Review of Identification

- What does identification mean, econometrically?
- Some relate "endogeneity" to identification, it's often imprecise.
- **Intuition:**
  - Data generated by two models have the same distribution $\Rightarrow$ Not identified
  - Two different model specifications lead to different data distributions $\Rightarrow$ Identified

Inference and Identification

- Sample $\xrightarrow{\text{(Statistical Inference)}}$ Population
- Population $\xrightarrow{\text{(Identification)}}$ Unobserved Parameters

Notations

- Data $X \sim P$, parameter θ
- Θ: parameter space
- Set of all distributions $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$
- Assume model is correctly specified: $P \in \mathcal{P}$
- The *identification set* $\equiv$ the set of all θ that could have generated the data P:

$$\Theta(P) := \{\theta \in \Theta : P_\theta = P\}$$

Remarks

- Identification is a property of the distribution of the data and the model
- Identification is not about sample size
- Identification can help build estimators
- Only after having identification do you discuss estimation

Example

- Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
$$E[\epsilon_i | X_i = x] = 0, \quad \forall x$$

- Goal: under non-collinearity, β is identified
- Proof: show that the identification set is a singleton
- Suppose not:

$$\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i}, \forall x$$

- This isn't always true. Thus, β is identified.

Identification of the Roy Model

- Above we show identification by contradiction
- Another way: express parameters as a function of data moments
- Consider the Roy model

$$Y_{fi} = \mu_f + \epsilon_{fi}$$
$$Y_{hi} = \mu_h + \epsilon_{hi}$$

- ϵ_{fi} and ϵ_{hi} follow a joint normal distribution.
- Let J_i denote the choice.

Identification of Roy Models

- One can express the parameters with the following moments
  - $Pr(J_i = f)$
  - $E[Y_i|J_i = f]$
  - $E[Y_i|J_i = h]$
  - $Var[Y_i|J_i = f]$
  - $Var[Y_i|J_i = h]$
  - $E[[Y_i - E(Y_i|J_i = f)]^3|J_i = f]$
  - $E[[Y_i - E(Y_i|J_i = h)]^3|J_i = h]$

Potential Outcome Framework

- Neyman-Fisher-Roy-Quandt-Rubin Causal Model
- Another set of notations and languages
- Random variable: D is the actual state.
- For each state $d \in \mathcal{D}$, there's a random variable Y_d.
- *"What would have happened if we are in state d?"*
- We observe:

$$Y = DY_{d=1} + (1 - D)Y_{d=0}$$

- We only observe Y, not Y_d.

Remarks

- Parameters of interest
  - ATE: $E[Y_{d=1} - Y_{d=0}]$
  - ATT: $E[Y_{d=1} - Y_{d=0}|D = 1]$
- How to identify these?
- Potential Outcome Framework $=$ Roy Model

Selection

- Selection from the potential outcome point of view:

$$Y_d | D = d \text{ doesn't distribute the same as } Y_d | D = d'.$$

- Think of this back in the Roy model language:

$$\begin{aligned} Y_1 &= \mu_1 + \epsilon_1 \\ Y_0 &= \mu_0 + \epsilon_0 \\ D &= \mathbf{1}\{Y_1 \geqslant Y_0 \end{aligned}$$

Random Assignment

- Random assignment:
$$\{Y_d\}_{d\in\mathcal{D}} \perp D$$
- Under random assignment, the distribution of Y_d is point identified:

$$\begin{aligned} F_d(y) &:= P(Y_d \leqslant y), \text{(Definition)} \\ &= P(Y_d \leqslant y|D = d), \text{(Random assignment)} \\ &= P(Y \leqslant y|D = d), \text{(Definition)} \end{aligned}$$

- Any function of $F_d(Y)$ is point identified.
- E.g., ATE & ATT

Non-Identification of Joint Distribution

- Even with random assignment......
- The joint distribution of Y_1 and Y_0 is not (non-parametrically) identified.
- e.g., $Var(Y_1 - Y_0)$