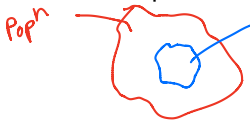
Tutorial 12: Sample Selection

Hammad Shaikh

April 8, 2021

Sample Selection Introduction

- Empirical analysis based on samples and not the population



↳ Empirics use sample to infer about population

- Sample used for analysis may not represent the population

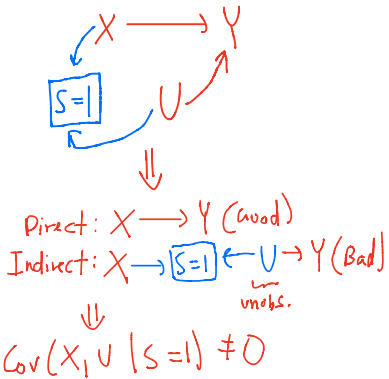
Sample of only females
↳ Empirics may not be informative about males

- Participants may select themselves to the sample of interest

Employed sample
↳ participant is choosing to be employed

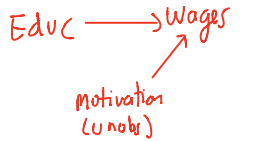
} Focus of today

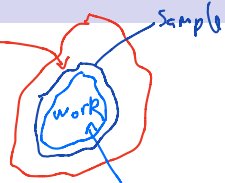
Formalization of the Selection Problem



- Population: $Y_i = \alpha_0 + \alpha_1 X_i + U_i, E(U_i|X_i) = 0$
  - ↳ $\{(X_i, Y_i)\}_{i=1}^{N}$ where N = population size
  - ↳ observe everyone
- Sample selection: $s_i = I(Z_i\gamma_1 + X_i\gamma_2 + V_i \geq 0)$
  - ↳ $S_i = \begin{cases} 1, (x_i, y_i) \text{ obs.} \\ 0, \text{ do not obs. } y_i \end{cases}$
  - Exog. Educ. unobs.
  - * Selection eqn charac. who is in sample ($s_i=1$) and who is not ($s_i=0$)
- Conditional expectation for selected sample: 0 if random sample
  $E(Y_i|X_i, Z_i, s_i = 1) = \beta_0 + X_i\beta_1 + E(u_i|X_i, Z_i, s_i = 1)$
  - in sample
  - $\rho \times \frac{\phi(Z_i\gamma_1 + X_i\gamma_2)}{\Phi(Z_i\gamma_1 + X_i\gamma_2)}$ } Inverse Mills Ratio $\Rightarrow$ X_i is correlated with $E[u_i|X_i, Z_i, s_i = 1]$
- Selection bias: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ if $Cov(X_i, \epsilon_i|s_i = 1) \neq 0$
  - contain inverse mills ratio (omited)
  - ↳ omited variable bias

Example of Sample Selection Bias





- Population: $\underbrace{ln(wage_i)}_{Y} = \beta_0 + \beta_1 \underbrace{educ_i}_{X} + \epsilon_i$
  - $\hookrightarrow E[Y|X] = \beta_0 + \beta_1 X + \underbrace{E[\epsilon|X]}_{0}$
- Unobservables: $\epsilon_i = \theta_1 motiv_i + e_i$
  - $\hookrightarrow \theta_1 > 0 \Rightarrow \uparrow$ motivation $\Rightarrow \uparrow$ wages
    - $\hookrightarrow Y_i = \beta_0 + \beta_1 X_i + (\theta_1 motiv_i + e_i) \Rightarrow Cov(Educ_i, motiv_i) = 0$
- Sample selection: $work_i = I(\gamma educ_i + \underbrace{(\theta_2 motiv_i + \eta_i)}_{\text{unobs. } V_i} \geq 0)$
  - $\hookrightarrow \gamma > 0, \theta_2 > 0$
    - $\Rightarrow Cov(\epsilon_i, V_i) \neq 0 \Rightarrow$ Selection problem
- Selection bias: $Cov(educ_i, motiv_i | work_i = 1) < 0$
  - $\hookrightarrow$ Suppose I am given $\downarrow educ_i$ & $work_i = 1$
    - $\Rightarrow \uparrow motiv_i \Rightarrow E[\underbrace{motiv}_{\text{unobs.}} \times educ \mid work_i = 1] \neq 0$

↳ selection bias for $\log(wage) = \beta_0 + \beta_1 Educ_i + \eta_i$ only for $workR_i = 1$ (η_i: motivation)

Solutions to Selection Problem

Control variable →

1. Include control variable W_i so that $Cov(X_i, \epsilon_i, W_i | s_i = 1) = 0$
   - ↳ $\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 motiv_i + \varepsilon_i$ ($motiv_i$: observed)
   - ↳ No endog. issue since $Cov(educ_i, \varepsilon_i | motiv_i) = 0$

Control function →

2. Control for the estimated Inverse Mills Ratio (Heckman)
   - ↳ (i) Estimate the selection eqn (eg. probit)
     - ↳ $\hat{s}_i = I(z_i \hat{\gamma}_1 + x_i \hat{\gamma}_2)$
   - (ii) Control for estimated Inverse Mills Ratio

$$\frac{\phi(z_i \hat{\gamma}_1 + x_i \hat{\gamma}_2)}{\Phi(z_i \hat{\gamma}_1 + x_i \hat{\gamma}_2)} = \hat{\lambda}_i(z_i, x_i)$$

$$Y_i = \beta_0 + \beta_1 X_i + \rho \cdot \hat{\lambda}_i + \varepsilon_i$$

} Implement in Stata together

control for inverse mills ratio