Causal Inference

Zhentao Shi

The Chinese University of Hong Kong

Western Philosophers

- Aristotle
  - Material: The mask is made of gold
  - Formal: Mathematics
  - Effective: Change the status by external force
  - Final: I run because I want to be healthy
- Thomas Aquinas: The first cause?
- David Hume (Skepticism): Causation as a relationship between two impressions in the mind. Causality cannot be proven.
- Karl Popper: Falsification

Marxism

- Immanuel Kant: Causality is not a feature of the external world, but rather a way of organizing our experience of the world.
- Causality is not abstract ideas, but rather a result of interactions between material forces and **human activity**.
  - The economic base determines the superstructure.
  - Human agency plays a crucial role in shaping causality.
  - Class struggle is a key driver of historical change and causality.

Hinduism

Karma is the universal law of
CAUSE
&
EFFECT
You reap what you sow
Karma Quotes via Gecko&Fly

Chinese Ideas

- 反者，道之动；弱者，道之用。天下万物生于有，有生于无
- 积善之家，必有余庆
- 遂古之初，谁传道之？上下未形，何由考之？... 阴阳三合，何本何化？...

Section 1

Potential Outcomes

Comparison of Two means

- Sample 1: $(X_1, \ldots, X_{N_x}) \sim N(\mu_x, \sigma_x^2)$
- Sample 2: $(Y_1, \ldots, Y_{N_y}) \sim N(\mu_y, \sigma_y^2)$
- Difference in sample means $\Delta = \bar{X} - \bar{Y}$
- Hypothesis $H_0 : \mu_x = \mu_y$. Test statistic

$$z = \frac{\Delta}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

- Without normality, asymptotic theory helps in large sample
- Purely statistical exercise.

Potential Outcome Framework

- Add a story to the statistical exercise.
- A triple (y_{1i}, y_{0i}, D_i)
  - $D_i \in \{0, 1\}$ is a treatment (from biomedical)
  - Two potential outcomes (y_{1i}, y_{0i})
- Observed outcome

$$y_i = \begin{cases} y_{1i}, & \text{if } D_i = 1 \quad \text{treatment group} \\ y_{0i}, & \text{if } D_i = 0 \quad \text{control group} \end{cases}$$

Equivalently,

$$y_i = y_{1i} D_i + y_{0i} (1 - D_i)$$

Examples

- Clinical research
    - Effects of drugs
    - Surgical techniques
    - Diets
- Heraclitus: "A man cannot step into the same river twice, because it is not the same river, and he is not same man."
- Economics
    - Effects of monetary police
    - Effects of poverty alleviation
    - Effects of pension reform

Treatment Effect

- $\Delta_i = y_{1i} - y_{0i}$ is a random variable that varies with individuals
  - e.g. severity of side effects after people receiving the same vaccine.
- Δ_i is unobservable. Researchers only observe y_{1i} or y_{0i}, but not both
- Direct control experiment
- A funny video: 2:25–4:25, 5:55–7:10

ATE and ATET

- Average treatment effect

$$\mathsf{ATE} = E[\Delta_i]$$

- Average treatment effect on the treated

$$\mathsf{ATET} = E[\Delta_i \mid D_i = 1]$$

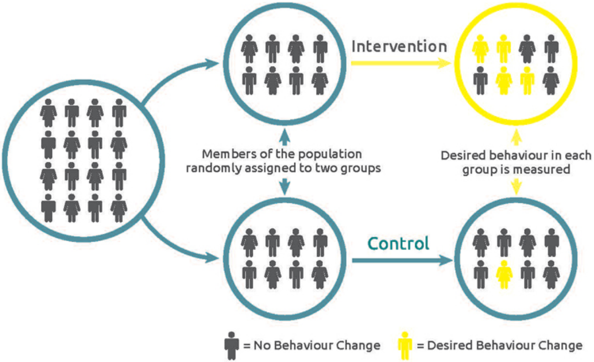
Section 2

Randomized Controlled Trials

RCT

- History: James Lind in 1747 identified a treatment of scurvy
- The “gold standard” for scientific discovery
- Given a random sample from the same population. Randomly split it into a **treatment group** and a **control group**.

Diagram of RCT



- Example: Zhongfei Xingnao Fang (link)

ATE Under RCT

- Random assignment implies

$$(y_{1i}, y_{0i}) \perp D_i.$$

The potential outcome is independent of the assignment.

- **Treatment group** "$\mathcal{T}$" (N_1 observations)
- **Control group** "$\mathcal{C}$" ($N_0 = N - N_1$ observations)

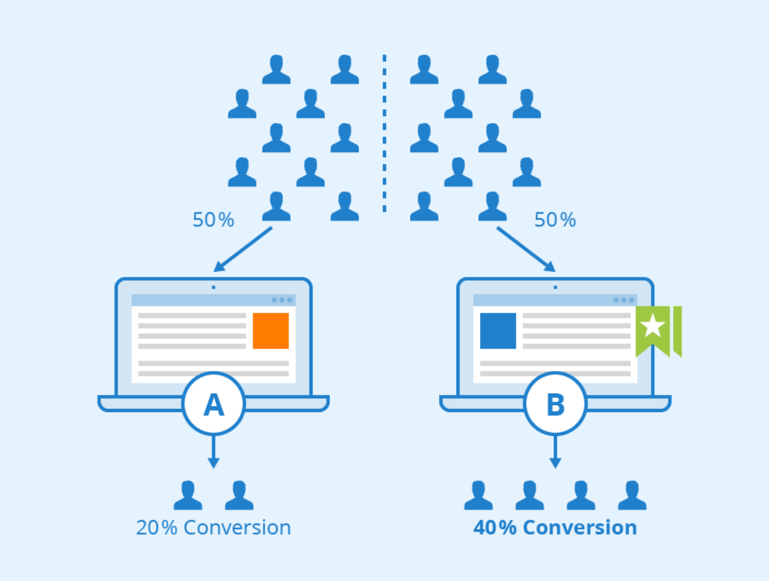
$$\begin{aligned}\widehat{ATE} &= \frac{1}{N_1}\sum_{i\in\mathcal{T}} y_i - \frac{1}{N_0}\sum_{i\in\mathcal{C}} y_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i y_i}{N_1/N} - \frac{(1-D_i)y_i}{N_0/N}\right]\end{aligned}$$

RCT in Development Economics

- Nobel prize 2019: Banerjee, Duflo, and Kremer
  - Deworming in Kenya (link)
  - Microcredit in India
- Example in Gansu, China (link)

- Very costly
- Few researchers have the resources

RCT in Tech Industry



Section 3

Observational Studies

Conditional ATE and ATET

- With control variables $X_i = x$, **conditional ATE** (CATE)

$$ATE(x) = E[\Delta_i | X_i = x]$$

- Similar, **conditional ATET** is defined as

$$ATE(x) = E[\Delta_i | D_i = 1, X_i = x]$$

- Straightforward if X_i is a discrete random variable

Unconfoundedness

- To mimic RCT, it requires **Conditional Independence**

$$(y_{1i}, y_{0i}) \perp D_i \mid X_i$$

  which is also called **Unconfoundedness**

- In an **observational study**, it means "Once X_i is controlled, the potential outcome is independent of the treatment"
- In principle, we should include all **confounding variables**
- Unconfoundedness is an untestable assumption!

- $ATET(x) = ATE(x)$ under unconfoundedness.

$$E[\Delta_i \mid D_i = 1, X_i] = E[\Delta_i \mid X_i]$$

Overlapping Condition

- A necessary condition

$$\Pr\left[D_i = 1 \mid X_i = x\right] \in (0, 1)$$

- In the subsample $\{X_i = x\}$, define $\mathcal{T}_x$, $\mathcal{C}_x$, N_x, $N_{x,1}$, and $N_{x,0}$ accordingly. Then

$$\begin{aligned}
ATE(x) &= E\left[y_{1i} - y_{0i} \mid X_i = x\right] \\
&\overset{\text{C.I.}}{=} E\left[y_{1i} - y_{0i} \mid D_i, X_i = x\right] \\
\widehat{ATE}(x) &= \frac{1}{N_{x,1}} \sum_{i \in \mathcal{T}_x} y_i - \frac{1}{N_{x,0}} \sum_{i \in \mathcal{C}_x} y_i \\
&= \frac{1}{N_x} \sum_{i=1}^{N} \left[\frac{D_i y_i}{N_{x,1} / N_x} - \frac{(1 - D_i) y_i}{N_{x,0} / N_x} \right]
\end{aligned}$$

Continuous X

- The above analysis is based on discrete X_i.
- If X is continuous, one way is to nonparametrically estimate

$$m_j(x) = E[y_{ji} \mid X_i = x], \text{ for } j \in \{0, 1\}$$

$$ATE(x) = m_1(x) - m_0(x)$$

- It involves nonparametric estimation techniques that we don't cover
- Average $ATE(X_i)$ over the support of X_i:

$$ATE = E[ATE(X_i)] = \int ATE(X_i) \, \mathrm{d}F(X_i)$$

Propensity Score

- **Propensity score**:

$$P[x] = \Pr[D_i = 1 \mid X_i = x] = E[D_i \mid X_i = x]$$

- In the treatment group

$$\begin{aligned} E\left[\frac{D_i y_i}{P[X_i]}\right] &\overset{\text{LIE}}{=} E\left[\frac{1}{P[X_i]} E[D_i y_{1i} \mid X_i]\right] \\ &\overset{\text{C.I.}}{=} E\left[\frac{1}{P[X_i]} E[D_i \mid X_i]\, E[y_{1i} \mid X_i]\right] \\ &= E[E[y_{1i} \mid X_i]] \overset{\text{LIE}}{=} E[y_{1i}] \end{aligned}$$

ATE Under Continuous X

- Similarly, in the control group $E\left[\frac{(1-D_i)y_i}{1-P[X_i]}\right] = E\left[y_{0i}\right]$
- The (unconditional) ATE is

$$ATE = E\left[\frac{D_i y_i}{P\left[X_i\right]} - \frac{\left(1-D_i\right)y_i}{1-P\left[X_i\right]}\right]$$

  - Important to ensure $P[X_i] \in (0,1)$. Logistic.

Linear Regression

- Linear regression

$$y_i = \alpha + X_i'\beta + \varepsilon_i$$

under the assumption $E[\varepsilon_i|X_i] = 0$ implies that if X_i is a scalar (the simplest case), then

$$\beta = \frac{E[y_i|X_i = x_1] - E[y_i|X_i = x_0]}{x_1 - x_0}$$

for any $x_0, x_1 \in \mathcal{X}$.

- The slope coefficient is the difference between two groups.

Interpretation of Linear Regression: I

- Linear regression is a purely statistical exercise, just like the comparison of two means.
- Example: Wage gap in gender

- $E[y_i|X_i]$ always exists, but the linear regression may not get the "causal" effect

Example: Omitted Variable Bias

- The causal model

$$y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i, \quad \text{with } E[\varepsilon_i \mid X_i, Z_i] = 0$$

but Z_i is omitted from regression

- Linear regression of y_i on X_i can still be implemented:

$$E[y_i \mid X_i] = \alpha + \beta X_i + \gamma E[Z_i \mid X_i] = \alpha + \theta X_i$$

where

$$\theta = \beta + \gamma \frac{\mathrm{Cov}(X_i, Z_i)}{\mathrm{Var}(X_i)}$$

Omitted Variable Bias (Cont.)

- The reduced form
$$y_i = \alpha + \theta X_i + u_i$$
ensures $E[u_i \mid X_i] = 0$ (under joint normality), but this is not a causal model.
- From the observational data, we cannot shift X_i without shifting u_i simultaneously.
- Causal question cannot be answered without further structures.

- Importance of the causal model: Only the causal relationship has policy implications.

Regression-based Causal Model

- Continue with the example: Change the year of education.
- "Keeping everything else equal, if a person's X_i is changed from x_0 to x_1, then β is the average change of y_i." This is a potential outcome claim.
- Given $D_i \in \{0, 1\}$ there are two potential outcomes

$$y_{0i} = \alpha_0 + X_i'\beta_0 + \varepsilon_{0i}$$
$$y_{1i} = \alpha_1 + X_i'\beta_1 + \varepsilon_{1i}$$

The linear assumption makes life easier under continuous X_i.

CATE

- The above model implies **heterogeneous treatment effect**

$$\Delta_i = (\alpha_1 - \alpha_0) + X_i'(\beta_1 - \beta_0) + (\varepsilon_{1i} - \varepsilon_{0i})$$

- By construction $E[\varepsilon_{1i} \mid X_i] = E[\varepsilon_{0i} \mid X_i] = 0$, and thus

$$ATE(X_i) = E[\Delta_i \mid X_i] = (\alpha_1 - \alpha_0) + X_i'(\beta_1 - \beta_0)$$

- If $\beta_1 = \beta_0$, then $ATE(X_i) = \alpha_1 - \alpha_0$ is a level change homogeneous to all people

Selection Bias

$$\begin{aligned} ATET(X_i) &= E[\Delta_i \mid D_i = 1, X_i] \\ &= ATE(X_i) + E[\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1, X_i] \end{aligned}$$

- Under unconfoundedness, $ATE(X_i) = ATET(X_i)$
- Otherwise, **selection bias** if

$$E[\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1, X_i] \neq 0$$

The individual knows $\varepsilon_{1i} - \varepsilon_{0i}$, and he elects to the treatment group because of that.

Self-Selection

- Example: College premium
  - Treatment: college entrance D_i
  - Unconfoundedness

$$(\varepsilon_{i,0}, \varepsilon_{i,1}) \perp D_i \mid X_i$$

  does not hold in general.

- The linear regression does not provide credible causal interpretation.
- Need other techniques to estimate causality.

Section 4

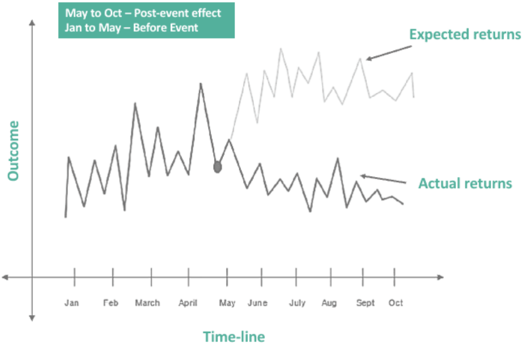
Quasi experiment

Regression Discontinuity Design

- A jump point due to some ad hoc policy cutoff
- Unconfoundedness is naturally satisfied: as if randomized.
- e.g. College premium
- Caveat: valid only for the subpopulation around the cutoff
- Suppose the cutoff is c
  - $y_{0i} = \alpha_0 + \beta_0(x_i - c) + \varepsilon_{0i}$
  - $y_{1i} = \alpha_1 + \beta_1(x_i - c) + \varepsilon_{1i}$

Event Study

- A time series topic, but very similar to treatment
- The same individual is observed over time $t = 1, 2, \ldots, T$
- An event happens at time $t = T_1$
  - Before event (control group)
  - After event (treatment group)



Implementing Event Study

- Let $D_t = \mathbb{I}(t \geq T_1)$
- Regression

$$y_t = \alpha + \beta D_t + \varepsilon_t$$

- Key assumption

$$E[\varepsilon_t | D_t] = 0 \text{ for all } t = 1, 2 \ldots, T$$

- Other control variables can be added into the regression
- My 2005 undergraduate thesis

Difference-in-Difference (DID)

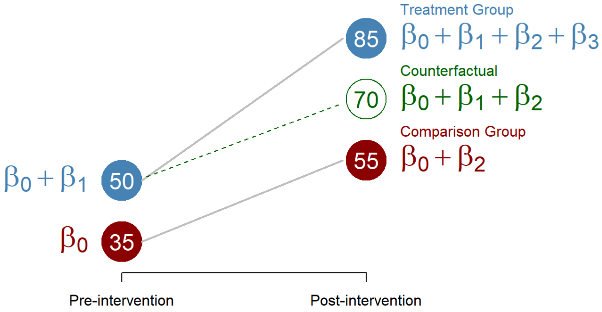
- Two groups, two periods (simple panel data)
  - The two groups are naturally different
  - **Parallel trend** over time. This is an assumption!
- One of the most popular empirical techniques
- Example: North Korea and South Korea

Implementing DID

- Two indicators D_i and D_t
- Regression is convenient for hypothesis testing

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 D_t + \beta_3 \cdot D_i D_t + \epsilon_{it}$$

- Other control variables can be added



Summary

- Potential outcome framework
- RCT
- Propensity score
- Regressions for CATE
- RDD
- DID

- Popular in academia as well as tech sector (link1), (link2)
    - Bytedance
    - Kuaishou