JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Causal Inference: Adjustments

Overview

- Recap:
  - Hypothesis Testing, when and how we can draw conclusions from measurements

- Today:
  - Causal inference definitions
  - Adjustment methods

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Definitions

Recap: What is causal inference?

- Classic setup: is cold medicine effective?

Treatment T
t = took cold medicine
t = did not take medicine

Outcome Y
y = had fever
y = did not have fever

Y(t): The *potential outcome* you would observe under treatment t
The observed outcome that you can *potentially* observe, but that you may not

Recap: Individual Treatment Effect (ITE)

- For each individual *i*,
  - ITE = $Y_i(t=1) - Y_i(t=0)$
  - [Outcome had person *i* taken cold medicine – outcome if they did not]

- ITE is often what we actually care about: should you take cold medicine?

- But we can't measure it!
  - Either you take the medicine or you don't: we can't observe both $Y_i(t=1)$ and $Y_i(t=0)$
  - **Fundamental problem of causal inference**
  - Once an outcome is observed, the unobserved outcome is called the *counterfactual*

Causal Estimand: Average Treatment Effect (ATE)

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

- ATE = $E[Y_i(T=1) - Y_i(T=0)] = E[Y(1)] - E[Y(0)]$
- Does ATE = $E[Y \mid T=1] - E[Y \mid T=0]$?
  - Can we just average over the data in the table, ignoring the missing values?

Image credit: https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf

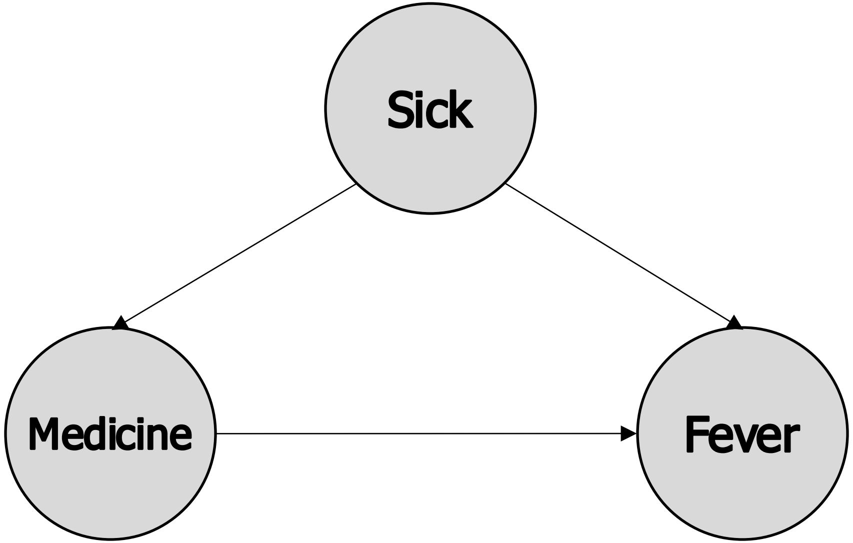
Does ATE = $E[Y \mid T = 1] - E[Y \mid T = 0]$?

- Let's pretend we surveyed a bunch of people. We asked them if they took medicine on Sunday and if they had a fever on Monday

		Has Fever (Y)		
		Yes	**No**	**Sum**
Took Medicine (T)	**Yes**	61	43	104
	No	12	80	92
	Sum	73	113	186

- $E[Y = fever \mid T = medicine] - E[Y = fever \mid T = no\ medicine]$
- $\frac{61}{104} - \frac{12}{92} = 0.4561$
- Value is positive → taking medicine causes fevers?!

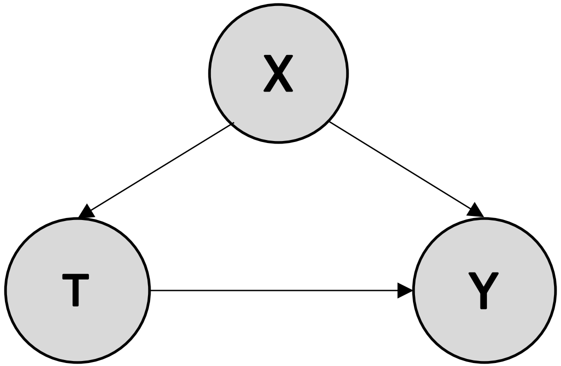
Causal graph and Confounder

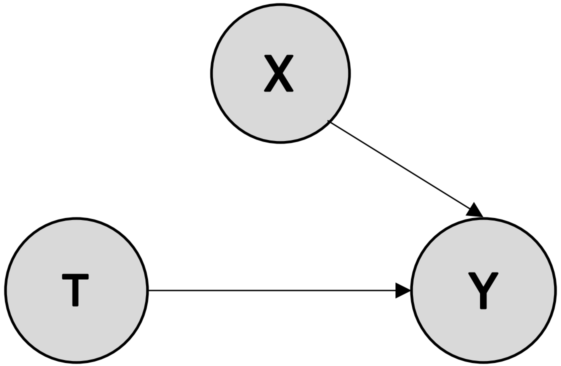


- People only took medicine if they were already feeling sick
- **Confounder:** Variable that affects both probability of receiving treatment and outcome

How can we measure ATE without this problem?

- Randomized control trial (RCT)
- More realistic scenario:
  - We'll probably study effects of medicine on someone who is sick
  - If we survey people, there still might be differences: lower income person may not be able to afford medicine and may also have worse nutrition that leads to more severe illness: income is a confounder (X)
- Instead of surveying people, we take a group of people and randomly assign them to "treatment" or "control" group





Randomized Control Trials

- How can we conduct randomized control trials for:
  - Effects of smoking on lung cancer
    - We randomly assign people to smoke or not smoke?
  - Effect of gender on hiring decisions: do men get more higher paying job offers?
    - We randomly assign people genders??

- Many of the questions we may want to study are not possible to investigate through randomized control trials
- We have to rely on ***observational*** data

The rest of this class

- When/how can we estimate ATE directly from the data?
- How do we adjust for confounders?

- Properties/assumptions of causal inference
- Adjustment methods:
  - Regression
  - Propensity scores: matching, weighting, stratification
  - Additional notes

- We're discussing fundamental concepts in more abstract terms
- Next class, we'll look at more concrete examples involving text

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Properties/Assumptions and Regression

When/how can we estimate ATE from the data?

1. Conditional Exchangeability / Unconfoundedness
2. Positivity
3. No interference
4. Consistency

Ignorabilty/Exchangeability

$$(Y(1), Y(0)) \perp\!\!\!\perp T$$

- The potential outcomes of an individual does not depend on whether or not they really have been treated
- Potential outcome Y(1) and potential outcome Y(0) have the same values, whether or not they were treated
- We can <u>ignore</u> the missing data
- Alternative view: <u>exchangeability</u> if we swap the treatment and control groups, the new treatment group would observe the same outcomes as the old one

$$\begin{aligned}\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] &= \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] \\ &= \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]\end{aligned}$$

Conditional Exchangeability / Unconfoundedness (#1)

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

- Within levels of X, potential outcomes and treatment are not associated.
- Controlling for X makes the treatment and control groups comparable

- [Main assumption needed for causal inference]
- [More on how to "control for X" in a few moments]

Positivity (#2)

- For all values of covariates X present in the population of interest (i.e. x such that $P(X = x) > 0$):

$$0 < P(T = 1 \mid X = x) < 1$$

- Example: Imagine that the treatment group is all men, $P(T = 1 \mid X = \text{woman}) = 0$. Can we really estimate effects of treatment on all people?
- Mathematically, we end up conditioning on a zero-probability event and dividing by zero.
- Alternative view: *overlap*, we only can estimate causal effects where there is overlap between treatment and control group
- [Can be hard to satisfy for high-dimensional covariates]

No interference (#3)

$$Y_i(t_1, \ldots, t_{i-1}, t_i, t_{i+1}, \ldots, t_n) = Y_i(t_i)$$

- Outcome of one individual is unaffected by anyone else's treatment
- Example: I took cold medicine but roommates didn't → I had a fever because I caught a new infection from them
- Commonly difficult to satisfy in network studies

Consistency (#4)

$$T = t \implies Y = Y(t)$$

- If the treatment is T, then the observed outcome Y is the potential outcome under treatment T
- Example:
  - Individual had a fever in the morning but not the afternoon: we didn't specify what "having a fever the next day" means

How can we measure ATE?

- Given the assumptions of unconfoundedness, positivity, consistency, and no interference, we can identify the average treatment effect

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X\left[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\right]$$

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] && \text{(linearity of expectation)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]\right] \\
& && \text{(law of iterated expectations)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y(1) \mid T = 1, X] - \mathbb{E}[Y(0) \mid T = 0, X]\right] \\
& && \text{(unconfoundedness and positivity)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\right] \\
& && \text{(consistency)}
\end{aligned}$$

No interference justifies that this is the value we want to measure (instead of lefthand side of Slide 17)

How can we measure ATE?

- Given the assumptions of unconfoundedness, positivity, consistency, and no interference, we can identify the average treatment effect

$$\mathbb{E}[Y(1) - Y(0)] = \boxed{\mathbb{E}_X}\ [\boxed{\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]}]$$

Use a model to estimate E[Y | T = t, X = x]

$$\frac{1}{n}\sum_i [\mathbb{E}[Y \mid T = 1, X = x_i] - \mathbb{E}[Y \mid T = 0, X = x_i]]$$

Replace outer expectation with empirical mean over data

Pseudocode: Regression Adjustment

- X = [took medicine; felt sick yesterday; has underlying health condition]
- y = [had fever next day]
- Fit model (e.g. regression) over X, y
- Compute mean [model predictions for data where T = 1] – [model predictions for data where T = 0]

[This is your HW, except we use continuous values for some variables and directly look at coefficients instead of computing ATE]

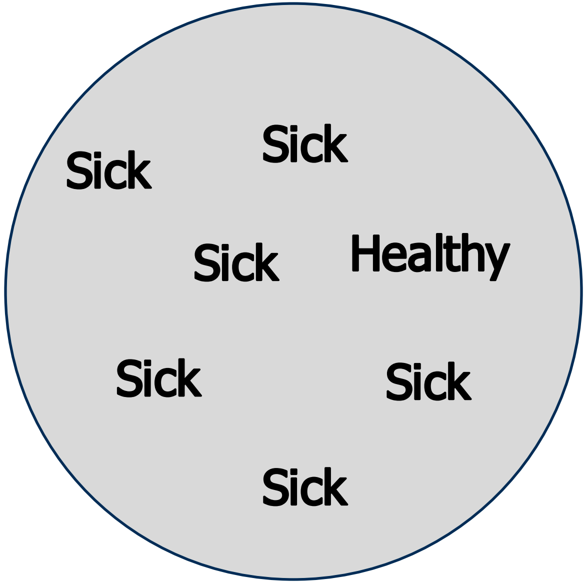
Takeaways

- We need to make assumptions about our data to do this type of estimation
- Lots of carelessness around assumptions in practice

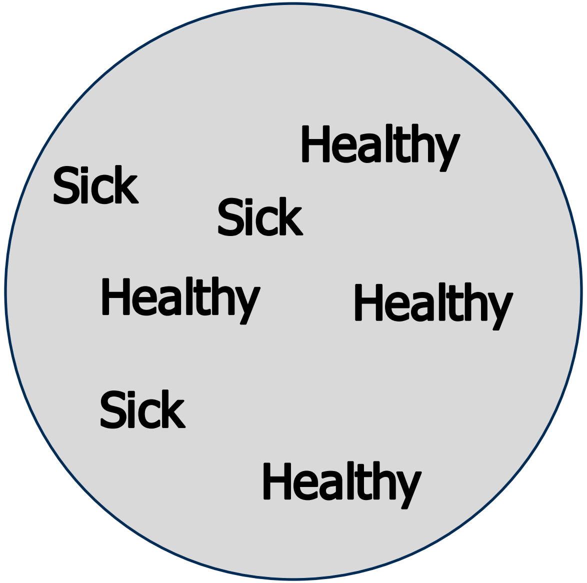
Matching and Propensity Scores

Johns Hopkins
Whiting School of Engineering

Direct Matching

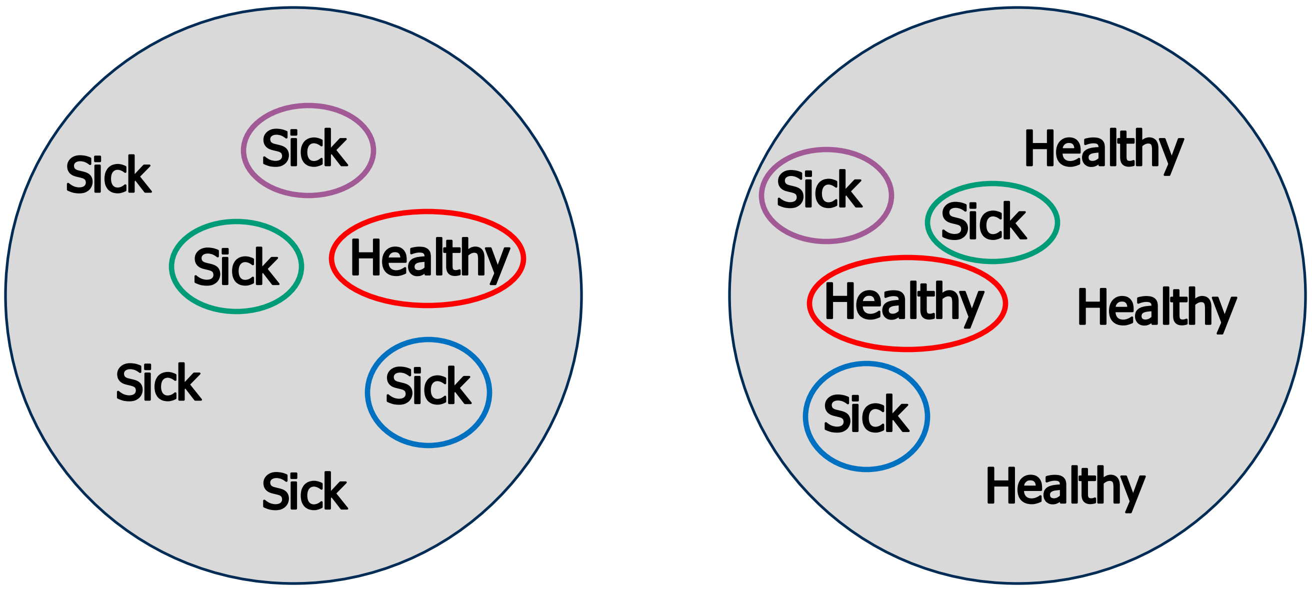


Took Medicine



Didn’t Take Medicine

Direct Matching



Took Medicine **Didn't Take Medicine**

Direct Matching

- We force treatment and control groups to be comparable by matching each person who received treatment with someone who did not but who otherwise had similar characteristics

- Lots of variants on how exactly to do this:
  - Greedy matching vs. optimal matching
  - Allowing multiple matches
  - Discarding bad matches

- Some data is better suited to matching approaches than others (e.g. matching is better if there are many more control individuals than treated individuals)

Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21. doi: 10.1214/09-STS313

4 Basic Steps to matching

1. Defining “closeness”: the distance measure used to determine whether an individual is a good match for another

2. Implementing a matching method, given that measure of closeness

3. Assessing the quality of the resulting matched samples, and perhaps iterating with Steps (1) and (2) until well-matched samples result

4. Analysis of the outcome and estimation of the treatment effect, given the matching done in Step (3)

Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21. doi: 10.1214/09-STS313

Propensity Score

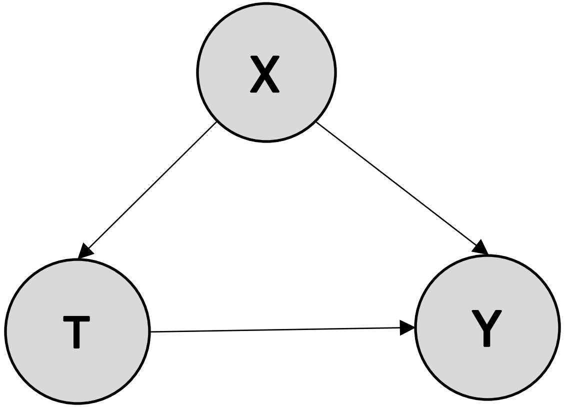
- X might be high dimensional, it it necessary to match on (or more generally adjust for) all of X?

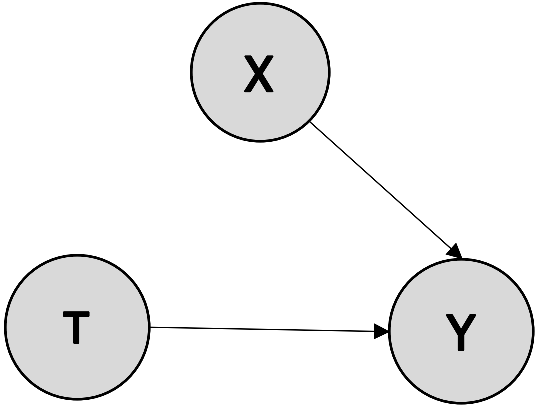
- Define the *propensity score* as the probability of receiving treatment, given confounders:
  - $e(X) = P(T = 1 \mid X = x)$

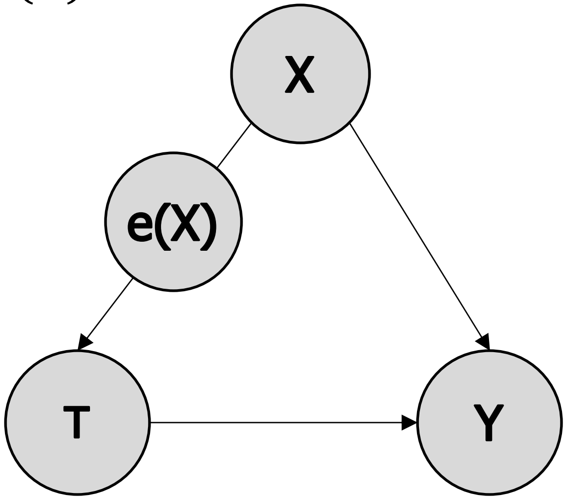
Propensity Score Theorem

- Given positivity, unconfoundedness given X implies unconfoundedness given the propensity score e(X)

$$Y(1), Y(0) \perp T \mid X \;\Rightarrow Y(1), Y(0) \perp T \mid e(X)$$







Propensity Score Theorem

- Given positivity, unconfoundedness given X implies unconfoundedness given the propensity score e(X)

$$Y(1), Y(0) \perp T \mid X \;\Rightarrow Y(1), Y(0) \perp T \mid e(X)$$

- When we are adjusting for X, we can swap in e(X) instead
- We don’t typically actually know e(X) but we can estimate it from the data

Estimating Propensity Scores

- Train a model (e.g. Logistic Regression) to predict T from X
  - Use output scores of model as propensity scores

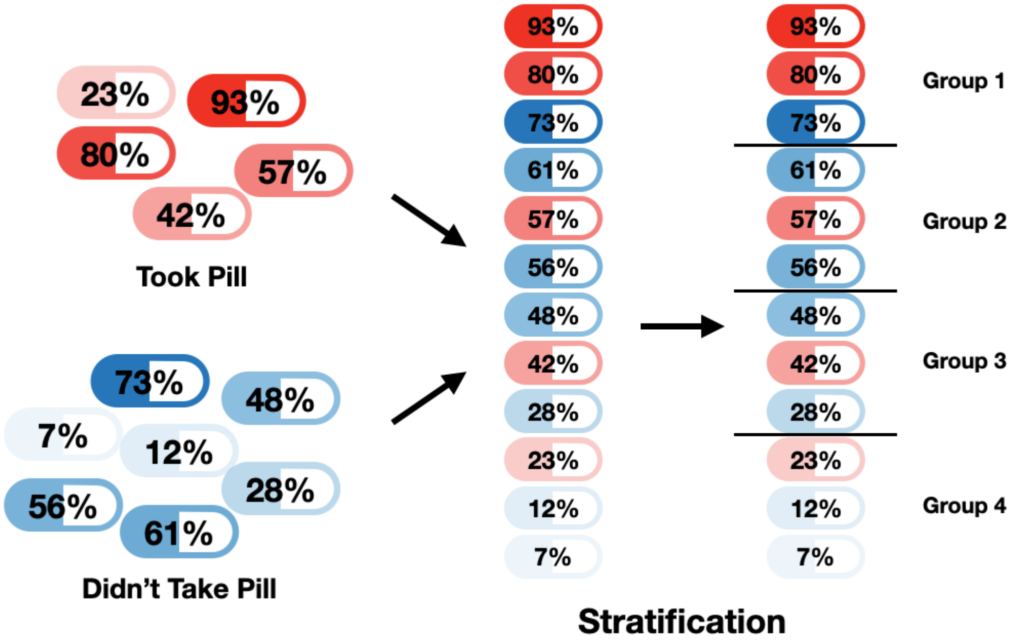
- It's easy to overfit, especially as X becomes higher-dimensional:
  - Use held-out data or cross validation approach so that you are not training and estimating on the same data

Propensity Matching

- We can match treatment and control groups using propensity scores instead of covariates directly
- We define ”closeness” as similar propensity scores
- Advantages (compared to direct matching):
  - Lower-dimensional data
  - Evidence that this works better than direct matching
  - Recall the definition of confounder: we only want to adjust for covariates that are predictive of treatment, propensity scores figures out which values those are for us
- Disadvantages (compared to direct matching)::
  - Matches are no longer meaningful (we can’t tell if they look reasonable from looking at them)

Propensity Stratification

- Stratify (bucket) individuals into mutually exclusive subsets with the same propensity score
- 5 subsets (quintiles) is a common choice
- Compute estimand for each strata and them pool them (typically weighted equally)



Rosenbaum P.R., Rubin D.B. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984;79:516–524
Image: https://towardsdatascience.com/propensity-score-5c29c480130c

IPW (Inverse probability weighting)

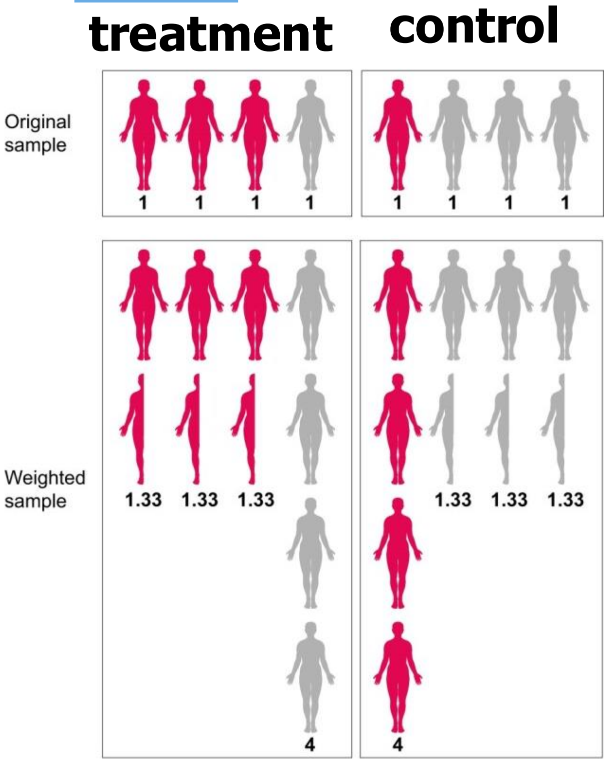
$$w_i = \frac{T_i}{e(X_i)} + \frac{1 - T_i}{1 - e(X_i)}$$

- Define weight: inverse estimate of the probability of the treatment that the individual actually received

ATE = weighted avg. of treated individuals – weighted avg. of untreated individuals

- [Also called IPTW, inverse probability of treatment weighting]

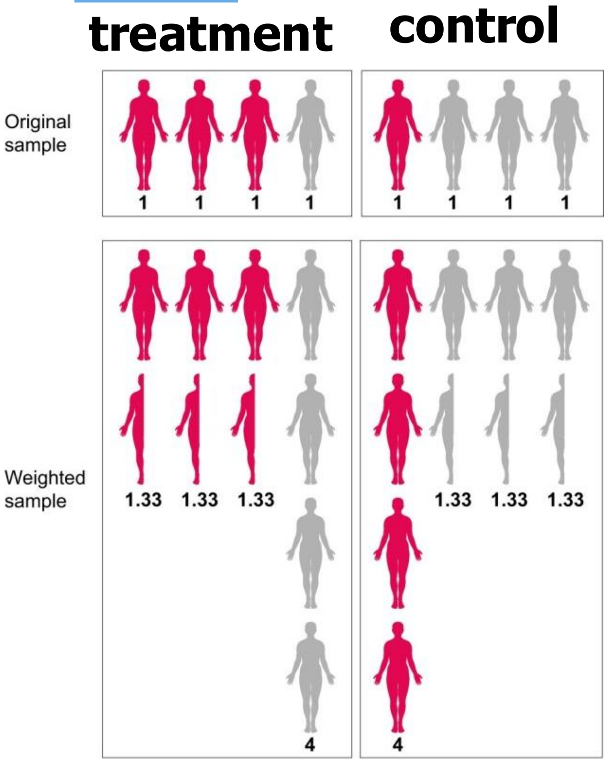
IPW (Inverse probability weighting)



- Setup: Red = felt sick
  - ¾ people who felt sick took medicine
  - P(taking medicine | feel sick) = 0.75
  - P(no medicine | feel sick) = 0.25
- Weights:
  - Took medicine, felt sick: 1/0.75 = 1.333
  - No medicine, felt sick: 1/.25 = 4
  - [similarly calculate weights for people who didn’t feel sick]
- When we apply weights, we’ve balanced feeling sick with not feeling sick

Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, Jager KJ. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2021 Aug 26;15(1):14-20. doi: 10.1093/ckj/sfab158

IPW (Inverse probability weighting)



- We're creating "pseudeopopulations"
- Similar concept: when collecting survey data, you may upweight respondents of particular demographics to match population statistics

How do propensity adjustment methods compare?

- Often choice depends on what model is best suited to data and analysis
- Several studies have demonstrated that propensity score **matching** eliminates a greater proportion of the systematic differences than **stratification** (Austin, 2009a; Austin, Grootendorst, & Anderson, 2007; Austin & Mamdani, 2006)
- In some settings propensity score **matching** and **IPTW** were shown to be comparable; in others propensity score matching was slightly better (Austin, 2009a)

Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786.

Quiz

1. We're trying to measure the effect of gender on Wikipedia article length, where occupation is a confounder. What would be a violation of positivity?

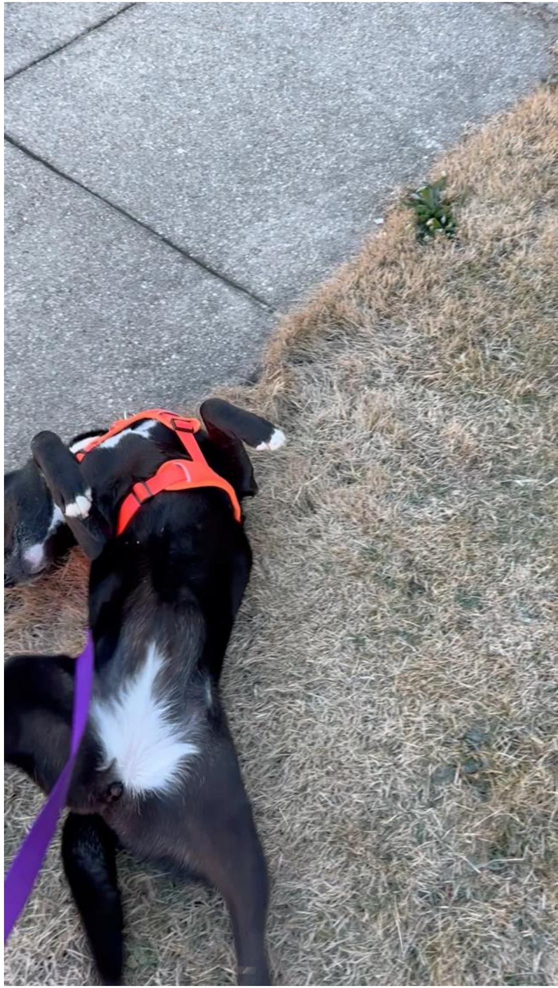
A. There are no women professional U.S. football players with Wikipedia articles

B. There is a higher bar for women to have Wikipedia articles than men

C. Articles about women are more likely to be written during gender equality Wikipedia edit-a-thons

2. In our running example of taking cold medicine, what would you expect calculated propensity scores to look like?

A. People who are sick all have high scores and others have low scores
B. People who took medicine all have high scores and others have low scores
C. People who are sick all have low scores and others have high scores
D. People who took medicine all have low scores and others have high scores



Regression vs. Matching?

- “matching methods should not be seen in conflict with regression adjustment and in fact the two methods are complementary and best used in combination”
  - E.g. you could stratify based on propensity scores and then use regression adjustment with each statum to adjust for lingering differences
- “matching methods highlight areas of the covariate distribution where there is not sufficient overlap between the treatment and control groups, such that the resulting treatment effect estimates would rely heavily on extrapolation”
- ”methods such as linear regression adjustment can actually increase bias in the estimated treatment effect when the true relationship between the covariate and outcome is even moderately non-linear”

Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21. doi: 10.1214/09-STS313

Some additional notes

Double Machine Learning

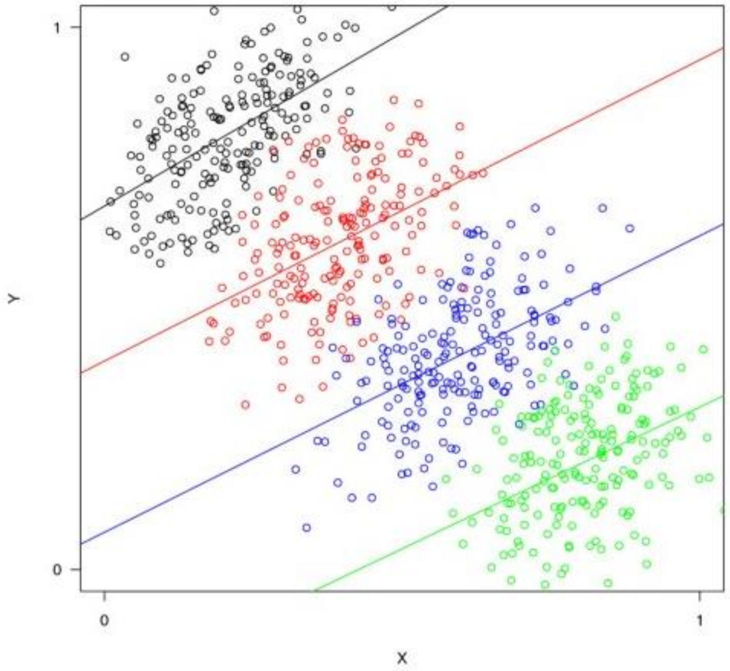
- General framework for estimating causal effects using ML (random forests, lasso or post-lasso, neural nets, boosted regression trees, and various hybrids and ensembles of these methods)

- Available in Python and R packages:
  - https://github.com/DoubleML

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68, https://doi.org/10.1111/ectj.12097

Mixed Effects Regression Models

- We discussed regression adjustment for confounders
- When data is hierarchical / non-independent we need a better regression model
- E.g. you examine if dosage of medicine affects fevers
- Your data is from hospitals in different countries where underlying health conditions that affect baseline health
- Recall Simpson's Paradox



- Data looks negatively correlated overall
- Subsetting data shows positive correlations

Mixed Effects Regression Models

- We can account for differences across subgroups by allowing subgroups to have different parameters (e.g. different intercepts in linear regression)
- Subgroup is a *random* effect
- Dosage is a *fixed* effect

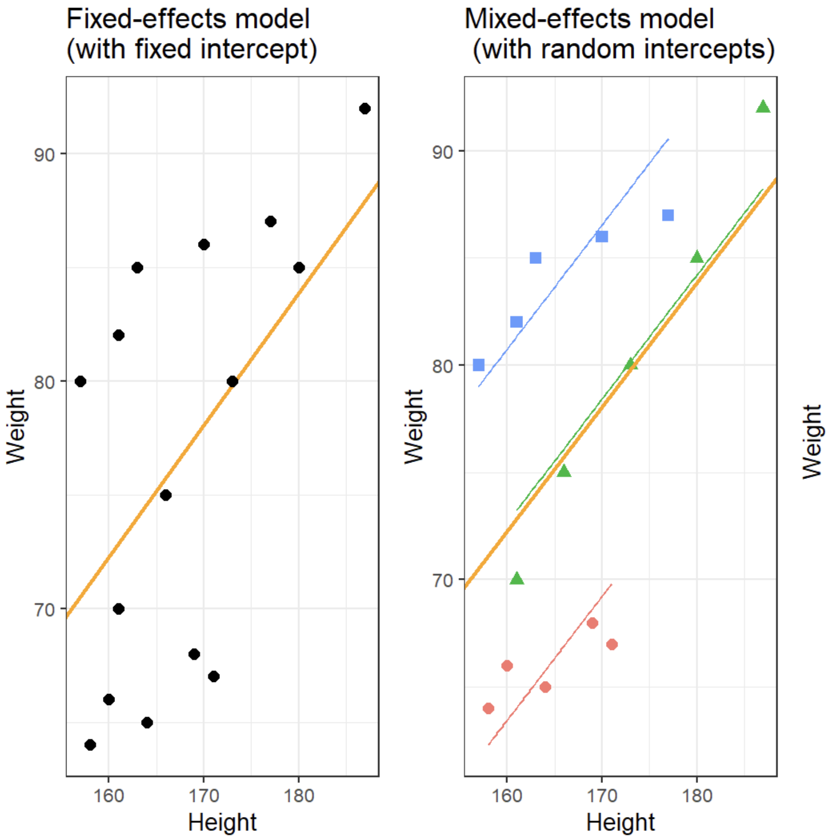
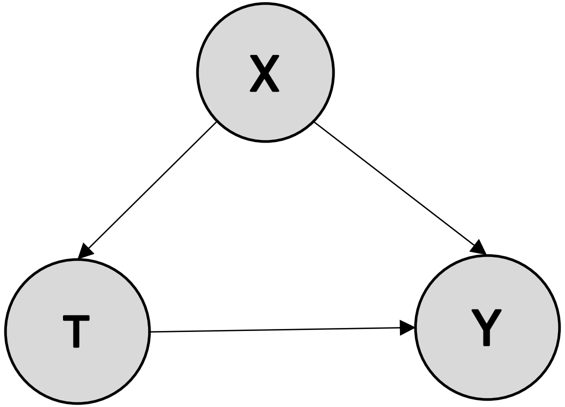
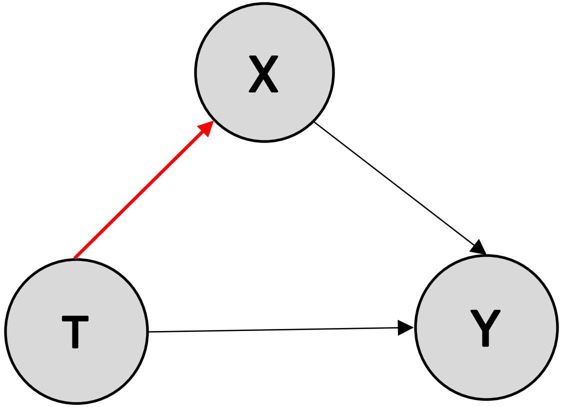


Image credit: https://ladal.edu.au/regression.html

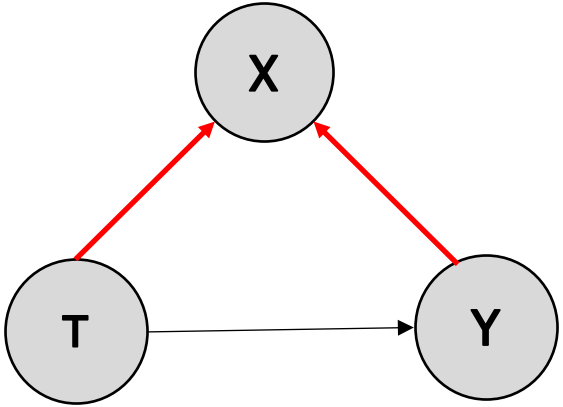
Confounders vs. Mediators vs. Colliders



confounder

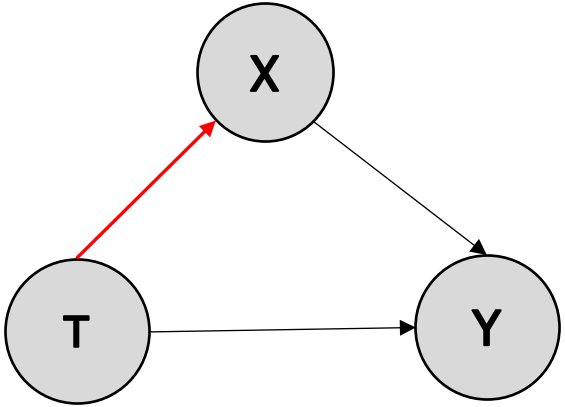


mediator



collider

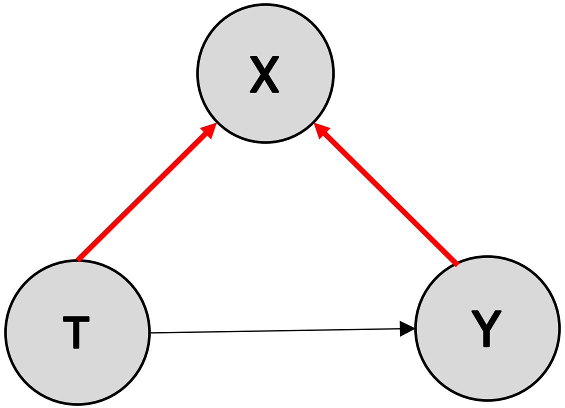
Confounders vs. Mediators vs. Colliders



mediator

- Example:
  - Estimating if gender has an effect on social media likes
  - Gender (T) influences the topic of posts (X)
  - Topic of posts (X) and gender (T) influence number of likes (Y)
- If we adjust for X, we may be removing some of the effect
- We may still choose to adjust for X if we specifically want to capture the direct effect and not the indirect effect
- We may want to separate out direct and indirect effects in a mediation analysis

Confounders vs. Mediators vs. Colliders



collider

- Example:
  - Studying if getting a dog makes people wake up earlier
  - Getting dog (T) influences wake up time (Y) and if you take morning walks (X)
  - People who happen to wake up early (Y) take morning walks too (X)
  - If you condition on X (e.g. restrict data to people who take morning walks), you're selecting for people who wake up early in your control group → you find that having a dog makes you get up later
- If we adjust for X, we are adding bias to our estimator!

Takeaways

- Methods for adjusting for confounders
    - Regression
    - Matching
    - Propensity scores (matching, weighting, and stratification)
- Confounders vs. Mediators vs. Colliders

- Next class:
    - Case studies of causal inference involving NLP and text

Logistics

- HW 1 and 2 grades
- HW 3
- Midterm

References

- Brady Neal, “Introduction to Causal Inference from a Machine Learning Perspective”, Course Lecture Notes, Chapter 2, https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf
- Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21. doi: 10.1214/09-STS313
- Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, Jager KJ. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2021 Aug 26;15(1):14-20. doi: 10.1093/ckj/sfab158
- Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786.