CAUSAL INFERENCE

Technical Track Session I

Sergio Urzua

University of Maryland

Policy questions are causal in nature

Does school decentralization **improve** school quality?

What is the **effect** of conditional cash transfers on school attendance ?

Does health insurance **decrease** health expenditures of the poor?

Do higher health worker salaries **lead to** better performance?

3 common but typically uniformative strategies

- Before and after comparisons
- Comparisons of participants and non-participants
- Measuring only statistical correlation (or association)

Before and after comparisons

- Compare outcome of interest pre-intervention (t=0) to outcome post-intervention (t=1)

- Examples
  - School decentralization: test scores
  - Conditional cash transfers: school attendance
  - Health insurance: health expenditure
  - Higher health worker salaries: absenteeism

Before and after comparisons

- Yields impact of intervention + whatever else happened between t=0 and t=1
  - **Concurrent trends**
  - Other interventions, labor market shocks, aggregate health shocks

- Health insurance and simultaneous heavy rains
  - Health insurance → increase utilization, decrease expenditure per utilization
  - Malaria epidemic → increase utilization and expenditure
  - Crop loss → decrease income and utilization
  - Impossible to disentangle changes between t=0 and t=1
    - Health insurance (- or +)
    - Malaria (+)
    - Cross loss (-)

- Could *underestimate* or *overestimate* true impact of intervention
  - Not just magnitude but also sign of the effect

Compare participants and non-participants

- Comparing units with intervention to units not part of intervention
- Gives intervention effect + whatever is different between participants and non-participants
  - **Selection bias**
  - Programs usually targeted to certain areas or people
  - Individuals voluntarily apply or join a program
- CCTs
  - Applicants more motivated than non-applicants → higher school attendance to begin with

Compare participants and non-participants

- Impossible to disentangle these unobservable characteristics of participants from intervention impact
- What differentiates participants and non-participants in a CCT?
  - CCT
  - Motivation, perceived importance of school
    - Unobserved and very difficult to measure
    - Could be higher or lower among participants
- Again could underestimate or overestimate intervention impact

Statistical correlations

$$s_{it} = \alpha + \beta_1 CCT_{it} + \sum_{j=1}^{J} \beta_j \vec{X}_i + \varepsilon_{it}$$

- Multivariate regression analysis alone does not take care of these problems
- Participant versus non-participants in CCT example with a vector *X* of *J* different characteristics for each household *i* (control variables).
- Motivation correlated with *CCT* but unobserved so part of ε.
- Induces correlation between ε and *CCT*
- β = biased estimator for impact of CCT.

How can we generalize this?

- Problems with causal inference for
  - Before and after comparisons
    - Common time effects
  - Comparisons of participants and non-participants
    - Selection bias
    - Holds even in a multivariate regression context
- No easy fix for selection bias
  - Topic of rest of the workshop!
  - Need a common framework or language that could apply to all examples…

Defining terms

- Define the population by *U*, and each unit in *U* by *u*.
  - Example: *U is a sample of households and u is particular household*
- *Y* is the outcome of interest, or response variable
- For each $u \in U$, there is an associated value of *Y(u)*
  - Example: *Y(u) is the realization of health expenditure for household u.*

The treatment variable

- Let ***D*** be a variable that indicates the state to which each unit in *U* is exposed.

$$D = \begin{cases} 1 & \text{If unit } u \text{ is in treatment group} \\ 0 & \text{If unit } u \text{ is in comparison group} \end{cases}$$

- Example:
  - D=1 → Household is covered by health insurance
  - D=0 → Household is not covered by health insurance

The treatment variable

- Let ***D*** be a variable that indicates the state to which each unit in *U* is exposed.

$$D = \begin{cases} 1 & \text{If unit } u \text{ is in treatment group} \\ 0 & \text{If unit } u \text{ is in comparison group} \end{cases}$$

- The response *Y* is potentially affected by whether *u* receives treatment or not.
  - Y is therefore a function of D.
  - $\boldsymbol{Y_1(u)}$ = treated outcome for unit *u*
  - $\boldsymbol{Y_0(u)}$ = comparison outcome for same unit *u*.

The effect of treatment on the outcome

- We are interested in the effect caused by treatment for unit u:

$$\delta_u = Y_1(u) - Y_0(u)$$

- Example: the difference in health expenditure when *u* has health insurance and when *u* has no health insurance

- **Fundamental problem** of causal inference
  - For a given unit *u*, we observe either $Y_1(u)$ or $Y_0(u)$
  - We never observe *u* both with and without health insurance at the same time.

The effect of treatment on the outcome

- Fundamental problem of causal inference
  - For a given unit *u*, we observe either $Y_1(u)$ or $Y_0(u)$
  - We never observe *u* both with and without health insurance at the same time.

- We cannot observe the **counterfactual**
  - If *u* is actually treated, we cannot observe what would have happened to *u* in the absence of treatment.

So what do we do?

- We can never measure treatment effect on a particular unit *u*

- Instead, we identify the average treatment effect for the population *U*

$$
\begin{aligned}
ATE_U &= E_U[Y_1(u) - Y_0(u)] \\
&= E_U[Y_1(u)] - E_U[Y_0(u)] \\
&= E_U[Y_1(u) \mid D = 1] - E_U[Y_0(u) \mid D = 0] \\
&= \overline{\delta}
\end{aligned}
$$

Let's re-arrange some terms

- Add and subtract $E_U[Y_0(u) \mid D=1]$

$$\begin{aligned}\overline{\delta} &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] \\ &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] + E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=1] \\ &= \underbrace{E_U[Y_1(u) - Y_0(u) \mid D=1]}_{\text{Treatment effect}} + \underbrace{E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=0]}_{\text{Selection bias}}\end{aligned}$$

Treatment effect: Average difference between treated and untreated outcomes for treatment group [TOT].

On average, among those who got health insurance, what difference did the insurance make?

Selection bias: Difference in average untreated outcomes between treatment and comparison groups

Besides effect of health insurance, there may be other differences between insured and uninsured group

Let's re-arrange some terms

- $E_U[Y_0(u) \mid D=1]$ is typically unobserved
- Objectives of empirical work
  - First best: Identify situations in which selection bias = 0
  - Second best: Correct for selection bias

$$\begin{aligned}\bar{\delta} &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] \\ &= E_U[Y_1(u) \mid D=1] - E_U[Y_0(u) \mid D=0] + E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=1] \\ &= \underbrace{E_U[Y_1(u) - Y_0(u) \mid D=1]}_{\text{Treatment effect}} + \underbrace{E_U[Y_0(u) \mid D=1] - E_U[Y_0(u) \mid D=0]}_{\text{Selection bias}}\end{aligned}$$

How can we remove or minimize selection bias?

- Objective of methods discussed throughout workshop
  - Randomization
  - Differences-in-differences
  - Matching
  - Instrumental variables
  - Regression discontinuity design

References

- James Heckman (2005) "The Scientific Model of Causality". Sociological Methodology, Volume 35, Issue 1, Pages 1-97.

- Esther Duflo, Rachel Glennerster, and Michael Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit," in T.Paul Schultz and John Strauss (eds.) Handbook of Development Economics, Vol 4.

- Joshua Angrist and Jorn-Steffen Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press

- Donald B. Rubin (1974): *"Estimating causal effects of treatments in randomized and nonrandomized experiments", Journal of Educational Psychology 66*, pp. 688-701.