Topics

Rómulo A. Chumacero

Introduction

- LRM has several limitations, some of which were previously discussed
- Omitting relevant variables, measurement errors, and simultaneity bring inconsistency
- Extend the framework to tackle some of these issues

Outline

1. Instrumental variables
2. Simultaneous equations
3. Treatment effects (experiments and quasi-experiments)

Instrumental Variables

- When $\mathcal{E}(x_t u_t) \neq 0$, OLS estimators are biased and inconsistent
- Consider the model:

$$y_t = \beta x_t^* + u_t$$

but $x^* \sim \left(0, \sigma_{x^*}^2\right)$ is not observed. Instead, we observe

$$x_t = x_t^* + v_t$$

$$\widehat{\beta} = \frac{\sum xy}{\sum x^2} = \beta \frac{\sum x x^*}{\sum x^2} + \frac{\sum xu}{\sum x^2} \xrightarrow{p} \beta \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} \right)$$

OLS estimator is closer to 0 than true β (attenuation bias)

- In general case ($\mathcal{E}(x_t u_t) = \Sigma_{xu} \neq 0$):

$$Y = X\beta + u$$

$$\widehat{\beta} \xrightarrow{p} \beta + \Sigma_{xx}^{-1} \Sigma_{xu} \neq \beta$$

- A consistent estimator can be obtained using instrumental variables
- To be a valid instrument, Z requires to satisfy two conditions:
  - Instrument relevance [Z is correlated with X]: $T^{-1} Z'X \xrightarrow{p} \Sigma_{ZX}$
  - Instrument exogeneity [Z is uncorrelated with u]: $T^{-1} Z'u \xrightarrow{p} 0$

Instrumental Variables

- As long as $\dim(X) = k \leq \dim(Z) = m$:

$$Z'Y = Z'X\beta + Z'u \quad \text{with} \quad \mathcal{V}(Z'u) = \sigma^2 \mathcal{E}(Z'Z)$$

- This suggests obtaining the GLS estimator:

$$\widehat{\beta}_{IV} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'Y = (X'P_Z X)^{-1} X'P_Z Y$$

$$\widehat{\mathcal{V}}\left(\widehat{\beta}_{IV}\right) = \widehat{\sigma}^2 (X'P_Z X)^{-1}$$

$$\widehat{\sigma}^2 = T^{-1}\left(Y - X\widehat{\beta}_{IV}\right)'\left(Y - X\widehat{\beta}_{IV}\right)$$

- Unlike OLS, the IV estimator is consistent:

$$\widehat{\beta}_{IV} = \beta + \left[T^{-1}X'Z(Z'Z)^{-1}Z'X\right]^{-1} T^{-1}X'Z(Z'Z)^{-1}Z'u$$

$$T^{-1}X'Z(Z'Z)^{-1}Z'X \xrightarrow{p} \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \quad \text{and} \quad T^{-1}X'Z(Z'Z)^{-1}Z'u \xrightarrow{p} \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{Zu} = 0$$

- Special case: $\dim(X) = \dim(Z)$: Standard IV estimator

- As long as $X'Z$ is $k \times k$ and nonsingular:

$$\widehat{\beta}_{SIV} = (Z'X)^{-1}Z'Y \quad \text{with } \widehat{\mathcal{V}}\left(\widehat{\beta}_{SIV}\right) = \widehat{\sigma}^2 (Z'X)^{-1}Z'Z(X'Z)^{-1}$$

IV and 2SLS

- The IV estimator can also be seen as the result of a double application of LS
  - Stage 1: Regress each variable in X on Z to obtain $\widehat{X}$:

$$\widehat{X} = Z\left(Z'Z\right)^{-1}Z'X = P_Z X$$

  - Stage 2: Regress Y on $\widehat{X}$ to obtain the 2SLS estimator:

$$\begin{aligned}\widehat{\beta}_{2SLS} &= \left(\widehat{X}'\widehat{X}\right)^{-1}\widehat{X}'Y \\ &= \left(X'P_Z X\right)^{-1}X'P_Z Y = \widehat{\beta}_{IV}\end{aligned}$$

- As IV estimator is obtained from the 2SLS procedure, variances are the same

Choice of Instruments

- Crucial question: where to find useful instruments?
- Some may be from X itself (those thought to be exogenous)
- Some are lagged variables
- Invalid instruments produce meaningless results. Essential to assess validity
  - Instrument relevance:
    - More variation of X due to Z: accurate estimators and asymptotic normality
    - Instruments that account for little variation are called **weak instruments**
    - With one endogenous regressor check if F<10 (1st stage)
    - Try to discard weak instruments or use more advanced tools to estimate
  - Instrument exogeneity
    - If Z is correlated with u, IV is inconsistent
    - Test for $Cov(z, u) = 0$ [Overidentifying restrictions test, $k < m$]
      - Obtain $\widehat{u}_{IV} = Y - X\widehat{\beta}_{IV}$
      - Regress $\widehat{u}_{IV}$ on constant and Z
      - Check $TR^2 \xrightarrow{D} \chi^2_{m-k}$

Test for Endogeneity

- IV estimation is called for when X and u are correlated
- We would like to have a test to evaluate $\mathrm{H}_0 : Cov\,(x, u) = 0$
  - If not rejected, although IV is consistent, OLS is more efficient
  - If rejected, IV is consistent and OLS is not
- Hausman Test:
  - Regress X on Z and obtain residuals for each X $(\widehat{v})$
  - Regress Y on X and the $\widehat{v}$´s
  - Test if coefficients associated with $\widehat{v}$´s are significant

Simultaneous Equations

- Most models contain systems of equations with more than one endogenous variable
- The simplest example of a structural model:

$$\begin{aligned} &\text{Demand:} \quad Q = \alpha_1 P + \alpha_2 X + u_d \\ &\text{Supply:} \quad Q = \beta_1 P + u_s \\ \mathcal{E}(u_d) \;&=\; \mathcal{E}(u_s) = 0 = Cov(u_d, u_s), \quad \mathcal{V}(u_d) = \sigma_d^2, \; \mathcal{V}(u_s) = \sigma_s^2 \end{aligned}$$

  - Equilibrium P and Q are endogenous, X (income) is considered exogenous
  - This means that P and Q are both correlated with u_d and u_s
  - Structural parameters $(\alpha_1, \alpha_2, \beta_1)$ can not be estimated consistently with usual methods
- Reduced form equations:

$$\begin{aligned} \beta_1 P + u_s &= \alpha_1 P + \alpha_2 X + u_d \\ P &= \frac{\alpha_2}{\beta_1 - \alpha_1} X + \frac{u_d - u_s}{\beta_1 - \alpha_1} = \delta_1 X + v_1 \\ Q &= \frac{\beta_1 \alpha_2}{\beta_1 - \alpha_1} X + \frac{\beta_1 u_d - \alpha_1 u_s}{\beta_1 - \alpha_1} = \delta_2 X + v_2 \end{aligned}$$

  - Reduced form parameters (δ_1, δ_2) can be estimated consistently using LS (why?)
  - Reduced form estimation is important because it summarizes the equilibrium outcomes
  - They can be used to forecast
  - They lack structural interpretation, as they are a combination of structural parameters

Identification

- Reduced-form parameters (RFP) can be consistently estimated
- Can we use them to obtain consistent estimators of structural parameters (SP)?
- Identification problem: SP is identified if it has unique representation with RFP
  - Order condition:
    - G: # of endogenous variables, K: # of exogenous variables
    - g: number of endogenous variables on the equation, k: exogenous variables on the equation
    - $K - k \geq g - 1$ (exo. variables excluded at least as great as endo. included -1)
      - With equality, identified
      - With inequality, overidentified
  - Rank condition
- On the example: $K = 1, G = 2$
  - Demand: $g = 2, k = 1 \rightarrow K - k = 0 < 1$ (unidentified)
  - Supply: $g = 2, k = 0 \rightarrow K - k = 1 = 1$ (identified)

Estimation

- IV techniques can be used to estimate simultaneous equations
- Consider the estimation of the SP of equation n

$$Y_n = Y_{\overline{n}}\beta_n + X_n\gamma_n + u_n = Z_n\alpha_n + u_n$$

  Y_n: $T \times 1$, $Y_{\overline{n}} : T \times (g-1)$, $X_n : T \times k$, $Z_n = [Y_{\overline{n}}\ X_n]$, $\alpha_n' = \left[\beta_n'\ \gamma_n'\right]$

- Assume that order condition for identification is satisfied
- Apply 2SLS:
  - Regress Z_n on X (all exogenous variables) and obtain:

$$\widehat{Z}_n = X\left(X'X\right)^{-1}X'Z_n = P_X Z_n$$

  - Regress Y_n on $\widehat{Z}_n$ to obtain the IV (2SLS) estimator:

$$\widehat{\alpha}_n = \left(Z_n'P_X Z_n\right)^{-1}Z_n'P_X Y_n$$

$$\widehat{\mathcal{V}}\left(\widehat{\alpha}_n\right) = \widehat{\sigma}_n^2\left(Z_n'P_X Z_n\right)^{-1}$$

$$\widehat{\sigma}^2 = T^{-1}\left(Y_n - Z_n\widehat{\alpha}_n\right)'\left(Y - Z_n\widehat{\alpha}_n\right)$$

- Inference can be conducted as usual

Treatment Effects

- Experiments can be used to assess causal effects
  - Treatment and control groups to assess effect of treatment
- True randomized controlled experiments are rare in economics
- Consider the question: Do hospitals make people healthier?
  - Information on people's health and on visits to hospitals
  - People that visit hospitals (treatment group) report poorer health
  - Post hoc, ergo procter hoc fallacy
  - Selection bias (treatment and control groups are not randomly assigned)
- Quasi-experiment (natural experiment): external events sometimes produce what appears to be randomization

Randomized Controlled Experiments

- Control and treatment groups are randomly assigned
- The causal effect of the treatment can be assessed directly

$$\begin{aligned} y_i &= \beta_0 + \beta_1 d_i + u_i \\ d_i &= \begin{cases} 1 & \text{treatment} \\ 0 & \text{control} \end{cases} \end{aligned}$$

- Treatment effect: β_1:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(d_i - \overline{d}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N} \left(d_i - \overline{d}\right)^2} = \overline{Y}_1 - \overline{Y}_0$$

  is also called the Difference estimator (the difference between means of treatment and control groups)

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N} \left(d_i - \overline{d}\right)\left(u_i - \overline{u}\right)}{\sum_{i=1}^{N} \left(d_i - \overline{d}\right)^2} = \beta_1 + \overline{u}_1 - \overline{u}_0$$

- For $\widehat{\beta}_1$ to be unbiased, we require $\mathcal{E}\left(\overline{u}_1 - \overline{u}_0\right) = \mathcal{E}\left(\overline{u}_1\right) - \mathcal{E}\left(\overline{u}_0\right) = 0$
  - Expected values of all other factors affecting the outcome must be the same for C-T (covariate balance)
  - Self-selection violates this requirement
- Falsely attributing the effect to treatment

Potential Problems with Experiments

- Threats to internal validity (is statistical inference valid for the population studied?)
  - Failure to randomize (treatment is based in part on characteristic or preference)
  - Failure to follow protocol (treatment assigned versus received)
  - Attrition (dropping out of the study is not random)
  - Experimental effects (being in experiment changes behavior, Hawthorne effect)
  - Small samples (valid inference)
- Threats to external validity (ability to generalize results to other population and settings)
  - Nonrepresentative sample
  - Nonrepresentative program or policy
  - General equilibrium effects (scale, duration, financing)
  - Treatment versus eligibility effects (participation in actual programs is voluntary)

Solutions to Problems

- Overt bias: the effect is (partially or fully) due to x, not treatment
  - Use Difference estimator with additional controls (x)

$$y_i = \beta_0 + \beta_1 d_i + \beta_2^{'} x_i + u_i$$

    Consistent under conditional mean independence of u wrt x and d

- Selection on observables (propensity score matching)

$$\Pr(d_i = 1 \,|x_i\,) = F\,(x_i, \delta)$$

  - Estimate $F\left(x_i, \widehat{\delta}\right)$
  - For each individual with $d_i = 1$, choose a 'clone' with similar $F\left(x_i, \widehat{\delta}\right)$, but with $d_i = 0$ (e.g., nearest neighbor)
  - Obtain the difference estimator between groups

- Consistent under conditional mean independence of u wrt x and d
  - If not, use IV

Quasi-Experiments

- Natural experiments
  - Real-world conditions approximate randomized controlled experiment
  - Treatment appears as if it were randomly assigned
  - Before vs after data
  - Omitted variables bias (unobservables), that are fixed to individual
- Suppose we observe two groups before and after a policy change
  - Treatment group affected, control group not affected
  - Assume a common trend in both groups

Quasi-Experiments

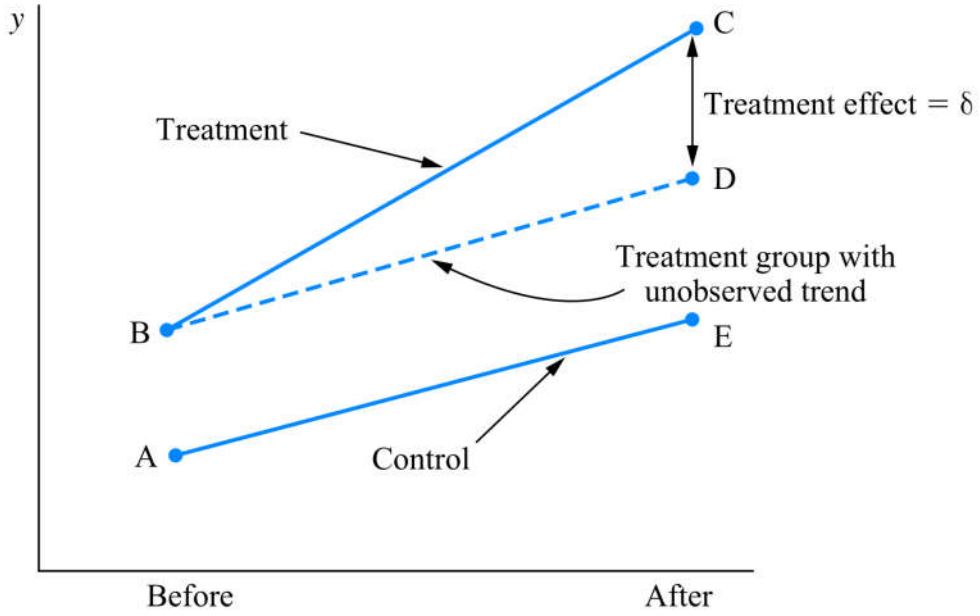

Figure 1: Difference-in-Difference Estimator

$$\widehat{\delta} = \left[\overline{y}_{Treated,After} - \overline{y}_{Control,After}\right] - \left[\overline{y}_{Treated,Before} - \overline{y}_{Control,Before}\right] = \left(\widehat{C} - \widehat{E}\right) - \left(\widehat{B} - \widehat{A}\right)$$

$$\begin{aligned} \Delta y_i &= \alpha + \delta d_i + u_i, \quad \text{DiD (DD)} \\ \Delta y_i &= \alpha + \delta d_i + x_i^{'}\beta + u_i, \quad \text{DiD with controls} \end{aligned}$$

Regression Discontinuity Design (RD)

- Exploits knowledge of rules determining treatment
- When arbitrary, they provide good experiments (discontinuity)
- Sharp RD
  - Treatment status; deterministic and discontinuous on observable g, with g_0 known

$$d_i = \begin{cases} 1 & \text{if } g_i \geq g_0 \\ 0 & \text{if } g_i < g_0 \end{cases}$$

  - Difference estimator with observations in the neighborhood of g_0

$$y_i = \beta_0 + \beta_1 d_i + u_i$$

  - Difference estimator with additional covariates

$$\beta_0 + \beta_1 d_i + \beta_2^{'} x_i + u_i$$

  - Generalized RD

$$\beta_0 + \beta_1 d_i + \beta_2^{'} x_i + \beta_3^{'} d_i x_i + u_i$$

- Fuzzy RD
  - Treatment status; probabilistic

$$\Pr(d_i = 1 \,|g_i) = \begin{cases} h_1(g_i) & \text{if } g_i \geq g_0 \\ h_0(g_i) & \text{if } g_i < g_0 \end{cases}$$

- Structural breaks