New Developments in Econometrics
Lecture 4: Linear Panel Data Models II

Jeff Wooldridge

Cemmap Lectures, UCL, June 2009

1. Estimating Production Functions Using Proxy Variables
2. Pseudo Panels from Pooled Cross Sections

1. Estimating Production Functions Using Proxy Variables

- Common approaches to production function estimation using firm-level panel data: fixed effects and first differencing. Typically, one assumes a Cobb-Douglas production function with additive firm heterogeneity.
- Problem: As we saw earlier, FE and FD estimators assume strict exogeneity of the inputs, conditional on firm heterogeneity. Generally rules out the possibility that inputs are chosen in response to curren or past productivity shocks, a severe restriction on firm behavior.

- Instrumental variables methods can be used to relax the strict exogeneity assumption: lagged inputs as IVs after differencing or quasi-differencing. [Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bover (1995), Blundell and Bond (2000).]
- Unfortunately, differencing removes much of the variation in the explanatory variables and can exacerbate measurement error in the inputs. Often, the instruments available after differencing often are only weakly correlated with the differenced explanatory variables. The extra moment conditions discussed earlier – Blundell and Bond, say – can help.

• Olley and Pakes (1996) (OP) suggest a different approach. Rather than allow for time-constant firm heterogeneity, OP show how investment can be used as a proxy variable for unobserved, time-varying productivity. Specifically, productivity can be expressed as an unknown function of capital and investment (when investment is strictly positive). OP present a two-step estimation method where, in the first stage, semiparametric methods are used to estimate the coefficients on the variable inputs. In a second step, the parameters on capital inputs can be identified under assumptions on the dynamics of the productivity process.

- Levinsohn and Petrin (2003) (LP) suggest using intermediate inputs to proxy for unobserved productivity (to avoid the zero investment problem). Still OP estimation.
- In implementing LP (or OP), convenient to use low-order polynomials. Petrin, Poi, and Levinsohn (2004) (PPL) suggest third-degree polynomials. Seems to give similar results to locally weighted estimation (and is now programmed in Stata).
- A unified approach that can be applied to various situations: estimate two equations simultaneously. Simplifies inference, more efficient, provides insights into identification.

- Set up as a two-equation system for panel data with the same dependent variable, but where the set of instruments differs across equation, as in Wooldridge (1996).
- Write a production function for firm i in time period t as

$$y_{it} = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + v_{it} + e_{it}, t = 1, \ldots, T, \tag{1}$$

where

y_{it} = natural logarithm of the firm's output

$\mathbf{l}_{it}$ = $1 \times J$ vector of variable inputs (labor)

$\mathbf{k}_{it}$ = $1 \times K$ vector of observed state variables (capital)

• The sequence $\{v_{it} : t = 1, \ldots, T\}$ is unobserved productivity, and $\{e_{it} : t = 1, 2, \ldots, T\}$ is a sequence of shocks.

• Key implication of the theory underlying OP and LP: for some function $g(\cdot, \cdot)$,

$$v_{it} = g(\mathbf{k}_{it}, \mathbf{m}_{it}), t = 1, \ldots, T, \tag{2}$$

where $\mathbf{m}_{it}$ is a $1 \times M$ vector of proxy variables. In OP, $\mathbf{m}_{it}$ consists of investment; in LP, $\mathbf{m}_{it}$ is intermediate inputs. In OP, representation (2) involves inverting a function relating investment to productivity and capital, but only for strictly positive investment; in LP, it is inverting a function relating intermediate inputs to productivity and capital.

• For simplicity, assume $g(\cdot,\cdot)$ is time invariant. Under the assumption

$$E(e_{it}|\mathbf{l}_{it},\mathbf{k}_{it},\mathbf{m}_{it}) = 0, t = 1,2,\ldots,T, \tag{3}$$

we have the following regression function:

$$E(y_{it}|\mathbf{l}_{it},\mathbf{k}_{it},\mathbf{m}_{it}) = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + g(\mathbf{k}_{it},\mathbf{m}_{it}). \tag{4}$$

Since $g(\cdot,\cdot)$ is allowed to be a general function – in particular, linearity in $\mathbf{k}$ is a special case, $g(\mathbf{k_{it}},\mathbf{m_{it}}) = \eta + \mathbf{k}_{it}\boldsymbol{\pi} + \mathbf{m}_{it}\boldsymbol{\psi}$ – the vector $\boldsymbol{\gamma}$ (and the intercept, α) are not identified from (4):

$$E(y_{it}|\mathbf{l}_{it},\mathbf{k}_{it},\mathbf{m}_{it}) = (\alpha+\eta) = \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}(\boldsymbol{\gamma}+\boldsymbol{\pi}) + \mathbf{m}_{it}\boldsymbol{\psi}. \tag{5}$$

- Equation (4) appears to identify $\boldsymbol{\beta}$. However, this need not be true, particularly when $\mathbf{m}_{it}$ contains intermediate inputs. As shown by Ackerberg, Caves, and Frazer (2006) (ACF), if labor inputs are chosen at the same time as intermediate inputs there is a fundamental identification problem in (4): $\mathbf{l}_{it}$ is a deterministic function of $(\mathbf{k}_{it}, \mathbf{m}_{it})$, which means $\boldsymbol{\beta}$ is nonparametrically unidentified.
- To make matters worse, ACF show that $\mathbf{l}_{it}$ actually drops out of $E(y_{it}|\mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{m}_{it})$ when the production function is Cobb-Douglas.

• Better to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ together. In $y_{it} = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + v_{it} + e_{it}$, assume

$$E(e_{it}|\mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{m}_{it}, \mathbf{l}_{i,t-1}, \mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1}, \ldots, \mathbf{l}_{i1}, \mathbf{k}_{i1}, \mathbf{m}_{i1}) = 0, t = 1, 2, \ldots, T. \quad (6)$$

Allows for serial dependence in the shocks $\{e_{it} : t = 1, 2, \ldots, T\}$: past values of e_{it} do not appear in the conditioning set.

• Next, restrict the dynamics in the productivity process:

$$\begin{aligned} E(v_{it}|\mathbf{k}_{it}, \mathbf{l}_{i,t-1}, \mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1}, \ldots) &= E(v_{it}|v_{i,t-1}) \\ &= f(v_{i,t-1}) \equiv f[g(\mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1})], \end{aligned} \quad (7)$$

where the latter equivalence holds for some $f(\cdot)$ because $v_{i,t-1} = g(\mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1})$.

- We can write

$$v_{it} = f(v_{i,t-1}) + a_{it},\ E(a_{it}|\mathbf{k}_{it}, \mathbf{l}_{i,t-1}\mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1}, \dots) = 0, \tag{8}$$

which allows the variable inputs in $\mathbf{l}_{it}$ and the intermediate inputs, $\mathbf{m}_{it}$, to be correlated with the productivity innovations, a_{it}. However, $\mathbf{k}_{it}$, past $(\mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{m}_{it})$, and functions of these are uncorrelated with a_{it}.

- Plugging (8) into the production function gives

$$y_{it} = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + f[g(\mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1})] + a_{it} + e_{it}. \tag{9}$$

• Now, we can specify the two equations that identify $(\boldsymbol{\beta}, \boldsymbol{\gamma})$:

$$y_{it} = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + g(\mathbf{k}_{it}, \mathbf{m}_{it}) + e_{it}, t = 1, \ldots, T \tag{10}$$

and

$$y_{it} = \alpha + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + f[g(\mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1})] + u_{it}, t = 2, \ldots, T, \tag{11}$$

where $u_{it} \equiv a_{it} + e_{it}$.

• When $g(\mathbf{k}_{it}, \mathbf{m}_{it}) = \eta + \mathbf{k}_{it}\boldsymbol{\pi} + \mathbf{m}_{it}\boldsymbol{\psi}$ and $v_{it} = \tau + v_{i,t-1} + a_{it}$ (random walk plus drift),

$$y_{it} = (\alpha + \eta) + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}(\boldsymbol{\gamma} + \boldsymbol{\pi}) + \mathbf{m}_{it}\boldsymbol{\psi} + e_{it} \tag{12}$$

$$y_{it} = (\alpha + \eta + \tau) + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \mathbf{k}_{i,t-1}\boldsymbol{\pi} + \mathbf{m}_{i,t-1}\boldsymbol{\psi} + u_{it} \tag{13}$$

• Importantly, the available orthogonality conditions differ across these two equations. In (10), the orthogonality conditions on the errors are

$$E(e_{it}|\mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{m}_{it}, \mathbf{l}_{i,t-1}, \mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1}, \ldots, \mathbf{l}_{i1}, \mathbf{k}_{i1}, \mathbf{m}_{i1}) \tag{14}$$

The orthogonality conditions for (11) are

$$E(u_{it}|\mathbf{k}_{it}, \mathbf{l}_{i,t-1}, \mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1}, \ldots, \mathbf{l}_{i1}, \mathbf{k}_{i1}, \mathbf{m}_{i1}) = 0, t = 2, \ldots, T. \tag{15}$$

• So, for example, in (12), $\mathbf{l}_{it}$, $\mathbf{k}_{it}$, and $\mathbf{m}_{it}$ act as their own instruments – and there are many extras obtained from lags of everything. In (13), $\mathbf{l}_{it}$ is endogenous but $\mathbf{k}_{it}$, $\mathbf{k}_{i,t-1}$, and $\mathbf{m}_{i,t-1}$ act as their own instruments. We can use $\mathbf{l}_{i,t-1}$ and other lags of everything as instruments for $\mathbf{l}_{it}$.

• When (10) does not identify $\boldsymbol{\beta}$, (11) would still generally identify $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In the special case just considered, could estimate

$$y_{it} = \eta_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \mathbf{k}_{i,t-1}\boldsymbol{\pi} + \mathbf{m}_{i,t-1}\boldsymbol{\psi} + u_{it} \tag{16}$$

by pooled IV, using instruments $(\mathbf{l}_{i,t-1}, \mathbf{k}_{it}, \mathbf{k}_{i,t-1}, \mathbf{m}_{i,t-1})$, or use pooled 2SLS with more instruments, or efficient GMM.

• Of course it is better to use (16) along with the first equation,

$$y_{it} = \alpha_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}(\boldsymbol{\gamma} + \boldsymbol{\pi}) + \mathbf{m}_{it}\boldsymbol{\psi} + e_{it}. \tag{17}$$

• To add flexibility, approximate $g(\cdot,\cdot)$ and $f(\cdot)$ functions by low-order polynomials, say, up to order three. If k_{it} and m_{it} are both scalars, $g(k,m)$ is linear in terms of the form $k^p m^q$, where p and q are nonnegative integers with $p + q \leq 3$. More generally, $g(\mathbf{k},\mathbf{m})$ contains all polynomials of order three or less. In any case, assume that we can write

$$g(\mathbf{k}_{it},\mathbf{m}_{it}) = \lambda_0 + \mathbf{c}(\mathbf{k}_{it},\mathbf{m}_{it})\boldsymbol{\lambda} \tag{18}$$

for a $1 \times Q$ vector of functions $\mathbf{c}(\mathbf{k}_{it},\mathbf{m}_{it})$. The function $\mathbf{c}(\mathbf{k}_{it},\mathbf{m}_{it})$ contains at least $\mathbf{k}_{it}$ and $\mathbf{m}_{it}$ separately, since a linear version of $g(\mathbf{k}_{it},\mathbf{m}_{it})$ should always be an allowed special case.

- Assume that $f(\cdot)$ can be approximated by a polynomial in v:

$$f(v) = \rho_0 + \rho_1 v + \ldots + \rho_G v^G. \tag{19}$$

- Now we can plug into the original production function to get

$$y_{it} = \alpha_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \mathbf{c}_{it}\boldsymbol{\lambda} + e_{it}, t = 1, \ldots, T \tag{20}$$

and

$$y_{it} = \eta_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \rho_1(\mathbf{c}_{i,t-1}\boldsymbol{\lambda}) + \ldots + \rho_G(\mathbf{c}_{i,t-1}\boldsymbol{\lambda})^G + u_{it}, t = 2, \ldots, T, \tag{21}$$

where α_0 and η_0 are new intercepts and $\mathbf{c}_{it} \equiv \mathbf{c}(\mathbf{k}_{it}, \mathbf{m}_{it})$.

• Can specify instrumental variables (IVs) for each of these two equations. The most straightforward choice of IVs for (20) is simply

$$\mathbf{z}_{it1} \equiv (1, \mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{c}^{o}_{it}), \tag{22}$$

where $\mathbf{c}^{o}_{it}$ is $\mathbf{c}_{it}$ but without $\mathbf{k}_{it}$. The choice in (22) corresponds to the regression analysis in OP and LP for estimating $\boldsymbol{\beta}$ in a first stage.

• Any nonlinear function of $(\mathbf{l}_{it}, \mathbf{k}_{it}, \mathbf{c}^{o}_{it})$ is also a valid IV, as are all lags and all functions of these lags. Adding a lag could be useful for generating overidentifying restrictions to test the model assumptions.

• Instruments for (21) would include $(\mathbf{k}_{it}, \mathbf{l}_{i,t-1}, \mathbf{c}_{i,t-1})$ and, especially if $G > 1$, nonlinear functions of $\mathbf{c}_{i,t-1}$ (probably low-order polynomials):

$$\mathbf{z}_{it2} = (1, \mathbf{k}_{it}, \mathbf{l}_{i,t-1}, \mathbf{c}_{i,t-1}, \mathbf{q}_{i,t-1}), \tag{23}$$

where $\mathbf{q}_{i,t-1}$ is a set of nonlinear functions of $\mathbf{c}_{i,t-1}$, probably consisting of low-order polynomials.

• Total of $2 + J + K + Q + G$ parameters in (21). $(\mathbf{k}_{it}, \mathbf{l}_{i,t-1}, \mathbf{c}_{i,t-1})$ act as their own instruments, and then we would include enough nonlinear functions in $\mathbf{q}_{i,t-1}$ to identify $\rho_1, \ldots, \rho_G$.

• A sensible choice for the instrument matrix for the two equations: for each (i,t),

$$\mathbf{Z}_{it} \equiv \begin{pmatrix} (\mathbf{l}_{it}, \mathbf{c}_{it}, \mathbf{z}_{it2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{it2} \end{pmatrix}, t = 2, \ldots, T. \tag{24}$$

This choice makes it clear that all instruments available for (21) are also valid for (20), and we have some additional moment restrictions in (20). Recall $\mathbf{c}_{it}$ is a function of $(\mathbf{k}_{it}, \mathbf{m}_{it})$.

• GMM estimation of all parameters at the same time is relatively straightforward. For each $t > 1$, define a 2×1 residual function as

$$\mathbf{r}_{it}(\boldsymbol{\theta}) = \begin{pmatrix} y_{it} - \alpha_0 - \mathbf{l}_{it}\boldsymbol{\beta} - \mathbf{k}_{it}\boldsymbol{\gamma} - \mathbf{c}_{it}\boldsymbol{\lambda} \\ y_{it} - \eta_0 - \mathbf{l}_{it}\boldsymbol{\beta} - \mathbf{k}_{it}\boldsymbol{\gamma} - \rho_1(\mathbf{c}_{i,t-1}\boldsymbol{\lambda}) - \ldots - \rho_G(\mathbf{c}_{i,t-1}\boldsymbol{\lambda})^G \end{pmatrix}, \quad (25)$$

so that

$$E[\mathbf{Z}_{it}'\mathbf{r}_{it}(\boldsymbol{\theta})] = \mathbf{0}, t = 2, \ldots, T. \quad (26)$$

• Wooldridge (2009, Economics Letters) contains more details.

• In practice, probably include a full set of time period dummies with coefficients unrestricted across equations to allow heterogeneity over time in the production function and productivity.

• In the case of random walk for productivity, even with polynomials for $g(\mathbf{k}_{it}, \mathbf{m}_{it})$, the system is linear in its parameters:

$$y_{it} = \alpha_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \mathbf{c}_{it}\boldsymbol{\lambda} + e_{it} \tag{27}$$

$$y_{it} = \eta_0 + \mathbf{l}_{it}\boldsymbol{\beta} + \mathbf{k}_{it}\boldsymbol{\gamma} + \mathbf{c}_{i,t-1}\boldsymbol{\lambda} + u_{it} \tag{28}$$

where $\mathbf{c}_{it}$ contains, at a minimum, linear functions of $(\mathbf{k}_{it}, \mathbf{m}_{it})$.

• Can write

$$\mathbf{y}_{it} = \mathbf{X}_{it}\boldsymbol{\theta} + \mathbf{r}_{it} \tag{29}$$

where $\mathbf{y}_{it} = (y_{it}, y_{it})'$ is 2×1,

$$\mathbf{X}_{it} = \begin{pmatrix} 1 & 0 & \mathbf{l}_{it} & \mathbf{k}_{it} & \mathbf{c}_{it} \\ 0 & 1 & \mathbf{l}_{it} & \mathbf{k}_{it} & \mathbf{c}_{i,t-1} \end{pmatrix}, \tag{30}$$

and $\boldsymbol{\theta} = (\alpha_0, \eta_0, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\lambda}')'$. $\mathbf{Z}_{it}$ as in (24).

- Acharya and Keller (2006) apply the two-step OP method and the joint GMM (OP/W) approach, and get more stable estimates of capital elasticity using the latter. The labor elasticity for OP is .548 (.054) and .557 (.018) for OP/W. For capital, the OP elasticity is .234 (.164) for OP and the OP/W elasticity is .513 (.105), with the OP estimate changing much more across specification.

2. Pseudo Panels from Pooled Cross Sections

• It is important to distinguish between the population model and the sampling scheme. We are interested in estimating the parameters of

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f + u_t,\ t = 1, \ldots, T, \tag{31}$$

which represents a population defined over T time periods.

• Normalize $E(f) = 0$. Assume all elements of $\mathbf{x}_t$ have some time variation. To interpret $\boldsymbol{\beta}$, contemporaneous exogeneity conditional on f:

$$E(u_t|\mathbf{x}_t, f) = 0, t = 1, \ldots, T. \tag{32}$$

• The current literature does not even use $E(u_t|\mathbf{x}_t, f) = 0$. We will use an implication of (32):

$$E(u_t|f) = 0, t = 1, \ldots, T. \tag{33}$$

Because f aggregates all time-constant unobservables, we should think of (32) as implying that $E(u_t|g) = 0$ for any time-constant variable g, whether unobserved or observed.

• Deaton (1985) considered the case of independently sampled cross sections. Assume that the population for which (31) holds is divided into G groups (or cohorts). Common is birth year. For a random draw i at time t, let g_i be the group indicator, taking on a value in $\{1, 2, \ldots, G\}$.

• By the previous discussion, for a random draw i,

$$E(u_{it}|g_i) = 0. \tag{34}$$

• Taking the expected value of (31) conditional on group membership and using only (34), we have

$$E(y_t|g) = \eta_t + E(\mathbf{x}_t|g)\boldsymbol{\beta} + E(f|g), t = 1, \ldots, T. \tag{35}$$

This is Deaton's starting point, and Moffitt's (1993). If we start with (31) under (33), there is no "randomness" in (35); it is a population moments equation.

• More recent authors have left $v_{gt} \equiv E(u_t|g)$ in (35) and then treat it as random, even though it is a population mean (which is zero for all g and t).

• Define the population means

$$\alpha_g = E(f|g),\ \mu_{gt}^y = E(y_t|g),\ \boldsymbol{\mu}_{gt}^{\mathbf{x}} = E(\mathbf{x}_t|g) \tag{36}$$

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

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g. \tag{37}$$

- Equation (37) holds without any assumptions restricting the dependence between $\mathbf{x}_t$ and u_r across t and r. In fact, $\mathbf{x}_t$ can contain lagged dependent variables or contemporaneously endogenous variables (omitted variables, simultaneity, measurement error). Should we be suspicious? Probably, but the source of identification is different from exogeneity assumptions.
- Equation (37) looks like a linear regression model in the population means, μ_{gt}^{y} and $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. One can use a "fixed effects" regression to "estimate" η_t, α_g, and $\boldsymbol{\beta}$.

• In the Deaton setup, with reasonable cell sizes, N_{gt}, treat as a minimum distance (MD) problem. One commonly used (inefficient) estimator is fixed effects applied to the sample means, based on the same relationship in the population:

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \mu_{gt}^{y} \right) \tag{38}$$

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

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} \text{ on } 1, d2, \ldots, dT, c2, \ ..., cG, \tag{39}$$

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

- From (38), clear that underlying population model cannot contain a full set of group/time interactions because then the $\ddot{\mu}_{gt}^{x}$ would all be identically zero. We *could* allow this feature with individual-level data.
- The absense of a full cohort/time effects in the population model is a key identifying restriction.

• $\boldsymbol{\beta}$ is not identified if we can write $\boldsymbol{\mu}^{\mathbf{x}}_{gt} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g$ for vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_g$, $t = 1, \ldots, T$, $g = 1, \ldots, G$. In other words, we must exclude a full set of group/time effects in the structural model but we need some interaction between them in the covariate means. Identification still might be weak if variation in $\{\ddot{\boldsymbol{\mu}}^{\mathbf{x}}_{gt} : t = 1, .., T, \ g = 1, \ldots, G\}$ is small: a small change in estimates of $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ can lead to large changes in $\hat{\boldsymbol{\beta}}$.

• Estimation by nonseparable MD because $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ are the restrictions on the structural parameters $\boldsymbol{\theta}$ given cell means $\boldsymbol{\pi}$ (Chamberlain, lecture notes). But given $\boldsymbol{\pi}$, restrictions are linear in $\boldsymbol{\theta}$.

• The optimal MD estimator is intuitive and easy to obtain. After "FE" estimation on the sample means (pooled across g and t), obtain the residual variances within each cell, $\hat{\tau}_{gt}^2$, based on $y_{igt} - \mathbf{x}_{igt}\check{\boldsymbol{\beta}} - \check{\alpha}_g - \check{\eta}_t$, where $(\check{\boldsymbol{\beta}}, \check{\alpha}_g, \check{\eta}_t)$ are the "FE" estimates.

• Define "regressors" $\hat{\boldsymbol{\omega}}_{gt} = (\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}, \mathbf{d}_t, \mathbf{c}_g)$, and let $\hat{\mathbf{W}}$ be the $GT \times (K + T + G - 1)$ stacked matrix (where we drop, say, the time dummy for the first period.). Let $\hat{\mathbf{C}}$ be the $GT \times GT$ diagonal matrix with $\hat{\tau}_{gt}^2/(N_{gt}/N)$ down the diagonal.

- The optimal MD estimator (minimum chi-square estimator), which is $\sqrt{N}$-asymptotically normal, is

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}^{y}. \tag{40}$$

As in separable cases, the efficient MD estimator looks like a "weighted least squares" estimator, where cell (g,t) is weighted by $(N_{gt}/N)/\hat{\tau}_{gt}^{2}$. The asymptotic variance is estimated as $(\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}/N$.

- Bootstrapping to account for "weak" identification might make sense.
- Inoue (2008) obtains a different limiting distribution, which is stochastic, because he treats estimation of $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ and μ_{gt}^{y} asymmetrically.

• There are $GT(K+1) - (K+T+G-1)$ overidentification conditions. Conveniently, the sum of squared residuals from the weighted least squares regression is distributed as $\chi^2_{GT(K+1)-(K+T+G-1)}$ if the restrictions are true.

• Deaton (1985), VN (1993), and Collado (1998), use a different asymptotic analysis: $GT \to \infty$ (Deaton) or $G \to \infty$, with fixed cell sizes. Some authors try to have it both ways (first ignoring the estimation error in the cell means, so N_{gt} is large, and then assuming a large number of groups and/or time periods).

• Allows for models with lagged dependent variables, but now the vectors of means contain redundancies. If

$$y_t = \eta_t + \rho y_{t-1} + \mathbf{z}_t\boldsymbol{\gamma} + f + u_t,\ E(u_t|g) = 0, \tag{41}$$

then the same moments are valid. But, now we would define the vector of means as $(\mu^y_{gt}, \boldsymbol{\mu}^{\mathbf{z}}_{gt})$, and appropriately pick off μ^y_{gt} in defining the moment conditions. We now have fewer moment conditions to estimate the parameters.

• The MD approach applies to extensions of the basic model. Random trend model (Heckman and Hotz, 1989):

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f_1 + f_2 t + u_t. \tag{42}$$

$$\mu^y_{gt} = \eta_t + \boldsymbol{\mu}^{\mathbf{x}}_{gt}\boldsymbol{\beta} + \alpha_g + \varphi_g t, \tag{43}$$

• We can even estimate models with time-varying factor loads on the heterogeneity:

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + \lambda_t f + u_t, \tag{44}$$

$$\mu^y_{gt} = \eta_t + \boldsymbol{\mu}^{\mathbf{x}}_{gt}\boldsymbol{\beta} + \lambda_t\alpha_g. \tag{45}$$

• How can we use a stronger assumption, such as $E(u_t|\mathbf{z}_t, f) = \mathbf{0}$, $t = 1, \ldots, T$, for instruments $\mathbf{z}_t$, to more precisely estimate $\boldsymbol{\beta}$? Gives lots of potentially useful moment conditions:

$$E(\mathbf{z}_t' y_t|g) = \eta_t E(\mathbf{z}_t'|g) + E(\mathbf{z}_t' \mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t' f|g), \tag{46}$$

using $E(\mathbf{z}_t' u_t|g) = \mathbf{0}$.