Inferring Causal Effects from Observational Data

This is Yixin's study note for the [Coursera course: A Crash Course in Causality: Inferring Causal Effects from Observational Data] by Professor Jason A. Roy from University of Pennsylvania.

Global Definition

These are definitions you will see through the entire course.

- Confounding: confounders(X) are variables that affect treatment(A) and the outcome(Y)
- Ignorability / Exchangeability: confouders (X) fully captures the confoundings
- Positivity assumption: the probability for each treatment for a given set of X's should be non-zero

I. Confounding and DAG (Directed Acyclic Graphs)

Motivation

- We want to have **a sufficient set of confounders to control for** when carrying out casual inference
- Graphs (DAG) are useful for causal inference: helpful for **indentifying** what variables to control for; make assumptions explicit

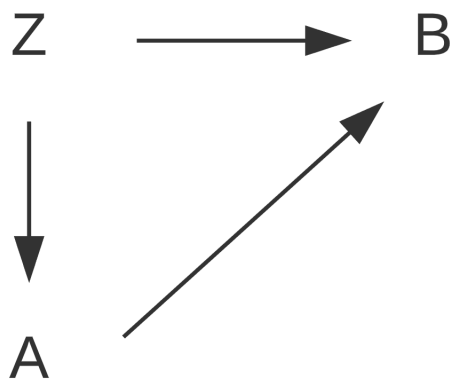
What (Definition)

DAG

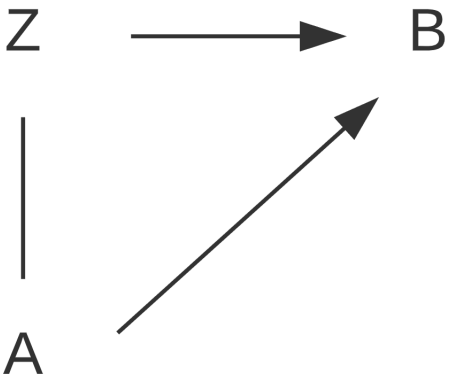
- A effects Y
- A and Y: nodes, vertices (our variables)

- Link (edge): directed path
- This is a DAG, because all links between variables are directed and not cycle

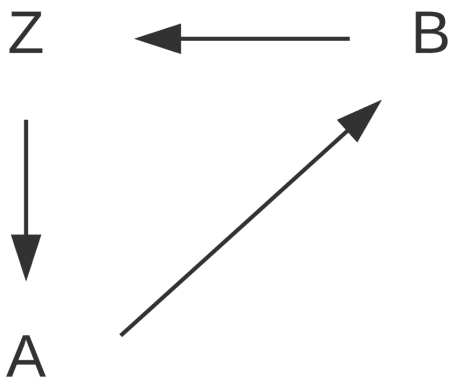
This is a DAG:



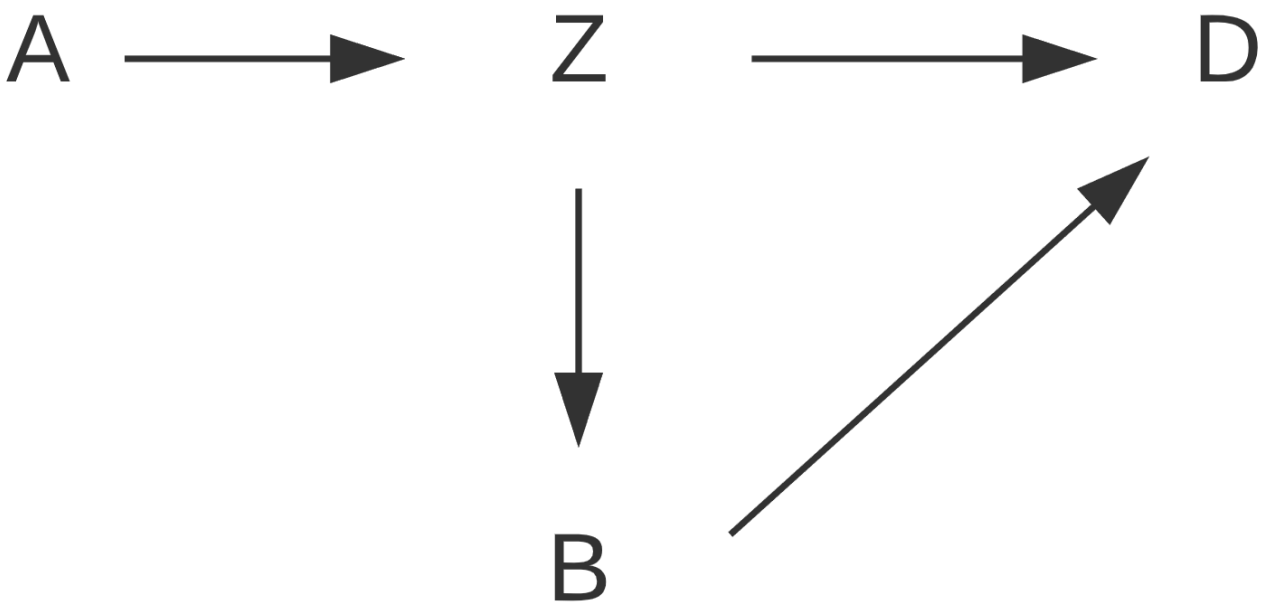
NOT a DAG :



NOT a DAG:



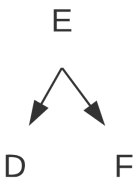
Even more terminology..



- A is Z's **parent**
- B is a **child** of Z
- D is a **descendant** of A
- Z is an **ancestor** of D
- D has 2 **parents**: B, Z

Types of Paths

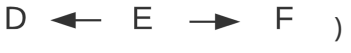
- **Fork**: D ← E → F
  - also written as:



- **Chain**: D → E → F
- **Inverted forks** (E is a '**collider**'): D → E ← F

Association

If node **D** and **F** are on the ends of a path, they're associated if :

1. some information flows to both of them (e.g. )
2. Information from one makes it to the other (e.g. )

Paths that do not induce association, inverted forks for example: (e.g. )

Conditional independence (d-seperation)

Blocking

1. Association on a **chain/ fork** can be *blocked* by conditioning on E.
2. But the opposite occurs if a **collider** is conditioned on: conditioning on E opens a path for D and F, and thus D and F becomes associated (they were originally independent).

D-seperation rules (🌟important)

Def: A and B are d-seperated by a set of nodes C if C **blocks every path** from A to B

- $A \backslash \text{independent} B | C$

A path is d-seperated by a set of nodes C if:

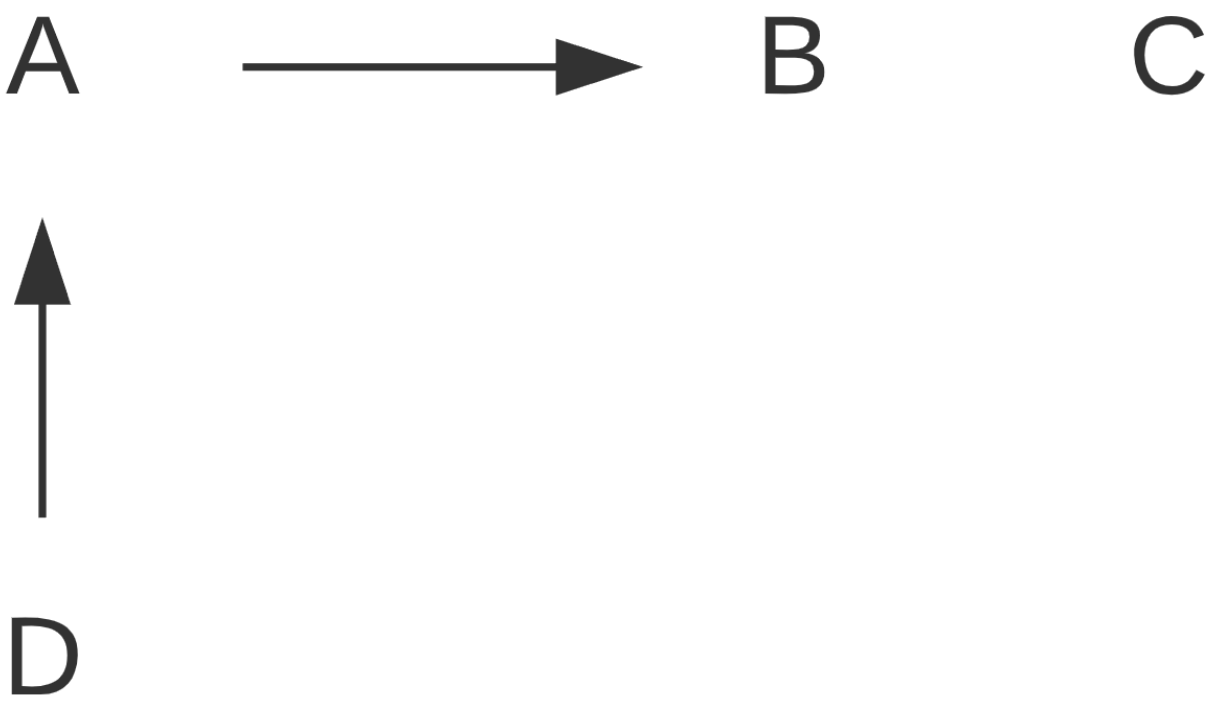
- the path contains a chain and the middle part is in C, OR
- the path contains a fork and the middle part is in C, OR
- the path contains an inverted fork and the middle part is NOT in C, NOR are any descendants of C

How (to use it)

1. DAG and Probability Distributions

A DAG indicates the independency, conditional independency between variables.

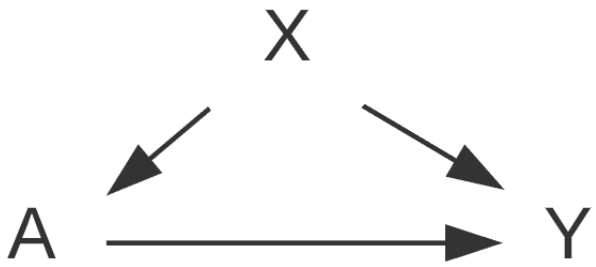
example 1:



1. C is independent of all other variables $P(C|A,B,D) = P(C)$
2. B depends on A conditioned on D, C $P(B|A,C,D) = P(B|A)$
3. B and D are marginally dependent: $P(B|D)! = P(B)$
4. $P(D|A,B,C) = P(D|A)$

2. DAG and Causal Relationship

Recall that a confounder(X) affects both the treatment(A) and the outcome(Y)



Backdoor paths

- There's no need for us to identify every specific confounders, we just need to identify a set of variables ***sufficient to control for confounding***. And ***blocking backdoor paths*** would do that.
- Frontdoor path: **A -> Y**, which captures the effect of treatment. This is what we want to observe, so we do not worry about frontdoor paths.
- Backdoor path: paths from A to Y that travel through arrows going into A: **A <- X -> Y**. This path

confound the relationship between A and Y (i.e. **A->Y** is left).

Then, what are the 2 criteria for indentifying sets of variables sufficents to control for confounding?

3. Backdoor path criterion

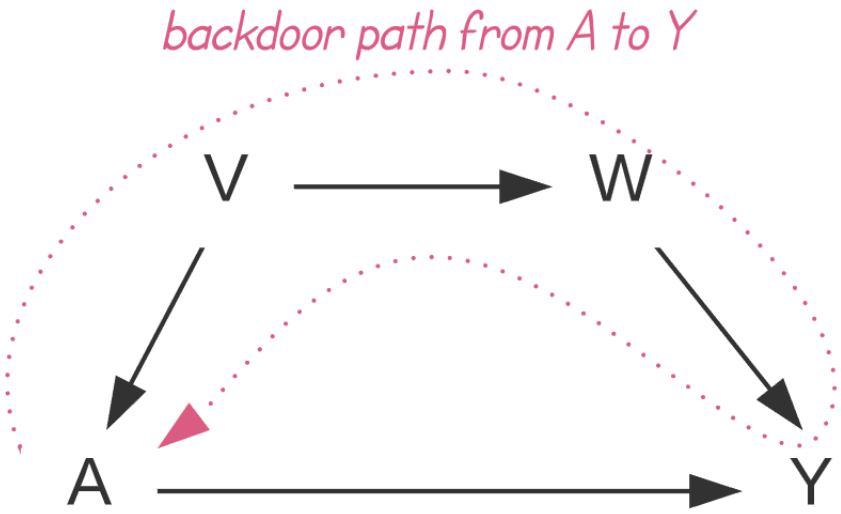
A set of variables ***X*** is sufficient to control for confounding if:

1. it **blocks all backdoor paths** from treatment to the outcome
2. it does not include any descendants of the treatment

Note: ***X*** is not necessarily unique.

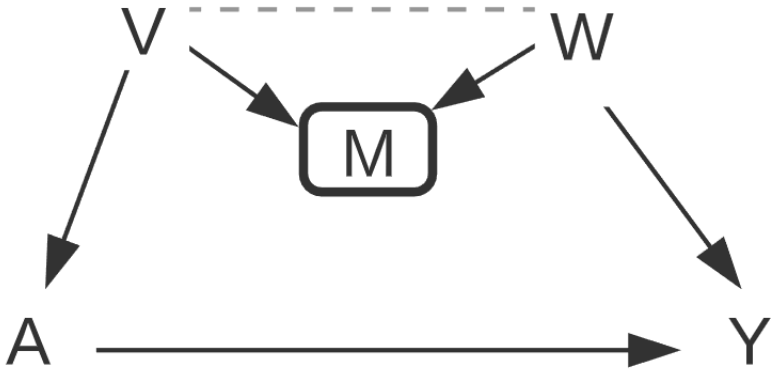
Using d-seperation rules mentioned above, let's work on the examples.

- example 1:



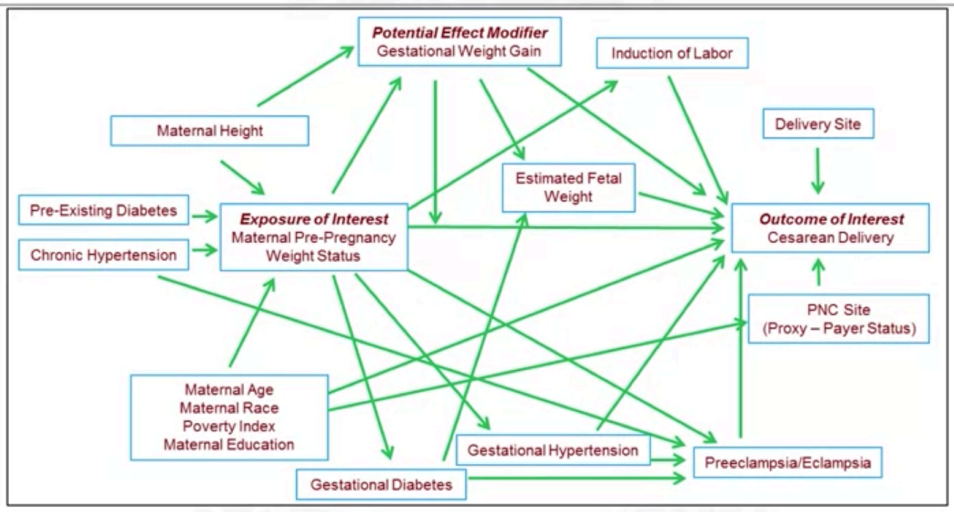
  - There is confounding, because info from ***V*** flows to both A and Y.
  - The only one backdoor path is not blocked by a collider. And according to d-seperation rules, the sets of variables sufficient to control for confounding are: {V}, {W}, {V,W}

- Example 2:



- There is no confounding. Because 1) from neither V or W flows info to A and Y 2) info from A(Y) does not flow into Y(A).
- ***M*** is a collider. Remember that conditioning on a collider would open a path for V and W. Thus, we either 1) not conditioning on M or 2) conditioning on M and one of V and W.
- sets of variables sufficient to control for confounding are: {}(since there's no confounding), {V}, {W},{V,M},{V,W},{V,M,W}

However, in reality, the actual framework could be as complicated as follows



Adapted from:

Figure 1. Directed acyclic graph (DAG) of the association between maternal pre-pregnancy weight status and the risk of cesarean delivery.

Anjel Vahratian, Anna Maria Siega-Riz, David A. Savitz, Jun Zhang. **Maternal Pre-pregnancy Overweight and Obesity and the Risk of Cesarean Delivery in Nulliparous Women.** Annals of Epidemiology, Volume 15, Issue 7, 2005, 467–474. http://dx.doi.org/10.1016/j.annepidem.2005.02.005.

PennMedOnline

and we need an alternative criterion that does not require knowledge of the whole DAG.

4. Disjunctive Cause Criterion

- The critierion is simple: control for all (observed) causes of the exposure, the outcome, or both.
- If a set of observed variables satisfy the backdoor path criteria: the variables selected based on the disjunctive cause criterion will be sufficient.
- Example:
  - {M,W,V}: observed pre-treatment variables
  - $\{U_1, U_2\}$: unobserved pre-treatment variables
  - Suppose we know that W&V are causes of A,Y, or both; M is not a cause of either A or Y
  - The point here is to compare 2 methods: 1) controlling all pre-treament variables {M,W,V}; 2) controlling variables based on disjunctive cause criterion {W,V}.

[TBC: hypothetical examples]

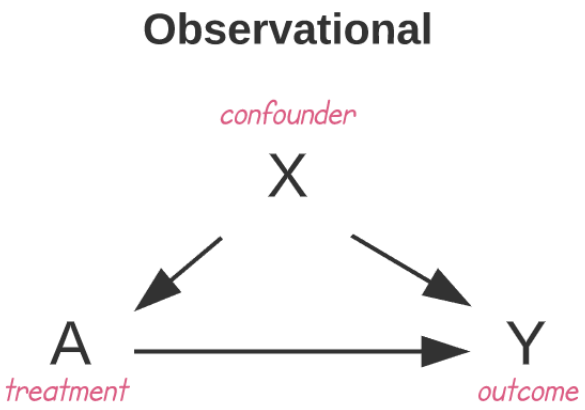
II. Matching and Propensity Scores

In the last chapter, we learn what variables should be controlled. In the following chapters, we will learn how to control confounding.

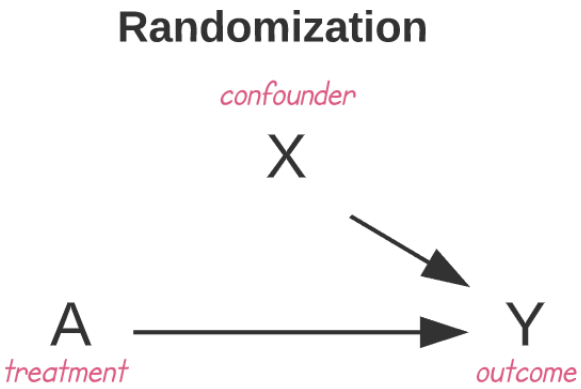
Motivation

Why observational, not randomized?

Recall the DAG above: This DAG represents observational studies.



A randomized trial effectively erase the arrow from X to A. At **design phase**, we make sure that the distribution of X that affects Y are the same across treatment groups. Thus, if the outcome Y distribution ends up differing, the differences will not be attributed to differences in X.



However, randomized trails could be **expensive** (both money and time) or sometimes even **unethical**. We need methods to enable us to carry causal analysis with observational data.

What (definition)

Matching

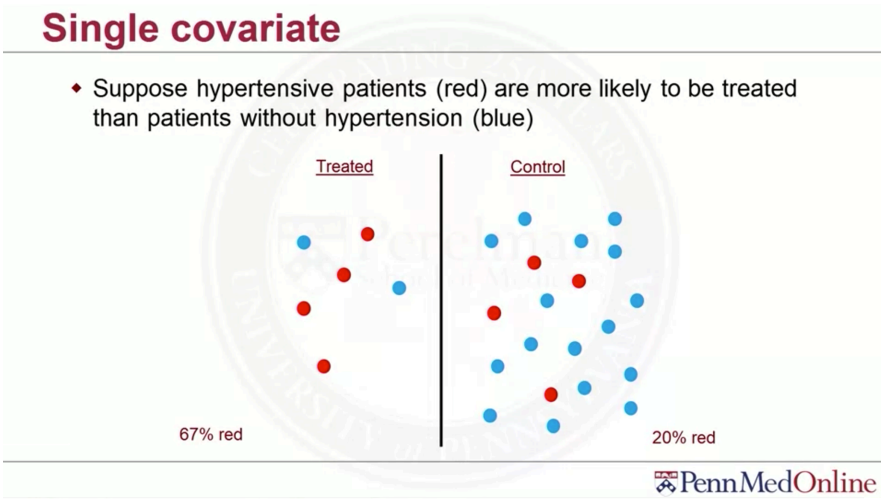
Matching is a method that attempts to make an observational study more like a randomized trail. The main idea is to match individuals in the treated group (A=1) to individuals in the control group (A=0) on the covariates (X). An example would be :

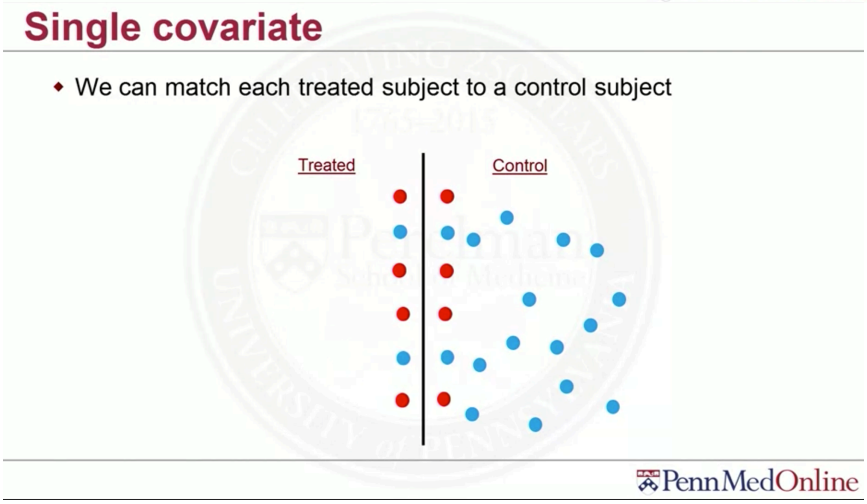
> Suppose older people are more likely to get A=1, meaning that at younger ages, there're more people with A=0, at older ages, there're more people with A=1. Age is the covariate.
>
> In a randomized trial, at any level of covariate (age), there should be about the same number of treated(A=1) and untreated(A=0) people.
>
> In observational studies, by matching treated people to control people of the same age, we make observational study more like a randomized trial.

It is done at the design phase so that's blinded to the oucomes.

Extact matching

An demonstration of matching on a single covariate:





Fine balance

When we can not find great matches, or have too many covariates, we accept non-ideal matches where distribution of covariates is balanced across groups.

Number of matches

- one-to-one
- Many-to-one: match k controls to evey treated subject
- variables: [1,k] depends on whether there're good matches available

How (general matching guidance)

Distance Metric

1. Match directly on confounders

There are two options for the closeness metrics: Mahalanobis distance (M distance), robust Mahalanobis distance (robust M distance).

- M distance: $D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$, where S is the covariance matrix. So M distance is the square root of the sum of squared distances between each covariate scaled by the convariance matrix.
- robust M distance: to elimiate outliers' impacts, we could replace each covariate value with its rank, constant diagonal on covariance matrix, and calculate the usual M distance

Matches Selection

1. Greedy

One-to-one greedy matching steps:

> 1. randomly order treated subjects, control subjects
> 2. start with the first treated subjects, match to the control with the smallest distance (greedy)
> 3. **remove the matched control from the available matches**
> 4. repeated step 2 and 3

For many(control)-to-one matching steps:

> For k:1 matching: after everyone has 1 match, go through the list again and find 2nd matachs.
>
> Repeat until k matches.

Greedy matching is computationally faster, but not optimal and can lead to some bad matches. And it is variant to the initial order of the (treated) list.

Caliper

It might be helpful to have 'caliper' - a maximum acceptable distance. We discard the treated subject if the best matches exceeds caliper.

Caliper also helps us examine the positivity assumption: a treated subject really didn't have much of a chance of being a control subject.

The *drawback* of caliper is that our matching might lose interpretibility: 'all treated subjectd excep those for which there is no match available'.

2. Optimal

On the contrary, optimal matching aims at minimizing the global distance measure (e.g. total distance). But it is **computationally demanding**.

Whether or not to perform optimal matching depends on the size of the problem: 1 million treatment-control pairings is feasible on most computers.

- 100 treated and 1,000 controls results in 100,000 possible pairings.

One way to make optimal matching more feasible is to **impose contraints**.

- example: imagine we might want to match within hospitals in a multi-site clinical study. We have a study that has several hospitals and then patients within hospital. Rather than do an optimal matching on the entire dataset all at once, you could do optimal matching within hospitals. So the idea would be that you only accept matched pairs, so matched treated to control pairs within a given hospital.

Accessing balance of matching

To access balance is to **determine whether matching was successful**. One way is to check covariate balance.

standardized difference

- Difference in means of covariates between groups, divided by the pooled standard deviation.
-

$$smd = \frac{\overline{X}_{treatment} - \overline{X}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

- Rules of thumb:
  - $smd < 0.1$: adequate balance
  - 0.1 ~ 0.2: not too alarming
  - $smd > 0.2$: imbalanced

After matching: analyzing data

1.Randomization tests (permutation tests)

- To examine null hypothesis by randomly permuting labels. If there's no actual treatment effect, it should not differ.
  - see wiki[[https://en.wikipedia.org/wiki/Resampling_(statistics)#:~:text=A%20permutation%20test%20(also%20called,under%20all%20possible%20rearrangements%20of](https://en.wikipedia.org/wiki/Resampling_(statistics)#:~:text=A) permutation test (also called,under all possible rearrangements of)]
- works for continuous data as well

2. McNemar test for binary data

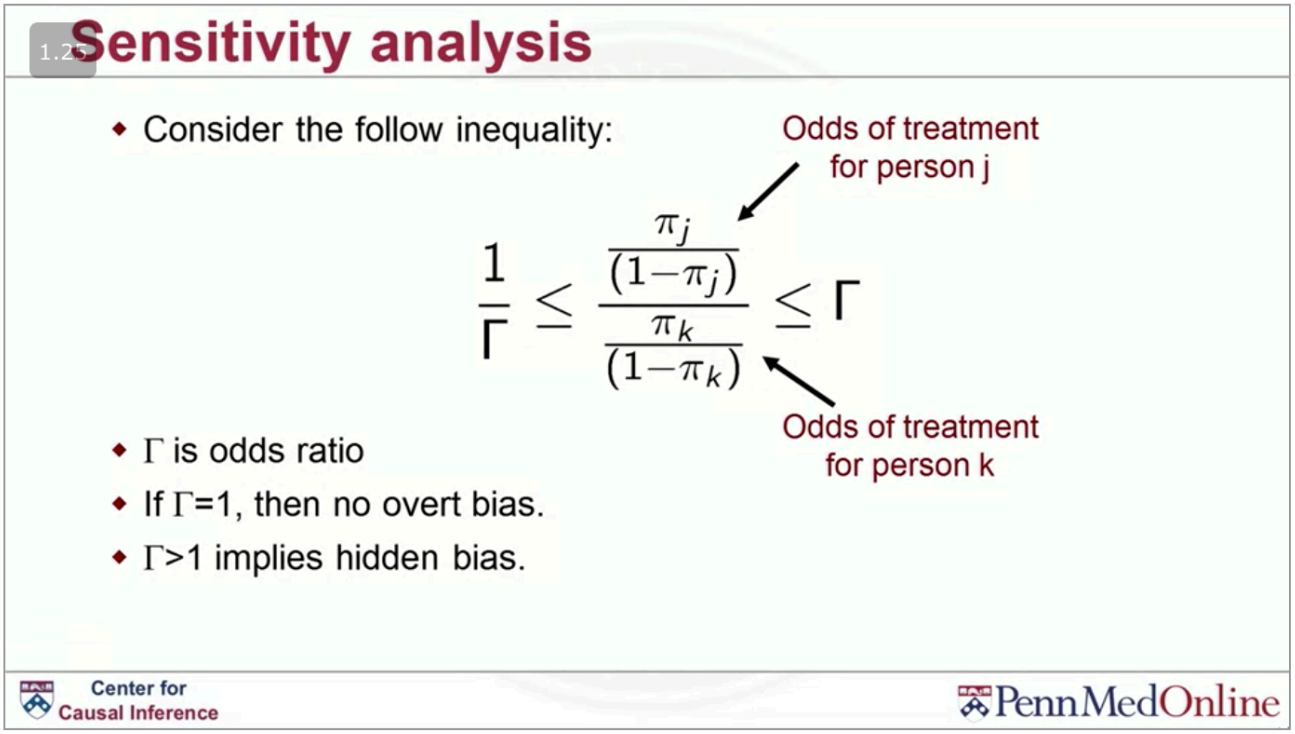
3. paired t test for continous data

Sensitivity Analysis

The goal of sensitivity analysis is to determine whether there is possible **hidden bias** (unmeasured confounders). Matching cannot guarantee balance on variables that we did not match on. (while randomized trials can).

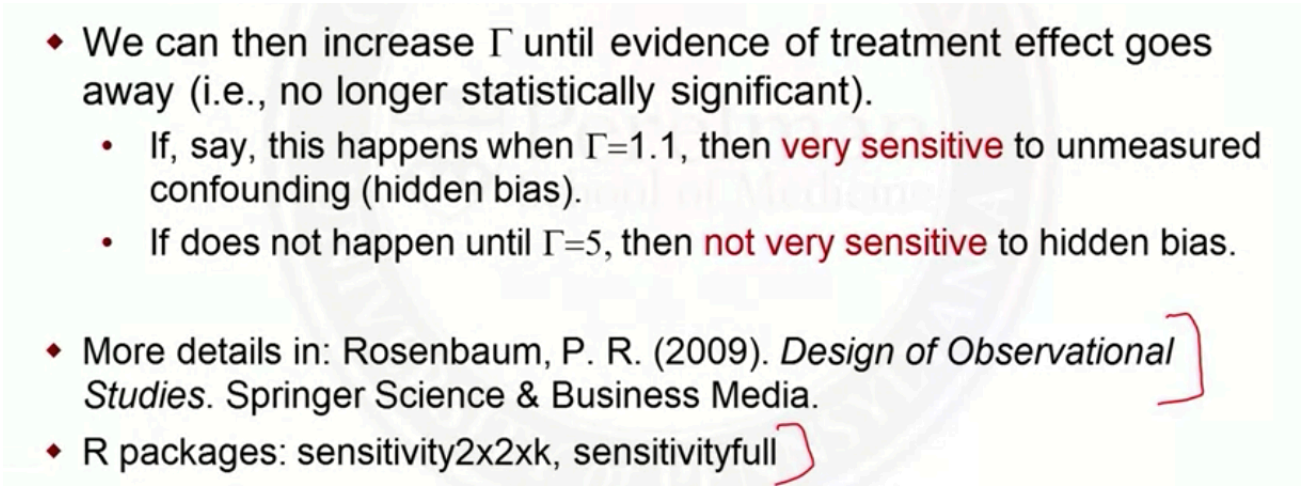
If there's hidden bias, how severe it would have to be to change conclusions? (staitstically significant, or effect direction)

> 1. let π_j π_k be the prob that person j and k received treatment
> 2. suppose person j and k perfectly matched and their covariate X_j X_k are the same
> 3. $\pi_j = \pi_k$ means no hidden bias



Γ quantifies how much 'no bias' assumption is violated. We could assume $\Gamma = 1$.

How sensitive are conclusion is to hidden bias?



How (Propensity Score Matching)

Propensity Score

Probability of receiving treatment given covariate X:

$$\pi_i = P(A = 1|X_i)$$

Balancing Score

- Propensity Score is a balancing score.
- Despite the different X, 2 subjects with the same propensity score are **equally likely** to be treated.
    - $P(X = x|\pi(X) = p, A = 1) = P(X = x|\pi(X) = p, A = 0)$
        - the distributions of covariate X given p score are the same across treatment v.s. control, 'either type of X about as often in the treatment as in the controller'
- Implication: match on the propensity score should achieve balance

Estimated Propensity Score

In a randomized trial, the propensity score is generally know. e.g. in a coin flipping trial, P score is 0.5

In an observational study, when we are talking about 'P score', which is actually unknow, we are referring to **estimated** P score: $\pi_i = P(A = 1|X_i)$

Match

- Using P score to match simplified the problem
- Before matching, we compare P score distribution across treatment and control group and check their overlap. Good one overlap everywhere.

Trimming tails

- Removing subjects with extreme values of P score
    - remove control subjects with P score < min P score for treatment group
    - remove treatment subjects with P score > max P score for control group
- Prevents extrapolation

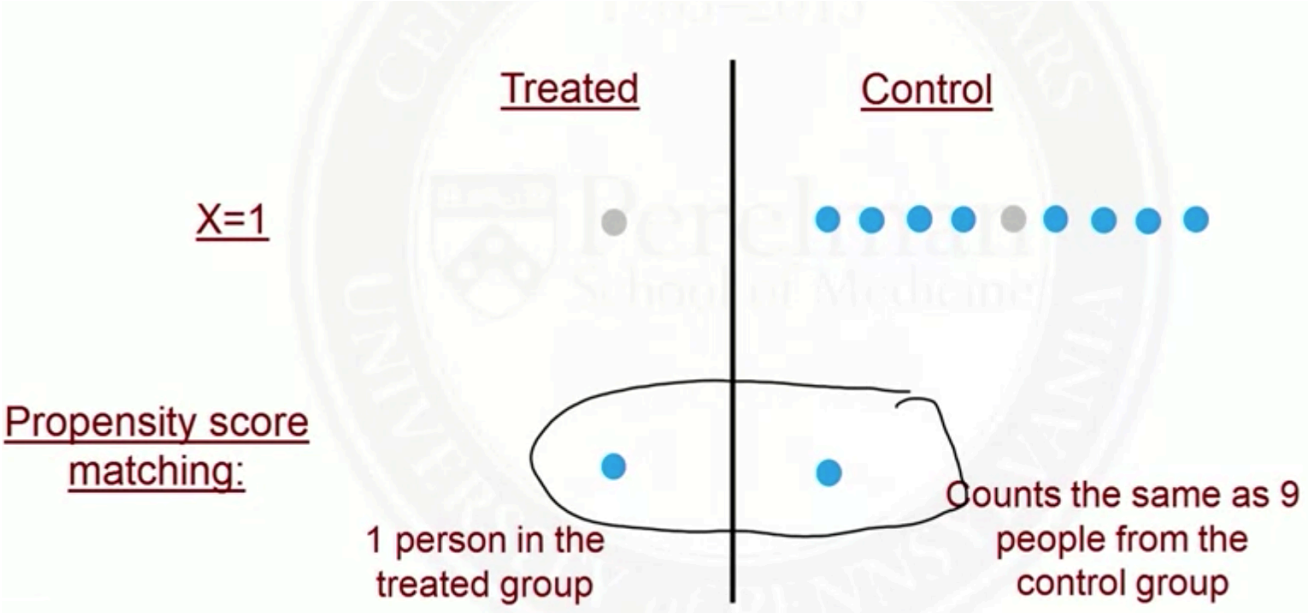
Match

- Pick a distance, do greedy or optimal matching as we discussed previsouly
- **Logit (log-odds) of P score** is often used because we want to stretch the original distribution (which is between 0 and 1), making the difference more obvious while preserving the rank
- Caliper: max acceptable distance
    - 0.2 * standard deviation of logit-transformed propensity score
    - Trade-off: maller caliper: more variance but less bias

III. IPTW (Inverse Probability of Treatment Weighting)

Motivation

Suppose $P(A = 1|X = 1) = 0.1$, meaning that on average, out of every 10 subjects with $X = 1$ there will be 1 treated. Whereas in a randomized trial, the probability should be 50%; in Propensity score one-to-one matching, the 1 person in the treated group represents 9 persons in the control group.



However, we want both groups contribute equal amount collectively. And we could achieve that by weighing by the inverse of probability of treatment received (treatment / control).

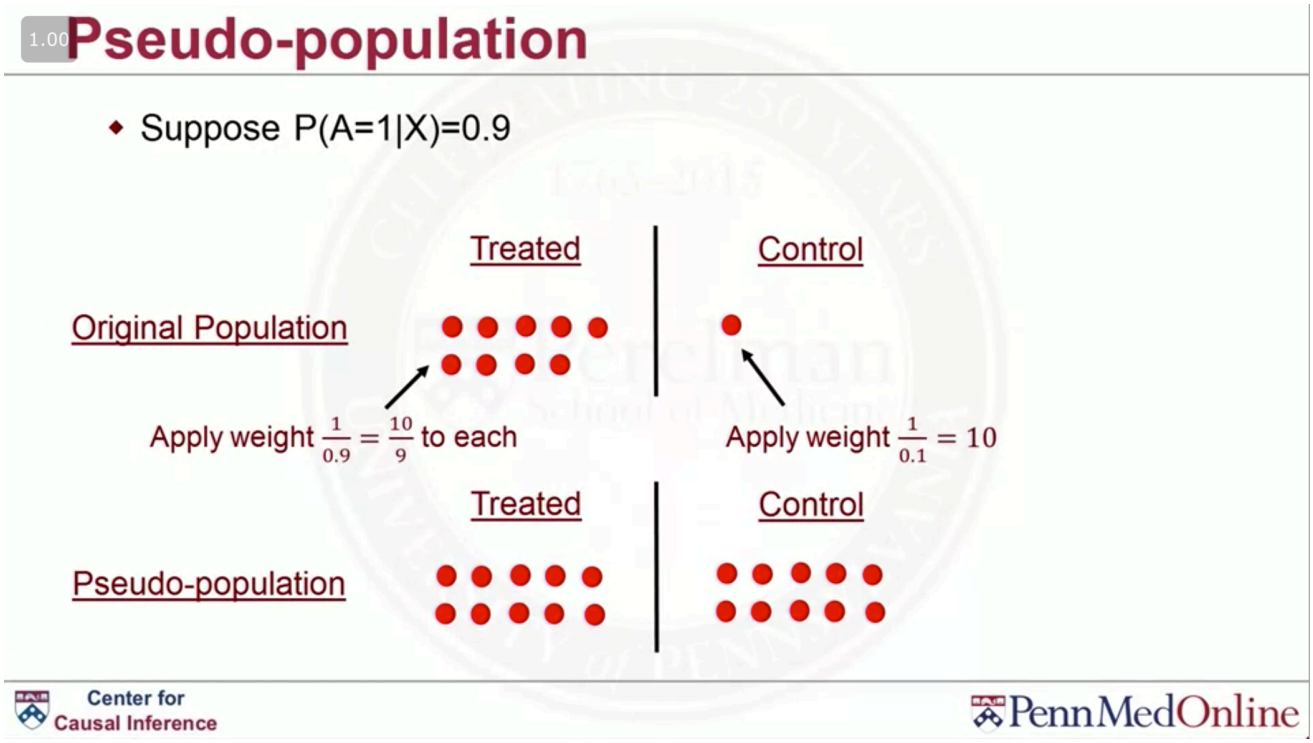
- For treated subjects, weight by $\frac{1}{P(A=1|X)}$
- For control subjects, weight by $\frac{1}{P(A=0|X)}$

This is called IPTW.

More intuition...

In a survey, sometimes we would oversample the minority group in order to get a large enough sample size of that group; and when estimating the population mean, we need to account for the oversample by weighting (e.g. Horvitz-Thompson estimator).

In an observed study, where you have confoundings, at various values of confounders, you will have *oversampling* of either the treated group or the control group. **IPTW creats a *Pseudo-population* where treatment assignment no longer depends on X.**



In a pseudo-population, everyone is equally likely to be treated, regardless of their covariate X.

What

IPTW

- For treated subjects, weight by $\frac{1}{P(A=1|X)}$
- For control subjects, weight by $\frac{1}{P(A=0|X)}$

IPTW Estimator

$E(Y^1)$: if everyone is treated, it is the hypothetical mean of the potential outcome.

But in a **pseudo-population**, we sum the Y's in treated pseudo-population (adjusted by (π_i)) and divide it by number of subjects in treated pseudo-population.

$$E(Y^1) = \frac{\sum_{i=1}^{n} I(A_i = 1)\frac{Y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{I(A_i=1)}{\pi_i}}$$

where π_i is the propensity score,

assuming exchangeability (ignorability, X fully captures all confoundeings) and positivity (propensity score being strictly between 0 and 1).

Marginal Structural Models

> a more interesting causal model where IPTW estimators can work

Definition

MSM is a model for the mean of the **potential outcomes**. 'Marginal' means 'population average' which is **not conditional on the confounders**; 'structural' means it is a model for potential outcomes not observed outcomes.

Linear MSM

$$E(Y^a) = \psi_o + \psi_1 a, a = 0, 1$$

$E(Y^0) = \psi_o$

$E(Y^1) = \psi_o + \psi_1$

So ψ_1 is the average causal effect $E(Y_1) - E(Y_0)$

Logistic MSM

$$logitE(Y^a) = \psi_o + \psi_1 a, a = 0, 1$$

where $E(Y^a)$ is the odds of $Y = a$

So $exp(\psi_1)$ is the causal odds ratio

$$exp(\psi_i) = \frac{\frac{P(Y^1=1)}{1-P(Y^1=1)}}{\frac{P(Y^0=1)}{1-P(Y^0=1)}}$$

MSM with effect modification

i.e. interaction terms

Usually we just want the marginal mean over the entire population, but there might be come small subset that we care about in terms of modification. Suppose V is a variable that modifies the effect of A.

A linear MSM with effect modification:

$$E(Y^a|V) = \psi_0 + \psi_1 a + \psi_2 V + \psi_3 aV, a = 0, 1$$

So $E(Y^1|V) - E(Y^0|V) = \psi_1 + \psi_3 V,$ the potential outcomes given the values of V.

General MSM

$$gE(Y^a|V) = h(a, V; \psi)$$

where $g()$ is a link function.

The right side looks similar to general parametric model, but the key issue is that potential outcomes are not observed outcome, and the left hand side of an MSM is not observed data, but rather, involves with potential outcomes (Y^a).

So how do we estimate potential outcomes?

How

General linear regression v.s. MSM

Estimation in MSM: we **set** a to a certain value and estimate potential outcome:

$$E(Y^a) = g^{-1}(\psi_o + \psi_1 a)$$

v.s. the regression model, we **condition** on A_i-- restricting to the subgroup with A_i:

$$E(Y_i|A_i) = g^{-1}(\psi_o + \psi_1 A_i)$$

A randomized trial does not have confounding and we could use regressio model.

For observed studies, recall that pseudo-population (obtained from IPTW) is free from confounding. Therefore, we can estimate MSM params by solving estimating equations for the observed data of pseudo-population.

$$\sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \psi} V_i^{-1} W_i \{Y_i - \mu_i(\psi)\} = 0$$

Where $W_i = \frac{1}{A_i P(A=1|X_i) + (1-A_i) P(A=0|X_i)}$, which is the inverse of propensity score for treated subjects, and inverse of probability of getting control for control subjects.

MSMs estimation steps

1. estimate propensity score
2. Create weights $W_i = \frac{1}{A_i P(A=1|X_i) + (1-A_i) P(A=0|X_i)}$
    1. the inverse of propensity score for treated subjects
    2. the inverse of probability of getting control for control subjects.

3. Specitfy the MSM of interests (e.g. Linear, Poisson)
4. Use software to fit a weighted generalized linear model (if you pick a linear MSM in the 3rd step)
5. Use asymptotic (sandwich) variance estimator (or boostrapping)
    1. to account for fact the pseudo-population might be larger than sample size

Assessing balance

> Check covariate balance on weighted sample using standardized differences.

Recall the standardized difference:

$$smd = |\frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}|$$

On our pseudo-population, we instead use **weighted means** and **weighted variance**. And then follow the instructions we have for accessing balance.

If imbalanced...

- Refine propensity score model
- Non-linearity
- ANd re-access balance

Large Weights

Issues

Large weights could create a problem(i.e. a large standard error). It could lead to noisier est of causal effects. e.g. imagine a person with a weight of 10,000, this person would have a big impact on the est.

B.T.W Bootstrapping is a way to estimate standard errors.

And an extremely large weight - the inverse of probability of receiving treatment- indicate a near violation of positivity assumption.

Remedies

Investigation steps:

1. look into why:
    1. unusual situation?
    2. problem with data?

    3. problem with the propensity score model?
2. Trimming: remove the extremes following the rules we have for propensity score tail trimming
    1. remove control subjects with P score < min P score for treatment group
    2. remove treatment subjects with P score > max P score for control group

Doubly Robust Estimators (Augmented IPTW)

Propensity score model v.s. outcome regression model

Propensity score model: $E(Y^1) = \frac{1}{n}\sum \frac{A_i Y_i}{\pi_i(X_i)}$ where $\pi_i(X_i)$ is the propensity score because we only look at treated subjects. **If the p score is correctly specified, this est is unbiased**. And you can think of it as a weighted mean of a pseudo-population.

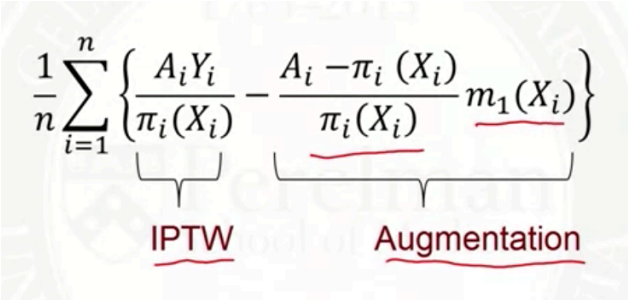
Outcome model: $m_1(X) = E(Y|A=1, X)$, an outcome model **restricted to treated subjects**. So it looks like a regression model.

And the $E(Y^1) = \frac{1}{n}\sum_{i=1}^{n} A_i Y_i + (1-A_i) m_i(X_i)$

- For subjects with A=1, use observed Y
- For control subjects, use **predicted value** of Y given their X's and if their had been 1 (the outcome model was fit using treated subjects only)

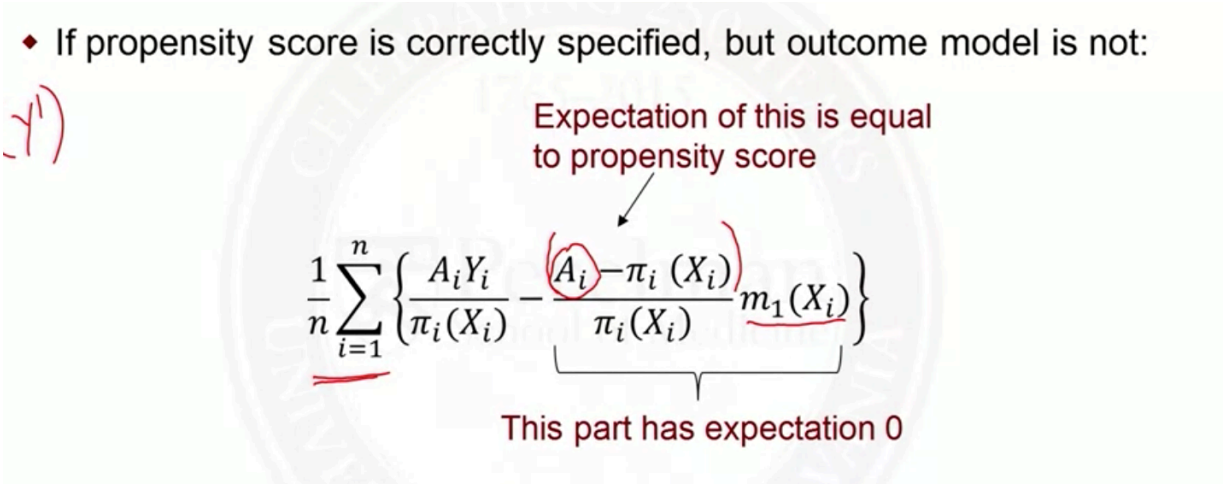
Doubly Robust Estimators

> An estimator that would be unbiased if either this propensity score model is correct or the outcome regression model is correct

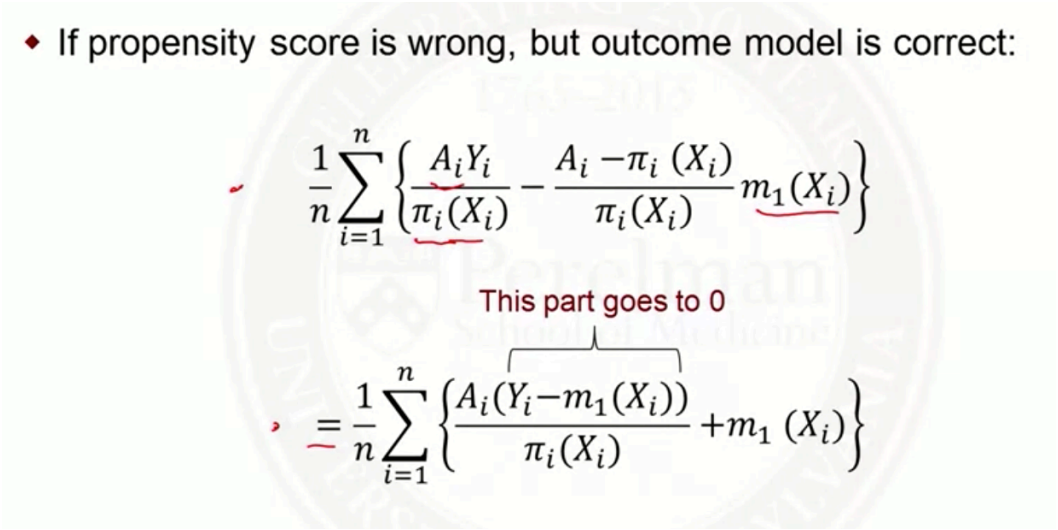


Intuition

- If propensity score is correctly specified, but outcome model is not:



- If propensity score is wrong, but outcome model is correct:



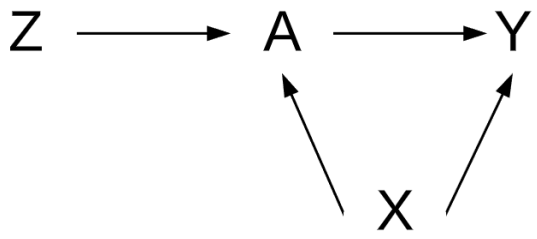
IV. Instrumental Variables Methods

Motivation

There will be case when we are sure that there's unmeasured confounding, and these sort of standard methods that control for the confounding are not going to be good enough. We would need Instrumental Variables, which is an alternative causal inference method that does not rely on ignorability assumption.

What

Suppose the DAG looks like this:



Here Z is an IV, affecting treatment A but not directly on Y. We could think of Z as an '**encouragement**' of treatment.

- An example would be: A-smoking during pregnancy, Y: birthweight, X: mother's age, weight, etc; Z: randomize to either receive encouragement to STOP smoking ($Z = 1$) or receive usual care ($Z = 0$)

So **causal effect of encouragement** is measured by:

$$E(Y^{z=1}) - E(Y^{z=0})$$

The focus of IV methods is to inder the causal effect of treatment itself.

IV would be assigned or randomized in nature (e.g. quarter of birth, distance to hospital, etc.)

How

Randomized trials with noncompliance

Design

Important definition 🌟

- **Z: ramdomization to treatment (1,0)** - we hope people with Z=1 would actually receive treatment, but that is no guarantee
- **A: treatment received (1,0)**
- Y: outcome

Let's denote potential values of treatment:

- $A^{Z=1} = A^1$: value of treatment if randomized to Z=1
- $A^{Z=0} = A^0$: value of treatment if randomized to Z=1
- A^1, A^0 Could either be 0 or 1

So causal effect of treatment assignment (Z) on treatment received (A):

$$E(A^1 - A^0)$$

What could be estimated from the observed data: $E(A^1) = E(A|Z = 1)$, $E(A^0) = E(A|Z = 0)$

causal effect of treatment assignment (Z) on outcome (Y):

$$E(Y^{Z=1}) - E(Y^{Z=0})$$

What could be estimated from the observed data: $E(Y^{Z=1}) = E(Y|Z = 1)$,
$E(Y^{Z=1}) = E(A|Z = 0)$

What about causal effect of treatment received (A) on the outcome (Y)?

Compliance classes

Potential values of treatment (A)

A^0	A^1	Label
0	0	Never-takers
0	1	Compliers
1	0	Defiers
1	1	Always-takers

Compliance classes are those with $A^0 = 0, A^1 = 1$.

#####

A motivation for using IV in general is concerns about possible unmeasured confounding, which prevents us from marginalizing over all confounders(via matching, IPTW). And IV methods instead, focus on a ***local*** *average treatment effect.*

Local average treament effect

If we are inferring on the compliers subpopulation:

$E(Y^{Z=1}|A^0 = 0, A^1 = 1)p - E(Y^{Z=0}|A^0 = 0, A^1 = 1)$

$= E(Y^{Z=1} - Y^{Z=0}|compliers)$

$= E(Y^{a=1} - Y^{a=0}|compliers)$ (because they are the compliers!)

Known as **complier average causal effect (CACE)**

Observed Data - Identification Challenge

How do we know who the compliers are given that our observed data only have Z=0 OR Z=1. Compliance classes are also known as principal strata. They are latent.

Z	A	A0	A1	Class
0	0	0	?	Never-takers or compliers
0	1	1	?	Always-takers or defiers
1	0	?	0	Never-takers or defiers
1	1	?	1	Always-takers or compliers

How can we estimate CACE from observed data? What assumptions do we need?

Assumptions

Exclusion restriction

A variabe is an IV only:

1. If it is associated with the outcome (could be checked via data)
2. IV affects the outcome **only through treatment** (**exclusion assumption**) (rely on subject matter knowledge)

Monotonicity assumption

> The assumption says: there are **no defiers**.

It is called monotonicity because the assumption is that the probability of treatment should increase with more encouragement.

With monotonicity, we solved the identification challenge:

Z	A	A^0	A^1	Class
0	0	0	?	Never-takers or compliers
0	1	1	1	Always-takers ~~or defiers~~
1	0	0	0	Never-takers ~~or defiers~~
1	1	?	1	Always-takers or compliers

Causal effect identification and estimation

$$E(Y|Z=1)$$

$$= E(Y|Z=1, alwaystakers) * P(alwaystakers)$$

$$+E(Y|Z=1) = E(Y|Z=1, nevertakers) * P(nevertakers)$$

$$+E(Y|Z=1) = E(Y|Z=1, compliers) * P(compliers)$$

And among always takers and never takers- Z does not do anything.

$$E(Y|Z=1)$$

$$= E(Y|always-takers) * P(always-takers)$$

$$+E(Y|never-takers) * P(never-takers)$$

$$+E(Y|Z=1, compliers) * P(compliers)$$

Same for Z=0

$$E(Y|Z=0)$$

$$= E(Y|always-takers) * P(always-takers)$$

$$+E(Y|never-takers) * P(never-takers)$$

$$+E(Y|Z=0, compliers) * P(compliers)$$

And we founf the E for always-takers and never-takers are the same.

So:

$$E(Y|Z=1) - E(Y|Z=0)$$

$$= E(Y|Z=1, compliers) * P(compliers)$$

$$-E(Y|Z=0, compliers) * P(compliers)$$

which implies:

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{P(compliers)}$$

$$= E(Y|Z=1, compliers) - E(Y|Z=0, compliers)$$

$$= E(Y^{a=1}|compliers) - E(Y^{a=0}|compliers)$$

$$= CACE$$

And note that $P(compliers) = E(A = 1|Z = 1) - E(A = 1|Z = 0)$, because $E(A|Z = 1)$ is the proportion of always-takers or compliers; E(A|Z=0) is the proportion of always-takers.

$$CACE = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(A|Z = 1) - E(A|Z = 0)}$$

- If perfect compliance ($P(compliance) = 1$): $CACE = ITT$
- But the denominator always between 0 and 1, so $CECE >= ITT$

Two Stage Least Squares (2SLS)

Illustration

A method for estimating a causal effect in the IV setting.

Assuming Z is a valid IV

Stage 1: $A_i = \alpha_0 + Z_i\alpha_1 + \epsilon_i$

We obtain $\hat{A}_i = \hat{\alpha_0} + Z_i\hat{\alpha_1}$, this is the predicted A_i for a given Z_i

$\hat{A}_i = E(A|Z)$

Stage 2: with $\hat{A}_i$ from stage 1, $Y_i = \beta_0 + \hat{A}_i\beta_1 + \epsilon_i$

- $\hat{A}_i$ is the projection of A onto space of Z
- $\beta_1 = E(Y|\hat{A} = 1) - E(Y|\hat{A} = 0)$

β_1 is the change of Y by one unit of $\hat{A}$ change. And A only has 2 values: 0,1.

From Z=0 to Z=1, we observe change in $\hat{A}$ by $\hat{\alpha_1}$.

From $\hat{A} = \hat{\alpha_o}$ to $\hat{A} = \hat{\alpha_o} + \hat{\alpha_1}$, we observe change in $\hat{Y}$ by $E(Y|Z = 1) - E(Y|Z = 0)$. So, 1 unit change in $\hat{A}$ leads to $\frac{E(Y|Z=1)-E(Y|Z=0)}{\hat{\alpha_1}}$ unit change in $\bar{Y}$.

- And from stage 1 we know that $\hat{\alpha_1} = E(A|Z = 1) - E(A|Z = 0)$
- So 1 unit change in $\hat{A}$ leads to $\frac{E(Y|Z=1)-E(Y|Z=0)}{E(A|Z=1)-E(A|Z=0)}$ unit change in $\bar{Y}$
- which is CACE
- So $\beta_1 = CACE$

The 2SLS estimator is a consistent estimator of the complier average causal effect.

2SLS more generally

It can also be used with convariates and for non-binary data/

Stage 1: regress A ~ Z, X and obtained fitted value of A

Stage 2: regress Y ~ $\hat{A}, Z$, and the coefficient of $\hat{A}$ is the causal effect.

Sensitivity Analysis

To examine the IV assumptions:

- exclusion restrition: if Z does directly affect Y by an amout ρ, would my conclusion change?
- Monotonicity: if the proportion of defiers was π would my conclusion change?

Weak Instruments

What if the instrument is weakly predictable of treatment?

Measure of strength of an instrument

- proportion of compliers: $E(A|Z=1) - E(A|Z=0)$

Suppose only 1% are compliers, then the causal effects are very unstable.

If the IV is weak, an IV analysis might not be the best option.

Strengthening IV

- **near/far matching**: match so that covariate are similiar but the instrument is very different