Modern Methods of Causal Inference

Cause & Effect

The effect of a cause

Cause of an effect X

- Smoking causes lung cancer at the age of 65 for men

  i $\quad z_i = \begin{cases} 1 & \text{if the person smokes} \\ 0 & \text{otherwise} \end{cases}$

  $y_i = \begin{cases} 1 & \text{if the person developed lung cancer} \\ 0 & \text{otherwise} \end{cases}$

  $(z_i = 1, y_i = 0)$

- Potential outcomes model (for mutable causes)

  i has two potential outcomes

  $y_i(1)$ if they were "treated"

  $y_i(0)$ if they were "not treated"

$(z_i = 1, y_i = 0) \longrightarrow y_i(1) = 0$

$(z_i = 0, y_i = 0) \longrightarrow y_i(0) = 0$

$\Big\} \; y_i(1) - y_i(0) = 0$

$(z_j = 0, y_j = 1) \longrightarrow y_j(0) = 1$

$(z_j = 1, y_j = 0) \longrightarrow y_j(1) = 0$

$\Big\} \; y_j(1) - y_j(0) = -1$

Individual causal effect

$\Big\} \; \frac{1}{2}\left[(y_i(1) - y_i(0)) + (y_j(1) - y_j(0))\right]$

Average of two ICEs

Assumption: Each individual smokes with probability $\frac{1}{2}$ $(P(z_i) = P(z_j))$ and z_i, z_j are independent of $y_i(0), y_i(1), y_j(0), y_j(1)$

$z_i = \begin{cases} 1 & \text{w/p } \frac{1}{2} \\ 0 & \text{w/p } \frac{1}{2} \end{cases}$

Consider the case where $z_i + z_j = 1$.

$$y_i z_i + y_j z_j = \left(y_i(1) z_i + y_i(0)(1 - z_i)\right) \cdot z_i + \left(y_j(1) z_j + y_j(0)(1 - z_j)\right) z_j$$

$$= y_i(1) z_i + y_j(1) z_j$$

$$E\left[y_i z_i + y_j z_j \,\middle|\, y_i(1), y_i(0), y_j(1), y_j(0), z_i + z_j = 1\right] = \frac{y_i(1) + y_j(1)}{2}$$

Try: $E\left[y_i(1 - z_i) + y_j(1 - z_j) \,\middle|\, y_i(1), y_i(0), y_j(1), y_j(0), z_i + z_j = 1\right] = \frac{y_i(0) + y_j(0)}{2}$

(*) $y_i z_i + y_j z_j - \left(y_i(1 - z_i) + y_j(1 - z_j)\right)$ is in expectation equal to $\frac{1}{2}\left[(y_i(1) - y_i(0)) + (y_j(1) - y_j(0))\right]$

Living in Atlanta:

Smoking	Lung cancer at 65: Yes	Lung cancer at 65: No	
Yes	150	5	155
No	2500	122	2622

$\frac{5}{150+5} - \frac{122}{2500+122} = -0.014$

proportion of smokers $\frac{155}{155+2622} = 0.056$

Living in Orlando:

Smoking	Lung cancer at 65: Yes	Lung cancer at 65: No	
Yes	580	350	930
No	2270	2000	4270

$\frac{350}{350+580} - \frac{2000}{2000+2270} = -0.092$

proportion of smokers $\frac{930}{930+4270} = 0.178$

Adding up the two tables above:

Smoking	Lung cancer at 65: Yes	Lung cancer at 65: No
Yes	730	355
No	4770	2122

$\frac{355}{355+730} - \frac{2122}{2122+4770} = 0.019$ → You should smoke (LOL)

Independence assumption fails

// leading to a "ridiculous" result

Covariates: Measurement not subject to the choice of the treatment the individual makes (X_i)

Assumption: Conditioned on X_i, the treatment is independent of the potential outcomes

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

("$\perp\!\!\!\perp$" denotes Philip Daneid's (1979) notation for independence)

Consider K binary covariates. If possible, 2n independently chosen people, exactly half of them are in $\{Z=1\}$

$$D_k = \frac{1}{n}\sum_{i=1}^{2n} z_i x_i^{(k)} - \frac{1}{n}\sum_{i=1}^{2n}(1-z_i)x_i^{(k)}, \quad k=1,2,\cdots,K$$

$$Pr\left(\max_{1\le k\le K} |D_k| \ge t\right) \le 2K\exp\left(-\frac{2t^2}{2/n}\right)$$

In theory we just need $X_i = (y_{i(1)}, y_{i(0)})$

Instead collect information about confounders X_i's that are informative about your potential outcomes and related to the treatment assignment

i, j exactly one is treated and the other is control

Denote the treated unit by t and the control unit by c $\quad y_t - y_c = (*)$

Pair Matching

$\mathcal{T} = \{t_1, \cdots, t_T\}$ are my treated units

$\mathcal{C} = \{c_1, \cdots, c_C\}$ are my control units

Assume $C = |\mathcal{C}| > T = |\mathcal{T}|$ (For example, more ppl don't smoke)

Then a pair matching is a function $m: \mathcal{T} \to \mathcal{C}$ which take t_i to $m(t_i) \in \mathcal{C}$.

δ_{tc} is a measure of the distance between X_t and X_c

- $\|X_t - X_c\|_2$
- $(X_t - X_c)^T S^{-1} (X_t - X_c)$ (Mahalanobis distance)
- $|F^{-1}(X_t) - F^{-1}(X_c)|$

$$\min \sum_{t \in \mathcal{T}} \delta_{t\, m(t)}$$

→ Nearest-neighbor approach – one with replacement match

→ Start with the smallest pair (t, c), then remove it and repeat T times [Greedy approach]
– one without replacement match

e.g. ε is a small number ($\varepsilon > 0$)

	c_1	c_2	c_3
t_1	ε	0	2ε
t_2	6ε	ε	0
t_3	6ε	2ε	ε

matrix of δ_{tc} values

Greedy approach: (t_1, c_2), (t_2, c_3), (t_3, c_1) | Total 6ε → Not ideal (6ε can be any large number)

600 pair matches $\frac{1}{600} \sum_{t \in \mathcal{T}} (y_t - y_{m(t)})$

$\mathrm{Var}(y_{m(t)}$ part) without repeat $\frac{\sigma^2}{600}$ < with repeat $\frac{\sigma^2}{600^2} \sum_{c \in \mathcal{C}} r_c^2$ $\quad r_c \geq 0, \sum_{\mathcal{C}} r_c = T = 600$

We don't want repeat control

The optimal pair matching problem

$$a_{tc} = \begin{cases} 0, & \text{if } t \text{ is not matched to } c \\ 1, & \text{if } t \text{ is matched to } c \end{cases}$$

$$\min \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} \delta_{tc} \cdot a_{tc} \quad \text{s.t.} \quad \begin{cases} \sum_t a_{tc} \leq 1 \text{ for all } c \\ \sum_c a_{tc} = 1 \text{ for all } t \end{cases}$$

Solving via the minimum network flow problem

① Network of warehouses

② Connection between the warehouses

③ Cost of moving a product between warehouses

④ Capacity constraint of each connection

$\mathcal{N}$ $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$

f_e is the cost per unit in connection $e \in \mathcal{E}$

W_e is the capacity constraint

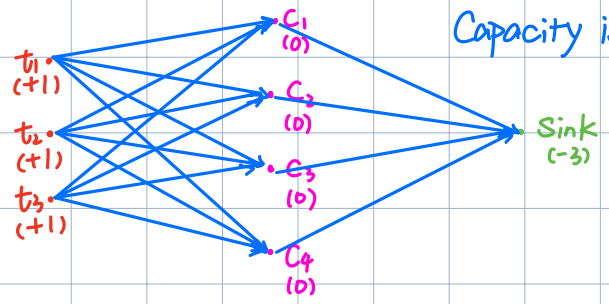
S_n is the $\overset{(+)}{\text{supply}}/\overset{(-)}{\text{demand}}$ of warehouse $n \in \mathcal{N}$

K_e for $e \in \mathcal{E}$ are integers (amount of product to be moved for connection e)

$$\min \sum f_e K_e \quad \text{s.t.} \begin{cases} K_e \le W_e \\ -S_n = \sum_{e=(n,m) \text{ for any } m} K_e - \sum_{e=(m,n) \text{ for any } m} K_e \end{cases}$$

(going out: $e=(n,m)$ for any m; coming in: $e=(m,n)$ for any m)

$\mathcal{T} = \{t_1, t_2, t_3\}$ $\mathcal{C} = \{c_1, c_2, c_3, c_4\}$



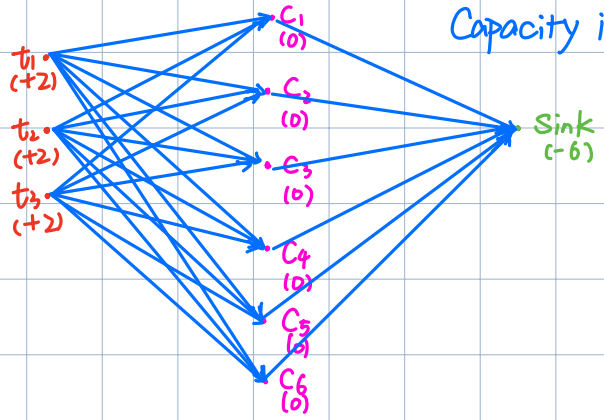
RELAX-IV
- Optmatch in R
- C call relax

$$\left(|\mathcal{N}| \times |\mathcal{E}_a| + (|\mathcal{E}_a| + |\mathcal{N}| \log |\mathcal{N}|) \sum_{n \in \mathcal{N}} |S_n|\right) \le \text{const} \cdot C^3$$

$|\mathcal{N}|$: T; $|\mathcal{E}_a|$: $TC+C \le 4C^3$; $|S_n|$: $\le C^2$; $\sum |S_n|$: $2T \le \sim C^3$

1-k matching

$\mathcal{T} = \{t_1, t_2, t_3\}$ $\mathcal{C} = \{c_1, c_2, c_3, c_4, c_5, c_6\}$



Variable ratio matching

(one to various numbers)

$\lfloor \hat{P}(z_{t_1} = 0 \mid X_{t_1}) \times C \rfloor$

Full match

matching is observational design
matching is non-experimental design

$(\tau_s, \mathcal{C}_s)$ for $s = 1, \cdots, S$

$|\tau_s| = 1$ or $|\mathcal{C}_s| = 1$

$\bigcup_{s=1}^{S} \tau_s = \tau \quad \bigcup_{s=1}^{S} \mathcal{C}_s = \mathcal{C}$

Consider $\tau = \{t_1, \cdots, t_T\}$, $\mathcal{C} = \{c_1, \cdots, c_C\}$

$m: \tau \to \mathcal{C}$ defines a pair matching

- 1-1 match:

$(t_1, m(t_1)), (t_2, m(t_2)), \cdots, (t_T, m(t_T))$

$y_{t_1} - y_{m(t_1)}, y_{t_2} - y_{m(t_2)}, \cdots, y_{t_T} - y_{m(t_T)}$

- 1-2 match

$(t_1, m_1(t_1), m_2(t_1)), \cdots, (t_T, m_1(t_T), m_2(t_T))$

$y_{t_1} - \frac{1}{2}(y_{m_1(t_1)} + y_{m_2(t_1)}), \cdots, y_{t_T} - \frac{1}{2}(y_{m_1(t_T)} + y_{m_2(t_T)})$

- 1 - variable ratio

Potential Outcomes notation

$y_i(1), y_i(0) \to y_i(1) - y_i(0)$ individual treatment effect for i

by matching, we try to impute this unknown term

Think of $\hat{y}_{t_i}(0) = y_{m(t_i)}$

$y_{t_i}(0)$ is <u>counterfactual outcome</u>

We constructed a counterfactual estimate using matching

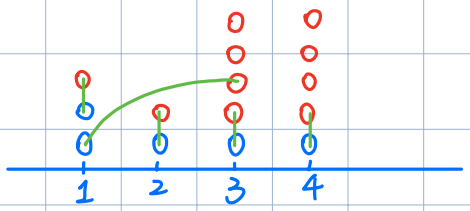
$\frac{1}{T}\sum_{t=1}^{T}(y_t - y_{m(t)})$ is our final matching estimator. What is it estimating?

$\frac{1}{T}\sum_{t=1}^{T}(y_{t(1)} - y_{t(0)})$ corresponds to $E(y_i(1) - y_i(0) | z_i = 1)$

When the treatment effect is different across the population,
$E(y_i(1) - y_i(0))$, $E(y_i(1) - y_i(0) | z_i = 1)$, $E(y_i(1) - y_i(0) | z_i = 0)$ are different

△ $\frac{1}{T}\sum_{t=1}^{T}(y_t - y_{m(t)})$ is NOT a consistent estimator of $E(y_{i(1)} - y_{i(0)} \mid z_i = 1)$

$T < C$, $2T = C$



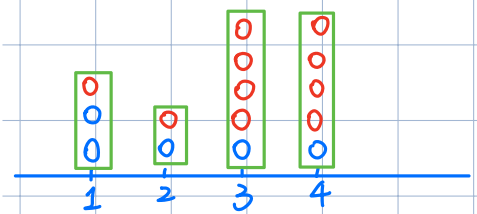
$T = 5$ $C = 10$

We need $Pr(z_i = 1 \mid X) < \frac{1}{2}$ for all X.

Reference: Savje (2022) On inconsistency of matching without replacement Biometrika

Full Matching

In each matched set, there is either exactly one treatment unit or exactly one control unit. It will give consistent estimator for $E(y_{i(1)} - y_{i(0)})$



$T = 5$ $C = 10$

$E(y_{i(1)} - y_{i(0)})$ → Average treatment effect (ATE)

$E(y_{i(1)} - y_{i(0)} \mid z_i = 1)$ → Average treatment effect on the treated (ATT)

$E(y_{i(1)} - y_{i(0)} \mid z_i = 0)$ → Average treatment effect on the control (ATC)

Balance

$\frac{1}{T}\sum g(X_t) - \frac{1}{T}\sum g(X_{m(t)})$ is close to zero for many functions g.

- $g(x) = x \Rightarrow \bar{X}_t - \bar{X}_{m(t)}$

Cardinality Matching targeting for balance.

$$\mathcal{B}_k : -b_k \sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc} \le \sum_{t=1}^{T}\sum_{c=1}^{C}(g_k(X_t) - g_k(X_c))a_{tc} \le b_k \sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc}$$

where $a_{tc} = 1_{\{t \text{ is matched to } c\}}$

Solve for a_{tc}

$$\max \sum_{t=1}^{T}\sum_{c=1}^{C} a_{tc} \quad \text{s.t.} \begin{cases} \mathcal{B}_k\text{'s} \\ a_{tc}\text{'s are } 0,1 \\ \sum_{c=1}^{C} a_{tc} \le 1, \ \sum_{t=1}^{T} a_{tc} \le 1 \end{cases}$$

Standardized Mean Difference

$$\left| \frac{\bar{X}_t - \bar{X}_{m(t)}}{\sqrt{\frac{S_t^2 + S_c^2}{2}}} \right|$$

More on matching

Assumption: $(y_i(1), y_i(0)) \perp\!\!\!\perp z_i \mid X_i$ (1)

Suppose I can find a lower dim $b(X_i)$ so that $(y_i(1), y_i(0)) \perp\!\!\!\perp z_i \mid b(X_i)$ (2)

Theorem: There is a one dim covariate $b(X_i)$ for which (1) implies (2).

Lemma1: Suppose $z_i \perp\!\!\!\perp X_i \mid b(X_i)$ for some $b(X_i)$, then (1) implies (2).

Lemma2: Let $\lambda(x) = Pr(z_i = 1 \mid X_i = x)$. Then $z_i \perp\!\!\!\perp X_i \mid b(X_i)$ for some function $b(X_i)$ iff $\lambda(X_i) = f(b(X_i))$ for some f

* In light of the two lemmas, the theorem is proved by showing that

$$(y_i(1), y_i(0)) \perp\!\!\!\perp z_i \mid \lambda(X_i)$$

Pf of lemma2:

- "if": Suppose $\lambda(X_i) = f(b(X_i))$.

  $Pr(z_i = 1 \mid X_i, b(X_i)) = Pr(z_i = 1 \mid X_i) = \lambda(X_i)$

  $Pr(z_i = 1 \mid b(X_i)) = E(Pr(z_i = 1 \mid X_i, b(X_i)) \mid b(X_i)) = E(\lambda(X_i) \mid b(X_i)) = f(b(X_i)) = \lambda(X_i)$

  Hence $z_i \perp\!\!\!\perp X_i \mid b(X_i)$

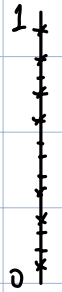
- "only if": Suppose $z_i \perp\!\!\!\perp X_i \mid b(X_i)$ but $\lambda(x)$ is NOT $f(b(x))$ for any f.

  Then there is a pair x_1, x_2 so that $b(x_1) = b(x_2)$ but $\lambda(x_1) \neq \lambda(x_2)$, i.e., $Pr(z_i = 1 \mid X_i = x_1) \neq Pr(z_i = 1 \mid X_i = x_2)$ Hence the contradiction.

Pf of lemma 1:

Sufficient to prove $Pr(z_i = 1 \mid y_i(1), y_i(0), b(X_i)) = Pr(z_i = 1 \mid b(X_i))$.

$$\begin{aligned} LHS &= E[Pr(z_i = 1 \mid y_i(1), y_i(0), X_i, b(X_i)) \mid y_i(1), y_i(0), b(X_i)] \\ &= E[Pr(z_i = 1 \mid X_i, b(X_i)) \mid y_i(1), y_i(0), b(X_i)] \\ &= E[\lambda(X_i) \mid y_i(1), y_i(0), b(X_i)] \overset{\text{lemma2}}{=\!=} \lambda(X_i) \end{aligned}$$



m data points (X_i, z_i, y_i)

Fit a logistic regression of z_i on X_i

$\hat{\lambda}(X_i) = \hat{Pr}(z_i = 1 | X_i)$

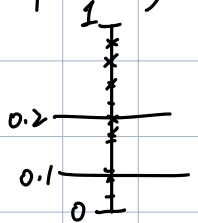
How to incorporate the fitted values in matching?

$$\delta'_{tc} = \delta_{tc} / \mathbb{1}(|\hat{\lambda}(X_t) - \hat{\lambda}(X_{m(t)})| < a2) = \begin{cases} \delta_{tc}, & \text{if } |\hat{\lambda}(X_t) - \hat{\lambda}(X_{m(t)})| < 0.2 \\ \infty, & \text{o.w.} \end{cases}$$

$\lambda(X_i) = Pr(z_i = 1 | X_i)$ is "Propensity Score"

δ'_{tc} is called "Propensity Score Caliper"

Propensity Score Stratification



If using 0.2, we have $\frac{1}{0.2} = 5$ groups (strata)

The strata are defined according to the quantiles of propensity score

Weighting

$$ATT = E(y_i(1) - y_i(0) | z_i = 1)$$

$$\underbrace{\frac{1}{T}\sum_{t \in \mathcal{I}} y_t}_{\hat{E}(y_i(1) | z_i = 1)} - \underbrace{\frac{1}{T}\sum_{t \in \mathcal{I}} y_{m(t)}}_{\hat{E}(y_i(0) | z_i = 1)}$$

$$\frac{1}{T}\sum_{c \in \mathcal{C}} w_c y_c$$

Suppose we want an estimator of $E(y_i(0) | z_i = 1)$ of the form $\frac{1}{T}\sum_{c \in \mathcal{C}} w_c y_c$.

$$\mu_0 = E(z_i y_i(0)) \left[= E(z_i y_i(0) | z_i = 1) Pr(z_i = 1) + E(z_i y_i(0) | z_i = 0) Pr(z_i = 0) \right.$$
$$\left. = E(y_i(0) | z_i = 1) \cdot p + 0 \right]$$

using $\hat{\mu}_0 = \frac{1}{n}\sum_{c \in \mathcal{C}} w_c y_c = \frac{1}{n}\sum_{i=1}^{n} (1 - z_i) w_i y_i$ when $n = T + C$.

Let $m_0(X_i) = E(y_i(0) | X_i)$.

$$\mu_0 = E(z_i y_i(0)) = E[E(z_i y_i(0) | X_i)] = E[Pr(z_i = 1 | X_i) E(y_i(0) | X_i)] = E(\lambda(X_i) m_0(X_i))$$

Let $\varepsilon_i = y_i(0) - m_0(X_i)$.

$$\hat{\mu}_0 - \mu_0 = \frac{1}{n}\sum_{i=1}^{n}(1-z_i)w_i m_0(X_i) - \frac{1}{n}\sum_{i=1}^{n} z_i m_0(X_i)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}(1-z_i)w_i \varepsilon_i \quad \text{---> has } E(\cdot|X_i)=0$$
$$+ \frac{1}{n}\sum_{i=1}^{n} z_i m_0(X_i) - E(z_i y_i(0)) \quad \text{---> has } E(\cdot|X_i)=0$$

We can estimate the above bias by minimizing

$$\left|\frac{1}{n}\sum_{i=1}^{n}(1-z_i)w_i m_0(X_i) - \frac{1}{n}\sum_{i=1}^{n} z_i m_0(X_i)\right| \text{, if } m_0(\cdot) \text{ is known}$$

$$\max_{m\in\mathcal{M}} \left|\frac{1}{n}\sum_{i=1}^{n}(1-z_i)w_i m(X_i) - \frac{1}{n}\sum_{i=1}^{n} z_i m(X_i)\right| \text{, over } m \in \mathcal{M}$$

Suppose $m(X_i) = X_i$ one dimensional.

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(1-z_i)w_i X_i}_{\frac{C}{n}\bar{X}_{w,\text{control}}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} z_i X_i}_{\frac{T}{n}\bar{X}_t}$$

Covariate Balancing Methods

We say that $f(x)$ is balanced in a pair match if

$$\frac{1}{T}\sum_{t\in\mathcal{T}} f(x_t) \approx \frac{1}{T}\sum_{t\in\mathcal{T}} f(x_{m(t)})$$

In full match we want to balance for the population average.

Balance targeted procedure

- Maximum Cardinality Matching

$$B_k: \; -b_k \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{tc} \le \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{tc} g_{tck} \le b_k \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{tc} \qquad a_{tc} = 1_{\{t \text{ is matched to } c\}}$$

$$g_{tck} = f_k(x_t) - f_k(x_c)$$

$$\max \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{tc} \text{ subject to } B_k \text{ for } k=1,\cdots,K \text{ and } a_{tc}\in\{0,1\},\ \sum_c a_{tc} \le 1,\ \sum_t a_{tc} \le 1$$

Recall propensity score $\pi(X)$

We have data $(y_i(1), y_i(0), X_i, z_i)$ i.i.d. from some population. $y_i = \begin{cases} y_i(1), & \text{if } z_i=1 \\ y_i(0), & \text{if } z_i=0 \end{cases}$

$\pi(X_i) = P_r(z_i = 1 | X_i)$ → propensity score

$z_i \perp\!\!\!\perp X_i \mid \pi(X_i)$

Ideally we can estimate it from the data using some flexible regression model of z_i on X_i.

$$E\left(\frac{z_i f(x_i)}{\pi(x_i)}\right) = E\left(E\left(\frac{z_i f(x_i)}{\pi(x_i)} \middle| X_i\right)\right)$$
$$= E\left(\frac{f(x_i)}{\pi(x_i)} \Pr(z_i=1|X_i)\right)$$
$$= E(f(x_i)) \text{ as long as } \pi(x_i)>0 \text{ with prob. 1}$$

For any function f, $\frac{1}{n}\sum_{i=1}^{n} \frac{z_i f(x_i)}{\pi(x_i)} \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

Consider $f(x_i) = E(y_i|X_i, z_i=1)$.

$$E\left(\frac{z_i E(y_i|x_i, z_i=1)}{\pi(x_i)}\right) = E\left(E(y_i|X_i, z_i=1)\right)$$
$$\| \qquad\qquad \|$$
$$E\left(\frac{z_i y_i}{\pi(x_i)}\right) \qquad E(y_i(1)) \quad \text{(exercise)}$$

<u>Lemma.</u> $E\left(\frac{z_i y_i}{\pi(x_i)}\right) = E(y_i(1))$ for $\pi(x_i)>0$ a.s.

If we want to estimate $E(y_i(1))$, we use $\frac{1}{n}\sum_{i=1}^{n} \frac{z_i y_i}{\hat{\pi}(x_i)}$

Similarly we can estimate $E(y_i(0))$ by $\frac{1}{n}\sum_{i=1}^{n} \frac{(1-z_i)y_i}{1-\hat{\pi}(x_i)}$ for $\pi(x_i)<1$ a.s.

$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{z_i y_i}{\hat{\pi}(x_i)} - \frac{(1-z_i)y_i}{1-\hat{\pi}(x_i)}\right]$ is a reasonable estimator for $ATE = E(y_i(1)-y_i(0))$ when $\pi(x_i)\in(0,1)$ a.s.

Not stable when $\hat{\pi}(x_i)$ is close to 0 or 1.

<u>Inverse Propensity Weighted estimator (IPW estimator)</u>

Suppose we estimate $\hat{\pi}(x_i)$ to make sure that $\frac{1}{n}\sum_{i=1}^{n} \frac{z_i f(x_i)}{\hat{\pi}(x_i)} \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

Choose a model $\pi_\beta(x)$ for $\Pr(z_i=1|x)$

The corresponding likelihood is $\prod_{i=1}^{n}[\pi_\beta(x_i)]^{z_i}[1-\pi_\beta(x_i)]^{1-z_i}$

$$\text{s.t.} \quad \left|\frac{1}{n}\sum_{i=1}^{n} \frac{z_i f_k(x_i)}{\pi_\beta(x_i)} - \frac{1}{n}\sum_{i=1}^{n} f_k(x_i)\right| < b_k, \quad k=1,\cdots,K$$

→ <u>Covariate Balancing Propensity Score</u>

- A solution to very small or very large PS estimator is <u>trimming</u> the estimated propensity score.

- $\sum_{i=1}^n z_i w_i f_k(x_i) \approx \frac{1}{n}\sum_{i=1}^n f_k(x_i)$ for $w_i = \frac{1}{n\pi(x_i)}$

Stable weight: $w' = (\frac{1}{T}, \cdots, \frac{1}{T})$ is a target stable weight

If we can find w that is NOT too far from the stable weights and it satisfies

B_{kw}: $\left|\sum_{i=1}^n z_i w_i f_k(x_i) - \frac{1}{n}\sum_{i=1}^n f_k(x_i)\right| \le b_k$ for $k = 1, \cdots, K$.

Minimize $\sum_{i=1}^n (w_i - w_i')^2$ such that B_{kw} holds for $k = 1, \cdots, K$

Then estimate $E(y_{i(1)})$ by $\sum_{i=1}^n z_i w_i y_i$

Similarly, we can estimate $E(y_{i(0)})$

An alternative objective is to maximize $\sum_{i=1}^n -w_i \log w_i$

※ 09/14/23 · Identification in causal inference

· Instrumental variables

n individuals drawn i.i.d. from a population. We want to estimate a population quantity using our data

Full data: $(y_{i(1)}, y_{i(0)}, X_i, z_i) \overset{iid}{\sim} P$

CATE

$\tau = E(y_{i(1)} - y_{i(0)} | z_i = 1) \overset{or}{=} E(y_{i(1)} - y_{i(0)} | z_i = 0) \overset{or}{=} E(y_{i(1)} - y_{i(0)} | X_i = n) \overset{or}{=} E(y_{i(1)} - y_{i(0)})$

Observed data: $(y_i, X_i, z_i) \overset{iid}{\sim} P_{obs}$ where $y_i = y_{i(1)} z_i + y_{i(0)}(1 - z_i)$

Identification: A parameter is identifiable if it can be expressed as a function of the observed data distribution P_{obs}.

$\tau = g(P_{obs})$ In that case, g is the identifier function. Then we can give a point estimate of τ as $\hat{\tau} = g(\hat{P}_{obs})$

We can get the sampling distribution of $\hat{\tau}$ by bootstrapping.

In general, none of those are identifiable without additional assumptions.

Assumption: $(y_{i(1)}, y_{i(0)}) \perp\!\!\!\perp z_i | X_i$ (Ignorability, Exchangeability, No-??? confounders)

$\tau = E(y_{i(1)} - y_{i(0)}) = E(y_{i(1)}) - E(y_{i(0)})$

$E(y_{i(1)}) = E[E(y_{i(1)} | X_i)] \overset{\text{assumption}}{=\!=} E[E(y_{i(1)} | X_i, z_1 = 1)] = E[E(y_i | X_i, z_1 = 1)]$

{ $y_{i(1)} \perp\!\!\!\perp z_i | X_i$, $P(z_i = 1 | X_i) > 0$ a.s.

Lemma: $\tau = E(y_{i(1)} - y_{i(0)})$ is identifiable under the assumptions { $E(y_{i(1)} | X_i, z_i = 1) = E(y_{i(1)} | X_i)$, $E(y_{i(0)} | X_i, z_i = 0) = E(y_{i(0)} | X_i)$

as $\tau = E[E(y_i|x_i, z_i=1) - E(y_i|x_i, z_i=0)]$

* $0 < Pr(z_i=1|x_i) < 1$ a.s. overlap Since it is equivalent to say that $f(x_i|z_i=1)$ & $f(x_i|z_i=0)$ have the same support.

g-formula, back-door adjusted,
outcome model best identification

$$E\left[\frac{z_i y_i}{\pi(x_i)}\right] = E\left[E\left(\frac{z_i y_i}{\pi(x_i)}|x_i\right)\right] = E\left[\frac{1}{\pi(x_i)}E(z_i y_i|x_i)\right]$$
$$= E\left[\frac{1}{\pi(x_i)}E(z_i y_{i(1)}|x_i)\right]$$
$$= E\left[\frac{1}{\pi(x_i)}E(z_i|x_i)E(y_{i(1)}|x_i)\right]$$
$$= E[E(y_{i(1)}|x_i)]$$

Proposition: $\tau = E(y_{i(1)} - y_{i(0)})$ is identifiable under the assumption $(y_{i(1)}, y_{i(0)}) \perp\!\!\!\perp z_i | x_i$ and $0 < Pr(z_i=1|x_i) < 1$ as $E[E(y_i|z_i=1, x_i) - E(y_i|z_i=0, x_i)]$

or $E\left(\frac{z_i y_i}{\pi(x_i)} - \frac{(1-z_i)y_i}{1-\pi(x_i)}\right)$

* $E(y_{i(1)}) = E\left[E(y_i|z_i=1, x_i) + \frac{z_i(y_i - E(y_i|z_i=1, x_i))}{\pi(x_i)}\right]$ → the third formula exercise

Theorem: Consider any function $h(x)$ and $e(x)$,

$$E(y_{i(1)}) = E\left[h(x_i) + \frac{z_i(y_i - h(x_i))}{e(x_i)}\right]$$ → Doubly Robust Identification

as long as one of the following is true:

(i) $h(x_i) = E(y_i|z_i=1, x_i)$; (ii) $e(x_i) = \pi(x_i)$

Pf: Assume (ii), $E\left[h(x_i) + \frac{z_i(y_i - h(x_i))}{\pi(x_i)}\right] = E\left[\frac{h(x_i)(\pi(x_i) - z_i)}{\pi(x_i)} + \frac{z_i y_i}{\pi(x_i)}\right]$

$$= E\left[\frac{h(x_i)(\pi(x_i) - z_i)}{\pi(x_i)}\right] + E\left[\frac{z_i y_i}{\pi(x_i)}\right]$$
$$= 0 + E(y_{i(1)})$$

Proposition: $\tau = E(y_{i(1)} - y_{i(0)}|\tilde{x}_i = \tilde{x})$ where $\tilde{x}_i$ is a subvector of x_i. Under Ignorability and overlap, we can identify τ by

$$\int_{x:\tilde{x}} [E(y_i|z_i=1, x_i=x) - E(y_i|z_i=0, x_i=x)] dP_{\text{marginal}}(x)$$
(integrating out $\tilde{x}$)

$(y_i(1), y_i(0)) \perp\!\!\!\perp z_i \mid X_i$ might fail to hold:

Example: Effect of NICU on mortality after birth

X_i = mother choose to give birth in a hospital with nicu

U_i = the intent of the mother to go hospital with a nicu

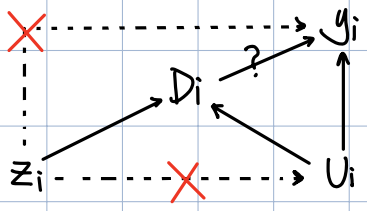
$(y_i(1), y_i(0)) \perp\!\!\!\perp z_i \mid X_i, U_i$ → unmeasurable

$(y_i(1), y_i(0)) \not\perp\!\!\!\perp z_i \mid X_i$

Example: Effect of strong/harsh sentencing in a criminal case on recitivization

Instrumental Variables Method

(D_i, X_i, y_i) $z_i \leftarrow$ an instrument



{ (a) the instrument is associated with the treatment
(b) the instrument is NOT associated with the unmeasured covariates
(c) the instrument has no direct effect on the outcome

y_i = whether the baby was healthy

X_i = observed covariates

U_i = mother's intent to go to a hospital with nicu

D_i = whether the baby was treated in a nicu

z_i = whether the mother lived more than 30 min away from the nearest hospital with nicu

Notation: $z_i = 0$ or 1

$$D_i(z_i) \begin{cases} D_i(1) & \text{if } z_i = 1 \\ D_i(0) & \text{if } z_i = 0 \end{cases}$$

$$y_i(z_i, d_i) \begin{cases} y_i(0,0) \\ y_i(0,1) \\ y_i(1,0) \\ y_i(1,1) \end{cases}$$

$$(z_i, D_i(1), D_i(0), y_i(1,1), y_i(1,0), y_i(0,1), y_i(0,0), X_i) \overset{iid}{\sim} P$$

$$(z_i, D_i, y_i, X_i) \overset{iid}{\sim} P_{obs}$$

Assumptions: ① If $z_i = z_i'$, then $D_i(z_i) = D_i(z_i')$

If $z_i = z_i'$, $d_i = d_i'$, then $y_i(z_i, d_i) = y_i(z_i', d_i')$

② $E(D_i(1) \mid X_i) \neq E(D_i(0) \mid X_i)$ a.s.

③ $z_i \perp\!\!\!\perp (y_i(1,1), y_i(1,0), y_i(0,1), y_i(0,0)) \mid X_i$

④ $y_i(z_i, d_i) = y_i(z_i', d_i)$

Examples:

- Randomized Encouragement

  The effect of drinking while pregnant on infant birthweight

  Instrument $\{1, \cdots, n\}$ $z_i = \{1, 0$

- Natural Experiment

  The effect of going to war on future wages.

  Vietnam war lottery

  $z_i = \begin{cases} 1, & \text{if their name came up in lottery} \\ 0, & \text{o.w.} \end{cases}$

- Distance to speciality care

  The effect of being treated in a NICU on infant health.

  $z_i = \begin{cases} 1, & \text{if the mother lived less than 30 mins from a nicu hospital} \\ 0, & \text{o.w.} \end{cases}$

- Preference for the treatment

  The effect of harsh sentencing a recidivism.

  $z_i = \begin{cases} 1, & \text{if the judge in the case had a history of harsh sentencing} \\ 0, & \text{o.w.} \end{cases}$

- The effect of food stamp on CVD

  $z_i = \begin{cases} 1, & \text{if the person is from Florida} \\ 0, & \text{if the person is from California} \end{cases}$

- Mendallian Randomization

z_i = gene expression

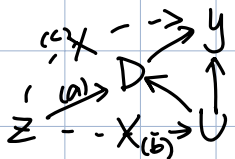
The effect of drinking on infant birthweight

$z_i = \begin{cases} 1, & \text{if ALDH2 has null expression} \\ 0, & \text{o.w.} \end{cases}$

Compliance Classification

z_i	$D_i(z_i)$			
1	1	1	0	0
0	1	0	1	0
	Always takers	Compliers	Defiers	Never takers
	at	co	de	nt

※ 09/21/23 · Instrumental variables



- Identification theorem
- Violation of assumptions
- Additional discussion

Potential Outcomes

$y(z,d) = y(z',d)$ for all z, z', d.

We simplify our notation to $y(d)$

$$ATE = E(y(1)-y(0))$$
$$= Pr(nt)\,E(y(1)-y(0)|nt) + Pr(co)\,E(y(1)-y(0)|co) + Pr(de)\,E(y(1)-y(0)|de) + Pr(at)\,E(y(1)-y(0)|at)$$

"—" unable to measure b/c changing z doesn't change D for nt and at

Observed data:

Y, Z, D all are binary. There is no X.

We have data based on a multinomial distribution on $\{0,1\}^3$. We have 8 parameters we can estimate.

Parameters of the full data

$Pr(z=1)$

$Pr(co)$ $Pr(de)$ $Pr(at)$ ~~$Pr(nt)$~~

$E(y^1|at)$ ~~$E(y^0|at)$~~

$E(y^1|co)$ $E(y^0|co)$

~~$E(y^1|nt)$~~ $E(y^0|nt)$

$E(y^1|de)$ $E(y^0|de)$

} 10 in total

↳ are ruled out by the monotonicity assumption

- Monotonicity: $D(z) \geq D(z')$ for any $z \geq z$

  For binary z, $D(1) \geq D(0)$. This implies no defiers.

Our estimand $E(y(1)-y(0)|co)$ → Complier ATE (CATE) → Local ATE (LATE)

Identification Theorem: Assume there is no X.

$$E(y(1)-y(0)|co) = \frac{E(y|z=1) - E(y|z=0)}{E(D|z=1) - E(D|z=0)}$$

Pf: $E(y|z=1) - E(y|z=0)$

$= E(y(1)D(1) + y(0)(1-D(1)) \mid z=1) - E(y(1)D(0) + y(0)(1-D(0)) \mid z=0)$

$\overset{z \perp\!\!\!\perp y(1),y(0),D(1),D(0)}{=} E(y(1)D(1) + y(0)(1-D(1))) - E(y(1)D(0) + y(0)(1-D(0)))$

$= E[(y(1)-y(0))(D(1)-D(0))]$

$\overset{\text{Monotonicity}}{=} E(y(1)-y(0)|D(1)-D(0)=1) \, Pr(D(1)-D(0)=1)$

$\overset{\text{Using association of } z \text{ and } D}{=} E(y(1)-y(0)|co) \cdot (E(D|z=1) - E(D|z=0))$ □

For any given person i, we only see z_i, $D_i(z_i)$.

$D_i(1)$ and $D_i(0)$ are not simultaneously observable. Hence we do not know their compliance group.

Understanding LATE

Option 1: If the treatment effect is homogeneous across compliance groups

Option 2: If we can explain who compliers are.

Lemma: For any function $h(y, x)$,

$$E(h(y(1), X) \mid co) = \frac{E(h(y, X)D \mid z=1) - E(h(y, X)D \mid z=0)}{E(D \mid z=1) - E(D \mid z=0)}$$

$$E(h(y(0), X) \mid co) = \frac{E(h(y, X)(1-D) \mid z=1) - E(h(y, X)(1-D) \mid z=0)}{E(1-D \mid z=1) - E(1-D \mid z=0)}$$

Pf: $E(h(y,X)D \mid z=1) - E(h(y,X)D \mid z=0)$

$= E(h(y(1),X)D(1) \mid z=1) - E(h(y(0),X)D(0) \mid z=0)$

$= E[E(h(y(1),X)D(1) \mid z=1, X) - E(h(y(0),X)D(0) \mid z=0, X)]$

$= E(h(y(1),X)D(1)) - E(h(y(1),X)D(0))$

$= E(h(y(1),X)(D(1)-D(0)))$

$= E(h(y(1),X) \mid co)(Pr(D=1 \mid z=1) - Pr(D=1 \mid z=0))$ □

Using the lemma, we can estimate

$E(y(1) \mid co)$

$E(y(1) \mid at) = E(y \mid z=0, D=1)$

$E(y(0) \mid co)$

$E(y(0) \mid nt) = E(y \mid z=1, D=0)$

Consider a linear regression model,

$$y(d) = \alpha_0 + \alpha_1 d + \alpha_2 X + \alpha_3 U + \varepsilon_d$$

Assume $\alpha_2 = 0$.

$$y(d) = \alpha_0 + \alpha_1 d + \alpha_3 U + \varepsilon_d$$

What we see: $y = \alpha_0 + \alpha_1 D + \alpha_3 U + \varepsilon_0 + (\varepsilon_1 - \varepsilon_0)D$

Assume $E(\varepsilon_d \mid U) = 0$. $E(\varepsilon_d \mid D, U) = 0$

$E(y \mid D=1) = \alpha_0 + \alpha_1 + \alpha_3 E(U \mid D=1) + E(\varepsilon_1 \mid D=1)$

$E(y \mid D=0) = \alpha_0 + \alpha_3 E(U \mid D=0) + E(\varepsilon_0 \mid D=0)$

$$E(y|D=1) - E(y|D=0) = \alpha_1 + \underbrace{\alpha_3[E(U|D=1) - E(U|D=0)]}_{\text{bias}} + \underbrace{E(\varepsilon_1|D=1) - E(\varepsilon_0|D=0)}_{=0}$$

Z is an instrument $E(U|Z=1) - E(U|Z=0) = 0$ (*)

$$E(\varepsilon_d|Z=1) = E(\varepsilon_d|Z=0) = 0$$

Then, $E(y|Z=1) - E(y|Z=0)$

$$= \alpha_1[E(D|Z=1) - E(D|Z=0)] + \alpha_3[E(U|Z=1) - E(U|Z=0)]$$

In this linear model, if assumption (b) does not hold (so that (*) does not necessarily hold) the bias of LATE estimator is $\alpha_3 \cdot \dfrac{E(U|Z=1) - E(U|Z=0)}{E(D|Z=1) - E(D|Z=0)}$

The bias of the previous estimator is $\alpha_3[E(U|D=1) - E(U|D=0)]$

$$y_{(d,z)} = \alpha_0 + \alpha_1 d + \alpha_2 z + \alpha_3 U + \varepsilon_d$$

(a) $E(D|Z=1) \neq E(D|Z=0)$

$$\frac{1}{\sum_i 1_{(Z_i=1)}} \sum_i 1_{(D_i=1, Z_i=1)} - \frac{1}{\sum_i 1_{(Z_i=0)}} \sum_i 1_{(D_i=1, Z_i=0)}$$

Theorem: (y_i, D_i, Z_i) are iid from a population.

(a) $$\frac{E(y_i|Z_i=1) - E(y_i|Z_i=0)}{E(D_i|Z_i=1) - E(D_i|Z_i=0)} = \frac{Cov(y_i, Z_i)}{Cov(D_i, Z_i)}$$

(b) $$\sqrt{n}\left[\frac{\sum_i Z_i(y_i - \bar{y})}{\sum_i Z_i(D_i - \bar{D})} - \frac{Cov(y_i, Z_i)}{Cov(D_i, Z_i)}\right] \xrightarrow{d} N\left(0, \frac{E[\eta_i^2 (Z_i - EZ_i)^2]}{[Cov(D_i, Z_i)]^2}\right)$$

where $\eta_i = y_i - Ey_i - \dfrac{Cov(y_i, Z_i)}{Cov(D_i, Z_i)}(D_i - ED_i)$.

09/26/23 · Classical IV methods

$$y_i = D_i\beta + X_i'\alpha + \varepsilon_i, \quad E(\varepsilon_i)=0$$

$\beta \leftarrow$ estimand

$$\Rightarrow \begin{cases} E[y(1)|X_i] = \beta + X_i'\alpha \\ E[y(0)|X_i] = X_i'\alpha \end{cases} \quad \text{assuming } E(\varepsilon_i|X_i)=0$$

Run a least square of y on D_i, X_i

Get the coefficient of D_i from the regression

Is the coefficient consistent for β?

Only consistent if $E(\varepsilon_i|D_i,X_i)=0$ (Exogeneity)

If not, aka $E(\varepsilon_i|D_i,X_i)\neq 0$, it is called Endogeneity.

$$\begin{cases} (2) y_i = D_i\beta + X_i'\alpha + \varepsilon_i \\ (1) D_i = Z_i'\lambda + X_i'\tau + \eta_i \end{cases} \quad \text{assuming} \quad \begin{cases} E(\varepsilon_i|Z_i,X_i)=0 \\ E(\eta_i|Z_i,X_i)=0 \end{cases}$$

(2): Second stage or structured model

(1): First stage model

Goal is to estimate or make inference about β.

Procedure 1:

Get $\hat{\lambda}$ and $\hat{\tau}$ using LS of the first equation.

Write $y_i = Z_i'\delta + X_i'\theta + V_i$ where $\delta = \lambda\beta$, $\theta = \alpha + \tau\beta$, $V_i = \varepsilon_i + \eta_i\beta$

Now we have $E(V_i|Z_i,X_i)=0$, so we can estimate $\hat{\delta}$ and $\hat{\theta}$ consistently by LS of the reduced form equation.

$$\hat{\beta} = \left(\hat{\lambda}'\left(\frac{1}{n}\sum Z_iZ_i'\right)\hat{\lambda}\right)^{-1}\hat{\lambda}'\left(\frac{1}{n}\sum Z_iZ_i'\right)\hat{\delta}$$ (weighted LS of $\hat{\delta}$ on $\hat{\lambda}$)

Procedure 2:

From $\hat{\lambda}$ and $\hat{\tau}$ calculate $\hat{D}_i = Z_i'\hat{\lambda} + X_i'\hat{\tau}$

$$y_i = \hat{D}_i'\beta + X_i'\alpha$$

and estimating β using LS.

The estimators from the two procedures are equivalent.
This is called 2SLS (2 stage least square estimator)

$$\sqrt{n}\begin{pmatrix}\hat{\delta}-\delta\\ \hat{\pi}-\pi\end{pmatrix} \xrightarrow{d} N\left[0, \begin{pmatrix}Q_z^{*-1} & 0\\ 0 & Q_z^{*-1}\end{pmatrix}\Lambda^*\begin{pmatrix}Q_z^{*-1} & 0\\ 0 & Q_z^{*-1}\end{pmatrix}\right]$$

where $Q_z^* = E(z_i z_i')$, $\Lambda^* = \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_i \varepsilon_i z_i', \frac{1}{\sqrt{n}}\sum_i \eta_i z_i'\right)$

$\delta = \pi\beta$

① We can test for $H_0: \pi = 0$

Reject H_0, then use 2SLS

Check the F statistic value is very large (>40). Then proceed to 2SLS.

② Consider testing $H_0: \beta = \beta_0$

Equivalently, $H_0: \delta - \pi\beta_0 = 0$

$g(\beta_0) := \delta - \pi\beta_0 \qquad \hat{g}(\beta_0) = \hat{\delta} - \hat{\pi}\beta_0$

$\sqrt{n}(\hat{g}(\beta_0) - g(\beta_0)) \xrightarrow{d} N(0, \Sigma_{g(\beta_0)})$

Under H_0, $n\hat{g}(\beta_0)\Sigma_{g(\beta_0)}^{-1}\hat{g}(\beta_0) \xrightarrow{d} \chi_k^2$

Invert this test to get confidence set for β.

※ 09/28/23 · Causal inference in Panel data
· Difference-in-differences

$t = 1, \cdots, T \quad i = 1, \cdots, N$

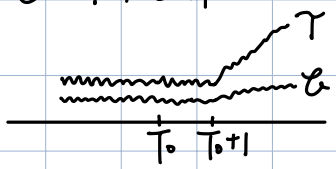
$y_{it} \quad D_{it} = \begin{cases} 0 & \text{if unit } i \text{ is not treated at time } t \\ 1 & \text{if unit } i \text{ is treated at time } t \end{cases}$

Treatment is staggered - $D_{it} = 1$ if $D_{i(t-1)} = 1$

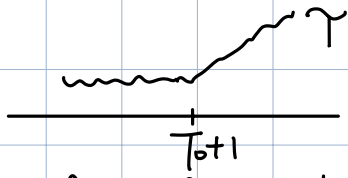
It may happen that the units that are treated are all treated at the same time

$T = \{i : D_{iT} = 1\}$ T_0 = the last time period when units in T are not treated

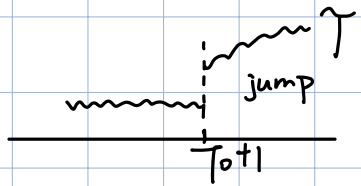
$C = \{i : D_{iT} = 0\}$



Difference-in-difference



Before-after analysis



Structure break or interrupted time series model

Dit	t = 1	2	...			T
i = 1	0	0	...	0	1	1
2	0	0	...	0	1	1
⋮						
N	0	0	...			0

units might be treated at different time

Two units who are observed for two time periods.
One of the two units is treated in the second time period.
The other unit remains untreated.

$y_{it}(1)$ for $t = 1, 2$ - potential outcomes for unit i at time 1 and 2 if the unit was treated

$y_{it}(0)$ for $t = 1, 2$ - potential outcomes if unit i remained untreated

$$y_{it} = \begin{cases} y_{it}(1) & \text{if } D_{i2} = 1 \\ y_{it}(0) & \text{if } D_{i2} = 0 \end{cases}$$

Treatment anticipation: y_{i1} depends on whether the unit gets treatment at time 2

Assumption: No anticipation of the treatment, i.e., $y_{i1}(0) = y_{i1}(1)$.

Under the assumption $y_{i1} = y_{i1}(1) = y_{i1}(0)$

$$y_{i2} = D_{i2} y_{i2}(1) + (1 - D_{i2}) y_{i2}(0)$$

	$y_{i1}(\cdot)$	$y_{i2}(\cdot)$
1	✓	✓
0	✓	X

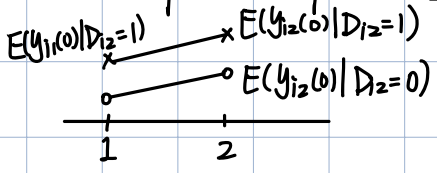
if $D_{i2} = 1$

$$ATT = E(y_{i2}(1) - y_{i2}(0) \mid D_{i2} = 1)$$

Assumption: $D_{i2} \perp (y_{i2}(0) - y_{i1}(0))$

the trend of unit i under no treatment

This assumption implies $E(y_{i2}(0) - y_{i1}(0) \mid D_{i2} = 1) = E(y_{i2}(0) - y_{i1}(0) \mid D_{i2} = 0)$



parallel

parallel trends in mean between the groups

$$\begin{aligned} ATT &= E(y_{i2}(1) - y_{i2}(0) \mid D_{i2} = 1) \\ &= E(y_{i2}(1) \mid D_{i2} = 1) - E(y_{i2}(0) \mid D_{i2} = 1) \\ &= E(y_{i2} \mid D_{i2} = 1) - [E(y_{i1}(0) \mid D_{i2} = 1) + E(y_{i2}(0) - y_{i1}(0) \mid D_{i2} = 0)] \\ &= E(y_{i2} - y_{i1} \mid D_{i2} = 1) - E(y_{i2} - y_{i1} \mid D_{i2} = 0) \end{aligned}$$

→ Difference-in-difference identification

For iid data (y_{i1}, y_{i2}, D_{i2}), the estimator is

$$\frac{1}{\sum(D_{i2}=1)} \sum_i 1(D_{i2}=1)(y_{i2} - y_{i1}) - \frac{1}{\sum(D_{i2}=0)} \sum_i 1(D_{i2}=0)(y_{i2} - y_{i1})$$

Challenges with DiD:

① Only restricted to two time periods

② It assumes no anticipation

③ Parallel trend only holds given covariates

④ The scale dependence of the parallel trend assumption

$Z_i \perp (y_{i(1)}, y_{i(0)}) \mid X_i$ → This is invariant to the scale

$D_{i2} \perp y_{i2(0)} - y_{i1(0)} \mid X_i$ → But this is NOT (For example, $\log y_{i2(0)} - \log y_{i1(0)}$)

③ says it might only hold that $E(y_{i2(0)} - y_{i1(0)} \mid D_{i2}=1, X_i) = E(y_{i2(0)} - y_{i1(0)} \mid D_{i2}=0, X_i)$ a.s.

→ Parallel trend given covariates

Theorem: Under no-anticipation assumption and parallel trends given covariates,

$$ATT = E\left[\left(\frac{D_{i2}}{E(D_{i2})} - \frac{\frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)}}{E\left(\frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)}\right)}\right)(y_{i2}-y_{i1})\right]$$

assuming $\lambda(X_i) = Pr(D_{i2}=1 \mid X_i) \in (0,1)$ almost surely.

Pf: $E\left(\frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)}\right) = E\left[E\left(\frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)} \middle| X_i\right)\right]$

$$= E\left[\frac{\lambda(X_i)}{1-\lambda(X_i)} E(1-D_{i2} \mid X_i)\right]$$

$$= E(\lambda(X_i)) = E(D_{i2})$$

Hence, RHS becomes

$$\frac{1}{E(D_{i2})} E\left[\left(D_{i2} - \frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)}\right)(y_{i2}-y_{i1})\right]$$

- $E[D_{i2}(y_{i2}-y_{i1})] = E[y_{i2}-y_{i1} \mid D_{i2}=1] \cdot E(D_{i2})$
- $E\left[\frac{\lambda(X_i)(1-D_{i2})}{1-\lambda(X_i)}(y_{i2}-y_{i1})\right] = E\left[\frac{\lambda(X_i)}{1-\lambda(X_i)} E\left((1-D_{i2})(y_{i2}-y_{i1}) \mid X_i\right)\right]$

$$= E\left[\frac{\lambda(X_i)}{1-\lambda(X_i)} \cdot E(y_{i2}-y_{i1} \mid X_i, D_{i2}=0) Pr(D_{i2}=0 \mid X_i)\right]$$

$$ATT = E(y_{i2(1)} - y_{i2(0)} \mid D_{i2}=1)$$

$$= E_X\left[E(y_{i2(1)} - y_{i2(0)} \mid X_i, D_{i2}=1) \middle| D_{i2}=1\right]$$

$$= E_X\left[E(y_{i2(1)} - y_{i1(0)} + y_{i1(0)} - y_{i2(0)} \mid X_i, D_{i2}=1) \middle| D_{i2}=1\right]$$

$$= E_X\left[E(y_{i2(1)} - y_{i1(0)} \mid X_i, D_{i2}=1) - E(y_{i2(0)} - y_{i1(0)} \mid X_i, D_{i2}=1) \middle| D_{i2}=1\right]$$

↓ Parallel trend given covariates

$$= E_X\left[E(y_{i2(1)} - y_{i1(0)} \mid X_i, D_{i2}=1) - E(y_{i2(0)} - y_{i1(0)} \mid X_i, D_{i2}=0) \middle| D_{i2}=1\right]$$

$$= E_X\left[E(y_{i2}-y_{i1} \mid X_i, D_{i2}=1) - E(y_{i2}-y_{i1} \mid X_i, D_{i2}=0) \,\middle|\, D_{i2}=1 \right]$$

exercise: complete the last two steps of the calculation.

In light of "~" in the proof, the estimator $\widehat{ATT}$ is

$$\text{1st item} = \frac{1}{\frac{\sum(D_{i2}=1)}{n}} \cdot \frac{1}{n}\sum_i D_{i2}(y_{i2}-y_{i1}) = \frac{1}{\sum(D_{i2}=1)} \sum_{i:D_{i2}=1} (y_{i1}-y_{i2})$$

(minus)

$$\text{2nd item} = \frac{1}{\frac{\sum(D_{i2}=1)}{n}} \cdot \frac{1}{n}\sum_i \frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}(y_{i2}-y_{i1}) = \frac{1}{\sum(D_{i2}=1)} \sum_i \frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}(y_{i2}-y_{i1})$$

※ 10/03/23

Issues of Difference-in-difference

① multiple time periods

② possible anticipation

③ parallel trends conditioned on the X_i

④ Scale restriction on parallel trends

Assumptions: (1) $y_{i1}(1) = y_{i1}(0)$

(2) $y_{i2}(0) - y_{i1}(0) \perp D_{i2} \mid X_i$

$$E(y_{i2}(0) - y_{i1}(0) \mid D_{i2}=1, X_i) = E(y_{i2}(0) - y_{i1}(0) \mid D_{i2}=0, X_i) \text{ a.s.}$$ ③

(3) $Pr(D_{i2}=1 \mid X_i) \in (0,1)$

$\tau = E(y_{i2}(1) - y_{i2}(0) \mid D_{i2}=1)$ is identified as

$$E\left[\frac{D_{i2}}{E(D_{i2})}(y_{i2}-y_{i1}) - \frac{\frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}}{E\left(\frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}\right)}(y_{i2}-y_{i1}) \right]$$

We can also identify τ by

$$E\left[\frac{D_{i2}}{E(D_{i2})}\left(y_{i2}-y_{i1} - E(y_{i2}-y_{i1} \mid X_i, D_{i2}=0) \right) \right]$$ exercise

or $$E\left[\frac{D_{i2}}{E(D_{i2})}(y_{i2}-y_{i1} - E(y_{i2}-y_{i1} \mid X_i, D_{i2}=0)) - \frac{\frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}}{E\left(\frac{\hat{\pi}(X_i)(1-D_{i2})}{1-\hat{\pi}(X_i)}\right)}(y_{i2}-y_{i1} - E(y_{i2}-y_{i1} \mid X_i, D_{i2}=0)) \right]$$

For Issue ①:

$y_{i1}(1), y_{i2}(1), \cdots, y_{iT}(1)$

$y_{i1}(0), y_{i2}(0), \cdots, y_{iT}(0)$

$D_{i1}, D_{i2}, \cdots, D_{iT}$

Assume T_0 is the last time the unit remains untreated

$T_0 < T$ or $T_0 = T$

$$ATT_g = E(y_{ig}(1) - y_{ig}(0) \mid D_{iT}=1), \quad g=1,\cdots,T$$

Assumption: $E[y_{ig}(0) - y_{i1}(0) \mid D_{iT}=1] = E[y_{ig}(0) - y_{i1}(0) \mid D_{iT}=0]$

Assumption: $y_{i1}(1) = y_{i1}(0)$, $y_{iT_0}(1) = y_{iT_0}(0)$

Proposition: $ATT_g = E(y_{ig} - y_{i1} \mid D_{iT}=1) - E(y_{ig} - y_{i1} \mid D_{iT}=0)$

$E(y_{it}(0) - y_{i(t-1)}(0) \mid D_{iT}=1) = E(y_{it}(0) - y_{i(t-1)}(0) \mid D_{iT}=0)$ for all $t \geq 2$

$\Rightarrow E(y_{ig}(0) - y_{i(g-1)}(0) \mid D_{iT}=1) = E(y_{ig}(0) - y_{i(g-1)}(0) \mid D_{iT}=0)$

In fact, $ATT_g = E(y_{ig} - y_{ig'} \mid D_{iT}=1) - E(y_{ig} - y_{ig'} \mid D_{iT}=0)$ for $g' = g-1, \cdots, T_0$

For Issue ②:

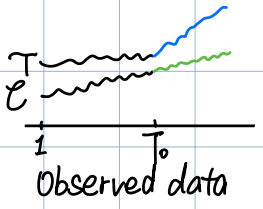
Assumption: $y_{i1}(0) = y_{i1}(1), \cdots, y_{ik}(0) = y_{ik}(1)$ for some $k < T_0$ (Limited anticipation)

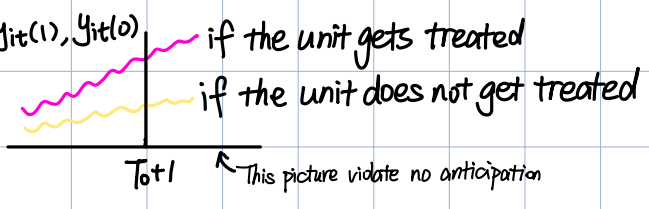
Under limited anticipation,

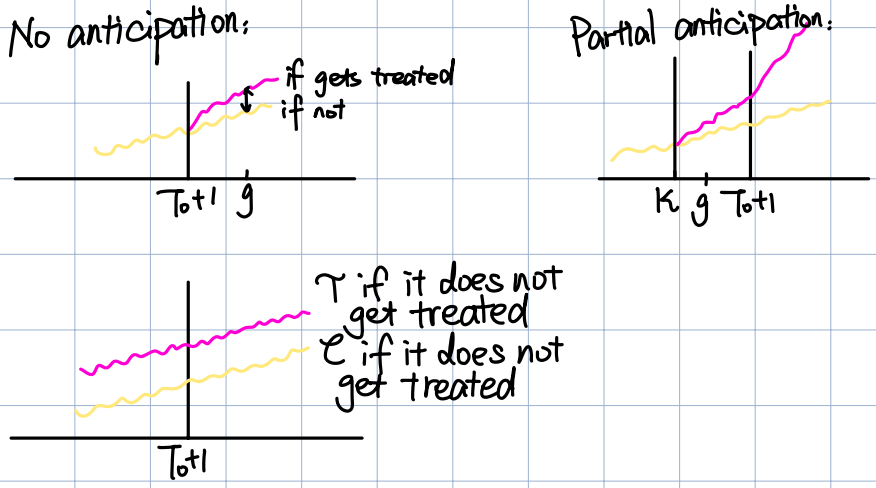
$$ATT_g = E(y_{ig} - y_{ig'} \mid D_{iT}=1) - E(y_{ig} - y_{ig'} \mid D_{iT}=0) \text{ for } g' = g-1, \cdots, k$$

※ 10/05/23

T ← treated unit C ← control unit







We have two time periods = 1, 2, and y_{i1}, y_{i2}, D_{i2}

Potential outcomes: $y_{i1}(0), y_{i2}(0), y_{i2}(1)$ $[y_{i1}(1) = y_{i1}(0)]$

Standard parellel trend:

$$E(y_{i2}(0) - y_{i1}(0) \mid D_{i2} = 1) = E(y_{i2}(0) - y_{i1}(0) \mid D_{i2} = 0)$$

We say that a parallel trend that is <u>invariant under scale transformation</u> would satisfy $E(g(y_{i2}(0)) - g(y_{i1}(0)) \mid D_{i2} = 1) = E(g(y_{i2}(0)) - g(y_{i1}(0)) \mid D_{i2} = 0)$]→(1)

for all strictly monotone function g.

<u>Proposition</u>: (1) is true if and only if

(2) $F_{y_{i2}(0)|D_{i2}=1}(y) - F_{y_{i1}(0)|D_{i2}=1}(y) = F_{y_{i2}(0)|D_{i2}=0}(y) - F_{y_{i1}(0)|D_{i2}=0}(y)$ for all y.

<u>Pf.</u> (2)⇒(1): Take the derivative by d operation then multiply by $g(y)$ on both side. Then integrate.

(1)⇒(2): Fix $\bar{y}$. $g_1(y) = y$, $g_2(y) = y - 1(y \le \bar{y})$

Apply g_1 and g_2 to (1), and then substract to get (2).

<u>Proposition</u>: (2) is true if and only if

(3) $F_{y_{it}(0)|D_{i2}=d}(y) = \theta C_t(y) + (1-\theta) H_d(y)$

where $t = 1, 2$, $d = 0, 1$, θ is a number between 0 and 1, C_t and H_d are CDF.

<u>Pf.</u> (3)⇒(2) by simple subtraction of the distribution function.

Lemma: Given two distribution functions F_1 and F_2 (with densities f_1 and f_2).

$$F_1 = \theta G + (1-\theta)\tilde{F}_1 \quad \text{and} \quad F_2 = \theta G + (1-\theta)\tilde{F}_2$$

where θ and $\tilde{F}_1, \tilde{F}_2$ only depends on $f_1 - f_2$. $G, \tilde{F}_1, \tilde{F}_2$ are CDFs, $\theta \in [0,1]$.

(2) $\Rightarrow$ (3):

$$\begin{cases} F_{y_{i2}(0)|D_{i2}=1} = \theta_1 G_1 + (1-\theta_1)\tilde{F}_{y_{i2}(0)|D_{i2}=1} \\ F_{y_{i1}(0)|D_{i2}=1} = \theta_1 G_1 + (1-\theta_1)\tilde{F}_{y_{i1}(0)|D_{i2}=1} \end{cases}$$

$$\begin{cases} F_{y_{i2}(0)|D_{i2}=0} = \theta_0 G_0 + (1-\theta_0)\tilde{F}_{y_{i2}(0)|D_{i2}=0} \\ F_{y_{i1}(0)|D_{i2}=0} = \theta_0 G_0 + (1-\theta_0)\tilde{F}_{y_{i1}(0)|D_{i2}=0} \end{cases}$$

By (2) and the lemma, $\theta_0 = \theta_1$, $\tilde{F}_{y_{i2}(0)|D_{i2}=1} = \tilde{F}_{y_{i2}(0)|D_{i2}=0}$ & $\tilde{F}_{y_{i1}(0)|D_{i2}=1} = \tilde{F}_{y_{i1}(0)|D_{i2}=0}$

So, $F_{y_{it}|D_{i2}=d} = \theta G_d + (1-\theta)\tilde{F}_t$ where $\tilde{F}_t = \tilde{F}_{y_{it}(0)|D_{i2}=d}$

- $i \rightarrow y_{i1}, D_{i2}$ Panel Data
- $i \rightarrow (y_i, D_i, T_i)$ Repeated Cross-sectional Data

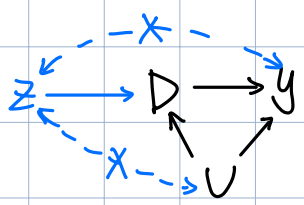
$y_i = h(D_i, T_i, X_i)$

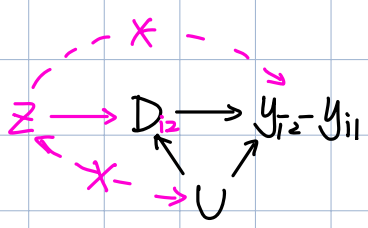
$y_i = \alpha + \beta D_i + \tau T_i + \varepsilon_i$

$y_{i2}(0) - y_{i1}(0) \perp D_{i2} \leftarrow (y_{i2}(0), y_{i1}(0), y_{i2}(1), y_{i1}(1)) \perp D_{i2}$

$\downarrow$ $\qquad$ $\downarrow$

Instrumented parallel trends $\leftarrow$ z_i





$$LATE = E(y_{i2}(1) - y_{i1}(0) \mid \text{Complier})$$

$$= \frac{E(y_{i2} - y_{i1} \mid z_i = 1) - E(y_{i2} - y_{i1} \mid z_i = 0)}{E(D_{i2} \mid z_i = 1) - E(D_{i2} \mid z_i = 0)}$$

David Campbell:

$$y_i = \alpha + \tau C_i + \beta I_i + \eta T_i + \text{error}_i$$

$C_i = \begin{cases} 1, & \text{if it is in the control group} \\ 0, & \text{if it is in the treatment group} \end{cases}$

$I_i = \begin{cases} 1, & \text{if } C_i = 1 \text{ and time} = 2 \\ 0, & \text{if } C_i = 0 \text{ and time is 1 or 2.} \end{cases}$

$I_i = C_i \times T_i$

$T_i = \begin{cases} 0, & \text{at time 1} \\ 1, & \text{at time 2} \end{cases}$

$i \rightarrow (y_i, C_i, T_i)$

Note that this model implies

$$E(y_i \mid T_i=1, C_i=1) - E(y_i \mid T_i=0, C_i=1) = (\alpha+\tau+\beta+\eta)-(\alpha+\tau) = \beta+\eta$$

$$E(y_i \mid T_i=1, C_i=0) - E(y_i \mid T_i=0, C_i=0) = (\alpha+\eta)-\alpha = \eta$$

$(y_{i(0)}, y_{i(1)}, C_i, T_i)$ ← Full data

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

Interested in estimating $ATT = E(y_{i(1)} - y_{i(0)} \mid C_i=1, T_i=1)$

U_i : unmeasured confounders

$y_{i(0)} = h(U_i, T_i)$, U_i is also related to C_i

For particular h and U, this model reduces to our previous linear model

[e.g. $U_i = \alpha + \tau C_i$, h a linear function]