Basic

SUTVA

Potential outcomes. $Y_i(1)$, $Y_i(0)$

Causal effect: comparison between the potential outcomes under treatment and under control for the same unit or a common set of units

Causal estimands: functions of PO

$$\text{ITE}: \tau_i = Y_i(1) - Y_i(0)$$

(condition) $$\text{CATE}: \tau(x) = E(Y_i(1) - Y_i(0) \mid X = x)$$

$$\text{ATE}: \tau = E(Y_i(1) - Y_i(0)) = E_x(\tau(x))$$

(for treated units) $$\text{ATT}: \tau = E(Y_i(1) - Y_i(0) \mid Z_i = 1)$$

(for controlled units) $$\text{ATC}: \tau = E(Y_i(1) - Y_i(0) \mid Z_i = 0)$$

Ratios: $\gamma_i = \frac{Y_i(1)}{Y_i(0)}$

$\gamma = \frac{E(Y_i(1))}{E(Y_i(0))}$

$\tau^{ATE} = \Pr(Z_i=1)\tau^{ATT} + \Pr(Z_i=0)\tau^{ATC}$

In CRE: $\tau^{ATE} = \tau^{ATT} = \tau^{ATC}$ Observe: τ

Assumptions → assignment mechanism

→ $Pr(Z_i=1 \mid X_i, Y_i(1), Y_i(0))$

Unconfounded Assigment: propensity score

$$Pr(Z_i=1 \mid X_i, Y_i(0), Y_i(1)) = Pr(Z_i=1 \mid X_i)$$

$$Z_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid X_i$$

PO

? $Pr(Y_i(z) \mid X_i) = Pr(Y_i(z) \mid X_i, Z_i)$

$= Pr(Y_i(z) \mid X_i, Z_i=z)$

$= Pr(Y_i(Z_i) \mid X_i, Z_i=z)$

$= Pr(Y_i^{obs} \mid X_i, Z_i=z)$

$Y_i^{obs} = Y_i = Z_i Y_i(1) + (1-Z_i) Y_i(0) = Y_i(Z_i)$

e.g. $Y(\cdot) = (\cdot)^2 + X$

$\hat{Y}_i = f(Z_i, Y_i(1), Y_i(0))$

$\hat{Y}_i = \begin{cases} Y_i(1), & \text{if } Z_i=1 \\ Y_i(0), & \text{if } Z_i=0 \end{cases}$

Randomized experiments (RE)

$$Z \perp\!\!\!\perp X, \quad Z \perp\!\!\!\perp V, \quad Z \perp\!\!\!\perp Y(1), Y(0)$$

Under Randomization, confoundness ✓

types of RE

- Completely RE
  - Bernoulli: $P(Z_i = 1 \mid X, Y(0), Y(1)) = \frac{1}{2}$
  - $P(Z_i = 1 \mid X, Y(0), Y(1)) = \frac{N_1}{N}$
- Stratified
- Paired
- Clustered.

CRE

Fisher's Randomized Test

assumptions: $Y(0), Y(1)$ fixed

sharp null hypotheses

$$H_0: Y_i(1) = Y_i(0) \text{ for all units } i=1,\dots,n$$

test statistic $T = T(Z, Y)$

- DIM $\hat{\tau} = \bar{Y}(1) - \bar{Y}(0)$
- T statistic
- Wilcoxon Rank
- ...

Z: 分布已知 → T: 分布已知

exact p-value $Pr(T \geq T^{obs})$

CRE

Neyman's Repeated Sampling Approach

estimand : ATE

1. find unbiased estimator of ATE
2. obtain variance and interval estimates of the unbiased estimator (derive the estimator distribution under repeated sampling based on the (randomized) distribution of Z)

estimator :

DIM $\hat{\tau}^{unadj} = \sum_{i=1}^{N}\left(\frac{Z_i Y_i}{N_1}\right) - \sum_{i=1}^{N}\left(\frac{(1-Z_i) Y_i}{N_0}\right) = \bar{Y}_1 - \bar{Y}_0$

unbiasedness $E_{Z:\text{Sampling}}\left(\hat{\tau}^{unadj} \mid Y(0), Y(1)\right) = \tau$

$$V\left(\hat{\tau}^{unadj}\right) = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N}$$

$$V^{Neyman} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1}$$

$$\sqrt{V}^{neyman} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1}$$

CRE unbiasedness 被RE保证. covariance balance 是*为更小的 variance

Covariate Balance and Adjustment

a secondary analysis

Option 1. by design

- stratified RE
- paired RE
- rerandomization

Option 2. by analysis

- regression (ANCOVA)
- propensity score weighting
- model-based imputation

Observational Studies

In observational studies, some structural (often untestable) assumptions must be made.

Assumption 1. Unconfoundness (untestable)

$$Z_i \perp\!\!\!\perp Y_i(1), Y_i(0) \mid X_i$$

$$Pr(Z_i \mid Y_i(0), Y_i(1), X_i) = Pr(Z_i \mid X_i)$$

Assumption 2. Overlap (directly checked from the data)

$$0 < Pr(Z_i = 1 \mid X_i) < 1 \quad \text{for all } i$$

(Assumptions for estimating ATT: 1. $Z_i \perp\!\!\!\perp Y_i(0) \mid X_i$ 2. $Pr(Z_i = 1 \mid X_i) < 1$)

Under Assumptions, we have

$$Pr(Y(z) \mid X) = Pr(Y \mid X, Z = z)$$

the distribution of p.o. $Y(z)$ = observed distribution of Y in treatment arm $Z = z$

05

so we have

$$\tau^{ATE} = E_x\{\tau(x)\}$$

令 $\mu_z(x) = E(Y(z)|x) = E(Y|Z=z, X)$

$$= E\{\mu_1(x) - \mu_0(x)\}$$

$$= E\{E(Y|Z=1, x) - E(Y|Z=0, x)\}$$

连续 $= \int \{E(Y|Z=1, X=x) - E(Y|Z=0, X=x)\} f(x) dx$

离散 $= \sum_x \{E(Y|Z=1, X=x) - E(Y|Z=0, X=x)\} Pr(x)$

$Pr(Y(z)|X) = Pr(Y|Z=z, X) = f_z(x)$

$$\begin{cases} Pr(Y(z)) = \sum_x Pr(Y(z)|X=x) \cdot Pr(X=x) \\ Pr(Y|Z=z) = \sum_x Pr(Y|Z=z, X=x) Pr(X=x|Z=z) \end{cases}$$

Two estimation strategies for τ^{ATE}

① Outcome modeling | regression

要估计 $\hat{\mu}_z(x)$

$$\hat{\tau}^{ATE} = \hat{\tau}^{reg} = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right\}$$

② Inverse probability weighting IPW

要估计 $\hat{e}(x)$

$$\hat{\tau}^{ATE} = \hat{\tau}^{ipw} = \frac{\sum_{i=1}^{N} Z_i Y_i / \hat{e}(X_i)}{\sum_{i=1}^{N} Z_i / \hat{e}(X_i)} - \frac{\sum_{i=1}^{N} (1-Z_i) Y_i / \{1-\hat{e}(X_i)\}}{\sum_{i=1}^{N} (1-Z_i) / \{1-\hat{e}(X_i)\}}$$

(离散的估计见 by Book. 10.4.1)

③ Matching

④ Stratification (离散的 X)

Outcome Modeling | Regression

estimands 都是 $\mu_z(x)$ 的函数

$$\tau^{ATE} = \frac{1}{N}\sum_{i=1}^{N}\{\mu_1(x_i) - \mu_0(x_i)\}$$

$$\tau^{ATT} = \frac{1}{N_1}\sum_{i=1}^{N} z_i\{y_i - \mu_0(x_i)\}$$

$$\tau^{ATC} = \frac{1}{N_0}\sum_{i=1}^{N}(1-z_i)\{\mu_1(x_i) - y_i\}$$

$$\tau^{CATE}(x) = \mu_1(x) - \mu_0(x) \qquad CATE = E(Y(1)-Y(0)\mid X=x)$$

估计 $\mu_z(x)$

To estimate $\mu_z(x)$, we have many choices of model.

① Linear regression, we assume

(注：这种方法不一定可信，因为 Z 和 X 有可能共线性)

$$E(Y\mid Z=1,X) = \mu_1(x) = \beta_0 + \beta_z + \beta_x x$$

$$E(Y\mid Z=0,X) = \mu_0(x) = \beta_0 + \beta_x x$$

$$E(Y\mid Z,X) = \beta_0 + \beta_z \cdot Z + \beta_x^T X$$

$Y \sim 1 + Z + X$

$$\therefore \tau(X) = \mu_1(x) - \mu_0(x) = \beta_0 + \beta_z + \beta_x^T x - (\beta_0 + \beta_x^T x) = \beta_z$$

$$\therefore \tau^{ATE} = E\{\tau(x)\} = \beta_z$$

Z 的系数 β_z

② We can also assume

$$E(Y|Z=1,X) = \mu_1(X) = \beta_0 + \beta_Z + \beta_X^T X + \beta_{ZX}^T X$$
$$= (\beta_0 + \beta_Z) + (\beta_X^T + \beta_{ZX}^T)X$$

$$E(Y|Z=0,X) = \mu_0(X) = \beta_0 + \beta_X^T X$$

$$E(Y|Z,X) = \beta_0 + \beta_Z Z + \beta_X^T X + \beta_{ZX}^T XZ$$

$$\therefore \tau(x) = \mu_1(x) - \mu_0(x)$$
$$= \beta_Z + \beta_{ZX}^T X$$

$$\therefore \tau^{ATE} = E\{\tau(X)\} = \beta_Z + \beta_{ZX}^T E(X)$$

$$\therefore \hat{\tau}^{ATE} = \hat{\beta}_Z + \hat{\beta}_{ZX}^T \cdot \bar{X}$$

$Y \sim 1 + Z + X + ZX$

(或者先中心为 X 为 $\tilde{X}$, $Y \sim 1 + Z + \tilde{X} + Z\tilde{X}$)

$\hat{\tau}^{ATE}$

Why balance is important?

"If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the region with little overlap, which is sensitive to the model specification assumption."

Balance $\frac{Pr(X|Z=1)}{Pr(X|Z=0)}$

$$= \frac{\frac{Pr(Z=1|X)\,Pr(X)}{Pr(Z=1)}}{\frac{Pr(Z=0|X)\,Pr(X)}{Pr(Z=0)}}$$

$$= \frac{Pr(Z=1|X)}{Pr(Z=0|X)} \cdot \frac{Pr(Z=0)}{Pr(Z=1)}$$

Overlap

metrics of balance:

ASD

visualize balance:

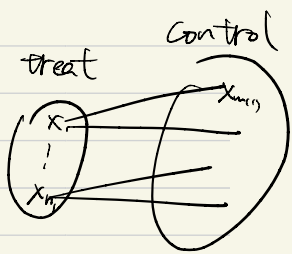
Love plot

often be applied in settings where
the interest is ATT
a large reservior of potential controls

Matching

Nearest-Neighbor Matching

$$\vec{Y}_i(0) = \begin{cases} \sum_{j \in M_i} Y_j / M, & Z_i = 1 \\ Y_i, & Z_i = 0 \end{cases}$$



$$\vec{Y}_i(1) = \begin{cases} Y_i, & Z_i = 1 \\ \sum_{j \in M_i} Y_j / M, & Z_i = 0 \end{cases}$$

$$\vec{\tau}^{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left(\vec{Y}_i(1) - \vec{Y}_i(0) \right)$$

$$\vec{\tau}^{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} Z_i \left(\vec{Y}_i(1) - \vec{Y}_i(0) \right)$$

can be used as a pre-processing step

Stratification

Suppose we have a K-level covariate X

$$E(Y(1)) = \sum_k E(Y \mid X = k, Z = 1) \Pr(X = k)$$

$$\widehat{E(Y(1))} = \sum_k \bar{Y}_{k,1} \frac{n_k}{n}$$

$$\widehat{E(Y(0))} = \sum_k \bar{Y}_{k,0} \frac{n_k}{n}$$

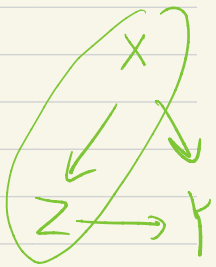
$$\hat{\tau} = \sum_k (\bar{Y}_{k,1} - \bar{Y}_{k,0}) \frac{n_k}{n}$$

Propensity Score

a probability mechanism

a summary statistic of assignment

$$e(X) = Pr(Z=1|X) = E(Z|X)$$



a summary score of the covariates

Property 1. $Z \perp X \mid e(X)$

Property 2. $Z_i \perp\!\!\!\perp \{Y_{i(0)}, Y_{i(1)}\} \mid X_i \Rightarrow Z_i \perp\!\!\!\perp \{Y_{i(0)}, Y_{i(1)}\} \mid e(X_i)$

Propensity score analysis

Stage 1. estimate the propensity score $\hat{e}(X)$

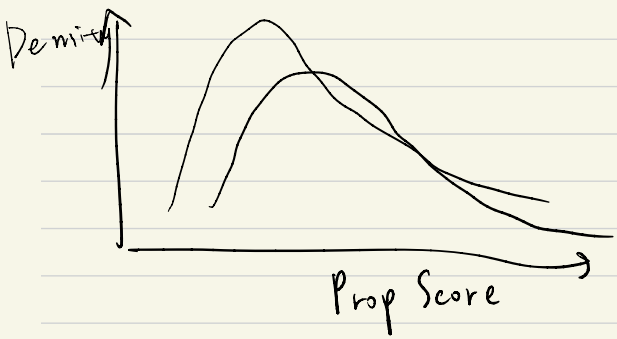
Stage 2. given the $\hat{e}(X)$, estimate the causal effects through ~~stratification~~, weighting, matching, regression and etc.

ensure overlap and balance

estimate propensity score

initial fit → discard outliers → check balance → re-fit

$$\text{logit} \quad Pr(Z_i = 1 \mid X_i) = \beta X_i$$



Propensity Score Weighting

IPW Inverse Probability Weighting

$$E\left(\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right) = \tau^{ATE}$$

IPW creates a weighted population

the covariates distributions of 2 groups are balanced

proof.

$$E\left(\frac{ZY}{e(X)}\right) = E\left(E\left(\frac{ZY}{e(X)}\middle|X\right)\right)$$

$$= E\left(\frac{1}{e(X)}E(ZY|X)\right) \qquad Y = ZY(1) + (1-Z)Y(0), \quad e(X) = E(Z|X)$$

$$= E\left(\frac{1}{e(X)}\cdot E\left(Z^2Y(1) + Z(1-Z)Y(0)\middle|X\right)\right)$$

$$= E\left(\frac{1}{e(X)}E(Z^2|X)E(Y(1)|X)\right) = E\left(E(Y(1)|X)\right) = Y(1)$$

$$E\left(\frac{XZ}{e(X)}\right) = E\left(\frac{X(1-Z)}{1-e(X)}\right)$$

Define inverse probability weights

$$\begin{cases} w_1(x_i) = \frac{1}{e(x_i)}, & \text{for } z_i = 1 \\ w_0(x_i) = \frac{1}{1-e(x_i)}, & \text{for } z_i = 0 \end{cases}$$

moment estimator

$$\hat{\tau}_{ipw,1} = \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{Y_i Z_i}{e(x_i)} - \sum_{i=1}^{N}\frac{Y_i(1-Z_i)}{1-e(x_i)}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{Y_i Z_i w_1(x_i) - Y_i(1-Z_i)w_0(x_i)\right\}$$

Normalize weights (sum of weights within group should be 1)

$$\hat{\tau}_{ipw,2} = \frac{\sum_{i=1}^{N} Y_i Z_i w_1(x_i)}{\sum_{i=1}^{N} Z_i w_1(x_i)} - \frac{\sum_{i=1}^{N} Y_i(1-Z_i)w_0(x_i)}{\sum_{i=1}^{N}(1-Z_i)w_0(x_i)}$$

$$V(\hat{\tau}_{ipw,2}) < V(\hat{\tau}_{ipw,1})$$

Propensity Score Weighting

Extension Beyond IPW: Move the Goal post

use specified target populations.

"often units who are "on the fence" are of interest"

Balancing Weights

$f(x)$: density of the observed covariates

$g(x)$: density of a target population's covariates

$$h(x) = \frac{g(x)}{f(x)}$$ a tilting function

A new class of estimands: ATE over the target population

$$\tau_h \equiv E_g[Y_i(1) - Y_i(0)] = \frac{\int \tau(x)\, g(x)\, \mu(dx)}{\int g(x)\, \mu(dx)}$$

$$= \frac{\int \tau(x) f(x) h(x) \mu(dx)}{\int f(x) h(x) \mu(dx)} = \frac{E\{h(x)\tau(x)\}}{E\{h(x)\}}$$

choice of $h(x)$ determines the target population, estimand, weights

Let $f_z(x) = Pr(X=x \mid Z=z)$ covariate distribution within each group z

$$f_1(x) = Pr(X \mid Z=1) = \frac{Pr(Z=1 \mid X)\, Pr(X)}{Pr(Z=1)}$$

$$\propto e(x) f(x)$$

$$f_0(x) = Pr(X \mid Z=0) = \frac{Pr(Z=0 \mid X)\, Pr(X)}{Pr(Z=0)}$$

$$\propto (1-e(x)) f(x)$$

weight $f_z(x)$ to the target population g

$$w_1(x) f_1(x) \propto g(x) = f(x) h(x)$$

$$w_0(x) f_0(x) \propto g(x) = f(x) h(x)$$

$$\Rightarrow w_1(x) \propto \frac{f(x)h(x)}{f_1(x)} \propto \frac{f(x)h(x)}{e(x)f(x)} = \frac{h(x)}{e(x)}$$

balancing weights

$$w_0(x) \propto \frac{f(x)h(x)}{f_0(x)} \propto \frac{f(x)h(x)}{(1-e(x))f(x)} = \frac{h(x)}{1-e(x)}$$

Sample estimator of WATE

$$\hat{\tau}_h = \frac{\sum_i w_1(x_i) z_i Y_i}{\sum_i w_1(x_i) z_i} - \frac{\sum_i w_0(x_i)(1-z_i) Y_i}{\sum_i w_0(x_i)(1-z_i)}$$

"each unit is weighted by its probability of being assigned to the opposite group."

Overlap Weights

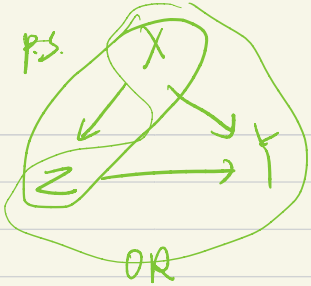
$h(x) = e(x)(1-e(x))$

$$\begin{cases} w_1(x) \propto 1 - e(x), & \text{for } z=1 \\ w_0(x) \propto e(x), & \text{for } z=0 \end{cases}$$

$$h(x) = e(x)(1-e(x))$$

$$= \left(\frac{1}{e} + \frac{1}{1-e}\right)^{-1} \qquad \left(\frac{1}{e_1} + \frac{1}{e_2} + \frac{1}{e_3}\right)^{-1}$$

Doubly Robust Estimation



estimand: ATE

OR estimator:

$$E(Y|z)|X) = m_z(X)$$

$$\hat{\tau}_{OR} = \frac{1}{N}\sum_{i=1}^{N}\{z_i(Y_i - \hat{m}_0(X_i)) + (1-z_i)(\hat{m}_1(X_i) - Y_i)\}$$

? can be sensitive when overlap is weak

PS Weighting estimator:

$$\hat{\tau}_{IPW} = \frac{\sum Y_i z_i / e(X_i)}{\sum z_i / e(X_i)} - \frac{\sum Y_i (1-z_i)/(1-e(X_i))}{\sum (1-z_i)/(1-e(X_i))}$$

DR estimator:

$$\tau = E(Y(1)) - E(Y(0)) = E\left\{\frac{zY}{e(X)} - \frac{z-e(X)}{e(X)} m_1(X)\right\} -$$

$$E\left\{\frac{(1-z)Y}{1-e(X)} + \frac{z-e(X)}{1-e(X)} m_0(X)\right\}$$

$$= E\left\{m_1(X_i) + \frac{z_i(Y_i - m_1(X_i))}{e(X_i)}\right\}$$

用 $\hat{m}_z$ 代替式

$$-E\left\{m_0(X_i) + \frac{(1-z_i)(Y_i - m_0(X_i))}{1-e(X_i)}\right\}$$

$$\hat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{z_i Y_i}{\hat{e}(X_i)} - \frac{z_i - \hat{e}(X_i)}{\hat{e}(X_i)} \hat{m}_1(X_i)\right\}$$

$$-\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(1-z_i)Y_i}{1-\hat{e}(X_i)} + \frac{z_i - \hat{e}(X_i)}{1-\hat{e}(X_i)} \hat{m}_0(X_i)\right\}$$

$$= \cdots$$

let's see $\hat{\mu}_{1,dr} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{Y_i Z_i}{\hat{e}_i} - \frac{(Z_i - \hat{e}_i)\hat{m}_i}{\hat{e}_i}\right\}$

$m_1(x) = E(Y(1)|X)$

$e(x) = E(Z=1|X)$

$$\hat{\mu}_{1,dr} - E(Y(1)) = E\left\{\frac{Z(Y - m_1(X))}{e(X)} + m_1(X)\right\} - E(Y(1))$$

$$= E\left\{\frac{Z\left(ZY(1) + (1-Z)Y(0) - m_1(X)\right)}{e(X)} + m_1(X) - Y(1)\right\}$$

$$= E\left\{\frac{Z(Y(1) - m_1(X))}{e(X)} - (Y(1) - m_1(X))\right\}$$

$$= E\left\{\frac{Z - e(X)}{e(X)}(Y(1) - m_1(X))\right\} \checkmark$$

$Z \perp\!\!\!\perp Y(1), Y(0) \mid X$

$$= E\left(E\left\{\frac{Z - e(X)}{e(X)}(Y(1) - m_1(X)) \,\middle|\, X\right\}\right)$$

$$= E\left(E\left\{\frac{Z - e(X)}{e(X)} \,\middle|\, X\right\} \times E\left\{Y(1) - m_1(X) \,\middle|\, X\right\}\right)$$

$e_{true}(X)$ $\qquad$ $m_{1,true}(X)$

$$= E\left(\frac{E(Z|X) - e(X)}{e(X)} \times \left(E(Y(1)|X) - m_1(X)\right)\right)$$

So $\hat{\mu}_{1,dr} - E(Y(1)) = 0$ if either $E(Z=1|X) = e(X)$ or $E(Y(1)|X) = m_1(X)$

(asymptotically unbiased)

$\hat{\tau}_{dr}$ is a consistent estimator of ATE if either model is correctly specified. (large sample property)

Overlap and Balance

$$Y(1), Y(0) \perp Z \mid X = x$$

$$f(0,x) = E(Y \mid Z=0, X=x)$$

$$= E(ZY(1) + (1-Z)Y(0) \mid Z=0, X=x)$$

$$= E(Y(0) \mid Z=0, X=x) = E(Y(0) \mid X=x)$$

Treatment Effect Heterogeneity

Machine Learning Approaches

$Y = f(z, x) + \epsilon$

$$f(z, x) = E(Y \mid Z = z, X = x)$$ ← model this

CATE $\tau(x) = f(1, x) - f(0, x)$

S-Learner (single)

use all data to fit model $f(z, x)$

e.g. BART $f(z, x) = \sum_{t=1}^{T} g(x, z; T_t, M_t)$

T-Learner (two)

fit separate models to the treated/control groups

$E(Y \mid Z = 1, X = x) = f_1(x)$

$E(Y \mid Z = 0, X = x) = f_0(x)$

CATE $\tau(x) = f_1(x) - f_0(x)$.

R-Learner

$Y_i = \mu(x_i) + \tau(x_i) z_i + \varepsilon_i$ $\quad f(z, x) = \mu(x) + \tau(x) \cdot z$

$m(x_i) = E(Y_i \mid x_i) = \mu(x_i) + \tau(x_i) e(x_i)$

$Y_i - m(x_i) = (z_i - e(x_i)) \tau(x_i) + \varepsilon_i$

$\tau(\cdot) = \arg\min_{\tau} \{ E[((Y_i - m(x_i)) - (z_i - e(x_i)) \tau(x_i))^2] \}$

Causal Forests

s-learner for CATE, estimate $f(z, x)$

→ split to maximize heterogeneity of the treatment effect

$$\frac{|\hat{\tau}_L - \hat{\tau}_R|}{\sqrt{\widehat{Var}(\hat{\tau}_L) + \widehat{Var}(\hat{\tau}_R)}}$$

hope: within leaf k, as if randomized, similar covariates

leaves: $L_1(x), \ldots, L_k(x)$

In leaf k: estimated treatment effect

$$\hat{\tau}_k = \frac{1}{|\{i : X_i \in L_k(x), T_i = 1\}|} \sum Y_i - \frac{1}{\{i : X_i \in L_k(x), j T_i = 0\}} \sum Y_i$$

→ tuning cross-validation

$$MSE \quad L(\hat{\tau}) = E\left[\left(Y_i(1) - Y_i(0) - \hat{\tau}(X_i) \right)^2 \right]$$ unknown

Honesty criterion: divide samples into three subsamples (train, estimate causal effect, test)

S-learner for CATE. estimate $f(z,x)$

BART (Bayesian Additive Regression Trees)

partition the space of x and z

$$f(z,x) = \sum_{t=1}^{T} g(x,z; \mathcal{T}_t, \mathcal{M}_t)$$

$\mathcal{T}_t$ represents the tree structure

$\mathcal{M}_t$ parameters for predictions in each leaf

$\mathcal{M}_t = (\mu_{t1}, \cdots, \mu_{tL_t})$ L_t: number of leafs

Bayesian approach

prior probability of splitting decreases with depth

shrinkage of mean parameters $\mu_{tl} \sim N(0, \frac{6\sigma^2}{T})$

Do we expect CATE to be as complex as $f(z,x)$?

→ re-parameterize

$$f(z,x) = \mu(x) + \tau(x)z$$

further we can include p.s. score

$$f(z,x) = \mu(x, \hat{e}(x)) + \tau(x)z$$

use ML for both outcome model and propensity model

Double Machine Learning (DML) for high-dimensional data

→ An early example : use ML in DR estimator

→ A canonical example : Partial Linear Model

intuition: relation between Y and X is usually more complex than the relation between Y and Z

idea: use ML model for $Y \sim X$ and linear for $Y \sim Z$

DML for ATE

$$Y = Z\tau + m(X) + U \qquad E(U|X,Z) = 0$$

$$Z = e(X) + V \qquad E(V|X) = 0$$

steps:

1. split sample into K parts $\{\mathcal{I}_k\}_{k=1}^K$. $\mathcal{I}_k, \mathcal{I}_k^c$
2. estimate Z from X using $\mathcal{I}_k^c$
3. estimate Y from X using $\mathcal{I}_k^c$
4. estimate ATE in $\mathcal{I}_k$. regress residuals.
5. aggregate over K folds. repeat 2-3 for $k = 2, \dots, K$

DML for CATE

$$Y = Z\tau(X) + m(X) + U \qquad E(U|X,Z) = 0$$

$$Z = e(X) + V \qquad E(V|X) = 0$$

$$E(UV|X,Z) = 0$$

$$\hat{\tau}(x) = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}} \, E_n[(\hat{U} - \tau(x)\hat{V})^2]$$

Sensitivity Analysis

perform sensitivity analysis to assess how sensitive the causal analysis if unconfound is violated

two key assumption
- overlap (testable)
- unconfoundedness (untestable)

Unconfoundedness and Balance

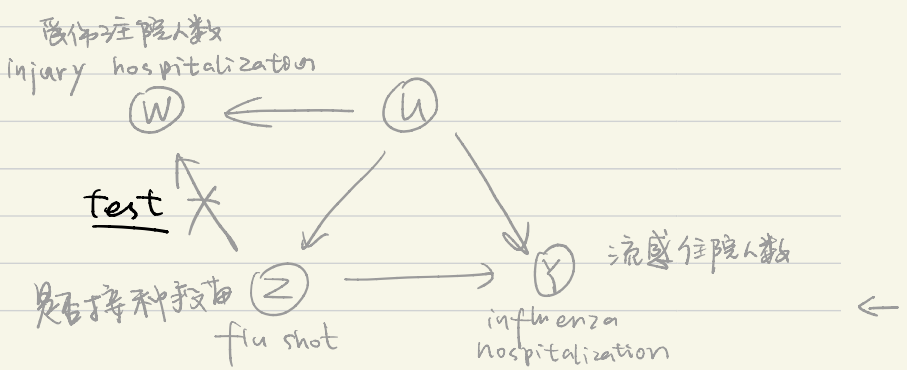
$$\{Y_{(1)}, Y_{(0)}\} \perp Z \mid X$$

implies: within strata of observed covariates, potential outcomes corresponding to both treatment conditions would be balanced between groups

e.g. $P_r(Y_{(0)} \mid Z=1)$ vs. $P_r(Y_{(0)} \mid Z=0)$

In practice, we use balance in covariates $P_r(X \mid Z=1)$ $P_r(X \mid Z=0)$ (?)

as a proxy to balance in potential outcomes.



Assess unconfoundedness indirectly

Falsification tests

Method 1. use multiple control groups

using data from alternative control group

indicator $T_i \in \{-1, 0, 1\}$

suppose $Y_i = \begin{cases} Y_i(0) & \text{if } T_i \in \{-1, 0\} \\ Y_i(1) & \text{if } T_i = 1 \end{cases}$

extend the unconfoundedness to $\{Y_i(0), Y_i(1)\} \perp T_i \mid X_i$

testable $Y_i(0) \perp \mathbb{1}\{T_i = 0\} \mid X_i, \; T_i \in \{-1, 0\}$

$Y_i^{obs} \perp \mathbb{1}\{T_i = 0\} \mid X_i, \; T_i \in \{-1, 0\}$

using data from prior periods

Method 2. using lagged outcomes (observed before the treatment)

Y_{lag} as a proxy for $Y(0)$ $\quad V = (X \mid Y_{lag})$

test H_0: $E(Y_{lag,2=1} - Y_{lag,2=0} \mid V = v) = 0 \; \forall v$

special case

vs. H_1: $\exists v$ s.t. $E(Y_{lag,2=1} - Y_{lag,2=0} \mid V = v) \neq 0$

using alternative placebo outcome supposed not affected by treatment

Method 3. using negative control outcomes (NCO)

select another outcome variable known not be causally affected by the treatment of interest

$W = 2W(1) + (1-2)W(0)$ unconfoundedness implies $\{W(1), W(0)\} \perp 2 \mid X$

$\rightarrow E[W(1) - W(0) \mid X] = 0$

e.g. flu shot (Z), influenza hospital (Y), injury hospitalization (W)

Sensitivity analysis in causal

assess the bias of causal effect estimates when the unconfoundedness is assumed to fail

Assumption $p(z \mid Y(0), Y(1), X) \neq p(z \mid X)$

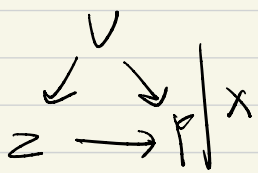
$z \not\perp \{Y(1), Y(0)\} \mid X$

$p(z \mid Y(0), Y(1), X, U) = p(z \mid X, U)$

$z \perp \{Y(1), Y(0)\} \mid (X, U)$

Method 1. Rubin $(\pi, \alpha, \delta_1, \delta_0)$

$$\begin{cases} U \sim \text{Bern}(\pi) \\ \text{logit}(p(z=1 \mid u)) = \gamma + \alpha u \\ \text{logit}(p(Y(z)=1 \mid u)) = \beta_z + \delta_z u \end{cases}$$



Method 2. Rosenbaum's bounds (only for testing no causal effect)

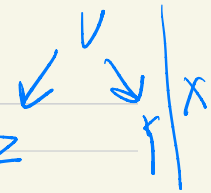
$$\frac{1}{\Gamma} \leq \frac{Pr(z=1 \mid V=1) / Pr(z=0 \mid V=1)}{Pr(z=1 \mid V=0) / Pr(z=0 \mid V=0)} \leq \Gamma.$$

Method 3. E-value

$Y(1) = Y, \quad z=1$

$Y(0) = Y, \quad z=0.$

bp's work (assume binary outcomes)



observed relative risk

$$RR^{obs}_{zy|x} = \frac{p(y=1|z=1,x)}{p(y=1|z=0,x)}$$

$$= \frac{\sum_u p(y=1|z=1,x,u)\,p(u|z=1,x)}{\sum_u p(y=1|z=0,x,u)\,p(u|z=0,x)}$$

Causal RR

$$RR^{true}_{zy|x} = \frac{p(y_u=1|x)}{p(y_0=1|x)}$$

$$= \frac{\sum_u p(y=1|z=1,x,u)\,p(u|x)}{\sum_u p(y=1|z=0,x,u)\,p(u|x)}$$

(?) why not $V \to Z$

treatment–confounder association $(Z \to U)$

RR of level $V=u$ (not necessarily binary)

$$RR_{zv|x}(u) = \frac{p(u|z=1,x)}{p(u|z=0,x)}$$

measures of confounding

$$RR_{zv|x} = \max_u RR_{zu|x}(u)$$

confounder–outcome association $(V \to Y)$

$$RR_{vy(z)|x} = \frac{\max_u p(y(z)=1|x,u)}{\min_u p(y(z)=0|x,u)} \quad z=0,1$$

$$RR_{uy|x} = \max\{RR_{uy(0)|x}, RR_{vy(1)|x}\}$$

$$RR^{obs}_{Y1/x} = \frac{Pr(Y=1 \mid Z=1, X=x)}{Pr(Y=1 \mid Z=0, X=x)}$$

其中 $Pr(Y=1 \mid Z=1, X=x)$

$= Pr(Y=1 \mid Z=1, X=x, V=1)\, Pr(V=1 \mid Z=1, X=x) +$

$Pr(Y=1 \mid Z=1, X=x, V=0)\, Pr(V=0 \mid Z=1, X=x)$

$= Pr(Y=1 \mid X=x, V=1)\, Pr(V=1 \mid Z=1, X=x) +$

$Pr(Y=1 \mid X=x, V=0)\, Pr(V=0 \mid Z=1, X=x)$

let $Pr(V=1 \mid Z=1, X=x) = f_{1,x}$

$Pr(V=1 \mid Z=0, X=x) = f_{0,x}$

$= Pr(Y=1 \mid X=x, V=1) \cdot f_{1,x} + (1-f_{1,x}) \cdot Pr(Y=1 \mid X=x, V=0)$

$= Pr(Y=1 \mid V=0, X=x)\left(RR_{V1/x} \cdot f_{1,x} + (1-f_{1,x})\right)$

同理 分母 $= Pr(Y=1 \mid V=0, X=x)\left(RR_{V1/x} f_{0,x} + (1-f_{0,x})\right)$

Drop $X = x$ for simplicity (conduct analysis within strata of covariates or propensity scores)

Goal of sensitivity analysis:

recover RR_{zy}^{true} from RR_{zy}^{obs} and (RR_{zu}, RR_{uy})

$$\rightarrow \quad RR_{zy}^{true} \geq RR_{zy}^{obs} \Big/ \frac{RR_{zu} \times RR_{uy}}{RR_{zu} + RR_{uy} - 1}$$

Stronger $\rightarrow$ $\max(RR_{zu}, RR_{uy})$

$$\geq \left\{ RR_{zy}^{obs} + \sqrt{RR_{zy}^{obs}\left(RR_{zy}^{obs} - RR_{zy}^{true}\right)} \right\} \Big/ RR_{zy}^{true}$$

to explain away the observed RR ($RR_{zy}^{true} = 1$)

$$\rightarrow \max(RR_{zu}, RR_{uy}) \geq RR + \sqrt{RR(RR-1)}$$

E-value

Instrument Variables

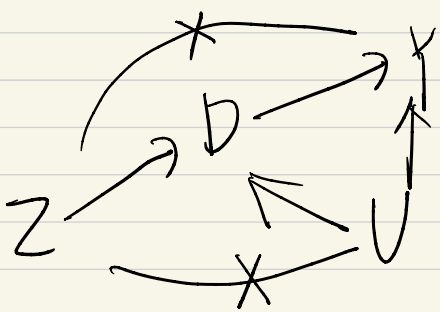
handle unmeasured confounding

Main idea

1. find an IV that influences treatment assignment but is independent of unmeasured confounders and has no direct effect on the outcome (except through its effect on treatment).
(natural experiment, random)

2. use this variable to extract variation in the treatment that is free of the unmeasured confounders

3. use this confounder-free variation in the treatment to estimate the causal effect of the treatment



$$\hat{\beta}_{1,iv} = \frac{\widehat{cov}(Y_i, Z_i)}{\widehat{cov}(D_i, Z_i)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^{N}(D_i - \bar{D})(Z_i - \bar{Z})}$$

When Z binary (Wald estimator)

$$\hat{\beta}_{1,iv} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$$

Interpretation I (indirect least squares)

$$Y_i = \pi_{10} + \pi_{11} \cdot Z_i + \varepsilon_{1i} \quad (+ \pi_{12}' X_i)$$

$$D_i = \pi_{20} + \pi_{21} \cdot Z_i + \varepsilon_{2i} \quad (+ \pi_{22}' X_i)$$

$$\hat{\beta}_{1,iv} = \frac{\hat{\pi}_{11}}{\hat{\pi}_{21}}$$

Interpretation II (2sls)

Stage 1. $\hat{D}_i = \hat{\pi}_{20} + \hat{\pi}_{21} \cdot Z_i \quad (+ \hat{\pi}_{12}' X_i)$

Stage 2. $Y_i = \beta_0 + \beta_1 \hat{D}_i + \eta_i \quad (+ \beta_2' X_i)$

$\hat{\beta}_{1,2sls}$

Discuss IV using potential outcome notation

(focusing on the case of binary treatment and binary IV)

IV $Z=0,1$; treatment $D=0,1$; outcome Y ; covariates X

Potential outcomes: $D(z)$. $Y(z,d)$

Motivating context: randomized experiments with noncompliance ($Z_i \neq D_i$ for some units)

ITT approach:

$$\tau^{ITT} = E[Y_i(1) - Y_i(0)]$$

$$\hat{\tau}^{ITT} = \sum Y_i Z_i / \sum Z_i - \sum Y_i (1-Z_i) / \sum (1-Z_i)$$

estimates the causal effect of the assignment on outcome, but not the effect of treatment received

Effectiveness: "Does the vaccine help people?"

Efficacy: "Does the vaccine work?"

2V Approach to Noncompliance

D is a post-assignment. $b(z)$, $z=0,1$

$Y(z)$ for $z=0,1$ (omit the double index $Y(z,d)$ for simplicity)

Observed data: Z_i, $D_i = D(Z_i)$, $Y_i = Y(Z_i)$

central idea: (i) random assignment Z is an IV

(ii) divide units into latent subgroups based on compliance behavior

Compliance type: $S_i = (D_i(0), D_i(1))$

$D_i(1)$ \ $D_i(0)$	0	1
0	NT	b
1	C	AT

The observed cells of Z, D are mixture

Z	D	S
0	0	[C, NT]
0	1	[AT, D]
1	0	[NT, b]
1	1	[C, AT]

Define 2TT for each compliance type ($s = c, n, a, d$)

$$\tau_s^{ITT} = E[Y_i(1) - Y_i(0) \mid S_i = s]$$

Global 2TT

$$\tau_y^{ITT} = \pi_c \tau_c^{2TT} + \pi_n \tau_n^{2TT} + \pi_a \tau_a^{2TT} + \pi_d \tau_d^{2TT}$$

Assumptions

equivalent to IV estimand when the random assignment be viewed as IV

A1. SUTVA

A2. $cov(Z_i, D_i) > 0$

A3. IV is randomized

A4. ER ($\tau_n^{2TT} = \tau_a^{2TT} = 0$)

A5. monotonicity ($\pi_d = 0$)

$\therefore \tau_y^{2TT} = \pi_c \tau_c^{2TT}$

(global 2TT may be viewed as a conservative estimate of treatment effect $\tau^{2TT} \le \tau_c^{2TT}$)

Complier Average Causal Effect (CACE)

$$\tau^{CACE} = \tau_c^{2TT} = E[Y_i(1) - Y_i(0) \mid S_i = c]$$

$$= \frac{\tau_y^{2TT}}{\pi_c} = \frac{\tau_y^{2TT}}{\tau_D^{2TT}} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$$

Principal Stratification

Post-treatment confounding

$$Z \longrightarrow D \longrightarrow Y$$

Examples:

Non compliance in randomised experiments

$$\tau^{CACE} = E[Y_i(1) - Y_i(0) \mid D_i(0)=0, D_i(1)=1]$$

Selection Bias in cluster randomised trials

$$\tau^{overall} = E[Y_i(1) - Y_i(0)]$$

$$\tau^{recruited} = E[Y_i(1) - Y_i(0) \mid D_i=1]$$

Censoring by death

$$\tau^{SACE} = E[Y_i(1) - Y_i(0) \mid D_i(0)=1, D_i(1)=1]$$

Survival Average Causal Effect

The individual principle strata memberships are not observed → need additional assumption for identifying PCE.

Assumption 1:

$$\{Y_i(0), Y_i(1), D_i(0), D_i(1)\} \perp Z_i \mid X_i$$

The observed (Z, D) consist of mixtures of principal strata.

Estimation of PS inherently involves latent mixture models

Estimation { moment-based
mixture model approach

$$\prod_i \Pr(Y_i(0), Y_i(1), D_i(0), D_i(1), Z_i, X_i, \theta) \quad (?)$$

$$= \prod_i \Pr(Z_i \mid Y_i(0), Y_i(1), S_i, X_i, \theta) \Pr(Y_i(0), Y_i(1) \mid S_i, X_i, \theta) \Pr(S_i \mid X_i, \theta) \Pr(X_i \mid \theta)$$

$$\propto \prod_i \Pr(Y_i(0) \mid S_i, X_i, \theta)^{1-Z_i} \Pr(Y_i(1) \mid S_i, X_i, \theta)^{Z_i} \Pr(S_i \mid X_i, \theta)$$

Regression Discontinuity Designs

RDD. quasi-experimental design

The treatment status changes discontinuously according to some underlying pre-treatment variable (so-called forcing / running variable)

Basic idea: in these studies, comparing units with similar values of this variable, but different levels of treatment would lead to causal effect of the treatment at the threshold.

Formulation

$$Pr(Z=1 \mid S_0^+) \neq Pr(Z=1 \mid S_0^-)$$

- Sharp RD $Pr(Z=1 \mid S_0^+) = 1 \quad Pr(Z=1 \mid S_0^-) = 0$
- Fuzzy RD $Pr(Z=1 \mid S_0^+) > Pr(Z=1 \mid S_0^-)$

Two framework to RDD { continuity-based / local randomization

Sharp RD continuity approach

S_i : the running variable ; S_0 is the cutoff

Z_i : binary treatment indicator $Z_i = I\{S_i \le S_0\}$

Assumption 1 (continuity of conditional Regression functions)

$E[Y(0) \mid S=s]$ and $E[Y(1) \mid S=s]$ are continuous in s

$$\Rightarrow E[Y(0) \mid S=s_0] = \lim_{s \to s_0^-} E[Y(0) \mid S=s]$$

$$= \lim_{s \to s_0^-} E[Y(0) \mid Z=0, S=s]$$

$$= \lim_{s \to s_0^-} E[Y \mid S=s]$$

$$E[Y(1) \mid S=s_0] = \lim_{s \to s_0^+} E[Y \mid S=s]$$

Estimand :

$$\tau_{RD}^{true} = \lim_{s \to s_0^-} E[Y \mid S=s] - \lim_{s \to s_0^+} E[Y \mid S=s]$$

↓

the causal effect at the threshold

Estimation :

$$\hat{\tau}_{RD}^{true} = \hat{f}_+(s_0) - \hat{f}_-(s_0)$$

→ Sharp RD Remarks

Sharp RD completely violates the overlap assumption: there is zero overlap in the treatment probability at the threshold

continuity can be indirectly tested. e.g. distributions of all pre-treatment variables should be very similar on either side of the cutoff

→ Bandwidth Selection

In local linear regression, choice of h boils down to a bias-variance tradeoff.

Smaller h: closer to threshold so less bias.

Smaller sample size so more variance

Fuzzy RD continuity approach

S_i : the running variable ; S_0 is the cutoff

Z_i : binary assignment indicator $Z_i = I\{S_i \le S_0\}$

D_i : treatment received

Estimand :

$$\tau_{FRD} = \frac{\lim_{s \to s_0^-} E[Y|S=s] - \lim_{s \to s_0^+} E[Y|S=s]}{\lim_{s \to s_0^-} E[D|S=s] - \lim_{s \to s_0^+} E[D|S=s]}$$

Assumption 1 (continuity of conditional Regression functions)

$$E[Y(0)|S=s_0] = \lim_{s \to s_0^-} E[Y|S=s]$$

$$E[Y(1)|S=s_0] = \lim_{s \to s_0^+} E[Y|S=s]$$

$\Big\}$ $\tau_{FRD} = \tau_{LACE}(s_0)$

Assumption 2 (monotonicity)

$D_i(s)$ is monotone in s at $s = s_0$

Estimation :

$$\hat{\tau}_{FRD} = \frac{\hat{f}_+(s_0) - \hat{f}_-(s_0)}{\hat{g}_+(s_0) - \hat{g}_-(s_0)}$$

Sharp RD Local Randomization Approach

the running variable as a random variable

Assumption 1. Local overlap: There exists a subset of units, V_{S_0}, in the random sample (or population) in the study such that for each $i \in V_{S_0}$, $\Pr(S_i \le S_0) > \xi$ and $\Pr(S_i > S_0) > \xi$ for some sufficient large $\xi > 0$.

Assumption 2. Local SRD-SUTVA: For each $i \in V_{S_0}$.

$D_i' = \mathbb{1}\{S_i' \le S_0\}$ $D_i'' = \mathbb{1}\{S_i'' \le S_0\}$ with possibly $S_i' \neq S_i''$

If $D_i' = D_i''$, that is, if either $S_i' \le S_0$ and $S_i'' \le S_0$ or $S_i' > S_0$ and $S_i'' > S_0$, then $Y_i(D') = Y_i(D'')$

Estimand:

$$\tau_{S_0} = E[Y_i(1) - Y_i(0) \mid i \in V_{S_0}]$$

Assumption 3. Local randomization: within V_{S_0}

$$\Pr(S_i \mid Y_i(0), Y_i(1), X_i) = \Pr(S_i)$$

$$\tau_{S_0} = E[Y_i \mid D_i = 1, i \in V_{S_0}] - E[Y_i \mid D_i = 0, i \in V_{S_0}]$$