Control Function and Related Methods: Linear Models

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. Models Linear in Endogenous Variables
2. Models Nonlinear in Endogenous Variables
3. Correlated Random Coefficient Models
4. Endogenous Switching
5. Random Coefficients in Reduced Forms

1. Models Linear in Endogenous Variables

- Most models that are linear in parameters are estimated using standard IV methods – two stage least squares (2SLS).
- An alternative, the control function (CF) approach, relies on the same kinds of identification conditions.
- In models with nonlinearities or random coefficients, the form of exogeneity is stronger and more restrictions are imposed on the reduced forms.

CF Methods Based on Linear Projections

- 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 $z_1 = 1$):

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_1, \tag{1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of the $1 \times L$ exogenous variables $\mathbf{z}$.

- Weakest exogeneity assumption:

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

- Reduced form for y_2 is a linear projection:

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

- The linear projection of u_1 on v_2 in error form is

$$u_1 = \rho_1 v_2 + e_1, \tag{4}$$

where $\rho_1 = E(v_2 u_1)/E(v_2^2)$ is the population regression coefficient.

- By construction, $E(v_2 e_1) = 0$ and $E(\mathbf{z}'e_1) = \mathbf{0}$.

• Plug $u_1 = \rho_1 v_2 + e_1$ into $y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$:

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

where v_2 is now an explanatory variable in the equation. The new error, e_1, is uncorrelated with y_2 as well as with v_2 and $\mathbf{z}$.

• Two-step procedure: (i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the reduced form residuals, $\hat{v}_{i2}$; (ii) Regress

$$y_{i1} \text{ on } y_{i2},\ \mathbf{z}_{i1}, \text{and } \hat{v}_{i2}. \tag{6}$$

• OLS estimators from (6) are consistent for $\boldsymbol{\delta}_1, \alpha_1$, and ρ_1. These are the *control function* estimators.

• Implicit error in (6) is $e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, so asymptotic variance depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$.

• Can use heteroskedasticity-robust t statistic to test $H_0 : \rho_1 = 0$ (y_2 exogenous). Regression-based Hausman test.

• Algebra: The OLS estimates of $\boldsymbol{\delta}_1$ and α_1 from (6) are *identical* to the 2SLS estimates of

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_1$$

CF Methods Based on Conditional Expectations

- Start again with the basic equation

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_1$$

We can derive a CF approach based on $E(y_1|y_2, \mathbf{z})$ rather than $L(y_1|y_2, \mathbf{z})$.

- The estimating equation is based on

$$E(y_1|y_2, \mathbf{z}) = \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + E(u_1|y_2, \mathbf{z}). \tag{7}$$

- The linear projection approach imposes no distributional assumptions, even about a conditional mean. (Second moments finite.) Using the CE approach we may have to impose a lot of structure to get $E(u_1|y_2, \mathbf{z})$.
- As an example, suppose

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0] \tag{8}$$

where (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

$$E(u_1|y_2, \mathbf{z}) = \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \tag{9}$$

where $\lambda(\cdot)$ is the inverse Mills ratio (IMR).

• Heckman two-step approach (for endogeneity, not sample selection):

(i) Probit to get $\hat{\boldsymbol{\delta}}_2$ and compute the *generalized residuals*,

$\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$.

(ii) Regress y_{i1} on $\mathbf{z}_{i1},\ y_{i2},\ \widehat{gr}_{i2},\ i = 1,\ldots,N$.

• The Stata command `treatreg` effectively implements this procedure (two-step or full MLE).

- Consistency of the CF estimator hinges on the probit model for $D(y_2|\mathbf{z})$ being correctly specified along with $E(u_1|e_2) = \rho_1 e_2$, where $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$.
- Instead we can apply 2SLS directly to $y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + u_1$. We need make no distinction among cases where y_2 is discrete, continuous, or some mixture.
- If $\mathbf{y}_2$ is a vector the CF approach based on $E(y_1|\mathbf{y}_2, \mathbf{z})$ can be much harder than 2SLS. We need $E(u_1|\mathbf{y}_2, \mathbf{z})$.

• How might we use the binary nature of y_2 in IV estimation in a robust manner?

(i) Obtain the fitted probabilities, $\hat{\Phi}_{i2} = \Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, from the first stage probit.

(ii) Estimate $y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_{i1}y_2 + u_{i1}$ by IV using $(\mathbf{z}_{i1}, \hat{\Phi}_{i2})$ as instruments (not regressors!)

• If $E(u_1|\mathbf{z}) = 0$, this IV estimator is fully robust to misspecification of the probit model, usual standard errors from IV asymptotically valid. Efficient IV estimator if $P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$.

2. Models Nonlinear in Endogenous Variables

- Adding nonlinear functions of EEVs produces differences between IV and CF approaches. For example, add y_2^2:

$$y_1 = \alpha_1 y_2 + \gamma_1 y_2^2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_1 \tag{10}$$

$$E(u_1|\mathbf{z}) = 0. \tag{11}$$

- Assumption (11) is stronger than $E(\mathbf{z}' u_1) = \mathbf{0}$ and is essential for nonlinear models (so that nonlinear functions of EEVs come with their own IVs).
- Suppose z_2 is a scalar not in $\mathbf{z}_1$. We can use z_2^2 as an instrument for y_2^2. So the IVs would be $(\mathbf{z}_1, z_2, z_2^2)$ for $(\mathbf{z}_1, y_2, y_2^2)$.

- A linear projection CF approach would regress y_2 and y_2^2 separately on $(\mathbf{z}_1, z_2, z_2^2)$, obtain two sets of residuals, and add these as controls in an OLS regression. This is identical to the IV estimate. (Can add $z_1\mathbf{z}_2$ to IV list.)

- If we make a stonger assumption then a single control function suffices. In particular, *assume*

$$E(u_1|\mathbf{z}, y_2) = E(u_1|v_2) = \rho_1 v_2, \tag{12}$$

where $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2$.

- Independence of (u_1, v_2) and $\mathbf{z}$ is sufficient for the first equality, which is a substantive restriction. Linearity of $E(u_1|v_2)$ is also a substantive restriction.

• Assumption (12) imposes real restrictions; not just a linear projection. It would be hard to justify for discrete y_2 (or discrete y_1).

• If we assume (12),

$$E(y_1|\mathbf{z}, y_2) = \alpha_1 y_2 + \gamma_1 y_2^2 + \mathbf{z}_1\boldsymbol{\delta}_1 + \rho_1 v_2, \tag{13}$$

and a CF approach is immediate.

(i) Get the OLS residuals, $\hat{v}_{i2}$, from the first-stage regression y_{i2} on $\mathbf{z}_i$.

(ii) OLS of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, y_{i2}^2, \hat{v}_{i2}$.

• A single control function suffices.

• This CF method *not* equivalent to a 2SLS estimate. CF likely more efficient but less robust.

- Similar comments hold in a model such as

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

- We could use IVs of the form $(\mathbf{z}_1, z_2, z_2\mathbf{z}_1)$ and add squares, too.
- If we assume $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2$ with $E(u_1|y_2, \mathbf{z}) = \rho_1 v_2$ then just add one CF.
- In general, CF approach imposes extra assumptions when we base it on $E(y_1|y_2, \mathbf{z})$. In a parametric context, often half to for models nonlinear in parameters and random coefficient models.

• Heckman and Vytlacil (1998) suggest "plug-in" estimators in (14) (and also with random coefficients). Key assumption along with $E(u_1|\mathbf{z}) = 0$ is

$$E(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$$

• Estimating equation is based on $E(y_1|\mathbf{z})$:

$$E(y_1|\mathbf{z}) = \alpha_1(\mathbf{z}\boldsymbol{\pi}_2) + \mathbf{z}_1\boldsymbol{\delta}_1 + (\mathbf{z}\boldsymbol{\pi}_2)\mathbf{z}_1\boldsymbol{\gamma}_1$$

(i) Regress y_{i2} on $\mathbf{z}_i$, get fitted values $\hat{y}_{i2}$. (ii) Regress y_{i1} on $\hat{y}_{i2}$, $\mathbf{z}_{i1}$, $\hat{y}_{i2}\mathbf{z}_{i1}$.

- As with CF approach must deal with generated regressors. CF approach gives simple test of exogeneity of y_2.
- Plug-in approach less robust than the estimator that uses nonlinear functions of $\mathbf{z}$ as IVs [because such methods do not restrict $E(y_2|\mathbf{z})$].
- Can use IV with instruments $(\mathbf{z}_i, \hat{y}_{i2}\mathbf{z}_i)$.

3. Correlated Random Coefficient Models

- Suppose we allow y_2 to have a random slope:

$$y_1 = \eta_1 + a_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_1, \tag{15}$$

where a_1, the "random coefficient" on y_2. Heckman and Vytlacil (1998) call (15) a "correlated random coefficient" (CRC) model.

- For a random draw i from the population:

$$y_{i1} = \eta_1 + a_{i1} y_{i2} + \mathbf{z}_1 \boldsymbol{\delta}_1 + u_{i1} \tag{16}$$

- Write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = E(a_1)$ (the average partial effect) is (initially) the object of interest.
- Rewrite the equation as

$$
\begin{aligned}
y_1 &= \eta_1 + \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + v_1 y_2 + u_1 &(17)\\
&\equiv \eta_1 + \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + e_1. &(18)
\end{aligned}
$$

- Potential problem with applying IV: the error term $v_1 y_2 + u_1$ is not necessarily uncorrelated with the instruments $\mathbf{z}$, even if we maintain

$$E(u_1|\mathbf{z}) = E(v_1|\mathbf{z}) = 0. \tag{19}$$

- We want to allow y_2 and v_1 to be correlated, $Cov(v_1, y_2) \equiv \tau_1 \neq 0$, along with $Cov(y_2, u_1) \neq 0$.
- Suppose the conditional covariate is constant:

$$Cov(v_1, y_2|\mathbf{z}) = Cov(v_1, y_2), \tag{20}$$

which is sufficient along with (19) for standard IV estimators to consistently estimate $(\alpha_1, \boldsymbol{\delta}_1)$ (not intercept).

- The CF approach due to Garen (1984) requires more assumptions, but is more efficient and delivers more:

$$\begin{aligned} y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + v_2 \\ E(y_1|\mathbf{z}, v_2) &= \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + E(v_1|\mathbf{z}, v_2)y_2 + E(u_1|\mathbf{z}, v_2) \\ &= \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + \theta_1 v_2 y_2 + \rho_1 v_2 \end{aligned}$$

- CF estimator: After getting residuals $\hat{v}_{i2}$ from y_{i2} on $\mathbf{z}_i$ run

$$y_{i1} \text{ on } 1,\ y_{i2},\ \mathbf{z}_{i1},\ \hat{v}_{i2}y_{i2},\ \hat{v}_{i2}$$

- Joint Wald test for null that y_2 is exogenous (two degrees of freedom).

• Neither $Cov(v_1, y_2|\mathbf{z}) = Cov(v_1, y_2)$ nor Garen's CF assumptions $[E(v_1|\mathbf{z}, v_2) = \theta_1 v_2, E(u_1|\mathbf{z}, v_2) = \rho_1 v_2]$ can be obtained if y_2 follows standard discrete response models.

• Card (2001) shows (20) can can be violated even if y_2 is continuous. Wooldridge (2005) shows how to allow parametric heteroskedasticity in the reduced form equation.

4. Endogenous Switching

- Suppose y_2 is binary and interacts with an unobservable. If y_2 also interacts with $\mathbf{z}_1$ we have an unrestricted "endogenous switching regression" model:

$$y_1 = \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\psi}_1)\boldsymbol{\gamma}_1 + u_1 + y_2 v_1 \qquad (21)$$

where $\boldsymbol{\psi}_1 = E(\mathbf{z}_1)$ and α_1 is the average treatment effect.

- If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 > 0]$ follows a probit model,

$$E(u_1|e_2,\mathbf{z}) = \rho_1 e_2, E(v_1|e_2,\mathbf{z}) = \xi_1 e_2$$

then

$$E(y_1|\mathbf{z},e_2) = \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\psi}_1)\boldsymbol{\gamma}_1 + \rho_1 e_2 + \xi_1 y_2 e_2$$

- By iterated expectations,

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

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 for the probit model.

• The two-step estimation method is the one due to Heckman (1976). Centering $\mathbf{z}_{i1}$ before interacting with y_{i2} ensures $\hat{\alpha}_1$ is the estimated ATE:

$$y_{i1} \text{ on } 1,\ y_{i2},\ \mathbf{z}_{i1},\ y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1),\ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ y_{i2}h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2) \tag{22}$$

where $\hat{\boldsymbol{\delta}}_2$ is from the probit of y_{i2} on $\mathbf{z}_i$.

• Note: We do not get any interesting treatment effects by taking changes or derivatives of $E(y_1|\mathbf{z}, y_2)$.

• The average treatment effect on the treated (ATT) for a given $\mathbf{z}$ is estimated as

$$\hat{\tau}_{att}(\mathbf{z}) = \hat{\alpha}_1 + (\mathbf{z}_1 - \bar{\mathbf{z}}_1)\hat{\boldsymbol{\gamma}}_1 + \hat{\xi}_1\lambda(\mathbf{z}\hat{\boldsymbol{\delta}}_2).$$

Can average out $\mathbf{z}$ over the treated group to get the unconditional ATT.

- Extension to random coefficients everywhere:

$$y_1 = \eta_1 + a_1 y_2 + \mathbf{z}_1\mathbf{d}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\mathbf{g}_1 + u_1. \tag{23}$$

- If we assume that $E(a_1|v_2)$, $E(\mathbf{d}_1|v_2)$, and $E(\mathbf{g}_1|v_2)$ are linear in e_2, then

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\xi}_1 + \rho_1 E(e_2|\mathbf{z},y_2) \\ &\quad + \xi_1 y_2 E(e_2|\mathbf{z},y_2) + \mathbf{z}_1 E(e_2|\mathbf{z},y_2)\boldsymbol{\psi}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)E(e_2|\mathbf{z},y_2)\boldsymbol{\omega}_1 \\ &= \eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \xi_1 y_2 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) \\ &\quad + h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)\mathbf{z}_1\boldsymbol{\psi}_1 + y_2 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\omega}_1. \end{aligned}$$

- After the first-stage probit, the second-stage regression can be obtained as

$$y_{i1} \text{ on } 1,\ y_{i2},\ \mathbf{z}_{i1}, y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1),\ \hat{h}_{i2}, y_{i2}\hat{h}_{i2}, \hat{h}_{i2}\mathbf{z}_{i1},\ y_{i2}\hat{h}_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1) \qquad (24)$$

across all observations i, where $\hat{h}_{i2} = h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$

- So IMR appears by itself, interacted with y_{i2} and $\mathbf{z}_{i1}$, and also in a triple interaction.
- Can bootstrap standard errors or use delta method.
- Null test of exogenous is joint significance of all terms containing

- If apply linear IV to the endogenous switching regression model get "local average treatment effect" interpretation under weak assumptions.
- Under the assumption that each regime has the same unobservable, IV estimates the treatment effect conditional on $\mathbf{z}_{i1}$.
- If we believe the CF assumptions, we can estimate treatment effects conditional on $(y_{i2}, \mathbf{z}_i)$, and so ATT and ATU as special cases.

• Let $\mathbf{x}_1$ be a general function of $(y_2, \mathbf{z}_1)$, including an intercept. Then the general model can be written as

$$y_1 = \mathbf{x}_1 \mathbf{b}_1$$

where $\mathbf{b}_1$ is a $K_1 \times 1$ random vector. If y_2 follows a probit

$$y_1 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 > 0]$$

then under multivariate normality (or weaker assumptions) the CF approach allows us to estimate

$$E(\mathbf{b}_1 | y_2, \mathbf{z})$$

• IV approaches allow us to estimate $E(\mathbf{b}_1)$ under some assumptions and only LATE under others.

5. Random Coefficients in Reduced Forms

- Random coefficients in reduced forms ruled out in Blundell and Powell (2003) and Imbens and Newey (2006).
- Hoderlein, Nesheim, and Simoni (2012) show cannot generally get point identification, even in simple model.
- Of interest because the reaction of individual units to changes in the instrument may differ in unobserved ways.
- Under enough assumptions can obtain new CF methods in linear models that allow for slope heterogeneity everywhere.

• For simplicity, a single EEV, y_2. $\mathbf{x}_1$ a function of $(y_2, \mathbf{z}_1)$, and has an intercept. IV vector $\mathbf{z}$ also contains unity:

$$y_1 = \mathbf{x}_1\mathbf{b}_1 \equiv \mathbf{x}_1\boldsymbol{\beta} + \mathbf{x}_1\mathbf{a}_1 \equiv \mathbf{x}_1\boldsymbol{\beta} + u_1 \tag{25}$$

$$y_2 = \mathbf{z}\mathbf{g}_2 = \mathbf{z}\boldsymbol{\gamma}_2 + \mathbf{z}\mathbf{c}_2 \equiv \mathbf{z}\boldsymbol{\gamma}_2 + v_2 \tag{26}$$

where $\mathbf{b}_1 = \boldsymbol{\beta}_1 + \mathbf{a}_1$, $E(\mathbf{a}_1) = \mathbf{0}$, $\mathbf{g}_2 = \boldsymbol{\gamma}_2 + \mathbf{c}_2$, $E(\mathbf{c}_2) = \mathbf{0}$, and

$$u_1 = \mathbf{x}_1\mathbf{a}_1$$

$$v_2 = \mathbf{z}\mathbf{c}_2$$

• Assume $(\mathbf{a}_1, \mathbf{c}_2)$ independent of $\mathbf{z}$.

• We can estimate v_2 because $E(v_2|\mathbf{z}) = 0$. Assuming joint normality (and somewhat weaker), can obtain a CF approach.

$$E(\mathbf{a}_1|v_2, \mathbf{z}) = \frac{Cov(\mathbf{a}_1, v_2|\mathbf{z})}{Var(v_2|\mathbf{z})} \cdot v_2 = \frac{E(\mathbf{a}_1\mathbf{c}_2)\mathbf{z}_i'}{Var(v_2|\mathbf{z})} \cdot v_2 \tag{27}$$

Now

$$\begin{aligned} Var(v_2|\mathbf{z}) &= \mathbf{z}'\boldsymbol{\Omega}_{\mathbf{c}}\mathbf{z} \\ Cov(\mathbf{a}_1, v_2|\mathbf{z}) &= E(\mathbf{a}_1 v_2|\mathbf{z}) = E(\mathbf{a}_1\mathbf{c}_2)\mathbf{z}' = \boldsymbol{\Omega}_{\mathbf{ac}}\mathbf{z}'. \end{aligned}$$

• Combining gives

$$E(\mathbf{a}_1|v_2,\mathbf{z}) = \frac{\mathbf{\Omega_{ac}z}'}{\mathbf{z}'\mathbf{\Omega_c z}} \cdot v_2 \tag{28}$$

and so

$$\begin{aligned} E(y_1|v_2,\mathbf{z}) &= \mathbf{x}_1\boldsymbol{\beta}_1 + \mathbf{x}_1 E(\mathbf{a}_1|v_2,\mathbf{z}) = \mathbf{x}_1\boldsymbol{\beta}_1 + \frac{\mathbf{x}_1\mathbf{\Omega_{ac}z}'}{\mathbf{z}'\mathbf{\Omega_c z}} v_2 \\ &= \mathbf{x}_1\boldsymbol{\beta}_1 + [(\mathbf{x}_1 \otimes \mathbf{z})vec(\mathbf{\Omega_{ac}})]v_2/h(\mathbf{z},\mathbf{\Omega_c}) \\ &\equiv \mathbf{x}_1\boldsymbol{\beta}_1 + [(\mathbf{x}_1 \otimes \mathbf{z})v_2/h(\mathbf{z},\mathbf{\Omega_c})]\boldsymbol{\xi}_1 \end{aligned} \tag{29}$$

where $\boldsymbol{\xi}_1 = vec(\mathbf{\Omega_{ac}})$ and $h(\mathbf{z},\mathbf{\Omega_c}) \equiv \mathbf{z}'\mathbf{\Omega_c z}$.

• Can operationalize (29) by noting that $\boldsymbol{\gamma}_2$ and $\boldsymbol{\Omega}_{\mathbf{c}}$ are identified from the reduced form for y_2:

$$E(y_2|\mathbf{z}_i) = \mathbf{z}\boldsymbol{\gamma}_2$$
$$Var(y_2|\mathbf{z}) = \mathbf{z}'\boldsymbol{\Omega}_{\mathbf{c}}\mathbf{z}$$

• Can use two-step estimation for $\boldsymbol{\gamma}_2$ and $\boldsymbol{\Omega}_{\mathbf{c}}$ but also the quasi-MLE using the normal distribution.

• Given consistent estimators of $\boldsymbol{\gamma}_2$ and $\boldsymbol{\Omega}_{\mathbf{c}}$, we can form

$$\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\gamma}}_2, \hat{h}_{i2} = \mathbf{z}_i'\hat{\boldsymbol{\Omega}}_{\mathbf{c}}\mathbf{z}_i \tag{30}$$

• Can use OLS on the second-stage estimating equation:

$$y_{i1} = \mathbf{x}_{i1}\boldsymbol{\beta}_1 + (\mathbf{x}_{i1} \otimes \mathbf{z}_i)(\hat{v}_{i2}/\hat{h}_{i2})\boldsymbol{\xi}_1 + error_i \tag{31}$$

• Need to adjust the asymptotic variance of $(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\xi}}_1')'$ for the first-stage estimation, possibly via bootstrapping or the delta method.

• The population equation underlying (31) has heteroskedasticity. Account for in inference, maybe estimation (GMM).

• Notice that no terms in $(\mathbf{x}_{i1} \otimes \mathbf{z}_i)$ appears by itself in the equation; each is interacted with $(\hat{v}_{i2}/\hat{h}_{i2})$, which is necessary to preserve identification.

Control Function and Related Methods: Nonlinear Models

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. General Approach
2. Nonlinear Models with Additive Errors
3. Models with Intrinsic Nonlinearity
4. “Special Regressor” Methods for Binary Response

1. General Approach

- With models that are nonlinear in parameters, the linear projection approach to CF estimation rarely works (unless the model happens to be linear in the EEVs).
- If $\mathbf{u}_1$ is a vector of "structural" errors, $\mathbf{y}_2$ is the vector of EEVs, and $\mathbf{z}$ is the vector of exogenous variables, we at least have to model $E(\mathbf{u}_1|\mathbf{y}_2,\mathbf{z})$ and often $D(\mathbf{u}_1|\mathbf{y}_2,\mathbf{z})$ (in a parametric context).
- An important simplification is when

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z},\boldsymbol{\delta}_2) + \mathbf{v}_2 \tag{1}$$

where $\mathbf{v}_2$ is independent of $\mathbf{z}$. Unfortunately, this rules out discrete $\mathbf{y}_2$.

- With discreteness in $\mathbf{y}_2$, difficult to get by without modeling $D(\mathbf{y}_2|\mathbf{z})$.
- In many cases – particularly when $\mathbf{y}_2$ is continuous – one has a choice between two-step control function estimators and one-step estimators that estimate parameters at the same time. (Typically these have a quasi-LIML flavor.)
- More radical suggestions are to use generalized residuals in nonlinear models as an approximate solution to endogeneity.

2. Nonlinear Models with Additive Errors

- Suppose

$$y_1 = g_1(\mathbf{y}_2, \mathbf{z}_1, \boldsymbol{\gamma}_1) + u_1$$
$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}, \boldsymbol{\gamma}_2) + \mathbf{v}_2$$

and

$$E(u_1|\mathbf{v}_2, \mathbf{z}) = E(u_1|\mathbf{v}_2) = \mathbf{v}_2\boldsymbol{\rho}_1$$

- Assume we have enough relevant elements in $\mathbf{z}_2$ so identification holds.

- We can base a CF approach on

$$E(y_1|\mathbf{y}_2,\mathbf{z}) = g_1(\mathbf{y}_2,\mathbf{z}_1,\boldsymbol{\gamma}_1) + \mathbf{v}_2\boldsymbol{\rho}_1 \tag{2}$$

- Estimate $\boldsymbol{\gamma}_2$ by multivariate nonlinear least squares, or an MLE, to get

$$\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \mathbf{g}_2(\mathbf{z}_i,\hat{\boldsymbol{\gamma}}_2).$$

- In second step, estimate $\boldsymbol{\gamma}_1$ and $\boldsymbol{\rho}_1$ by NLS using the mean function in (2).
- Easiest when $\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Gamma}_2 + \mathbf{v}_2$ so can use linear estimation in first stage.
- Can allow a vector $\mathbf{y}_1$, as in Blundell and Robin (1999, Journal of Applied Econometrics): expenditure share system with total expenditure endogenous.

- In principle can have $\mathbf{y}_2$ discrete provided we can find $E(u_1|\mathbf{y}_2, \mathbf{z})$.
- If y_2 is binary, can have nonlinear switching regression – but with additive noise.
- Example: Exponential function:

$$
\begin{aligned}
y_1 &= \exp(\eta_1 + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_1\boldsymbol{\psi}_1) + u_1 + y_2 v_1 \\
y_2 &= 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 > 0]
\end{aligned}
$$

- Usually more natural for the unobservables to be inside the exponential function. And what if y_1 is something like a count variable?
- Same issue arises in share equations.

3. Models with Intrinsic Nonlinearity

- Typically three approaches to nonlinear models with EEVs.

(1) Plug in fitted values from a first step estimation in an attempt to mimic 2SLS in linear model. Usually does not produce consistent estimators because the implied form of $E(y_1|\mathbf{z})$ or $D(y_1|\mathbf{z})$ is incorrect.

(2) CF approach: Plug in residuals in an attempt to obtain $E(y_1|y_2,\mathbf{z})$ or $D(y_1|y_2,\mathbf{z})$.

(3) Maximum Likelihood (often limited information): Use models for $D(y_1|y_2,\mathbf{z})$ and $D(y_2|\mathbf{z})$ jointly.

- All strategies are more difficult with nonlinear models when y_2 is discrete.

Binary and Fractional Responses

Probit model:

$$y_1 = 1[\alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + u_1 \geq 0], \tag{3}$$

where $u_1|\mathbf{z} \sim Normal(0,1)$. Analysis goes through if we replace $(\mathbf{z}_1, y_2)$ with any known function $\mathbf{x}_1 \equiv \mathbf{g}_1(\mathbf{z}_1, y_2)$.

• The Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for y_2,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2,\ \ v_2|\mathbf{z} \sim Normal(0, \tau_2^2). \tag{4}$$

- RV approach comes close to requiring

$$(u_1, v_2) \text{ independent of } \mathbf{z}. \tag{5}$$

If we also assume

$$(u_1, v_2) \sim \text{Bivariate Normal} \tag{6}$$

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

$$P(y_1 = 1|y_2, \mathbf{z}) = \Phi(\alpha_{\rho 1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{\rho 1} + \theta_{\rho 1} v_2) \tag{7}$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$.

The Rivers-Vuong CF approach is

(i) OLS of y_{i2} on $\mathbf{z}_i$, to obtain the residuals, $\hat{v}_{i2}$.

(ii) Probit of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$ to estimate the scaled coefficients. A simple t test on $\hat{v}_2$ is valid to test $H_0 : \rho_1 = 0$.

• Can recover the original coefficients, which appear in the partial effects – see Wooldridge (2010, Chapter 15). Or, obtain average partial effects by differentiating the estimated "average structural function":

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1} \hat{v}_{i2}), \tag{8}$$

that is, we average out the reduced form residuals, $\hat{v}_{i2}$.

• Cover the ASF in more detail later.

- The two-step CF approach easily extends to fractional responses: $0 \leq y_1 \leq 1$. Modify the model as

$$E(y_1|y_2, \mathbf{z}, q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \tag{9}$$

where $\mathbf{x}_1$ is a function of $(y_2, \mathbf{z}_1)$ and q_1 contains unobservables.

- Assume $q_1 = \rho_1 v_2 + e_1$ where $D(e_1|\mathbf{z}, v_2) = Normal(0, \sigma_{e_1}^2)$.
- Use the *same* two-step estimator as for probit. (In Stata, `glm` command in second stage.) In this case, must obtain APEs from the ASF in (8).

- In inference, assume only that the mean is correctly specified. (Use sandwich in Bernoulli quasi-LL.)
- To account for first-stage estimation, the bootstrap is convenient.
- No IV procedures available unless assume that, say, the log-odds transform of y_1 is linear in $\mathbf{x}_1\boldsymbol{\beta}_1$ and an additive error independent of $\mathbf{z}$.

• CF has clear advantages over "plug-in" approach, even in binary response case. Suppose rather than conditioning on v_2 along with $\mathbf{z}$ (and therefore y_2) to obtain $P(y_1 = 1|\mathbf{z}, y_2)$ we use

$$\begin{aligned} P(y_1 = 1|\mathbf{z}) &= \Phi\{[\alpha_1(\mathbf{z}\boldsymbol{\pi}_2) + \mathbf{z}_1\boldsymbol{\delta}_1]/\omega_1\} \\ \omega_1^2 &= Var(\alpha_1 v_2 + u_1) \end{aligned}$$

(i) OLS on the reduced form, and get fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$. (ii) Probit of y_{i1} on $\hat{y}_{i2}$, $\mathbf{z}_{i1}$. 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 $y_2\mathbf{z}_1$, and use probit in second stage. Does *not* result in consistent estimation of the scaled parameters or the partial effects.
- Adding the CF $\hat{v}_2$ solves the endogeneity problem regardless of how y_2 appears.

Example: Married women's fraction of hours worked.

```
. use mroz

. gen frachours = hours/8736

. sum frachours

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   frachours |       753    .0847729    .0997383          0   .5666209
```

```
. reg nwifeinc educ exper expersq kidslt6 kidsge6 age huseduc husage

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  8,   744) =   23.77
       Model |  20722.898     8  2590.36225           Prob > F      =  0.0000
    Residual | 81074.2176   744  108.970723           R-squared     =  0.2036
-------------+------------------------------           Adj R-squared =  0.1950
       Total | 101797.116   752  135.368505           Root MSE      =  10.439

------------------------------------------------------------------------------
    nwifeinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6721947   .2138002     3.14   0.002     .2524713    1.091918
       exper |  -.3133239   .1383094    -2.27   0.024    -.5848472   -.0418007
     expersq |  -.0003769   .0045239    -0.08   0.934    -.0092581    .0085043
     kidslt6 |   .9004389   .8265936     1.09   0.276    -.7222947    2.523172
     kidsge6 |   .4462001   .3225308     1.38   0.167    -.1869788    1.079379
         age |   .2819309   .1075901     2.62   0.009     .0707146    .4931472
     huseduc |   1.188289   .1617589     7.35   0.000     .8707307    1.505847
      husage |   .0681739   .1047836     0.65   0.515    -.1375328    .2738806
       _cons |  -15.46223     3.9566    -3.91   0.000    -23.22965   -7.694796
------------------------------------------------------------------------------

. predict v2h, resid
```

```
. glm frachours educ exper expersq kidslt6 kidsge6 age nwifeinc v2h, fam(bin)
      link(probit) robust
note: frachours has noninteger values

Generalized linear models                          No. of obs      =       753
Optimization     : ML                              Residual df     =       744
                                                   Scale parameter =         1
Deviance         =  77.29713199                    (1/df) Deviance =   .103894
Pearson          =  83.04923963                    (1/df) Pearson  =  .1116253

Variance function: V(u) = u*(1-u/1)                [Binomial]
Link function    : g(u) = invnorm(u)               [Probit]

                                                   AIC             =  .4338013
Log pseudolikelihood = -154.3261842                BIC             = -4851.007

------------------------------------------------------------------------------
             |               Robust
   frachours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0437229   .0169339     2.58   0.010       .010533    .0769128
       exper |   .0610646   .0096466     6.33   0.000      .0421576    .0799717
     expersq |    -.00096   .0002691    -3.57   0.000     -.0014875   -.0004326
     kidslt6 |  -.4323608   .0782645    -5.52   0.000     -.5857565   -.2789651
     kidsge6 |  -.0149373   .0202283    -0.74   0.460     -.0545841    .0247095
         age |  -.0219292   .0043658    -5.02   0.000      -.030486   -.0133725
    nwifeinc |  -.0131868   .0083704    -1.58   0.115     -.0295925    .0032189
         v2h |   .0102264   .0085828     1.19   0.233     -.0065957    .0270485
       _cons |  -1.169224   .2397377    -4.88   0.000     -1.639102   -.6993472
------------------------------------------------------------------------------
```

```
. fracivp frachours educ exper expersq kidslt6 kidsge6 age
           (nwifeinc = huseduc husage)

Fitting exogenous probit model
note: frachours has noninteger values

Probit model with endogenous regressors           Number of obs   =        753
                                                  Wald chi2(7)    =     240.78
Log pseudolikelihood = -3034.3388                 Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0131207   .0083969    -1.56   0.118    -.0295782    .0033369
        educ |   .0434908   .0169655     2.56   0.010     .0102389    .0767427
       exper |    .060721   .0098379     6.17   0.000      .041439    .0800029
     expersq |  -.0009547   .0002655    -3.60   0.000    -.0014751   -.0004342
     kidslt6 |  -.4299361   .0802032    -5.36   0.000    -.5871314   -.2727408
     kidsge6 |  -.0148507   .0205881    -0.72   0.471    -.0552025    .0255012
         age |  -.0218038   .0045701    -4.77   0.000     -.030761   -.0128465
       _cons |  -1.162787   .2390717    -4.86   0.000    -1.631359   -.6942151
-------------+----------------------------------------------------------------
     /athrho |   .1059984   .0906159     1.17   0.242    -.0716056    .2836024
    /lnsigma |   2.339528   .0633955    36.90   0.000     2.215275    2.463781
-------------+----------------------------------------------------------------
         rho |   .1056032   .0896054                     -.0714835    .2762359
       sigma |   10.37633    .657813                      9.163926    11.74915
------------------------------------------------------------------------------
Instrumented:  nwifeinc
Instruments:   educ exper expersq kidslt6 kidsge6 age huseduc husage
------------------------------------------------------------------------------
Wald test of exogeneity (/athrho = 0): chi2(1) =     1.37 Prob > chi2 = 0.2421
```

- What are the limits to the CF approach? Consider

$$E(y_1|\mathbf{z}, y_2, q_1) = \Phi(\alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + q_1) \tag{10}$$

where y_2 is discrete. Rivers-Vuong approach does not generally work (even if y_1 is binary).

- Neither does plugging in probit fitted values, assuming

$$P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2) \tag{11}$$

In other words, do *not* try to mimic 2SLS as follows: (i) Do probit of y_2 on $\mathbf{z}$ and get the fitted probabilities, $\hat{\Phi}_2 = \Phi(\mathbf{z}\hat{\boldsymbol{\delta}}_2)$. (ii) Do probit of y_1 on $\mathbf{z}_1, \hat{\Phi}_2$, that is, just replace y_2 with $\hat{\Phi}_2$.

- The only strategy that works under *traditional* assumptions is maximum likelihood estimation based on $f(y_1|y_2,\mathbf{z})f(y_2|\mathbf{z})$. [Perhaps this is why some, such as Angrist (2001), promote the notion of just using linear probability models estimated by 2SLS.]
- "Bivariate probit" software can be used to estimate the probit model with a binary endogenous variable. Wooldridge (2011) shows that the same quasi-LIML is consistent when y_1 is fractional if (10) holds.
- Can also do a full switching regression when y_1 is fractional. Use "heckprobit" quasi-LLs.

• A CF approach based on generalize residuals can be justified for "small" amounts of endogeneity. Consider

$$E(y_1|y_2, \mathbf{z}, q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1) \tag{11}$$

and

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 > 0] \tag{12}$$

• (q_1, e_1) jointly normal and independent of $\mathbf{z}$.

• Let

$$\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) \tag{13}$$

be the generalized residuals from the probit estimation.

• The variable addition test (essentially score test) for the null that q_1 and e_2 are uncorrelated can be obtained by "probit" of y_{i1} on the mean function

$$\Phi(\mathbf{x}_{i1}\boldsymbol{\beta}_{\rho 1} + \eta_{\rho 1}\widehat{gr}_{i2}) \tag{14}$$

and use a robust t statistic for $\hat{\eta}_{\rho 1}$. (Get scaled estimates of $\boldsymbol{\beta}_1$ and η_1.)

• Wooldridge (2011) suggests that this can approximate the APEs, obtained from the estimated average structural function:

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\eta}_{\rho 1} \widehat{gr}_{i2}) \tag{15}$$

• Simulations suggest this can work pretty well, even if the amount of endogeneity is not "small."

- If we have two sources of unobservables, add an interaction:

$$E(y_1|y_2,\mathbf{z}) \approx \Phi(\mathbf{x}_1\boldsymbol{\beta}_{\rho 1} + \eta_{\rho 1} gr_2 + \omega_{\rho 1} y_2 gr_2) \tag{16}$$

$$\widehat{ASF}(y_2,\mathbf{z}_1) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\eta}_{\rho 1}\widehat{gr}_{i2} + \hat{\omega}_{\rho 1} y_2 \widehat{gr}_{i2}) \tag{17}$$

- Two df test of null that y_2 is exogenous.

Multinomial Responses

• Recent push by Petrin and Train (2010), among others, to use control function methods where the second step estimation is something simple – such as multinomial logit, or nested logit – rather than being derived from a structural model. So, if we have reduced forms

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2, \tag{18}$$

then we jump directly to convenient models for $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$. The average structural functions are obtained by averaging the response probabilities across $\hat{\mathbf{v}}_{i2}$.

- Can use the same approach when we have a vector of shares, say $\mathbf{y}_1$, adding up to unity. (Nam and Wooldridge, 2012.) The multinomial distribution is in the linear exponential family.
- No generally acceptable way to handle discrete $\mathbf{y}_2$, except by specifying a full set of distributions.
- Might approximate by adding generalized residuals as control functions to standard models (such as MNL).

Exponential Models

- IV and CF approaches available for exponential models. For $y_1 \geq 0$ (could be a count) write

$$E(y_1|y_2,\mathbf{z},r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \tag{19}$$

where q_1 is the omitted variable independent of $\mathbf{z}$. $\mathbf{x}_1$ can be any function of $(y_2,\mathbf{z}_1)$.

- CF method can be based on

$$E(y_1|y_2,\mathbf{z}) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)E[\exp(q_1)|y_2,\mathbf{z}]. \tag{20}$$

• For continuous y_2, can find $E[\exp(q_1)|y_2, \mathbf{z}]$ when $D(y_2|\mathbf{z})$ is homoskedastic normal (Wooldridge, 1997) and when $D(y_2|\mathbf{z})$ follows a probit (Terza, 1998).

• In the probit case,

$$E(y_1|y_2, \mathbf{z}) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1) \tag{21}$$

$$\begin{aligned} h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1) = {} & \exp(\theta_1^2/2)\{y_2\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2) \\ & + (1 - y_2)[1 - \Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}. \end{aligned} \tag{22}$$

• Can use two-step NLS, where $\hat{\boldsymbol{\pi}}_2$ is obtained from probit. If y_1 is count, use a QMLE in the linear exponential family, such as Poisson or geometric.

- Can show the VAT score test is obtained from the mean function

$$\exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \eta_1 \widehat{gr}_i) \tag{23}$$

where

$$\widehat{gr}_{i2} = y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$$

- Convenient to use Poisson QMLE. Computationally very simple. At a minimum might as well test $H_0 : \eta_1 = 0$ first.

• As in binary/fractional case, adding the GR to the exponential mean might account for endogeneity, too.

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \exp(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_1 + \hat{\eta}_1 \widehat{gr}_{i2})$$

$$= \left[N^{-1} \sum_{i=1}^{N} \exp(\hat{\eta}_1 \widehat{gr}_{i2}) \right] \exp(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_1)$$

• Add $y_{i2}\widehat{gr}_{i2}$ for a swithing regression version:

$$\text{“}E(y_{i1}|y_{i2}, \mathbf{z}_i) = \exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \eta_1 \widehat{gr}_i + \omega_1 y_{i2} \widehat{gr}_{i2}) \quad (24)$$

- IV methods that work for any $\mathbf{y}_2$ without distributional assumptions are available [Mullahy (1997)]. If

$$E(y_1|\mathbf{y}_2, \mathbf{z}, q_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1) \tag{25}$$

and q_1 is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(q_1)|\mathbf{z}] = 1, \tag{26}$$

where $E[\exp(q_1)] = 1$ is a normalization. The moment conditions are

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0. \tag{27}$$

- Requires nonlinear IV methods. How to approximate the optimal instruments?

Quantile Regression

- Suppose

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + u_1, \tag{28}$$

where y_2 is endogenous and $\mathbf{z}$ is exogenous, with $\mathbf{z}_1 \subset \mathbf{z}$.

- Amemiya's (1982) two-stage LAD estimator is a plug-in estimator.

Reduced form for y_2,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2. \tag{29}$$

First step applies OLS or LAD to (29), and gets fitted values,

$y_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$. These are inserted for y_{i2} to give LAD of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$.

2SLAD relies on symmetry of the composite error $\alpha_1 v_2 + u_1$ given $\mathbf{z}$.

• If $D(u_1, v_2|\mathbf{z})$ is "centrally symmetric" can use a control function approach, as in Lee (2007). Write

$$u_1 = \rho_1 v_2 + e_1, \tag{30}$$

where e_1 given $\mathbf{z}$ would have a symmetric distribution. Get LAD residuals $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i \hat{\boldsymbol{\pi}}_2$ and do LAD of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$. Use t test on $\hat{v}_{i2}$ to test null that y_2 is exogenous.

• Interpretation of LAD in context of omitted variables is difficult unless lots of symmetry assumed.

• See Lee (2007) for discussion of general quantiles.

4. "Special Regressor" Methods for Binary Response

- Lewbel (2000) showed how to semi-parametrically estimate parameters in binary response models if a regressor with certain properties is available. Dong and Lewbel (2012) have recently relaxed those conditions somewhat.
- Let y_1 be a binary response:

$$\begin{aligned} y_1 &= 1[w_1 + \mathbf{y}_2\boldsymbol{\alpha}_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + u_1 > 0] \\ &= 1[w_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0] \end{aligned} \tag{31}$$

where w_1 is the "special regressor" normalized to have unity coefficient and assumed to be continuously distributed.

• In willingness-to-pay applications, $w_1 = -cost$, where $cost$ is the amoung that a new project will cost. Then someone prefers the project if

$$y_1 = 1[wtp > cost]$$

• In studies that elicit WTP, $cost$ is often set completely exogenously: independent of everything else, including $\mathbf{y}_2$.

• Dong and Lewbel (2012) assume w_1, like $\mathbf{z}_1$, is exogenous in (31), and that there are suitable instruments:

$$E(w_1' u_1) = 0, E(\mathbf{z}' u_1) = \mathbf{0} \tag{32}$$

• Need usual rank condition if we had linear model without w_1: *rank* $E(\mathbf{z}'\mathbf{x}_1) = K_1$.

• Setup is more general but they also write a linear equation

$$w_1 = \mathbf{y}_2\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + r_1 \tag{33}$$
$$E(\mathbf{y}_2'r_1) = \mathbf{0},\ E(\mathbf{z}'r_1) = 0$$

and then require (at a minimum)

$$E(r_1u_1) = 0. \tag{34}$$

• Condition (34), along with previous assumptions, means w_1 must be excluded from the reduced form for y_2 (which is a testable restriction).

- To see this, multiply (33) by u_1, take expectations, impose exogeneity on w_1 and $\mathbf{z}$, and use (31):

$$E(u_1 w_1) = E(u_1 \mathbf{y}_2)\boldsymbol{\pi}_1 + E(u_1 \mathbf{z})\boldsymbol{\pi}_2 + E(u_1 r_1)$$

or

$$0 = E(u_1 \mathbf{y}_2)\boldsymbol{\pi}_1 \qquad (35)$$

For this equation to hold except by fluke we need $\boldsymbol{\pi}_1 = \mathbf{0}$ (and in the case of a scalar y_2 this is the requirement). From (33) this means $\mathbf{y}_2$ and w_1 are uncorrelated after $\mathbf{z}$ has been partialled out. This implies w_1 does not appear in the reduced form for y_2 once $\mathbf{z}$ is included.

- Can easily test the Dong-Lewbel identification assumption on the special regressor. Can hold if w_1 depends on $\mathbf{z}$ provided that w_1 is independent of $\mathbf{y}_2$ conditional on $\mathbf{z}$.
- In WTP studies, means we can allow w_1 to depend on $\mathbf{z}$ but not $\mathbf{y}_2$.
- Dong and Lewbel application: y_1 is decision to migrate, y_2 is home ownership dummy. The special regressor is *age*. But does *age* really have no partial effect on home ownership given the other exogenous variables?

• If the assumptions hold, D-L show that, under regularity conditions (including wide support for w_1),

$$\begin{aligned} s_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + e_1 \\ E(\mathbf{z}'e_1) &= 0 \end{aligned} \tag{36}$$

where

$$s_1 = \frac{(y_1 - 1[w_1 \geq 0])}{f(w_1|\mathbf{y}_2, \mathbf{z})} \tag{37}$$

where $f(\cdot|\mathbf{y}_2, \mathbf{z})$ is the density of w_1 given $(\mathbf{y}_2, \mathbf{z})$.

- Estimate this density by MLE or nonparametrics:

$$\hat{s}_{i1} = \frac{(y_{i1} - 1[w_{i1} \geq 0])}{\hat{f}(w_{i1}|\mathbf{y}_{i2}, \mathbf{z}_i)}$$

- Requirement that $f(\cdot|\mathbf{y}_2, \mathbf{z})$ is continuous means the special regressor must appear additively and nowhere else. So no quadratics or interactions.

Semiparametric and Nonparametric Control Function Methods

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. The Average Structural Function
2. Nonparametric Estimation Approaches
3. Semiparametric Approaches
4. Parametric Approximations, Reconsidered

1. The Average Structural Function

- In nonlinear models it can be counterproductive to focus on parameters. Sometimes parameters cannot be identified but average partial effects can.
- Example: Suppose

$$P(y = 1|\mathbf{x}, q) = \Phi(\mathbf{x}\boldsymbol{\beta} + q) \quad (1)$$
$$q \sim Normal(0, \sigma_q^2)$$

- Even if we assume q is independent of $\mathbf{x}$, $\boldsymbol{\beta}$ is not identified. But $\boldsymbol{\beta}_q = \boldsymbol{\beta}/\sqrt{1 + \sigma_q^2}$ is. These scaled parameters index the average partial effects.

• In fact, $\boldsymbol{\beta}_q$ appears in the average structural function:

$$ASF(\mathbf{x}) = E_q[\Phi(\mathbf{x}\boldsymbol{\beta} + q)] = \Phi\left(\mathbf{x}\boldsymbol{\beta}/\sqrt{1+\sigma_q^2}\right). \tag{2}$$

• $\boldsymbol{\beta}_q$ is exactly what is estimated from probit of y on $\mathbf{x}$ when $D(q|\mathbf{x}) = D(q)$.

• Blundell and Powell (2003) define the notion of the ASF in a very general setting.

$$y_1 = g_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{u}_1) \equiv g_1(\mathbf{x}_1, \mathbf{u}_1) \tag{3}$$

where $\mathbf{u}_1$ is a vector of unobservables.

• The ASF averages out the unobservables for given values of $(\mathbf{y}_2, \mathbf{z}_1)$:

$$ASF_1(\mathbf{y}_2, \mathbf{z}_1) = \int g_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{u}_1) dF_1(\mathbf{u}_1), \quad (4)$$

where F_1 is the distribution of $\mathbf{u}_1$.

• Notice that $\mathbf{y}_2$ and $\mathbf{z}_1$ are treated symetrically in the definition. Endogeneity of $\mathbf{y}_2$ is irrelvant for the definition.

• Typically approach: Parameterize $g_1(\cdot)$, make distributional assumptions about $\mathbf{u}_1$, make identification assumptions (including restrictions on $D(\mathbf{y}_2|\mathbf{z})$.

- Sometimes useful to start with a weaker assumption:

$$E(y_1|\mathbf{y}_2,\mathbf{z}_1,\mathbf{q}_1) = g_1(\mathbf{y}_2,\mathbf{z}_1,\mathbf{q}_1) \tag{5}$$

Allows more natural treatment of models for counts, fractional reponses when only conditional means are specified.

- Can write as

$$\begin{aligned} y_1 &= g_1(\mathbf{y}_2,\mathbf{z}_1,\mathbf{q}_1) + e_1 \\ \mathbf{u}_1 &= (\mathbf{q}_1, e_1) \end{aligned} \tag{6}$$

but may only wish to maintain $E(e_1|\mathbf{y}_2,\mathbf{z}_1,\mathbf{q}_1) = 0$ (and not stronger forms of independence).

- Key insight of Blundell and Powell. Suppose $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2 \tag{7}$$

$$(\mathbf{u}_1, \mathbf{v}_2) \text{ is independent of } \mathbf{z} \tag{8}$$

- Next, define

$$\begin{aligned} E(y_1|\mathbf{y}_2, \mathbf{z}) &= E(y_1|\mathbf{v}_2, \mathbf{z}) \equiv h_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2) \\ &= \int g_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{u}_1) dG_1(\mathbf{u}_1|\mathbf{v}_2) \end{aligned} \tag{9}$$

- Using iterated expectations,

$$ASF(\mathbf{y}_2, \mathbf{z}) = E_{\mathbf{v}_2}[h_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2)] \tag{10}$$

- To identify $ASF(\mathbf{y}_2, \mathbf{z})$ we can shift attention from $g_1(\bullet)$ to $h_1(\bullet)$, and the latter depends on (effectively) observed variables: $\mathbf{v}_2 = \mathbf{y}_2 - \mathbf{g}_2(\mathbf{z})$.
- Wooldridge (2011) makes the argument slightly more general. Start with the "structural" conditional mean specification

$$E(y_1|\mathbf{y}_2, \mathbf{z}, \mathbf{q}_1) = g_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{q}_1).$$

- Suppose that for $\mathbf{r}_2 = \mathbf{k}_2(\mathbf{y}_2, \mathbf{z})$ we assume

$$D(\mathbf{q}_1|\mathbf{y}_2, \mathbf{z}) = D(\mathbf{q}_1|\mathbf{r}_2) \tag{11}$$

for a vector of "generalized residuals" $\mathbf{r}_2$, which we assume can be estimated.

- We can still recover the ASF:

$$E(y_1|\mathbf{y}_2, \mathbf{z}_1, \mathbf{r}_2) = \int g_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{q}_1) dF_1(\mathbf{q}_1|\mathbf{r}_2) \equiv h_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{r}_2) \tag{12}$$

$$ASF(\mathbf{y}_2, \mathbf{z}_1) = E_{\mathbf{r}_2}[h_1(\mathbf{y}_2, \mathbf{z}_1, \mathbf{r}_2)] \tag{13}$$

- Where might $\mathbf{r}_2$ come from if not an additive, reduced form error? Perhaps generalized residuals when $\mathbf{y}_2$ is discrete, or even standardized residuals if heteroskedasticity is present in a reduced form.
- When $\mathbf{y}_2$ is discrete (11) is nonstandard. Typically the assumption is made on the underlying continuous variables in a discrete response model.
- Generally, when $\mathbf{y}_2$ is continuous and has wide support we have many good options to choose from. Much harder when $\mathbf{y}_2$ is discrete.

- Focus on the ASF can have some surprising implications. Suppose

$$y = 1[\mathbf{x}\boldsymbol{\beta} + u > 0]$$
$$u|\mathbf{x} \sim Normal[0, \exp(\mathbf{x}_1\boldsymbol{\gamma})]$$

where $\mathbf{x}_1$ excludes an intercept. This is the so-called "heteroskedastic probit" model.

- The response probability is

$$P(y = 1|\mathbf{x}) = \Phi[\exp(-\mathbf{x}_1\boldsymbol{\gamma}/2)\mathbf{x}\boldsymbol{\beta}]; \tag{14}$$

all parameters identified. Can use MLE.

- The partial derivatives of $P(y = 1|\mathbf{x})$ are complicated; not proportional to β_j.
- The partial effects on the ASF are proportional to the β_j:

$$ASF(\mathbf{x}) = 1 - G(-\mathbf{x}\boldsymbol{\beta}) \tag{15}$$

where $G(\cdot)$ is the unconditional distribution of u.

- Is the focus on the ASF “superior” in such examples (there are a lot of them)? Maybe, but we cannot really tell the difference between heteroskedasticity in u and random coefficients,

$$y_i = 1[\mathbf{x}_i\mathbf{b}_i > 0]$$

with $\mathbf{b}_i$ independent of $\mathbf{x}_i$.

2. Nonparametric Estimation Approaches

- In Blundell and Powell's (2003) general setup, use a two-step CF approach. In the first step, the function $\mathbf{g}_2(\bullet)$ is estimated:

$$\begin{aligned} \mathbf{y}_2 &= \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2 \\ E(\mathbf{v}_2|\mathbf{z}) &= \mathbf{0} \end{aligned} \tag{16}$$

- Can use kernel regression or or series estimation or impose, say, index restrictions.
- Need the residuals,

$$\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \hat{\mathbf{g}}_2(\mathbf{z}_i). \tag{17}$$

• In the second step, use nonparametric regression of y_{i1} on $(\mathbf{x}_{i1}, \hat{\mathbf{v}}_{i2})$ to obtain $\hat{h}_1(\bullet)$.

• The ASF is consistently estimated as

$$\widehat{ASF}(\mathbf{x}_1) = N^{-1} \sum_{i=1}^{N} \hat{h}_1(\mathbf{x}_1, \hat{\mathbf{v}}_{i2}) \tag{18}$$

• Need to choose bandwidths in kernels or rates in series suitably.

• Inference is generally difficult. With series, can treat as flexible parametric models that are misspecified for any particular N. Ackerberg, Chen, Hahn (2009, Review of Economics and Statistics).

• Note how $\mathbf{g}_1(\bullet)$ is not estimated (and is not generally identified).

Quantile Structural Function

• Like Blundell and Powell (2003), Imbens and Newey (2006) consider a triangular system.

• As before, the structural equation is

$$y_1 = g_1(y_2, \mathbf{z}_1, \mathbf{u}_1).$$

• Now the reduced form need not have an additive error but needs to satisfy monoticity in the error:

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

where $g_2(\mathbf{z}, \bullet)$ is strictly monotonic.

• Monotoncity rules out discrete y_2 but allows some interaction between the single unobserved heterogeneity in y_2 and the exogenous variables.

• One useful result: Imbens and Newey show that, if $(\mathbf{u}_1, e_2)$ is independent of $\mathbf{z}$, then a valid control function that can be used in a second stage is $v_2 \equiv F_{y_2|\mathbf{z}}(y_2, \mathbf{z})$, where $F_{y_2|\mathbf{z}}$ is the conditional distribution of y_2 given $\mathbf{z}$.

• One can use parametric or nonparametric estimates $\hat{v}_{i2} = \hat{F}_{y_2|\mathbf{z}}(y_{i2}, \mathbf{z}_i)$ in a second-step nonparametric estimation, and then average to get the ASF.

- Imbens and Newey described identification of other quantities of interest, including the quantile structural function. When u_1 is a scalar and monotonically increasing in u_1, the QSF is

$$QSF_\tau(\mathbf{x}_1) = g_1(\mathbf{x}_1, Quant_\tau(u_1)),$$

where $Quant_\tau(u_1)$ is the τ^{th} quantile of u_1.

3. Semiparametric Approaches

- Full nonparametric estimation can lead to the "curse of dimensionality," especially if the dimensions of $\mathbf{z}$ and/or $\mathbf{y}_2$ are large. Semiparametric approaches can help.

- Blundell and Powell (2004) show how to relax distributional assumptions on (u_1, v_2) in the specification

$$y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0] \tag{19}$$

$$y_2 = g_2(\mathbf{z}) + v_2 \tag{20}$$

$$(u_1, v_2) \text{ is independent of } z \tag{21}$$

where $\mathbf{x}_1$ can be any function of $(y_2, \mathbf{z}_1)$. (19) is semiparametric because no distributional assumptions are made on u_1.

- Under these assumptions,

$$P(y_1 = 1|\mathbf{z}, v_2) = E(y_1|\mathbf{z}, v_2) = H_1(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{22}$$

for some (generally unknown) function $H_1(\cdot,\cdot)$. The average structural function is just $ASF(\mathbf{x}_1) = E_{v_{i2}}[H_1(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$.

- 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_1 and $\boldsymbol{\beta}_1$ (up to scale) and $G_1(\cdot)$, the distribution of u_1.

- Estimated ASF is obtained from $\hat{G}_1(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1)$ or

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

• In some cases, an even more parametric approach suggests itself. Suppose we have the exponential regression

$$E(y_1|y_2, \mathbf{z}, q_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \qquad (24)$$

where q_1 is the unobservable.

• If $y_2 = \mathbf{g}_2(\mathbf{z})\boldsymbol{\pi}_2 + v_2$ and (q_1, v_2) is independent of $\mathbf{z}$, then

$$E(y_1|y_2, \mathbf{z}_1, v_2) = h_2(v_2)\exp(\mathbf{x}_1\boldsymbol{\beta}_1), \qquad (25)$$

where now $h_2(\cdot) > 0$ is an unknown function. It can be approximated using a sieve with an log link function to ensure nonnegativity. First-stage residuals $\hat{v}_2$ replace v_2.

- To handle certain cases where $\mathbf{y}_2$ is discrete, Wooldridge (2011) suggests making the model for $\mathbf{y}_2$ parametric or semiparametric, leaving $E(y_1|\mathbf{y}_2, \mathbf{z}_1, \mathbf{q}_1)$ unspecified.
- Suppose $\mathbf{r}_2$ is a vector of estimable "generalized residuals" – $\mathbf{r}_2 = \mathbf{k}_2(\mathbf{y}_2, \mathbf{z}, \boldsymbol{\theta}_2)$ for known function $\mathbf{k}_2(\cdot)$ and identified parameters $\boldsymbol{\theta}_2$ – and we are willing to assume $\mathbf{r}_2$ acts as a "sufficient statistic" for endogeneity of $\mathbf{y}_2$:

$$D(\mathbf{q}_1|\mathbf{y}_2, \mathbf{z}) = D(\mathbf{q}_1|\mathbf{r}_2).$$

• We can use nonparametric, semiparametric, or flexible parametric approaches to estimate

$$E(y_1|\mathbf{y}_2, \mathbf{z}_1, \mathbf{r}_2) = h_1(\mathbf{x}_1, \mathbf{r}_2)$$

in a second stage by inserting $\hat{\mathbf{r}}_{i2}$ in place of $\mathbf{r}_{i2}$. The $\hat{\mathbf{r}}_{i2}$ would often come from an MLE.

• Smoothness in $\mathbf{r}_2$ is critical, and it must vary enough separately from $(\mathbf{y}_2, \mathbf{z}_1)$.

• As before,

$$\widehat{ASF}(\mathbf{x}_1) = N^{-1}\sum_{i=1}^{N}\hat{h}_1(\mathbf{x}_1, \hat{\mathbf{r}}_{i2})$$

• Suppose y_2 is binary. We might model y_2 as flexible probit, or heteroskedatic probit. In the probit case $\hat{r}_{i2}$ are the GRs; an extension holds for heteroskedastic probit.

• We could use full nonparametric in the second stage or assume something like

$$E(y_1|\mathbf{x}_1, r_2) = H_1(\mathbf{x}_1\boldsymbol{\beta}_1, r_2)$$

similar to BP (2004).

4. Parametric Approximations, Reconsidered

- One implication of the Blundell and Powell approach: It is liberating even if we focus on parametric analysis at one or both stages. The problem is reduced to getting a good approximation to $E(y_1|\mathbf{y}_2,\mathbf{z}_1,\mathbf{v}_2)$ and a reliable way to obtain residuals $\hat{\mathbf{v}}_{i2}$.
- Sometimes flexible parametric may be preferred to obtain more precise estimators and make computation simpler.

- Example: Suppose we start with

$$
\begin{aligned}
y_1 &= 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0] \\
y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + e_2 \\
e_2 &= \sqrt{h_2(\mathbf{z})}\, v_2
\end{aligned}
$$

where (u_1, v_2) is independent of $\mathbf{z}$ and $h_2(\mathbf{z}) > 0$ is a heteroskedasticity function.

- Under joint normality can write

$$u_1 = \rho_1 v_2 + a_1$$

where a_1 is independent of $(v_2, \mathbf{z})$ (and therefore $\mathbf{x}_1$).

• The control function is

$$\hat{v}_{i2} = \hat{e}_{i2}/\sqrt{\hat{h}_2(\mathbf{z}_i)}$$

and this can be used in a Blundell-Powell analysis or just a flexible probit. Typically $\hat{h}_2(\mathbf{z}_i)$ would be a flexible exponential function.

• Because we know that the ASF can be obtained from averaging v_2 out of $E(y_1|y_2, \mathbf{z}_2, v_2)$, we can use a very flexible parametric model in the second stage. For example, estimate the equation

$$\Phi(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \rho_1\hat{v}_{i2} + \eta_1\hat{v}_{i2}^2 + \hat{v}_{i2}\mathbf{x}_{i1}\boldsymbol{\omega}_1)$$

and then, for fixed $\mathbf{x}_1$, average out $\hat{v}_{i2}$.

• Might even use a "hetprobit" model where $\hat{v}_{i2}$ can appear in the variance.

• All works when y_1 is fractional, too.

• Generally, if $\mathbf{v}_2$ is the control function, use models in the second stage that reflect the nature of y_1, and use sensible (but robust) estimation methods. If y_1 is a count, and y_2 a scalar, might use Poisson regression with mean function

$$\exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \rho_1\hat{v}_{i2} + \eta_1\hat{v}_{i2}^2 + \hat{v}_{i2}\mathbf{x}_{i1}\boldsymbol{\omega}_1)$$

- Even if y_{i2} is discrete, or we have a vector $\mathbf{y}_{i2}$, we might obtain generalized residuals from MLE estimation and use similar schemes.
- At a minimum, flexible parametric approaches are simple ways to allow sensitivity analysis. Also the tests for exogeneity are valid quite generally.
- Simple empirical example: Women's LFP and fertility. Maybe a more serious nonparametric analysis is needed.

```
. use labsup

. probit morekids age agesq nonmomi educ samesex

Probit regression                               Number of obs   =      31857
                                                LR chi2(5)      =    2365.72
                                                Prob > chi2     =     0.0000
Log likelihood = -20893.576                     Pseudo R2       =     0.0536

------------------------------------------------------------------------------
    morekids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1190289   .0307712     3.87   0.000     .0587185    .1793393
       agesq |  -.0010264   .0005285    -1.94   0.052    -.0020623    9.44e-06
     nonmomi |  -.0028068   .0003653    -7.68   0.000    -.0035228   -.0020908
        educ |  -.0882257   .0023305   -37.86   0.000    -.0927935    -.083658
     samesex |   .1458026   .0143639    10.15   0.000       .11765    .1739552
       _cons |  -1.652074   .4418017    -3.74   0.000    -2.517989   -.7861586
------------------------------------------------------------------------------

. predict zd2, xb

. gen gr2 = morekids*normalden(zd2)/normal(zd2) -
                   (1 - morekids)*normalden(-zd2)/normal(-zd2)

. sum gr2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         gr2 |     31857    1.42e-10    .7802979  -1.854349   1.638829
```

```
. probit worked morekids age agesq nonmomi educ gr2

Probit regression                                 Number of obs   =      31857
                                                  LR chi2(6)      =    2069.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -20530.203                       Pseudo R2       =     0.0480

------------------------------------------------------------------------------
      worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.7692097   .2375535    -3.24   0.001    -1.234806   -.3036134
         age |   .1694555   .0324621     5.22   0.000     .1058309    .2330802
       agesq |  -.0022156   .0005353    -4.14   0.000    -.0032648   -.0011663
     nonmomi |  -.0047247   .0004426   -10.67   0.000    -.0055922   -.0038572
        educ |   .0614195   .0081478     7.54   0.000     .0454501     .077389
         gr2 |   .2985435    .146857     2.03   0.042      .010709     .586378
       _cons |  -2.961497   .4402391    -6.73   0.000     -3.82435   -2.098645
------------------------------------------------------------------------------

. margeff

Average marginal effects on Prob(worked==1) after probit

------------------------------------------------------------------------------
      worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.2769055   .0764453    -3.62   0.000    -.4267356   -.1270754
         age |   .0624673    .011951     5.23   0.000     .0390437    .0858909
       agesq |  -.0008167   .0001972    -4.14   0.000    -.0012033   -.0004302
     nonmomi |  -.0017417   .0001623   -10.73   0.000    -.0020598   -.0014236
        educ |   .0226414   .0029956     7.56   0.000     .0167702    .0285126
         gr2 |   .1100537   .0541261     2.03   0.042     .0039684    .2161389
------------------------------------------------------------------------------


. probit worked morekids age agesq nonmomi educ

Probit regression                                 Number of obs   =      31857
```

```
                                                  LR chi2(5)      =    2065.65
                                                  Prob > chi2     =     0.0000
Log likelihood =  -20532.27                       Pseudo R2       =     0.0479

------------------------------------------------------------------------------
      worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.2872582   .0149444   -19.22   0.000    -.3165487   -.2579677
         age |   .1478441   .0306711     4.82   0.000     .0877298    .2079583
       agesq |  -.0020299   .0005275    -3.85   0.000    -.0030637   -.0009961
     nonmomi |  -.0042178   .0003656   -11.54   0.000    -.0049343   -.0035012
        educ |   .0772888   .0023425    32.99   0.000     .0726977    .0818799
       _cons |  -2.912884   .4395955    -6.63   0.000    -3.774475   -2.051293
------------------------------------------------------------------------------

. margeff

Average marginal effects on Prob(worked==1) after probit

------------------------------------------------------------------------------
      worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.1070784   .0055612   -19.25   0.000    -.1179783   -.0961786
         age |   .0545066    .011295     4.83   0.000     .0323687    .0766445
       agesq |  -.0007484   .0001944    -3.85   0.000    -.0011293   -.0003674
     nonmomi |   -.001555    .000134   -11.61   0.000    -.0018175   -.0012924
        educ |   .0284945   .0008184    34.82   0.000     .0268905    .0300986
------------------------------------------------------------------------------



. reg morekids age agesq nonmomi educ samesex

      Source |       SS       df       MS              Number of obs =   31857
-------------+------------------------------           F(  5, 31851) =  490.14
       Model |  568.823401     5  113.76468            Prob > F      =  0.0000
    Residual |  7392.85001 31851  .232107313           R-squared     =  0.0714
```

```
-------------+------------------------------           Adj R-squared =  0.0713
       Total |  7961.67342 31856  .249926966           Root MSE      =  .48178

------------------------------------------------------------------------------
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0440951   .0114907     3.84   0.000      .021573    .0666173
       agesq |  -.0003733   .0001975    -1.89   0.059    -.0007604    .0000137
    nonmomi |  -.0010596   .0001372    -7.72   0.000    -.0013285   -.0007907
        educ |  -.0328862   .0008425   -39.03   0.000    -.0345376   -.0312348
     samesex |   .0549188   .0053988    10.17   0.000      .044337    .0655006
       _cons |  -.1173924   .1649003    -0.71   0.477    -.4406034    .2058187
------------------------------------------------------------------------------

. predict v2, resid
```

```
. reg worked morekids age agesq nonmomi educ v2, robust

Linear regression                                      Number of obs =   31857
                                                       F(  6, 31850) =  386.10
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0634
                                                       Root MSE      =  .47607

------------------------------------------------------------------------------
             |               Robust
      worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.2134551   .0971395    -2.20   0.028    -.4038523   -.0230578
         age |    .062004   .0122329     5.07   0.000      .038027    .0859811
       agesq |  -.0008351   .0001998    -4.18   0.000    -.0012268   -.0004435
     nonmomi |  -.0016831   .0001729    -9.74   0.000    -.0020219   -.0013443
        educ |   .0253881   .0033018     7.69   0.000     .0189164    .0318598
          v2 |   .1066713   .0973191     1.10   0.273    -.0840778    .2974204
       _cons |  -.6258267    .164867    -3.80   0.000    -.9489724   -.3026811
------------------------------------------------------------------------------
```

```
. reg worked morekids age agesq nonmomi educ, robust

Linear regression                                      Number of obs =   31857
                                                       F(  5, 31851) =  463.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0634
                                                       Root MSE      =  .47607

------------------------------------------------------------------------------
             |               Robust
      worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.1071292   .0055642   -19.25   0.000    -.1180352   -.0962232
         age |   .0572601     .01145     5.00   0.000     .0348177    .0797025
       agesq |  -.0007945   .0001964    -4.05   0.000    -.0011794   -.0004095
     nonmomi |  -.0015715   .0001399   -11.23   0.000    -.0018457   -.0012973
        educ |   .0288871   .0008471    34.10   0.000     .0272268    .0305473
       _cons |  -.6154823   .1646188    -3.74   0.000    -.9381415   -.2928231
------------------------------------------------------------------------------
```

```
. biprobit (worked morekids age agesq nonmomi educ) (morekids = age agesq nonmomi educ samesex)

Comparison:    log likelihood = -41425.846

Fitting full model:

Seemingly unrelated bivariate probit              Number of obs   =      31857
                                                  Wald chi2(10)   =    4547.96
Log likelihood = -41423.859                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
worked       |
    morekids |  -.7217122   .1996235    -3.62   0.000    -1.112967   -.3304573
         age |   .1640598   .0309654     5.30   0.000     .1033687    .2247509
       agesq |  -.0021513   .0005238    -4.11   0.000     -.003178   -.0011245
     nonmomi |  -.0045819   .0003826   -11.98   0.000    -.0053318    -.003832
        educ |   .0610621   .0087203     7.00   0.000     .0439706    .0781535
       _cons |  -2.888228   .4378617    -6.60   0.000    -3.746421   -2.030035
-------------+----------------------------------------------------------------
morekids     |
         age |   .1194871   .0307611     3.88   0.000     .0591965    .1797777
       agesq |  -.0010345   .0005284    -1.96   0.050    -.0020701    1.10e-06
     nonmomi |   -.002818   .0003658    -7.70   0.000    -.0035349   -.0021011
        educ |  -.0884098   .0023329   -37.90   0.000    -.0929822   -.0838374
     samesex |   .1443195   .0144174    10.01   0.000     .1160619     .172577
       _cons |  -1.654577   .4416112    -3.75   0.000    -2.520119   -.7890352
-------------+----------------------------------------------------------------
     /athrho |   .2809234   .1385037     2.03   0.043     .0094611    .5523856
-------------+----------------------------------------------------------------
         rho |   .2737595   .1281236                      .0094608    .5023061
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =  3.97332    Prob > chi2 = 0.0462


. probit worked morekids age agesq nonmomi educ gr2 gr2morekids
```

```
Probit regression                                 Number of obs   =      31857
                                                  LR chi2(7)      =    2085.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -20522.231                       Pseudo R2       =     0.0484

------------------------------------------------------------------------------
      worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -.6711282   .2387995    -2.81   0.005    -1.139167   -.2030898
         age |   .1704885   .0324784     5.25   0.000      .106832    .2341449
       agesq |  -.0022739   .0005358    -4.24   0.000     -.003324   -.0012239
     nonmomi |  -.0046295   .0004433   -10.44   0.000    -.0054984   -.0037607
        educ |   .0656273   .0082146     7.99   0.000     .0495269    .0817276
         gr2 |   .3796138   .1482436     2.56   0.010     .0890617     .670166
 gr2morekids |  -.2779973   .0696335    -3.99   0.000    -.4144764   -.1415181
       _cons |  -2.932433   .4405598    -6.66   0.000    -3.795914   -2.068951
------------------------------------------------------------------------------
```

Panel Data Models with Heterogeneity and Endogeneity

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. Introduction
2. General Setup and Quantities of Interest
3. Assumptions with Neglected Heterogeneity
4. Models with Heterogeneity and Endogeneity
5. Estimating Some Popular Models

1. Introduction

- When panel data models contain unobserved heterogeneity and omitted time-varying variables, control function methods can be used to account for both problems.
- Under fairly week assumptions can obtain consistent, asymptotically normal estimators of average structural functions – provided suitable instruments are available.
- Other issues with panel data: How to treat dynamics? Models with lagged dependent variables are hard to estimate when heterogeneity and other sources of endogeneity are present.

- Approaches to handling unobserved heterogeneity:

1. Treat as parameters to estimate. Can work well with large T but with small T can have incidental parameters problem. Bias adjustments are available for parameters and average partial effects. Usually weak dependence or even independence is assumed across the time dimension.

2. Remove heterogeneity to obtain an estimating equation. Works for simple linear models and a few nonlinear models (via conditional MLE or a quasi-MLE variant). Cannot be done in general. Also, may not be able to identify interesting partial effects.

- Correlated Random Effects: Mundlak/Chamberlain. Requires some restrictions on distribution of heterogeneity, although these can be nonparametric. Applies generally, does not impose restrictions on dependence over time, allows estimation of average partial effects. Can be easily combined with CF methods for endogeneity.
- Can try to establish bounds rather than estimate parameters or APEs. Chernozhukov, Fernández-Val, Hahn, and Newey (2009) is a recent example.

2. General Setup and Quantities of Interest

• Static, unobserved effects probit model for panel data with an omitted time-varying variable r_{it}:

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

What are the quantities of interest for most purposes?

(i) The element of $\boldsymbol{\beta}$, the β_j. These give the directions of the partial effects of the covariates on the response probability. For any two continuous covariates, the ratio of coefficients, β_j/β_h, is identical to the ratio of partial effects (and the ratio does not depend on the covariates or unobserved heterogeneity, c_i).

(ii) The magnitudes of the partial effects. These depend not only on the value of the covariates, say $\mathbf{x}_t$, but also on the value of the unobserved heterogeneity. In the continuous covariate case,

$$\frac{\partial P(y_t = 1|\mathbf{x}_t, c, r_t)}{\partial x_{tj}} = \beta_j \phi(\mathbf{x}_t \boldsymbol{\beta} + c + r_t). \tag{2}$$

• Questions: (a) Assuming we can estimate $\boldsymbol{\beta}$, what should we do about the unobservables (c, r_t)? (b) If we can only estimate $\boldsymbol{\beta}$ up-to-scale, can we still learn something useful about magnitudes of partial effects? (c) What kinds of assumptions do we need to estimate partial effects?

- Let $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$ be a random draw from the cross section. Suppose we are interested in

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i, \mathbf{r}_{it}) = m_t(\mathbf{x}_{it}, \mathbf{c}_i, \mathbf{r}_{it}). \tag{3}$$

$\mathbf{c}_i$ can be a vector of unobserved heterogeneity, $\mathbf{r}_{it}$ a vector of omitted time-varying variables.

- Partial effects: if x_{tj} is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}, \mathbf{r}_t) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c}, \mathbf{r}_t)}{\partial x_{tj}}, \tag{4}$$

or discrete changes.

• How do we account for unobserved $(\mathbf{c}_i, \mathbf{r}_{it})$? If we know enough about the distribution of $(\mathbf{c}_i, \mathbf{r}_{it})$ we can insert meaningful values for $(\mathbf{c}, \mathbf{r}_t)$. For example, if $\boldsymbol{\mu}_{\mathbf{c}} = E(\mathbf{c}_i)$, $\boldsymbol{\mu}_{\mathbf{r}_t} = E(\mathbf{r}_{it})$ then we can compute the partial effect at the average (PEA),

$$PEA_j(\mathbf{x}_t) = \theta_j(\mathbf{x}_t, \boldsymbol{\mu}_{\mathbf{c}}, \boldsymbol{\mu}_{\mathbf{r}_t}). \tag{5}$$

Of course, we need to estimate the function m_t and $(\boldsymbol{\mu}_{\mathbf{c}}, \boldsymbol{\mu}_{\mathbf{r}_t})$. If we can estimate the distribution of $(\mathbf{c}_i, \mathbf{r}_{it})$, or features in addition to its mean, we can insert different quantiles, or a certain number of standard deviations from the mean.

- Alternatively, we can obtain the average partial effect (APE) (or population average effect) by averaging across the distribution of $\mathbf{c}_i$:

$$APE(\mathbf{x}_t) = E_{(\mathbf{c}_i, \mathbf{r}_{it})}[\theta_j(\mathbf{x}_t, \mathbf{c}_i, \mathbf{r}_{it})]. \tag{6}$$

The difference between (5) and (6) can be nontrivial. In some leading cases, (6) is identified while (5) is not. (6) is closely related to the notion of the average structural function (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$ASF_t(\mathbf{x}_t) = E_{(\mathbf{c}_i, \mathbf{r}_{it})}[m_t(\mathbf{x}_t, \mathbf{c}_i, \mathbf{r}_{it})]. \tag{7}$$

- Passing the derivative through the expectation in (7) gives the APE.

3. Assumptions with Neglected Heterogeneity

Exogeneity of Covariates

- Cannot get by with just specifying a model for the contemporaneous conditional distribution, $D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$.
- The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \tag{8}$$

Chamberlain (1984) labeled (8) *strict exogeneity conditional on the unobserved effects* $\mathbf{c}_i$. Conditional mean version:

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

- The *sequential exogeneity* assumption is

$$D(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{it},\mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \tag{10}$$

Much more difficult to allow sequential exogeneity in in nonlinear models. (Most progress has been made for lagged dependent variables or specific functional forms, such as exponential.)

- Neither strict nor sequential exogeneity allows for contemporaneous endogeneity of one or more elements of $\mathbf{x}_{it}$, where, say, x_{itj} is correlated with unobserved, time-varying unobservables that affect y_{it}.

Conditional Independence

• In linear models, serial dependence of idiosyncratic shocks is easily dealt with, either by "cluster robust" inference or Generalized Least Squares extensions of Fixed Effects and First Differencing. With strictly exogenous covariates, serial correlation never results in inconsistent estimation, even if improperly modeled. The situation is different with most nonlinear models estimated by MLE.

• *Conditional independence* (CI) (under strict exogeneity):

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \qquad (11)$$

- In a parametric context, the CI assumption reduces our task to specifying a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$.
- In random effects and correlated random frameworks (next section), CI plays a critical role in being able to estimate the "structural" parameters and the parameters in the distribution of $\mathbf{c}_i$ (and therefore, in estimating PEAs). In a broad class of popular models, CI plays no essential role in estimating APEs.

Assumptions about the Unobserved Heterogeneity

Random Effects

- Generally stated, the key RE assumption is

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{12}$$

Under (12), the APEs are actually nonparametrically identified from

$$E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{13}$$

- In some leading cases (RE probit and RE Tobit with heterogeneity normally distributed), if we want PEs for different values of **c**, we must assume more: strict exogeneity, conditional independence, and (12) with a parametric distribution for $D(\mathbf{c}_i)$.

Correlated Random Effects

A CRE framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but restricted in some way. In a parametric setting, we specify a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$, as in Chamberlain (1980,1982), and much work since. Distributional assumptions that lead to simple estimation – homoskedastic normal with a linear conditional mean — can be restrictive.

- Possible to drop parametric assumptions and just assume

$$D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i), \tag{14}$$

without restricting $D(c_i|\bar{\mathbf{x}}_i)$. Altonji and Matzkin (2005, Econometrica).

- Other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ are possible.

• APEs are identified very generally. For example, under (14), a consistent estimate of the average structural function is

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N} q_t(\mathbf{x}_t, \bar{\mathbf{x}}_i), \tag{15}$$

where $q_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$.

• Need a random sample $\{\bar{\mathbf{x}}_i : i = 1, \ldots, N\}$ for the averaging out to work.

Fixed Effects

• The label "fixed effects" is used differently by different researchers. One view: $\mathbf{c}_i,\ i = 1,\ldots,N$ are parameters to be estimated. Usually leads to an "incidental parameters problem."

• Second meaning of "fixed effects": $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted and we look for objective functions that do not depend on $\mathbf{c}_i$ but still identify the population parameters. Leads to "conditional MLE" if we can find "sufficient statistics" $\mathbf{s}_i$ such that

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{s}_i). \tag{16}$$

• Conditional Independence is usually maintained.

• Key point: PEAs and APEs are generally unidentified.

4. Models with Heterogeneity and Endogeneity

- Let y_{it1} be a scalar response, $\mathbf{y}_{it2}$ a vector of endogenous variables, $\mathbf{z}_{it1}$ exogenous variables, and we have

$$E(y_{it1}|\mathbf{y}_{it2},\mathbf{z}_{it1},\mathbf{c}_{i1},\mathbf{r}_{it1}) = m_{t1}(\mathbf{y}_{it2},\mathbf{z}_{it1},\mathbf{c}_{i1},\mathbf{r}_{it1}) \tag{17}$$

- $\mathbf{y}_{it2}$ is allowed to be correlated with $\mathbf{r}_{it1}$ (as well as with $\mathbf{c}_{i1}$).
- The vector of exogenous variables $\{\mathbf{z}_{it} : t = 1,\ldots,T\}$ with $\mathbf{z}_{it1} \subset \mathbf{z}_{it}$ are strictly exogenous in the sense that

$$E(y_{it}|\mathbf{y}_{it2},\mathbf{z}_{i},\mathbf{c}_{i1},\mathbf{r}_{it1}) = E(y_{it}|\mathbf{y}_{it2},\mathbf{z}_{it1},\mathbf{c}_{i1},\mathbf{r}_{it1}) \tag{18}$$

$$D(\mathbf{r}_{it1}|\mathbf{z}_{i},\mathbf{c}_{i1}) = D(\mathbf{r}_{it1}) \tag{19}$$

- Sometimes we can eliminate $\mathbf{c}_i$ and obtain an equation that can be estimated by IV (linear, exponential). Generally not possible.
- Now a CRE approach involves modeling $D(\mathbf{c}_{i1}|\mathbf{z}_i)$.
- Generally, we need to model how $\mathbf{y}_{it2}$ is related to $\mathbf{r}_{it1}$.
- Control Function methods are convenient for allowing both.
- Suppose y_{it2} is a scalar and

$$\begin{aligned} y_{it2} &= m_{it2}(\mathbf{z}_{it}, \bar{\mathbf{z}}_i, \boldsymbol{\delta}_2) + v_{it2} \\ E(v_{it2}|\mathbf{z}_i) &= 0 \\ D(r_{it1}|v_{it2}, \mathbf{z}_i) &= D(r_{it1}|v_{it2}) \end{aligned} \tag{20}$$

• With suitable time-variation in the instruments, the assumptions in (20) allow identification of the ASF if we assume a model for

$$D(\mathbf{c}_{i1}|\mathbf{z}_i, v_{it2})$$

Generally, we can estimate

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, v_{it2}) = E(y_{it1}|y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, v_{it2}) \equiv g_{t1}(y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, v_{it2}) \tag{21}$$

• The ASF is now obtained by averaging out $(\bar{\mathbf{z}}_i, v_{it2})$:

$$ASF(y_{t2}, \mathbf{z}_{t1}) = E_{(\bar{\mathbf{z}}_i, v_{it2})}[g_{t1}(y_{t2}, \mathbf{z}_{t1}, \bar{\mathbf{z}}_i, v_{it2})]$$

• Most of this can be fully nonparametric (Altonji and Matzkin, 2005; Blundell and Powell, 2003) although some restriction is needed on $D(\mathbf{c}_{i1}|\mathbf{z}_i, v_{it2})$, such as

$$D(\mathbf{c}_{i1}|\mathbf{z}_i, v_{it2}) = D(\mathbf{c}_{i1}|\bar{\mathbf{z}}_i, v_{it2})$$

• With T sufficiently large we can add other features of $\{\mathbf{z}_{it} : t = 1, \ldots, T\}$ to $\bar{\mathbf{z}}_i$.

5. Estimating Some Popular Models

Linear Model with Endogeneity

- Simplest model is

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + u_{it1} \equiv \mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + u_{it1} \tag{22}$$

$$E(u_{it1}|\mathbf{z}_i, c_{i1}) = 0$$

- The fixed effects 2SLS estimator is common. Deviate variables from time averages to remove c_{i1} then apply IV:

$$\ddot{y}_{it1} = \ddot{\mathbf{x}}_{it1}\boldsymbol{\beta}_1 + \ddot{u}_{it1}$$

$$\ddot{\mathbf{z}}_{it} = \mathbf{z}_{it} - \bar{\mathbf{z}}_i$$

- Easy to make inference robust to serial correlation and heteroskedasticity in $\{u_{it1}\}$. ("Cluster-robust inference.")
- Test for (strict) exogeneity of $\{y_{it2}\}$:

(i) Estimate the reduced form of y_{it2} by usual fixed effects:

$$y_{it2} = \mathbf{z}_{it}\boldsymbol{\delta}_1 + c_{i2} + u_{it2}$$

Get the FE residuals, $\hat{\ddot{u}}_{it2} = \ddot{y}_{it2} - \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\delta}}_1$.

- Estimate the augment equation

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \rho_1 \hat{\ddot{u}}_{it2} + c_{i1} + error_{it} \tag{23}$$

by FE and use a cluster-robust test of $H_0 : \rho_1 = 0$.

• The random effects IV approach assumes c_{i1} is uncorrelated with $\mathbf{z}_i$, and nominally imposes serial independence on $\{u_{it1}\}$.

• Simple way to test the null whether REIV is sufficient. (Robust Hausman test comparing REIV and FEIV.)

Estimate

$$y_{it1} = \eta_1 + \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + u_{it1} \tag{24}$$

by REIV, using instruments $(1, \mathbf{z}_{it}, \bar{\mathbf{z}}_i)$. The estimator of $\boldsymbol{\beta}_1$ is the FEIV estimator.

• Test $H_0 : \boldsymbol{\xi}_1 = \mathbf{0}$, preferably using a fully robust test. A rejection is evidence that the IVs are correlated with c_i, and should use FEIV.

• Other than the rank condition, the key condition for FEIV to be consistent is that the instruments, $\{\mathbf{z}_{it}\}$, are strictly exogenous with respect to $\{u_{it}\}$. With $T \geq 3$ time periods, this is easily tested – as in the usual FE case.

• The augmented model is

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \mathbf{z}_{i,t+1}\boldsymbol{\psi}_1 + c_{i1} + u_{it1},\ t = 1,\ldots,T-1$$

and we estimate it by FEIV, using instruments $(\mathbf{z}_{it}, \mathbf{z}_{i,t+1})$.

• Use a fully robust Wald test of $H_0 : \boldsymbol{\psi}_1 = \mathbf{0}$. Can be selective about which leads to include.

Example: Estimating a Passenger Demand Function for Air Travel

$N = 1,149$, $T = 4$.

- Uses route concentration for largest carrier as IV for log(*fare*).

```
. use airfare

. * Reduced form for lfare; concen is the IV.

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

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

```
. xtivreg lpassen ldist ldistsq y98 y99 y00 (lfare = concen), re theta

G2SLS random-effects IV regression              Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

R-sq:  within  = 0.4075                         Obs per group: min =         4
       between = 0.0542                                        avg =       4.0
       overall = 0.0641                                        max =         4

                                                Wald chi2(6)       =    231.10
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
theta              = .91099494

------------------------------------------------------------------------------
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.5078762    .229698    -2.21   0.027     -.958076   -.0576763
       ldist |  -1.504806   .6933147    -2.17   0.030    -2.863678   -.1459338
     ldistsq |   .1176013   .0546255     2.15   0.031     .0105373    .2246652
         y98 |   .0307363   .0086054     3.57   0.000     .0138699    .0476027
         y99 |   .0796548     .01038     7.67   0.000     .0593104    .0999992
         y00 |   .1325795   .0229831     5.77   0.000     .0875335    .1776255
       _cons |   13.29643   2.626949     5.06   0.000     8.147709    18.44516
-------------+----------------------------------------------------------------
     sigma_u |  .94920686
     sigma_e |  .16964171
         rho |  .96904799   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Instrumented:   lfare
Instruments:    ldist ldistsq y98 y99 y00 concen
------------------------------------------------------------------------------

. * The quasi-time-demeaning parameter is quite large: .911 ("theta").
```

```
. xtivreg2 lpassen ldist ldistsq y98 y99 y00 (lfare = concen), fe cluster(id)
Warning - collinearities detected
Vars dropped:  ldist ldistsq

FIXED EFFECTS ESTIMATION
------------------------
Number of groups =      1149                    Obs per group: min =          4
                                                               avg =        4.0
                                                               max =          4

Number of clusters (id) = 1149                        Number of obs =      4596
                                                      F(  4,  1148) =     26.07
                                                      Prob > F      =    0.0000
Total (centered) SS     =  128.0991685                Centered R2   =    0.2265
Total (uncentered) SS   =  128.0991685                Uncentered R2 =    0.2265
Residual SS             =   99.0837238                Root MSE      =     .1695

------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.3015761   .6124127    -0.49   0.622    -1.501883    .8987307
         y98 |   .0257147   .0164094     1.57   0.117    -.0064471    .0578766
         y99 |   .0724166   .0250971     2.89   0.004     .0232272    .1216059
         y00 |   .1127914   .0620115     1.82   0.069    -.0087488    .2343316
------------------------------------------------------------------------------
Instrumented:         lfare
Included instruments: y98 y99 y00
Excluded instruments: concen
------------------------------------------------------------------------------
```

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

. xtivreg lpassen ldist ldistsq y98 y99 y00 concenb (lfare = concen), re theta

G2SLS random-effects IV regression              Number of obs      =      4596
Group variable: id                              Number of groups   =      1149

                                                Wald chi2(7)       =    218.80
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
theta              = .90084889

------------------------------------------------------------------------------
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.3015761   .2764376    -1.09   0.275    -.8433838    .2402316
       ldist |  -1.148781   .6970189    -1.65   0.099    -2.514913    .2173514
     ldistsq |   .0772565   .0570609     1.35   0.176    -.0345808    .1890937
         y98 |   .0257147   .0097479     2.64   0.008     .0066092    .0448203
         y99 |   .0724165   .0119924     6.04   0.000     .0489118    .0959213
         y00 |   .1127914   .0274377     4.11   0.000     .0590146    .1665682
     concenb |  -.5933022   .1926313    -3.08   0.002    -.9708527   -.2157518
       _cons |    12.0578   2.735977     4.41   0.000     6.695384    17.42022
-------------+----------------------------------------------------------------
     sigma_u |  .85125514
     sigma_e |  .16964171
         rho |  .96180277   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Instrumented:   lfare
Instruments:    ldist ldistsq y98 y99 y00 concenb concen
------------------------------------------------------------------------------
```

```
. ivreg lpassen ldist ldistsq y98 y99 y00 concenb (lfare = concen), cluster(id)

Instrumental variables (2SLS) regression               Number of obs =    4596
                                                       F(  7,  1148) =   20.28
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0649
                                                       Root MSE      =  .85549

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.3015769   .6131465    -0.49   0.623     -1.50459    .9014366
       ldist |  -1.148781   .8809895    -1.30   0.193    -2.877312    .5797488
     ldistsq |   .0772566   .0811787     0.95   0.341    -.0820187    .2365319
         y98 |   .0257148   .0164291     1.57   0.118    -.0065196    .0579491
         y99 |   .0724166   .0251272     2.88   0.004     .0231163    .1217169
         y00 |   .1127915   .0620858     1.82   0.070    -.0090228    .2346058
     concenb |  -.5933019   .2963723    -2.00   0.046    -1.174794   -.0118099
       _cons |   12.05781   4.360868     2.77   0.006      3.50164    20.61397
------------------------------------------------------------------------------
Instrumented:  lfare
Instruments:   ldist ldistsq y98 y99 y00 concenb concen
------------------------------------------------------------------------------
```

```
. * Now test whether instrument (concen) is strictly exogenous.

. xtivreg2 lpassen y98 y99 concen_p1 (lfare = concen), fe cluster(id)

FIXED EFFECTS ESTIMATION
------------------------
Number of groups =      1149                    Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3

Number of clusters (id) =         1149                Number of obs =     3447
                                                      F(  4,  1148) =    33.41
                                                      Prob > F      =   0.0000
Total (centered) SS     =  67.47207834                Centered R2   =   0.4474
Total (uncentered) SS   =  67.47207834                Uncentered R2 =   0.4474
Residual SS             =  37.28476721                Root MSE      =    .1274

------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -.8520992   .3211832    -2.65   0.008    -1.481607   -.2225917
         y98 |   .0416985   .0098066     4.25   0.000     .0224778    .0609192
         y99 |   .0948286    .014545     6.52   0.000      .066321    .1233363
   concen_p1 |   .1555725   .0814452     1.91   0.056    -.0040571    .3152021
------------------------------------------------------------------------------

Instrumented:         lfare
Included instruments: y98 y99 concen_p1
Excluded instruments: concen
------------------------------------------------------------------------------
```

```
. * What if we just use fixed effects without IV?

. xtreg lpassen lfare y98 y99 y00, fe cluster(id)

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

R-sq:  within  = 0.4507                         Obs per group: min =         4
       between = 0.0487                                        avg =       4.0
       overall = 0.0574                                        max =         4

                                                F(4,1148)          =    121.85
corr(u_i, Xb)  = -0.3249                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 1149 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |  -1.155039   .1086574   -10.63   0.000    -1.368228   -.9418496
         y98 |   .0464889   .0049119     9.46   0.000     .0368516    .0561262
         y99 |   .1023612   .0063141    16.21   0.000     .0899727    .1147497
         y00 |   .1946548   .0097099    20.05   0.000     .1756036     .213706
       _cons |   11.81677     .55126    21.44   0.000     10.73518    12.89836
-------------+----------------------------------------------------------------
     sigma_u |  .89829067
     sigma_e |  .14295339
         rho |   .9753002   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. * Test formally for endogeneity of lfare in FE:

. qui areg lfare concen y98 y99 y00, absorb(id)

 . predict u2h, resid

. xtreg lpassen lfare y98 y99 y00 v2h, fe cluster(id)

------------------------------------------------------------------------------
             |               Robust
     lpassen |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lfare |   -.301576   .4829734    -0.62   0.532    -1.249185    .6460335
         y98 |   .0257147   .0131382     1.96   0.051    -.0000628    .0514923
         y99 |   .0724165   .0197133     3.67   0.000     .0337385    .1110946
         y00 |   .1127914    .048597     2.32   0.020     .0174425    .2081403
         u2h |  -.8616344   .5278388    -1.63   0.103    -1.897271    .1740025
       _cons |   7.501007   2.441322     3.07   0.002     2.711055    12.29096
-------------+----------------------------------------------------------------

. * p-value is about .10, so not strong evidence even though FE and
. * FEIV estimatoestimates are uite different.
```

• Turns out that the FE2SLS estimator is robust to random coefficients on $\mathbf{x}_{it1}$, but one should include a full set of time dummies. (Murtazashvili and Wooldridge, 2005).

• Can model random coefficients and use a CF approach.

$$
\begin{aligned}
y_{it1} &= c_{i1} + \mathbf{x}_{it1}\mathbf{b}_{i1} + u_{it1} \\
y_{it2} &= \eta_2 + \mathbf{z}_{it}\boldsymbol{\delta}_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}
\end{aligned}
$$

• Assume $E(c_{i1}|\mathbf{z}_i, v_{it2})$ and $E(\mathbf{b}_{i1}|\mathbf{z}_i, v_{it2})$ are linear in $(\bar{\mathbf{z}}_i, v_{it2})$ and $E(u_{it1}|\mathbf{z}_i, v_{it2})$ is linear in v_{it2}, can show

$$
\begin{aligned}
E(y_{it1}|\mathbf{z}_i, v_{it2}) &= \tau_1 + \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \rho_1 v_{it2} \\
&\quad + [(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_{\bar{\mathbf{z}}}) \otimes \mathbf{x}_{it1}]\boldsymbol{\omega}_1 + v_{it2}\mathbf{x}_{it1}\boldsymbol{\zeta}_1
\end{aligned} \tag{25}
$$

(1) Regress y_{it2} on 1, $\mathbf{z}_{it}$, $\bar{\mathbf{z}}_i$ and obtain residuals $\hat{v}_{it2}$.

(2) Regress

$$y_{it1} \text{ on } 1, \mathbf{x}_{it1}, \bar{\mathbf{z}}_i, \hat{v}_{it2}, [(\bar{\mathbf{z}}_i - \bar{\mathbf{z}}) \otimes \mathbf{x}_{it1}], \hat{v}_{it2}\mathbf{x}_{it1}$$

- Probably include time dummies in both stages.

Binary and Fractional Response

• Unobserved effects (UE) "probit" model – exogenous variables. For a binary or fractional y_{it},

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1, \ldots, T. \tag{26}$$

Assume strict exogeneity (conditional on c_i) and Chamberlain-Mundlak device:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i,\ a_i|\mathbf{x}_i \sim Normal(0, \sigma_a^2). \tag{27}$$

• In binary response case under serial independence, all parameters are identified and MLE (Stata: xtprobit) can be used. Just add the time averages $\bar{\mathbf{x}}_i$ as an additional set of regressors. Then $\hat{\mu}_c = \hat{\psi} + \bar{\mathbf{x}}\hat{\boldsymbol{\xi}}$ and $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}' \left[N^{-1} \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) \right] \hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Can evaluate PEs at, say, $\hat{\mu}_c \pm k\hat{\sigma}_c$.

• Only under restrictive assumptions does c_i have an unconditional normal distribution, although it becomes more reasonable as T gets large.

• Simple to test $H_0 : \boldsymbol{\xi} = \mathbf{0}$ as null that c_i, $\bar{\mathbf{x}}_i$ are independent.

• The APEs are identified from the ASF, estimated as

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{28}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$.

• For binary or fractional response, APEs are identified without the conditional serial independence assumption. Use pooled Bernoulli quasi-MLE (Stata: glm) or generalized estimating equations (Stata: xtgee) to estimate scaled coefficients based on

$$E(y_{it}|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a). \tag{29}$$

(Time dummies have been supressed for simplicity.)

• A more radical suggestion, but in the spirit of Altonji and Matzkin (2005), is to just use a flexible model for $E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly, say,

$$\begin{aligned} E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[&\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + \\ &(\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}]. \end{aligned} \tag{30}$$

Just average out over $\bar{\mathbf{x}}_i$ to get APEs.

• If we have a binary response, start with

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{31}$$

and assume CI, we can estimate $\boldsymbol{\beta}$ by FE logit without restricting $D(c_i|\mathbf{x}_i)$.

- In any nonlinear model using the Mundlak assumption $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$, if $T \geq 3$ can include lead values, $\mathbf{w}_{i,t+1}$, to simply test strict exogeneity.
- Example: Married Women's Labor Force Participation: $N = 5,663$, $T = 5$ (four-month intervals).
- Following results include a full set of time period dummies (not reported).
- The APEs are directly comparable across models, and can be compared with the linear model coefficients.

LFP	(1)	(2)		(3)		(4)		(5)
Model	Linear	Probit		CRE Probit		CRE Probit		FE Logit
Est. Method	FE	Pooled MLE		Pooled MLE		MLE		MLE
	Coef.	Coef.	APE	Coef.	APE	Coef.	APE	Coef.
kids	−.0389	−.199	−.0660	−.117	−.0389	−.317	−.0403	−.644
	(.0092)	(.015)	(.0048)	(.027)	(.0085)	(.062)	(.0104)	(.125)
lhinc	−.0089	−.211	−.0701	−.029	−.0095	−.078	−.0099	−.184
	(.0046)	(.024)	(.0079)	(.014)	(.0048)	(.041)	(.0055)	(.083)
$\overline{kids}$	—	—	—	−.086	—	−.210	—	—
	—	—	—	(.031)	—	(.071)	—	—
$\overline{lhinc}$	—	—	—	−.250	—	−.646	—	—
	—	—	—	(.035)	—	(.079)	—	—

Probit with Endogenous Explanatory Variables

- Represent endogeneity as an omitted, time-varying variable, in addition to unobserved heterogeneity:

$$
\begin{aligned}
P(y_{it1} = 1|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) &= P(y_{it1} = 1|y_{it2}, \mathbf{z}_{it1}, c_{i1}, r_{it1}) \\
&= \Phi(\mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + r_{it1})
\end{aligned}
$$

- Elements of $\mathbf{z}_{it}$ are assumed strictly exogenous, and we have at least one exclusion restriction: $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$.

• Papke and Wooldridge (2008, Journal of Econometrics): Use a Chamberlain-Mundlak approach, but only relating the heterogeneity to all strictly exogenous variables:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + a_{i1},\ D(a_{i1}|\mathbf{z}_i) = D(a_{i1}).$$

• Even before we specify $D(a_{i1})$, this is restrictive because it assumes, in particular, $E(c_i|\mathbf{z}_i)$ is linear in $\bar{\mathbf{z}}_i$ and that $Var(c_i|\mathbf{z}_i)$ is constant. Using nonparametrics can get by with less, such as $D(c_{i1}|\mathbf{z}_i) = D(c_{i1}|\bar{\mathbf{z}}_i)$.

- Only need

$$E(y_{it1}|y_{it2},\mathbf{z}_i,c_{i1},v_{it1}) = \Phi(\mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + v_{it1}), \tag{32}$$

so applies to fractional response.

- Need to obtain an estimating equation. First, note that

$$\begin{aligned} E(y_{it1}|y_{it2},\mathbf{z}_i,a_{i1},r_{it1}) &= \Phi(\mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + r_{it1}) \\ &\equiv \Phi(\mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + v_{it1}). \end{aligned} \tag{33}$$

- Assume a linear reduced form for y_{it2}:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\delta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T \tag{34}$$
$$D(v_{it2}|\mathbf{z}_i) = D(v_{it2})$$

(and we might allow for time-varying coefficients).

- Next, assume

$$v_{it1}|(\mathbf{z}_i, v_{it2}) \sim Normal(\eta_1 v_{it2}, \kappa_1^2), t = 1, \ldots, T.$$

[Easy to allow η_1 to change over time; just have time dummies interact with v_{it2}.]

- Assumptions effectively rule out discreteness in y_{it2}.

- Write

$$v_{it1} = \eta_1 v_{it2} + e_{it1}$$

where e_{it1} is independent of $(\mathbf{z}_i, v_{it2})$ (and, therefore, of y_{it2}) and normally distributed. Again, using a standard mixing property of the normal distribution,

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, v_{it2}) = \Phi(\mathbf{x}_{it1}\boldsymbol{\beta}_{\kappa 1} + \psi_{\kappa 1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{\kappa 1} + \eta_{\kappa 1} v_{it2}) \quad (35)$$

where the "κ" denotes division by $(1+\kappa_1^2)^{1/2}$.

- Identification comes off of the exclusion of the time-varying exogenous variables $\mathbf{z}_{it2}$.

• Two step procedure (Papke and Wooldridge, 2008):

(1) Estimate the reduced form for y_{it2} (pooled or for each t separately). Obtain the residuals, $\hat{v}_{it2}$.

(2) Use the probit QMLE to estimate $\boldsymbol{\beta}_{\kappa 1}, \psi_{\kappa 1}, \boldsymbol{\xi}_{\kappa 1}$ and $\eta_{\kappa 1}$.

• How do we interpret the scaled estimates? They give directions of effects. Conveniently, they also index the APEs. For given y_2 and $\mathbf{z}_1$, average out $\bar{\mathbf{z}}_i$ and $\hat{v}_{it2}$ (for each t):

$$\hat{\alpha}_{\kappa 1} \cdot \left[N^{-1} \sum_{i=1}^{N} \phi(\hat{\alpha}_{\kappa 1} y_{t2} + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{\kappa 1} + \hat{\psi}_{\kappa 1} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_{\kappa 1} + \hat{\eta}_{\kappa 1}\hat{v}_{it2}) \right].$$

- Application: Effects of Spending on Test Pass Rates
- $N = 501$ school districts, $T = 7$ time periods.
- Once pre-policy spending is controlled for, instrument spending with the "foundation grant."
- Initial spending takes the place of the time average of IVs.

```
. * First, linear model:

. ivreg math4 lunch alunch lenroll alenroll y96-y01 lexppp94 le94y96-le94y01
       (lavgrexp = lfound lfndy96-lfndy01), cluster(distid)

Instrumental variables (2SLS) regression               Number of obs =    3507
                                                       F( 18,   500) =  107.05
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4134
                                                       Root MSE      =  .11635

                               (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   .5545247   .2205466     2.51   0.012     .1212123     .987837
       lunch |  -.0621991   .0742948    -0.84   0.403    -.2081675    .0837693
      alunch |  -.4207815   .0758344    -5.55   0.000    -.5697749   -.2717882
     lenroll |   .0463616   .0696215     0.67   0.506    -.0904253    .1831484
    alenroll |   -.049052    .070249    -0.70   0.485    -.1870716    .0889676
         y96 |  -1.085453   .2736479    -3.97   0.000    -1.623095   -.5478119
...
         y01 |   -.704579   .7310773    -0.96   0.336    -2.140941    .7317831
    lexppp94 |  -.4343213   .2189488    -1.98   0.048    -.8644944   -.0041482
     le94y96 |   .1253255   .0318181     3.94   0.000     .0628119    .1878392
...
     le94y01 |   .0865874   .0816732     1.06   0.290    -.0738776    .2470524
       _cons |   -.334823   .2593105    -1.29   0.197    -.8442955    .1746496
------------------------------------------------------------------------------
Instrumented:  lavgrexp
Instruments:   lunch alunch lenroll alenroll y96 y97 y98 y99 y00 y01
               lexppp94 le94y96 le94y97 le94y98 le94y99 le94y00 le94y01
               lfound lfndy96 lfndy97 lfndy98 lfndy99 lfndy00 lfndy01
------------------------------------------------------------------------------
```

```
. * Get reduced form residuals for fractional probit:

. reg lavgrexp lfound lfndy96-lfndy01 lunch alunch lenroll alenroll y96-y01
       lexppp94 le94y96-le94y01, cluster(distid)

Linear regression                                      Number of obs =    3507
                                                       F( 24,   500) = 1174.57
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9327
                                                       Root MSE      =  .03987

                               (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
    lavgrexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lfound |   .2447063   .0417034     5.87   0.000      .1627709    .3266417
     lfndy96 |   .0053951   .0254713     0.21   0.832      -.044649    .0554391
     lfndy97 |  -.0059551   .0401705    -0.15   0.882     -.0848789    .0729687
     lfndy98 |   .0045356   .0510673     0.09   0.929     -.0957972    .1048685
     lfndy99 |   .0920788   .0493854     1.86   0.063     -.0049497    .1891074
     lfndy00 |   .1364484   .0490355     2.78   0.006      .0401074    .2327894
     lfndy01 |   .2364039   .0555885     4.25   0.000       .127188    .3456198
       ...
       _cons |   .1632959   .0996687     1.64   0.102     -.0325251     .359117
------------------------------------------------------------------------------

. predict v2hat, resid
(1503 missing values generated)
```

```
. glm math4 lavgrexp v2hat lunch alunch lenroll alenroll y96-y01 lexppp94
      le94y96-le94y01, fa(bin) link(probit) cluster(distid)
note: math4 has non-integer values

Generalized linear models                          No. of obs      =      3507
Optimization     : ML                              Residual df     =      3487
                                                   Scale parameter =         1
Deviance         =  236.0659249                    (1/df) Deviance =  .0676989
Pearson          =  223.3709371                    (1/df) Pearson  =  .0640582

Variance function: V(u) = u*(1-u/1)                [Binomial]
Link function    : g(u) = invnorm(u)               [Probit]

                               (Std. Err. adjusted for 501 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   1.731039   .6541194     2.65   0.008     .4489886    3.013089
       v2hat |  -1.378126    .720843    -1.91   0.056    -2.790952    .0347007
       lunch |  -.2980214   .2125498    -1.40   0.161    -.7146114    .1185686
      alunch |  -1.114775   .2188037    -5.09   0.000    -1.543623    -.685928
     lenroll |   .2856761    .197511     1.45   0.148    -.1014383    .6727905
    alenroll |  -.2909903   .1988745    -1.46   0.143    -.6807771    .0987966
         ...
       _cons |  -2.455592   .7329693    -3.35   0.001    -3.892185   -1.018998
------------------------------------------------------------------------------
```

```
. margeff

Average partial effects after glm
      y  = Pr(math4)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   .5830163   .2203345     2.65   0.008     .1511686    1.014864
       v2hat |  -.4641533    .242971    -1.91   0.056    -.9403678    .0120611
       lunch |  -.1003741   .0716361    -1.40   0.161    -.2407782      .04003
      alunch |  -.3754579   .0734083    -5.11   0.000    -.5193355   -.2315803
     lenroll |   .0962161   .0665257     1.45   0.148    -.0341719    .2266041
    alenroll |  -.0980059   .0669786    -1.46   0.143    -.2292817    .0332698
         ...
------------------------------------------------------------------------------

. * These standard errors do not account for the first-stage estimation.
. * Can use the panel bootstrap. Might also look for partial effects at
. * different parts of the spending distribution.
```

Count and Other Multiplicative Models

- Conditional mean with multiplicative heterogeneity:

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{36}$$

where $c_i \geq 0$. Under strict exogeneity in the mean,

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i), \tag{37}$$

the "fixed effects" Poisson estimator is attractive: it does not restrict $D(y_{it}|\mathbf{x}_i, c_i)$, $D(c_i|\mathbf{x}_i)$, or serial dependence.

- The FE Poisson estimator is the conditional MLE derived under a Poisson and conditional independence assumptions. It is one of the rare cares where treating the c_i as parameters to estimate gives a consistent estimator of $\boldsymbol{\beta}$.
- The FE Poisson estimator is fully robust to any distributional failure and serial correlation. y_{it} does not even have to be is not a count variable! Fully robust inference is easy (`xtpqml` in Stata).

- For endogeneity there are control function and GMM approaches, with the former being more convenient but imposing more restrictions.
- CF uses same approach as before.
- Start with an omitted variables formulation:

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, r_{it1}) = \exp(\mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + r_{it1}). \quad (38)$$

- The $\{\mathbf{z}_{it}\}$ – including the excluded instruments – are assumed to be strictly exogenous here.

- If y_{it2} is (roughly) continuous we might specify

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\boldsymbol{\pi}_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}.$$

- Also write

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}$$

so that

$$E(y_{it1}|y_{it2},\mathbf{z}_i,v_{it1}) = \exp(\psi_1 + \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + v_{it1}),$$

where $v_{it1} = a_{i1} + r_{it1}$.

• Reasonable (but not completely general) to assume (v_{it1}, v_{i2}) is independent of $\mathbf{z}_i$.

• If we specify $E[\exp(v_{it1})|v_{it2}] = \exp(\eta_1 + \rho_1 v_{it2})$ (as would be true under joint normality), we obtain the estimating equation

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, v_{it2}) = \exp(\kappa_1 + \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \rho_1 v_{it2}). \quad (39)$$

• Now apply a simple two-step method. (1) Obtain the residuals $\hat{v}_{it2}$ from the pooled OLS estimation y_{it2} on 1, $\mathbf{z}_{it}$, $\bar{\mathbf{z}}_i$ across t and i. (2) Use a pooled QMLE (perhaps the Poisson or NegBin II) to estimate the exponential function, where $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ are explanatory variables along with $(\mathbf{x}_{it1})$. (As usual, a fully set of time period dummies is a good idea in the first and second steps).

• Note that y_{it2} is not strictly exogenous in the estimating equation. and so GLS-type methods account for serial correlation should not be used. GMM with carefully constructed moments could be.

• Estimating the ASF is straightforward:

$$\widehat{ASF}_t(y_{t2}, \mathbf{z}_{t1}) = N^{-1} \sum_{i=1}^{N} \exp(\hat{\kappa}_1 + \mathbf{x}_{t1}\hat{\boldsymbol{\beta}}_1 + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_1 + \hat{\rho}_1\hat{v}_{it2});$$

that is, we average out $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$.

• Test the null of contemporaneous exogeneity of y_{it2} by using a fully robust t statistic on $\hat{v}_{it2}$.

• Can allow more flexibility by iteracting $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ with $\mathbf{x}_{it1}$, or even just year dummies.

- A GMM approach – which slightly extends Windmeijer (2002) – modifies the moment conditions under a sequential exogeneity assumption on instruments and applies to models with lagged dependent variables.
- Write the model as

$$y_{it} = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) r_{it} \tag{40}$$

$$E(r_{it}|\mathbf{z}_{it}, \ldots, \mathbf{z}_{i1}, c_i) = 1, \tag{41}$$

which contains the case of sequentially exogenous regressors as a special case ($\mathbf{z}_{it} = \mathbf{x}_{it}$).

- Now start with the transformation

$$\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} = c_i(r_{it} - r_{i,t+1}). \tag{42}$$

- Can easily show that

$$E[c_i(r_{it} - r_{i,t+1})|\mathbf{z}_{it}, \ldots, \mathbf{z}_{i1}] = 0, t = 1, \ldots, T - 1.$$

• Using the moment conditions

$$E\left[\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})}\middle|\mathbf{z}_{it},\ldots,\mathbf{z}_{i1}\right] = 0,\ t = 1,\ldots,T-1 \qquad (43)$$

generally causes computational problems. For example, if $x_{itj} \geq 0$ for some j and all i and t – for example, if x_{itj} is a time dummy – then the moment conditions can be made arbitarily close to zero by choosing β_j larger and larger.

• Windmeijer (2002, Economics Letters) suggested multiplying through by $\exp(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta})$ where $\boldsymbol{\mu}_{\mathbf{x}} = T^{-1}\sum_{r=1}^{T} E(\mathbf{x}_{ir})$.

- So, the modified moment conditions are

$$E\left[\frac{y_{it}}{\exp[(\mathbf{x}_{it}-\boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]}-\frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1}-\boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]}\bigg|\mathbf{z}_{it},\ldots,\mathbf{z}_{i1}\right]=0. \quad (44)$$

- As a practical matter, replace $\boldsymbol{\mu}_{\mathbf{x}}$ with the overall sample average,

$$\bar{\mathbf{x}}=(NT)^{-1}\sum_{i=1}^{N}\sum_{r=1}^{T}\mathbf{x}_{ir}. \quad (45)$$

- The deviated variables, $\mathbf{x}_{it}-\bar{\mathbf{x}}$, will always take on positive and negative values, and this seems to solve the GMM computational problem.

Difference-in-Differences Estimation

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. The Basic Methodology
2. How Should We View Uncertainty in DD Settings?
3. Inference with a Small Number of Groups
4. Multiple Groups and Time Periods
5. Individual-Level Panel Data
6. Semiparametric and Nonparametric Approaches
7. Synthetic Control Methods for Comparative Case Studies

1. The Basic Methodology

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

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u, \tag{1}$$

where y is the outcome of interest.

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

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

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

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

- Instead, use the same two groups from another (“untreated”) state as an additional control. Let *dE* be a dummy equal to one for someone over 65 and *dB* be the dummy for living in the “treatment” state:

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \tag{3}$$

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

$$\begin{aligned}\hat{\delta}_3 = & [(\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1})] \\ & - [(\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{A,N,2} - \bar{y}_{A,N,1})]\end{aligned} \tag{4}$$

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. Even if the intervention is independent of observed covariates, adding those covariates may improve precision of the DD or DDD estimate.

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 suggested 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, in the "new" view, the additional uncertainty swamps the sampling error in estimating group/time period means.

- One way to view the uncertainty introduced in the DL framework – a perspective explicitly taken by ADH – is that our analysis should better reflect the uncertainty in the quality of the control groups.
- ADH show how to construct a synthetic control group (for California) using pre-training characteristics of other states (that were not subject to cigarette smoking restrictions) to choose the “best” weighted average of states in constructing the control.

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

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

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

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

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

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

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

3. Inference with a Small Number of Groups

- Suppose we have aggregated data on few groups (small G) and large group sizes (each M_g is large). Some of the groups are subject to a policy intervention.
- How is the sampling done? With random sampling from a large population, no clustering is needed.
- Sometimes we have random sampling within each segment (group) of the population. Except for the relative dimensions of G and M_g, the resulting data set is essentially indistinguishable from a data set obtained by sampling entire clusters.

- The problem of proper inference when M_g is large relative to G – the "Moulton (1990) problem" – has been recently studied by Donald and Lang (2007).
- DL treat the parameters associated with the different groups as outcomes of random draws.

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

$$
\begin{aligned}
y_{gm} &= \alpha + \beta x_g + c_g + u_{gm} && (5) \\
&= \delta_g + \beta x_g + u_{gm}. && (6)
\end{aligned}
$$

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

- Donald and Lang focus on (5), where c_g is assumed to be independent of x_g with zero mean. Define the composite error $v_{gm} = c_g + u_{gm}$.

- Standard pooled OLS inference applied to (5) can be badly biased because it ignores the cluster correlation. And we cannot use fixed effects.

- DL propose studying the regression in averages:

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G. \tag{7}$$

If we add some strong assumptions, we can perform inference on (7) using standard methods. In particular, assume that $M_g = M$ for all g, $c_g | x_g \sim Normal(0, \sigma_c^2)$ and $u_{gm} | x_g, c_g \sim Normal(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim Normal(0, \sigma_c^2 + \sigma_u^2/M)$. Because we assume independence across g, (7) satisfies the classical linear model assumptions.

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

$$\bar{y}_g \text{ on } 1,\ x_g,\ g = 1, \ldots, G. \tag{8}$$

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

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

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

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

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

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

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

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

$$\bar{y}_g = \alpha + \mathbf{x}_g\boldsymbol{\beta} + \bar{\mathbf{z}}_g\boldsymbol{\gamma} + \bar{v}_g, g = 1, \ldots, G, \tag{9}$$

provided $G > K + L + 1$.

- If c_g is independent of $(\mathbf{x}_g, \bar{\mathbf{z}}_g)$ with a homoskedastic normal distribution, and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution. Regressions like (9) are reasonably common, at least as a check on results using disaggregated data, but usually with larger G then just a handful.

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

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

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

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

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

- $\hat{\beta} = \overline{\Delta w}_2 - \overline{\Delta w}_1$ has been a workhorse in the quasi-experiemental literature, and obtaining inference in the traditional setting is straightforward [Card and Krueger (1994), for example.]
- Even when DL approach can be applied, should we? Suppose $G = 4$ with two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). DL involves the OLS regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$; inference is based on the t_2 distribution.

- Can show the DL estimate is

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{10}$$

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

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

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

or weight by population frequencies.

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

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

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

$$y_{gm} = \delta_g + \mathbf{z}_{gm}\boldsymbol{\gamma}_g + u_{gm}. \tag{11}$$

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

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

$$\delta_g = \alpha + \mathbf{x}_g \boldsymbol{\beta}, g = 1, \ldots, G, \tag{12}$$

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

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

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\delta}} \tag{13}$$

- Asymptotics are as the M_g get large, and $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution; its estimated asymptotic variance is $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.
- When separate group regressions are used, the $\hat{\delta}_g$ are independent and $\hat{\mathbf{V}}$ is diagonal.
- Estimator looks like "GLS," but inference is with G (number of rows in $\mathbf{X}$) fixed with M_g growing.

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

$$\hat{\delta}_g = \alpha + \mathbf{x}_g \boldsymbol{\beta} + c_g, g = 1, \ldots, G \tag{14}$$

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

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

4. Multiple Groups and Time Periods

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

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \quad (15)$$
$$i = 1,\ldots,M_{gt},$$

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

- As in cluster sample cases, can write

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt},\ i = 1,\ldots,M_{gt}; \qquad (16\ )$$

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

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \qquad (17)$$

Think of (17) as a model at the group/time period level.

- As discussed by BDM, a common way to estimate and perform inference in the individual-level equation

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma} + v_{gt} + u_{igt}$$

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.
- We cannot replace $\lambda_t + \alpha_g$ a full set of group/time interactions because that would eliminate $\mathbf{x}_{gt}$.

- If we view $\boldsymbol{\beta}$ in $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}$ as ultimately of interest – which is usually the case because $\mathbf{x}_{gt}$ contains the aggregate policy variables – 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}.
- But we can use OLS on the individual-level data to estimate the δ_{gt} in

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

assuming $E(\mathbf{z}_{igt}'u_{igt}) = \mathbf{0}$ and the group/time period sample 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 across groups and/or time can be used to impose the restrictions.
- However we obtain the $\hat{\delta}_{gt}$, 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 $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}$.
- The minimum distance (MD) approach effectively drops v_{gt} 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, proceed ignoring estimation error, and act *as if*

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{18}$$

- We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (18) by OLS – which means full year and group effects, along with $\mathbf{x}_{gt}$ – then the OLS estimator has satisfying large-sample properties as G and T both increase, provided $\{v_{gt} : t = 1, 2, \ldots, T\}$ is a weakly dependent time series for all g.
- Simulations in BDM and Hansen (2007a) indicate cluster-robust inference works reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and G is moderately large.

- Hansen (2007b) shows how to improve efficiency by using feasible GLS – by modeling $\{v_{gt}\}$ as, say, an AR(1) process.
- Naive estimators of ρ are seriously biased due to panel structure with group fixed effects. Can remove much of the bias and improve FGLS.
- Important practical point: FGLS estimators that exploit serial correlation require strict exogeneity of the covariates, even with large T. Policy assignment might depend on past shocks.

5. 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

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it},\ t = 1,2, \tag{19}$$

where $d2_t = 1$ if $t = 2$ and zero otherwise, c_i is an observed effect τ is the treatment effect. Remove c_i by first differencing:

$$(y_{i2} - y_{i1}) = \eta + \tau(w_{i2} - w_{i1}) + (u_{i2} - u_{i1}) \tag{20}$$

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \tag{21}$$

If $E(\Delta w_i \Delta u_i) = 0$, OLS applied to (21) is consistent.

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

$$\hat{\tau}_{FD} = \Delta\bar{y}_{treat} - \Delta\bar{y}_{control}, \tag{22}$$

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

• It is *not* more general to regress y_{i2} on $1, w_{i2}, y_{i1}$, $i = 1, \ldots, N$, even though this appears to free up the coefficient on y_{i1}. Why? Under (19) with $w_{i1} = 0$ we can write

$$y_{i2} = \eta + \tau w_{i2} + y_{i1} + (u_{i2} - u_{i1}). \tag{23}$$

Now, if $E(u_{i2}|w_{i2}, c_i, u_{i1}) = 0$ then u_{i2} is uncorrelated with y_{i1}, and y_{i1} and u_{i1} are correlated. So y_{i1} is correlated with $u_{i2} - u_{i1} = \Delta u_i$.

• In fact, if we add the standard no serial correlation assumption, $E(u_{i1}u_{i2}|w_{i2}, c_i) = 0$, and write the linear projection $w_{i2} = \pi_0 + \pi_1 y_{i1} + r_{i2}$, then can show that

$$plim(\hat{\tau}_{LDV}) = \tau + \pi_1(\sigma_{u_1}^2/\sigma_{r_2}^2)$$

where

$$\pi_1 = Cov(c_i, w_{i2})/(\sigma_c^2 + \sigma_{u_1}^2).$$

• For example, if w_{i2} indicates a job training program and less productive workers are more likely to participate ($\pi_1 < 0$), then the regression y_{i2} (or Δy_{i2}) on 1, w_{i2}, y_{i1} underestimates the effect.

- If more productive workers participate, regressing y_{i2} (or Δy_{i2}) on 1, w_{i2}, y_{i1} overestimates the effect of job training.
- Following Angrist and Pischke (2009), suppose we use the FD estimator when, in fact, unconfoundedness of treatment holds conditional on y_{i1} (and the treatment effect is constant). Then we can write

$$y_{i2} = \gamma + \tau w_{i2} + \psi y_{i1} + e_{i2}$$
$$E(e_{i2}) = 0,\ Cov(w_{i2}, e_{i2}) = Cov(y_{i1}, e_{i2}) = 0.$$

- Write the equation as

$$\begin{aligned}\Delta y_{i2} &= \gamma + \tau w_{i2} + (\psi - 1)y_{i1} + e_{i2} \\ &\equiv \gamma + \tau w_{i2} + \lambda y_{i1} + e_{i2}\end{aligned}$$

Then, of course, the FD estimator generally suffers from omitted variable bias if $\psi \neq 1$. We have

$$plim(\hat{\tau}_{FD}) = \tau + \lambda \frac{Cov(w_{i2}, y_{i1})}{Var(w_{i2})}$$

- If $\lambda < 0$ ($\psi < 1$) and $Cov(w_{i2}, y_{i1}) < 0$ – workers observed with low first-period earnings are more likely to participate – the $plim(\hat{\tau}_{FD}) > \tau$, and so FD overestimates the effect.

• We might expect ψ to be close to unity for processes such as earnings, which tend to be persistent. (ψ measures persistence without conditioning on unobserved heterogeneity.)

• As an algebraic fact, if $\hat{\lambda} < 0$ (as it usually will be even if $\psi = 1$) and w_{i2} and y_{i1} are negatively correlated in the sample, $\hat{\tau}_{FD} > \hat{\tau}_{LDV}$. But this does not tell us which estimator is consistent.

• If either $\hat{\lambda}$ is close to zero or w_{i2} and y_{i1} are weakly correlated, adding y_{i1} can have a small effect on the estimate of τ.

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

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it},\ t = 1,\ldots,T, \tag{24}$$

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 advantages for larger T if u_{it} is weakly dependent).

• What if designation is correlated with unit-specific trends?

"Correlated random trend" model:

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it} \tag{25}$$

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 $t = 2, \ldots, T$,

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}. \tag{26}$$

Can difference again or estimate (26) by FE.

• Can derive 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

$$E[y_{it}(0)|\mathbf{w}_i, \mathbf{c}_i] = E[y_{it}(0)|\mathbf{c}_i] \tag{27}$$

$$E[y_{it}(1)|\mathbf{w}_i, \mathbf{c}_i] = E[y_{it}(1)|\mathbf{c}_i], \tag{28}$$

where $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})$ is the time sequence of all treatments.

Suppose the gain from treatment only depends on t,

$$E[y_{it}(1)|\mathbf{c}_i] = E[y_{it}(0)|\mathbf{c}_i] + \tau_t. \tag{29}$$

Then

$$E(y_{it}|\mathbf{w}_i, \mathbf{c}_i) = E[y_{it}(0)|\mathbf{c}_i] + \tau_t w_{it} \tag{30}$$

where $y_{i1} = (1 - w_{it})y_{it}(0) + w_{it}y_{it}(1)$. If we assume

$$E[y_{it}(0)|\mathbf{c}_i] = \alpha_{t0} + c_{i0}, \tag{31}$$

then

$$E(y_{it}|\mathbf{w}_i, \mathbf{c}_i) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{32}$$

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

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

$$E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, \mathbf{c}_i) = \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + w_{it} \cdot (\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i \cdot w_{it}, \tag{33}$$

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 $\tau_i = \tau + a_i$ and then use

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}_i. \tag{34}$$

• With $T \geq 3$, can also get to a random trend model, where $g_i t$ is added to (25). Then, can difference followed by a second difference or fixed effects estimation on the first differences. With $\tau_t = \tau$,

$$\Delta y_{it} = \psi_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it} \boldsymbol{\gamma}_0 + [\Delta w_{it} \cdot (\mathbf{x}_{it} - \boldsymbol{\xi}_t)] \boldsymbol{\delta} + a_i \cdot \Delta w_{it} + g_i + \Delta u_{it}. \quad (35)$$

• Might ignore $a_i \Delta w_{it}$, using the results on the robustness of the FE estimator in the presence of certain kinds of random coefficients, or, again, estimate $\tau_i = \tau + a_i$ for each i and form (34).

- As in the simple $T = 2$ case, using unconfoundedness conditional on unobserved heterogeneity and strictly exogenous covariates leads to different strategies than assuming unconfoundedness conditional on past responses and outcomes of other covariates.
- In the latter case, we might estimate propensity scores, for each t, as $P(w_{it} = 1|y_{i,t-1},\ldots,y_{i1},w_{i,t-1},\ldots,w_{i1},\mathbf{x}_{it})$.

6. Semiparametric and Nonparametric Approaches

- Consider the setup of Heckman, Ichimura, Smith, and Todd (1997) and Abadie (2005), with two time periods. No units treated 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,

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

$W = 1$ means treatment in the second time period.

- Along with $Y_0(1) = Y_0(0)$ (no counterfactual in time period zero), key unconfoundedness assumption:

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

Also the (partial) overlap assumption is critical for τ_{att}

$$P(W = 1|X) < 1 \tag{38}$$

or the full overlap assumption for $\tau_{ate} = E[Y_1(1) - Y_1(0)]$, $0 < P(W = 1|X) < 1$.

Under (37) and (38),

$$\tau_{att} = E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{\rho[1 - p(X)]} \right\} \tag{39}$$

where Y_t, $t = 0, 1$ are the observed outcomes (for the same unit),

$\rho = P(W = 1)$ is the unconditional probability of treatment, and

$p(X) = P(W = 1|X)$ is the propensity score.

• All quantities are observed or, in the case of $p(X)$ and ρ, 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

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

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

- If we add

$$E[Y_1(1) - Y_0(1)|X, W] = E[Y_1(1) - Y_0(1)|X], \tag{41}$$

a similar approach works for τ_{ate}.

$$\hat{\tau}_{ate} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{\hat{p}(X_i)[1 - \hat{p}(X_i)]} \right\} \tag{42}$$

7. Synthetic Control Methods for Comparative Case Studies

- Abadie, Diamond, and Hainmueller (2007) argue that in policy analysis at the aggregate level, 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 control units, 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

$$y_{12} - \sum_{j=2}^{J+1} w_j y_{j2}, \tag{43}$$

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.

Missing Data

Jeff Wooldridge
Michigan State University

Programme Evaluation for Policy Analysis
Institute for Fiscal Studies
June 2012

1. When Can Missing Data be Ignored?
2. Regressing on Missing Data Indicators
3. Inverse Probability Weighting
4. Imputation
5. Heckman-Type Selection Corrections

1. When Can Missing Data be Ignored?

- Linear model with IVs:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + u_i, \tag{1}$$

where $\mathbf{x}_i$ is $1 \times K$, instruments $\mathbf{z}_i$ are $1 \times L$, $L \geq K$. Let s_i is the selection indicator, $s_i = 1$ if we can use observation i.

- With $L = K$, the "complete case" estimator is

$$\hat{\boldsymbol{\beta}}_{IV} = \left(N^{-1} \sum_{i=1}^{N} s_i \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i \mathbf{z}_i' y_i \right) \tag{2}$$

$$= \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} s_i \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i \mathbf{z}_i' u_i \right). \tag{3}$$

- For consistency, *rank* $E(\mathbf{z}_i'\mathbf{x}_i|s_i = 1) = K$ and

$$E(s_i\mathbf{z}_i'u_i) = \mathbf{0}, \tag{4}$$

which is implied by

$$E(u_i|\mathbf{z}_i, s_i) = 0. \tag{5}$$

Sufficient for (5) is

$$E(u_i|\mathbf{z}_i) = 0,\ \ s_i = h(\mathbf{z}_i) \tag{6}$$

for some function $h(\bullet)$.

• Zero covariance assumption in the population, $E(\mathbf{z}_i' u_i) = \mathbf{0}$, is not sufficient for consistency when $s_i = h(\mathbf{z}_i)$.

• If $\mathbf{x}_i$ contains elements correlated with u_i, we cannot select the sample based on those endogenous elements even though we are instrumenting for them.

• Special case is when $E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$ and selection s_i is a function of $\mathbf{x}_i$.

- Nonlinear models/estimation methods:

  Nonlinear Least Squares: $E(y|\mathbf{x}, s) = E(y|\mathbf{x})$.

  Least Absolute Deviations: $Med(y|\mathbf{x}, s) = Med(y|\mathbf{x})$

  Maximum Likelihood: $D(\mathbf{y}|\mathbf{x}, s) = D(\mathbf{y}|\mathbf{x})$ or $D(s|\mathbf{y}, \mathbf{x}) = D(s|\mathbf{x})$.

- All of these allow selection on **x** but not generally on **y**. For estimating $\mu = E(y_i)$, unbiasedness and consistency of the sample mean computed using the selected sample requires $E(y|s) = E(y)$.

• Panel data: If we model $D(\mathbf{y}_t|\mathbf{x}_t)$, and s_t is the selection indicator, the sufficient condition to ignore selection is

$$D(s_t|\mathbf{x}_t, \mathbf{y}_t) = D(s_t|\mathbf{x}_t),\ t = 1, \ldots, T. \tag{7}$$

Let the true conditional density be $f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \boldsymbol{\gamma})$. Then the partial log-likelihood function for a random draw i from the cross section can be written as

$$\sum_{t=1}^{T} s_{it} \log f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{g}) \equiv \sum_{t=1}^{T} s_{it} l_{it}(\mathbf{g}). \tag{8}$$

Can show under (7) that

$$E[s_{it} l_{it}(\mathbf{g})|\mathbf{x}_{it}] = E(s_{it}|\mathbf{x}_{it}) E[l_{it}(\mathbf{g})|\mathbf{x}_{it}]. \tag{9}$$

- If $\mathbf{x}_{it}$ includes $\mathbf{y}_{i,t-1}$, (7) allows selection on $\mathbf{y}_{i,t-1}$, but not on "shocks" from $t - 1$ to t.
- Similar findings for NLS, quasi-MLE, quantile regression.
- Methods to remove time-constant, unobserved heterogeneity: for a random draw i,

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \tag{10}$$

with IVs $\mathbf{z}_{it}$ for $\mathbf{x}_{it}$. Random effects IV methods (unbalanced panel):

$$E(u_{it}|\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT},s_{i1},\ldots,s_{iT},c_i) = 0,\ t = 1,\ldots,T \tag{11}$$

$$E(c_i|\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT},s_{i1},\ldots,s_{iT}) = E(c_i) = 0. \tag{12}$$

Selection in any time period cannot depend on u_{it} or c_i.

• FE on unbalanced panel: can get by with just (11). Let $\ddot{y}_{it} = y_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$ and similarly for and $\ddot{\mathbf{x}}_{it}$ and $\ddot{\mathbf{z}}_{it}$, where $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods for observation i. The FEIV estimator is

$$\hat{\boldsymbol{\beta}}_{FEIV} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{z}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it}' \ddot{\mathbf{z}}_{it}' y_{it} \right).$$

Weakest condition for consistency is $\sum_{t=1}^{T} E(s_{it} \ddot{\mathbf{z}}_{it}' u_{it}) = 0$.

• One important violation of (11) is when units drop out of the sample in period $t + 1$ because of shocks (u_{it}) realized in time t. This generally induces correlation between $s_{i,t+1}$ and u_{it}.

- A simple variable addition test is to estimate the auxiliary model

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \rho s_{i,t+1} + c_i + u_{it}$$

by FE2SLS, where $s_{i,t+1}$ acts as its own instrument, and test $\rho = 0$.

Lose a time period, so need $T \geq 3$ initially.

- Similar to test of strict exogeneity of instruments: include leads $\mathbf{z}_{i,t+1}$ and estimate by FE2SLS.

- Consistency of FE (and FEIV) on the unbalanced panel under breaks down if the slope coefficients are random and one ignores this in estimation. The error term contains the term $\mathbf{x}_i \mathbf{d}_i$ where $\mathbf{d}_i = \mathbf{b}_i - \boldsymbol{\beta}$.
- Simple test based on the alternative

$$E(\mathbf{b}_i | \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, s_{i1}, \ldots, s_{iT}) = E(\mathbf{b}_i | T_i). \tag{13}$$

Add interaction terms of dummies for each possible sample size (with $T_i = T$ as the base group):

$$1[T_i = 2]\mathbf{x}_{it},\ 1[T_i = 3]\mathbf{x}_{it},\ \ldots,\ 1[T_i = T - 1]\mathbf{x}_{it}. \tag{14}$$

Estimate equation by FE or FEIV. (In latter case, IVs are $1[T_i = r]\mathbf{z}_{it}$.)

- Can use FD in basic model, too, which is very useful for attrition problems. Generally, if

$$\Delta y_{it} = \varphi_t + \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it},\ t = 2, \ldots, T \tag{15}$$

and, if $\mathbf{z}_{it}$ is the set of IVs at time t, we can use

$$E(\Delta u_{it}|\mathbf{z}_{it}, s_{it}) = 0 \tag{16}$$

as being sufficient to ignore the missingess. Again, can add $s_{i,t+1}$ to test for attrition.

- Nonlinear models with unosberved effects are more difficult to handle. Certain conditional MLEs (logit, Poisson) can accomodate selection that is arbitrarily correlated with the unobserved effect.

2. Regression on Missing Data Indicators

- When data are missing on the covariates, it is common in empirical work it is common to see the data used when covariates are observed and otherwise to include a missing data indicator.
- Not clear that this is that helpful. It does not generally produce consistent estimators when the data are missing as a function of the covariates (above).
- Suppose we start with the standard population model

$$
\begin{aligned}
y &= \alpha + \mathbf{x}\boldsymbol{\beta} + u \\
E(u|\mathbf{x}) &= 0
\end{aligned}
$$

- Assume we always observe y. Let s be the selection indicator for observing $\mathbf{x}$ (all or nothing for simplicity). Then $m = 1 - s$ is the missing data indicator.
- If (u, s) is independent of $\mathbf{x}$ then we can assume $E(\mathbf{x}) = \mathbf{0}$ for identification [because $E(\mathbf{x}) = E(\mathbf{x}|s = 1)$].
- Note that s is allowed to be correlated with u but not with any of the observables.
- Write

$$\begin{aligned} y &= \alpha + s\mathbf{x}\boldsymbol{\beta} + (1 - s)\mathbf{x}\boldsymbol{\beta} + u \\ &= \alpha + s\mathbf{x}\boldsymbol{\beta} + m\mathbf{x}\boldsymbol{\beta} + u \end{aligned}$$

- Using the independence assumption,

$$\begin{aligned} E(y|\mathbf{x},m) &= \alpha + s\mathbf{x}\boldsymbol{\beta} + m\mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{x},m) \\ &= \alpha + s\mathbf{x}\boldsymbol{\beta} + m\mathbf{x}\boldsymbol{\beta} + E(u|m) \\ &= (\alpha + \eta) + s\mathbf{x}\boldsymbol{\beta} + m\mathbf{x}\boldsymbol{\beta} + \rho m \end{aligned}$$

- The proper population regression with missing data is the linear projection of y on $(1, s\mathbf{x}, m)$:

$$\begin{aligned} L(y|1,s\mathbf{x},m) &= (\alpha + \eta) + s\mathbf{x}\boldsymbol{\beta} + L(m\mathbf{x}|1,s\mathbf{x},m)\boldsymbol{\beta} + \rho m \\ &= (\alpha + \eta) + s\mathbf{x}\boldsymbol{\beta} + \rho m \end{aligned}$$

because $L(m\mathbf{x}|1,s\mathbf{x},m) = \mathbf{0}$. (Use $E(\mathbf{x}) = \mathbf{0}$, $sm = 0$, and m independent of $\mathbf{x}$.)

- We have shown that the slopes on $s\mathbf{x}$ are correct: $\boldsymbol{\beta}$ from the population model. The intercept is not the population intercept. When we allow for $E(\mathbf{x}) \neq \mathbf{0}$ the intercept will be different yet.
- Not obvious that there are interesting situations where $L(m\mathbf{x}|1, s\mathbf{x}, m) = L(m\mathbf{x}|1, m)$, which means adding m solves the missing data problem.

- Key point: The assumption $E(u|\mathbf{x}, s) = 0$ is sufficient for complete-case OLS to be consistent for $\boldsymbol{\beta}$; it allows arbitrary correlation between s and $\mathbf{x}$. Adding s (or m) as a regressor and using all data uses something like independence between s and $\mathbf{x}$ (but u and s can be related).

3. Inverse Probability Weighting

Weighting for Cross Section Problems

• When selection is not on conditioning variables, can try to use probability weights to reweight the selected sample to make it representative of the population. Suppose y is a random variable whose population mean $\mu = E(y)$ we would like to estimate, but some observations are missing on y. Let $\{(y_i, s_i, \mathbf{z}_i) : i = 1, \ldots, N\}$ indicate independent, identically distributed draws from the population, where $\mathbf{z}_i$ is always observed (for now).

- Missingness is "ignorable" or "selection on observables" assumption:

$$P(s = 1|y, \mathbf{z}) = P(s = 1|\mathbf{z}) \equiv p(\mathbf{z}) \tag{17}$$

where $p(\mathbf{z}) > 0$ for all possible values of $\mathbf{z}$. Consider

$$\tilde{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{p(\mathbf{z}_i)} \right) y_i, \tag{18}$$

where s_i selects out the observed data points. Using (17) and iterated expectations, can show $\hat{\mu}_{IPW}$ is consistent (and unbiased) for y_i. (Same kind of estimate used for treatment effects.)

- Sometimes $p(\mathbf{z}_i)$ is known, but mostly it needs to be estimated. Let $\hat{p}(z_i)$ denote the estimated selection probability:

$$\hat{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) y_i. \tag{19}$$

Can also write as

$$\hat{\mu}_{IPW} = N_1^{-1} \sum_{i=1}^{N} s_i \left(\frac{\hat{\rho}}{\hat{p}(\mathbf{z}_i)} \right) y_i \tag{20}$$

where $N_1 = \sum_{i=1}^{N} s_i$ is the number of selected observations and $\hat{\rho} = N_1/N$ is a consistent estimate of $P(s_i = 1)$.

- A different estimate is obtained by solving the least squares problem

$$\min_m \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) (y_i - m)^2.$$

- Horowitz and Manski (1998) study estimating population means using IPW. HM focus on bounds in estimating $E[g(y)|\mathbf{x} \in A]$ for conditioning variables $\mathbf{x}$. Problem with certain IPW estimators based on weights that estimate $P(s = 1)/P(s = 1|\mathbf{z})$: the resulting estimate of the mean can lie outside the natural bounds. One should use $P(s = 1|\mathbf{x} \in A)/P(s = 1|\mathbf{x} \in A, \mathbf{z})$ if possible. Unfortunately, cannot generally estimate the proper weights if x is sometimes missing.

- The HM problem is related to another issue. Suppose

$$E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}. \tag{21}$$

Let **z** be a variables that are always observed and let $p(\mathbf{z})$ be the selection probability, as before. Suppose at least part of x is not always observed, so that **x** is not a subset of **z**. Consider the IPW estimator of α, $\boldsymbol{\beta}$ solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) (y_i - a - \mathbf{x}_i \mathbf{b})^2. \tag{22}$$

• The problem is that if

$$P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x}), \tag{23}$$

the IPW is generally inconsistent because the condition

$$P(s = 1|\mathbf{x}, y, \mathbf{z}) = P(s = 1|\mathbf{z}) \tag{24}$$

is unlikely. On the other hand, if (23) holds, we can consistently estimate the parameters using OLS on the selected sample.

• If $\mathbf{x}$ always observed, case for weighting is much stronger because then $\mathbf{x} \subset \mathbf{z}$. If selection is on $\mathbf{x}$, this should be picked up in large samples in flexible estimation of $P(s = 1|\mathbf{z})$.

- If selection is exogenous and $\mathbf{x}$ is always observed, is there a reason to use IPW? Not if we believe $E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}$ along with the homoskedasticity assumption $Var(y|\mathbf{x}) = \sigma^2$. Then, OLS is efficient and IPW is less efficient. IPW can be more efficient with heteroskedasticity (but WLS with the correct heteroskedasticity function would be best).

• Still, one can argue for weighting under (23) as a way to consistently estimate the linear projection. Write

$$L(y|1,x) = \alpha^* + \mathbf{x}\boldsymbol{\beta}^* \tag{25}$$

where $L(\cdot|\cdot)$ denotes the linear projection. Under under $P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x})$, the IPW estimator is consistent for $\boldsymbol{\theta}^* = (\alpha^*, \boldsymbol{\beta}^{*\prime})'$. The unweighted estimator has a probabilty limit that depends on $p(\mathbf{x})$.

- Parameters in LP show up in certain treatment effect estimators, and are the basis for the "double robustness" result of Robins and Ritov (1997) in the case of linear regression.
- The double robustness result holds for certain nonlinear models, but must choose model for $E(y|\mathbf{x})$ and the objective function appropriately; see Wooldridge (2007). [For binary or fractional response, use logistic function and Bernoulli quasi-log likelihood (QLL). For nonnegative response, use exponential function with Poisson QLL.]

• Return to the IPW regression estimator under $P(s = 1|\mathbf{x}, y, \mathbf{z}) = P(s = 1|\mathbf{z}) = G(\mathbf{z}, \boldsymbol{\gamma})$, with

$$E(u) = 0, E(\mathbf{x}'u) = 0, \tag{26}$$

for a parametric function $G(\bullet)$ (such as flexible logit), and $\hat{\gamma}$ is the binary response MLE. The asymptotic variance of $\hat{\boldsymbol{\theta}}_{IPW}$, using the estimated probability weights, is

$$Avar\sqrt{N}(\hat{\boldsymbol{\theta}}_{IPW} - \boldsymbol{\theta}) = [E(\mathbf{x}_i'\mathbf{x}_i)]^{-1}E(\mathbf{r}_i\mathbf{r}_i')[E(\mathbf{x}_i'\mathbf{x}_i)]^{-1}, \tag{27}$$

where $\mathbf{r}_i$ is the $P \times 1$ vector of population residuals from the regression $(s_i/p(\mathbf{z}_i))\mathbf{x}_i'u_i$ on $\mathbf{d}_i'$, and $\mathbf{d}_i$ is the $M \times 1$ score for the MLE used to obtain $\hat{\boldsymbol{\gamma}}$.

• Variance in (27) is always smaller than the variance if we knew $p(\mathbf{z}_i)$.

Leads to a simple estimate of $Avar(\hat{\boldsymbol{\theta}}_{IPW})$:

$$\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'\right)\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)\mathbf{x}_i'\mathbf{x}_i\right)^{-1} \tag{28}$$

If selection is estimated by logit with regressors $\mathbf{h}_i = \mathbf{h}(\mathbf{z}_i)$,

$$\hat{\mathbf{d}}_i = \mathbf{h}_i'(s_i - \Lambda(\mathbf{h}_i\hat{\boldsymbol{\gamma}})), \tag{29}$$

where $\Lambda(a) = \exp(a)/[1 + \exp(a)]$.

- Illustrates an interesting finding of RRZ (1995): Can never do worse for estimating the parameters of interest, $\boldsymbol{\theta}$, and usually do better, when adding irrelevant functions to a logit selection model in the first stage. The Hirano, Imbens, and Ridder (2003) estimator keeps expanding $\mathbf{h}_i$.
- Adjustment in (27) carries over to general nonlinear models and estimation methods. Ignoring the estimation in $\hat{p}(\mathbf{z})$, as is standard, is asymptotically conservative. When selection is exogenous in the sense of $P(s = 1|\mathbf{x}, y, \mathbf{z}) = P(s = 1|\mathbf{x})$, the adjustment makes no difference.

- Nevo (2003) studies the case where population moments are $E[\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ and selection depends on elements of $\mathbf{w}_i$ not always observed.
- Approach: Use information on population means $E[\mathbf{h}(\mathbf{w}_i)]$ such that $P(s = 1|\mathbf{w}) = P[s = 1|h(\mathbf{w})]$ and use method of moments.

- For a logit selection model,

$$E\left\{\frac{s_i}{\Lambda[\mathbf{h}(\mathbf{w}_i)\boldsymbol{\gamma}]}\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta})\right\} = 0 \tag{30}$$

$$E\left\{\frac{s_i\mathbf{h}(\mathbf{w}_i)}{\Lambda[\mathbf{h}(\mathbf{w}_i)\boldsymbol{\gamma}]}\right\} = \bar{\boldsymbol{\mu}}_h \tag{31}$$

where $\bar{\boldsymbol{\mu}}_h$ is known. Equation (31) generally identifies $\boldsymbol{\gamma}$, and $\hat{\boldsymbol{\gamma}}$ can be used in a second step to choose $\hat{\boldsymbol{\theta}}$ in a weighted GMM procedure.

Attrition in Panel Data

• Inverse probability weighting can be applied to the attrition problem in panel data. Many estimation methods can be used, but consider MLE. We have a parametric density, $f_t(y_t|\mathbf{x}_t, \boldsymbol{\theta})$, and let s_{it} be the selection indicator. Pooled MLE on on the observed data:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}), \tag{32}$$

which is consistent if $P(s_{it} = 1|y_{it}, \mathbf{x}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$. If not, maybe we can find variables $\mathbf{r}_{it}$, such that

$$P(s_{it} = 1|y_{it}, \mathbf{x}_{it}, \mathbf{r}_{it}) = P(s_{it} = 1|\mathbf{r}_{it}) \equiv p_{it} > 0. \tag{33}$$

• The weighted MLE is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T}(s_{it}/p_{it})\log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta}). \tag{34}$$

Under (33), $\hat{\theta}_{IPW}$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it},\boldsymbol{\theta})] \tag{35}$$

where $q_t(\mathbf{w}_{it},\boldsymbol{\theta}) = \log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta})$.

• How do we choose $\mathbf{r}_{it}$ to make (33) hold (if possible)? RRZ (1995) propose a sequential strategy,

$$\pi_{it} = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1), t = 1,\ldots,T. \tag{36}$$

Typically, $\mathbf{z}_{it}$ contains elements from $(\mathbf{w}_{i,t-1},\ldots,\mathbf{w}_{i1})$.

• How do we obtain p_{it} from the π_{it}? Not without some strong assumptions. Let $\mathbf{v}_{it} = (\mathbf{w}_{it}, \mathbf{z}_{it}), t = 1, \ldots, T$. An ignorability assumption that works is

$$P(s_{it} = 1|\mathbf{v}_i, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1). \quad (37)$$

That is, given the entire history $\mathbf{v}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT})$, selection at time t depends only on variables observed at $t - 1$. RRZ (1995) show how to relax it somewhat in a regression framework with time-constant covariates. Using (37), can show that

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1}\cdots\pi_{i1}. \quad (38)$$

- So, a consistent two-step method is: (i) In each time period, estimate a binary response model for $P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1)$, which means on the group still in the sample at $t - 1$. The fitted probabilities are the $\hat{\pi}_{it}$. Form $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1} \cdots \hat{\pi}_{i1}$. (ii) Replace p_{it} with $\hat{p}_{it}$ in (34), and obtain the weighted pooled MLE.
- As shown by RRZ (1995) in the regression case, it is more efficient to estimate the p_{it} than to use know weights, if we could. See RRZ (1995) and Wooldridge (2010) for a simple regression method for adjusting the score.

• IPW for attrition suffers from a similar drawback as in the cross section case. Namely, if $P(s_{it} = 1|\mathbf{w}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$ then the unweighted estimator is consistent. If we use weights that are not a function of $\mathbf{x}_{it}$ in this case, the IPW estimator is generally inconsistent.

• Related to the previous point: would rarely apply IPW in the case of a model with completely specified dynamics. Why? If we have a model for $D(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0})$ or $E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0})$, then our variables affecting attrition, $\mathbf{z}_{it}$, are likely to be functions of $(y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0})$. If they are, the unweighted estimator is consistent. For misspecified models, we might still want to weight.

4. Imputation

- So far, we have discussed when we can just drop missing observations (Section 1) or when the complete cases can be used in a weighting method (Section 2). A different approach to missing data is to try to fill in the missing values, and then analyze the resulting data set as a complete data set. Little and Rubin (2002) provide an accessible treatment to *imputation* and *multiple imputation* methods, with lots of references to work by Rubin and coauthors.

- Imputing missing values is not always valid. Most methods depend on a *missing at random* (MAR) assumption. When data are missing on the response variable, y, MAR is essentially the same as $P(s = 1|y, \mathbf{x}) = P(s = 1|\mathbf{x})$. *Missing completely at random* (MCAR) is when s is independent of $\mathbf{w} = (\mathbf{x}, y)$.
- MAR for general missing data patterns. Let $\mathbf{w}_i = (\mathbf{w}_{i1}, \mathbf{w}_{i2})$ be a random draw from the population. Let $r_i = (r_{i1}, r_{i2})$ be the "retention" indicators for $\mathbf{w}_{i1}$ and $\mathbf{w}_{i2}$, so $r_{ig} = 1$ implies $\mathbf{w}_{ig}$ is observed. MCAR is that $\mathbf{r}_i$ is independent of $\mathbf{w}_i$. The MAR assumption is that $P(r_{i1} = 0, r_{i2} = 0|\mathbf{w}_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}$ and so on.

• MAR is more natural with monotone missing data problems; we just saw the case of attrition. If we order the variables so that if $\mathbf{w}_{ih}$ is observed the so is $\mathbf{w}_{ig}$, $g < h$. Write

$f(\mathbf{w}_1,\ldots,\mathbf{w}_G) = f(\mathbf{w}_G|\mathbf{w}_{G-1},\ldots,\mathbf{w}_1)$

$\cdot f(\mathbf{w}_{G-1}|\mathbf{w}_{G-1},\ldots,\mathbf{w}_1)\cdots f(\mathbf{w}_2|\mathbf{w}_1)f(\mathbf{w}_1)$. Partial log likelihood:

$$\sum_{g=1}^{G} r_{ig}\log f(\mathbf{w}_{ig}|\mathbf{w}_{i,g-1},\ldots,\mathbf{w}_{i1},\boldsymbol{\theta}), \tag{39}$$

where we use $r_{ig} = r_{ig}r_{i,g-1}\cdots r_{i2}$. Under MAR,

$$E(r_{ig}|\mathbf{w}_{ig},\ldots,\mathbf{w}_{i1}) = E(r_{ig}|\mathbf{w}_{i,g-1},\ldots,\mathbf{w}_{i1}). \tag{40}$$

(39) is the basis for filling in data in monotonic MAR schemes.

• Simple example of imputation. Let $\mu_y = E(y)$, but data are missing on some y_i. Unless $P(s_i = 1|y_i) = P(s_i = 1)$, the complete-case average is not consistent for μ_y. Suppose that the selection is ignorable conditional on $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}). \tag{41}$$

NLS using selected sample is consistent for $\boldsymbol{\beta}$. Obtain a fitted value, $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, for any unit it the sample. Let $\hat{y}_i = s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the imputed data. Imputation estimator:

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\}. \tag{42}$$

• From $plim(\hat{\mu}_y) = E[s_i y_i + (1 - s_i)m(\mathbf{x}_i, \boldsymbol{\beta})]$ we can show consistency of $\hat{\mu}_y$ because under (41),

$$E[s_i y_i + (1 - s_i)m(\mathbf{x}_i, \boldsymbol{\beta})] = E[m(\mathbf{x}_i, \boldsymbol{\beta})] = \mu_y. \tag{43}$$

• Danger in using imputation methods: we might be tempted to treat the imputed data as real random draws. Generally leads to incorrect inference because of inconsistent variance estimation. (In linear regression, easy to see that estimated variance is too small.)

• Little and Rubin (2002) call (43) the method of "conditional means." In Table Table 4.1 they document the downward bias in variance estimates.

- LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ – assuming that we can estimate $D(y|\mathbf{x})$. If we assume $D(u_i|\mathbf{x}_i) = Normal(0, \sigma_u^2)$, draw $\check{u}_i$ from a Normal$(0, \hat{\sigma}_u^2)$, distribution, where $\hat{\sigma}_u^2$ is estimated using the complete case nonlinear regression residuals, and then use $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) + \check{u}_i$ for the missing data. Called the "conditional draw" method of imputation (special case of stochastic imputation).
- Generally difficult to quantity the uncertainty from single-imputation methods, where a single imputed values is obtained for each missing variable. Can bootstrap the entire estimation/imputation steps, but this is computationally intensive.

• Multiple imputation is an alternative. Its theoretical justification is Bayesian, based on obtaining the posterior distribution – in particular, mean and variance – of the parameters conditional on the observed data. For general missing data patterns, the computation required to impute missing values is intensive, and involves simulation methods of estimation. See also Cameron and Trivedi (2005).

• General idea: rather than just impute one set of missing values to create one "complete" data set, create several imputed data sets. (Often the number is fairly small, such as five or so.) Estimate the parameters of interest using each imputed data set, and average to obtain a final parameter estimate and sampling error.

• Let $\mathbf{W}_{mis}$ denote the matrix of missing data and $\mathbf{W}_{obs}$ the matrix of observations. Assume that MAR holds. MAR used to estimate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$, the posterier mean of $\boldsymbol{\theta}$ given $\mathbf{W}_{obs}$. But by iterated expectations,

$$E(\boldsymbol{\theta}|\mathbf{W}_{obs}) = E[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{44}$$

If $\hat{\boldsymbol{\theta}}_d = E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis}^{(d)})$ for imputed data set d, then approximate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$ as

$$\bar{\boldsymbol{\theta}} = D^{-1}\sum_{d=1}^{D}\hat{\boldsymbol{\theta}}_d. \tag{45}$$

• Further, we can obtain a "sampling" variance by estimating $Var(\theta|\mathbf{W}_{obs})$ using

$$Var(\theta|\mathbf{W}_{obs}) = E[Var(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}] + Var[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}], \tag{46}$$

which suggests

$$\widehat{Var}(\theta|\mathbf{W}_{obs}) = D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})' \equiv \bar{\mathbf{V}} + \mathbf{B}. \tag{47}$$

• For small number of imputations, a correction is usually made, namely, $\bar{\mathbf{V}} + (1 + D)^{-1}\mathbf{B}$. assuming that one trusts the MAR assumption and the underlying distributions used to draw the imputed values, inference with multiple imputations is fairly straightforward. D need not be very large so estimation using nonlinear models is relatively easy, given the imputed data.

• Use caution when applying to models with missing conditioning variables. Suppose $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are interested in $D(y|\mathbf{x})$, data are missing on y and $\mathbf{x}_2$, and selection is a function of $\mathbf{x}_2$. Using the complete cases will be consistent. Imputation methods would not be, as they require $D(s|y, \mathbf{x}_1, \mathbf{x}_2) = D(s|\mathbf{x}_1)$.

5. Heckman-Type Selection Corrections

- With random slopes in the population, get a new twist on the usual Heckman procedure.

$$y_1 = a_1 + \mathbf{x}_1\mathbf{b}_1 \equiv \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 + \mathbf{x}_1\mathbf{e}_1$$

where $u_1 = a_1 - \alpha_1$ and $\mathbf{e}_1 = \mathbf{b}_1 - \boldsymbol{\beta}_1$. Let $\mathbf{x}$ be the full set of exogenous explanatory variables with $\mathbf{x}_1$ a strict subset of $\mathbf{x}$.

- Assume selection follows a standard probit:

$$y_2 = [\eta_2 + \mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$$
$$D(v_2|\mathbf{x}) = Normal(0,1)$$

- Also, $(u_1, \mathbf{e}_1, v_2)$ independent of $\mathbf{x}$ with $E(u_1, \mathbf{e}_1 | v_2)$ linear in v_2. Then

$$E(y_1|\mathbf{x}, v_2) = \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2 + \mathbf{x}_1 v_2 \boldsymbol{\psi}_1$$

and so

$$E(y_1|\mathbf{x}, y_2 = 1) = \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1\lambda(\eta_2 + \mathbf{x}\boldsymbol{\delta}_2) + \lambda(\eta_2 + \mathbf{x}\boldsymbol{\delta}_2) \cdot \mathbf{x}_1\boldsymbol{\psi}_1$$

- Compared with the usual Heckman procedure, add the interactions $\hat{\lambda}_{i2} \cdot \mathbf{x}_{i1}$, where $\hat{\lambda}_{i2} = \lambda(\hat{\eta}_2 + \mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$ is the estimated IMR:

$$y_{i1} \text{ on } 1,\ \mathbf{x}_{i1},\ \hat{\lambda}_{i2},\ \hat{\lambda}_{i2} \cdot \mathbf{x}_{i1} \ \text{ using } y_{i2} = 1$$

- Bootstrapping is convenient for inference. Full MLE, where $(u_1, \mathbf{e}_1, v_2)$ is multivariate normal, would be substantially more difficult.
- Can test joint significance of $(\hat{\lambda}_{i2}, \hat{\lambda}_{i2} \cdot \mathbf{x}_{i1})$ to test null of no selection bias – no need to adjust for first-stage estimation.
- Be careful with functional form. Interactions might be significant because population model is not a true conditional mean.

- Back to constant slopes but endogenous explanatory variable.
- If can find IVs, has advantage of allowing missing data on explanatory variables in addition to the response variable. (A variable that is exogenous in the population model need not be in the selected subpopulation.)

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{48}$$

$$y_2 = \mathbf{z}_2\boldsymbol{\delta}_2 + v_2 \tag{49}$$

$$y_3 = 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. \tag{50}$$

- Assume (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) observed when $y_3 = 1$; (b) $E(u_1|\mathbf{z}, v_3) = \gamma_1 v_3$; (c) $v_3|\mathbf{z}\sim Normal(0,1)$; (d) $E(\mathbf{z}_2' v_2) = \mathbf{0}$ and $\boldsymbol{\delta}_{22} \neq \mathbf{0}$ where $\mathbf{z}_2\boldsymbol{\delta}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_{21}\boldsymbol{\delta}_{22}$.
- Then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{51}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. Selection is exogenous in (51) because $E(e_1|\mathbf{z}, y_3) = 0$. Because y_2 is not exogenous, we estimate (51) by IV, using the selected sample, with IVs $[\mathbf{z}_2, \lambda(\mathbf{z}\boldsymbol{\delta}_3)]$ because $g(\mathbf{z}, 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$.

• The two-step estimator is (i) Probit of y_3 on $\mathbf{z}$ to (using all observations) to get $\hat{\lambda}_{i3} \equiv \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$; (ii) IV (2SLS if overidentifying restrictions) of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{\lambda}_{i3}$ using IVs $(\mathbf{z}_{i2}, \hat{\lambda}_{i3})$.

• If y_2 is always observed, tempting to obtain the fitted values $\hat{y}_{i2}$ from the reduced form y_{i2} on $\mathbf{z}_{i2}$, and then use OLS of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}, \hat{\lambda}_{i3}$ in the second stage. But this effectively puts $\alpha_1 v_2$ in the error term, so we would need $u_1 + \alpha_2 v_2$ to be normally (or something similar). Rules out discrete y_2. The procedure just outlined uses the linear projection $y_2 = \mathbf{z}_2\boldsymbol{\pi}_2 + \eta_2\lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$ in the selected population, and does not care whether this is a conditional expectation.

- In theory, can set $\mathbf{z}_2 = \mathbf{z}$, although that usually means lots of collinearity in the (implicit) reduced form for y_2 in the selected sample.
- Choosing $\mathbf{z}_1$ a strict $\mathbf{z}_2$ and $\mathbf{z}_2$ a strict ssubset of $\mathbf{z}$ enforces discipline. Namely, we should have an exogenous variable that would be valid as an IV for y_2 in the absense of sample selection, and at least one more variable (in $\mathbf{z}$) that mainly affects sample selection.

• If an explanatory variable is not always observed, ideally can find an IV for it and treat it as endogenous even if it is exogenous in the population. The usual Heckman approach (like IPW and imputation) is hard to justify in the model $E(y|\mathbf{x}) = E(y|\mathbf{x}_1)$ if $\mathbf{x}_1$ is not always observed. The first-step would be estimation of $P(s = 1|\mathbf{x}_2)$ where $\mathbf{x}_2$ is always observed. But then we would be assuming $P(s = 1|\mathbf{x}) = P(s = 1|\mathbf{x}_2)$, effectively an exclusion restriction on a reduced form.