Lecture 11
Conditional randomized experiment, unconfoundedness

Outline

- Conditional randomized experiment
  - Unconfoundedness
  - Balancing score
  - Estimators: outcome regression, IPW, matching
- Imbens and Rubin Chapter 12, Peng's book Chapter 11.3

Conditional randomized experiment

- Treatment assignment mechanism depends on pre-treatment covariates $\boldsymbol{X}_i$
  - Example: stratified randomized experiment, proportion of treated units can be different in different strata
- Unconfoundedness property: $W_i \perp \big(Y_i(0), Y_i(1)\big) \mid \boldsymbol{X}_i$
  - Assignment mechanism does not depend any unobserved $\boldsymbol{U}$ pretreatment confounders
  - $\boldsymbol{X}_i$ can either be continuous or discrete
  - If $\boldsymbol{X}_i$ is discrete or discretized → stratified randomized experiment

- Propensity score: $e(\boldsymbol{X}_i) = P(W_i = 1 | \boldsymbol{X}_i) \in (0,1)$
  - Overlap assumption: $e(\boldsymbol{x}) \neq 0$ or 1 for any $\boldsymbol{x}$ (otherwise we won't have data to identify $\tau(\boldsymbol{x})$)
  - In stratified randomized experiment: $e(\boldsymbol{X}_i = j) = P(W_i = 1 | \boldsymbol{X}_i = j) = N_t(j)/N(j)$

- Identify conditional average treatment effect under unconfoundedness

$$\tau(\boldsymbol{x}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x})$$
$$= \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1) - \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0)$$
$$= \mathbb{E}\big(Y_i^{\text{obs}} \big| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1\big) - \mathbb{E}\big(Y_i^{\text{obs}} \big| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0\big)$$

Conditioning on confounded covariates

- (Population) average treatment effect

$$\tau = \mathbb{E}(\tau(\boldsymbol{X}_i)) = \mathbb{E}\left(\mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i, W_i = 1\right) - \mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i, W_i = 0\right)\right)$$
$$= \sum_{\boldsymbol{x}} \left(\mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1\right) - \mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0\right)\right) P(\boldsymbol{X}_i = \boldsymbol{x})$$

Shared weights

- Conditioning on the confounding covariates $\boldsymbol{X}_i$ is important

$$\mathbb{E}\left(Y_i^{\text{obs}} \middle| W_i = 1\right) - \mathbb{E}\left(Y_i^{\text{obs}} \middle| W_i = 0\right)$$
$$= \sum_{\boldsymbol{x}} \mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1\right) P(\boldsymbol{X}_i = \boldsymbol{x} | W_i = 1) - \sum_{\boldsymbol{x}} \mathbb{E}\left(Y_i^{\text{obs}} \middle| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1\right) P(\boldsymbol{X}_i = \boldsymbol{x} | W_i = 0)$$

Different weights

- If $e(\boldsymbol{X}_i) = P(W_i = 1 | \boldsymbol{X}_i) \equiv c$, then $W_i \perp \boldsymbol{X}_i \Longrightarrow P(\boldsymbol{X}_i = \boldsymbol{x} | W_i = 1) = P(\boldsymbol{X}_i = \boldsymbol{x} | W_i = 0)$

Review of Simpson's paradox

- Compare the success rates of two treatment of kidney stores
- Treatment A: open surgery; treatment B: small puctures

	Treatment A	Treatment B
Small stones	**93%** (81/87)	87% (234/270)
Large stones	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

$P(\boldsymbol{X}_i = \boldsymbol{x})$

(87 + 270)/700=0.51

(263 + 80)/700=0.49

- What is the confounder here? Size of the stone
  - Small stone: propensity score is $\frac{87}{87+270} = 0.24$
  - Large stone: propensity score is $\frac{263}{263+80} = 0.77$
- True average causal effect: $83.2\% - 78.2\% : (93\% \times 0.51\ +\ 73\% \times 0.49)\ -\ (87\% \times 0.51\ +\ 69\% \times 0.49)$

Simpson’s paradox: UC Berkeley gender bias

- In the early 1970s, the University of California, Berkeley was sued for gender discrimination over admission to graduate school.
- “Causal” effect of sex on application admission (data of Year 1973 admission)

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

- Confounding covariate: department

Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

Department	Men		Women		“$e(\boldsymbol{X}_i)$”	$P(\boldsymbol{X}_i)$
	Applicants	Admitted	Applicants	Admitted		
A	825	62%	108	82%	0.12	0.21
B	560	63%	25	68%	0.04	0.13
C	325	37%	593	34%	0.65	0.21
D	417	33%	375	35%	0.47	0.18
E	191	28%	393	24%	0.67	0.13
F	272	6%	341	7%	0.56	0.14

For data from departments A-F:

- Raw average admission rate between men and women:
  46% V.S. 30%
- After adjusting for department:
  40% V.S. 44%

Balancing score

- Under unconfoundedness, we can remove all biases in comparing treated and control units by conditioning on each level of $\boldsymbol{X}_i$
- Too few samples to compare at each level if too many variables in $\boldsymbol{X}_i$

- Balancing score $b(\boldsymbol{X}_i)$: lower-dimensional functions of $\boldsymbol{X}_i$ that remove differences between treatment and control groups

$$W_i \perp \boldsymbol{X}_i \mid b(\boldsymbol{X}_i)$$

  - Balancing scores are not unique: any one-to-one mapping of a balancing score is a balancing score
  - Propensity score $e(\boldsymbol{X}_i)$ is a balancing score
    - We want to show that $P(W_i = 1|\boldsymbol{X}_i, e(\boldsymbol{X}_i)) = P(W_i = 1|e(\boldsymbol{X}_i))$
      - $P(W_i = 1|\boldsymbol{X}_i, e(\boldsymbol{X}_i)) = P(W_i = 1|\boldsymbol{X}_i) = e(\boldsymbol{X}_i)$
      - By the law of total expectation
        $P(W_i = 1|e(\boldsymbol{X}_i)) = \mathbb{E}[W_i|e(\boldsymbol{X}_i)] = \mathbb{E}[\mathbb{E}[W_i|\boldsymbol{X}_i, e(\boldsymbol{X}_i)]|e(\boldsymbol{X}_i)]$
        $= \mathbb{E}[\mathbb{E}[W_i|\boldsymbol{X}_i]|e(\boldsymbol{X}_i)] = \mathbb{E}[e(\boldsymbol{X}_i)|e(\boldsymbol{X}_i)] = e(\boldsymbol{X}_i)$
  - Propensity score the coarsest balancing score (Lemma 12.3 of Imbens and Rubin book): $e(\boldsymbol{X}_i)$ is a function of any $b(\boldsymbol{X}_i)$

Unconfoundedness with balancing score

- Why do we care about balancing score?

$$W_i \perp \big(Y_i(0), Y_i(1)\big) \mid \boldsymbol{X}_i \Longrightarrow W_i \perp \big(Y_i(0), Y_i(1)\big) \mid b(\boldsymbol{X}_i)$$

  - Given a vector of covariates that ensure unconfoundedness, adjustment for differences in balancing scores removes all biases associated with differences in the covariates

  - For the propensity score $W_i \perp \big(Y_i(0), Y_i(1)\big) \mid e(\boldsymbol{X}_i)$

  - $e(\boldsymbol{X}_i)$ can be reviewed as a summary score of the pre-treatment covariate

$$\tau = \mathbb{E}\Big(\mathbb{E}\big(Y_i^{\text{obs}} \big| e(\boldsymbol{X}_i), W_i = 1\big) - \mathbb{E}\big(Y_i^{\text{obs}} \big| e(\boldsymbol{X}_i), W_i = 0\big)\Big)$$

  - The proof can be found on Page 267, Imbens and Rubin Chapter 12.3
    - $P(W_i = 1 | b(\boldsymbol{X}_i)) = P(W_i = 1 | \boldsymbol{X}_i, b(\boldsymbol{X}_i)) = P(W_i = 1 | \boldsymbol{X}_i) = e(\boldsymbol{X}_i)$
    - By the law of total expectation

$$P(W_i = 1 | Y_i(0), Y_i(1), b(\boldsymbol{X}_i)) = \mathbb{E}[\mathbb{E}[W_i | \boldsymbol{X}_i, Y_i(0), Y_i(1), b(\boldsymbol{X}_i)] | Y_i(0), Y_i(1), b(\boldsymbol{X}_i)]$$
$$= \mathbb{E}[\mathbb{E}[W_i | \boldsymbol{X}_i, Y_i(0), Y_i(1)] | Y_i(0), Y_i(1), b(\boldsymbol{X}_i)]$$
$$= \mathbb{E}[e(\boldsymbol{X}_i) | Y_i(0), Y_i(1), b(\boldsymbol{X}_i)] = e(\boldsymbol{X}_i)$$

Estimate ATE under unconfoundedness

- Adjust for confounding variables when estimating the average treatment effect τ
- Three strategies
  - Outcome regression
  - Inverse probability weighting
  - Matching
- We are not introducing new methods to estimate ATE for randomized experiments, we review the estimators we discuss in previous lectures from a different angle, to prepare us to perform causal inference in observation studies

Outcome regression estimator

- $\tau = \mathbb{E}\left(\mathbb{E}(Y_i^{\text{obs}} \mid \boldsymbol{X}_i, W_i = 1) - \mathbb{E}(Y_i^{\text{obs}} \mid \boldsymbol{X}_i, W_i = 0)\right)$
- Define the conditional expectations $\mu_w(\boldsymbol{x}) = \mathbb{E}(Y_i^{\text{obs}} \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = w)$
- We can estimate the conditional expectations via a regression model and obtain $\hat{\mu}_w(\boldsymbol{x})$
- Estimator for the ATE: $\hat{\tau}_{\text{reg}} = \frac{1}{N}\left\{\sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \hat{\mu}_0(\boldsymbol{X}_i)\right) + (1 - W_i)\left(\hat{\mu}_1(\boldsymbol{X}_i) - Y_i^{\text{obs}}\right)\right\}$
- For example, if we assume a linear regression model

$$\mathbb{E}(Y_i^{\text{obs}} \mid \boldsymbol{X}_i, W_i) = \alpha + \tau W_i + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{\gamma}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}})W_i$$

  - $\hat{\mu}_w(\boldsymbol{x}) = \hat{\alpha} + \hat{\tau} w + \hat{\boldsymbol{\beta}}^T \boldsymbol{x} + \hat{\boldsymbol{\gamma}}^T(\boldsymbol{X}_i - \overline{\boldsymbol{X}})w$ where the coefficients are estimated by OLS
  - This is equivalent to fitting two separate linear models on treated units and control units

- $\hat{\tau}_{\text{reg}} = \frac{1}{N}\{\sum_{i=1}^{N} W_i(\hat{\mu}_1(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_i)) + (1 - W_i)(\hat{\mu}_1(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_i))\} = \hat{\tau}$
  - As $\sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \hat{\mu}_1(\boldsymbol{X}_i)\right) = 0$ and $\sum_{i=1}^{N}(1 - W_i)\left(Y_i^{\text{obs}} - \hat{\mu}_0(\boldsymbol{X}_i)\right) = 0$

Outcome regression estimator

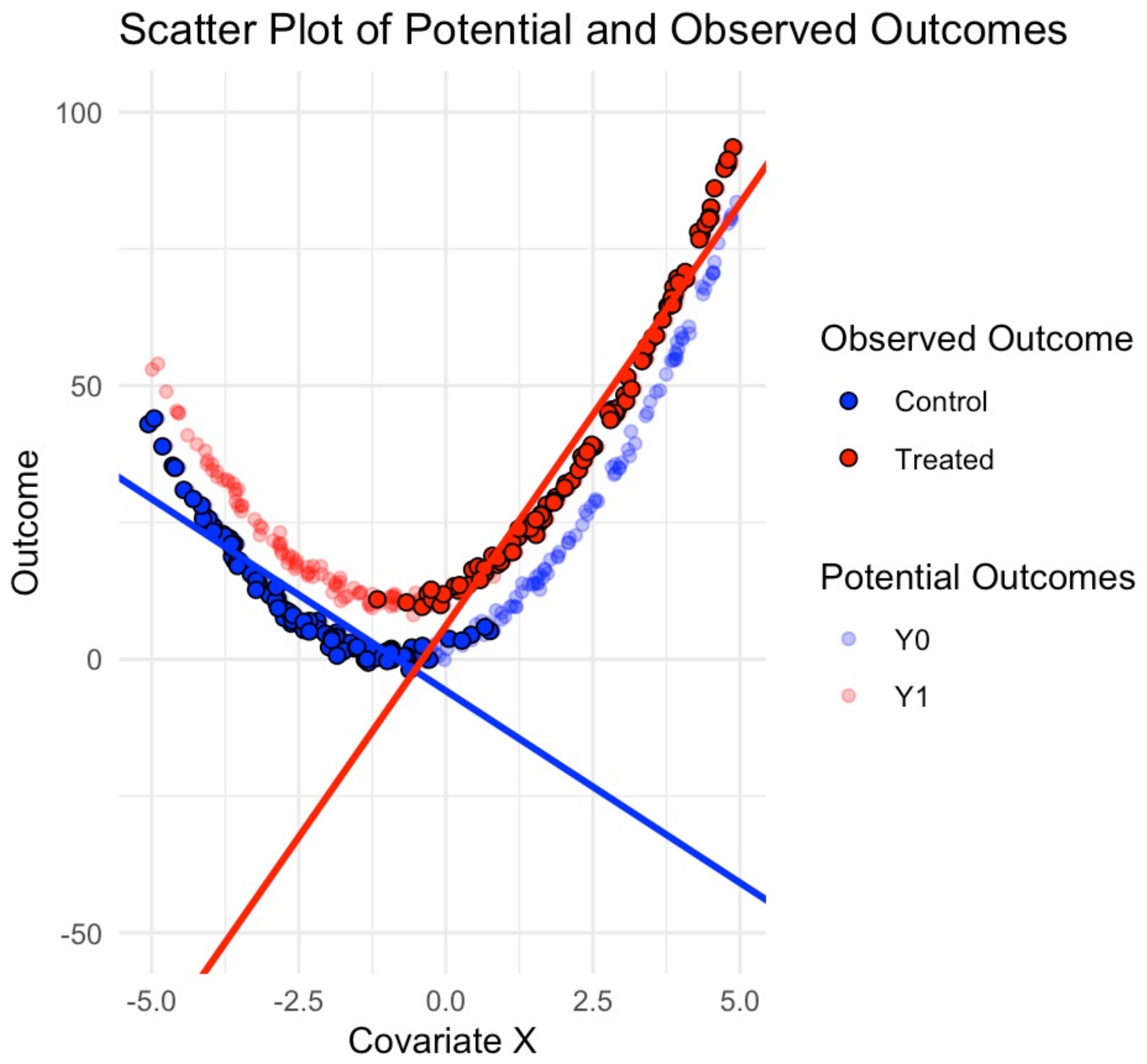
- Unlike in completely randomized experiment where covariates are not confounders, the estimator is not consistent if the linear model is incorrect

- Statistical inference: bootstrap

- In practice, we can use any kinds of machine learning approaches (linear regressions, logistic regression, random forest, SVM, deep learning, …) to obtain $\hat{\mu}_w(\boldsymbol{x})$

- Drawback: does not explicitly rely on overlapping, heavily relies on extrapolation in the region with little overlap

Sensitivity to model mis-specification



- Treatment assignment heavily depend on covariates
- Poor overlapping

- Adjust for X using linear regression for treated and control units separately

- Extrapolation is terribly biased
  - Lead to biased estimate of treatment effect

Inverse probability weighting (IPW)

- What if we don't want to put a model assumption on the observed (potential) outcome?
  - If $\boldsymbol{X}_i$ is unconfounded ($W_i \perp \boldsymbol{X}_i$) and the model assumption is wrong, we may lose efficiency, but $\hat{\tau}_{\text{reg}}$ is likely still unbiased for τ
  - If $\boldsymbol{X}_i$ are confounding covariates and the model assumption is wrong, $\hat{\tau}_{\text{reg}}$ is often be a biased estimator of τ

- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right], \quad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]$$

Proof:

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\left.\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right| X_i\right]\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\left.\frac{Y_i(1) \cdot W_i}{e(X_i)}\right| X_i\right]\right] = \mathbb{E}_{\text{sp}}\left[\frac{\mathbb{E}_{\text{sp}}[Y_i(1)|X_i] \cdot \mathbb{E}_W[W_i|X_i]}{e(X_i)}\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right]$$

Same derivation for the second equation.

Inverse probability weighting (IPW)

- What if we don't want to put a model assumption on the observed (potential) outcome?
  - If $\boldsymbol{X}_i$ is unconfounded ($W_i \perp \boldsymbol{X}_i$) and the model assumption is wrong, we may lose efficiency, but $\hat{\tau}_{\mathbf{reg}}$ is likely still unbiased for τ
  - If $\boldsymbol{X}_i$ are confounding covariates and the model assumption is wrong, $\hat{\tau}_{\mathbf{reg}}$ is often be a biased estimator of τ

- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right], \quad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]$$

  - We give a weight $\lambda_i = 1/P(W_i = w|\, \boldsymbol{X}_i)$ to each unit i, inversely proportional to the probability of being assigned to the group w
  - Intuitively, unit that has a smaller $e(\boldsymbol{X}_i)$ has less chance to appear in the treatment group, so we should give it a higher weight

Inverse probability weighting estimator

$$\hat{\tau}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)}$$

$$= \frac{1}{N}\sum_{i:W_i=1}\lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\lambda_i \cdot Y_i^{\text{obs}},$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i}\cdot(1-e(X_i))^{1-W_i}} = \begin{cases} 1/(1-e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

IVW estimator in stratified randomized experiment

- Propensity score in each strata is $e(\boldsymbol{X}_i = j) = P(W_i = 1 |\ \boldsymbol{X}_i = j) = \frac{N_t(j)}{N(j)}$
- $\hat{\tau}_{\text{IPW}} = \frac{1}{N}\sum_{j=1}^{K}\left(\sum_{i:\,B_i=j}\frac{N(j)}{N_t(j)}W_iY_i^{\text{obs}} - \sum_{i:\,B_i=j}\frac{N(j)}{N_c(j)}(1-W_i)Y_i^{\text{obs}}\right) = \frac{1}{N}\sum_{j=1}^{K}N(j)\left(\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right)$

- Same as the estimator from Neyman's repeated sampling approach

Matching estimator

- In conditional randomized experiments, the IVW estimator do not have any further assumptions as the propensity scores $e(\boldsymbol{X}_i)$ are known.

- Instead of weighting based on $e(\boldsymbol{X}_i)$, we can also perform matching based on $e(\boldsymbol{X}_i)$

- We can match treatment and control unit to form a pair if their propensity scores are very close to each other
  - To assess the effect of job-training program on a thirty-ear-old women with two children under the age of six, with a high school education and four months of work experience in the past 12 months, we want to compare her with a thirty-ear-old women with two children under the age of six, with a high school education and four months of work experience in the past 12 months, **who did not attend the program**

- As $W_i \perp \big(Y_i(0), Y_i(1)\big) \mid e(\boldsymbol{X}_i)$, we can treat the matched data as from a paired randomized experiment