CLASSICAL MINIMUM DISTANCE ESTIMATION

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Motivation
3. General Setup
4. Applications

1. INTRODUCTION

- Classical minimum distance (CMD) estimation is useful for combining different estimators of the same parameter. More generally, for obtaining “structural” estimates from “reduced form” estimates when a known relationship exists between the “structural” and “reduced form” parameters.
- The relationship is sometimes linear, but need not be. Often the restrictions are “separable” in a sense to be made precise, but they need not be.

- CMD is often an alternative to generalized method of moments estimation.
- Often it is easy to estimate some “reduced form” parameters, which we denote $\boldsymbol{\pi}$, and then use CMD to recover estimates of $\boldsymbol{\theta}$, the parameters of interest (“structural” parameters). In interesting cases, the dimension of $\boldsymbol{\pi}$ is strictly larger than that of $\boldsymbol{\theta}$.

2. MOTIVATION

- Consider a simple example. We have two populations, with distributions $y_1 \sim (\mu, \sigma_1^2)$ and $y_2 \sim (\mu, \sigma_2^2)$, that is, $E(y_1) = E(y_2) = \mu$ but the populations may have different variances. Suppose we have random samples of size N_1 and N_2 from each population, and let $\hat{\pi}_1 = \bar{y}_1$ and $\hat{\pi}_2 = \bar{y}_2$ be the sample averages.
- Question: How should we combine $\hat{\pi}_1$ and $\hat{\pi}_2$ in the most efficient way to estimate $\theta = \mu$?

- We can think of the class of linear combinations,

$$a_1\bar{y}_1 + a_2\bar{y}_2 \tag{1}$$

where $a_1 + a_2 = 1$ is needed for unbiasedness. So, write

$$\hat{\theta}_a = a\bar{y}_1 + (1-a)\bar{y}_2 \tag{2}$$

and then

$$\begin{aligned} Var(\hat{\theta}_a) &= a^2 Var(\bar{y}_1) + (1-a)^2 Var(\bar{y}_2) \\ &= a^2(\sigma_1^2/N_1) + (1-a)^2(\sigma_2^2/N_2) \equiv a^2 \upsilon_1 + (1-a)^2 \upsilon_2. \end{aligned} \tag{3}$$

- Can use calculus to minimize the variance as a function of a:

$$a^* = \frac{v_2}{v_1 + v_2} = \frac{(\sigma_2^2/N_2)}{(\sigma_1^2/N_1) + (\sigma_2^2/N_2)}, \tag{4}$$

which simply says that the weight for each estimator depends on its variance relative to the other estimator. Naturally, we should give more weight to the estimator with the smallest variance (which, in turn, depends on the population variance and the sample size).

- This simple example suggests that a general framework, which allows nonlinear restrictions and correlation across estimators, is useful.

3. GENERAL SETUP

- Let $\boldsymbol{\theta}$ denote a $P \times 1$ vector of parameters that we are ultimately interested in estimating, and let $\boldsymbol{\pi}$ be $S \times 1$ with $S > P$. (If $S = P$ the problem is just one of solving P equations – possibly nonlinear – in P unknowns.)

Separable Case

- For a function $\mathbf{h} : \mathbb{R}^P \to \mathbb{R}^S$, assume the population values satisfy

$$\boldsymbol{\pi}_o = \mathbf{h}(\boldsymbol{\theta}_o). \tag{5}$$

- Assume $\mathbf{h}$ is continuously differentiable on the interior of $\boldsymbol{\Theta}$, and that $\boldsymbol{\theta}_o \in int(\boldsymbol{\Theta})$.

- This setup is actually too restrictive in some cases because it assumes the restrictions are separable in the two sets of parameters.
- Let $\hat{\boldsymbol{\pi}}$ be a $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\pi}_o$ (which is often easy to obtain):

$$\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o) \overset{a}{\sim} Normal(0, \boldsymbol{\Xi}_o) \tag{6}$$

and assume that $\boldsymbol{\Xi}_o$ $(S \times S)$ is positive definite.

- When $S > P$, there is generally no solution to the S equations

$$\hat{\boldsymbol{\pi}} = \mathbf{h}(\boldsymbol{\theta}). \tag{7}$$

- The general idea is to estimate $\boldsymbol{\theta}_o$ by minimizing the "distance" between $\hat{\boldsymbol{\pi}}$ and $\mathbf{h}(\boldsymbol{\theta})$. Essentially, we choose $\boldsymbol{\theta}$ to minimize the "length" of the vector

$$\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta}). \tag{8}$$

- In the simple example above, $\theta = \mu$, $\pi_1 = \mu_1$, $\pi_2 = \mu_2$. That is, the elements of $\boldsymbol{\pi}$ are estimated ignoring the restriction that they are the same.

$$\mathbf{h}(\theta) = \begin{pmatrix} \theta \\ \theta \end{pmatrix}, \boldsymbol{\pi} - \mathbf{h}(\boldsymbol{\theta}) = \begin{pmatrix} \pi_1 - \theta \\ \pi_2 - \theta \end{pmatrix} \tag{9}$$

and

$$\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta}) = \begin{pmatrix} \hat{\pi}_1 - \theta \\ \hat{\pi}_2 - \theta \end{pmatrix}. \tag{10}$$

- In the general case, we can define a so-called **classical minimum distance** (**CMD**) estimator for a wide class of weighting matrices, but one almost always (eventually) uses the asymptotically efficient version.
- Let $\hat{\mathbf{\Xi}}$ be a consistent estimator of $\mathbf{\Xi}_o$, that is, $plim_{N\to\infty}(\hat{\mathbf{\Xi}}) = \mathbf{\Xi}_o$.
- Then $\hat{\boldsymbol{\theta}}$ solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} [\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta})]'\hat{\mathbf{\Xi}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta})]. \tag{11}$$

- The objective function is effectively a weighted Euclidean distance, with the weighting matrix the inverse of the estimated $Avar[\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o)]$.
- In the previous example,

$$\boldsymbol{\Xi}_o \equiv Avar[\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o)] \; = \begin{pmatrix} \sigma_{o1}^2/\rho_{o1} & 0 \\ 0 & \sigma_{o2}^2/\rho_{o2} \end{pmatrix} \tag{12}$$

where we define $\rho_{o1} = \lim_{N\to\infty} N_1/N$ and $\rho_{o2} = \lim_{N\to\infty} N_2/N$ and $N = N_1 + N_2$ (so $\rho_{o1} + \rho_{o2} = 1$).

- The minimum distance estimator in this case minimizes

$$[\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta})]'\hat{\boldsymbol{\Xi}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\theta})] \;= N\{(\hat{\pi}_1 - \theta)^2(N_1/\hat{\sigma}_1^2) + (\hat{\pi}_2 - \theta)^2(N_2/\hat{\sigma}_2^2)\}, \quad (13)$$

with solution

$$\hat{\theta} = \frac{\hat{\omega}_1}{(\hat{\omega}_1 + \hat{\omega}_2)}\hat{\pi}_1 + \frac{\hat{\omega}_2}{(\hat{\omega}_1 + \hat{\omega}_2)}\hat{\pi}_2 \quad (14)$$

where $\hat{\omega}_1 = N_1/\hat{\sigma}_1^2$ and $\hat{\omega}_2 = N_2/\hat{\sigma}_2^2$.

- Simple algebra shows that that this is the same estimator we derived earlier (where we replace σ_g^2 with $\hat{\sigma}_g^2$, $g = 1,2$).

- Back to the general case. Let $\mathbf{H}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{h}(\boldsymbol{\theta})$ be the $S \times P$ Jacobian of $\boldsymbol{\theta}$. Then the first order condition for $\hat{\boldsymbol{\theta}}$ is

$$\mathbf{H}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Xi}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})] = \mathbf{0}, \tag{15}$$

which is P equations in the P unknowns, $\hat{\boldsymbol{\theta}}$.

- By a standard mean value expansion,

$$\sqrt{N}[\mathbf{h}(\hat{\boldsymbol{\theta}}) - \mathbf{h}(\boldsymbol{\theta}_o)] = \mathbf{H}(\boldsymbol{\theta}_o)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1). \tag{16}$$

- Also, remember $\boldsymbol{\pi}_o = \mathbf{h}(\boldsymbol{\theta}_o)$.

- Therefore,

$$\begin{aligned} \mathbf{0} &= \mathbf{H}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Xi}}^{-1}\{\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o) - \sqrt{N}\,[\mathbf{h}(\hat{\boldsymbol{\theta}}) - \mathbf{h}(\boldsymbol{\theta}_o)]\} && (17) \\ &= \mathbf{H}(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Xi}}^{-1}\{\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o) - \mathbf{H}(\boldsymbol{\theta}_o)\,\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1)\}. && (18) \end{aligned}$$

- Now, $\mathbf{H}(\hat{\boldsymbol{\theta}}) = \mathbf{H}(\boldsymbol{\theta}_o) + o_p(1)$ (because $\mathbf{H}(\cdot)$ is continuous) and we assume $\hat{\boldsymbol{\Xi}} = \boldsymbol{\Xi}_o + o_p(1)$.
- Therefore,

$$\mathbf{H}(\boldsymbol{\theta}_o)'\boldsymbol{\Xi}_o^{-1}\mathbf{H}(\boldsymbol{\theta}_o)\,\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{H}(\boldsymbol{\theta}_o)'\boldsymbol{\Xi}_o^{-1}\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o) + o_p(1). \qquad (19)$$

- Because $Avar[\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o)] = \boldsymbol{\Xi}_o$,

$$\mathbf{H}(\boldsymbol{\theta}_o)'\boldsymbol{\Xi}_o^{-1}\mathbf{H}(\boldsymbol{\theta}_o)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \overset{d}{\to} Normal(\mathbf{0}, \mathbf{H}(\boldsymbol{\theta}_o)'\boldsymbol{\Xi}_o^{-1}\mathbf{H}(\boldsymbol{\theta}_o)) \tag{20}$$

and so

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \overset{d}{\to} Normal(\mathbf{0}, (\mathbf{H}_o'\boldsymbol{\Xi}_o^{-1}\mathbf{H}_o)^{-1}) \tag{21}$$

where $\mathbf{H}_o \equiv \mathbf{H}(\boldsymbol{\theta}_o)$.

- To get $Avar(\hat{\boldsymbol{\theta}})$ we divide $Avar[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$ by N, as usual:

$$Avar(\hat{\boldsymbol{\theta}}) = (\mathbf{H}_o'\boldsymbol{\Xi}_o^{-1}\mathbf{H}_o)^{-1}/N. \tag{22}$$

- We get same asymptotic variance whether we use $\boldsymbol{\Xi}_o$ or estimate it consistently. In other words, we can derive the limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ acting as if we know $\boldsymbol{\Xi}_o$, and then estimate the asymptotic variance by plugging in $\hat{\boldsymbol{\Xi}}$ for $\boldsymbol{\Xi}_o$.
- We estimate the asymptotic variance as

$$Avar(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{H}}'(\hat{\boldsymbol{\Xi}}/N)^{-1}\hat{\mathbf{H}}]^{-1} = \left\{\hat{\mathbf{H}}'\left[\widehat{Avar}(\hat{\boldsymbol{\pi}})\right]^{-1}\hat{\mathbf{H}}\right\}^{-1} \qquad (23)$$

and use this to construct standard errors, confidence intervals, and Wald tests of multiple hypotheses.

- Can also use a test based on the difference in criterion functions.

- Can show that using a consistent estimator of $\boldsymbol{\Xi}_o^{-1}$ produces the estimator with the smallest asymptotic variance in the class of all CMD estimators.
- The efficient CMD estimator is often called the **minimum chi-square estimator**. This name comes from the fact that the objective function, properly scaled, has an asymptotic chi-square distribution if $\boldsymbol{\pi}_o = \mathbf{h}(\boldsymbol{\theta}_o)$.

- Precisely,

$$
\begin{aligned}
& N[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})]'\hat{\boldsymbol{\Xi}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})] \\
= \; & \{\sqrt{N}\,[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})]\}'\hat{\boldsymbol{\Xi}}^{-1}\{\sqrt{N}\,[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})]\} \qquad (24) \\
= \; & [\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})]'\left\{\widehat{Avar}(\hat{\boldsymbol{\pi}})\right\}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}})] \;\overset{d}{\rightarrow}\; \chi^2_{S-P}. \qquad (25)
\end{aligned}
$$

- (If do not multiply by N, the objective function converves in probability to zero.)
- The statistic is conveniently used to test the $S - P$ overidentifying restrictions.
- If we have different choices for $\hat{\boldsymbol{\pi}}$, it is better to use the most efficient estimator.

Linear Case

- If the restrictions can be expressed as

$$\boldsymbol{\pi}_o = \mathbf{H}\boldsymbol{\theta}_o \tag{26}$$

for a known $S \times P$ known, nonrandom matrix $\mathbf{H}$ with rank P, then the minimum chi-square estimator is obtained in closed form:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}'\hat{\boldsymbol{\Xi}}^{-1}\mathbf{H})^{-1}\mathbf{H}'\hat{\boldsymbol{\Xi}}^{-1}\hat{\boldsymbol{\pi}} \tag{27}$$

- This has the form of a generalized least squares estimator of $\hat{\boldsymbol{\pi}}$ on $\mathbf{H}$ using estimated variance matrix $\hat{\boldsymbol{\Xi}}$.

- Nothing changes of we replace $\hat{\mathbf{\Xi}}$ with $\hat{\mathbf{\Xi}}/N$, so it looks like GLS using $\widehat{Avar}(\hat{\boldsymbol{\pi}})$ as the variance matrix.
- Sometimes, $\hat{\mathbf{\Xi}}$ is diagonal, and then the MCS estimator looks like a weighted least squares estimator.
- Running GLS of $\hat{\boldsymbol{\pi}}$ on $\mathbf{H}$ using variance matrix $\hat{\mathbf{\Xi}}/N$ also gives the correct estimate of $Avar(\hat{\boldsymbol{\theta}})$.
- Viewing MCS as GLS or WLS is fine for computation, but it is misleading for statistical inference. The number of rows in $\mathbf{H}$, S, is fixed. It is not growing with N. The MCS estimator inherits its asymptotic distribution from that of $\hat{\boldsymbol{\pi}}$.

- Already saw one linear example. As another simple linear example, suppose we have a single population described by $y \geq 0$ where, for some $\theta_o > 0$,

$$E(y) = \theta_o \tag{28}$$

$$Var(y) = \theta_o \tag{29}$$

$$\pi_{o1} = E(y), \hat{\pi}_1 = \bar{y} \tag{30}$$

$$\pi_{o2} = Var(y), \hat{\pi}_2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{y})^2. \tag{31}$$

- Now, the 2 × 2 asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o)$ is diagonal only in the special case $E[(y-\theta_o)^3] = 0$ (symmetric distribution). For the Poisson distribution, where the mean and variance are the same, the central third moment is not zero.
- Generally, can show

$$Avar[\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_o)] = \begin{pmatrix} \sigma_o^2 & \eta_o \\ \eta_o & \kappa_o \end{pmatrix} \tag{32}$$

where $\sigma_o^2 = Var(y)$, $\eta_o = E[(y-\theta_o)]^3$, and $\kappa_o = E[(y-\theta_o)^4] - \sigma_o^4$. Of course, these are all easily estimable using the sample counterparts.

- The resulting minimum chi-square estimator is not necessarily more efficient than just the sample average. (For example, in the Poisson case the MLE is the sample average.)

Nonseparable Case

- Sometimes the relationship between $\boldsymbol{\pi}_o$ and $\boldsymbol{\theta}_o$ cannot be written in separable form. Instead, suppose Q restrictions can be written as

$$\mathbf{g}(\boldsymbol{\pi}_o, \boldsymbol{\theta}_o) = \mathbf{0}. \tag{33}$$

- Chamberlain (Harvard lecture notes) has shown that the optimal weighting matrix in this case is (a consistent estimator of)

$$[\nabla_{\boldsymbol{\pi}}\mathbf{g}(\boldsymbol{\pi}_o, \boldsymbol{\theta}_o)\Xi_o\nabla_{\boldsymbol{\pi}}\mathbf{g}(\boldsymbol{\pi}_o, \boldsymbol{\theta}_o)']^{-1}, \tag{34}$$

where $\nabla_{\boldsymbol{\pi}}\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is the $Q \times S$ Jacobian of $\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$. (In the separable case, $\nabla_{\boldsymbol{\pi}}\mathbf{g}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is $\mathbf{I}_S$.)

- So, we need to obtain a prelimary estimator of $\boldsymbol{\theta}_o$, say $\breve{\boldsymbol{\theta}}$. Likely this is obtained using $\hat{\boldsymbol{\pi}}$ with the $Q \times Q$ identity matrix as the weighting matrix. Then the MCS estimator solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{g}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})' [\nabla_{\boldsymbol{\pi}} \mathbf{g}(\hat{\boldsymbol{\pi}}, \breve{\boldsymbol{\theta}}) \hat{\boldsymbol{\Xi}} \nabla_{\boldsymbol{\pi}} \mathbf{g}(\hat{\boldsymbol{\pi}}, \breve{\boldsymbol{\theta}})']^{-1} \mathbf{g}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}). \tag{35}$$

- Useful for estimation with pseudo panel data, where independent cross sections are turned into pseudo panels by grouping units (say, individuals by birth year).

4. APPLICATIONS

4.1. Chamberlain's Approach to Unobserved Effects Models

- Consider the usual unobserved effects panel data model under strict exogeneity:

$$y_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \tag{36}$$

$$E(u_{it}|\mathbf{x}_i, c_i) = 0, t = 1, \ldots, T \tag{37}$$

where no restrictions are put on $D(c_i|\mathbf{x}_i)$. We know fixed effects, first differencing, or FGLS versions of these are consistent for $\boldsymbol{\beta}$, assuming time-varying $\{\mathbf{x}_{it}\}$.

- Explicitly including different intercepts makes the setup more realistic, but, of course, does not change the main point.
- Drop “*o*” subscript on parameters for simplicity.
- FE, FD efficient under different assumptions. Both are inefficient under general serial correlation patterns in $\{u_{it} : t = 1, \ldots, T\}$.

- FEGLS and FDGLS are equally efficient under the system homoskedasticity requirement

$$Var(\mathbf{u}_i|\mathbf{x}_i, c_i) = Var(\mathbf{u}_i) = \mathbf{\Omega}. \tag{38}$$

- But FEGLS and FDGLS are not generally efficient if system homoskedasticity does not hold. Rather than model, say, $Var(\ddot{\mathbf{u}}_i|\ddot{\mathbf{X}}_i)$, can do better than FE or FD by using minimum distance.

- Use the linear projection of c_i on $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$:

$$L(c_i|1, \mathbf{x}_i) = \psi + \mathbf{x}_{i1}\boldsymbol{\lambda}_1 + \mathbf{x}_{i2}\boldsymbol{\lambda}_2 + \ldots + \mathbf{x}_{iT}\boldsymbol{\lambda}_T \tag{39}$$

or

$$c_i = \psi + \mathbf{x}_{i1}\boldsymbol{\lambda}_1 + \mathbf{x}_{i2}\boldsymbol{\lambda}_2 + \ldots + \mathbf{x}_{iT}\boldsymbol{\lambda}_T + a_i \tag{40}$$

$$E(a_i) = 0,\ E(\mathbf{x}_i' a_i) = \mathbf{0}. \tag{41}$$

- Absorb ψ into the α_t:

$$\begin{aligned} y_{it} &= \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_{i1}\boldsymbol{\lambda}_1 + \mathbf{x}_{i2}\boldsymbol{\lambda}_2 + \ldots + \mathbf{x}_{iT}\boldsymbol{\lambda}_T + a_i + u_{it} && (42) \\ &\equiv \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_i\boldsymbol{\lambda} + v_{it} && (43) \\ E(v_{it}) &= 0,\ E(\mathbf{x}_i' v_{it}) = \mathbf{0}. && (44) \end{aligned}$$

- We could estimate all parameters by pooled OLS. Turns out, as in the Mundlak approach, this delivers the FE estimate for $\boldsymbol{\beta}$. Same for random effects.

- Instead, use minimum distance. Write the equation as

$$y_{it} = \alpha_t + \mathbf{x}_{i1}\boldsymbol{\lambda}_1 + \mathbf{x}_{i2}\boldsymbol{\lambda}_2 + ..+ \mathbf{x}_{it}(\boldsymbol{\beta} + \boldsymbol{\lambda}_t) + \ldots + \mathbf{x}_{iT}\boldsymbol{\lambda}_T + v_{it} \tag{45}$$

$$\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_T, \boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_T', \boldsymbol{\beta}')' \tag{46}$$

for $T + K + TK$ parameters.

- Write an unrestricted system as

$$y_{it} = \pi_{t0} + \mathbf{x}_i\boldsymbol{\pi}_t + v_{it},\ t = 1, \ldots, T \tag{47}$$

- $\boldsymbol{\pi}$ has dimension $T + T^2K$.

- The restrictions are linear. When $T = 2$, the restrictions can be written as

$$\pi_{10} = \alpha_1,\ \boldsymbol{\pi}_{11} = \boldsymbol{\beta} + \boldsymbol{\lambda}_1, \boldsymbol{\pi}_{12} = \boldsymbol{\lambda}_2 \tag{48}$$

$$\pi_{20} = \alpha_2,\ \boldsymbol{\pi}_{21} = \boldsymbol{\lambda}_1, \boldsymbol{\pi}_{22} = \boldsymbol{\beta} + \boldsymbol{\lambda}_2 \tag{49}$$

or

$$\begin{pmatrix} \pi_{10} \\ \boldsymbol{\pi}_{11} \\ \boldsymbol{\pi}_{12} \\ \pi_{20} \\ \boldsymbol{\pi}_{21} \\ \boldsymbol{\pi}_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{I}_K \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \\ 0 & 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{I}_K \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \boldsymbol{\beta} \end{pmatrix} \tag{50}$$

- So the matrix **H** is $(4K + 2) \times (3K + 2)$ in this case. (There are K overidentifying restrictions, where K is the dimension of $\mathbf{x}_{it}$.)

- Generally, estimate $(\pi_{t0}, \boldsymbol{\pi}_t)$ by OLS of y_{it} on $1, \mathbf{x}_i$, $i = 1, \ldots, N$, separately for each t. A GLS approach does not help with efficiency because the set of regressors is the same in each time period.
- Need to estimate the variance-covariance matrix of the entire vector, $\hat{\boldsymbol{\pi}}$, using a fully robust variance matrix (that allows heteroskedasticity and serial correlation in $\{v_{it} : t = 1, \ldots, T\}$):

$$\left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \mathbf{X}_i\right) \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \tag{51}$$

where $\mathbf{X}_i = \mathbf{I}_T \otimes (1, \mathbf{x}_i)$ and $\hat{\mathbf{v}}_i$ is the $T \times 1$ vector of OLS residuals for each i.

4.2. Models with Time-Varying Factor Loads

- Now consider the model

$$y_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i + u_{it} \tag{52}$$

$$E(u_{it}|\mathbf{x}_i, c_i) = 0 \tag{53}$$

- Normalize $\eta_1 = 1$.
- Use Mundlak:

$$E(c_i|\mathbf{x}_i) = E(c_i|\bar{\mathbf{x}}_i) = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} \tag{54}$$

(or could use linear projection).

- Now plug in and absorb $\eta_t\psi$ into time intercepts:

$$E(y_{it}|\mathbf{x}_i) = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}) \equiv \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\eta_t\boldsymbol{\xi} \tag{55}$$

Now, we can write the conditional expectation without imposing the restrictions:

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta}_t + \bar{\mathbf{x}}_i\boldsymbol{\gamma}_t,\ t = 1, \ldots, T. \tag{56}$$

- Let $\boldsymbol{\pi}_t = (\alpha_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t)$. Use OLS for each t to estimate α_t, $\boldsymbol{\beta}_t$, and $\boldsymbol{\gamma}_t$, and then impose the restrictions using CMD.

- The “structural” parameters are

$$\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_T, \boldsymbol{\beta}', \boldsymbol{\xi}', \eta_2, \ldots, \eta_T,)'. \tag{57}$$

- The mapping from $\boldsymbol{\theta}$ to $\boldsymbol{\pi}$ is nonlinear. There are $T + 2TK$ elements of $\boldsymbol{\pi}$, and $2T - 1 + 2K$ elements of $\boldsymbol{\theta}$.
- Estimation of the η_t along with the other parameters requires some sort of nonlinear estimation. One could try pooled nonlinear least squares, but that is generally less efficient than minimum chi-square estimation.
- Recall that the usual FE estimator is consistent for $\boldsymbol{\beta}$ if $E(\ddot{\mathbf{x}}_{it}' c_i) = \mathbf{0}$ even if the η_t vary.

- It is easier to test $H_0 : \eta_2 = \eta_3 = \ldots = \eta_T = 1$. Write $\delta_t = \eta_t - 1$, $t = 2, \ldots, T$, so we have, in error form,

$$y_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \delta_t\bar{\mathbf{x}}_i\boldsymbol{\xi} + v_{it}, t = 1, \ldots, T. \quad (58)$$

- Under the null, $\delta_t = 0$, $t = 2, \ldots, T$. The estimator under the null, if we use either pooled OLS or RE, is the FE estimator of $\boldsymbol{\beta}$; we also get $\hat{\alpha}_t$, $\hat{\boldsymbol{\xi}}$.

- Applying the score principle, the gradient of the mean function, with parameters written as $(\alpha_1, \ldots, \alpha_T, \boldsymbol{\beta}', \boldsymbol{\xi}', \boldsymbol{\delta}')'$, is

$$(d1_t, \ldots, dT_t, \mathbf{x}_{it}, \bar{\mathbf{x}}_i + \delta_t \bar{\mathbf{x}}_i, d2_t \bar{\mathbf{x}}_i \boldsymbol{\xi}, \ldots, dT_t \bar{\mathbf{x}}_i \boldsymbol{\xi}) \tag{59}$$

Evaluated at the null this becomes

$$(d1_t, \ldots, dT_t, \mathbf{x}_{it}, \bar{\mathbf{x}}_i, d2_t \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}, \ldots, dT_t \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}) \tag{60}$$

- Therefore, the score test is a variable addition test, where we add the $T - 1$ regressors

$$(d2_t\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}},\ldots,dT_t\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}) \tag{61}$$

which is just the $T - 1$ time dummies interacted with the scalar $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$.

- So, first use the Mundlak regression

$$y_{it} \text{ on } 1,\ d2_t,\ \ldots,\ dT_t,\ \mathbf{x}_{it},\ \bar{\mathbf{x}}_i \tag{62}$$

to get $\hat{\boldsymbol{\xi}}$. Then use POLS or RE of

$$y_{it} \text{ on } 1,\ d2_t\ ,\ldots,\ dT_t,\ \mathbf{x}_{it},\ \bar{\mathbf{x}}_i,\ d2_t\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}},\ \ldots,\ dT_t\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}} \tag{63}$$

and test the last $T - 1$ regressors for joint significance.

- Even if RE is used in (63), a fully robust test should be used.
- If the test rejects can then use minimum distance estimation.

Alternatively, hope that FE is consistent for estimating $\boldsymbol{\beta}$ (assuming the η_t are not of interest).

- Adding time-constant regressors $\mathbf{z}_i$ causes no important changes.

Also, note that some elements of $\mathbf{x}_{it}$ might be interactions with the time dummies. It is the significance of the terms $dr_t\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ that signals time-varying factor loads.

- Under the Mundlak approach, $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ is an estimated proxy of c_i, and the idea is to see whether the coefficients on $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$.

4.3. Pseudo Panel Data

- Following is excerpted from Imbens and Wooldridge (2007, NBER Lectures).

- A pseudo panel data set is created from repeated cross sections. Units are grouped, often by birth year or geography or some other observable feature.

- For example, we might have cross sections of individuals from 2000 through 2009, and suppose we have people born from 1931 through 1970. A pseudo panel data set computes averages for each birth cohort in each year, resulting in $G = 40$ groups and $T = 10$ time periods. But the underlying model is at the individual level.

- Several different asymptotic frameworks have been proposed. Most seem to be against the spirit of obtaining large random samples from cross sections over time.
- Important to specify the underlying population model – specified at the individual (more generally, unit) level. Then, study how aggregation affects our ability to estimate population parameters.

- Write the standard, additive unobserved effects model, written for a generic unit in the population:

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f + u_t, \, t = 1, \ldots, T. \tag{64}$$

Notice how we assume that a model over T time periods. For this setup to make sense, it must be the case that we can think of a stationary population, so that the same units are represented in each time period. (Deaton, 1985)

- Because of the presence of the η_t, can set $E(f) = 0$.

- The random quantities in (64) are the response variable, y_t, the covariates, $\mathbf{x}_t$ (a $1 \times K$ vector), the unobserved effect, f, and the unobserved idiosyncratic errors, $\{u_t : t = 1, \ldots, T\}$.
- Subsequent analysis is for “small” T, so the η_t are parameters. Consider case where all elements of $\mathbf{x}_t$ have some time variation.
- What restrictions should we make? Contemporaneous exogeneity conditional on f, that is,

$$E(u_t|\mathbf{x}_t, f) = 0,\ t = 1, \ldots, T \tag{65}$$

is one possibility.

We will use an implication of (2):

$$E(u_t|f) = 0, t = 1,\ldots,T. \tag{66}$$

Think of (64) as representing $E(y_t|\mathbf{x}_t, f)$ where any time constant factors are lumped into f.

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

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + f_i + u_{it},\ t = 1,\ldots,T. \tag{67}$$

- Deaton (1985): Assume that the population for which (64) holds is divided into G groups (or cohorts). This designation cannot depend on time. Birth year, or ranges of birth years, our county of residence, are common.
- Condition (66) then implies

$$E(u_{it}|g_i) = 0,\ t = 1,\ldots,T. \tag{68}$$

The η_t account for any change in the average unobservables over time and f_i accounts for any time-constant factors.

- Take expected value of (67) conditional on group membership and use only (68):

$$E(y_{it}|g_i = g) = \eta_t + E(\mathbf{x}_{it}|g_i = g)\boldsymbol{\beta} + E(f_i|g_i = g), t = 1, \ldots, T. \tag{69}$$

- Should we be suspicious that we do not even need $E(u_t|\mathbf{x}_t, f) = 0$ to identify the parameters? (Yes.)
- Later we will see that the key assumption is that the structural model (64) does not require a full set of group/time effects. If such effects are required, then one way to think about the resulting misspecification is that $E(u_{it}|g_i = g)$ is not zero.

- If we define the population means

$$\begin{aligned} \alpha_g &= E(f_i | g_i = g) \\ \mu^y_{gt} &= E(y_{it} | g_i = g) \\ \boldsymbol{\mu}^{\mathbf{x}}_{gt} &= E(\mathbf{x}_{it} | g_i = g) \end{aligned} \tag{70}$$

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

$$\mu^y_{gt} = \eta_t + \boldsymbol{\mu}^{\mathbf{x}}_{gt} \boldsymbol{\beta} + \alpha_g, \, g = 1, \ldots, G, \ t = 1, \ldots, T. \tag{71}$$

- No restriction on the dependence between $\mathbf{x}_{it}$ and u_{ir} across t and r. $\mathbf{x}_{it}$ can contain lagged dependent variables, most commonly $y_{i,t-1}$, or contemporaneously endogenous variables (Angrist (1991), measurement error).
- Taking (71) as starting point for estimating $\boldsymbol{\beta}$ (along with η_t and α_g) makes the issues pretty clear. If we have sufficient observations in the group/time cells, then the means μ_{gt}^{y} and $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ can be estimated fairly precisely, and these can be used in a minimum distance estimation framework to estimate $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ consists of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$ (where, say, we set $\eta_1 = 0$ as the normalization).

- Suppose we knew population means. How well does (71) identify $\boldsymbol{\beta}$?

If we apply “pooled OLS” to the moments,

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \boldsymbol{\mu}_{gt}^{\mathbf{x}\prime} \boldsymbol{\mu}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \boldsymbol{\mu}_{gt}^{\mathbf{x}\prime} \mu_{gt}^{y} \right). \tag{72}$$

- Now add in the group and time effects:

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \mu_{gt}^{y} \right), \tag{73}$$

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

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} \text{ on } 1, d2, \ldots, dT, c2, \ ..., cG. \tag{74}$$

- Key point: Equation (73) shows that underlying model cannot contain a full set of group/time interactions. We *could* allow this feature with individual-level data. This is the key identifying restriction.
- $\boldsymbol{\beta}$ is not identified if we can write

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g$$

for vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_g$. Therefore, while we must exclude a full set of group/time effects in the structural model, we need some interaction between them in the distribution of the covariates across group/time.

- Even if we accept identification strategy, variation in $\{\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} : t = 1,..,T, g = 1,\ldots,G\}$ might not be sufficient to learn much about $\boldsymbol{\beta}$: we may be removing almost all of the variation in the mean of the covariates across group and time.

Estimation

- Assume we have a random sample on $(\mathbf{x}_t, y_t)$ of size N_t, and we have specified the G groups or cohorts. Write $\{(\mathbf{x}_{it}, y_{it}) : i = 1, \ldots, N_t\}$. For each random draw i, it is useful to let $\mathbf{r}_i = (r_{it1}, r_{it2}, \ldots, r_{itG})$ be a vector of group indicators, so $r_{itg} = 1$ if observation i is in group g (drawn at time t).

- The sample average on the response variable in group/time cell (g, t) can be written as

$$\hat{\mu}_{gt}^{y} = N_{gt}^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} = (N_{gt}/N_t)^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it}, \tag{75}$$

where $N_{gt} = \sum_{i=1}^{N_t} r_{itg}$ is properly treated as a random outcome.

- $\hat{\mu}_{gt}^{y}$ is generally consistent for μ_{gt}^{y}. First, $\hat{\rho}_{gt} = N_{gt}/N_t$ converges in probability to $\rho_g = P(r_{itg} = 1)$ – the fraction of the population in group or cohort g. So

$$\hat{\rho}_{gt}^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} \overset{p}{\to} \rho_g^{-1} E(r_{itg} y_{it}) = E(y_{it}|r_{itg} = 1) = \mu_{gt}^{y}.$$

- Let $\mathbf{w}_{it} = (y_{it}, \mathbf{x}_{it})'$. Then

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

where $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ is the sample average for group/time cell (g, t) and $\boldsymbol{\Omega}_{gt}^{\mathbf{w}} = Var(\mathbf{w}_t|g)$is the $(K + 1) \times (K + 1)$ variance matrix for group/time cell (g, t).

- When we stack the means across groups and time periods, it is helpful to have the result

$$\sqrt{N}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}} - \boldsymbol{\mu}_{gt}^{\mathbf{w}}) \rightarrow Normal(\mathbf{0}, (\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}_{gt}^{\mathbf{w}}), \tag{76}$$

where $N = \sum_{t=1}^{T} N_t$ and $\kappa_t = \lim_{N\rightarrow\infty}(N_t/N)$ is, essentially, the fraction of all observations accounted for by cross section t.

- $\rho_g\kappa_t$ is consistently estimated by N_{gt}/N, and so the sample average for cell (g,t) gets weighted by N_{gt}/N, the fraction of all observations accounted for by cell (g,t).

- Need a consistent estimator of $\mathbf{\Omega}_{gt}^{\mathbf{w}}$, and the group/time sample variance serves that purpose:

$$\hat{\mathbf{\Omega}}_{gt}^{\mathbf{w}} = N_{gt}^{-1} \sum_{i=1}^{N_t} r_{itg} (\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})(\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})' \xrightarrow{p} \mathbf{\Omega}_{gt}^{\mathbf{w}}. \tag{77}$$

- Let $\boldsymbol{\pi}$ be the vector of all cell means. For each (g,t), there are $K+1$ means, and so $\boldsymbol{\pi}$ is a $GT(K+1) \times 1$ vector. Stack $\boldsymbol{\pi}$ starting with the $K+1$ means for $g = 1$, $t = 1$, $g = 1$, $t = 2$, ..., $g = 1$, $t = T$, ..., $g = G$, $t = 1$, ..., $g = G$, $t = T$. Now, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are always independent across g because we assume random sampling for each t.

- When $\mathbf{x}_t$ does not contain lags or leads, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are independent across t, too. For now, assume this. Then,

$$\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \rightarrow \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}), \tag{78}$$

where $\boldsymbol{\Omega}$ is the $GT(K+1) \times GT(K+1)$ block diagonal matrix with (g, t) block $\boldsymbol{\Omega}_{gt}^{\mathbf{w}}/(\rho_g \kappa_t)$. Note that $\boldsymbol{\Omega}$ incorporates both different cell variance matrices as well as the different frequencies of observations. The set of equations in (8) constitute the restrictions on $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$. Let $\boldsymbol{\theta}$ be the $(K + T + G - 1)$ vector of these parameters, written as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$.

- There are GT restrictions in equations (71):

$$\mu_{gt}^{y} = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g,\ g = 1,\ldots,G,\ t = 1,\ldots,T;$$

there can be many overidentifying restrictions. Write these restrictions as

$$\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \mathbf{0}, \tag{79}$$

where $\mathbf{h}(\cdot,\cdot)$ is a $GT \times 1$ vector.

- In this MD problem, the parameters are not separable, but $\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is linear in each argument, which means MD estimators of $\boldsymbol{\theta}$ are in closed form.

- Do need an initial consistent estimator of $\boldsymbol{\theta}$. Straightforward is the "fixed effects" estimator described above.
- With the restrictions written as $\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \mathbf{0}$, Chamberlain (2007) shows that the optimal weighting matrix is the inverse of

$$\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})', \tag{80}$$

where $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is the $GT \times GT(K+1)$ Jacobian of $\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$. An estimator of $\boldsymbol{\pi}$ is just the cell averages. Can use the "fixed effects" estimator $\check{\boldsymbol{\theta}}$ as the initial consistent estimator of $\boldsymbol{\theta}$.

- Can show $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})'$ is a block diagonal matrix with blocks

$$(-1,\boldsymbol{\beta}')(\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}_{gt}^{w}(-1,\boldsymbol{\beta}')'. \tag{81}$$

But

$$\tau_{gt}^{2} \equiv (-1,\boldsymbol{\beta}')\boldsymbol{\Omega}_{gt}^{w}(-1,\boldsymbol{\beta}')' = Var(y_t - \mathbf{x}_t\boldsymbol{\beta}|g), \tag{82}$$

the error variance at time *t* of group *g*.

- A consistent estimator is

$$\hat{\tau}_{gt}^{2} = N_{gt}^{-1} \sum_{i=1}^{N_{gt}} r_{itg}(y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - \check{\eta}_{t} - \check{\alpha}_{g})^{2},$$

which is just the residual variance estimated within cell (g, t), using the preliminary estimates of $\boldsymbol{\beta}$, η_t, and α_g.

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

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

and so

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}^{y}. \qquad (83)$$

Here, $\hat{\mathbf{C}}$ is diagonal with entries $\hat{\tau}_{gt}^{2}/(N_{gt}/N)$.

- Efficient MD estimator looks like a "weighted least squares" estimator. The estimated asymptotic variance of $\hat{\boldsymbol{\theta}}$ is $(\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}/N$, where $\hat{\mathbf{C}}^{-1}$ is diagonal with entries $(N_{gt}/N)/\hat{\tau}_{gt}^2$. Easy to weight each cell (g,t) and then compute both $\hat{\boldsymbol{\theta}}$ and its asymptotic standard errors via a weighted regression.
- Must compute the $\hat{\tau}_{gt}^2$ using the individual-level data. Or, assume τ_{gt}^2 is constant, in which case the weight for cell (g,t) is simply N_{gt}/N.

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

$$\check{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\boldsymbol{\mu}}^{y}. \tag{84}$$

- It appears that we cannot get the correct asymptotic variance of $\check{\boldsymbol{\theta}}$ by using "heteroskedasticity-robust" estimator in the regression $\hat{\mu}_{gt}^{y}$ on $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, $\mathbf{d}_t$, $\mathbf{c}_g$.
- Simulation study could see how these standard errors behave. Results of Stock and Watson (*Econometrica*, 2008) seem to imply they cannot be correct with "small" T.)

- (1) Several papers, including Deaton (1985), Verbeek and Nijman (1993), and Collado (1998), use a different asymptotic analysis. In the current notation, $GT \rightarrow \infty$ (Deaton) or $G \rightarrow \infty$, with the cell sizes fixed. Seems unnatural. $T \rightarrow \infty$ makes conceptual sense but T is usually small.

(2) McKenzie (2004) shows estimators derived under large G asymptotics can have good properties under the MD asymptotics. Turns out IV estimators proposed by Collado (1998), Verbeek and Vella (2005), are just different ways of using the population moment conditions.

- Inoue (2008) comes closer, but gets nonnormal limiting distribution. (Asymmetry in treating moments of y and $\mathbf{x}$.)

- Application to models with lags relatively straightforward. The only difference now is that the vectors of means, $\{\boldsymbol{\mu}_{gt}^{\mathbf{w}} : g = 1, \ldots, G; t = 1, \ldots, T)$ contain redundancies, so modify the moment conditions. Suppose

$$y_t = \eta_t + \rho y_{t-1} + \mathbf{z}_t \boldsymbol{\gamma} + f + u_t \tag{85}$$
$$E(u_t | g) = 0, \; g = 1, \ldots, G$$

Original moments are still valid, but the vector of means would be $(\mu_{gt}^{y}, \boldsymbol{\mu}_{gt}^{\mathbf{z}})$, and then appropriately pick off μ_{gt}^{y} in defining the moment conditions.

- MD approach exposes how pseudo panels identify population parameters. Seems tenuous. Need a careful simulation study, where individual-level data are generated from the population model, and where g_i – the group identifier – is randomly drawn, too. Underlying model should have full time effects. Verbeek and Vella (2005) come close, but omit aggregate time effects in the main model while generating the explanatory variables to have means that differ by group/time cell.

- A key point is that even if we can get precise estimates of the cell means – which is often the case with survey data – the nature of the variation in $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ across g and t might not be enough to precisely estimate $\boldsymbol{\beta}$. At a minimum, if we write

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g + \boldsymbol{\psi}_{gt}$$

then we need $\boldsymbol{\psi}_{gt} \neq \mathbf{0}$ for at least some pairs (g, t).

- At the same time we cannot allow unrestricted group/time interactions in the individual-level model.