldridge (2005) discusses related issues in the context of probit models with exogenous variables and heteroskedasticity. In one example, he shows that, depending on whether heteroskedasticity in the probit is due to heteroskedasticity in $Var(u_1|\mathbf{x}_1)$, where u_1 is the latent error, or due to random slopes, the APEs are completely different in general. The same is true here: the APE when the coefficient on y_2 is random is generally very different from the APE obtained if we maintain $a_1 = \alpha_1$ but allow $Var(q_1|v_2)$ to be heteroskedastic. In the latter case, the APE is a positive multiple of α_1.

Incidentally, we can estimate the APE in (3.14) fairly generally. A parametric approach is to assume joint normality of (d_1, q_1, v_2) (and independence with $\mathbf{z}$). Then, with a normalization restriction, it can be shown that

$$E(y_1|\mathbf{z}, v_2) = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2)/(1 + \eta_1 y_2 + \lambda_1 y_2^2)^{1/2}], \tag{3.15}$$

which can be estimated by inserting $\hat{v}_2$ for v_2 and using nonlinear least squares or Bernoulli QMLE. (The latter is often called "heteroskedastic probit" when y_1 is binary.) This procedure can be viewed as an extension to Garen's method for linear models with correlated random coefficients.

Estimation, inference, and interpretation would be especially straightforward (the latter possibly using the bootstrap) if we squint and pretend the term $(1 + \eta_1 y_2 + \lambda_1 y_2^2)^{1/2}$ is not present. Then, estimation would simply be Bernoulli QMLE of y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, $\hat{v}_{i2}$, and $y_{i2}\hat{v}_{i2}$, which means that we just add the interaction to the usual Rivers-Vuong procedure. The APE for y_2 would be estimated by taking the derivative with respect to y_2 and averaging out $\hat{v}_{i2}$, as usual:

$$N^{-1}\sum_{i=1}^{N}(\hat{\alpha}_1 + \hat{\psi}_1\hat{v}_{i2}) \cdot \phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2 + \hat{\theta}_1\hat{v}_{i2} + \hat{\psi}_1 y_2\hat{v}_{i2}), \tag{3.16}$$

and evaluating this at chosen values for $(\mathbf{z}_1, y_2)$ (or using further averaging across the sample values). This simplification cannot be reconciled with (3.9), but it is in the spirit of adding flexibility to a standard approach and treating functional forms as approximations. As a practical matter, we can compare this with the APEs obtained from the standard Rivers-Vuong approach, and a simple test of the null hypothesis that the coefficient on y_2 is constant is $H_0 : \psi_1 = 0$ (which should account for the first step estimation of $\hat{\boldsymbol{\pi}}_2$). The null hypothesis that y_2 is exogenous is the joint test $H_0 : \theta_1 = 0, \psi_1 = 0$, and in this case no adjustment is needed for the first-stage estimation. And why stop here? If we, add, say, y_2^2 to the structural model, we might add $\hat{v}_2^2$ to the estimating equation as well. It would be very difficult to relate parameters estimated from the CF method to parameters in an underlying structural model; indeed, it would be difficult to find a structural model given rise to this particular CF approach. But if the object of interest are the average partial effects, the focus on flexible models for $E(y_1|\mathbf{z}_1, y_2, v_2)$ can be liberating (or disturbing, depending on one's point of view about "structural" parameters).

Lewbel (2000) has made some progress in estimating parameters up to scale in the model $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 > 0],$ where y_2 might be correlated with u_1 and $\mathbf{z}_1$ is a $1 \times L_1$ vector of exogenous variables. Lewbel's (2000) general approach applies to this situation as well. Let $\mathbf{z}$ be the vector of all exogenous variables uncorrelated with u_1. Then Lewbel requires a continuous element of $\mathbf{z}_1$ with nonzero coefficient – say the last element, z_{L_1}– that does not appear in $D(u_1|y_2, \mathbf{z})$. (Clearly, y_2 cannot play the role of the variable excluded from $D(u_1|y_2, \mathbf{z})$ if y_2 is thought to be endogenous.) When might Lewbel's exclusion restriction

hold? Sufficient is $y_2 = g_2(\mathbf{z}_2) + v_2,$ where (u_1, v_2) is independent of $\mathbf{z}$ and $\mathbf{z}_2$ does not contain z_{L_1}. But this means that we have imposed an exclusion restriction on the reduced form of y_2, something usually discouraged in parametric contexts. Randomization of z_{L_1} does *not* make its exclusion from the reduced form of y_2 legitimate; in fact, one often hopes that an instrument for y_2 is effectively randomized, which means that z_{L_1} does *not* appear in the structural equation but does appear in the reduced form of y_2 – the opposite of Lewbel's assumption. Lewbel's assumption on the "special" regressor is suited to cases where a quantity that only affects the response, y_1, is randomized. A randomly generated project cost presented to subjects in a willingness-to-pay study is one possibility. Even in such scenarios, one cannot identify the effects of covariates on willingness to pay because coefficients are identified only up to scale.

Returning to the probit response function in (3.9), we can understand the limits of the CF approach for estimating nonlinear models with discrete EEVs. The Rivers-Vuong approach, and its extension to fractional responses, cannot be expected to produce consistent estimates of the parameters or APEs for discrete y_2. The problem is that we cannot write

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{3.17}$$

$$D(v_2|\mathbf{z}) = D(v_2) = \text{Normal}(0, \tau_2^2). \tag{3.18}$$

In other words, unlike when we estimate a linear structural equation, the reduced form in the RV approach is not just a linear projection – far from it. In the extreme we have completely specified $D(y_2|\mathbf{z})$ as homoskedastic normal, which is clearly violated if y_2 is a binary variable, a count variable, or a corner solution (commonly called a "censored" variable). Unfortunately, even just assuming independence between v_2 and $\mathbf{z}$ rules out discrete y_2, an assumption that plays an important role even in fully nonparametric approaches. The bottom line is that there

are no known two-step estimation methods that allow one to estimate a probit model or fractional probit model with discrete y_2, even if we make strong distributional assumptions.

Possibly because of the absense of valid two-step methods with discrete EEVs, some poor strategies still linger. For example, suppose y_1 and y_2 are both binary, (3.1) holds, y_2 follows the index model

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0], \tag{3.19}$$

and we maintain joint normality of (u_1, v_2) – now both with unit variances – and, of course, independence between the errors and $\mathbf{z}$. Because $D(y_2|\mathbf{z})$ follows a standard probit, it is tempting to try to mimic 2SLS as follows: (i) Run probit of y_2 on $\mathbf{z}$ and get the fitted probabilities, $\hat{\Phi}_2 = \Phi(\mathbf{z}\hat{\boldsymbol{\delta}}_2)$. (ii) Run probit of y_1 on $\mathbf{z}_1, \hat{\Phi}_2$; that is, just replace each y_{i2} with its fitted probability, $\hat{\Phi}_{i2}$. This does not work, as it would require passing the expected value passes through a nonlinear function. Some have called prodedures like this a "forbidden regression." We could find $E(y_1|\mathbf{z},y_2)$ as a function of the structural and reduced form parameters, insert the first-stage estimates of the RF parameters, and then use binary response estimation in the second stage. But the estimator is not probit with the fitted probabilities plugged in for y_2. Currently, the only strategy we have is maximum likelihood estimation based on $f(y_1|y_2,\mathbf{z})f(y_2|\mathbf{z})$, which is not difficult. Wooldridge (2010, Section 15.7.3) contains the likelihood function. [The dearth of options that allow some robustness to distributional assumptions on y_2 helps explain why some authors, notably Angrist (2001), have promoted the idea of just using linear probability models estimated by 2SLS. This strategy seems to provide good estimates of the average treatment effect in many applications. But it also seems true that MLE based on joint normality might yield useful approximations to the APEs, too, even if the distributional functions are not entirely correct. Such a view argues for fully robust inference

in the context of misspecified maximum likelihood, as in White (1982).]

An issue that comes up occasionally is whether "bivariate probit" software can be used to estimate the probit model with a binary endogenous variable. In fact, the answer is yes, and the endogenous variables can appear in any way in the model, particularly interacted with exogenous variables. The key is that the likelihood function is constructed from $f(y_1|y_2,\mathbf{x}_1)f_2(y_2|\mathbf{x}_2)$, and so its form does not change if $\mathbf{x}_1$ includes y_2. (Of course, one should have at least one exclusion restriction in the case $\mathbf{x}_1$ does depend on y_2.) MLE, of course, has all of its desirable properties, and the parameter estimates needed to compute APEs are provided directly.

If y_1 is a fractional response satisfying (3.9), y_2 follows (3.19), and (q_1, v_2) are jointly normal and independent of $\mathbf{z}$, a two-step method based on $E(y_1|\mathbf{z},y_2)$ is possible; the expectation is not in closed form, and estimation cannot proceed by simply adding a control function to a Bernoulli QMLE. But it should not be difficult to implement. Full MLE for a fractional response is more difficult than for a binary response, particularly if y_1 takes on values at the endpoints with positive probability.

An essentially parallel discussion holds for ordered probit response models, where y_1 takes on the ordered values $\{0,1,\ldots,J\}$. The RV procedure, and its extensions, applies immediately. In computing partial effects on the response probabilities, we simply average out the reduced for residuals, as in equation (3.8). The comments about the forbidden regression are immediately applicable, too: one cannot simply insert, say, fitted probabilities for the binary EEV y_2 into an ordered probit model for y_1 and hope for consistent estimates of anything of interest.

Likewise, methods for Tobit models when y_1 is a corner solution, such as labor supply or

charitable contributions, are analyzed in a similar fashion. If y_2 is a continuous variable, CF methods for consistent estimation can be obtained, at least under the assumptions used in the RV setup. Smith and Blundell (1986) and Wooldridge (2010, Chapter 17) contain treatments. The embellishments described above, such as letting $D(u_1|v_2)$ be a flexible normal distribution, carry over immediately to Tobit case, as do the cautions in looking for simple two-step methods when $D(y_2|\mathbf{z})$ is discrete. Maximum likleihood estimation of all parameters jointly is also quite feasible.

3.2. Multinomial and Ordered Responses

Allowing endogenous explanatory variables (EEVs) in multinomial response models is notoriously difficult, even for continuous endogenous variables. There are two basic reasons. First, multinomial probit (MNP), which mixes well well a reduced form normality assumption for $D(y_2|\mathbf{z})$, is still computationally difficult for even a moderate number of choices. Apparently, no one has undertaken a systematic treatment of MNP with EEVs, including how to obtain partial effects.

The multinomial logit (MNL) model and its extensions, such as nested logit and random coefficient versions, are much simpler computationally with lots of alternatives. Unfortunately, the normal distribution does not mix well with the extreme value distribution, and so, if we begin with a structural MNL model (or conditional logit), the estimating equations obtained from a CF approach are difficult to obtain, and MLE is very difficult, too, even if we assume a normal distribution in the reduced form(s).

Recently, some authors have suggested taking a practical approach to allowing continuous EEVs in multinomial response. The suggestions for binary and fractional responses in the

previous subsection – namely, use probit, or even logit, with flexible functions of both the observed variables and the reduced form residuals – is in this spirit.

Again it is convenient to model the source of endogeneity as an omitted variable. Let y_1 be the (unordered) multinomial response taking values $\{0, 1, \ldots, J\}$, let $\mathbf{z}$ be the vector of endogenous variables, and let $\mathbf{y}_2$ be a vector of endogenous variables. If r_1 represents omitted factors that the researcher would like to control for, then the structural model consists of specifications for the response probabilities

$$\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1), j = 0, 1, \ldots, J. \tag{3.20}$$

The average partial effects, as usual, are obtained by averaging out the unobserved heterogeneity, r_1. Assume that $\mathbf{y}_2$ follows the linear reduced form

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2. \tag{3.21}$$

Typically, at least as a first attempt, we would assume a convenient joint distribution for $(r_1, \mathbf{v}_2)$, such as multivariate normal and independent of $\mathbf{z}$. This approach has been applied when the response probabilities, conditional on r_1, have the conditional logit form. For example, Villas-Boas and Winer (1999) apply this approach to modeling brand choice, where prices are allowed to correlated with unobserved tastes that affect brand choice. In implementing the CF approach, the problem in starting with a multinomial or conditional logit model for (3.20) is computational. Nevertheless, estimation is possible, particular if one uses simulation methods of estimation briefly mentioned in the previous subsection.

A much simpler control function approach is obtained if we skip the step of modeling $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ and jump directly to convenient models for $\mathrm{P}(y_1 = j|\mathbf{z}_{i1}, \mathbf{y}_2, \mathbf{v}_2) = \mathrm{P}(y_1 = j|\mathbf{z}, \mathbf{y}_2)$. Petrin and Train (2006) are proponents of this solution.

The idea is that any parametric model for $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ is essentially arbitrary, so, if we can recover quantities of interest directly from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, why not specify these probabilities directly? If we assume that $D(r_1|\mathbf{z}, \mathbf{y}_2) = D(r_1|\mathbf{v}_2)$, and that $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ can be obtained from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ by integrating the latter with respect to $D(r_1|\mathbf{v}_2)$ then we can estimate the APEs directly from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ by averaging out across the reduced form residuals, as in previous cases.

Once we have selected a model for $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, which could be multinomial logit, conditional logit, or nested logit, we can apply a simple two-step procedure. First, estimate the reduced form for $\mathbf{y}_{i2}$ and obtain the residuals, $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \mathbf{z}_i\hat{\mathbf{\Pi}}_2$. (Alternatively, we can use strictly monotonic transformations of the elements of $\mathbf{y}_{i2}$.) Then, we estimate a multinomial response model with explanatory variables $\mathbf{z}_{i1}, \mathbf{y}_{i2},$ and $\hat{\mathbf{v}}_{i2}$. As always with control function approaches, we need enough exclusion restrictions in $\mathbf{z}_{i1}$ to identify the parameters and APEs. We can include nonlinear functions of $(\mathbf{z}_{i1}, \mathbf{y}_{i2}, \hat{\mathbf{v}}_{i2})$, including quadratics and interactions for more flexibility.

Given estimates of the probabilities $p_j(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, we can estimate the average partial effects on the structural probabilities by estimating the average structural function:

$$\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2) = N^{-1} \sum_{i=1}^{N} p_j(\mathbf{z}_1, \mathbf{y}_2, \hat{\mathbf{v}}_{i2}). \tag{3.22}$$

Then, we can take derivatives or changes of $\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2)$ with respect to elements of $(\mathbf{z}_1, \mathbf{y}_2)$, as usual. While the delta method can be used to obtain analytical standard errors, the bootstrap is simpler and feasible if one uses, say, conditional logit.

In an application to choice of television service, Petrin and Train (2006) find the CF

approach gives remarkably similar parameter estimates to the approach proposed by Berry, Pakes, and Levinsohn (1995), which we touch on in the cluster sample notes.

When the EEVs are discrete, the CF arguments above do not apply. One can often implement maximum likelihood without too much difficulty. For example, Adams, Chiang, and Jensen (2003) use MLE when the scalar y_2 follows an ordered probit.

3.3. Exponential Models

Exponential models represent a middle ground between linear models and discrete response models: to allow for EEVs in an exponential model, we need to impose more assumptions than needed for standard linear models but fewer assumptions than discrete response models. Both IV approaches and CF approaches are available for exponential models, the latter having been worked out for continuous and binary EEVs. With a single EEV, write

$$E(y_1|\mathbf{z},y_2,r_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + r_1), \tag{3.23}$$

where r_1 is the omitted variable. (Extensions to general nonlinear functions of $(\mathbf{z}_1, y_2)$ are immediate; we just add those functions with linear coefficients to (3.23). Leading cases are polynomials and interactions.) Suppose first that y_2 has a standard linear reduced form with an additive, independent error:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{3.24}$$

$$D(r_1, v_2|\mathbf{z}) = D(r_1, v_2), \tag{3.25}$$

so that (r_1, v_2) is independent of $\mathbf{z}$. Then

$$E(y_1|\mathbf{z},y_2) = E(y_1|\mathbf{z}, v_2) = E[\exp(r_1)|v_2]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2). \tag{3.26}$$

If (r_1, v_2) are jointly normal, then $E[\exp(r_1)|v_2] = \exp(\theta_1 v_2)$, where we set the intercept to

zero, assuming $\mathbf{z}_1$ includes an intercept. This assumption can hold more generally, too. Then

$$E(y_1|\mathbf{z},y_2) = E(y_1|\mathbf{z},v_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2), \tag{3.27}$$

and this expectation immediately suggest a two-step estimation procedure. The first step, as before, is to estimate the reduced form for y_2 and obtain the residuals. Then, include $\hat{v}_2$, along with $\mathbf{z}_1$ and y_2, in nonlinear regression or, especially if y_1 is a count variable, in a Poisson QMLE analysis. Like NLS, it requires only (3.27) to hold. A t test of $H_0 : \theta_1 = 0$ is valid as a test that y_2 is exogenous. Average partial effects on the mean are obtained from

$$\left[N^{-1} \sum_{i=1}^{N} \exp(\hat{\theta}_1 \hat{v}_{i2}) \right] \exp(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2).$$

Proportionate effects on the expected value, that is elasticities and semi-elasticities, do not depend on the scale factor out front.

Like in the binary case, we can use a random coefficient model to suggest more flexible CF methods. For example, if we start with

$$\begin{aligned} E(y_1|\mathbf{z},y_2,a_1,r_1) &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + r_1) \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + d_1 y_2 + r_1) \end{aligned} \tag{3.28}$$

and assume trivariate normality of (d_1, r_1, v_2) (and independence from $\mathbf{z}$), then it can be shown that

$$\begin{aligned} E(y_1|\mathbf{z},v_2) = \exp(&\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 \\ &+ (\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2). \end{aligned} \tag{3.29}$$

Therefore, the estimating equation involves a quadratic in y_2 and an interaction between y_2 and v_2. Notice that the term $(\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2$ is present even if y_2 is exogenous, that is, $\theta_1 = \psi_1 = 0$. If $\sigma_{dr} = Cov(d_1, r_1) \neq 0$ then (3.29) does not even identify $\alpha_1 = E(a_1)$ (we

would have to use higher-order moments, such as a variance assumption). But (3.29) *does* identify the average structural function (and, therefore, APEs). We just absorb σ_r^2 into the intercept, combine the linear terms in y_2, and add the quadratic in y_2. So, we would estimate

$$E(y_1|\mathbf{z}, v_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \rho_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 + \eta_1 y_2^2) \tag{3.30}$$

using a two-step QMLE. The ASF is more complicated, and estimated as

$$\widehat{ASF}(\mathbf{z}_1, y_2) = \left[N^{-1} \sum_{i=1}^{N} \exp(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\rho}_1 y_2 + \hat{\theta}_1 \hat{v}_{i2} + \hat{\psi}_1 y_2 \hat{v}_{i2} + \hat{\eta}_1 y_2^2) \right], \tag{3.31}$$

which, as in the probit example, implies that the APE with respect to y_2 need not have the same sign as α_1.

Our inability to estimate α_1 even in this very parametric setting is just one example of how delicate identification of parameters in standard index models can be. Natural extensions to models with random slopes generally cause even the mean heterogeneity (α_1 above) to be unidentified. Again, it must be emphasized that the loss of identification holds even if y_2 is assumed exogenous.

If y_2 is a binary model following a probit, then a CF approach due to Terza (1998) can be used. We return to the model in (3.23) where, for simplicity, we assume y_2 is not interacted with elements of $\mathbf{z}_1$; the extension is immediate.We can no longer assume (3.24) and (3.25). Instead, replace (3.24)

$$y_2 = 1[\mathbf{z}\boldsymbol{\pi}_2 + v_2 > 0] \tag{3.32}$$

and still adopt (3.25). In fact, we assume (r_1, v_2) is jointly normal. To implement a CF approach, we need to find

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= E[E(y_1|\mathbf{z},v_2)|\mathbf{z},y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)E[\exp(\eta_1 + \theta_1 v_2)|\mathbf{z},y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2,\mathbf{z}\boldsymbol{\pi}_2,\theta_1), \end{aligned} \tag{3.34}$$

where we absorb η_1 into the intercept in $\mathbf{z}_1$ without changing notation and

$$\begin{aligned} h(y_2,\mathbf{z}\boldsymbol{\pi}_2,\theta_1) = \exp(\theta_1^2/2)\{&y_2\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2) \\ &+ (1-y_2)[1-\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1-\Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}, \end{aligned} \tag{3.35}$$

as shown by Terza (1998). Now, $\boldsymbol{\pi}_2$ is estimated by a first-stage probit, and then NLS or, say, Poisson QMLE can be applied to the mean function

$$\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2,\mathbf{z}\hat{\boldsymbol{\pi}}_2,\theta_1). \tag{3.36}$$

As usual, unless $\theta_1 = 0$, one must account for the estimation error in the first step when obtaining inference in the second. Terza (1998) contains analytical formulas, or one may use the bootstrap.

In the exponential case, an alternative to either of the control function approaches just presented is available – and, it produces consistent estimators regardless of the nature of y_2. Write $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1,\mathbf{y}_2)$ as any function of exogenous and endogenous variables. If we start with

$$E(y_1|\mathbf{z},\mathbf{y}_2,r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \tag{3.37}$$

then we can use a transformation due to Mullahy (1997) to consistently estimate $\boldsymbol{\beta}_1$ by method of moments. By definition, and assuming only that $y_1 \geq 0$, we can write

$$\begin{aligned} y_1 &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1)a_1 \\ &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1)\exp(r_1)a_1,\ E(a_1|\mathbf{z},\mathbf{y}_2,r_1) = 1. \end{aligned}$$

If r_1 is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(r_1)|\mathbf{z}] = E[\exp(r_1)] = 1, \tag{3.38}$$

where the last equality is just a normalization that defines the intercept in $\boldsymbol{\beta}_1$. Therefore, we have conditional moment conditions

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0, \tag{3.39}$$

which depends on the unknown parameters $\boldsymbol{\beta}_1$ and observable data. Any function of $\mathbf{z}$ can be used as instruments in a nonlinear GMM procedure. An important issue in implementing the procedure is choosing instruments. See Mullahy (1997) for further discussion.

4. Semiparametric and Nonparametric Approaches

Blundell and Powell (2004) show how to relax distributional assumptions on (u_1, v_2) in the model model $y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0]$, where $\mathbf{x}_1$ can be any function of $(\mathbf{z}_1, y_2)$. The key assumption is that y_2 can be written as $y_2 = g_2(\mathbf{z}) + v_2,$ where (u_1, v_2) is independent of $\mathbf{z}$. The independence of the additive error v_2 and $\mathbf{z}$ pretty much rules out discreteness in y_2, even though $g_2(\cdot)$ can be left unspecified. Under the independence assumption,

$$\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{E}(y_1|\mathbf{z}, v_2) = H(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{4.1}$$

for some (generally unknown) function $H(\cdot, \cdot)$. The average structural function is just $\mathrm{ASF}(\mathbf{z}_1, y_2) = \mathrm{E}_{v_{i2}}[H(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$. We can estimate H and $\boldsymbol{\beta}_1$ quite generally by first estimating the function $g_2(\cdot)$ and then obtaining residuals $\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i)$. Blundell and Powell (2004) show how to estimate H and $\boldsymbol{\beta}_1$ (up to scale) and $G(\cdot)$, the distribution of u_1. The ASF is obtained from $G(\mathbf{x}_1\boldsymbol{\beta}_1)$. We can also estimate the ASF by averaging out the reduced form residuals,

$$\widehat{\text{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \hat{H}(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_1, \hat{v}_{i2}); \tag{4.2}$$

derivatives and changes can be computed with respect to elements of $(\mathbf{z}_1, y_2)$.

Blundell and Powell (2003) allow $P(y_1 = 1|\mathbf{z}, y_2)$ to have the general form $H(\mathbf{z}_1, y_2, v_2)$, and then the second-step estimation is entirely nonparametric. They also allow $\hat{g}_2(\cdot)$ to be fully nonparametric. But parametric approximations in each stage might produce good estimates of the APEs. For example, y_{i2} can be regressed on flexible functions of $\mathbf{z}_i$ to obtain $\hat{v}_{i2}$. Then, one can estimate probit or logit models in the second stage that include functions of $\mathbf{z}_1$, y_2, and $\hat{v}_2$ in a flexible way – for example, with levels, quadratics, interactions, and maybe even higher-order polynomials of each. Then, one simply averages out $\hat{v}_{i2}$, as in equation (4.2). Valid standard errors and test statistics can be obtained by bootstrapping or by using the delta method.

In certain cases, an even more parametric approach suggests itself. Suppose we have the exponential regression

$$E(y_1|\mathbf{z}, y_2, r_1) = \exp(\mathbf{x}_1 \boldsymbol{\beta}_1 + r_1), \tag{4.3}$$

where r_1 is the unobservable. If $y_2 = \mathbf{g}_2(\mathbf{z})\boldsymbol{\pi}_2 + v_2$ and (r_1, v_2) is independent of $\mathbf{z}$, then

$$E(y_1|\mathbf{z}_1, y_2, v_2) = h_2(v_2) \exp(\mathbf{x}_1 \boldsymbol{\beta}_1), \tag{4.4}$$

where now $h_2(\cdot)$ is an unknown function. It can be approximated using series, say, and, of course, first-stage residuals $\hat{v}_2$ replace v_2.

Blundell and Powell (2003) consider a very general setup, which starts with $y_1 = g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)$, and then discuss estimation of the ASF, given by

$$ASF_1(\mathbf{z}_1, \mathbf{y}_2) = \int g_1(\mathbf{z}_1, \mathbf{y}_2, u_1) dF_1(u_1), \tag{4.5}$$

where F_1 is the distribution of u_1. The key restrictions are that $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2, \tag{4.6}$$

where $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$. The additive, independent reduced form errors in (4.6) effectively rule out applications to discrete $\mathbf{y}_2$. Conceptually, Blundell and Powell's method is straightforward, as it is a nonparametric extenstion of parametric approaches. First, estimate $\mathbf{g}_2$ nonparametrically (which, in fact, may be done via a flexible parametric model, or kernel estimators). Obtain the residuals $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \hat{\mathbf{g}}_2(\mathbf{z}_i)$. Next, estimate $E(y_1|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2) = h_1(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ using nonparametrics, where $\hat{\mathbf{v}}_{i2}$ replaces $\mathbf{v}_2$. Identification of h_1 holds quite generally, provided we have sufficient exclusion restrictions (elements in $\mathbf{z}$ not in $\mathbf{z}_1$). BP discuss some potential pitfalls. Once we have $\hat{h}_1$, we can consistently estimate the ASF. For given $\mathbf{x}_1^o = (\mathbf{z}_1^o, \mathbf{y}_2^o)$, the ASF can always be written, using iterated expectations, as

$$E_{\mathbf{v}_2}\{E[g_1(\mathbf{x}_1^o, u_1)|\mathbf{v}_2]\}.$$

Under the assumption that $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$, $E[g_1(\mathbf{x}_1^o, u_1)|\mathbf{v}_2] = h_1(\mathbf{x}_1^o, \mathbf{v}_2)$ – that is, the regression function of y_1 on $(\mathbf{x}_1, \mathbf{v}_2)$. Therefore, a consistent estimate of the ASF is

$$N^{-1}\sum_{i=1}^{N} \hat{h}_1(\mathbf{x}_1, \hat{\mathbf{v}}_{i2}). \tag{4.7}$$

While semiparametric and parametric methods when y_2 (or, more generally, a vector $\mathbf{y}_2$) are continuous – actually, have a reduced form with an additive, independent error – they do not currently help us with discrete EEVs.

With univariate y_2, it possible to relax the additivity of v_2 in the reduced form equation under monotonicity assumptions. Like Blundell and Powell (2003), Imbens and Newey (2006) consider the triangular system, but without additivity in the reduced form of y_2. The structureal

equation is

$$y_1 = g_1(\mathbf{z}_1, y_2, \mathbf{u}_1), \tag{4.8}$$

where $\mathbf{u}_1$ is a vector heterogeneity (whose dimension may not even be known), and the reduced form for y_2 is

$$y_2 = g_2(\mathbf{z}, e_2), \tag{4.9}$$

where $g_2(\mathbf{z}, \cdot)$ is strictly monotonic. This assumption rules out discrete y_2 but allows some interaction between the unobserved heterogeneity in y_2 and the exogenous variables. As one special case, Imbens and Newey show that, if $(\mathbf{u}_1, e_2)$ is assumed to be independent of $\mathbf{z}$, then a valid control function that can be used in a second stage is $v_2 \equiv F_{y_2|\mathbf{z}}(y_2, \mathbf{z})$, where $F_{y_2|\mathbf{z}}$ is the conditional distribution of y_2 given $\mathbf{z}$. Imbens and Newey described identification of various quantities of interest, including the quantile structural function. When u_1 is a scalar and monotonically increasing in u_1, the QSF is

$$QSF_\tau(\mathbf{x}_1) = g_1(\mathbf{x}_1, \text{Quant}_\tau(u_1)), \tag{4.10}$$

where $\text{Quant}_\tau(u_1)$ is the τ^{th} of u_1. We consider quantile methods in more detail in the quantile methods notes.

5. Methods for Panel Data

We can combine methods for handling correlated random effects models with control function methods to estimate certain nonlinear panel data models with unobserved heterogeneity and EEVs. Here we provide as an illustration a parametric approach used by Papke and Wooldridge (2008), which applies to binary and fractional responses. The

manipulations are routine but point to more flexible ways of estimating the average marginal effects. It is important to remember that we currently have no way of estimating, say, unobserved effects models for fractional response variables, either with or without endogenous explanatory variables, without imposing some restrictions on the distribution of heterogeneity given the exogenous variables. Even the approaches that treat the unobserved effects as parameters – and use large T approximations – to not allow endogenous regressors. Plus, recall from the nonlinear panel data notes that most results are for the case where the data are assumed independent across time. Jackknife approaches further assume homogeneity across time.

We write the model with time-constant unobserved heterogeneity, c_{i1}, and time-varying unobservables, v_{it1}, as

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) = E(y_{it1}|y_{it2}, \mathbf{z}_{it1}, c_{i1}, v_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1}). \tag{5.1}$$

Thus, there are two kinds of potential omitted variables. We allow the heterogeneity, c_{i1}, to be correlated with y_{it2} and $\mathbf{z}_i$, where $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ is the vector of strictly exogenous variables (conditional on c_{i1}). The time-varying omitted variable is uncorrelated with $\mathbf{z}_i$ – strict exogeneity – but may be correlated with y_{it2}. As an example, y_{it1} is a female labor force participation indicator and y_{it2} is other sources of income. Or, y_{it1} is a test pass rate, and the school leve, and y_{it2} is a measure of spending per student.

When we write $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$, we are assuming $\mathbf{z}_{it2}$ can be excluded from the "structural" equation (4.1). This is the same as the requirement for fixed effects two stage least squares estimation of a linear model.

To proceed, we first model the heterogeneity using a Chamberlain-Mundlak approach:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}, a_{i1}|\mathbf{z}_i \sim \text{Normal}(0, \sigma^2_{a_1}). \tag{5.2}$$

We could allow the elements of $\mathbf{z}_i$ to appear with separate coefficients, too. Note that only exogenous variables are included in $\bar{\mathbf{z}}_i$. Plugging into (5.1) we have

$$\begin{aligned} E(y_{it1}|y_{it2}, \mathbf{z}_i, a_{i1}, v_{it1}) &= \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + v_{it1}) \\ &\equiv \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + r_{it1}). \end{aligned} \tag{5.3}$$

Next, we assume a linear reduced form for y_{it2}:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\delta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T, \tag{5.4}$$

where, if necessary, we can allow the coefficients in (5.4) to depend on t. The addition of the time average of the strictly exogenous variables in (5.4) follows from the Mundlak (1978) device. The nature of endogeneity of y_{it2} is through correlation between $r_{it1} = a_{i1} + v_{it1}$ and the reduced form error, v_{it2}. Thus, y_{it2} is allowed to be correlated with unobserved heterogeneity and the time-varying omitted factor. We also assume that r_{it1} given v_{it2} is conditionally normal, which we write as

$$r_{it1} = \eta_1 v_{it2} + e_{it1}, \tag{5.5}$$

$$e_{it1}|(\mathbf{z}_i, v_{it2}) \sim \text{Normal}(0, \sigma^2_{e_1}), t = 1, \ldots, T. \tag{5.6}$$

Because e_{it1} is independent of $(\mathbf{z}_i, v_{it2})$, it is also independent of y_{it2}. Using a standard mixing property of the normal distribution,

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = \Phi(\alpha_{e1} y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2}) \tag{5.7}$$

where the "e" subscript denotes division by $(1 + \sigma^2_{e_1})^{1/2}$. This equation is the basis for CF estimation.

The assumptions used to obtain (5.7) would not hold for y_{it2} having discreteness or

substantively limited range in its distribution. It is straightfoward to include powers of v_{it2} in (5.7) to allow greater flexibility. Following Wooldridge (2005) for the cross-sectional case, we could even model r_{it1} given v_{it2} as a heteroskedastic normal.

In deciding on estimators of the parameters in (5.7), we must note that the explanatory variables, while contemporaneous exogenous by construction, are not usually strictly exogenous. In particular, we allow y_{is2} to be correlated with v_{it1} for $t \neq s$. Therefore, generalized estimation equations, that assume strict exogeneity – see the notes on nonlinear panel data models – will not be consistent in general. We could apply method of moments procedures. A simple approach is to use use pooled nonlinear least squares or pooled quasi-MLE, using the Bernoulli log likelihood. (The latter fall under the rubric of generalized linear models.) Of course, we want to allow arbitrary serial dependence and $Var(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2})$ in obtaining inference, which means using a robust sandwich estimator.

The two step procedure is (i) Estimate the reduced form for y_{it2} (pooled across t, or maybe for each t separately; at a minimum, different time period intercepts should be allowed). Obtain the residuals, $\hat{v}_{it2}$ for all (i,t) pairs. The estimate of $\boldsymbol{\delta}_2$ is the fixed effects estimate. (ii) Use the pooled probit QMLE of y_{it1} on $y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, \hat{v}_{it2}$ to estimate $\alpha_{e1}, \boldsymbol{\delta}_{e1}, \psi_{e1}, \boldsymbol{\xi}_{e1}$ and η_{e1}.

Because of the two-step procedure, the standard errors in the second stage should be adjusted for the first stage estimation. Alternatively, bootstrapping can be used by resampling the cross-sectional units. Conveniently, if $\eta_{e1} = 0$, the first stage estimation can be ignored, at least using first-order asymptotics. Consequently, a test for endogeneity of y_{it2} is easily obtained as an asymptotic t statistic on $\hat{v}_{it2}$; it should be make robust to arbitrary serial correlation and misspecified variance. Adding first-stage residuals to test for endogeneity of an explanatory variables dates back to Hausman (1978). In a cross-sectional contexts, Rivers and

Vuong (1988) suggested it for the probit model.

Estimates of average partial effects are based on the average structural function

$$E_{(c_{i1},v_{it1})}[\Phi(\alpha_1 y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1})] \tag{5.8}$$

with respect to the elements of $(y_{t2}, \mathbf{z}_{t1})$. It can be shown that

$$E_{(\bar{\mathbf{z}}_i,v_{it2})}[\Phi(\alpha_{e1} y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2})]; \tag{5.9}$$

that is, we "integrate out" $(\bar{\mathbf{z}}_i, v_{it2})$ and then take derivatives or changes with respect to the elements of $(\mathbf{z}_{t1} y_{t2})$. Because we are not making a distributional assumption about $(\bar{\mathbf{z}}_i, v_{it2})$, we instead estimate the APEs by averaging out $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ across the sample, for a chosen *t*:

$$N^{-1}\sum_{i=1}^{N} \Phi(\hat{\alpha}_{e1} y_{t2} + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{e1} + \hat{\psi}_{e1} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{e1} + \hat{\eta}_{e1}\hat{v}_{it2}). \tag{5.10}$$

APEs computed from (5.10) – typically with further averaging out across *t* and the values of y_{t2} and $\mathbf{z}_{t1}$ – can be compared directly with linear model estimates, particular fixed effects IV estimates.

We can use the approaches of Altonji and Matzkin (2005) and Blundell and Powell (2003) to make the analysis less parametric. For example, we might replace (5.4) with $y_{it2} = g_2(\mathbf{z}_{it}, \bar{\mathbf{z}}_i) + v_{it2}$ (or use functions in addition to $, \bar{\mathbf{z}}_i$, as in AM). Then, we could maintain

$$D(c_{i1} + v_{it1}|\mathbf{z}_i, y_{it2}) = D(c_{i1} + v_{it1}|\bar{\mathbf{z}}_i, v_{it2}).$$

In the first estimation step, $\hat{v}_{it2}$ is obtained from a nonparametric or semiparametric pooled estimation. Then the function

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, v_{it2}) = h_1(\mathbf{x}_{it1}\boldsymbol{\beta}_1, \bar{\mathbf{z}}_i, v_{it2})$$

can be estimated in a second stage, with the first-stage residuals, $\hat{v}_{it2}$, inserted. Generally,

identification holds because the v_{it2} varying over time separately from $\mathbf{x}_{it1}$ due to time-varying exogenous instruments $\mathbf{z}_{it2}$. The inclusion of $\bar{\mathbf{z}}_i$ requires that we have at least one time-varying, strictly exogenous instrument for y_{it2}.

References

Adams, J.D., E.P. Chiang, and J.L. Jensen (2003), "The Influence of Federal Laboratory R&D on Industrial Research," *Review of Economics and Statistics* 85, 1003-1020.

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

Angrist, J.D. (1991), "Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics* 19, 2-16.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

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

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

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69, 1127-1160.

Garen, J. (1984), "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica* 52, 1199-1218.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251-1271.

Heckman, J.J. (1976), "The Common Structure of Statistical Models of Truncation, Sample

Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475-492.

Heckman, J.J. and T.E. MaCurdy (1986), "Labor Econometrics," in *Handbook of Econometrics*, Volume 3. Z. Griliches and M.D. Intriligator (eds.), 1918-1977. Amsterdam: Elsevier.

Heckman, J.J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling," *Journal of Human Resources* 33, 974-987.

Imbens, G.W. and W.K. Newey (2006), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," mimeo, MIT Department of Economics.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics* 97, 145-177.

Mullahy, J. (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics* 79, 586-593.

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

Newey, W.K. (1988), "Adaptive Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics* 38, 301-339.

Papke, L.E. and J.M. Wooldridge (2008), "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," forthcoming, *Journal of Econometrics*.

Petrin, A. and K. Train (2006), "Control Function Corrections for Unobserved Factors in Differentiated Product Models," mimeo, University of Minnesota Department of Economics.

Powell, J.L. (1994), "Estimation of Semiparametric Models," in *Handbook of Econometrics*, Volume 4. R.F. Engle and D.L. McFadden (eds.), 2443-2521. Amsterdam: Elsevier.

Rivers, D. and Q.H. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics* 39, 347-366.

Smith, R.J., and R.W. Blundell (1986), "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica* 54, 679-685.

Terza, J.V. (1998), "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects," *Journal of Econometrics* 84, 129-154.

Villas-Boas, J.M. and R.S. Winer (1999), "Endogeneity in Brand Choice Models," *Management Science* 45, 1324-1338.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 1-25.

Wooldridge, J.M. (1997), "On Two Stage Least Squares Estimation of the Average Treatment Effect in Random Coefficient Models," *Economics Letters* 56, 129-133.

Wooldridge, J.M. (2003), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters* 79, 185-191.

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

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. MIT Press: Cambridge, MA.

Cross-Section Econometrics
Lecture 8: Control Function Methods

Jeff Wooldridge
Michigan State University

AEA Lectures, Chicago, January 2012

1. Linear-in-Parameters Models: IV versus Control Functions
2. Correlated Random Coefficient Models
3. Nonlinear Models
4. Semiparametric and Nonparametric Approaches
5. Methods for Panel Data

1. Linear-in-Parameters Models: IV versus Control Functions

- Most models that are linear in parameters are estimated using standard IV methods – two stage least squares (2SLS) or generalized method of moments (GMM).
- An alternative, the control function (CF) approach, relies on the same kinds of identification conditions. In models with nonlinearities or random coefficients, the form of exogeneity is stronger and more restrictions are imposed on the reduced forms.

- Let y_1 be the response variable, y_2 the endogenous explanatory variable (EEV), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (with $z_1 = 1$):

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$.

- First consider the exogeneity assumption

$$E(\mathbf{z}'u_1) = \mathbf{0}. \tag{2}$$

Reduced form for y_2:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, \; E(\mathbf{z}'v_2) = \mathbf{0} \tag{3}$$

where $\boldsymbol{\pi}_2$ is $L \times 1$. Write the linear projection of u_1 on v_2, in error form, as

$$u_1 = \rho_1 v_2 + e_1, \tag{4}$$

where $\rho_1 = E(v_2u_1)/E(v_2^2)$ is the population regression coefficient. By construction, $E(v_2e_1) = 0$ and $E(\mathbf{z}'e_1) = \mathbf{0}$.

Plug $u_1 = \rho_1 v_2 + e_1$ into $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{5}$$

where v_2 is an explanatory variable in the equation. The new error, e_1, is uncorrelated with y_2 as well as with v_2 and $\mathbf{z}$.

• Two-step procedure: (i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the reduced form residuals, $\hat{v}_{i2}$; (ii) Regress

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, \text{ and } \hat{v}_{i2}. \tag{6}$$

The implicit error in (6) is $e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, which depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$ (exogeneity test). OLS estimators from (6) will be consistent for $\boldsymbol{\delta}_1, \alpha_1$, and ρ_1.

• The OLS estimates from (6) are *control function* estimates.

• The OLS estimates of $\boldsymbol{\delta}_1$ and α_1 from (6) are *identical* to the 2SLS estimates of

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$$

• Now extend the model:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{7}$$

$$E(u_1|\mathbf{z}) = 0. \tag{8}$$

Let z_2 be a scalar not also in $\mathbf{z}_1$. Under (8) – which is stronger than $E(\mathbf{z}'u_1) = \mathbf{0}$, and is essential for nonlinear models – we can use, say, z_2^2 as an instrument for y_2^2. So the IVs would be $(\mathbf{z}_1, z_2, z_2^2)$ for $(\mathbf{z}_1, y_2, y_2^2)$.

• What does CF approach entail? Now *assume*

$$E(u_1|\mathbf{z}, y_2) = E(u_1|v_2) = \rho_1 v_2, \tag{9}$$

where independence of (u_1, v_2) and $\mathbf{z}$ is sufficient for the first equality; linearity is a substantive restriction. Then,

$$E(y_1|\mathbf{z}, y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2, \tag{10}$$

and a CF approach is immediate: replace v_2 with $\hat{v}_2$ and use OLS on (10). *Not* equivalent to a 2SLS estimate. CF likely more efficient but less robust.

• In general, CF approaches impose extra assumptions when we base it on $E(y_1|\mathbf{z},y_2)$. The estimating equation is

$$E(y_1|\mathbf{z},y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z},y_2). \quad (11)$$

If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$, (u_1, e_2) is independent of $\mathbf{z}$, $E(u_1|e_2) = \rho_1 e_2$, and $e_2 \sim Normal(0,1)$, then

$$E(u_1|\mathbf{z},y_2) = \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \quad (12)$$

where $\lambda(\cdot)$ is the inverse Mills ratio (IMR).

• Heckman two-step approach (for endogeneity, not sample selection):

(i) Probit to get $\hat{\boldsymbol{\delta}}_2$ and compute the *generalized residuals*,

$\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$.

(ii) Regress y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, $\widehat{gr}_{i2}$, $i = 1,\ldots,N$.

• Consistency of the CF estimators hinges on the model for $D(y_2|\mathbf{z})$ being correctly specified, along with linearity in $E(u_1|e_2)$, where $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$. If we just apply 2SLS directly to $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$, it makes no distinction among discrete, continuous, or some mixture for y_2.

• How might we use the binary nature of y_2 in IV estimation in a robust manner?

(i) Obtain the fitted probabilities, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, from the first stage probit.

(ii) Estimate $y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_{i1}y_2 + u_{i1}$ by IV using $[\mathbf{z}_{i1}, \Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)]$ as instruments (not regressors!)

• Fully robust to misspecification of the probit model, usual standard errors from IV asymptotically valid. Efficient IV estimator if $P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$.

2. Correlated Random Coefficient Models

• Modify the original equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + u_1, \tag{13}$$

where a_1, the "random coefficient" on y_2. Heckman and Vytlacil (1998) call (13) a correlated random coefficient (CRC) model.

• Write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = E(a_1)$ is the object of interest. We can rewrite the equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \tag{14}$$

$$\equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1. \tag{15}$$

• Potential problem with applying IV: the error term $v_1 y_2 + u_1$ is not necessarily uncorrelated with the instruments $\mathbf{z}$, even under

$$E(u_1|\mathbf{z}) = E(v_1|\mathbf{z}) = 0. \tag{16}$$

We want to allow y_2 and v_1 to be correlated, $Cov(v_1, y_2) \equiv \tau_1 \neq 0$. A sufficient condition that allows for any *unconditional* correlation is

$$Cov(v_1, y_2|\mathbf{z}) = Cov(v_1, y_2), \tag{17}$$

and this is sufficient for IV to consistently estimate $(\alpha_1, \boldsymbol{\delta}_1)$.

• The usual IV estimator that ignores the randomness in a_1 is more robust than Garen's (1984) CF estimator, which adds $\hat{v}_2$ and $\hat{v}_2 y_2$ to the original model, or the Heckman/Vytlacil (1998) "plug-in" estimator, which replaces y_2 with $\hat{y}_2 = \mathbf{z}\hat{\boldsymbol{\pi}}_2$.

• The condition $Cov(v_1, y_2|\mathbf{z}) = Cov(v_1, y_2)$ cannot really hold for discrete y_2. Further, Card (2001) shows how it can be violated even if y_2 is continuous. Wooldridge (2005) shows how to allow parametric heteroskedasticity.

• In the case of binary y_2, we have what is often called the "switching regression" model. If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$ and $e_2|\mathbf{z} \sim Normal(0,1)$, then

$$E(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) y_2,$$

where

$$h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) = y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)$$

is the generalized residual function.

3. Nonlinear Models

• Typically three approaches to nonlinear models with EEVs.

(1) Plug in fitted values from a first step regression (in an attempt to mimic 2SLS in linear model). More generally, try to find $E(y_1|\mathbf{z})$ or $D(y_1|\mathbf{z})$ and then impose identifying restrictions.

(2) CF approach: plug in residuals in an attempt to obtain $E(y_1|y_2,\mathbf{z})$ or $D(y_1|y_2,\mathbf{z})$.

(3) Maximum Likelihood (often limited information): Use models for $D(y_1|y_2,\mathbf{z})$ and $D(y_2|\mathbf{z})$ jointly.

• All strategies are more difficult with nonlinear models when y_2 is discrete. Some poor practices have lingered.

Binary and Fractional Responses

Probit model:

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \geq 0], \tag{18}$$

where $u_1|z \sim \text{Normal}(0,1)$. Analysis goes through if we replace $(\mathbf{z}_1, y_2)$ with any known function $\mathbf{x}_1 \equiv \mathbf{g}_1(\mathbf{z}_1, y_2)$.

• The Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for y_2,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, \;\; v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2). \tag{19}$$

• RV approach comes close to requiring

$$(u_1, v_2) \text{ independent of } \mathbf{z}. \tag{20}$$

If we also assume

$$(u_1, v_2) \sim \text{Bivariate Normal} \tag{21}$$

with $\rho_1 = Corr(u_1, v_2)$, then we can proceed with MLE based on $f(y_1, y_2|\mathbf{z})$. A CF approach is available, too, based on

$$P(y_1 = 1|\mathbf{z}, y_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\rho 1} + \alpha_{\rho 1} y_2 + \theta_{\rho 1} v_2) \tag{22}$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$.

The Rivers-Vuong two-step approach is

(i) OLS of y_{i2} on $\mathbf{z}_i$, to obtain the residuals, $\hat{v}_2$.

(ii) Probit of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$ to estimate the scaled coefficients. A simple t test on $\hat{v}_2$ is valid to test $H_0 : \rho_1 = 0$.

• Can recover the original coefficients, which appear in the partial effects. Or,

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1} \hat{v}_{i2}), \tag{23}$$

that is, we average out the reduced form residuals, $\hat{v}_{i2}$. This formulation is useful for more complicated models.

• The two-step CF approach easily extends to fractional responses:

$$E(y_1|\mathbf{z}, y_2, q_1) = \Phi(\mathbf{x}_1 \boldsymbol{\beta}_1 + q_1), \tag{24}$$

where $\mathbf{x}_1$ is a function of $(\mathbf{z}_1, y_2)$ and q_1 contains unobservables.

• Use the *same* two-step estimator as for probit. APEs must be obtained from (23). In inference, we should only assume the mean is correctly specified.method can be used in the binary and fractional cases. To account for first-stage estimation, the bootstrap is convenient.

• Wooldridge (2005) describes some simple ways to make the analysis starting from (24) more flexible, including allowing $Var(q_1|v_2)$ to be heteroskedastic.

• CF has advantages over "fitted value" approach. Rather than conditioning on v_2 along with $\mathbf{z}$ (and therefore y_2) to obtain $P(y_1 = 1|\mathbf{z}, v_2) = P(y_1 = 1|\mathbf{z}, y_2, v_2)$, use

$$P(y_1 = 1|\mathbf{z}) = \Phi\{[\mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2)]/\omega_1\}$$
$$\omega_1^2 = Var(\alpha_1 v_2 + u_1)$$

(i) OLS on the reduced form, and get fitted values, $\hat{y}_{i2} = \mathbf{z}_i \hat{\boldsymbol{\delta}}_2$. (ii) Probit of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. Harder to estimate APEs and test for endogeneity.

• Danger with plugging in fitted values for y_2 is that one might be tempted to plug $\hat{y}_2$ into nonlinear functions, say y_2^2 or $y_2 \mathbf{z}_1$. This does *not* result in consistent estimation of the scaled parameters or the partial effects. Adding the CF $\hat{v}_2$ solves the endogeneity problem regardless of how y_2 appears.

• There are limits to the CF approach. Consider

$$E(y_1|\mathbf{z},y_2,q_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + q_1)$$

where y_2 is discrete. Rivers-Vuong approach does not generally work.

• Neither does plugging in probit fitted values, assuming

$$P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2) \tag{25}$$

In other words, do *not* try to mimic 2SLS as follows: (i) Do probit of y_2 on $\mathbf{z}$ and get the fitted probabilities, $\hat{\Phi}_2 = \Phi(\mathbf{z}\hat{\boldsymbol{\delta}}_2)$. (ii) Do probit of y_1 on $\mathbf{z}_1, \hat{\Phi}_2$, that is, just replace y_2 with $\hat{\Phi}_2$.

• The only strategy that works under traditional assumptions is maximum likelihood estimation based on $f(y_1|y_2,\mathbf{z})f(y_2|\mathbf{z})$. [Perhaps this is why some, such as Angrist (2001), promote the notion of just using linear probability models estimated by 2SLS.]

• "Bivariate probit" software can be used to estimate the probit model with a binary endogenous variable.

• A CF approach might work well for "small" amounts of endogeneity.

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \geq 0]$$
$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]$$

Can show – Wooldridge (2011) – that the score test for $H_0 : Cov(u_1, v_2) = 0$ is obtained as

(i) Probit of y_{i2} on $\mathbf{z}_i$ to get the generalized residuals,

$$\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$$

(ii) Probit of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \widehat{gr}_{i2}$ and use t statistic on $\widehat{gr}_{i2}$.

Might this account for endogeneity if it is not "too severe"?

Multinomial Responses

• Recent push by Petrin and Train (2010), among others, to use control function methods where the second step estimation is something simple – such as multinomial logit, or nested logit – rather than being derived from a structural model. So, if we have reduced forms

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2, \tag{26}$$

then we jump directly to convenient models for $P(y_1 = j|\mathbf{z}_1,\mathbf{y}_2,\mathbf{v}_2)$. The average structural functions are obtained by averaging the response probabilities across $\hat{\mathbf{v}}_{i2}$.

• No generally acceptable way to handle discrete $\mathbf{y}_2$, except by specifying a full set of distributions.

• Might approximate by adding generalized residuals as control functions to standard models (such as MNL).

4. Semiparametric and Nonparametric Approaches

• Blundell and Powell (2004) show how to relax distributional assumptions on (u_1, v_2) in the model $y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0]$, where $\mathbf{x}_1$ can be any function of $(\mathbf{z}_1, y_2)$. Their key assumption is that y_2 can be written as $y_2 = g_2(\mathbf{z}) + v_2$, where (u_1, v_2) is independent of $\mathbf{z}$, which rules out discreteness in y_2. Then

$$P(y_1 = 1|\mathbf{z}, v_2) = E(y_1|\mathbf{z}, v_2) = H(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{31}$$

for some (generally unknown) function $H(\cdot, \cdot)$. The average structural function is just $ASF(\mathbf{z}_1, y_2) = E_{v_{i2}}[H(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$.

• Two-step estimation: Estimate the function $g_2(\cdot)$ and then obtain residuals $\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i)$. BP (2004) show how to estimate H and $\boldsymbol{\beta}_1$ (up to scaled) and $G(\cdot)$, the distribution of u_1. The ASF is obtained from $G(\mathbf{x}_1\boldsymbol{\beta}_1)$ or

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\hat{H}(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1, \hat{v}_{i2}); \tag{32}$$

• Blundell and Powell (2003) allow $P(y_1 = 1|\mathbf{z}, y_2)$ to have general form $H(\mathbf{z}_1, y_2, v_2)$, and the second-step estimation is entirely nonparametric. Further, $\hat{g}_2(\cdot)$ can be fully nonparametric. Parametric approximations might produce good estimates of the APEs.

• BP (2003) consider a very general setup: $y_1 = g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)$, with

$$ASF_1(\mathbf{z}_1, \mathbf{y}_2) = \int g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)dF_1(u_1), \tag{33}$$

where F_1 is the distribution of u_1. Key restrictions are that $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2, \tag{34}$$

where $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$.

• Key Result:

$$ASF_1(\mathbf{z}_1, \mathbf{y}_2) = E_{\mathbf{v}_{i2}}[h_1(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_{i2})]$$

where $E(y_1|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2) = h_1(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$.

$$\widehat{ASF}_1(\mathbf{z}_1, \mathbf{y}_2) = N^{-1} \sum_{i=1}^{N} h_1(\mathbf{z}_1, \mathbf{y}_2, \hat{\mathbf{v}}_{i2})$$

• Fully nonparametric two-step estimates are available.

• Can justify flexible parametric approaches and just skip modeling $g_1(\cdot)$.

5. Methods for Panel Data

• Combine methods for correlated random effects models with CF methods for nonlinear panel data models with unobserved heterogeneity and EEVs.

• Illustrate a parametric approach used by Papke and Wooldridge (2008), which applies to binary and fractional responses.

• Nothing appears to be known about applying "fixed effects" probit to estimate the fixed effects while also dealing with endogeneity. Likely to be poor for small T.

• Model with time-constant unobserved heterogeneity, c_{i1}, and time-varying unobservables, v_{it1}, as

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1}). \tag{35}$$

Allow the heterogeneity, c_{i1}, to be correlated with y_{it2} and $\mathbf{z}_i$, where $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ is the vector of strictly exogenous variables (conditional on c_{i1}). The time-varying omitted variable, v_{it1}, is uncorrelated with $\mathbf{z}_i$ – strict exogeneity – but may be correlated with y_{it2}. As an example, y_{it1} is a female labor force participation indicator and y_{it2} is other sources of income.

• Write $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$, so that the time-varying IVs $\mathbf{z}_{it2}$ are excluded from the "structural."

• Chamberlain approach:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}, a_{i1}|\mathbf{z}_i \sim Normal(0, \sigma^2_{a_1}). \tag{36}$$

Next step:

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, r_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + r_{it1})$$

where $r_{it1} = a_{i1} + v_{it1}$. Next, assume a linear reduced form for y_{it2}:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\delta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T. \tag{37}$$

• Rules out discrete y_{it2} because

$$r_{it1} = \eta_1 v_{it2} + e_{it1}, \tag{38}$$

$$e_{it1} | (\mathbf{z}_i, v_{it2}) \sim Normal(0, \sigma^2_{e_1}), t = 1, \ldots, T. \tag{39}$$

Then

$$E(y_{it1} | \mathbf{z}_i, y_{it2}, v_{it2}) = \Phi(\alpha_{e1} y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i \boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2}) \tag{40}$$

where the "e" subscript denotes division by $(1 + \sigma^2_{e_1})^{1/2}$. This equation is the basis for CF estimation.

• Simple two-step procedure: (i) Estimate the reduced form for y_{it2} (pooled across t, or maybe for each t separately; at a minimum, different time period intercepts should be allowed). Obtain the residuals, $\hat{v}_{it2}$ for all (i,t) pairs. The estimate of $\boldsymbol{\delta}_2$ is the fixed effects estimate. (ii) Use the pooled probit (quasi)-MLE of y_{it1} on $y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, \hat{v}_{it2}$ to estimate $\alpha_{e1}, \boldsymbol{\delta}_{e1}, \psi_{e1}, \boldsymbol{\xi}_{e1}$ and η_{e1}.

• Delta method or bootstrapping (resampling cross section units) for standard errors. Can ignore first-stage estimation to test $\eta_{e1} = 0$ (but test should be fully robust to variance misspecification and serial independence).

• Estimates of average partial effects are based on the average structural function,

$$E_{(c_{i1}, v_{it1})}[\Phi(\alpha_1 y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1})], \tag{41}$$

which is consistently estimated as

$$N^{-1} \sum_{i=1}^{N} \Phi(\hat{\alpha}_{e1} y_{t2} + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{e1} + \hat{\psi}_{e1} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_{e1} + \hat{\eta}_{e1} \hat{v}_{it2}). \tag{42}$$

• These APEs, typically with further averaging out across t and perhaps over y_{t2} and $\mathbf{z}_{t1}$, can be compared directly with linear FEIV estimates.

Model:	**Linear**	**Fractional Probit**	
Estimation Method:	**Instrumental Variables**	**Pooled QMLE**	
	Coefficient	Coefficient	APE
log(*arexppp*)	.555	1.731	.583
	(.221)	(.759)	(.255)
lunch	−.062	−.298	−.100
	(.074)	(.202)	(.068)
log(*enroll*)	.046	.286	.096
	(.070)	(.209)	(.070)
$\hat{v}_2$	−.424	−1.378	—
	(.232)	(.811)	—
Scale Factor	—	.337	

```
. use meap92_01

. xtset distid year
       panel variable:  distid (strongly balanced)
        time variable:  year, 1992 to 2001
                delta:  1 unit

. des math4 avgrexp lunch enroll found

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
math4           double %9.0g                  fraction satisfactory, 4th
                                                grade math
avgrexp         float  %9.0g                  (rexppp + rexppp_1 + rexppp_2 +
                                                rexppp_3)/4
lunch           float  %9.0g                  fraction eligible for free lunch
enroll          float  %9.0g                  district enrollment
found           int    %9.0g                  foundation grant, $: 1995-2001

. sum math4 rexppp lunch

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       math4 |      5010    .6149834    .1912023       .059          1
      rexppp |      5010     6331.99    1168.198   3553.361   15191.49
       lunch |      5010    .2802852    .1571325      .0087   .9126999
```



```
. * Use foundation allowance interactions, as IVs.
. * First, linear model:

. ivreg math4 lunch alunch lenroll alenroll y96-y01 lexppp94 le94y96-le94y01
        (lavgrexp = lfound lfndy96-lfndy01), cluster(distid)

Instrumental variables (2SLS) regression              Number of obs =    3507
                                                      F( 18,   500) =  107.05
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4134
                                                      Root MSE      =  .11635

                                (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   .5545247   .2205466     2.51   0.012     .1212123     .987837
       lunch |  -.0621991   .0742948    -0.84   0.403    -.2081675    .0837693
      alunch |  -.4207815   .0758344    -5.55   0.000    -.5697749   -.2717882
     lenroll |   .0463616   .0696215     0.67   0.506    -.0904253    .1831484
    alenroll |   -.049052    .070249    -0.70   0.485    -.1870716    .0889676
         y96 |  -1.085453   .2736479    -3.97   0.000    -1.623095   -.5478119
         y97 |  -1.049922    .376541    -2.79   0.005     -1.78972   -.3101244
         y98 |  -.4548311   .4958826    -0.92   0.359    -1.429102    .5194394
         y99 |  -.4360973   .5893671    -0.74   0.460    -1.594038    .7218439
         y00 |  -.3559283   .6509999    -0.55   0.585    -1.634961     .923104
         y01 |   -.704579   .7310773    -0.96   0.336    -2.140941    .7317831
    lexppp94 |  -.4343213   .2189488    -1.98   0.048    -.8644944   -.0041482
    le94y96 |   .1253255   .0318181     3.94   0.000     .0628119    .1878392
    le94y97 |      .11487   .0425422     2.70   0.007     .0312865    .1984534
    le94y98 |   .0599439   .0554377     1.08   0.280    -.0489757    .1688636
    le94y99 |   .0557854   .0661784     0.84   0.400    -.0742367    .1858075
    le94y00 |    .048899   .0727172     0.67   0.502    -.0939699    .1917678
    le94y01 |   .0865874   .0816732     1.06   0.290    -.0738776    .2470524
```



```
       _cons |   -.334823   .2593105    -1.29   0.197    -.8442955    .1746496
------------------------------------------------------------------------------
Instrumented:  lavgrexp
Instruments:   lunch alunch lenroll alenroll y96 y97 y98 y99 y00 y01
               lexppp94 le94y96 le94y97 le94y98 le94y99 le94y00 le94y01
               lfound lfndy96 lfndy97 lfndy98 lfndy99 lfndy00 lfndy01
------------------------------------------------------------------------------
```



```
. * Get reduced form residuals for fractional probit:

. reg lavgrexp lfound lfndy96-lfndy01 lunch alunch lenroll alenroll y96-y01
      lexppp94 le94y96-le94y01, cluster(distid)

Linear regression                                     Number of obs =    3507
                                                      F( 24,   500) = 1174.57
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.9327
                                                      Root MSE      =  .03987

                                (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
    lavgrexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lfound |   .2447063   .0417034     5.87   0.000     .1627709    .3266417
     lfndy96 |   .0053951   .0254713     0.21   0.832     -.044649    .0554391
     lfndy97 |  -.0059551   .0401705    -0.15   0.882    -.0848789    .0729687
     lfndy98 |   .0045356   .0510673     0.09   0.929    -.0957972    .1048685
     lfndy99 |   .0920788   .0493854     1.86   0.063    -.0049497    .1891074
     lfndy00 |   .1364484   .0490355     2.78   0.006     .0401074    .2327894
     lfndy01 |   .2364039   .0555885     4.25   0.000      .127188    .3456198
         ...
       _cons |   .1632959   .0996687     1.64   0.102    -.0325251     .359117
------------------------------------------------------------------------------

. predict v2hat, resid
(1503 missing values generated)
```

```
. glm math4 lavgrexp v2hat lunch alunch lenroll alenroll y96-y01 lexppp94
      le94y96-le94y01, fa(bin) link(probit) cluster(distid)
note: math4 has non-integer values

Generalized linear models                         No. of obs      =       3507
Optimization     : ML                             Residual df     =       3487
                                                  Scale parameter =          1
Deviance         =  236.0659249                   (1/df) Deviance =  .0676989
Pearson          =  223.3709371                   (1/df) Pearson  =  .0640582

Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = invnorm(u)              [Probit]

                                (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   1.731039   .6541194     2.65   0.008      .4489886    3.013089
       v2hat |  -1.378126    .720843    -1.91   0.056    -2.790952    .0347007
       lunch |  -.2980214   .2125498    -1.40   0.161    -.7146114    .1185686
      alunch |  -1.114775   .2188037    -5.09   0.000    -1.543623    -.685928
     lenroll |   .2856761    .197511     1.45   0.148    -.1014383    .6727905
    alenroll |  -.2909903   .1988745    -1.46   0.143    -.6807771    .0987966
         ...
       _cons |  -2.455592   .7329693    -3.35   0.001    -3.892185   -1.018998
------------------------------------------------------------------------------
```



```
. margeff

Average partial effects after glm
      y  = Pr(math4)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   .5830163   .2203345     2.65   0.008      .1511686    1.014864
       v2hat |  -.4641533    .242971    -1.91   0.056    -.9403678    .0120611
       lunch |  -.1003741   .0716361    -1.40   0.161    -.2407782      .04003
      alunch |  -.3754579   .0734083    -5.11   0.000    -.5193355   -.2315803
     lenroll |   .0962161   .0665257     1.45   0.148    -.0341719    .2266041
    alenroll |  -.0980059   .0669786    -1.46   0.143    -.2292817    .0332698
         ...
------------------------------------------------------------------------------

. * These standard errors do not account for the first-stage estimation. Should
. * use the panel bootstrap accounting for both stages.
. * Only marginal evidence that spending is endogenous, but the negative sign
. * fits the story that districts increase spending when performance is
. * (expected to be) worse, based on unobservables (to us).
```