A Crash course in causality

• Confusion

- Spurious correlation (clearly unlinked things w/ high correlation)
- Anedocte: a man who lives 105y and believe the cause was eating one turnip a day
- Science Reporting: headlines don't use the form of the word cause, but do get interpreted causally: "Portive link between video game and academic performance", "Tennis reduce risk of death at any age"
- Reverse causality: Green space and exercise People that like to exercise they move closer to green spaces. But we would be interested in checking if creating more green space it'd cause more exercise.

- How to clear up confusion? (1)
- Formal definitions of causal effects
- Assumptions necessary to identify causal effects from data
- Rules about what variables need to be controlled for
- Sensitivity analyses to determine the impact of violations of assumptions on conclusions;

-> brief history at 14 minutes: Wright 1921, neyman 1923
- Rubin causal model 1974 (Potential outcomes)
- Causal diagrams (Robins 1986, Pearl 2000)
- Propensity scores (Rosenbaum and Rubin 1983)
- Time dependent confounding (Robins 1986, Robins 1997)
- Optimal dynamic treatment strategies (murphy 2003, Robins 2004).
- Targeted learning (van der Laan 2009)

* Causal inference requires making some untestable assumptions (causal assumptions). You can't validate it using the data you have.
↳ Knowledge from observational studies => going towards the truth.

-11-

- Potential Outcomes and counterfactuals

Treatments and outcomes

Treatment A and outcome Y

E.g. 1) A=1 if receive influenza vaccine; A=0 otherwise
2) A=1 if receive active drug; A=0 if placebo

Outcome e.g. 1) Y=1 if develop cardiovascular disease within 2 years; Y=0 otherwise
2) Y= time until death (continuous)

Potential outcomes

The outcomes we would see under each possible treatment option.

Y^a is the outcome that would be observed if treatment was set to A=a.

Classically: Y^0 and Y^1 : untreated and treated

Counterfactuals

Sometimes used interchangeably w/ potential outcome, though they are slightly different

The ones that would have been observed if had the treatment been different

I.e. the vaccine prevent me from getting the flu?

-> I got the vaccine and did not get sick
-> my actual exposure was A=1
-> my observed outcome was $Y = Y^1$

↳ What would have happened (contrary fact):
-> Have I not gotten the vaccine, would I've gotten sick?
-> my counterfactual exposure is A=0
-> my counterfactual outcome is Y^0

A crash course in Causality

Before treatment decision is made, any outcome is a potential outcome: Y^0 and Y^1.

After the study, there is an observed outcome, $Y = Y^A$, and counterfactual outcomes Y^{1-A}

↳ Anyway, it's used interchangeably

- Hypothetical Interventions

Causal effects of interventions and actions

It's common to assume there's no hidden version of the treatment

Take care when you can't change something directly or if you have more than one way of changing it, as the many ways you have to change may affect the ~~treatment~~ outcome in different ways. E.g losing weight: exercise, diet change?

Immutable variables It's also less clear what a causal effect of an immutable variable would mean. Causal effect of race, age, gender?

We can't manipulate them directly, but we can do interventions like changing name on resume, bariatric surgery, gift or money for imputing race, obesity and social economic status

Good cases: intervention we think can be randomized in a hypothetical trial

What are causal effect? In general: A head a causal effect on Y if Y^1 differs from Y^0

I can't say that something cause an effect if I don't know the counterfactual.

fundamental problem of causal effect / inference is that we can only observe one potential outcome for each person.

However, w/ certain assumptions, we can estimate population level (average) causal effects

↳ hopeless: what would have happened to me had I not taken ibuprofen? (unit level causal effect)

possible: What would the rate of headache remission be if everyone took ibuprofen when they had a headache versus if no one did?

A Crash course in Causality

• Causal Effect

Hypothetical worlds - Average Causal Effect

World
Population of Interest

World I
Everyone gets A=0

mean(Y)

World II
Everyone gets A=1

mean(Y)

Difference is average causal effect

We want $E[Y^1 - Y^0]$, if Y is binary it's a risk/probability

Suppose $E(Y^1 - Y^0) = -0.1$, no risk of outcome (3) is 10% lower for treatment (A=1).

If Y is continuous, it's fine

Conditioning on, versus setting, treatment

In general, $E(Y^1 - Y^0) \neq E(Y|A=1) - E(Y|A=0)$

Why? $E(Y|A=1)$ reads as "expected value of Y given A=1". This is restricting to the subpopulation of people who actually had A=1

- they might differ from the whole population in important ways;
- for example, people at higher risk for flu might be more likely to choose to get a flu shot

Pop of Int A = 0 A = 1

We can't do the mean(Y) diff, because we couldn't isolate the treatment effect.

- $E(Y|A=1)$: mean of Y among people w/ A=1
- $E(Y^1)$: mean of Y if the whole pop was treated A=1
- $E(Y|A=1) - E(Y|A=0)$ is generally not a causal effect, because it's comparing two diffent populations of people.
- $E(Y^1 - Y^0)$ is a causal effect, because it is comparing what would have happen if the same people were treated with A=1 versus if the same people were treated with A=0.

Other causal effects

- $E(Y^1/Y^0)$: Causal relative risk
- $E(Y^1 - Y^0 | A=1)$: Causal effect of treatment on the treated.

might be interested in how well treatment works among treated people.

- $E(Y^1 - Y^0 | V=v)$: average causal effect in the subpopulation w/ covariate V=v.

Challenge

We only observe one treatment and one outcome for each person: the fundamental problem.

How do we use observed data to link observed outcomes to potential outcomes?

What assumptions are necessary to estimate causal effects from observed data?

A Crash Course in Causality

- Causal Assumptions

- Identifiability

Identifiability of causal effects requires making some untestable assumptions. These are generally called causal assumptions.

The most common are:

- Stable unit treatment Value Assumption (SUTVA)
- Consistency
- Ignorability
- Positivity

They are all about the observed data: Y, A, and a set of pre-treatment covariates X

SUTVA

No interference:

- units do not interfere w/ each other;
- Treatment assignment of one unit does not affect that outcome of another unit;
- Spillover or contagion are also terms of interference

One version of treatment;

=> SUTVA allows us to write potential outcome for the i^{th} person in terms of only that person's treatments

Consistency

The potential outcome under treatment $A=a$, Y^a, is equal to the observed outcome if the actual treatment received is $A=a$.

$$Y = Y^a \text{ if } A=a, \text{ for all } a$$

Ignorability (Very important!!!)

Given pre-treatment covariates X, treatment assignments is independent from the potential outcome.

$$Y^0, Y^1 \perp\!\!\!\perp A \mid X$$

Also refered as the "no unmeasured confounders' assumption.

Among people w/ the same values of X, we can think of treatment A as being randomly assigned.

Ex: X = age - young, old

Old are more likely to treat A=1, also to have the outcome Y (big fracture)

So Y^1 and Y^0 are not independent of A. However, within levels of X, treatment might be randomly assigned.

Positivity

For every set of values for X, treatment assingment was not deterministic:

$$P(A=a \mid X=x) > 0 \text{ for all } X$$

We need data for the counterfactual, to learn what would have happen. If all x get A=a we would not have X data for A=b.

If so, exclude this X population where A is deterministic.

A crash course in Causality

• Causal Assumptions

We can check these assumptions on observed data: $E(Y \mid A=a, X=x)$: where only observed data.

$E(Y \mid A=a, X=x) = E(Y^a \mid A=a, X=x)$ by consistency

$= E(Y^a \mid X=x)$ by ignorability

If we want a marginal causal effect, we can average over X. ↳ not conditioned on X

• Stratification

Or standardization

Under certain causal assumptions:

$$E[Y^a] = E[Y \mid A=a, X=x]$$

$$E[Y^a] = \sum_x E[Y \mid A=a, X=x] P(X=x)$$

-> Average Potential Outcome : standardize mean

It's basically stratifying and then averaging! Obtain the treatment effect within each stratum then pool across strata, weighting by the probability of each stratum (it's ryz).

The effect is just the mean of Y within the strata for treated cases

=> Problems : Typically, there will be many X variables needed to achieve ignorability.

So it's common to have no data for a particular combination of X set values

(for all ages I'd need all blood pressures)

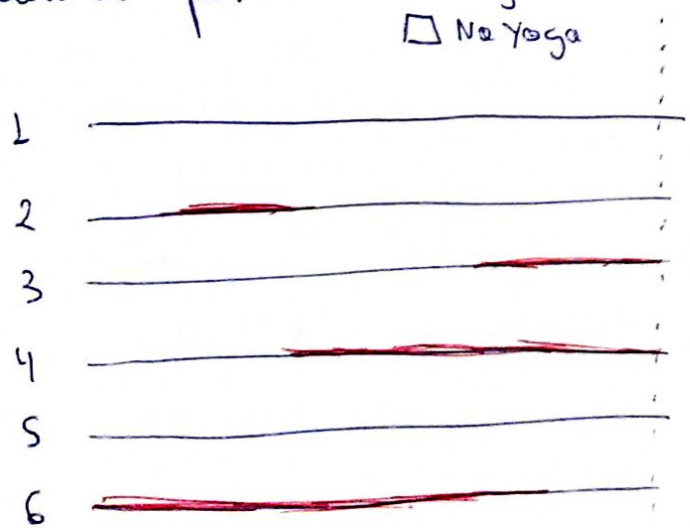
↳ We need alternatives!

• Incident user and active comparator design

A cross-sectional look (a snapshot) at treatments will reveal a spectrum of the treatment. Using the yoga example:

- Long time practioners, beginners
- Those who quit because it wasn't work
- Used to do in the past
- ...

↳ A lot of selection bias, it's hard to control for



Incident user design: restrict the treated population to those newly assigned, initiating treatment (also known as new user design)

is cleaner problem

Even if they quit, we're just measuring the effect of initiating

When measuring against no-treatment, it's not clear when one should initiate the follow-up.

Active Comparator: tackles this problem by comparing treatment with an alternative, like Zumba dancing for Yoga. usually, involves less confounding: ppl that practices zumba are more ~~likely~~ alike ~~them~~ than who do not exercise at all, but the causal question is also more narrow.

A crash course in causality

Incident and active comparator.

- Sometimes it's impossible to set an incident user design;
- maybe you want the no-treatment comparison
- there are alternatives to handle the time-varying treatments.

A crash course in causality

• Confounding

We're interested in the relationship between means of two different outcomes $E[Y^1 - Y^0]$

Confounding relates w/ ignorability,

$$Y^1, Y^0 \perp\!\!\!\perp A \mid X$$

So a confounder variable affects both outcome and treatment probability

If I decide treatment w/ a dumby not related w/ the outcome var, like flipping a coin, there's no problem.

If X impacts the outcome, but the treatment A is not assigned following it (family career history)

If X is age and older people are at a higher risk band for a disease and also are more likely to receive treatment, then age is a confounder.

Confounder control:

1. Identify a set of X variables that will make the ignorability assumption hold.

If we do this then set of variables is sufficient to control for confounding

2. Using statistical methods to control for these variables and estimate causal effect

Causal Graph

Which variables to control so we avoid confounding?

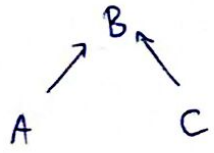
Causal graphs helps us to make our assumptions clear and makes it possible to identify the minimum set of variables able to deconfound

the treatment and effect ones, we want to d-separate them so we can achieve ignorability, $Y^0, Y^1 \perp\!\!\!\perp A \mid X$.

So d-separation is about the minimum set of variables that can make two of them independent from it when conditioned on this set.

$A \to B \to C$: Chain, $C \perp\!\!\!\perp A \mid B$

: Fork, $C \perp\!\!\!\perp A \mid B$



: Collider, $C \not\perp\!\!\!\perp A$

$C \not\perp\!\!\!\perp A \mid B$

Works if a sufficient set to control for confounding exists and if we correctly identify all observed causes of A and Y.

Controlling is all about blocking information leakage.

Front door path is the causal path between treat and outcome

Back door path is the not causal path between two variables

What to control?

Back door criteria block all the back door path between A and Y. You need a well specified causal graph to do it though. It's enough and guaranteed the min set.

Control for all the pre-treatment variables may work sometimes, but it's brute force and with a lot of bad cases.

Disjunctive causal criteria control for all (observed) causes of exposure, the outcome, or both. Don't need the whole causal graph, but you may control unnecessarily

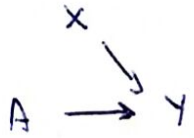
A Crash course in causality

- Matching

Obs

Vs

RCT



If a coin flip define A, we would have covariate balance:

$$P(X \mid A=0) \sim P(X \mid A=1)$$

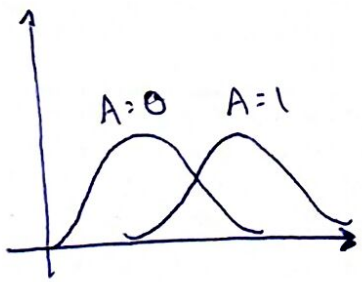
→ Why not always RCT?

- Expensive
- Unethical
- RCT takes time (wait to see outcome)
- Refuse to join it

⇒ Observational

- Pros: - Observing real world environment
  - Broader population eligible for study
  - Large sample size, inexpensive, rapid analysis
- Cons: - Regulations much weaker
  - measures without a pattern
  - Data quality is lower, no uniform standard of collection

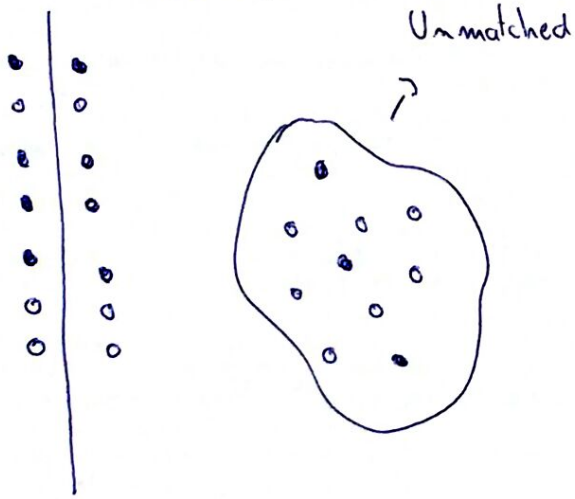
↳ How to solve? ⇒ Matching!



make Obs look like RCT!

Controlling by confounders at the design phase (blind to the outcome). After matching do the analysis as it is RCT.

matching will reveal a lack of overlap in covariates distribution, you may have $P(A=1 \mid Age < 50) = 0$. You can exclude them.



If you start with the treated population and you match w/ the control so it's skewed to the distribution in the treated population, then what you find is the average treatment effect on the treated

But there are techniques to target the population.

Fine balance you may relax similarity to make T / U features distribution to be more balanced

The number of matches may vary: 1-to-1, 1-to-K.

matching directly on confounders

We can use diffent distance metrics to find out the most similar untreated to match a treated example

Malahanobis Distance: $D(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$

S: covariance matrix to scale the distance.

Robust a Malahanobis Distance replace each covariate by its rank.

- A crash course in causality

- Greedy - NN matching

match every treated to the untreated w/ the minimum distance

Randomly order treated and start to match everyone - If 1-to-K, do it by rounds.

It's greedy because it does not guarantee total distance minimization. We can do it using optimal min distance algorithms

- Assessing Balance

After match we need to check how well we did in our task of making covariate distribution balanced between T / U.

We can do a t-test to the difference of each covariate mean between T / U

An alternative is to do a Standardize difference analysis. It's the difference in means between groups, divided by the (pooled) standard deviation

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

- Does not depend on sample size
- Often absolute value of smd is reported
- Calculate for each variable you match on
- Rules of thumb:

  < 0.1 adequate balance

  $0.1 - 0.2$ are not too alarming

  > 0.2 indicate serious imbalance

Analyzing data after matching

So smg analysis says our data is reliably balance, now we proceed with outcome analysis:

- Test for treatment effect
- Estimate a treatment effect and confidence interval
- methods should take matching into account

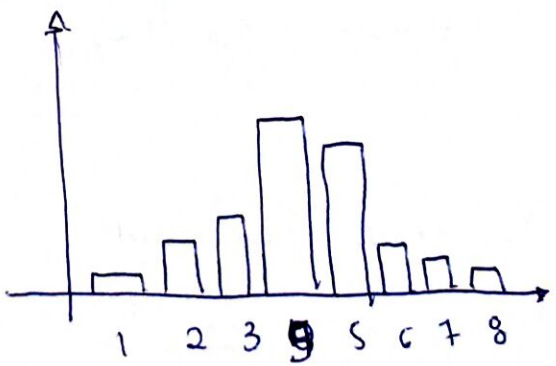
Randomization Tests / Permutation / Exact

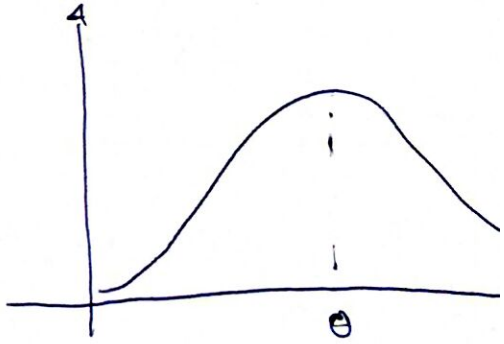
main idea:

- Compute test statistic from observed data
- Assume null hypothesis of no treatment effect is true
- Randomly permute treatment assignment within pairs and recompute test statistic
- Repeat many times and see how unusual observed statistic is.

This test is equivalent to the McNemar test for paired data.

You can count binary outcome or do difference in mean for continuous outcome





A crash course in causality

- Sensitivity Analysis

We want to find hidden bias, it's an uncontrolled confounder that would disregard ignorability

Overt bias: imbalance on observed variables

If there is hidden bias, determine how sensitive it would have to be to change conclusions

- change from statistically significant to not
- change in direction or effect

π_j: person j receives the treatment

π_K: " K " " "

Imagine j and K perfectly matched, the observed covariates x_j and x_K are the same.

if $\pi_K = \pi_j$ then there is no hidden bias.

$$\frac{1}{\Gamma} \leq \frac{\frac{\pi_j}{(1-\pi_j)}}{\frac{\pi_K}{(1-\pi_K)}} \leq \Gamma$$

Odds: probability divided by 1 minus the probability

Γ: Odds ratio

if $\Gamma = 1$, then no overt bias

$\Gamma >$ implies hidden bias

We can then increase Γ until evidence of treatment effect goes away. If it happens when $\Gamma = 1.1$, then very sensitive to unmeasured confounding (hidden bias). If does not happen until $\Gamma = 5$, then not very sensitive to hidden bias.

See Rosenbaum, P. R (2009). Design of Observational Studies.

- Propensity Score

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

Balance score: if we restrict to a subpopulation of subjects who have the same propensity score, there should be balance in the two treatment groups.

$$P(X=x \mid \pi(X)=p, A=1) = P(X=x \mid \pi(X)=p, A=0)$$

Conditioning on the propensity score is conditioning on an allocation probability

X distribution would be equal for A=0, 1

So it's another way to achieve balance

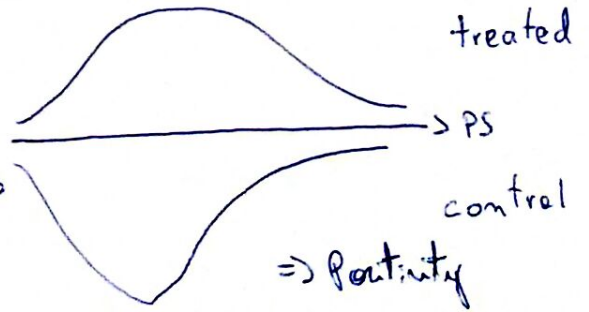
It's like being in a RCT world w/ chance of treatment being p.

In RCT it's known (usually 0.5). In a observational study it's unknown. But we can estimate it using A and X.

So propensity score would be an easier way for matching since we'd just to need to match in respect to it instead of doing it for all the covariates.

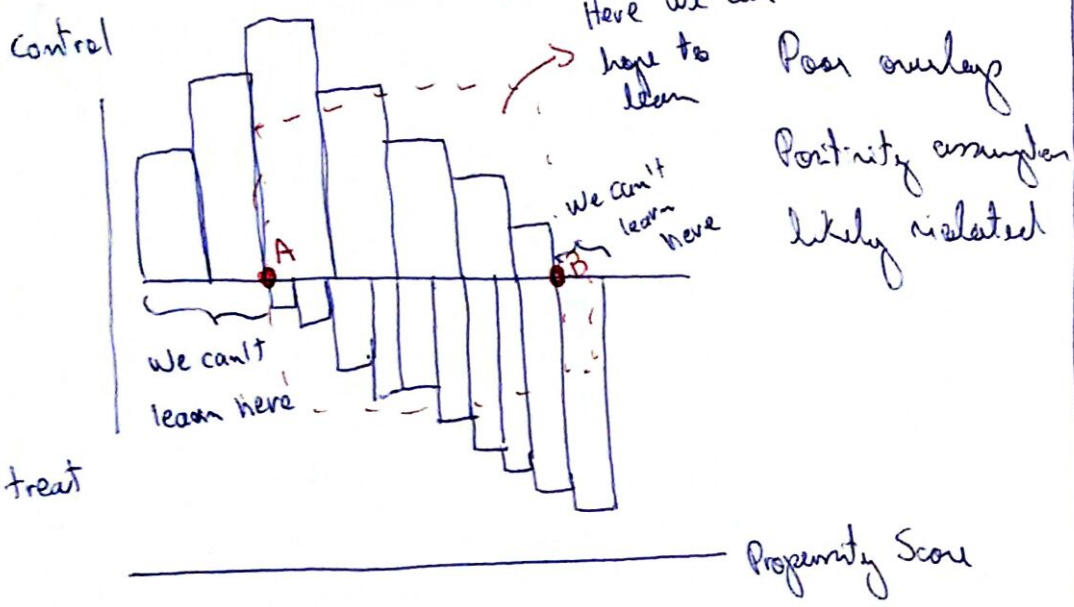
Overlap plotting PS for treated and untreated

we want to have pp' from both sides for all PS.



=> Positivity

A crash course in causality



Poor overlap

Positivity assumption likely violated

Trimming tails to fix the lack of overlap

Example: exclude control w/ PS less than the min for ~~control~~ treated group (A), or/and eliminate treated w/ PS greater than max in control

Matching

We can proceed by computing a distance between the propensity score for each treat subject w/ every control, then use NN or optimal like before

In practice, logit (log-odds) of the propensity score is often used, rather than the propensity score itself. match on logit(π) instead of π (unbounded, stretched distribution, rank preserved)

Common caliper $0.2 * \text{std}(\text{logit}(\pi))$

Smaller caliper: less bias, more variance

A Crash Course in Causality

- IPTW

Example

$P(A=1 \mid X=1) = 0.1$, X is the only confounder

So we would have one treated for 9 not treated in our sample.

Instead of matching one by one, we can weight them differently:

- For treated, weight by the inverse of $P(A=1 \mid X)$
- For untreated, weight by the inverse of $P(A=0 \mid X)$

This is known as inverse probability of treatment weighting (IPTW)

treated	Control
$X=1$ ●	●●●●●●●●● X
weight: $\frac{1}{P(A=1 \mid X=1)} = \frac{1}{0.1} = 10$	weight: $\frac{1}{P(A=0 \mid X=1)} = \frac{1}{0.9} = \frac{10}{9}$
$X=0$ ● ● ● ●	● $P(A=0 \mid X=0) = \frac{1}{0.2} = 5$
$P(A=1 \mid X=0) = \frac{1}{0.8} = \frac{5}{4}$	

Motivation: Survey

To estimate the population mean, can weight the data to account for the oversample.

=> Horvitz – Thompson estimator

Observational data

Certain groups are oversampled relative to the hypothetical sample from a randomized trial.

There's confounding in the original data (population);

- IPTW creates a pseudo-population where treatment assignment no longer depends on X

is no confounding in the pseudo pop.

Pseudo pop

Suppose $p(A=1 \mid X) = 0.9$

...

Estimator

Under the assumption of exchangeability and positivity, we can estimate $E(Y^1)$ as:

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

→ Sum of the Y^1's in the treated pseudo pop.

→ no of subjects in treated pseudo.

where $\pi_i = p(A=1 \mid X_i)$ is the propensity score.

- Marginal Structured model
- linear MSM

$$E[Y^a] = \Psi_0 + \Psi_1 a$$

$$E[Y^0] = \Psi_0 \qquad E[Y^1] = \Psi_0 + \Psi_1$$

$$E[Y^1] - E[Y^0] = \Psi_1$$

marginal = not conditioned on the confounders. Structured = ~~predicts~~ models potential outcomes and not observed outcomes

A Crash course in causality

- Logistic MSM for binary outcome

$$\text{logit}\{E(Y^a)\} = \psi_0 + \psi_1 a, \quad a = 0, 1$$

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

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

$\leftarrow$ Odds $Y^1 = 1$

$\leftarrow$ Odds $Y^0 = 1$

- MSM with effect modification (modified)

Suppose V can modify the effect of A

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

- General MSM

$$g\{\underset{Y^a|V}{E[Y^a]}\} = h(a, V; \psi),$$

g is a link function

h is a function specifying parametric form of a and V (typically additive, linear)

is not like a regression model, because it uses potential outcome and not observed

- Estimation

For generalized linear models

$$E(Y_i|X_i) = \mu_i = g^{-1}(X_i^T \beta),$$

estimation involves solving:

$$\sum_{i=1}^{n} \frac{d\mu_i^T}{d\beta} V_i^{-1} \{ Y_i - \mu_i(\beta) \} = 0 \quad \text{for } \beta$$

- MSM look like a GLM:

$$E(Y_i^a) = g^{-1}(\Psi_0 + \Psi_1 a),$$

this is NOT EQUIVALENT to the regression model:

$$E(Y_i | A_i) = g^{-1}(\Psi_0 + \Psi_1 A_i),$$ because of confounding. In a RCT they are the same.

However, the pseudo-population from IPTW is free from confounding (assuming ignorability and positivity)

we can therefore estimate MSM parameters by solving estimating equations for the observed data of the pseudo population

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

where $W_i = \dfrac{1}{A_i P(A=1|X_i) + (1-A_i) P(A=0|X_i)}$

A crash course in causality

• Assessing Balance

Try to measure the pseudo population quality.

Using standardized differences: the difference in means between groups, divided by the pooled standard deviation

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

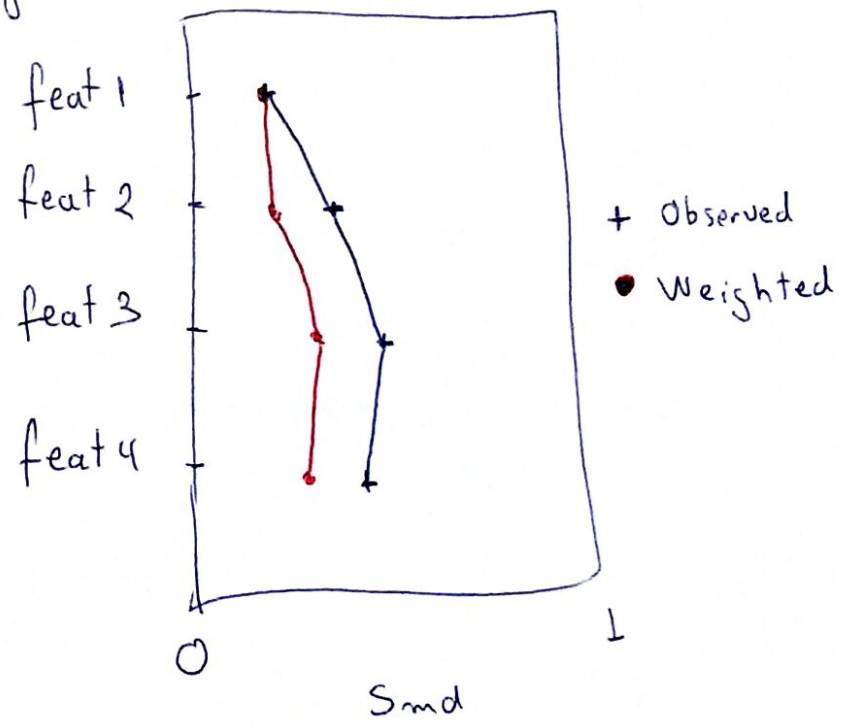
We are mostly interested in the magnitude instead of the direction of the differences.

When using weighted examples, use weighted means and weighted variances.

If imbalance after weighting:

Can refine propensity score model: interactions? non-linearity? Can then reassess balance

via plot



- Weights distributions

Large weights lead to noisier estimates of causal effects.

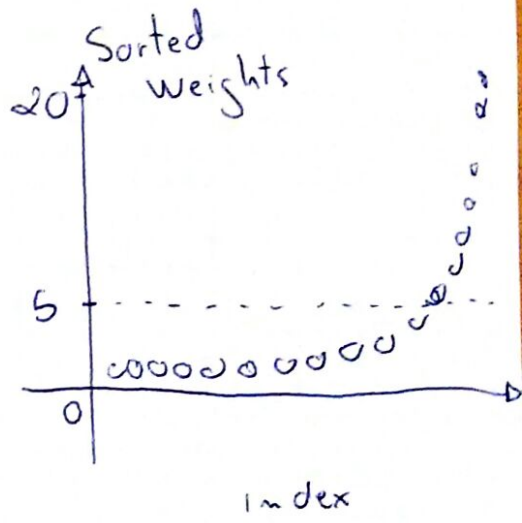
Bootstrapping: the variance will be high if we have individuals w/ high weights.

→ Relationship w/ positivity assumption:

- High weights indicates a very low probability of being treated. Thus, large weights indicate near violations of the positivity assumption.

So people w/ certain values for the covariates have just a small chance to be treated

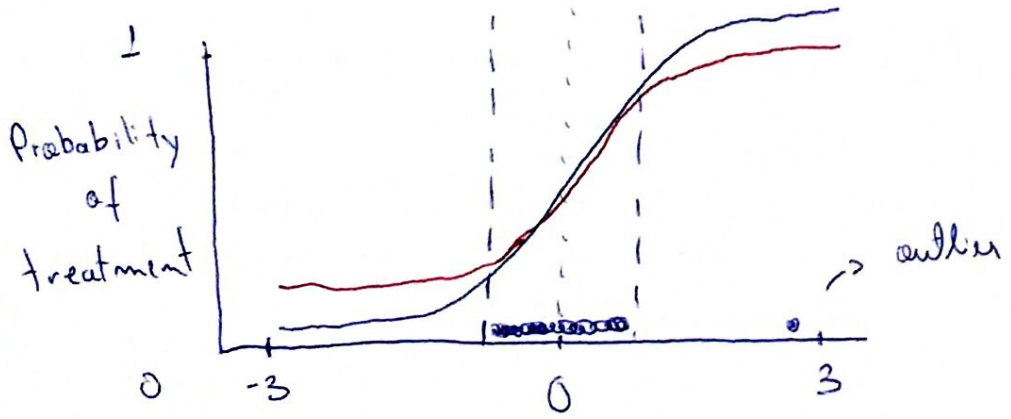
Checking Weights



Also, you can check summary stats about the weights

• Remedies for large weights

First of all: why do we have large weights? What is driving it? A combination of features or just a value? Do this person have data that look reasonable? Problem w/ the propensity score model?



Alternative curve that fits well observed data as well as the blue one and is not damaged as much by the outlier

- Trimming the tails

Large weights happen at the tails of the propensity score distributions.

Trimming the tails can eliminate some of the extreme values.

A common trimming strategy:

- Remove treated subjects whose the propensity scores are above the 98th percentile from the dist among controls
- Remove control subjects whose propensity scores are below the 2nd percentile from the distribution of treated subjects.

Reminder: trimming the tails changes the population

- Another option: weight truncation

Steps:

1) Determine the maximum allowed weight
   - Could be specific value (ex 100)
   - Could be based on a percentile (ex 99th)
2) If a weight is greater than the maximum allowable, ~~[illegible]~~ set it to the max allowable value
   - So if max is 100 and someone is 10K, set it to 100.

↳ Bias - variance trade-off:

Truncation: bias, but smaller variance

no truncation: unbiased, large variance

Truncating extremely large weights can improve the MSE.

A crash course in Causality

- Doubly robust estimators

Background: • IPTW

$$E(Y^1) = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\pi_1(X_i)}$$

- Regression based estimation

Estimate $E(Y^1)$ by specifying an outcome model $m_1(X) = E(Y \mid A=1, X)$, then averaging over the distribution of X:

$$\frac{1}{n}\sum_{i=1}^{n}\{A_i Y_i + (1-A_i)\, m_1(X_i)\}$$

↑ For subjects w/ $A=1$, use observed Y

↑ For other subjects, use predicted value of Y given their X's if their A had been 1.

→ A doubly robust estimator is an estimator that is unbiased if either the propensity score model OR the outcome regression model are correctly specified.

Ex:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\frac{A_i Y_i}{\pi_i(X_i)}}_{\text{IPTW}} - \underbrace{\frac{A_i - \pi_i(X_i)}{\pi_i(X_i)}\, m_1(X_i)}_{\text{Augmentation}}\right\}$$

If propensity score is correctly specified, but the model is not, $E[A_i] = \pi_i(X_i)$, so augmentation part is zero. We want to

find out $E(Y_1)$. Since we have a sample average ($\frac{1}{n}\sum^{n}\ldots$), as n grows we expect to have $E[\ldots]$, so $E[\ldots]$ should match $E[Y_1]$.

We can rearrange like:

$$\frac{1}{n}\sum^{n}\left\{\underbrace{\frac{A_i(Y_i - m_1(X_i))}{\pi_1(X_i)}}_{=0} + \underbrace{m_1(X_i)}_{\to E[Y_1]}\right\}$$

So if propensity score is wrong, but outcome is correct, the final eq will keep only the outcome model.

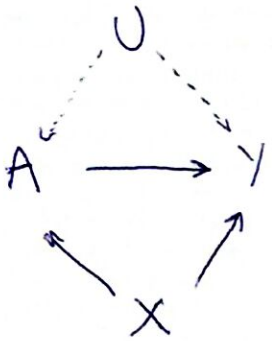
They are also known as augmented IPTW (AIPTW)

Can use semiparametric theory to identify best estimators.

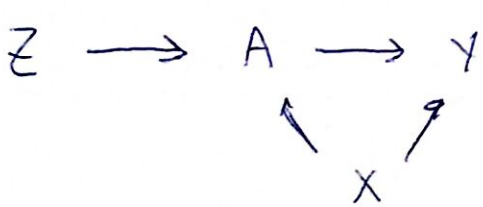
AIPTW should be more efficient than IPTW.

A crash course in causality

• Instrumental Variables



To solve confounding we can use matching, PS matching, IPTW and etc. But what if the confounder is unmeasured



An instrumental variable affects A, think about it as an encouragement, since it doesn't affect Y directly.

• Encouragement design

For smoking, birthweight problem, IV could be an encouragement to stop smoking, it does not impact the weight.

An intention to treat analysis would focus on the causal effect of encouragement

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

It tries to use the randomization provided by the IV to calculate the causal effect of smoking itself

IV can be random assigned as part of the experiment or it's believed to be randomized in nature (natural experiment)

- Randomized trials with noncompliance

Z: Randomization to treatment

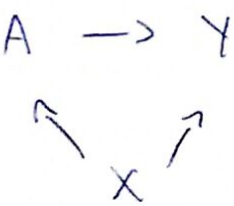
A: Treatment received

Y: Outcome

not everyone assigned will receive treatment (non-compliance)

non-compliance make RCT look like observational

Data: (Z, A, Y)



- Potential values of treatment:

$A^{Z=1} = A^1$

$A^{Z=0} = A^0$

⇒ Causal effect of assignment on receipt

CE of treatment assignment on treatment received as $E(A^1 - A^0)$

This is the proportion treated if everyone had been assigned to received treatment, minus the proportion treated if no one had been assigned to receive treatment

If perfect compliance, this would be equal to 1.

This is estimable from the observed data:

$E(A^1) = E(A|Z=1)$, $E(A^0) = E(A|Z=0)$

A Crash course in causality

CE of treatment assignment on outcome

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

Intention-to-treat effect. If perfect compliance, this would be equal to the causal effect of treatment.

Estimable from the observed data:

$$E(Y^{Z=1}) = E(Y \mid Z=1), \quad E(Y^{Z=0}) = E(Y \mid Z=0)$$

→ What about CE of treatment received?

Z is an IV!

• Compliance classes

Potential value of outcomes: Y^0, Y^1

Classify people based on the potential treatment:

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

Never-takers: we'd never observe the treatment, we can't say anything about CE on them.

Compliers: treatment received is randomized, we can estimate CE.

Defiers: It's randomized, but in the opposite way (unusual).

Always takers: again, no variable

A motivation for using IV is when there are unmeasured confounders (we cannot marginalize over all confounders via matching, IPTW ...)

IV methods don't focus on the ACE for the population, but on a local ↓ sub pop ATE.

The target of inference is:

$$E(Y^{Z=1} \mid \underbrace{A^0=0, A^1=1}_{\text{Same sub pop}}) - E(Y^{Z=0} \mid \underbrace{A^0=0, A^1=1}_{\text{Same sub pop}})$$

$$= E(Y^{Z=1} - Y^{Z=0} \mid \text{Compliers})$$

$$E(Y^{A=1} - Y^{A=0} \mid \text{Compliers})$$

} from treat assigned to treat received

This is causal because it contrasts counterfactuals in a common population. Known as complier average causal effect (CACE). no inference about the other sub population.

We can observe Z and A (assigned and taken)

Z	A	A^0	A^1	Class
0	0	0	?	never-taker or complier
0	1	1	?	Always-taker or defier
1	0	?	0	never-taker or defier
1	1	?	1	Always-taker or complier

- Identifiability

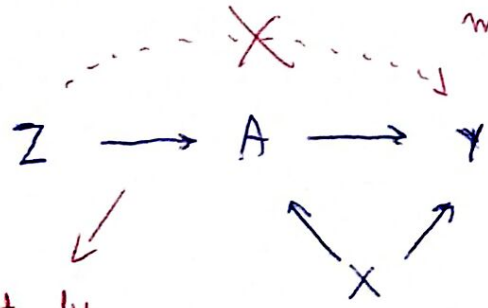
Compliance classes are also known as principal strata. There are latent (not directly observed)

- A crash course in causality

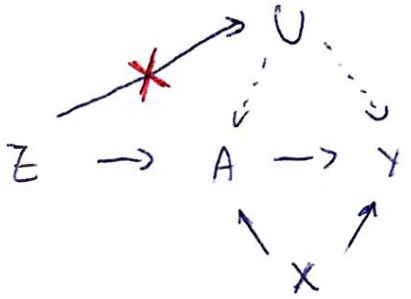
- IV Assumptions

A variable is an instrumental variable if:

1. It's associated with the treatment
2. It affects the outcome only through its effect on treatment;
   - This is known as the exclusion restriction



Imagine unmeasured confounder



Z cannot directly, or indirectly via U, affect Y.

Realistic?

As a coin flip, randomization as IV, impact the A, but not Y or U.

But if subjects are not blinded, knowledge of what they were assigned could affect them

need to exam this assumption for any given study.

The monotonicity assumption is that there are no defiers. No one consistently does the opposite of what they are told. It's called monotonicity because the assumption

is that the probability of treatment should increase w/ more encouragement

- Causal Effect identification and estimation

Goal is to estimate $E(Y^{a=1} - Y^{a=0} \mid \text{compliers})$

Starting w/ the ITT effect:

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

$$E(Y \mid Z=1) = E(Y \mid Z=1, \text{always-takers}) P(AT)$$

$$+ E(Y \mid Z=1, \text{never takers}) P(NT)$$

$$+ E(Y \mid Z=1, \text{compliers}) P(CO)$$

-> It's a weighted value / average in the 3 subpopulation

Among AT and NT, Z does nothing

$$E(Y \mid Z=1, AT) = E(Y \mid AT)$$

$$E(Y \mid Z=1, NT) = E(Y \mid NT)$$

$$P(AT \mid Z) = P(AT)$$

So: $E(Y \mid Z=1) = E(Y \mid AT) P(AT)$

$$+ E(Y \mid NT) P(NT)$$

$$+ E(Y \mid Z=1, CO) P(CO)$$

$$E(Y \mid Z=0) = E(Y \mid AT) P(AT)$$

$$+ E(Y \mid NT) P(NT)$$

$$+ E(Y \mid Z=0, CO) P(CO)$$

$$\Rightarrow E(Y \mid Z=1) - E(Y \mid Z=0) = E(Y \mid Z=1, CO) P(CO)$$

$$- E(Y \mid Z=0, CO) P(CO)$$

• A crash course in Causality

which implies:

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

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

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

$$= CACE$$

note that $P(CO) = E(A|Z=1) - E(A|Z=0)$

$E(A|Z=1)$: prop of always takers or compliers

$E(A|Z=0)$ = Always takers

$$CACE = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)}$$ 1 / 2

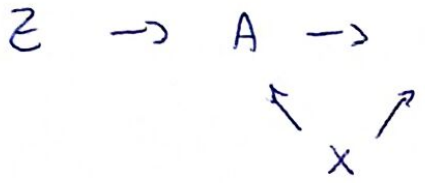
1: ITT, causal effect of treatment assignment on the outcome

2: Causal effect of treatment assignment on the treatment received.

If perfect compliance, ITT = CACE

CACE : at least ITT. ITT understates the effect because there are ppl that don't take it

• IVs in observational studies



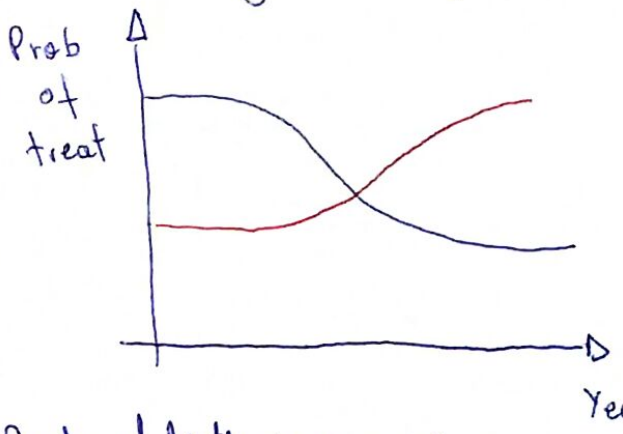
We can think of Z as encouragement, binary: yes or no; continuous: dose of encouragement

The assumption of influence in treat can be tested in data. The exclusion restriction assumption will largely need to rely on subject matter knowledge

Examples:

① Sometimes treatment preference change over a short period of time

Z: early time period Vs late time period



Quarter of birth → years → Income
School

If they are close in time, you have a reason to believe the restriction assumption

② Distance as IV

- Short distance is stronger encouragement
- Is distance likely associated with outcome

Provider preference: use treatment prescribed to previous patients as an IV for current patient. Idea: previous patient should be associated w/ current decision, but previous decision should not affect outcome.

A crash course in Causality

- Two stage least squares

OLS

Treatment A and outcome Y:

$$Y_i = \beta_0 + A_i \beta_1 + \epsilon_i$$

Should be independent

If there is confounding, A_i and ϵ_i will be correlated, so OLS fails here. β_1 won't reflect a causal effect. Even w/ some confounders in the model, if there's any unmeasured ones, it will fail.

- 2 Stages

① Regress treatment received, A, on the instrumental variable, Z

$$A_i = \alpha_0 + Z_i \alpha_1 + \epsilon_i$$

- where the error term is mean zero, constant variance
- By randomization, Z_i and ϵ_i are independent

Obtain $\hat{A}_i = \hat{\alpha}_0 + Z_i \hat{\alpha}_1$ for each subject.

- This is the predicted value of A given Z

② Regress the outcome, Y, on the fitted value from stage 1, $\hat{A}_i$:

$$Y_i = \beta_0 + \hat{A}_i \beta_1 + \epsilon_i$$

- where the error term is mean zero, constant variance
- By exclusion restriction Z independent of Y, given A
- $\hat{A}$ is projection of A onto space spanned by Z

• The estimate of β_1 is estimate of the causal effect

$\hat{A}$ is fully determinated by Z!

$\hat{A}_i$ is estimate of $E(A|Z)$

② $Y_i = \beta_0 + \hat{A}_i \beta_1 + \varepsilon_i$

$\beta_1 = E(Y|\hat{A}=1) - E(Y|\hat{A}=0)$

Suppose Z and A binary.

There are two values of $\hat{A}$ in the 2nd stage model: $\hat{\alpha}_0$ and $\hat{\alpha}_0 + \hat{\alpha}_1$.

If some non-compliance, there will not be 0 and 1

What we observe is going from $\hat{\alpha}_0$ to $\hat{\alpha}_0 + \hat{\alpha}_1$ in $\hat{A}$.

↳ This happens when we go from $Z=0$ to $Z=1$

We observe a mean difference of $\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)$ with a $\hat{\alpha}_1$ unit change in $\hat{A}$.

If we see a change in the mean of Y for a $\hat{\alpha}_1$ unit change in $\hat{A}$, then we should see a $\frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{\alpha}_1}$ unit change in the mean of Y for a 1 unit change in $\hat{A}$.

$\beta_1 = CACE = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)}$

↳ We can add X covariates in both stages

* Sensitivity analysis: if Z does influence Y directly by an amount α. If proportion of defiers is Π ... what happens?

A Crash course in Causality

• Weak Instruments

How to measure the strength of a IV

↳ How well does it predict the probability of treatment.

Estimate the proportion of compliers:

$E(A|Z=1) - E(A|Z=0)$, if close to 1, it's strong, if close to zero, it's weak.

A weak encouragement will lead to having just a few examples to estimate CE, large variance, unstable. Confidence intervals are too large to be useful

-> methods on strengthening IVs.

↳ near / far matching: match on that covariates are similar, but the instrument is very different