What's New in Econometrics?
Lecture 6
Control Functions and Related Methods

Jeff Wooldridge
NBER Summer Institute, 2007

1. Linear-in-Parameters Models: IV versus Control Functions

2. Correlated Random Coefficient Models

3. Some Common Nonlinear Models and Limitations of the CF Approach

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.
- 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

$$\mathrm{E}(\mathbf{z}'u_1) = \mathbf{0}. \tag{2}$$

Reduced form for y_2:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2,\ \mathrm{E}(\mathbf{z}'v_2) = \mathbf{0} \tag{3}$$

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 $\rho_1 = \mathrm{E}(v_2 u_1)/\mathrm{E}(v_2^2)$ is the population regression coefficient. By construction, $\mathrm{E}(v_2 e_1) = 0$ and $\mathrm{E}(\mathbf{z}'e_1) = \mathbf{0}$.

Plug (4) into (1):

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{5}$$

where we now view v_2 as an explanatory variable in the equation. By controlling for v_2, the error e_1 is uncorrelated with y_2 as well as with v_2 and $\mathbf{z}$.

- Two-step procedure: (i) Regress y_2 on $\mathbf{z}$ and

obtain the reduced form residuals, $\hat{v}_2$; (ii) Regress

$$y_1 \text{ on } \mathbf{z}_1, y_2, \text{ and } \hat{v}_2. \tag{6}$$

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 $\rho_1 = 0$. OLS estimators from (6) will be consistent for δ_1, α_1, and ρ_1. Simple test for null of exogeneity is (heteroskedasticity-robust) t statistic on $\hat{v}_2$.

- 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 starting from (1).
- Now extend the model:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{7}$$

$$\mathrm{E}(u_1|\mathbf{z}) = 0. \tag{8}$$

Let z_2 be a scaler not also in $\mathbf{z}_1$. Under the (8) – which is stronger than (2), 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? We require an assumption about $\mathrm{E}(u_1|\mathbf{z}, y_2)$, say

$$\mathrm{E}(u_1|\mathbf{z}, y_2) = \mathrm{E}(u_1|v_2) = \rho_1 v_2, \tag{9}$$

where the first equality would hold if (u_1, v_2) is independent of $\mathbf{z}$ – a nontrivial restriction on the reduced form error in (3), not to mention the structural error u_1. Linearity of $\mathrm{E}(u_1|v_2)$ is a substantive restriction. Now,

$$\mathrm{E}(y_1|\mathbf{z}, y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2, \tag{10}$$

and a CF approach is immediate: replace v_2 with $\hat{v}_2$

and use OLS on (10).

- These CF estimates are *not* the same as the 2SLS estimates using any choice of instruments for (y_2, y_2^2). CF approach likely more efficient, but less robust. For example, (8) implies $\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$.
- CF approaches can impose extra assumptions even in the simple model (1). For example, if y_2 is a binary response, the CF approach based on $E(y_1|\mathbf{z}, y_2)$ involves estimating

$$E(y_1|\mathbf{z}, y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z}, y_2). \tag{11}$$

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

$$E(u_1|\mathbf{z}, y_2) = \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \tag{12}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio (IMR). This leads to the Heckman two-step

estimate (for endogeneity, not sample selection). Obtain the probit estimate $\hat{\boldsymbol{\delta}}_2$ and 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} 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 $E(u_1|v_2)$. If we just apply 2SLS directly to (1), it makes no distinction among discrete, continuous, or some mixture for y_2.
- How might we robustly use the binary nature of y_2 in IV estimation? Obtain the fitted probabilities, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, from the first stage probit, and then use these as IVs for y_{i2}. This is fully robust to misspecification of the probit model and the usual standard errors from IV are asymptotically valid. It

is the efficient IV estimator if

$P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$.

2. Correlated Random Coefficient Models

Modify (1) 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. Think of a_1 as an omitted variable that interacts with y_2.Following Heckman and Vytlacil (1998), we refer to (13) as a correlated random coefficient (CRC) model.

- 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

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \tag{14}$$

$$\equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1, \tag{15}$$

- The potential problem with applying instrumental variables to (15) is that the error term $v_1 y_2 + u_1$ is

not necessarily uncorrelated with the instruments $\mathbf{z}$, even under

$$\mathrm{E}(u_1|\mathbf{z}) = \mathrm{E}(v_1|\mathbf{z}) = 0. \tag{16}$$

We want to allow y_2 and v_1 to be correlated, $\mathrm{Cov}(v_1, y_2) \equiv \tau_1 \neq 0$. A suffcient condition that allows for any *unconditional* correlation is

$$\mathrm{Cov}(v_1, y_2|\mathbf{z}) = \mathrm{Cov}(v_1, y_2), \tag{17}$$

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 $\hat{y}_2 = \mathbf{z}\hat{\boldsymbol{\pi}}_2$. See notes.

- Condition (17) cannot really hold for discrete y_2. 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 $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]$ and $v_2|\mathbf{z}$ is Normal$(0,1)$, then

$$E(y_1|\mathbf{z},y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2,$$

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. The two-step estimation method is the one due to Heckman (1976).

- Can also interact the exogenous variables with $h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$. Or, allow $E(v_1|v_2)$ to be more

flexible, as in Heckman and MaCurdy (1986).

3. Some Common Nonlinear Models and Limitations of the CF Approach

● CF 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.

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 Blundell-Smith (1986) and Rivers-Vuong (1988) approach is to make a

homoskedastic-normal assumption on the reduced form for y_2,

$$y_2 = \mathbf{z}\pi_2 + v_2, \;\; v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2). \tag{19}$$

A key point is that the RV approach essentially requires

$$(u_1, v_2) \text{ independent of } \mathbf{z}. \tag{20}$$

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 RV two-step approach is

(i) OLS of y_2 on $\mathbf{z}$, to obtain the residuals, $\hat{v}_2$.

(ii) Probit of y_1 on $\mathbf{z}_1, y_2, \hat{v}_2$ to estimate the scaled coefficients. A simple t test on $\hat{v}_2$ is valid to test $\mathrm{H}_0 : \rho_1 = 0$.

- Can recover the original coefficients, which appear in the partial effects. Or,

$$\widehat{\mathrm{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1} \hat{v}_{i2}), \tag{23}$$

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. Can use the the *same* two-step because the Bernoulli log likelihood is in the linear exponential family. Still estimate scaled coefficients. 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.
- 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 normally distributed, $\mathrm{P}(y_1 = 1|\mathbf{z}) = \Phi\{[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2)]/\omega_1\}$. So first do OLS on the reduced form, and get fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$. Then, 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. 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.

- Extension to random coefficients:

$$E(y_1|\mathbf{z}, y_2, \mathbf{c}_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + q_1), \tag{25}$$

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, is

$$\alpha_1 \phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2). \tag{26}$$

The APE in this case is much messier.

- Could just implement flexible CF approaches without formally starting with a "structural" model.

For example, could just do Bernoulli QMLE of y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, $\hat{v}_{i2}$, and $y_{i2}\hat{v}_{i2}$. Even here, APE can be different sign from α_1.

- 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. 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$(u_1|y_2, \mathbf{z})$ or $D(y_2|\mathbf{z})$. (y_2 cannot play the role) Cannot be an instrument as we usually think of it. Can be a variable randomized to be independent of y_2 and $\mathbf{z}$.
- Returning to the response function

$E(y_1|\mathbf{z},y_2,q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1)$, we can understand the limits of the CF approach for estimating nonlinear models with discrete EEVs. The Rivers-Vuong approach does not work.We cannot write $D(y_2|\mathbf{z})$ =Normal$(\mathbf{z}\boldsymbol{\pi}_2, \tau_2^2)$. 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.

- There some poor strategies that still linger. Suppose y_1 and y_2 are both binary and

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0] \tag{27}$$

and we maintain joint normality of (u_1, v_2).We should *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 $\hat{\Phi}_2$.

- Currently, the only strategy we have 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.)
- Yes, "bivariate" probit software be used to estimate the probit model with a binary endogenous variable. In fact, with any function of $\mathbf{z}_1$ and y_2 as explanatory variables.
- Parallel discussions hold for ordered probit, Tobit.

Multinomial Responses

- Recent push, by Villas-Boas (2005) and Petrin and Train (2006), 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

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2, \tag{28}$$

then we jump directly to convenient models for $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$. The average structural functions are obtained by averaging the response probabilities across $\hat{\mathbf{v}}_{i2}$. No convincing way to handle discrete $\mathbf{y}_2$, though.

Exponential Models

- Both IV approaches and CF approaches are available for exponential models. With a single EEV, write

$$E(y_1|\mathbf{z}, y_2, r_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + r_1), \tag{29}$$

where r_1 is the omitted variable. (Extensions to

general nonlinear functions $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)$ are immediate; we just add those functions with linear coefficients to (29). CF methods based on

$$E(y_1|\mathbf{z}, y_2, r_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)E[\exp(r_1)|\mathbf{z}, y_2]$$

This has been worked through when $D(y_2|\mathbf{z})$ is homoskedastic normal (Wooldridge, 1997 – see notes for a random coefficient version where α_1 becomes a_1 with $E(a_1) = \alpha_1$) and $D(y_2|\mathbf{z})$ follows a probit (Terza, 1998). In the latter case,

$$E(y_1|\mathbf{z}, y_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1)$$

$$h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1) = \exp(\theta_1^2/2)\{y_2\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2) + (1 - y_2)[1 - \Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}$$

- IV methods that work for any $\mathbf{y}_2$ are also available, as developed by Mullahy (1997). If

$$E(y_1|\mathbf{z}, \mathbf{y}_2, r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \qquad (30)$$

and r_1 is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(r_1)|\mathbf{z}] = 1, \tag{31}$$

where $E[\exp(r_1)] = 1$ is a normalization. The moment conditions are

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0. \tag{32}$$

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

$$\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{E}(y_1|\mathbf{z}, v_2) = H(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{33}$$

for some (generally unknown) function $H(\cdot,\cdot)$. The average structural function is just

$\text{ASF}(\mathbf{z}_1, y_2) = \text{E}_{v_{i2}}[H(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$.

- 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

$$\widehat{\text{ASF}}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N} \hat{H}(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1, \hat{v}_{i2}); \qquad (34)$$

- 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 also entirely nonparametric. They also allow $\hat{g}_2(\cdot)$ to be fully nonparametric. Parametric approximations in each stage might produce good estimates of the APEs.

- BP (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

$$ASF_1(\mathbf{z}_1, \mathbf{y}_2) = \int g_1(\mathbf{z}_1, \mathbf{y}_2, u_1) dF_1(u_1), \tag{35}$$

where F_1 is the distribution of u_1. The key restrictions are that $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2, \tag{36}$$

where $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$. The key is that the ASF can be obtained from $E(y_1|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2) = h_1(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ by averaging out $\mathbf{v}_2$, and fully nonparametric two-step estimates are available.

- Provides justification for the parametric versions discussed earlier, where the step of modeling $g_1(\cdot)$ in $y_1 = g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)$ can be skipped.

● Imbens and Newey (2006) consider the triangular system, but without additivity in the reduced form of y_2,

$$y_2 = g_2(\mathbf{z}, e_2), \tag{37}$$

where $g_2(\mathbf{z}, \cdot)$ is strictly monotonic. Rules out discrete y_2 but allows some interaction between the unobserved heterogeneity in y_2 and the exogenous variables. When (u_1, e_2) is independent of $\mathbf{z}$, a valid control function to 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 $\mathbf{z}$.

5. Methods for Panel Data

● Combine methods for handling correlated random effects models with control function methods to estimate certain nonlinear panel data models with unobserved heterogeneity and EEVs.

- Illustrate a parametric approach used by Papke and Wooldridge (2007), which applies to binary and fractional responses.
- In this model, 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. Perhaps jackknife methods can be adapted, but currently the assumptions are very strong (serial independence, homogeneity over time, exogenous regressors).
- Model with time-constant unobserved heterogeneity, c_{i1}, and time-varying unobservables, v_{it1}, as

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1}). \quad (38)$$

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:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + a_{i1}, a_{i1} | \mathbf{z}_i \sim \text{Normal}(0, \sigma^2_{a_1}). \qquad (39)$$

We could allow the elements of $\mathbf{z}_i$ to appear with separate coefficients, too. Note that only exogenous variables are included in $\bar{\mathbf{z}}_i$. Next step:

$$E(y_{it1} | y_{it2}, \mathbf{z}_i, r_{it}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1} \boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + r_{it1})$$

where $r_{it1} = a_{i1} + v_{it1}$. Next, we assume a linear reduced form for y_{it2}:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\delta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T. \quad (40)$$

Rules out discrete y_{it2} because

$$r_{it1} = \eta_1 v_{it2} + e_{it1}, \quad (41)$$

$$e_{it1}|(\mathbf{z}_i, v_{it2}) \sim \text{Normal}(0, \sigma^2_{e_1}), t = 1, \ldots, T. \quad (42)$$

Then

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = \Phi(\alpha_{e1}y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1}v_{it2}) \quad (43)$$

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,

$$E_{(c_{i1},v_{it1})}[\Phi(\alpha_1 y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1})], \qquad (44)$$

which is consistently estimated as

$$N^{-1}\sum_{i=1}^{N}\Phi(\hat{\alpha}_{e1}y_{t2}+\mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{e1}+\hat{\psi}_{e1}+\bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{e1}+\hat{\eta}_{e1}\hat{v}_{it2}). \quad (45)$$

These APEs, typically with further averaging out across t and the values of y_{t2} and $\mathbf{z}_{t1}$, can be compared directly with fixed effects IV estimates.

- We can use the approaches of Altonji and Matzkin (2005), Blundell and Powell (2003), and Imbens and Newey (2006) to make the analysis less parametric. For example, we might replace (40) with $y_{it2} = g_2(\mathbf{z}_{it}, \bar{\mathbf{z}}_i) + v_{it2}$ or $y_{it2} = g_2(\mathbf{z}_{it}, \bar{\mathbf{z}}_i, e_{it2})$ under monotonicity in e_2. Then a reasonable assumption is

$$D(c_{i1} + v_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = D(c_{i1} + v_{it1}|\bar{\mathbf{z}}_i, v_{it2}) \quad (46)$$

where, in the Imbens and Newey case, $v_{it2} = F_{y_{t2}|(\mathbf{z}_t,\bar{\mathbf{z}})}(y_{it2}|\mathbf{z}_{it}, \bar{\mathbf{z}}_i)$. After a first stage

estimation, the ASF can be obtained by estimating $E(y_{it1}|\mathbf{z}_{it1}, y_{it2}, \bar{\mathbf{z}}_i, v_{it2})$ and then averaging out across $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$.

"What's New in Econometrics"

Lecture 7

Bayesian Inference

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction

2. Basics

3. Bernstein-Von Mises Theorem

4. Markov-Chain-Monte-Carlo Methods

5. Example: Demand Models with Unobs Heterog in Prefer.

6. Example: Panel Data with Multiple Individual Specific Param.

7. Instrumental Variables with Many Instruments

8. Example: Binary Response with Endogenous Regressors

9. Example: Discrete Choice Models with Unobserved Choice Characteristics

1. Introduction

Formal Bayesian methods surprisingly rarely used in empirical work in economics.

Surprising, because they are attractive options in many settings, especially with many parameters (like random coefficient models), when large sample normal approximations are not accurate. (see examples below)

In cases where large sample normality does not hold, frequentist methods are sometimes awkward (e.g, confidence intervals that can be empty, such as in unit root or weak instrument cases).

Bayesian approach allows for conceptually straightforward way of dealing with unit-level heterogeneity in preferences/parameters.

Why are Bayesian methods not used more widely?

1. choice of methods does not matter (bernstein-von mises theorem)

2. difficulty in specifying prior distribution (not "objective")

3. need for fully parametric model

4. computational difficulties

2.A Basics: The General Case

Model:

$$f_{X|\theta}(x|\theta).$$

As a function of the parameter this is called the likelihood function, and denoted by $\mathcal{L}(\theta|x)$.

A prior distribution for the parameters, $p(\theta)$.

The posterior distribution,

$$p(\theta|x) = \frac{f_{X,\theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\theta}(x|\theta) \cdot p(\theta)}{\int f_{X|\theta}(x|\theta) \cdot p(\theta) d\theta}.$$

Note that, as a function of θ, the posterior is proportional to

$$p(\theta|x) \propto f_{X|\theta}(x|\theta) \cdot p(\theta) = \mathcal{L}(\theta|x) \cdot p(\theta).$$

2.B Example: The Normal Case

Suppose the conditional distribution of X given the parameter μ is $\mathcal{N}(\mu, 1)$.

Suppose the prior distribution for μ to be $\mathcal{N}(0, 100)$.

The posterior distribution is proportional to

$$f_{\mu|X}(\mu|x) \propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot 100}\mu^2\right)$$

$$= \exp -\frac{1}{2}(x^2 - 2x\mu + \mu^2 + \mu^2/100)$$

$$\propto \exp(-\frac{1}{2(100/101)}(\mu - (100/101)x)^2)$$

$$\sim \mathcal{N}(x \cdot 100/101, 100/101)$$

2.B The Normal Case with General Normal Prior Distribution

Model: $\mathcal{N}(\mu, \sigma^2)$
Prior distribution for μ is $\mathcal{N}(\mu_0, \tau^2)$.

Then the posterior distribution is:

$$f_{\mu|X}(\mu|x) \sim \mathcal{N}\left(\frac{x/\sigma^2 + \mu_0/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\tau^2 + 1/\sigma^2)}\right).$$

The result is quite intuitive: the posterior mean is a weighted average of the prior mean μ_0 and the observation x with weights proportional to the precision, $1/\sigma^2$ for x and $1/\tau^2$ for μ_0:

$$\mathbb{E}[\mu|X = x] = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \quad \frac{1}{\mathbb{V}(\mu|X)} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}.$$

Suppose we are really sure about the value of μ before we conduct the experiment. In that case we would set τ^2 small and the weight given to the observation would be small, and the posterior distribution would be close to the prior distribution.

Suppose on the other hand we are very unsure about the value of μ. What value for τ should we choose? We can let τ go to infinity. Even though the prior distribution is not a proper distribution anymore if $\tau^2 = \infty$, the posterior distribution is perfectly well defined, namely $\mu|X \sim \mathcal{N}(X, \sigma^2)$.

In that case we have an improper prior distribution. We give equal prior weight to any value of μ (flat prior). That would seem to capture pretty well the idea that a priori we are ignorant about μ.

This is not always easy to do. For example, a flat prior distribution is not always uninformative about particular functions of parameters.

2.C The Normal Case with Multiple Observations

N independent draws from $\mathcal{N}(\mu, \sigma^2)$, σ^2 known.

Prior distribution on μ is $\mathcal{N}(\mu_0, \tau^2)$.

The likelihood function is

$$\mathcal{L}(\mu|\sigma^2, x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right),$$

Then

$$\mu|X_1, \ldots, X_N$$

$$\sim \mathcal{N}\left(\overline{x} \cdot \frac{1}{1 + \sigma^2/(N \cdot \tau^2)} + \mu_0 \cdot \frac{\sigma^2/(N\tau^2)}{1 + \sigma^2/(N\tau^2)}, \frac{\sigma^2/N}{1 + \sigma^2/(N\tau^2)}\right)$$

3.A Bernstein-Von Mises Theorem: normal example

When N is large

$$\sqrt{N}(\bar{x} - \mu)|x_1, \ldots, x_N \approx \mathcal{N}(0, \sigma^2).$$

In large samples the prior does not matter.

Moreover, in a frequentist analysis, in large samples,

$$\sqrt{N}(\bar{x} - \mu)|\mu \sim \mathcal{N}(0, \sigma^2).$$

Bayesian probability and frequentiest confidence intervals agree:

$$\Pr\left(\mu \in \left[\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}, \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}\right] \middle| X_1, \ldots, X_N\right)$$

$$\approx \Pr\left(\mu \in \left[\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}, \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}\right] \middle| \mu\right) \approx 0.95;$$

3.B Bernstein-Von Mises Theorem: general case

This is known as the Bernstein-von Mises Theorem. Here is a general statement for the scalar case. Let the information matrix $\mathcal{I}_\theta$ at θ:

$$\mathcal{I}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f_X(x|\theta)\right] = -\int \frac{\partial^2}{\partial\theta\partial\theta'} \ln f_X(x|\theta) f_X(x|\theta) dx,$$

and let $\sigma^2 = \mathcal{I}_{\theta_0}^{-1}$.

Let $p(\theta)$ be the prior distribution, and $p_{\theta|X_1,\ldots,X_N}(\theta|X_1,\ldots,X_N)$ be the posterior distribution.

Now let us look at the distribution of a transformation of θ, $\gamma = \sqrt{N}(\theta - \theta_0)$, with density $p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) = p_{\theta|X_1,\ldots,X_N}(\theta_0 + \sqrt{N}\cdot\gamma|X_1,\ldots,X_N)/\sqrt{N}$.

Now let us look at the posterior distribution for γ if in fact the data were generated by $f(x|\theta)$ with $\theta = \theta_0$. In that case the posterior distribution of γ converges to a normal distribution with mean zero and variance equal to σ^2 in the sense that

$$\sup_{\gamma} \left| p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) - \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\gamma^2\right) \right| \longrightarrow 0.$$

See Van der Vaart (2001), or Ferguson (1996).

At the same time, if the true value is θ_0, then the mle $\widehat{\theta}_{mle}$ also has a limiting distribution with mean zero and variance σ^2:

$$\sqrt{N}(\widehat{\theta}_{ml} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

The implication is that we can interpret confidence intervals as approximate probability intervals from a Bayesian perspective.

Specifically, let the 95% confidence interval be $[\widehat{\theta}_{ml} - 1.96 \cdot \widehat{\sigma}/\sqrt{N}, \widehat{\theta}_{ml} + 1.96 \cdot \widehat{\sigma}/\sqrt{N}]$. Then, approximately,

$$\Pr\left(\widehat{\theta}_{ml} - 1.96 \cdot \widehat{\sigma}/\sqrt{N} \leq \theta \leq \widehat{\theta}_{ml} + 1.96 \cdot \widehat{\sigma}/\sqrt{N} \middle| X_1, \ldots, X_N\right)$$

$$\longrightarrow 0.95.$$

3.C When Bernstein-Von Mises Fails

There are important cases where this result does not hold, typically when convergence to the limit distribution is not uniform.

One is the unit-root setting. In a simple first order autoregressive example it is still the case that with a normal prior distribution for the autoregressive parameter the posterior distribution is normal (see Sims and Uhlig, 1991).

However, if the true value of the autoregressive parameter is unity, the sampling distribution is not normal even in large samples.

In such settings one has to take a more principled stand whether one wants to make subjective probability statements, or frequentist claims.

4. Numerical Methods: Markov-Chain-Monte-Carlo

The general idea is to construct a chain, or sequence of values, $\theta_0, \theta_1, \ldots$, such that for large k, θ_k can be viewed as a draw from the posterior distribution of θ given the data.

This is implemented through an algorithm that, given a current value of the parameter vector θ_k, and given the data $X_1, \ldots, X_N$ draws a new value θ_{k+1} from a distribution $f(\cdot)$ indexed by θ_k and the data:

$$\theta_{k+1} \sim f(\theta|\theta_k, \text{data}),$$

in such a way that if the original θ_k came from the posterior distribution, then so does θ_{k+1}

$$\theta_k|\text{data} \sim p(\theta|\text{data}), \quad \text{then} \quad \theta_{k+1}|\text{data} \sim p(\theta|\text{data}).$$

In many cases, irrespective of where we start, that is, irrespective of θ_0, as $k \longrightarrow \infty$, it will be the case that the distribution of the parameter conditional only on the data converges to the posterior distributionas $k \longrightarrow \infty$:

$$\theta_k|\text{data} \xrightarrow{d} p(\theta|\text{data}),$$

Then just pick a θ_0 and approximate the mean and standard deviation of the posterior distribution as

$$\widehat{\mathbb{E}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \theta_k,$$

$$\widehat{\mathbb{V}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \Big(\theta_k - \widehat{\mathbb{E}}[\theta|\text{data}]\Big)^2.$$

The first $K_0 - 1$ iterations are discarded to let algorithm converge to the stationary distribution, or "burn in."

4.A Gibbs Sampling

The idea being the Gibbs sampler is to partition the vector of parameters θ into two (or more) parts, $\theta' = (\theta_1', \theta_2')$. Instead of sampling θ_{k+1} directly from a conditional distribution of

$$f(\theta|\theta_k, \text{data}),$$

it may be easier to sample $\theta_{1,k+1}$ from the conditional distribution

$$p(\theta_1|\theta_{2,k}, \text{data}),$$

and then sample $\theta_{2,k+1}$ from

$$p(\theta_2|\theta_{1,k+1}, \text{data}).$$

It is clear that if $(\theta_{1,k}, \theta_{2,k})$ is from the posterior distribution, then so is $(\theta_{1,k}, \theta_{2,k})$.

4.B Data Augmentation

Suppose we are interested in estimating the parameters of a censored regression or Tobit model. There is a latent variable

$$Y_i^* = X_i'\beta + \varepsilon_i, \qquad \varepsilon_i | X_i \sim \mathcal{N}(0,1)$$

We observe

$$Y_i = \max(0, Y_i^*),$$

and the regressors X_i. Suppose the prior distribution for β is normal with some mean μ, and some covariance matrix Ω.

The posterior distribution for β does not have a closed form expression. The first key insight is to view both the vector $\mathbf{Y}^* = (Y_1^*, \ldots, Y_N^*)$ and β as unknown random variables.

The Gibbs sampler consists of two steps. First we draw all the missing elements of $\mathbf{Y}^*$ given the current value of the parameter β, say β_k

$$Y_i^* | \beta, \text{data} \sim \mathcal{TN}\left(X_i'\beta, 1, 0\right),$$

if observation i is truncated, where $\mathcal{TN}(\mu, \sigma^2, c)$ denotes a truncated normal distribution with mean μ, variance σ^2, and truncation point c (truncated from above).

Second, we draw a new value for the parameter, β_{k+1} given the data and given the (partly drawn) $\mathbf{Y}^*$:

$$p\left(\beta | \text{data}, \mathbf{Y}^*\right) \sim \mathcal{N}\left(\left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1} \cdot \left(\mathbf{X}'\mathbf{Y} + \Omega^{-1}\mu\right), \left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1}\right.$$

4.C Metropolis Hastings

We are again interested in $p(\theta|\text{data})$. In this case $\mathcal{L}(\theta|\text{data})$ is assumed to be easy to evaluate. Draw a new candidate value for the chain from a candidate distribution $q(\theta|\theta_k, \text{data})$. We will either accept the new value with probability The probability that the new draw θ is accepted is

$$\rho(\theta_k, \theta) = \min\left(1, \frac{p(\theta|\text{data}) \cdot q(\theta_k|\theta, \text{data})}{p(\theta_k|\text{data}) \cdot q(\theta|\theta_k, \text{data})}\right),$$

so that

$$\Pr\left(\theta_{k+1} = \theta\right) = \rho(\theta_k, \theta), \quad \text{and} \quad \Pr\left(\theta_{k+1} = \theta_k\right) = 1 - \rho(\theta_k, \theta).$$

The optimal (typically infeasible) choice for the candidate distribution is

$$q^*(\theta|\theta_k, \text{data}) = p(\theta|\text{data}) \quad \Longrightarrow \quad \rho(\theta_k, \theta) = 1$$

5. Example: Demand Models with Unobs Heterog in Prefer.

Rossi, McCulloch, and Allenby (1996, RMA) are interested in the optimal design of coupon policies. Supermarkets can choose to offer identical coupons for a particular product.

Alternatively, they may choose to offer differential coupons based on consumer's fixed characteristics.

Taking this ever further, they could tailoring the coupon value to the evidence for price sensitivity contained in purchase patterns.

Need to allow for household-level heterogeneity in taste parameters and price elasticities. Even with large amounts of data available, there will be many households for whom these parameters cannot be estimated precisely. RMA therefore use a hiearchical, or random coefficients model.

RMA model households choosing the product with the highest utility, where utility for household i, product j, $j = 0, 1, \ldots, J$, at purchase time t is

$$U_{ijt} = X'_{it}\beta_i + \epsilon_{ijt},$$

with the ϵ_{ijt} independent accross households, products and purchase times, and normally distributed with product-specific variances σ_j^2 (and σ_0^2 normalized to one).

The X_{it} are observed choice characteristics that in the RMA application include price, some marketing variables, as well as brand dummies.

All choice characteristics are assumed to be exogenous, although that assumption may be questioned for the price and marketing variables.

Because for some households we have few purchases, it is not possible to accurately estimate all β_i parameters. RMA therefore assume that the household-specific taste parameters are random draws from a normal distribution centered at $Z_i'\Gamma$:

$$\beta_i = Z_i'\Gamma + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \Sigma).$$

Now Gibbs sampling can be used to obtain draws from the posterior distribution of the β_i.

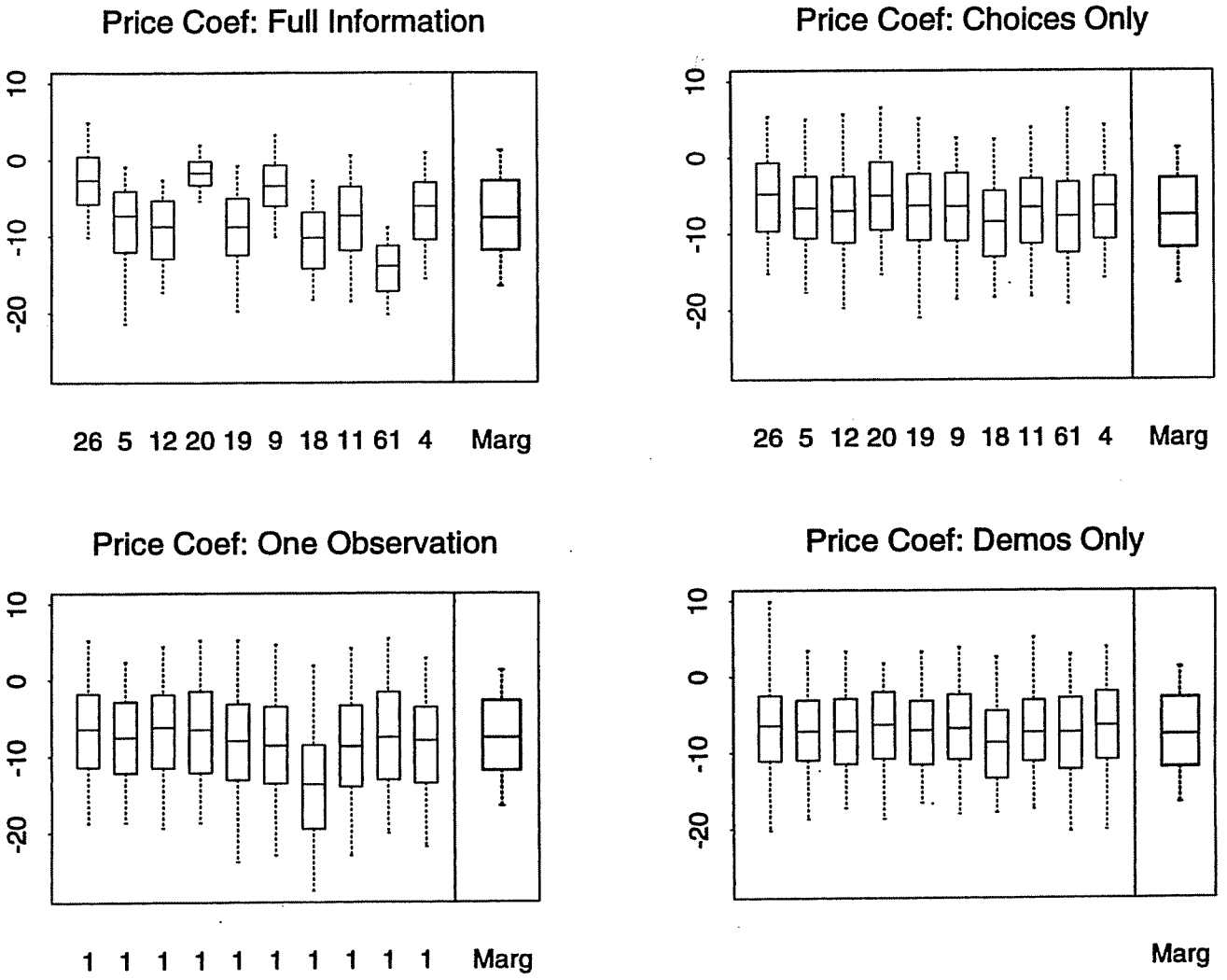
The **first** step is to draw the household parameters β_i given the utilities U_{ijt} and the common parameters Γ, Σ, and σ_j^2. This is straightforward, because we have a standard normal linear model for the utilities, with a normal prior distribution for β_i with parameters $Z_i'\Gamma$ and variance Σ, and T_i observations. We can draw from this posterior distribution for each household i.

In the **second** step we draw the σ_j^2 using the results for the normal distribution with known mean and unknown variance.

The **third** step is to draw from the posterior of Γ and Σ, given the β_i. This again is just a normal linear model, now with unknown mean and unknown variance.

The **fourth** step is to draw the unobserved utilities given the β_i and the data. Doing this one household/choice at a time, conditioning on the utilities for the other choices, this merely involves drawing from a truncated normal distribution, which is simple and fast.

Figure 2 **Boxplots of posterior distributions of household price coefficients. Various information sets. 10 selected households with the number of purchase occasions indicated along the *X* axis below each boxplot. The boxplot labelled "Marg" is the predictive distribution for a representative household from the model heterogeneity distribution. Note that these are the 11–20th households as ordered in our dataset.**



6. Example: Panel Data with Multiple Individual Specific Param.

Chamberlain and Hirano are interested in deriving predictive distributions for earnings using longitudinal data, using the model

$$Y_{it} = X_i'\beta + V_{it} + \alpha_i + U_{it}/h_i.$$

The second component in the model, V_{it}, is a first order autoregressive component,

$$V_{it} = \gamma \cdot V_{it-1} + W_{it},$$

$$V_{i1} \sim \mathcal{N}(0, \sigma_v^2), \qquad W_{it} \sim \mathcal{N}(0, \sigma_w^2).$$

$$U_{it} \sim \mathcal{N}(0, 1).$$

Analyzing this model by attempting to estimate the α_i and h_i directly would be misguided. From a Bayesian perspective this corresponds to assuming a flat prior distribution on a high-dimensional parameter space.

To avoid such pitfalls CH model α_i and h_i through a random effects specification.

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2). \quad \text{and} \quad h_i \sim \mathcal{G}(m/2, \tau/2).$$

In their empirical application using data from the Panel Study of Income Dynamics (PSID), CH find strong evidence of heterogeneity in conditional variances.

quantiles of the predictive dist. of $1/\sqrt{h_i}$

	Quantile						
Sample	0.05	0.10	0.25	0.50	0.75	0.90	0.95
All (N=813)	0.04	0.05	0.07	0.11	0.20	0.45	0.81
HS Dropouts (N=37)	0.06	0.08	0.11	0.16	0.27	0.49	0.79
HS Grads (N=100)	0.04	0.05	0.06	0.11	0.21	0.49	0.93
C Grads (N=122)	0.03	0.04	0.05	0.09	0.18	0.40	0.75

However, CH wish to go beyond this and infer individual-level predictive distributions for earnings.

Taking a particular individual, one can derive the posterior distribution of α_i, h_i, β, σ_v^2, and σ_w^2, given that individual's earnings as well as other earnings, and predict future earnings.

		0.90-0.10 quantile	
individual	sample std	1 year out	5 years out
321	0.07	0.32	0.60
415	0.47	1.29	1.29

The variation reported in the CH results may have substantial importance for variation in optimal savings behavior by individuals.

7. Example: Instrumental Variables with Many Instruments

Chamberlain and Imbens analyze the many instrument problem from a Bayesian perspective. Reduced form for years of education,

$$X_i = \pi_0 + Z_i'\pi_1 + \eta_i,$$

combined with a linear specification for log earnings,

$$Y_i = \alpha + \beta \cdot Z_i'\pi_1 + \varepsilon_i.$$

CI assume joint normality for the reduced form errors,

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}(0, \Omega).$$

This gives a likelihood function

$$\mathcal{L}(\beta, \alpha, \pi_0, \pi_1, \Omega|\text{data}).$$

The focus of the CI paper is on inference for β, and the sensitivity of such inferences to the choice of prior distribution in settings with large numbers of instruments.

A flat prior distribution may be a poor choice. One way to illustrate see this is that a flat prior on π_1 leads to a prior on the sum $\sum_{k=1}^{K} \pi_{ik}^2$ that puts most probability mass away from zero.

CI then show that the posterior distribution for β, under a flat prior distribution for π_1 provides an accurate approximation to the sampling distribution of the TSLS estimator.

As an alternative CI suggest a hierarchical prior distribution with

$$\pi_{1k} \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2).$$

In the Angrist-Krueger 1991 compulsory schooling example there is in fact a substantive reason to believe that σ_π^2 is small rather than the $\sigma_\pi^2 = \infty$ implicit in TSLS. If the π_{1k} represent the effect of the differences in the amount of required schooling, one would expect the magnitude of the π_{1k} to be less than the amount of variation in the compulsory schooling implying the standard deviation of the first stage coefficients should not be more than $\sqrt{1/12} = 0.289$.

Using the Angrist-Krueger data CI find that the posterior distribution for σ_π is concentrated close to zero, with the posterior mean and median equal to 0.119.

8. Example: Binary Response with Endogenous Regressors

Geweke, Gowrisankaran, and Town are interested in estimating the effect of hospital quality on mortality, taking into account possibly non-random selection of patients into hospitals. Patients can choose from 114 hospitals. Given their characteristics Z_i, latent mortality is

$$Y_i^* = \sum_{j=1}^{114} C_{ij}\beta_j + Z_i'\gamma + \epsilon_i,$$

where C_{ij} is an indicator for patient i going to hospital j. The focus is on the hospital effects on mortality, β_j. Realized mortality is

$$Y_i = 1\{Y_i^* \geq 0\}.$$

The concern is about selection into the hospitals, and the possibility that this is related to unobserved components of latent mortality GGT model latent the latent utility for patient i associated with hospital j as

$$C_{ij}^* = X_{ij}'\alpha + \eta_{ij},$$

where the X_{ij} are hospital-individual specific characteristics, including distance to hospital. Patient i then chooses hospital j if

$$C_{ij}^* \geq C_{ik}, \quad \text{for } k = 1, \ldots, 114.$$

The endogeneity is modelled through the potential correlation between η_{ij} and ϵ_i. Specifically, GGT asssume that as

$$\epsilon_i = \sum_{j=1}^{114} \eta_{ij} \cdot \delta_j + \zeta_i,$$

where the ζ_i is a standard normal random variable, independent of the other unobserved components.

GGT model the η_{ij} as standard normal, independent across hospitals and across individuals. This is a very strong assumption, implying essentially the independence of irrelevant alternatives property. One may wish to relax this by allowing for random coefficients on the hospital characteristics.

Given these modelling decisions GGT have a fully specified joint distribution of hospital choice and mortality given hospital and individual characteristics.

The log likelihood function is highly nonlinear, and it is unlikely it can be well approximated by a quadratic function.

GGT therefore use Bayesian methods, and in particular the Gibbs sampler to obtain draws from the posterior distribution of interest.

In their empirical analysis GGT find strong evidence for non-random selection. They find that higher quality hospitals attract sicker patients, to the extent that a model based on exogenous selection would have led to misleading conclusions on hospital quality.

9. Example: Discrete Choice Models with Unobserved Choice Characteristics

Athey and Imbens (2007, AI) study discrete choice models, allowing both for unobserved individual heterogeneity in taste parameters as well as for multiple unobserved choice characteristics.

In such settings the likelihood function is multi-modal, and frequentist approximations based on quadratic approximations to the log likelihood function around the maximum likelihood estimator are unlikely to be accurate.

The specific model AI use assumes that the utility for individual i in market t for choice j is

$$U_{ijt} = X'_{it}\beta_i + \xi'_j\gamma_i + \epsilon_{ijt},$$

where X_{it} are market-specific observed choice characteristics, ξ_j is a vector of unobserved choice characteristics, and ϵ_{ijt} is an idiosyncratic error term, with a normal distribution centered at zero, and with the variance normalized to unity.

The individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega),$$

with the Z_i observed individual characteristics.

AI specify a prior distribution on the common parameters, Δ, and Ω, and on the values of the unobserved choice characteristics ξ_j.

Using mcmc with the unobserved utilities as unobserved random variables makes sampling from the posterior distribution conceptually straightforward even in cases with more than one unobserved choice characteristic.

In contrast, earlier studies using multiple unobserved choice characteristics (Elrod and Keane, 1995; Goettler and Shachar, 2001), using frequentist methods, faced much heavier computational burdens.

What's New in Econometrics?
Lecture 8
Cluster and Stratified Sampling

Jeff Wooldridge
NBER Summer Institute, 2007

1. The Linear Model with Cluster Effects

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

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

4. Nonlinear Models

1. **The Linear Model with Cluster Effects**.

- 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.
- The linear model with an additive error is

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma + v_{gm} \tag{1}$$

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

- Key questions: Are we primarily interested in β or γ? Does v_{gm} contain a common group effect, as in

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

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error? Are the regressors (x_g, z_{gm}) appropriately exogenous? How big are the group sizes (M_g) and number of groups (G)?

- Two kinds of sampling schemes. First, 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. This is like the panel data setup we have covered. 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 or 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.

Large Group Asymptotics

- The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed. How should one use these methods? If

$$E(v_{gm}|x_g, z_{gm}) = 0 \tag{3}$$

then pooled OLS estimator of 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. In panel data case, (3)

does allow for non-strictly exogenous covariates, but only if there is no unobserved effect.

- Robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|x_g, z_{gm})$, or both. 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

$$\widehat{Avar}(\hat{\lambda}_{POLS}) = \left(\sum_{g=1}^{G} W_g' W_g \right)^{-1} \left(\sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g \right) \cdot \left(\sum_{g=1}^{G} W_g' W_g \right)^{-1} \quad (4)$$

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.

- If we strengthen the exogeneity assumption to

$$E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \quad (5)$$

where Z_g is the $M_g \times L$ matrix of unit-specific covariates, then we can use GLS. This is about the strongest assumption we can make. As discussed in the linear panel data notes, the 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,

$$Var(v_g) = \sigma_c^2 j'_{M_g} j_{M_g} + \sigma_u^2 I_{M_g}, \quad (6)$$

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. Plus, the usual assumptions include the "system homoskedasticity" assumption,

$$Var(v_g|x_g, Z_g) = Var(v_g). \quad (7)$$

- The random effects estimator $\hat{\lambda}_{RE}$ 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 is often overlooked: one can, and in many cases should, make inference completely robust to an unknown form of $Var(v_g|x_g, Z_g)$, whether we have a true cluster sample or panel data.
- Cluster sample example: random coefficient model,

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma_g + v_{gm}. \quad (8)$$

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.

- 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:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g, \qquad (9)$$

and this equation is estimated by pooled OLS. (Of course, the x_g get swept away by the within-group demeaning.) Often important to allow $Var(u_g|Z_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. Certainly should

for panel data (serial correlation), but also for cluster sampling. In linear panel data notes, we saw that FE can consistently estimate the average effect in the random coefficient case. But $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term. A fully robust variance matrix estimator is

$$\widehat{Avar}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g \right)^{-1} \left(\sum_{g=1}^{G} \ddot{Z}_g' \hat{u}_g \hat{u}_g' \ddot{Z}_g \right) \cdot \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g \right)^{-1}, \quad (10)$$

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.

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? 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. Generally, we 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. 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 (10) – 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).

- 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. But allowing time effects causes problems in this case. Really want "large" N and T to allow for a full set of time and unit-specific effects.

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

- When G is small and each M_g is large, thinking of sampling from different strata in a population, or even different populations, makes more sense. Alternatively, we might think that the clusters were randomly drawn from a large population, but only a small number were drawn. Either way, except for the relative dimensions of G and M_g, the resulting

data set is essentially indistinguishable from a data set obtained by sampling 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 (so it seems more like the second sampling experiment). Simplest case: a single regressor that varies only by group:

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

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

Notice how (12) is written as a model with 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.

They use this formulation to highlight the problems of applying standard inference to (11), leaving c_g as part of the error term, $v_{gm} = c_g + u_{gm}$.

- We know that standard pooled OLS inference can be badly biased because it ignores the cluster correlation. And Hansen's results do not apply. (We cannot use fixed effects here.)
- The DL solution is to study the OLS estimate in the regression "between" regresssion

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G, \tag{13}$$

which is identical to pooled OLS when the group sizes are the same. 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.

- If we add some strong assumptions, there is an exact solution to the inference problem. In addition

to assuming $M_g = M$ for all g, assume $c_g|x_g$ ~Normal$(0, \sigma_c^2)$ and assume $u_{gm}|x_g, c_g \sim$ Normal$(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim$ Normal$(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G \qquad (14)$$

satisfies the classical linear model assumptions. 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 $t_{M_1+M_2+...+M_G-2}$ distribution – which is what we would be doing if we use the usual pooled OLS statistics.
- Using (14) is *not* the same as using cluster-robust

standard errors for pooled OLS. Those are not even justified and, besides, we would use the wrong df in the t distribution.

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

$$\bar{y}_g = \alpha + x_g\beta + \bar{z}_g\gamma + \bar{v}_g, g = 1, \ldots, G, \qquad (15)$$

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. Regressions like (15) are reasonably common, at

least as a check on results using disaggregated data, but usually with larger G then just a few.

- If $G = 2$, should we give up? Suppose x_g is binary, indicating treatment and control. The DL estimate of β is the usual one: $\hat{\beta} = \bar{y}_1 - \bar{y}_0$. But in the DL setting, we cannot do inference (there are zero df). So, the DL setting rules out the standard comparison of means. It also rules out the typical setup for difference-in-differences, where there would be four groups, for the same reason.
- 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.
- Even when we can apply the approach, should we? Suppose there $G = 4$ 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

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

With $\hat{\beta}$ written as in (16), it is clearly it is approximately normal (for almost any underlying population distribution) provided the group sizes M_g are moderate. The DL approach would base inference on a t_2 distribution. 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 $\mu_3 = \mu_4$.

- Equation (16) hints at a different way to view the small G, large M_g setup. We estimated two parameters, α and β, given four moments that we can estimate with the data. The OLS estimates 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 (16) 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 probably generally 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

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}. \tag{17}$$

where we assume random sampling within group and independent sampling across groups. Generally, the OLS estimates withing group are $\sqrt{M_g}$ -asymptotically normal. The presence of x_g can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in (2.8). In particular,

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G, \tag{18}$$

where we now 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. In the first stage, we obtain the $\hat{\delta}_g$, either by group-specific regressions or pooling to impose some common slope elements in γ_g. 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

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

The asymptotics are as each group size gets large, and $\hat{\theta}$ has an asymptotic normal distribution; its estimated asymptotic variance is $(X'\hat{V}^{-1}X)^{-1}$. When separate regressions are used, the $\hat{\delta}_g$ are independent, and $\hat{V}$ is diagonal.

- Can test the overidentification restrictions. If reject, can go back to the DL approach (or find more elements of x_g). With large group sizes, can justify analyzing

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

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

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

If we have a reasonably large G in addition to large M_g, we have more flexibility. In addition to ignoring the estimation error in $\hat{\delta}_g$ (because of large M_g), we can also drop the normality assumption in c_g (because, as G gets large, we can apply the central limit theorem). But, of course, we are still assuming that the deviations, c_g, in $\delta_g = \alpha + x_g\beta + c_g$, are at least uncorrelated with x_g. We can apply IV methods in this setting, though, if we have suitable instruments.

4. Nonlinear Models

- Many of the issues for nonlinear models are the same as for linear models. The biggest difference is that, in many cases, standard approaches require distributional assumptions about the unobserved

group effects. In addition, it is more difficult in nonlinear models to allow for group effects correlated with covariates, especially when group sizes differ.

Large Group Asymptotics

We can illustrate many issues using an unobserved effects probit model. Let y_{gm} be a binary response, with x_g and z_{gm}, $m = 1, \ldots, M_g, g = 1, \ldots, G$ defined as in Section 1. Assume that

$$y_{gm} = 1[\alpha + x_g\beta + z_{gm}\gamma + c_g + u_{gm} \geq 0] \quad (21)$$

$$u_{gm}|x_g, Z_g, c_g \sim Normal(0, 1) \quad (22)$$

(where $1[\cdot]$ is the indicator function). Then

$$P(y_{gm} = 1|x_g, z_{gm}, c_g) = \Phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \quad (23)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf).

- We already discussed the issue of quantities of interest, including parameters and average partial effects.
- For estimation, if we assume c_g is independent of (x_g, Z_g) with a Normal$(0, \sigma_g^2)$ distribution, then pooled probit consistently estimates the scaled coefficients (multiplied by $(1 + \sigma_c^2)^{-1/2}$). The pooled or partial maximum likelihood estimator is sometimes called a *pseudo maximum likelihood* estimator
- If we add the conditional independence assumption that $\{u_{g1}, \ldots, u_{g,M_g}\}$ are independent conditional on (x_g, Z_g, c_g) then we can use random effects probit, albeit in an unbalanced case. As we saw before, all parameters are identified.
- A challenging task, and one that appears not to

have gotten much attention for true cluster samples, is allowing correlation between the unobserved heterogeneity, c_g, and the covariates that vary within group, z_{gm}. For linear models, we know that the fixed effects estimator allows arbitrary correlation, and does not restrict the within-cluster dependence of $\{u_{g1}, \ldots, u_{g,M_g}\}$. Unfortunately, allowing correlation between c_g and $(z_{g1}, z_{g2}, \ldots, z_{gM_g})$. Even if we assume normality and exchangeability in the mean, we must at least allow for difference variances:

$$c_g|(z_{g1}, \ldots, z_{g,M_g}) \sim \text{Normal}(\eta + \bar{z}_g\xi, \sigma^2_{a,M_g}), \qquad (24)$$

where σ^2_{a,M_g} denotes a different variance for each group size, M_g. Then the marginal distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma^2_{a,M_g})^{1/2}]. \qquad (25)$$

Any estimation must account for the different variances for different group sizes. With very large G and little variation in M_g, we might just use the unrestricted estimates $(\hat{\eta}_{M_g}, \hat{\xi}_{M_g}, \hat{\gamma}_{M_g})$, estimate the APEs for each group size, and then average these across group size. But more work needs to be done to see if such an approach loses too much in terms of efficiency.

- The methods of Altonji and Matzkin (2005) can be applied to allow more flexible relationships between c_g and $\bar{z}_g$, say, or other functions of $\{z_{g1}, \ldots, z_{g,M_g}\}$
- The logit conditional MLE applies to cluster samples without change, so we can estimate parameters without restricting $D(c_g | z_{g1}, \ldots, z_{gM_g})$.

A Small Number of Groups and Large Group

Sizes

- Unlike in the linear case, for nonlinear models exact inference is unavailable even under the strongest set of assumptions. But approximate inference is if the group sizes M_g are reasonably large
- With small G and random sampling of $\{(y_{gm}, z_{gm}) : m = 1, \ldots, M_g\}$ write

$$P(y_{gm} = 1|z_{gm}) = \Phi(\delta_g + z_{gm}\gamma_g) \tag{26}$$

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G. \tag{27}$$

Using a minimum distance approach, in a first step we estimate a series of G probits (or pool across g to impose common slopes), obtain the group "fixed effects" $\hat{\delta}_g, g = 1, \ldots, G$. Then, we impose the restrictions in (26) using linear MD estimation –

just as before. Now, the asymptotic variances $\widehat{Avar}(\hat{\delta}_g)$ come from the probits.

- The DL approach also applies with large M_g but we again must assume $\delta_g = \alpha + x_g\beta + c_g$ where c_g is independent of x_g and homoskedastic normal. As in the linear case, we just use classical linear model inference in the equation $\hat{\delta}_g = \alpha + x_g\beta + c_g$, provide $G > K + 1$.
- The same holds for virtually any nonlinear model with an index structure: the second step is linear regression.

Large G and Large M_g

- As in the linear case, more flexibility is afforded if G is somewhat large along with large M_g because we can relax the normality assumption in c_g in analyzing the regression $\hat{\delta}_g$ on $1, x_g, g = 1, \ldots, G$.

- A version of the method proposed by Berry, Levinsohn, and Pakes (1995) for estimating structural models using both individual-level and product-level data, or market-level data, or both can be treated in the large G, large M_g framework, where g indexes good or market and m indexes individuals within a market. Suppose there is a single good across many markets (so G is large) and we have many individuals within each market (the M_g are large). The main difference with what we have done up until now is that BLP must allow correlation between x_g (particularly, price) and the unobserved product attributes in market g, c_g. So, the second step involves instrumental variables estimation of the equation $\hat{\delta}_g = \alpha + x_g\beta + c_g$. If the M_g are large enough to ignore estimation error in

the $\hat{\delta}_g$, then the second step can be analyzed just like a standard cross section IV estimation. (BLP actually derive an asymptotic variance that accounts for estimation error in $\hat{\delta}_g$, along with the uncertainty in c_g, and simulation error – which comes from difficult computational problems in the first stage.)

"What's New in Econometrics"

Lecture 9

Partial Identification

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction
2. Example I: Missing Data
3. Example II: Returns to Schooling
4. Example III: Initial Conditions Problems in Panel Data
5. Example IV: Auction Data
6. Example V: Entry Models
7. Estimation and Inference

1. Introduction

Traditionally in constructing statistical or econometric models researchers look for models that are *(point-)identified*: given a large (infinite) data set, one can infer without uncertainty what the values are of the objects of interest.

It would appear that a model where we cannot learn the parameter values even in infinitely large samples would not be very useful.

However, it turns out that even in cases where we cannot learn the value of the estimand *exactly* in large samples, in many cases we can still learn a fair amount, even in finite samples. A research agenda initiated by Manski has taken this perspective.

Here we discuss a number of examples to show how this approach can lead to interesting answers in settings where previously were viewed as intractable.

We also discuss some results on inference.

1. Are we interested in confidence sets for parameters or for identified sets?

2. Concern about uniformity of inferences (confidence cant be better in partially identified case than in point-identified case).

2. I: Missing Data

If $D_i = 1$, we observe Y_i, and if $D_i = 0$ we do not observe Y_i. We always observe the missing data indicator D_i. We assume the quantity of interest is the population mean $\theta = \mathbb{E}[Y_i]$.

In large samples we can learn $p = \mathbb{E}[D_i]$ and $\mu_1 = \mathbb{E}[Y_i|D_i = 1]$, but nothing about $\mu_0 = \mathbb{E}[Y_i|D_i = 0]$. We can write:

$$\theta = p \cdot \mu_1 + (1 - p) \cdot \mu_0.$$

Since even in large samples we learn nothing about μ_0, it follows that without additional information there is no limit on the range of possible values for θ.

Even if p is very close to 1, the small probability that $D_i = 0$ combined with the possibility that μ_0 is very large or very small allows for a wide range of values for θ.

Now suppose we know that the variable of interest is binary: $Y_i \in \{0, 1\}$. Then natural (not data-informed) lower and upper bounds for μ_0 are 0 and 1 respectively. This implies bounds on θ:

$$\theta \in [\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}] = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)] \,.$$

These bounds are *sharp*, in the sense that without additional information we can not improve on them.

Formally, for all values θ in $[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]$, we can find a joint distribution of (Y_i, W_i) that is consistent with the joint distribution of the observed data and with θ.

We can also obtain informative bounds if we modify the object of interest a little bit.

Suppose we are interested in the median of Y_i, $\theta_{0.5} = \text{med}(Y_i)$.

Define $q_\tau(Y_i)$ to be the τ quantile of the conditional distribution of Y_i given $D_i = 1$. Then the median cannot be larger than $q_{1/(2p)}(Y_i)$ because even if all the missing values were large, we know that at least $p \cdot (1/(2p)) = 1/2$ of the units have a value less than or equal to $q_{1/(2p)}(Y_i)$.

Then, if $p > 1/2$, we can infer that the median must satisfy

$$\theta_{0.5} \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(2p-1)/(2p)}(Y_i), q_{1/(2p)}(Y_i)\right],$$

and we end up with a well defined, and, depending on the data, more or less informative identified interval for the median.

If fewer than 50% of the values are observed, or $p < 1/2$, then we cannot learn anything about the median of Y_i without additional information (for example, a bound on the values of Y_i), and the interval is $(-\infty, \infty)$.

More generally, we can obtain bounds on the τ quantile of the distribution of Y_i, equal to

$$\theta_\tau \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(\tau-(1-p))/p}(Y_i|D_i = 1), q_{\tau/p}(Y_i|D_i = 1)\right].$$

which is bounded if the probability of Y_i being missing is less than $\min(\tau, 1-\tau)$.

3. Example II: Returns to Schooling

Manski-Pepper are interested in estimating returns to schooling. They start with an individual level response function $Y_i(w)$.

$$\Delta(s,t) = \mathbb{E}[Y_i(t) - Y_i(s)],$$

is the difference in average outcomes (log earnings) given t rather than s years of schooling. Values of $\Delta(s,t)$ are the object of interest.

W_i is the actual years of school, and $Y_i = Y_i(W_i)$ be the actual log earnings.

If one makes an unconfoundedness/exogeneity assumption that

$$Y_i(w) \perp\!\!\!\perp W_i \mid X_i,$$

for some set of covariates, one can estimate $\Delta(s,t)$ consistently given some support conditions. MP relax this assumption.

Alternative Assumptions considered by MP

Increasing education does not lower earnings:

Assumption 1 (Monotone Treatment Response)
If $w' \geq w$, *then* $Y_i(w') \geq Y_i(w)$.

On average, individuals who choose higher levels of education would have higher earnings at each level of education than individuals who choose lower levels of education.

Assumption 2 (Monotone Treatment Selection)
If $w'' \geq w'$, *then for all* w, $\mathbb{E}[Y_i(w)|W_i = w''] \geq \mathbb{E}[Y_i(w)|W_i = w']$.

Under these two assumptions, bound on $\mathbb{E}[Y_i(w)]$ and $\Delta(s,t)$:

$$\mathbb{E}[Y_i|W_i = w] \cdot \mathsf{Pr}(W_i \geq w) + \sum_{v<w} \mathbb{E}[Y_i|W_i = v] \cdot \mathsf{Pr}(W_i = v)$$

$$\leq \mathbb{E}[Y_i(w)] \leq$$

$$\mathbb{E}[Y_i|W_i = w] \cdot \mathsf{Pr}(W_i \leq w) + \sum_{v>w} \mathbb{E}[Y_i|W_i = v] \cdot \mathsf{Pr}(W_i = v).$$

Using NLS data MP estimate the upper bound on the the returns to four years of college, $\Delta(12, 16)$ to be 0.397.

Translated into average yearly returns this gives us 0.099, which is in fact lower than some estimates that have been reported in the literature.

This analysis suggests that the upper bound is in this case reasonably informative, given a remarkably weaker set of assumptions.

4. Example III: Initial Conditions Problems in Panel Data (Honoré and Tamer)

$$Y_{it} = 1\{X_{it}'\beta + Y_{it-1} \cdot \gamma + \alpha_i + \epsilon_{it} \geq 0\},$$

with the ϵ_{it} independent $\mathcal{N}(0,1)$ over time and individuals. Focus on γ.

Suppose we also postulate a parametric model for the random effects α_i:

$$\alpha | X_{i1}, \ldots, X_{iT} \sim G(\alpha|\theta)$$

Then the model is almost complete.

All that is missing is:

$$p(Y_{i1}|\alpha_i, X_{i1}, \ldots, X_{iT}).$$

HT assume a discrete distribution for α, with a finite and known set of support points. They fix the support to be $-3, -2.8, \ldots, 2.8, 3$, with unknown probabilities.

In the case with $T = 3$ they find that the range of values for γ consistent with the data generating process (the identified set) is very narrow.

If γ is in fact equal to zero, the width of the set is zero. If the true value is $\gamma = 1$, then the width of the interval is approximately 0.1. (It is largest for γ close to, but not equal to, -1.) See Figure 1, taken from HT.

The HT analysis shows nicely the power of the partial identification approach: A problem that had been viewed as essentially intractable, with many non-identification results, was shown to admit potentially precise inferences. Point identification is not a big issue here.

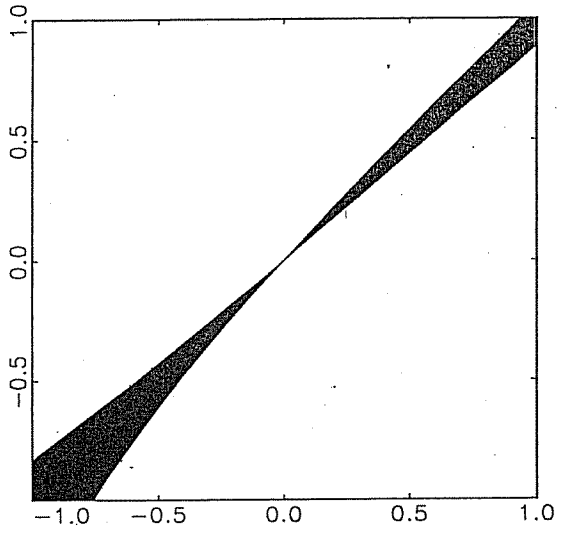


FIGURE 1.—Identified region for γ as a function of its true value.

5. Example IV: Auction Data

Haile and Tamer study English or oral ascending bid auctions. In such auctions bidders offer increasingly higher prices until only one bidder remains. HT focus on a symmetric independent private values model. In auction t, bidder i has a value ν_{it}, drawn independently from the value for bidder j, with cdf $F_\nu(v)$

HT are interested in the value distribution $F_\nu(v)$. This is assumed to be the same in each auction (after adjusting for observable auction characteristics).

One can imagine observing exactly when each bidder leaves the auction, thus directly observing their valuations. This is not what is typically observed. For each bidder we do not know at any point in time whether they are still participating unless they subsequently make a higher bid.

Haile-Tamer Assumptions

Assumption 3 *No bidder ever bids more than their valuation*

Assumption 4 *No bidder will walk away and let another bidder win the auction if the winning bid is lower than their own valuation*

Upper Bound on Value Distribution

Let the highest bid for participant i in auction t be b_{it}. We ignore variation in number of bidders per auction, and presence of covariates.

Let $F_b(b) = \Pr(b_{it} \leq b)$ be the distribution function of the bids (ignoring variation in the number of bidders by auction). This distribution can be estimated because the bids are observed.

Because no bidder ever bids more than their value, it follows that $b_{it} \leq \nu_{it}$. Hence, without additional assumptions,

$$F_\nu(v) \leq F_b(v), \quad \text{for all } v.$$

Lower Bound on Value Distribution

The second highest of the values among the n participants in auction t must be less than or equal to the winning bid. This follows from the assumption that no participant will let someone else win with a bid below their valuation.

Let $F_{\nu,m:n}(v)$ denote the mth order statistic in a random sample of size n from the value distribution, and let $F_{B,n:n}(b)$ denote the distribution of the winning bid in auctions with n participants. Then

$$F_{B,n:n}(v) \leq F_{\nu,n-1:n}(v).$$

The distribution of the any order statistic is monotonically related to the distribution of the parent distribution, and so a lower bound on $F_{\nu,n-1:n}(v)$ implies a lower bound on $F_{\nu}(v)$.

6. Example V: Entry Models (Cilberto & Tamer)

Suppose two firms, A and B, contest a set of markets. In market m, $m = 1, \ldots, M$, the profits for firms A and B are

$$\pi_{Am} = \alpha_A + \delta_A \cdot d_{Bm} + \varepsilon_{Am},\ \pi_{Bm} = \alpha_B + \delta_B \cdot d_{Am} + \varepsilon_{Bm}.$$

where $d_{Fm} = 1$ if firm F is present in market m, for $F \in \{A, B\}$, and zero otherwise.

Decisions assuming complete information satisfy Nash equilibrium condition

$$d_{Am} = 1\{\pi_{Am} \geq 0\}, \quad d_{Bm} = 1\{\pi_{Bm} \geq 0\}.$$

Incomplete Model

For pairs of values $(\varepsilon_{Am}, \varepsilon_{Bm})$ such that

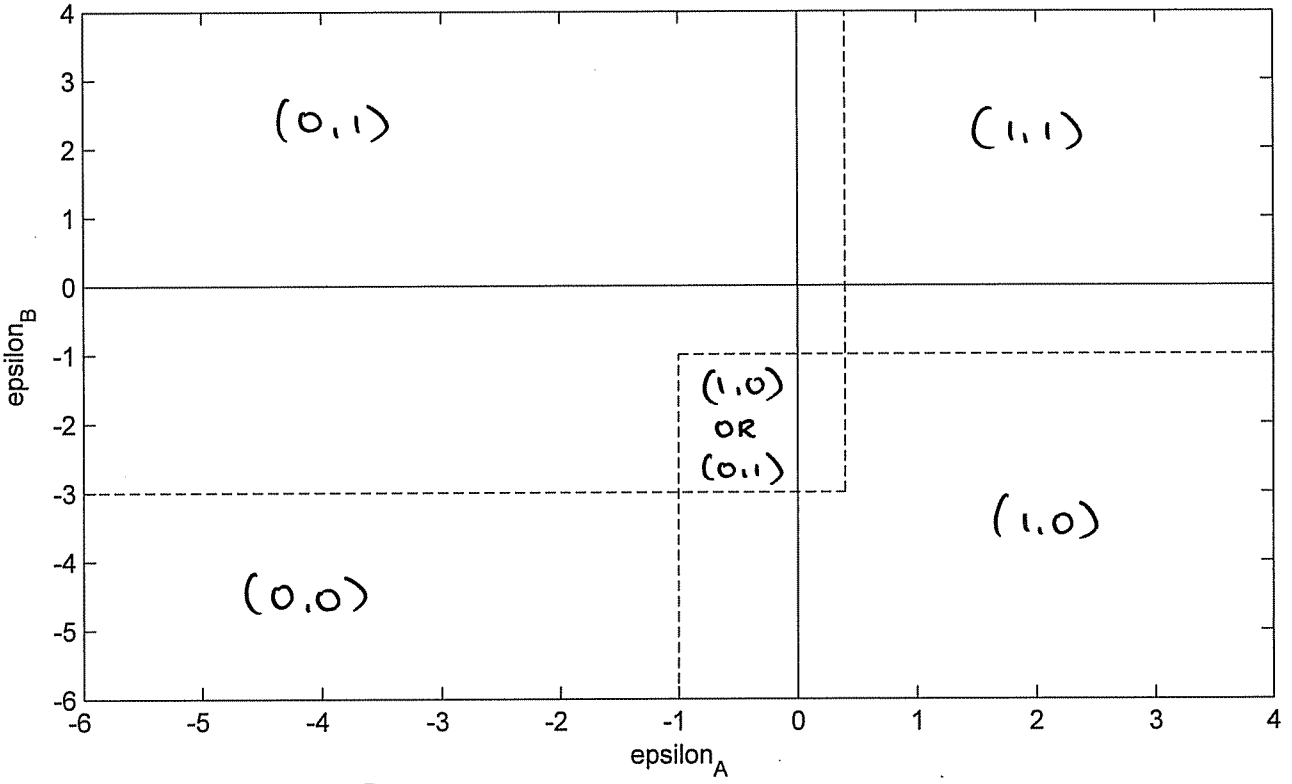
$$-\alpha_A < \varepsilon_A \leq -\alpha_A - \delta_A, \quad -\alpha_B < \varepsilon_B \leq -\alpha_B - \delta_B,$$

both $(d_A, d_B) = (0,1)$ and $(d_A, d_B) = (1,0)$ satisfy the profit maximization condition.

In the terminology of this literature, the model is *incomplete*. It does not specify the outcomes given the inputs. Missing is an equilibrium selection mechanism, which is typically difficult to justify.

Figure 1, adapted from CM, shows the different regions in the $(\varepsilon_{Am}, \varepsilon_{Bm})$ space.

Figure 1 (d_A, d_B)



$\alpha_A = 1$ $\delta_A = -1.4$

$\alpha_B = 3$ $\delta_B = -2$

Implication: Inequality Conditions

The implication of this is that the probability of the outcome $(d_{Am}, d_{Bm}) = (0,1)$ cannot be written as a function of the parameters of the model, $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, even given distributional assumptions on $(\varepsilon_{Am}, \varepsilon_{Bm})$.

Instead the model implies a lower and upper bound on this probability:

$$H_{L,01}(\theta) \leq \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \leq H_{U,01}(\theta).$$

Thus in general we can write the information about the parameters in large samples as

$$\begin{pmatrix} H_{L,00}(\theta) \\ H_{L,01}(\theta) \\ H_{L,10}(\theta) \\ H_{L,11}(\theta) \end{pmatrix} \leq \begin{pmatrix} \Pr\left((d_{Am}, d_{Bm}) = (0,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,1)\right) \end{pmatrix} \leq \begin{pmatrix} H_{U,00}(\theta) \\ H_{U,01}(\theta) \\ H_{U,11}(\theta) \\ H_{U,11}(\theta) \end{pmatrix}.$$

7.A Estimation

Chernozhukov, Hong, and Tamer study Generalized Inequality Restriction (GIR) setting:

$$\mathbb{E}[\psi(Z, \theta)] \geq 0,$$

where $\psi(z, \theta)$ is known. Fits CT entry example

Define for a vector x the vector $(x)_+$ to be the component-wise non-negative part, and $(x)_-$ to be the component-wise non-positive part, so that for all x, $x = (x)_- + (x)_+$.

For a given $M \times M$ non-negative definite weight matrix W, CHT consider the population objective function

$$Q(\theta) = \mathbb{E}[\psi(Z, \theta)]'_{-} W \mathbb{E}[\psi(Z, \theta)]_{-}.$$

For all $\theta \in \Theta_I$, we have $Q(\theta) = 0$, and for $\theta \notin \Theta_I$, we have $Q(\theta) > 0$

The sample equivalent to this population objective function is

$$Q_N(\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)'_{-} W \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_{-}.$$

We cannot simply estimate the identified set as

$$\tilde{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) = 0\}\,,$$

The reason is that even for θ in the identified set $Q_N(\theta)$ may be positive with high probability, and $\tilde{\Theta}_I$ can be empty when Θ_I is not, even in large samples.

A simple way to see that is to consider the standard GMM case with equalities and over-identification. If $\mathbb{E}[\psi(Z,\theta)] = 0$, the objective function will not be zero in finite samples in the case with over-identification.

This is the reason CHT suggest estimating the set Θ_I as

$$\hat{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) \leq a_N\}\,,$$

where $a_N \to 0$ at the appropriate rate.

7.B Inference

Fast growing literature, Beresteanu and Molinari (2006), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2006), Imbens and Manski (2004), Rosen (2006), and Romano and Shaikh (2007ab).

First issue: do we want a confidence set that includes each element of the identified set with fixed probability, or the entire identified set with that probability. First

$$\inf_{\theta \in [\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]} \Pr\left(\theta \in \mathsf{CI}_{\alpha}^{\theta}\right) \geq \alpha.$$

Second

$$\Pr\left([\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}] \subset \mathsf{CI}_{\alpha}^{[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]}\right) \geq \alpha.$$

The second requirement is stronger than the first, and so generally $\mathsf{CI}_{\alpha}^{\theta} \subset \mathsf{CI}_{\alpha}^{[\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}]}$.

7.B.I Well behaved Estimators for Bounds

Missing data example, (p, prob of missing data, known). Identified set:

$$\Theta_I = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)].$$

Standard interval for μ_1:

$$\text{CI}_\alpha^{\mu_1} = \left[\overline{Y} - 1.96 \cdot \sigma/\sqrt{N}_1, \overline{Y} + 1.96 \cdot \sigma/\sqrt{N}_1\right].$$

Three ways to construct 95% confidence intervals for θ.

$$\mathrm{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This is conservative. For each θ in the interior of Θ_I, the cov rate is 1. For $\theta \in \{\theta_{\mathsf{LB}}, \theta_{\mathsf{UB}}\}$, if $p < 1$, the cov rate is 0.975.

$$\mathrm{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.645 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.645 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This has the problem that if $p = 1$ (when θ is point-identified), the coverage is only 0.90. Imbens and Manski (2004) suggest modifying the confidence interval to

$$\mathrm{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - C_N \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + C_N \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right],$$

where the critical value C_N satisfies

$$\Phi\left(C_N + \sqrt{N} \cdot \frac{1-p}{\sigma/\sqrt{p}}\right) - \Phi\left(-C_N\right) = 0.95$$

This confidence interval has asymptotic coverage 0.95, uniformly over p, for $p \in [p_0, 1]$.

7.B.II Irregular Estimators for Bounds

Simple example of Generalized Inequality Restrictions (GIR) set up.

$$\mathbb{E}[X] \geq \theta, \quad \text{and} \quad \mathbb{E}[Y] \geq \theta.$$

The parameter space is $\Theta = [0, \infty)$. Let $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$. We have a random sample of size N of the pairs (X, Y). The identified set is

$$\Theta_I = [0, \min(\mu_X, \mu_Y)].$$

A naive 95% confidence interval would be

$$C_\alpha^\theta = [0, \min(\overline{X}, \overline{Y}) + 1.645 \cdot \sigma/N].$$

This confidence interval essentially ignores the moment inequality that is not binding in the sample. It has pointwise asymptotic 95% coverage for all values of μ_X, μ_Y, as long as $\min(\mu_X, \mu_Y) > 0$, and $\mu_X \neq \mu_Y$.

The first condition ($\min(\mu_X, \mu_Y) > 0$) is the same as the condition in the Imbens-Manski example. It can be dealt with in the same way by adjusting the critical value slightly based on an initial estimate of the width of the identified set.

The naive confidence interval essentially assumes that the researcher knows which moment conditions are binding. This is true in large samples, unless there is a tie.

However, in finite samples ignoring uncertainty regarding the set of binding moment inequalities may lead to a poor approximation, especially if there are many inequalities. One possibility is to construct conservative confidence intervals (e.g., Pakes, Porter, Ho, and Ishii, 2007). However, such intervals can be unnecessarily conservative if there are moment inequalities that are far from binding.

One would like construct confidence intervals that asymptotically ignore irrelevant inequalities, and at the same time are valid uniformly over the parameter space. Subsampling (but not bootstrapping) appears to work theoretically. See Romano and Shaikh (2007a), and Andrews and Guggenberger (2007). Little is known about finite sample properties in realistic settings.

What's New in Econometrics?
Lecture 10
Difference-in-Differences Estimation

Jeff Wooldridge
NBER Summer Institute, 2007

1. Review of the Basic Methodology

2. How Should We View Uncertainty in DD Settings?

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

4. Individual-Level Panel Data

5. Semiparametric and Nonparametric Approaches

6. Synthetic Control Methods for Comparative Case Studies

1. Review of the Basic Methodology

- The 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. In the case where the same units within a group are observed in each time period (panel data), 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.

- With repeated cross sections, let A be the control group and B the treatment group. Write

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \qquad (1)$$

where y is the outcome of interest. The dummy dB captures possible differences between the treatment and control groups prior to the policy change. The dummy $d2$ captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest is δ_1.

- The difference-in-differences estimate is

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

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.

- More convincing analysis sometimes available by refining 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, might use the over-65 population from another state as an additional control. Let dE be a dummy equal to one for someone over 65.

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (3)$$

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

$$\hat{\delta}_3 = (\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1}) \quad (4)$$

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.
- Can use multiple time periods and groups.

2. **How Should We View Uncertainty in DD Settings**?

- Standard 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, different approaches have been suggest that 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 (2007) argue for additional sources of uncertainty.
- In fact, for the most part, the additional uncertainty is assumed to swamp the sampling error in estimating group/time period means. (See DL

approach in 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.
- 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, the DL approach does not allow inference.
- 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 the standard

DD setting: a before and after period, where the policy change was to raise the cap on covered earnings; control group is low earners. Using Kentucky and a total sample size of 5,626, the DD estimate of the policy change is about 19.2% (longer time on workers' compensation) with $t = 2.76$. Using Michigan, with a total sample size of 1,524, the DD estimate is 19.1% with $t = 1.22$. (Adding controls does not help reduce the standard error.) There seems to be plenty of uncertainty in the estimate even with a pretty large sample size. Should we conclude that we really have no usable data for inference?

3. 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 DD case, add a full set of time dummies to the equation. This assumes the policy has the same effect in every year; 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 Meyer (1995) for applications.

- 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

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \qquad (5)$$
$$i = 1, \ldots, M_{gt},$$

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 $\boldsymbol{\beta}$.

- As in cluster sample cases, can write

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \qquad (6)$$

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

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \qquad (7)$$

We can think of (7) as a regression model at the group/time period level.

- As discussed by BDM, a common way to estimate and perform inference in (5) 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 (2007b) allow serial correlation in $\{v_{gt} : t = 1, 2, \ldots, T\}$ but assume independence across g.
- If we view (7) 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 (7). The MD approach from cluster sample notes effectively drops v_{gt} from (7) 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 MD estimator uses only sampling variation in the $\hat{\delta}_{gt}$. Here, we proceed ignoring estimation error, and so act as if (7) is, for $t = 1, \ldots, T, g = 1, \ldots, G$,

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt} \qquad (8)$$

We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (8) 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 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 (8) is inefficient when v_{gt} is serially uncorrelated, proposes feasible GLS. When T is small, estimating

the parameters in $\Omega_g = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. Estimates based on the FE residuals, $\hat{v}_{gt}$, disappear as $T \to \infty$, but can be substantial. In AR(1) case, $\hat{\rho}$ comes from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1}, \; t = 2, \ldots, T, g = 1, \ldots, G. \tag{9}$$

- One way to account for bias in $\hat{\rho}$: use fully robust inference. But, as Hansen (2007b) shows, this can be very inefficient relative to his suggestion to bias-adjust the estimator $\hat{\rho}$ and then use the bias-adjusted estimator in feasible GLS. (Hansen covers the general $AR(p)$ model.)
- 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 $T \to \infty$. Nothing new that GLS relies on strict exogeneity in serial correlation cases. In intervention analyis, might be concerned if the

policies can switch on and off over time.

- 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 (2003). Works pretty well with $G = 50$ and $T = 10$, but get substantial size distortions for $G = 50$ and $T = 20$.
- If the M_{gt} are not large, might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. Can instead aggregate the equations over individuals, giving

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \qquad (10)$$
$$t = 1,..,T, g = 1,\ldots,G.$$

Can estimate this by FE and use fully robust inference 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 (10). 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}.

4. Individual-Level Panel Data

- Let w_{it} be a binary indicator, which is unity if unit i participates in the program at time t. Consider

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it}, \; t = 1, 2, \qquad (11)$$

where $d2_t = 1$ if $t = 2$ and zero otherwise, c_i is an observed effect, and u_{it} are the idiosyncratic errors. The coefficient τ is the treatment effect. A simple estimation procedure is to first difference to remove

c_i :

$$(y_{i2} - y_{i1}) = \eta + \tau(w_{i2} - w_{i1}) + (u_{i2} - u_{i1}) \qquad (12)$$

or

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \qquad (13)$$

If $E(\Delta w_i \Delta u_i) = 0$, that is, the change in treatment status is uncorrelated with changes in the idiosyncratic errors, then OLS applied to (13) is consistent.

- If $w_{i1} = 0$ for all i, the OLS estimate is

$$\hat{\tau} = \Delta\bar{y}_{treat} - \Delta\bar{y}_{control}, \qquad (14)$$

which is a DD estimate except that we different the means of the same units over time.

- With many time periods and arbitrary treatment patterns, we can use

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \; t = 1, \ldots, T, \qquad (15)$$

which accounts for aggregate time effects and allows for controls, $\mathbf{x}_{it}$. Estimation by FE or FD to remove c_i is standard, provided the policy indicator, w_{it}, is strictly exogenous: correlation beween w_{it} and u_{ir} for any t and r causes inconsistency in both estimators (with FE having some advantages for larger T if u_{it} is weakly dependent)

- What if designation is correlated with unit-specific trends? "Correlated random trend" model:

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it} \qquad (16)$$

where g_i is the trend for unit i. A general analysis allows arbitrary corrrelation between (c_i, g_i) and w_{it}, which requires at least $T \geq 3$. If we first difference, we get, for $t = 2, \ldots, T$,

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}. \tag{17}$$

Can difference again or estimate (17) by FE.

- Can derive standard panel data approaches using the counterfactural framework from the treatment effects literature. For each (i, t), let $y_{it}(1)$ and $y_{it}(0)$ denote the counterfactual outcomes, and assume there are no covariates. Unconfoundedness, conditional on unobserved heterogeneity, can be stated as

$$E(y_{it0}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) \tag{18}$$

$$E(y_{it1}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it1}|c_{i1}), \tag{19}$$

where $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})$ is the time sequence of all treatments. If the gain from treatment only depends on t,

$$E(y_{it1}|c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t \tag{20}$$

and then

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t w_{it}. \tag{21}$$

If we assume

$$E(y_{it0}|c_{i0}) = \alpha_{t0} + c_{i0}, \tag{22}$$

then

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{23}$$

an estimating equation that leads to FE or FD (often with $\tau_t = \tau$).

- If add strictly exogenous covariates, and assume linearity of conditional expectations, and allow the gain from treatment to depend on $\mathbf{x}_{it}$ and an additive unobserved effect a_i, get

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, a_i) = \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 \\ + w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i w_{it}, \end{aligned} \tag{24}$$

a correlated random coefficient model because the

coefficient on w_{it} is $(\tau_t + a_i)$. Can eliminate a_i (and c_{i0}). Or, with $\tau_t = \tau$, can "estimate" the τ_i and then use

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}_i. \tag{25}$$

See Wooldridge (2002, Section 11.2) for standard error, or bootstrapping.

5. Semiparametric and Nonparametric Approaches

- Return to the setting with two groups and two time periods. Athey and Imbens (2006) generalize the standard DD model in several ways. Let the two time periods be $t = 0$ and 1 and label the two groups $g = 0$ and 1. Let $Y_i(0)$ be the counterfactual outcome in the absense of intervention and $Y_i(1)$ the counterfactual outcome with intervention. AI

assume that

$$Y_i(0) = h_0(U_i, T_i), \tag{26}$$

where T_i is the time period and

$$h_0(u,t) \text{ strictly increasing in } u \text{ for } t = 0, 1 \tag{27}$$

The random variable U_i represents all unobservable characteristics of individual i. Equation (26) incorporates the idea that the outcome of an individual with $U_i = u$ will be the same in a given time period, irrespective of group membership.

- The distribution of U_i is allowed to vary across groups, but not over time within groups, so that

$$D(U_i|T_i, G_i) = D(U_i|G_i). \tag{28}$$

The standard DD model can be expressed in this way, with

$$h_0(u,t) = u + \delta \cdot t \tag{29}$$

and

$$U_i = \alpha + \gamma G_i + V_i,\ V_i \perp (G_i, T_i) \tag{30}$$

although, because of the linearity, we can get by with the mean independence assumption $E(V_i|G_i, T_i) = 0$. With constant treatment effect,

$$Y_i = \alpha + \beta T_i + \gamma G_i + \tau G_i T_i + V_i, \tag{31}$$

Because $E(V_i|G_i, T_i) = 0$, the parameters in (31) can be estimated by OLS (usual DD analysis).

- Athey and Imbens call the extension of the usual DD model the *changes-in-changes* (CIC) model. Can recover

$$D(Y_i(0)|G_i = 1, T_i = 1), \tag{32}$$

under their assumptions (with an extra support condition). In fact, if $F^0_{gt}(y)$ the be cumulative distribution function of $D(Y_i(0)|G_i = g, T_i = t)$ for

$g = 1, 2$ and $t = 1, 2$, and $F_{gt}(y)$ is the cdf for the observed outcome Y_i conditional on $G_i = g$ and $T_i = t$, then

$$F_{11}^{(0)}(y) = F_{10}(F_{00}^{-1}(F_{01}(y))), \tag{33}$$

where $F_{00}^{-1}(\cdot)$ is the inverse function of $F_{00}(\cdot)$, which exists under the strict monotonicity assumption. Because $F_{11}^{(1)}(y) = F_{11}(y)$, we can estimate the entire distributions of both counterfactuals conditional on intervention, $G_i = T_i = 1$.

- Can apply to repeated cross sections or panel data. Of course, can also identify the average treatment effect

$$\tau_{CIC} = E(Y_{11}(1)) - E(Y_{11}(0)). \tag{34}$$

In particular,

$$\tau_{CIC} = E(Y_{11}) - E[F_{01}^{-1}(F_{00}(Y_{10}))]. \quad (35)$$

- Other approaches with panel data: Altonji and Matzkin (2005) under exchaneability in $D(U_i|W_{i1}, \ldots, W_{iT})$.
- Heckman, Ichimura, Smith, and Todd (1997) and Abadie (2005). Consider basic setup with two time periods, no treated units in first time period. Without an i subscript, $Y_t(w)$ is the counterfactual outcome for treatment level w, $w = 0, 1$, at time t. Parameter: the average treatment effect on the treated,

$$\tau_{ATT} = E[Y_1(1) - Y_1(0)|W = 1]. \quad (36)$$

Remember, in the current setup, no units are treated in the initial time period, so $W = 1$ means treatment in the second time period.

- Key unconfoundedness assumption:

$$E[Y_1(0) - Y_0(0)|X, W] = E[Y_1(0) - Y_0(0)|X] \qquad (37)$$

Also need

$$P(W = 1|X) < 1 \qquad (38)$$

is critical. Under (37) and (38),

$$\tau_{ATT} = E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{[1 - p(X)]} \right\} / P(W = 1), \qquad (39)$$

where Y_t, $t = 0, 1$ are the observed outcomes (for the same unit) and $p(X) = P(W = 1|X)$ is the propensity score. Dehejia and Wahba (1999) derived (39) for the cross-sectional case. All quantities are observed or, in the case of the $p(X)$ and $\rho = P(W = 1)$, can be estimated. As in Hirano, Imbens, and Ridder (2003), a flexible logit model can be used for $p(X)$; the fraction of units treated

would be used for $\hat{\rho}$. Then

$$\hat{\tau}_{ATT} = \hat{\rho}^{-1} N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{[1 - \hat{p}(X_i)]} \right\}. \qquad (40)$$

is consistent and $\sqrt{N}$-asymptotically normal. HIR discuss variance estimation. Imbens and Wooldridge (2007) provide a simple adjustment available in the case that $\hat{p}(\cdot)$ is treated as a parametric model.

- Similar approach works for τ_{ATE}.
- Regression version:

$$\Delta Y_i \text{ on } 1, W_i, \hat{p}(X_i), (W_i - \hat{\rho}) \cdot \hat{p}(X_i), i = 1, \ldots, N.$$

The coefficient on W_i is the estimated ATE.

Requires some functional form restrictions.

Certainly preferred to running the regression Y_{it} on $1, d1_t, d1_t \cdot W_i, \hat{p}(X_i)$. This latter regression

requires unconfoundedness in the levels, and as dominated by the basic DD estimate from ΔY_i on $1, W_i$

- Regression adjustment can also be used, as in HIST (1997).

6. Synthetic Control Methods for Comparative Case Studies

- Abadie, Diamond, and Hainmueller (2007) argue that in policy analysis at the aggregate leve, there is little or no estimation uncertainty: the goal is to determine the effect of a policy on an entire population, and the aggregate is measured without error (or very little error). Application: California's tobacco control program on state-wide smoking rates.
- ADH focus on the uncertainty with choosing a

suitable control for California among other states (that did not implement comparable policies over the same period).

- ADH suggest using many potential control groups (38 states or so) to create a single synthetic control group.
- Two time periods: one before the policy and one after. Let y_{it} be the outcome for unit i in time t, with $i = 1$ the treated unit. Suppose there are J possible controls, and index these as $\{2, \ldots, J+1\}$. Let $\mathbf{x}_i$ be observed covariates for unit i that are not (or would not be) affected by the policy; $\mathbf{x}_i$ may contain period $t = 2$ covariates provided they are not affected by the policy. Generally, we can estimate the effect of the policy as

$$y_{12} - \sum_{j=2}^{J+1} w_j y_{j2}, \tag{41}$$

where w_j are nonnegative weights that add up to one. How to choose the weights to best estimate the intervention effect?

- ADH propose choosing the weights so as to minimize the distance between $(y_{11}, \mathbf{x}_1)$ and $\sum_{j=2}^{J+1} w_j \cdot (y_{j1}, \mathbf{x}_j)$, say. That is, functions of the pre-treatment outcomes and the predictors of post-treatment outcomes.
- ADH propose permutation methods for inference, which require estimating a placebo treatment effect for each region, using the same synthetic control method as for the region that underwent the intervention.