

# Causal Inference I

MIXTAPE SESSION

---



# Roadmap

## Selecting Covariates Using Directed Acyclic Graphs

- Graph notation

- Backdoor criterion

- Collider bias

- What if you don't know the DAG?

## Unconfoundedness and Ignorable Treatment Assignment

- Aggregate target parameters

- Assumptions

## Estimators

- Subclassification

- Exact and Inexact Matching

- Propensity scores

- Regressions

## Concluding remarks

# Graphs

- Now we turn from potential outcomes modeling of causal effects to causal graphs
- Very important area, very common to see it in computer science intersections with data science, particularly tech, and often very advanced
- My focus is very narrow – I am using it mainly to help us carefully reason through design elements around matching and instrumental variables

# Adjusting for variables

- One of the first things you learn in a methods course is multivariate regression “controlling for  $X$ ”
- What is this? Why do we do this? What should  $X$  be? What causal parameter does it help identify?
- Unconfoundedness, selection on observables, ignorable treatment assignment are different terms describing the same thing – the RCT is still occurring, only within the dimensions of a conditioning set of confounders and covariates

Judea Pearl, 2011 Turing Award winner, drinking his first IPA



# Judea Pearl and DAGs

- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology
- They make causality concepts extremely clear, they provide a map to the estimation strategy, and maybe best of all, they communicate to others what must be true about the data generating process to recover the causal effect

# Design vs. Model

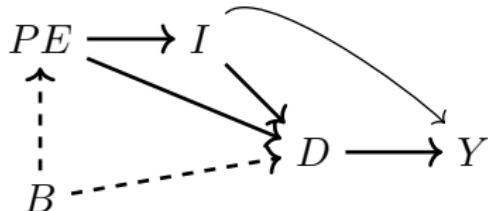
- DAGs require prior theory of treatment assignment in the world, making it more structural than design oriented methods which exploit randomization or simple treatment assignment mechanisms
- While DAGs are compatible with design based approaches, they are less likely to follow the same estimation methodologies
- DAGs are used in epidemiology a lot, and are also extremely common in industry and machine learning, including AI (e.g., causalens),
- My review of them will largely be used to service the design based approaches using covariate adjustment and instrumental variables

## Further reading

1. Pearl and MacKenzie (2018) The Book of Why: The New Science of Cause and Effect, Basic Books (*wildly popular*)
2. Pearl, Glymour and Jewell (2016)  
Causal Inference In Statistics: A Primer, Wiley Books (*accessible*)
3. Pearl (2009) Causality: Models, Reasoning and Inference, Cambridge, 2nd edition (*advanced*)
4. Morgan and Winship (2014)  
Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press, 2nd edition  
(*excellent*)

## Causal model

- The DAG is a sense “structural” because it describes (correctly) a system of equations that determines the treatment, the outcome, and all relevant routes between them
- Also called the structural model, but should not be confused with “structural econometrics” (e.g., Rust, Wolpin)
- Consider the following diagram representing the returns to education with simplified confounders



- $B$  is a **parent** of  $PE$  and  $D$
- $PE$  and  $D$  are **descendants** of  $B$
- There is a **direct (causal) path** from  $D$  to  $Y$
- There is a **mediated (causal) path** from  $B$  to  $Y$  through  $D$
- There are a lot of **paths** from  $PE$  to  $Y$  but none are direct, and some go “downstream” but others go “upstream”

# DAGs are harder than they seem

- DAGs are meant to represent the truth about *your data* – what determines the treatment in *your dataset* may be different from what determines in someone else's even if the topic is the same
- But you also need to be able to justify the choices
- DAGs have a complicated connection to the design methodologies which typically focus on randomization, not prior knowledge about the system of equations, and yet many design methods do in fact require prior statements that likely can only be stated with some willingness to commit

# Creating your DAG

- So, your goal is to, in good faith, create a DAG that is a reasonable and honest approximation of  $D$  and  $Y$  parents (confounders) as well as direct and indirect effects of  $D$  on  $Y$
- Where there is uncertainty, you have to acknowledge that in the DAG, and where there are convictions, you acknowledge that too
- We get ideas for DAGs from theory, models, observation, experience, prior studies, intuition, as well as conversations with domain experts, but keep in mind – these are not natural experiments; they are not exploiting randomized treatment assignment.

# Beware of Lazy DAGs

- Lazy DAG building is *extremely common* as the theory of DAGs and the practice of using DAGs is not the same thing
- You are engaging in *modeling the outcome* and if you are wrong, then the estimation is wrong
- Part of what led to these design approaches was because that modeling of the outcome had been unsuccessful in empirical labor

## Concluding caveats

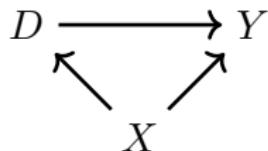
- This is not to turn you off but to impress upon you that this is a very different enterprise which is why I mainly use it for selecting covariates and justifying instruments
- But that said, it is an *extremely* important area of causal inference
- We will now turn to using DAGs for picking covariates using as "controls"

# Unconfoundedness and the backdoor criterion

- We will focus today on the unconfoundedness research design, which in my opinion is best described in causal graphs with the concept of the **backdoor criterion**
- As we will see, the DAG helps you solve the problem of choosing covariates for a model to resolve selection bias, but to do so requires confidence in your DAG

# Confounding

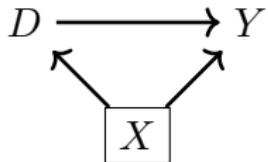
- Confounders create differences in  $D$  as well as differences in potential outcomes, and so therefore  $X$  creates selection bias



- Our knowledge about this will enable us to obtain unbiased and consistent average treatment effects by adjusting for the distribution of  $X$  in our samples

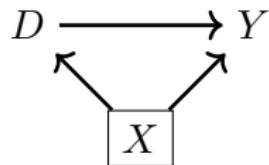
# Backdoor Paths

- Confounding creates **backdoor paths** between treatment and outcome ( $D \leftarrow X \rightarrow Y$ ) – i.e., spurious correlations
  - Distinct from something called a collider path ( $D \rightarrow X \leftarrow Y$ )
  - Distinct from something called a mediator path ( $D \rightarrow X \rightarrow Y$ )
- We can “block” any particular backdoor path by conditioning on variable  $X$  so long as it is not a collider (visualized here with a square over  $X$ )



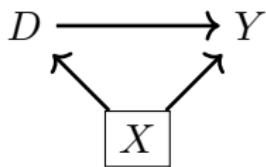
# Backdoor Paths

- Once we condition on  $X$ , we can calculate average differences in  $Y$  by treatment status to obtain an estimate of an aggregate causal parameter
- There are many methods for doing this and we cover them today – regression, matching, stratification, weights
- But all of them at their fundamental level are calculating simple differences in mean outcomes for given values of  $X$  and then taking weighted averages



# Backdoor Paths

- When all backdoor paths from  $D$  to  $Y$  are blocked, then the only remaining path between  $D$  to  $Y$  is the causal path
- We call this satisfying the backdoor criterion using Pearl's DAG terminology, and we call it unconfoundedness using Rubin's potential outcomes terminology (which I'll discuss later)



# Backdoor criterion

## Backdoor criterion

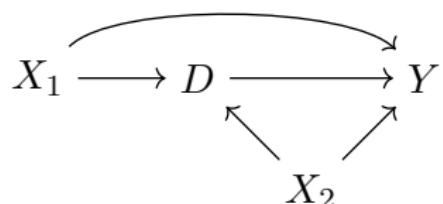
Conditioning on  $X$  satisfies the backdoor criterion with respect to  $(D, Y)$  directed path if:

1. All backdoor paths are blocked by  $X$
2. No element of  $X$  is a collider

Conditioning on a non-collider is sufficient to closing a backdoor path even if it's been opened by conditioning on a collider, so just be sure to close a backdoor path if you opened it with a collider

# What control strategy meets the backdoor criterion?

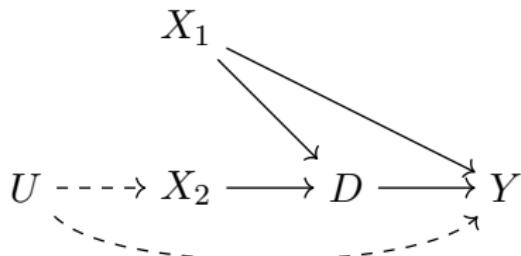
- List all backdoor paths from  $D$  to  $Y$ . I'll wait.



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?

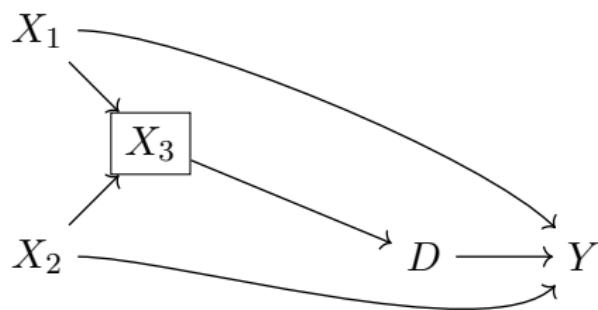
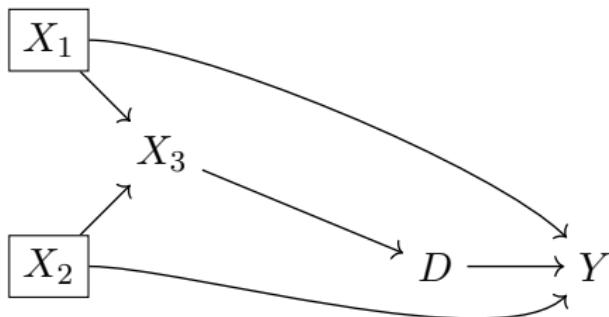
## What if you have an unobservable?

- List all the backdoor paths from  $D$  to  $Y$ .



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?
- What about the unobserved variable,  $U$ ?

## Multiple strategies



- Conditioning on the common causes,  $X_1$  and  $X_2$ , is sufficient
- ... but so is conditioning on  $X_3$

# Collider bias

- Backdoor paths can remain open in covariate adjustment strategies through two ways:
  1. You did not close the path because you did not condition on the confounder
  2. Your conditioning variable opened up a previously closed backdoor path because on that path the variable was a **collider**
- Colliders are “bad controls” which when you control for them, *create* new previously non-existent spurious correlations (not commonly discussed, even in economics)
- This is the risk of blindly controlling for variables – you may inadvertently include bad or irrelevant controls

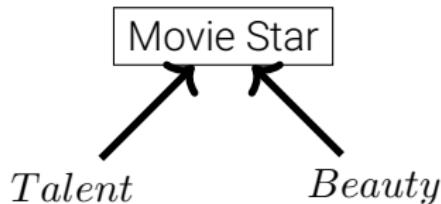
## Example 1: Movie stars

**Important:** Since unconditioned colliders block back-door paths, what exactly does conditioning on a collider do? Let's illustrate with a fun example and some made-up data

- CNN.com headline: Megan Fox voted worst – but sexiest – actress of 2009 ([link](#))
- Are these two things actually negatively correlated in the world?
- Assume talent and beauty are independent, but each causes someone to become a movie star. What's the correlation between talent and beauty for a sample of movie stars compared to the population as a whole (stars and non-stars)?

# Movie star DAG

Imagine casting directors pick movie stars based on talent and beauty



Talent and beauty can become correlated even though they are independent



*Figure:* Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent.

## Sample selection?

- Notice that this is clear when we are focused on sample selection
- But even a regression that included “star” would create the issue:

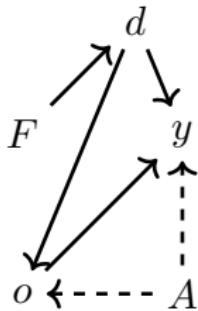
$$beauty_i = \alpha + \delta talent_i + \beta star_i + \varepsilon_i$$

- It's not just sample selection

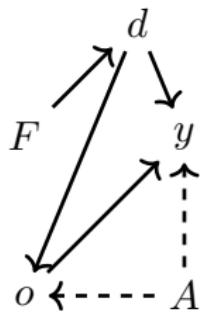
## Example 2: Discrimination

- Let's look at another example: very common for think tanks and journalists to say that the gender gap in earnings disappears once you control for occupation.
- But what if occupation is a collider, which it could be in a model with occupational sorting
- Then controlling for occupation in a wage regression searching for discrimination can lead to all kinds of crazy results even *in a simulation where we explicitly design there to be discrimination*

# DAG



$F$  is female,  $d$  is discrimination,  $o$  is occupation,  $y$  is earnings and  $A$  is ability. Dashed lines mean the variable cannot be observed. Note, by design, being a female has no effect on earnings or occupation, and has no relationship with ability. So earnings is coming through discrimination, occupation, and ability.



## Mediation and Backdoor paths

1.  $d \rightarrow o \rightarrow y$
2.  $d \rightarrow o \leftarrow A \rightarrow y$

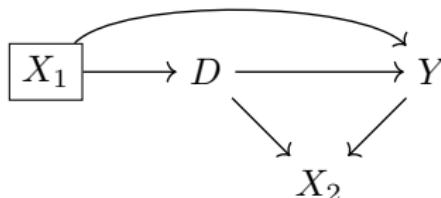
*Table:* Regressions illustrating collider bias with simulated gender disparity

Covariates:	<b>Unbiased combined effect</b>	<b>Biased</b>	<b>Unbiased wage effect only</b>
Female	-3.074*** (0.000)	0.601*** (0.000)	-0.994*** (0.000)
Occupation		1.793*** (0.000)	0.991*** (0.000)
Ability			2.017*** (0.000)
N	10,000	10,000	10,000
Mean of dependent variable	0.45	0.45	0.45

- Recall we designed there to be a discrimination coefficient of -1
- If we do not control for occupation, then we get the combined effect of  $d \rightarrow o \rightarrow y$  and  $d \rightarrow y$
- Because it seems intuitive to control for occupation, notice column 2 - the sign flips!
- We are only able to isolate the direct causal effect by conditioning on ability and occupation, but ability is unobserved

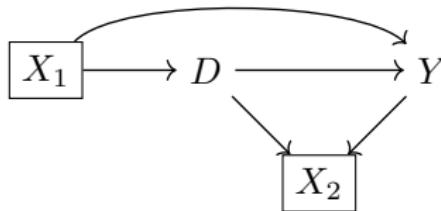
- **Colliders can be outcomes (and often those are the ones)**

→ There is only one backdoor path from  $D$  to  $Y$



→ Conditioning on  $X_1$  blocks the backdoor path

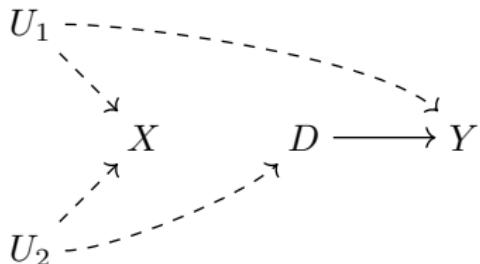
→ But what if we also condition on  $X_2$ ?



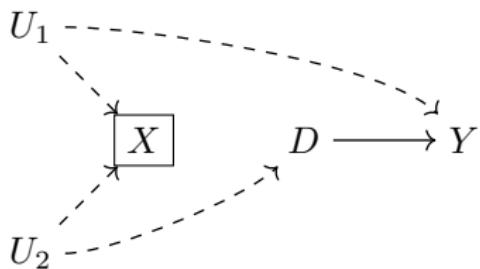
→ Conditioning on  $X_2$  opens up a new path, creating new spurious correlations between  $D$  and  $Y$

- Colliders could be pre-treatment covariates (called M-bias because it looks like an M)

→ Name the backdoor paths. Is it open or closed?

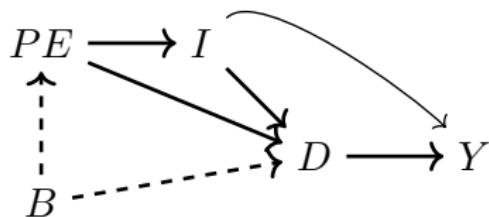


→ But what if we condition on  $X$ ?



# Testing the Validity of the DAG

- The DAG makes testable predictions
- Conditional on  $D$  and  $I$ , parental education ( $PE$ ) should no longer be correlated with  $Y$
- Can be hard to figure this out by hand, but software can help (e.g., Daggity.net is browser based, Causal Fusion is more advanced)
- Causal algorithms tend to be DAG based and are becoming popular in industry



## Contrast this with ordinary practices

- Person attempts to “control for omitted variable bias” by including as many “controls” as possible
- Person does not even attempt to think about treatment assignment mechanism and therefore has no idea what variables are colliders, covariates or confounders
- Machine learning can be “naive” too if the large dimension of features includes unknowingly colliders

# Double Machine Learning and Automated Model Selection: A Cautionary Tale

PAUL HÜNERMUND<sup>†</sup>

BEYERS LOUW<sup>‡</sup>

ITAMAR CASPI<sup>\*</sup>

<sup>†</sup>*Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK.*

E-mail: phu.si@cbs.dk

<sup>‡</sup>*Maastricht University, Tongersestraat 53, 6211 LM Maastricht, NL.*

E-mail: jb.louw@maastrichtuniversity.nl

<sup>\*</sup>*Bank of Israel, P.O.Box 780, 91007, Jerusalem, IL*

E-mail: itamar.caspi@boi.org.il

This version: May 25, 2023

First version: August 26, 2021

**Summary** Double machine learning (DML) has become an increasingly popular tool for automated variable selection in high-dimensional settings. Even though the ability to deal with a large number of potential covariates can render selection-on-observables assumptions more plausible, there is at the same time a growing risk that endogenous variables are included, which would lead to the violation of conditional independence. This paper demonstrates that DML is very sensitive to the inclusion of only a few “bad controls” in the covariate space. The resulting bias varies with the nature of the theoretical causal model, which raises concerns about the feasibility of selecting control variables in a data-driven way.

**Keywords:** *Double/Debiased Machine Learning, Directed Acyclic Graphs, Bad Controls, Backdoor Adjustment, Collider Bias, Causal Hierarchy*

## Covariate selection without DAGs

- What if you don't have a DAG you feel confident about?
  1. Include confounders that you feel pretty confident are there
  2. Include covariates that are *highly predictive* of the missing counterfactual (e.g.,  $Y^0$  for the ATT)
  3. Avoid outcomes (even though that still won't address M-bias colliders)
- While this approach may be less formalized, you are at least reasoning about the treatment assignment mechanism as opposed to just including whatever variables you have laying around (avoid the "kitchen sink regressions")

## Falsifications as a test

- Covariates should not be affected by the treatment, so examining them as falsifications can help establish the credibility of unconfoundedness
- Imbens and Rubin (2015) suggested using the lagged outcome (pre-treatment) as a way of checking, as those have similar confounder structures
- Falsifications too: One study questioned a finding that obesity was contagious in social networks by estimating the same model on things that cannot be contagious like acne, headaches and height and found the same things (likely confounding existed)

# Roadmap

## Selecting Covariates Using Directed Acyclic Graphs

- Graph notation

- Backdoor criterion

- Collider bias

- What if you don't know the DAG?

## Unconfoundedness and Ignorable Treatment Assignment

- Aggregate target parameters

- Assumptions

## Estimators

- Subclassification

- Exact and Inexact Matching

- Propensity scores

- Regressions

## Concluding remarks

## Shifting gears

- Having reviewed the fundamentals of DAGs, we will now move into the practical realm
- You have a non-experimental study.
- You have drawn your DAG or you have convinced yourself of your controls
- We now discuss parameters, assumptions and estimation

## Step 1: Pick the causal parameter of interest

- Which population are you interested in?
  - Everyone: ATE
  - Just the treated units: ATT
  - The control units only: ATU
- Interestingly, controlling for covariates can identify any of those (unlike IV which only identifies the LATE, diff-in-diff and synth which only identify the ATT)
- But the assumptions and specifications required to do so differ, so you have to start step 1 and decide which population and therefore which causal parameter you are targeting

## Aggregate causal parameters have missing potential outcomes

- Every aggregate causal parameter is missing some potential outcome (e.g., ATT is missing  $E[Y^0|D = 1]$ )
- If we are going to estimate the ATT, then it must be we are estimating  $E[Y^0|D = 1]$  somehow, but how?
- If we are using controls, then we in some way, shape or form “filling in” the missing counterfactual using the covariates the comparison group
- “At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others.” – Imbens and Rubin (2015)

# Assumptions

- When attempting to estimate aggregate causal parameters using covariate adjustment, there are two assumptions
  1. **Unconfoundedness:** fairly controversial to some because of what it implies about human behavior
  2. **Common support:** since most people stop at assumption 1, there's much less attention to common support
- Today we will assume you satisfy unconfoundedness, and therefore our focus is on violations of assumption 2

## Identifying assumption I: Unconfoundedness

$(Y^0, Y^1) \perp\!\!\!\perp D|X$ . There exists a set  $X$  of known and quantified confounders such that after adjusting for them, treatment assignment is *independent of potential outcomes*.

- Conditional on  $X$ , treatment assignment is randomly distributed (i.e., independent of both potential outcomes) – strong assumption
- For a large group of people within the same strata of  $X$ , this assumption states they choose treatments by flipping coins (not because they think the treatment helped them)
- Eliminating all backdoor paths on a DAG through blocking satisfies unconfoundedness; also called ignorability

## Identifying assumption I: Unconfoundedness

$(Y^0, Y^1) \perp\!\!\!\perp D | X$ . There exists a set  $X$  of known and quantified confounders such that after adjusting for them, treatment assignment is *independent of potential outcomes*.

$$\begin{aligned} E[Y^0 | D = 1, X = x] &= E[Y^0 | D = 0, X = x] \\ E[Y^1 | D = 1, X = x] &= E[Y^1 | D = 0, X = x] \end{aligned}$$

Unconfoundedness justifies substituting units in treatment for control based on  $X = x$  – but only if there are exact matches

# Economic meaning of backdoor criterion

- When people choose their own treatments, attempting to estimate aggregate causal parameters using covariates that conditional on those covariates, they are *randomizing choices*
- See my substack post, "Why do economists so dislike 'conditional independence'? (January 4, 2023) at <https://causalinf.substack.com/p/why-do-economists-so-dislike-conditional-independence>

## Why do economists so dislike "conditional independence"?

A humble theory from an economist



SCOTT CUNNINGHAM

JAN 4, 2023



25



11



1

Share

...

One of the first causal methods we ever learn in statistics class is to “control for X”. We can use a regression that controls for X, and if we wonder what that means, we need only review the Frisch-Waugh-Lovell theorem to see what a multivariate regression is equivalent to. As you progress, you may learn that there are a whole range of estimators, though, that use covariates to reconstruct a missing counterfactual when trying to estimate some average treatment effect. There’s matching methods which basically impute missing counterfactuals by finding units in the comparison group that have the same or almost the same covariate values. There’s even fixes for when you can’t find the exact matches, too, such as Abadie and Imbens (2011) bias correction methods for matching discrepancies. There’s propensity scores which can be used, also, to find matches, as well as used as weights in simple comparisons between treatment and control. And then there are things sort of in between like coarsened exact matching. The number of ways in which you can try to solve thorny causal inference problems through the use of covariates is very long, and very old, and if you enjoy causal inference and econometrics, many of these will the more you spend time with become interesting, and maybe even beautiful.

But ask an economist if these beautiful objects are fit for solving causal inference problems, and more times than not, the answer will be an unambiguous “no”. They won’t even flinch when they say it too! Many economists will just flat out look you in the face and say not only are these unlikely to be useful in real life — they just refuse to even accept the possibility that they’ll ever work.

# Common support visual



Common support is like having a bridge that allows you to move soldiers and equipment across a moat.

## Identifying assumption II: Common support

For ranges of  $X$ , there is a positive probability of being both treated and untreated

- There exists units in treatment and control with same values of  $X$  – you can't make the substitutions otherwise
- Dimension  $k$  means every specific combination of the conditioning set (e.g., not males and old, but adult males, adult females, youth male, youth female)
- Testable because common support is observable unlike unconfoundedness, but as you can imagine if the dimensions of  $X$  gets large (and with a continuous covariate it's infinite!) then it won't hold in any finite sample!

## Assumptions combined

But if we have them both (represented below), we can even outside of an RCT estimate the ATE through nonparametric matching

1.  $(Y^1, Y^0) \perp\!\!\!\perp D|X$  (strong unconfoundedness)
2.  $0 < Pr(D = 1|X) < 1$  with probability one (common support)

Comparing groups of individuals who have *the same values of  $X$* , treatment is no longer based gains,  $\delta$ .

The second term implies we have people in treatment and control for every strata of  $X$

# Estimating ATE with Assumptions

- Unconfoundedness lets you use  $Y_j^0$  from control group as  $i$ 's  $\textcolor{blue}{Y}_i^0$  and  $Y_i^1$  from treatment group as  $j$ 's  $\textcolor{blue}{Y}_j^1$  using  $X$  as the matching guide

$$\begin{aligned} E[Y^1 - Y^0 | X] &= E[Y^1 - Y^0 | X, D = 1] \\ &= E[Y | X, D = 1] - E[Y | X, D = 0] \end{aligned}$$

- Common support (the bridge metaphor) allows the match to take place as well as weight over the covariate distribution

$$\begin{aligned} \delta_{ATE} &= E[Y^1 - Y^0] = E\left[E[Y^1 - Y^0 | X]\right] \\ &= \int E[Y^1 - Y^0 | X, D = 1] dPr(X) \\ &= \int (E[Y | X, D = 1] - E[Y | X, D = 0]) dPr(X) \end{aligned}$$

## Maybe You Want the ATT

ATE requires conditional independence with respect to both  $Y^1$  and  $Y^0$  which would mean complete irrationality

If we want the ATT, we can go with strictly weaker assumptions – weak unconfoundedness and weak overlap

One kind of selection: person chooses the treatment based on what you gain,  $Y^1$ , but not what you lose,  $Y^0$ , and not net benefits,  $\delta = Y^1 - Y^0$

## ATT Identification

We can modify those assumptions and weaken both which helps a lot

1.  $Y^0 \perp\!\!\!\perp D|X$  (weak unconfoundedness)
2.  $Pr(D = 1|X) < 1$  (with  $Pr(D = 1) > 0$ ) (weak support)

We don't need full common support because we don't need to find counterfactuals for the control group – we only need units in the control group that match with our treatment group

Selection is weaker too, like I said – they are not entirely irrational, but who knows if it helps you

Z

## Estimating ATT

Weighted averages under both assumptions:

$$\delta_{ATT} = \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X|D = 1)$$

We match units in treatment and control because under weak unconfoundedness they're substitutable, and we use weak common support so that we can actually do it, then we take weighted averages over the differences.

What assumptions and method would you use to estimate the ATU?

# Estimators

- Now we will explore estimators that for lack of a better word “use covariates” to estimate aggregate causal parameters
- I will be bundling them around a few topics: exact matching, inexact matching, and regressions
- Themes about heterogeneous treatment effects, common support and correct (and incorrect) regression specifications will be common

# Roadmap

## Selecting Covariates Using Directed Acyclic Graphs

- Graph notation

- Backdoor criterion

- Collider bias

- What if you don't know the DAG?

## Unconfoundedness and Ignorable Treatment Assignment

- Aggregate target parameters

- Assumptions

## Estimators

- Subclassification

- Exact and Inexact Matching

- Propensity scores

- Regressions

## Concluding remarks

# Causal inference using covariates

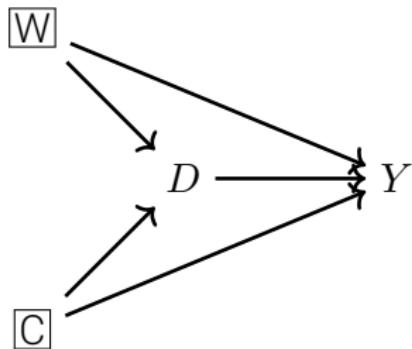
- Multivariate regression goes back to Yule (19th century) and Fisher (20th century)
- Subclassification was introduced by Cochran (late 60s)
- Propensity score matching introduced early 80s (Rosenbaum and Rubin)
- Regression based decomposition methods by Kitagawa (1950s), Oaxaca and Blinder (early 70s), and Kline, Wooldridge and others (in 00s)
- Machine learning methods like double debiased ML by Chernozhukov, et al. (2018) and causal forests (Athey and Wager) 2018

## Subclassification method

- Wanting to show subclassification as an early effort to tackle covariates, but which ultimately cannot handle large features due to common support issues before moving into regression then DML
- Titanic sank on maiden voyage April 15, 1912 after hitting an iceberg in North Atlantic
- 2200 on board, but only 700 survived, despite 20 lifeboats with 60 capacity (1200 potential lives could've been saved)
- Women and children first was a maritime rule to ration lifeboats, but there were different cabins (1st class, 2nd class, etc.) on different levels with different proximity to boats
- What was the causal effect of 1st class on survival adjusting for  $W$  and  $C$ ?

## Exercise: Titanic DAG

Figure: Women  $W$  and children  $C$  first maritime rule is a confounder for estimating first class  $D$  effect on surviving  $Y$



Backdoor criterion can be satisfied by blocking on  $W$  and  $C$ . These are our known confounders. Now we just need data to see if it's quantified.

## Titanic exercise

1. **Stratify the confounders:** Our age and sex variables are both binary, so we can only create four strata: male children, female children, male adults, female adults
2. **Calculate differences within strata:** Calculate average survival rates for each group within each of the four strata and difference within strata
3. **Calculate probability weights:** Count the number of people in each strata and divide by the total number of souls aboard (crew and passengers)
4. **Aggregate differences across strata using weights:** Estimate the ATE by aggregating the difference in survival rates over the four strata with each strata-specific difference weighted by that strata's weight

# Table 1: Typical balance table

*Table:* Differences in female and adult passengers by first class status on the Titanic.

<b>Variable name</b>	<b>First class</b>		<b>All other classes</b>	
	Obs	Mean	Obs	Mean
Percent adult	325	98.2%	1,876	94.5%
Percent female	325	44.6%	1,876	17.3%

## Table 2: Stratified sample

*Table:* Counts and Titanic survival rates by strata and first class status.

Strata	First class		All other classes		Total
	Obs	Mean	Obs	Mean	
Male adult	175	0.326	1,492	0.188	1,667
Female adult	144	0.972	281	0.626	425
Male child	5	1	59	0.407	64
Female child	1	1	44	0.613	45
Total observations	325		1,876		2,201

### Table 3: Estimates of aggregate parameters

*Table:* Differences in survival rates, stratification weights, and estimates of parameters

Strata	Differences in Survival Rates	$\text{Weight}_{k, ATE}$	$\text{Weight}_{k, ATT}$	$\text{Weight}_{k, ATU}$
Male adult	0.138	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
<b>No stratification</b>		<b>Stratification weighted estimates</b>		
	$\widehat{\text{SDO}}$	$\widehat{\text{ATE}}$	$\widehat{\text{ATT}}$	$\widehat{\text{ATU}}$
<b>Estimated coefficient</b>	0.35	0.20	0.24	0.19

## Drop the one female child

- We were able to estimate all three causal effect parameters because for all four strata there were units in both treatment and control
- But if we dropped the only female child in first class from the data, we'd be in trouble bc there wouldn't be any way to calculate a difference for that group
- But what could we identify?

Table: Counts and Titanic survival rates by strata and first class status.

Strata	First class		All other classes		Total
	Obs	Mean	Obs	Mean	
Male adult	175	0.326	1,492	0.188	1,667
Female adult	144	0.972	281	0.626	425
Male child	5	1	59	0.407	64
Female child	0	n/a	44	0.613	44
Total observations	324		1,876		2,200

ATT is the only one we can get

$$\begin{aligned}\hat{\delta}_{ATT} &= (0.137 \times 0.54) + (0.346 \times 0.44) + (0.593 \times 0.02) \\ &= 0.24 \text{ or } 24 \text{ percentage points}\end{aligned}\tag{1}$$

*Table:* Differences in survival rates, stratification weights, and estimates of parameters without perfect stratification

Strata	Differences in Survival Rates	$\text{Weight}_{k, \text{ATE}}$	$\text{Weight}_{k, \text{ATT}}$	$\text{Weight}_{k, \text{ATU}}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	n/a	n/a	n/a	0.02

	No stratification	Stratification weighted estimates		
	$\widehat{\text{SDO}}$	$\widehat{\text{ATE}}$	$\widehat{\text{ATT}}$	$\widehat{\text{ATU}}$
<b>Estimated coefficient</b>	0.35	n/a	0.24	n/a

Differences in survival rates, stratification weights, and estimated parameters. All coefficients should be multiplied by 100 to get a percentage point change in survival rate as a result of having a first class cabin. Note that the SDO is a simple difference in mean outcomes and therefore *not* a weighted average over the strata differences. But the estimated ATE, ATT and ATU parameters are weighted averages in difference in means using corresponding stratification weights.

## Why did this happen?

- Stratification requires having units in both groups for every value of  $X$  to get ATE
- If you want the ATT, you have to have units in the control group for every treated group based on its value of  $X$  (female children weren't treated, so didn't matter)
- If you want the ATU, you have to have units in the treatment group for every treated group based on its value of  $X$  (female children weren't treated, so did matter)
- This has a technical word we are going to learn more about called a "lack of common support"

# Curse of Dimensionality

- Stratification methods break down in finite samples because as increase the number of covariates, the "dimension" grows even faster – dimensions and covariates aren't the same in other words
- Assume we have  $k$  covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low, medium, high, etc.)
- The number of strata is  $3^k$ . For  $k = 10$ , then it's  $3^{10} = 59,049$
- The curse of dimensionality is based on the slices of all interactions of the covariates, not just the covariates, and that explodes fast

## Empty cells

- When some slice is empty, it means there's no one in that cell – so maybe you don't have any female children in first class, but you do have them in other cabins
- When you have empty cells, you can't match within that dimension of the data, and so the "curse of dimensionality" is one of the reasons why the kitchen sink approach breaks down so fast
- Matching methods really force us to see these curses; they're often hidden from OLS because OLS (as we'll see) overcomes the problem by just assuming a linearity (it doesn't match; it extrapolates)

## No overlap (common support violation)



No support is like **an incomplete bridge** which stops you from even being able to cross the moat even though the troops exist (i.e., unconfoundedness)

# Exact matching



## Exact Matching

Matching goes back at least to Rubin's early work on the propensity score, but we will start with nearest neighbor matching as there's ideas there we draw upon later with synth I want to emphasize

Matching will match a treated unit to a comparison unit that is identical on the known and quantified confounders

If we can't find one, it means common support failed, the estimate would need to use nearest neighbors (with matching bias), which we will discuss

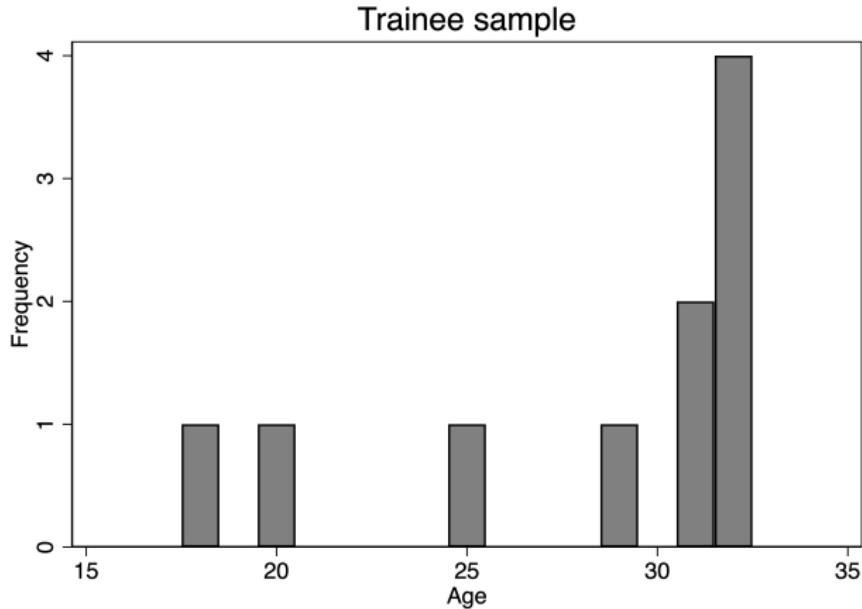
## Training example (unmatched)

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	31	\$ 26,629	1	29	\$ 23,178
2	31	\$ 26,633	2	39	\$ 33,817
3	18	\$ 15,324	3	33	\$ 27,061
4	32	\$ 27,717	4	46	\$ 43,109
5	32	\$ 27,725	5	32	\$ 26,040
6	25	\$ 20,762	6	39	\$ 33,815
7	32	\$ 27,716	7	31	\$ 25,052
8	32	\$ 27,719	8	33	\$ 27,060
9	20	\$ 16,723	9	25	\$ 19,787
10	29	\$ 24,552	10	29	\$ 23,173
			11	27	21,416
			12	32	26,040
			13	20	16,246
			14	41	36,316
			15	18	15,046
			16	29	23,178
			17	49	47,559
			18	32	26,040
			19	27	21,418
			20	46	43,109
Mean		28.2	\$24,150	Mean	32.85
					\$27,923

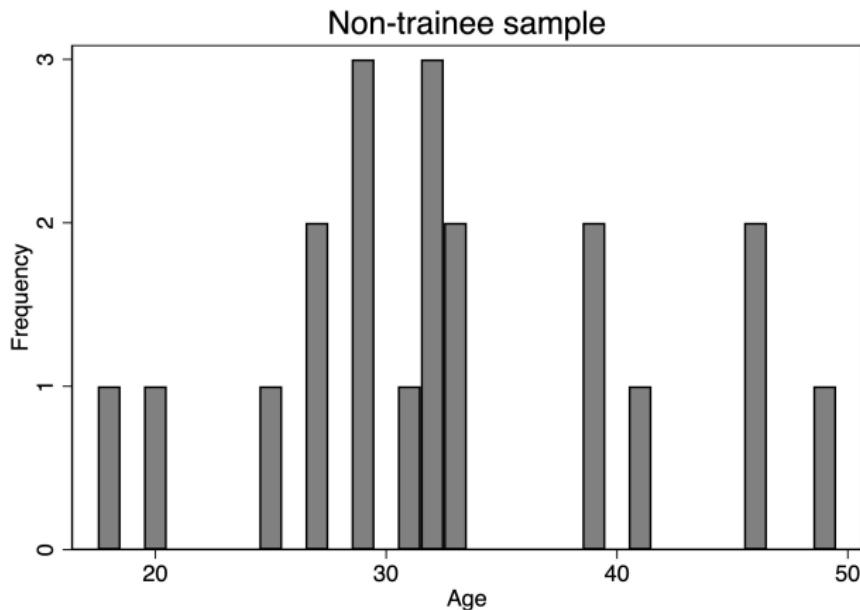
$$SDO = \$24,150 - 27,923 = -\$3,773$$

# Age Imbalance

*Figure:* Age distribution of a job training program's trainees (figure a) versus a sample of workers who were not enrolled in the trainee program (figure b).



# Age Imbalance



## Exact matching

- Exact matching finds a person in the control group whose value of  $X_j$  is exactly equal to each person in the treatment group  $i$
- Will not work if the conditioning set includes a continuous variable
- Will also not work if  $K$  gets large (curse of dimensionality we discuss later)

## ATT estimator

We will focus on the ATT for the rest of today and the equation is:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) \quad (2)$$

where  $Y_{j(i)}$  is the  $j^{\text{th}}$  unit matched to the  $i^{\text{th}}$  unit based on the  $j^{\text{th}}$  being "exactly equal to" the  $i^{\text{th}}$  unit with respect to the  $X$  conditioning set

## Number of matches

What if I find two or more  $M$  units with the identical  $X$  value? Then what?

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \left[ \frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right) \quad (3)$$

Notice that we are only dealing with  $Y_i^0$  by matching; The  $Y_i^1$  is fine as is.

## Matching algorithm

1. For each unit  $i$  in the treatment group with known and quantified confounder  $X = x_i$ , find all units  $j$  in the donor pool for whom  $x_i = x_j$ . These  $j$  units are our  $M$  matches and  $M$  can be one or it can be greater than one if you want it to be.
2. For each unit  $i$ , replace its missing potential outcome,  $Y_i^0$ , with the matched  $j$  units' realized outcomes,  $\frac{1}{M} \sum Y_{j(i)}$ , from Step 1. Do this for all  $i$  units in the treatment group.
3. For each unit  $i$ , calculate the difference between realized earnings and matched earnings,  $\hat{\delta}_i = Y_i - \frac{1}{M} \sum Y_{j(i)}$ .
4. Finally, estimate the sample ATT by averaging over all  $i$  differences in earnings from Step 3 as  $\frac{1}{N_T} \sum \hat{\delta}_i$ , where  $N_T$  is the number of treatment units.

## Matched sample

Table: Training example with matched sample using exact matching

Trainees			Matched Sample		
Unit	Age	Earnings	Matched Unit	Age	Earnings
1	31	\$26,693	2	31	\$25,052
2	31	\$26,691	2	31	\$25,052
3	18	\$15,392	18	18	\$15,046
4	32	\$27,776	5	32	\$26,045
5	32	\$27,779	5	32	\$26,045
6	25	\$20,821	4	25	\$19,787
7	32	\$27,778	5	32	\$26,045
8	32	\$27,780	5	32	\$26,045
9	20	\$16,781	8	20	\$16,246
10	29	\$24,610	6	29	\$23,178
Mean	28.2	\$24,210	Mean	28.2	\$22,854

## Estimated ATT using Exact Matching

Weak unconfoundedness of  $Y^0$  with respect to age justified substituting one group for another

But matching bias still exists if you fail common support – unconfoundedness is necessary but not sufficient

Even weak support is rare due to the curse of dimensionality

# Inexact matching



# Curse of Dimensionality

- If no matches can be found, it means many cells may contain either only treatment units or only control units but not both, and that violates our common support assumption
- We can always use “finer” classifications, but finer cells worsens the dimensional problem, so we don’t gain much from that. ex: using 10 variables and 5 categories for each, we get  $5^{10} = 9,765,625$ .
- Matching methods really force us to see these curses; they’re often hidden from OLS because OLS uses extrapolations based off functional form

## To Look Like Someone Else

- When we can make synthetic xerox copies of ourselves, that's exact matching
- But what if we can only make similar copies of ourselves, like fraternal, but not identical, twins? That's nearest neighbor matching – a form of "inexact matching", sort of like fraternal twins
- Introduces bias bc of inexact matching, but the magnitude of the bias depends on the severity of the discrepancy
- We can improve on nearest neighbor matching using bias adjustment (Abadie and Imbens 2011)

# Nearest Neighbor Matching

- Estimate  $\widehat{\delta}_{ATT}$  by *imputing* the missing potential outcome of each treatment unit  $i$  using the observed outcome from that outcome's "nearest" neighbor  $j$  in the control set using  $X$  for the matching

$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where  $Y_{j(i)}$  is the observed outcome of a control unit such that  $X_{j(i)}$  is the **closest** value to  $X_i$  among all of the control observations (eg match on  $X$ )

# Matching

- We could also use the average observed outcome over  $M$  closest matches:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \left[ \frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

- Works well when we can find good matches for each treatment group unit, so  $M$  is usually defined to be small (i.e.,  $M = 1$  or  $M = 2$ )

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_i$	$X_i$
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_i$	$X_i$
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

Match and plug in!

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_I$	$X_i$
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

$$\widehat{\delta}_{ATT} = \frac{1}{3} \cdot (6 - 9) + \frac{1}{3} \cdot (1 - 0) + \frac{1}{3} \cdot (0 - 9) = -3.7$$

## Measuring the matching discrepancy

- What does it mean to be close when I am working with a large number of covariates?
- What if we had a way of measuring a match in terms of how “close” each unit’s  $X_i$  value was to the matched  $X_j$
- Let’s do that and use the square root of the sum of all squared differences in each unit’s  $X_i - X_{j(i)}$  as a measure of how bad the match is
- This is called the Euclidean distance

# Euclidean distance

Definition: Euclidean distance

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

Let's do this together – sometimes it helps to manually calculate this

[https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY\\_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing)

# Inexact matching: Random match 1

Table 32: Matching on two covariates at random (first attempt)

Trainee sample				Non-Trainees				Matched sample #1			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	4	39	1.76	12775
2	29	2.80	12250	2	27	1.78	10075	20	48	1.87	14800
3	24	3.92	11000	3	21	1.84	8725	12	36	1.70	12100
4	27	2.29	11750	4	39	1.76	12775	8	33	1.97	11425
5	33	2.50	13250	5	38	1.61	12550	1	20	1.89	8500
6	22	1.34	10500	6	29	1.74	10525	15	43	1.45	13675
7	19	1.66	9750	7	39	1.57	12775	18	30	1.86	9000
8	20	2.60	10000	8	33	1.97	11425	7	39	1.57	12775
9	21	1.94	10250	9	24	1.81	9400	3	21	1.84	8725
10	30	3.37	12500	10	30	2.02	10750	11	33	1.64	11425
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	34.2	1.76	\$11,520

Euclidean distance: 45.8.

Estimated ATT equals \$11,075 - \$11,520 = -\$445.

# Inexact matching: Random match 2

Table 33: Matching on two covariates at random (second attempt)

Trainee sample				Non-Trainees				Matched sample #2			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	13	22	1.66	8950
2	29	2.80	12250	2	27	1.78	10075	5	38	1.61	12550
3	24	3.92	11000	3	21	1.84	8725	1	20	1.89	8500
4	27	2.29	11750	4	39	1.76	12775	20	48	1.87	14800
5	33	2.50	13250	5	38	1.61	12550	15	43	1.45	13675
6	22	1.34	10500	6	29	1.74	10525	9	24	1.81	9400
7	19	1.66	9750	7	39	1.57	12775	6	29	1.74	10525
8	20	2.60	10000	8	33	1.97	11425	17	19	1.86	8275
9	21	1.94	10250	9	24	1.81	9400	5	38	1.61	12550
10	30	3.37	12500	10	30	2.02	10750	18	30	1.86	9000
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	31	1.74	\$10,822.50

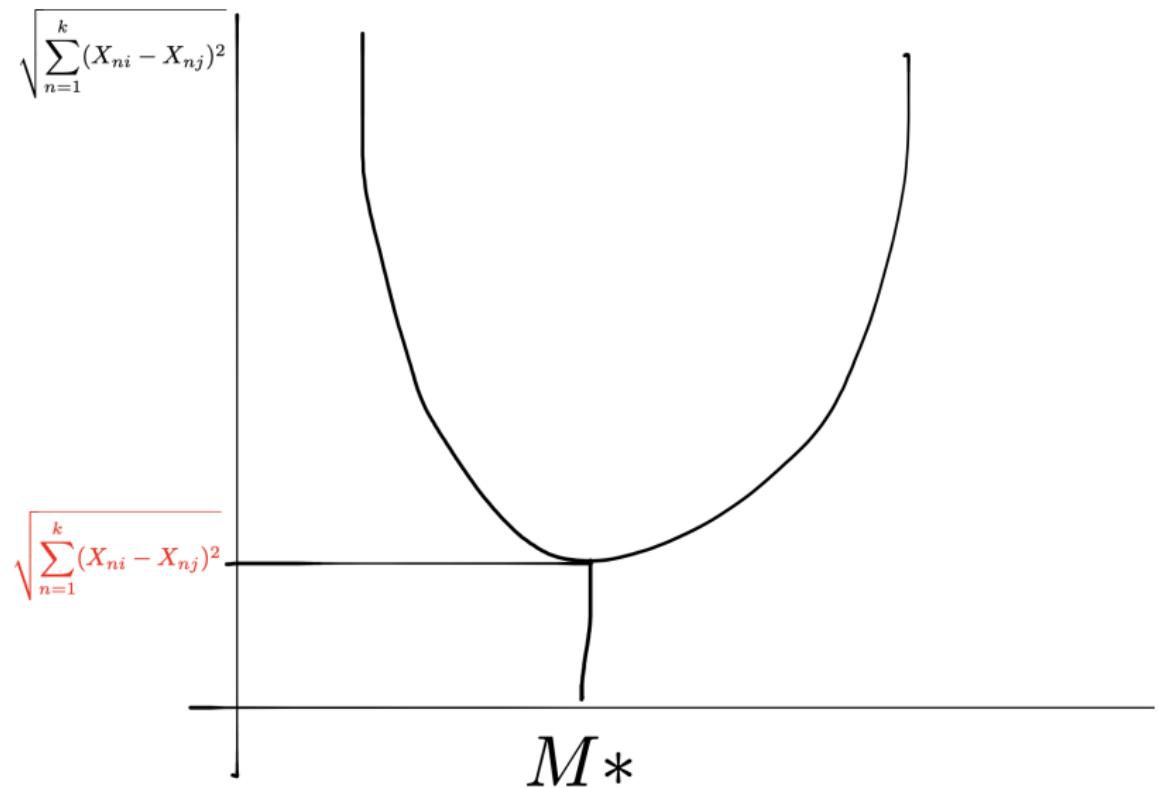
Euclidean distance: 32.53.

Estimated ATT equals \$11,075 - \$10,822.50 = \$252.50.

## Minimizing the Euclidean distance

- Abadie and Imbens (2006) show that there exists a unique solution to the matching problem that minimizes a given distance metric
- **Matching** in R and **teffects** in Stata (not sure in python)
- But the idea here is that any other match will always have a higher Euclidean distance so I've drawn a picture!

# Visualization of Optimal Match



# Inexact matching by minimizing the Euclidean distance

Table 34: Matching on two covariates with minimized Euclidian distance

Trainee sample				Non-Trainees				Optimal Match			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	14	18	1.89	8050
2	29	2.80	12250	2	27	1.78	10075	6	29	1.74	10525
3	24	3.92	11000	3	21	1.84	8725	9	24	1.81	9400
4	27	2.29	11750	4	39	1.76	12775	2	27	1.78	10075
5	33	2.50	13250	5	38	1.61	12550	8	33	1.97	11425
6	22	1.34	10500	6	29	1.74	10525	13	22	1.66	8950
7	19	1.66	9750	7	39	1.57	12775	17	19	1.86	8275
8	20	2.60	10000	8	33	1.97	11425	1	20	1.89	8500
9	21	1.94	10250	9	24	1.81	9400	3	21	1.84	8725
10	30	3.37	12500	10	30	2.02	10750	10	30	2.02	10750
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	24.3	1.85	\$9457.50

Minimized Euclidean distance: 3.00.

Estimated ATT\* equals \$11,075 - \$9457.50 = \$1,607.50.

## Other distance metrics

- Our example treated a one unit difference in age and one unit difference in GPA as the same, but those scales are different and matter a lot
- The Euclidean distance is not invariant to changes in the scale of the  $X$ 's.
- Alternative distance metrics that are invariant to changes in scale are more commonly used
- Normalized Euclidean distance and Mahalanobis distance both try to normalize it so that scale doesn't matter

# Normalized Euclidean distance

Definition: Normalized Euclidean distance

A commonly used distance is the normalized Euclidean distance:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

where

$$\hat{V}^{-1} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \dots, \hat{\sigma}_k^2)$$

## Normalized Euclidean distance

- Notice that the normalized Euclidean distance is equal to:

$$||X_i - X_j|| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}_n^2}}$$

- Thus, if there are changes in the scale of  $X_{ni}$ , these changes also affect  $\hat{\sigma}_n^2$ , and the normalized Euclidean distance does not change

# Mahalanobis distance

## Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$$

where  $\hat{\Sigma}_X$  is the sample variance-covariance matrix of  $X$ .

# Matching and the Curse of Dimensionality

- The larger the dimensions of the conditioning set, the less likely common support holds, and you can't not do it because you need these covariate dimensions to satisfy weak unconfoundedness!
- This problem is caused by the finite dataset, and it introduces a particular type of selection bias
- Curses are only overcome with new spells
- Abadie and Imbens (2011) derived a way to reduce the bias (bias adjustment or bias correction)

## Deriving the matching bias

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where each  $i$  and  $j(i)$  units are matched,  $X_i \approx X_{j(i)}$  and  $D_{j(i)} = 0$ .

Define potential outcomes and switching eq.

$$\mu^0(x) = E[Y|X = x, D = 0] = E[Y^0|X = x],$$

$$\mu^1(x) = E[Y|X = x, D = 1] = E[Y^1|X = x],$$

$$Y_i = \mu^{D_i}(X_i) + \varepsilon_i$$

# Deriving the matching bias

Substitute and distribute terms

$$\begin{aligned}\hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) \\ &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

# Deriving the matching bias

Difference between sample estimate and population parameter is:

$$\begin{aligned}\hat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Algebraic manipulation and simplification:

$$\begin{aligned}\hat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_i) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\mu^0(X_i) - \mu^0(X_{j(i)})) .\end{aligned}$$

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ .

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if  $k$  is large:

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D=1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D=1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly
- The difference  $\mu^0(X_i) - \mu^0(X_{j(i)})$  converges to zero very slowly

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly
- The difference  $\mu^0(X_i) - \mu^0(X_{j(i)})$  converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$  may not converge to zero and can be very large!

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly
- The difference  $\mu^0(X_i) - \mu^0(X_{j(i)})$  converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$  may not converge to zero and an be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$  may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly
- The difference  $\mu^0(X_i) - \mu^0(X_{j(i)})$  converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$  may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$  may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Bias is often an issue when we match in many dimensions

## Solutions to matching bias problem

The bias of the matching estimator is caused by large matching discrepancies  $\|X_i - X_{j(i)}\|$  which is virtually guaranteed by the curse of dimensionality. However:

1. But the matching discrepancies are observed. We can always check in the data how well we're matching the covariates.
2. For  $\widehat{\delta}_{ATT}$  we can sometimes make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making  $N_C$  large).
3. If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques

## Matching with bias correction

- Each treated observation contributes

$$\mu^0(X_i) - \mu^0(X_{j(i)})$$

to the bias.

- Bias-corrected (BC) matching:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[ (Y_i - Y_{j(i)}) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})) \right]$$

where  $\widehat{\mu^0}(X)$  is an estimate of  $E[Y|X = x, D = 0]$ . For example using OLS but other maybe too (neural nets?).

- Under some conditions, the bias correction eliminates the bias of the matching estimator without affecting the estimator's variance.

## Steps

1. Regress  $Y$  on  $X$  with OLS except only use the control sample:

$$Y_j = \alpha + \beta X_j + \varepsilon_j$$

where  $j$  are the units for which  $D_j = 0$ .

## Steps

2. Use the fitted values  $\hat{\alpha}$  and  $\hat{\beta}$  to predict  $\hat{\mu}^0(X)$  for both the  $i$  and the matched  $j(i)$  units:

$$\hat{\mu}_i^0 = \hat{\alpha} + \hat{\beta}X_i$$

$$\hat{\mu}_{j(i)}^0 = \hat{\alpha} + \hat{\beta}X_{j(i)}$$

## Steps

3. Subtract  $\hat{\mu}_i^0(X_i) - \hat{\mu}_{j(i)}^0(X_{j(i)})$ , our estimate of the selection bias caused by matching discrepancies, from the sample estimate of the *ATT*:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[ (Y_i - Y_{j(i)}) - (\hat{\mu}^0(X_i) - \hat{\mu}^0(X_{j(i)})) \right]$$

# Steps

4. Estimate Abadie-Imbens robust standard error (Abadie and Imbens 2006; 2008; 2011)

## Bias adjustment in matched data

---

unit	Potential Outcome		$D_i$	$X_i$
	under Treatment	under Control		
$i$	$Y_i^1$	$Y_i^0$		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

## Bias adjustment in matched data

---

unit	Potential Outcome			$X_i$
	under Treatment	under Control	$D_i$	
$i$	$Y_i^1$	$Y_i^0$		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate  $\widehat{\mu}^0(X) = \hat{\beta}_0 + \hat{\beta}_1 X = 2 + X$

## Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		$D_i$	$X_i$
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate  $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{(10 - 8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4 - 1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3} \\ &+ \frac{(10 - 9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33\end{aligned}$$

## Matching bias: Implications for practice

*Matching* bias arises because of the effect of large matching discrepancies on  $\mu^0(X_i) - \mu^0(X_{j(i)})$  due to a lack of common support. To minimize matching discrepancies:

1. Use a small  $M$  (e.g.,  $M = 1$ ). Larger values of  $M$  produce large matching discrepancies.
2. Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement.
3. Try to match covariates with a large effect on  $\mu^0(\cdot)$  particularly well.

## Large sample distribution for matching estimators

- Cannot use the bootstrap, so Abadie and Imbens derived the variance (Abadie and Imbens 2008)
- Matching estimators have a Normal distribution in large samples (provided the bias is small):

$$\sqrt{N_T}(\widehat{\delta}_{ATT} - \delta_{ATT}) \xrightarrow{d} N(0, \sigma_{ATT}^2)$$

- For matching without replacement, the “usual” variance estimator:

$$\widehat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2,$$

is valid.

# Large sample distribution for matching estimators

- For matching with replacement:

$$\begin{aligned}\widehat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2 \\ &+ \frac{1}{N_T} \sum_{D_i=0} \left( \frac{K_i(K_i-1)}{M^2} \right) \widehat{var}(\varepsilon|X_i, D_i = 0)\end{aligned}$$

where  $K_i$  is the number of times observation  $i$  is used as a match.

- $\widehat{var}(Y_i|X_i, D_i = 0)$  can be estimated also by matching. For example, take two observations with  $D_i = D_j = 0$  and  $X_i \approx X_j$ , then

$$\widehat{var}(Y_i|X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of  $\widehat{var}(\varepsilon_i|X_i, D_i = 0)$

## Curse of dimensionality, bias and heterogeneous treatment effects

- Recall the problem of many covariates for exact matching – the curse of dimensionality makes matching on  $K$  covariates implausible as the dimensions grow exponentially with  $K$
- This is problem because recall there are two assumptions needed to match
  1. Unconfoundedness: this gives you the right to match
  2. Common support: this gives you the ability to match
- Without both, then depending on the amount of heterogeneity in the treatment effects, matching will be biased

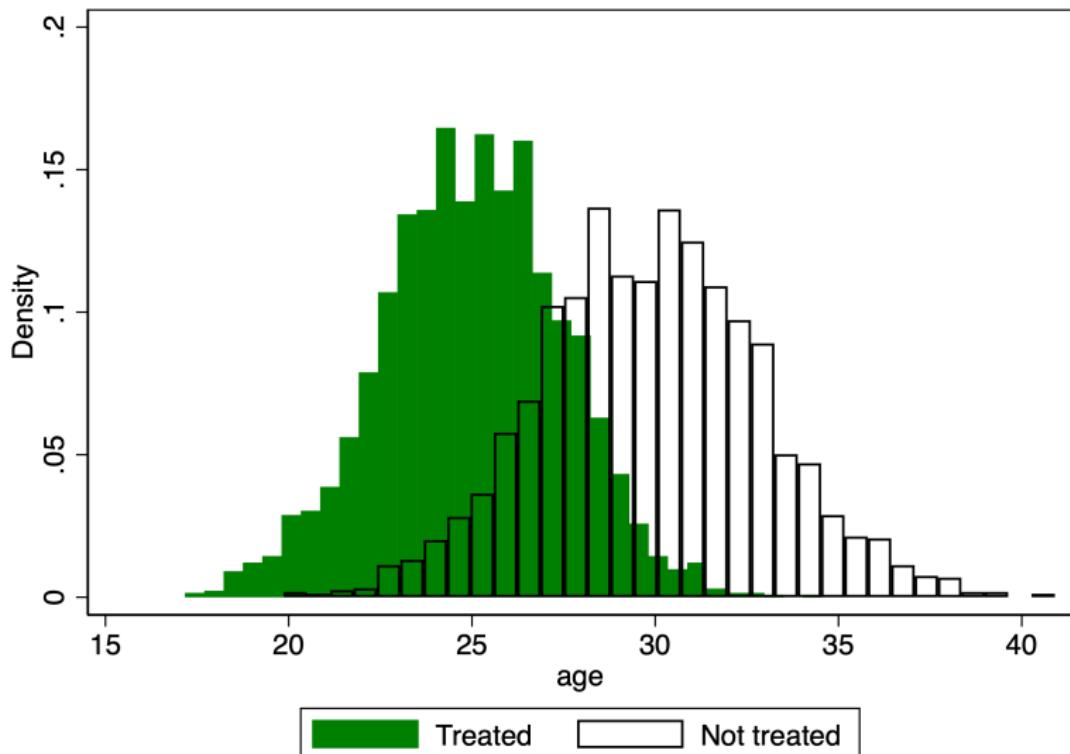
## Propensity score as dimension reduction

- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method
- They show that if treatment is independent of potential outcomes conditional on  $K$  covariates, then it will be independent of potential outcomes conditional on propensity score
- Main value of the propensity score is dimension reduction to reduce  $K$  covariates into a single scalar without loss of information
- Variety of ways to incorporate the propensity score – stratification, weighting and matching

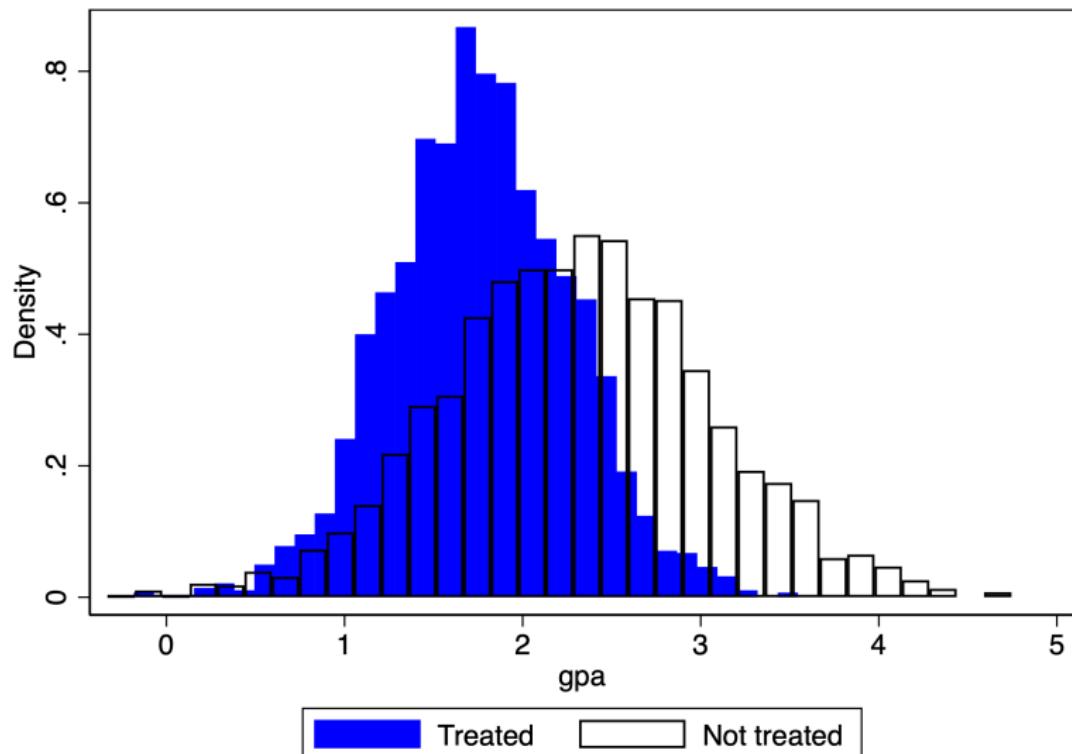
## Investigating overlap

- Once you obtain the propensity score, you can use it for estimation, but you can also use it for evaluating covariate balance
- It's an easy way to do it with histograms, and since the propensity score theorem holds for the dimensions of  $X$ , there's no loss of generality in investigating overlap that way versus one by one
- Remember: you need common support, not on  $X$  individual covariates alone, but within the  $K$  dimensions, so investigating with the propensity score is easier to do

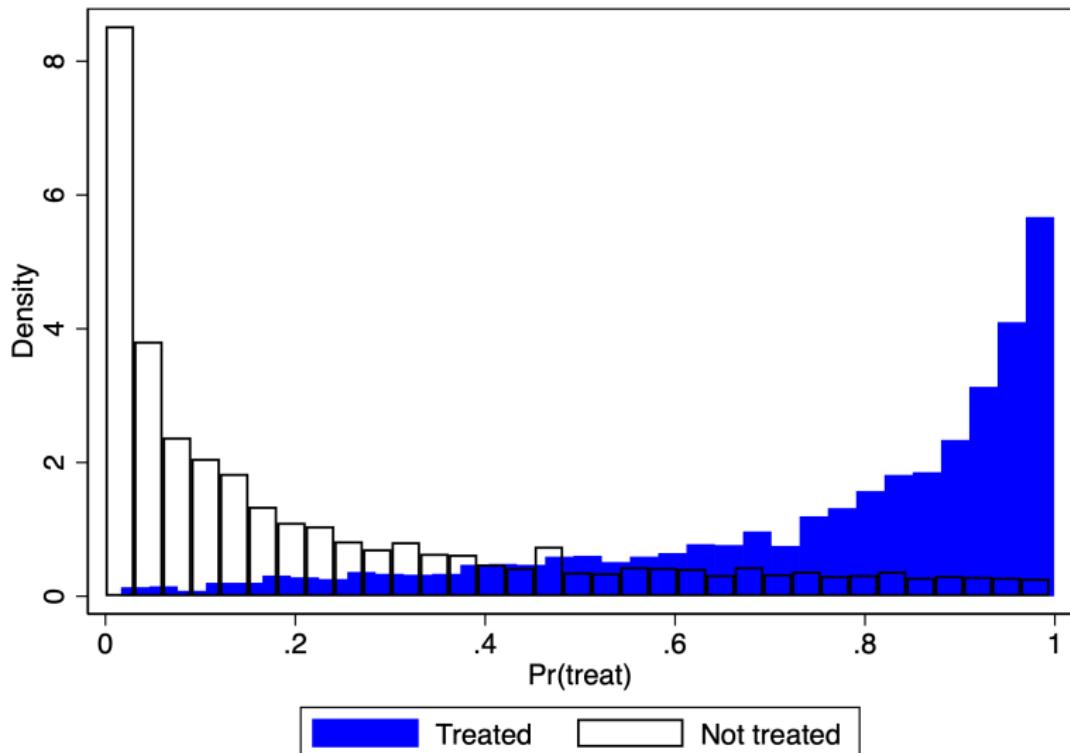
# Covariate 1 histograms



## Covariate 2 histograms



## Summarizing both with propensity score histogram



# Navigating the vastness of estimation

- Estimators abound and can be a little bewildering so to summarize them:
  1. Match units from one group to another using the propensity score (with various rules for finding how close to be)
  2. Weighting by the inverse propensity score
- Variety of techniques to derive standard errors from parametric methods to bootstrapping
- Can even introduce “doubly robust” methods to deal with matching bias like we did with nearest neighbor bias correction
- But all these estimators assume unconfoundedness, so really it’s all about addressing the lack of overlap; that’s the only bias that exists when you have unconfoundedness remember

# Formal Definition

## Definition of Propensity score

A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables:  $p(X) = Pr(D = 1|X)$

## Propensity score theorem

"We are interested in estimating the average effect of a binary treatment on a scalar outcome. If assignment to the treatment is exogenous or unconfounded, that is, independent of the potential outcomes given covariates, biases associated with simple treatment-control average comparisons can be removed by adjusting for differences in the covariates. Rosenbaum and Rubin (1983) show that adjusting solely for differences between treated and control units in the propensity score removes all biases associated with differences in covariates." – Kirano, Imbens and Ridder (2003, Econometrica)

# Propensity score theorem

## Propensity score theorem

If  $(Y^1, Y^0) \perp\!\!\!\perp D|X$  (full unconfoundedness), then  $(Y^1, Y^0) \perp\!\!\!\perp D|\rho(X)$  where  $\rho(X) = Pr(D = 1|X)$ , the propensity score

- Conditioning on the propensity score is enough to have independence between  $D$  and  $(Y^1, Y^0)$  (Rosenbaum and Rubin 1983)
- With full unconfoundedness, you can estimate the ATE, but you can estimate the ATT with weaker assumptions (independence with respect to  $(Y^0) \perp\!\!\!\perp D|\rho(X)$ )
- But remember – propensity scores don't solve unconfoundedness; if you don't have it with respect to  $K$  dimensions of  $X$ , you won't have it when you collapse it to the propensity score

# True vs Estimating Propensity Score

- Outside the randomized experiment, we don't know the "true propensity score"
  - We use DAGs or hunches to select covariates, careful not to include outcomes or colliders
  - But we still won't know the functional form (i.e., logit, probit, polynomials, interactions)
- Often then people will use higher order terms and interactions to provide flexibility but it does technically have misspecification biases
- This is an area most likely where machine learning could greatly enhance the estimation (take your pick)

## Step 1: Pick your parameter ATE vs ATT vs ATU

- Which population are you studying? Only those who were discriminated against? That's the ATT
- Do you want to imagine “what if blacks and whites were both discriminated against?” That's the ATE
- The more you can focus on one particular causal parameter, the easier and more justified it gets as it weakens both assumptions

## Step 2: Estimate the propensity score

- Estimate the conditional probability of treatment using probit or logit model (or ML)

$$Pr(D_i = 1|X_i) = F(\beta X_i)$$

- Note: don't use OLS because while it will get the mean right, it will not get correct values in the tails because of its linear projections
- OLS will give propensity scores outside the [0,1] bounds and probabilities cannot be negative or greater than one

## Step 2: Estimate the propensity score

- Use the estimated coefficients to predict the propensity score for each unit  $i$

$$\hat{\rho}_i(X_i) = \hat{\beta}X_i$$

- Note that each unit  $i$  now has a predicted probability of treatment given the values of their covariates relative to everyone else's

## Step 2: Estimate the propensity score

- Think of the propensity score as a frequentist concept of probability
- "If I drew someone from the sample with these characteristics, then how many of those are in the treatment group divided by the total with those characteristics"
- Or for each dimension of  $X$ , a ratio of  $\frac{N_T}{(N_T+N_C)}$

## Step 3a: Estimation with matching

- Most common method is to use matching
- Matching finds a unit in the comparison group with a similar  $\hat{\rho}_i(X)$  to service as counterfactual for the unit
- For the ATE, you'll need matches on both side; for the ATT, you'll need matches for the treatment group among controls
- Lack of overlap creates issues for matching which we'll note later

## Step 3a: Estimation with stratification

- Rare to see this done anymore, though it was one of the methods that Dehejia and Wahba (2002) tried
- Stratification is a kind of weighting method similar to Cochran's subclassification method where weights are group shares within certain ranges of the propensity score
- Four steps to doing this; I'll review again later, but let me briefly do it now

## Step 3a: Estimation with stratification

1. First, split the sample by ranges on the propensity score until you find covariate balance within each range
2. Next calculate weights for each region depend on the causal parameter you're seeking:
  - ATE: number of units in that region divided by total units in sample ;
  - ATT: number of treated units in that region divided by the number of treated units in the sample
  - ATU: number of control units in that region divided by the number of control units in the sample
3. Calculate simple difference in mean outcomes (i.e., without controls) for each propensity score region (Step 1)
4. Then take weighted average using weights from step 2 and difference in means from step 3

## Step 3b: Early weighting methods

- Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura, Smith, and Todd (1998) focus on the ATT
- Estimators based on local linear regressions of the outcome on treatment status and either covariates or the propensity score.
- They conclude that in general there is no clear ranking of their estimators
- Under some conditions the estimator based on adjustment for all covariates is superior to the estimator based on adjustment for the propensity score,
- Under other conditions the second estimator is to be preferred and lack of knowledge of the propensity score does not alter this conclusion

## Step 3b: IPW by Hirano, Imbens and Ridder

- Hirano, Imbens and Ridder (2003, *Econometrica*) considers estimation of the ATT
- Focus is on the efficient estimation and they show that while weighting on the inverse of the true propensity score is not an efficient estimator, paradoxically weighting each observation by the *inverse* of a nonparametric estimator of the propensity score is

## Step 3b: Estimation with inverse probability weighting

- IPW uses the estimated propensity score to reweight the outcomes for which there are several historical methods for doing so
- IPW is non-parametric – you are just taking averages and multiplying by the inverse of the propensity score weights depending on which parameter you want to estimate
- There are fewer implementation choices than in matching (i.e., no choice over distance, number of neighbors)
- There are bias adjustment methods called double robust where you combine imputing counterfactuals with weighting by the propensity score

## Step 3b: Estimation of ATT with IPW

### Estimating ATT with IPW

Given  $Y^0 \perp\!\!\!\perp D|X$  and weak common support, then

$$\begin{aligned}\delta_{ATT} &= E[Y^1 - Y^0|D = 1] \\ &= \frac{1}{Pr(D = 1)} \cdot E \left[ Y \cdot \frac{D - \rho(X)}{1 - \rho(X)} \right]\end{aligned}$$

Notice that when  $D = 1$ , the outcome is not weighted, but when  $D = 0$  it is. You're missing the  $Y^0$  for the treatment, not  $Y^1$  so you weight the treatment group  $Y$  values alone and weight "up" or "down" the comparison groups by their propensity scores

## Step 4: Standard Errors

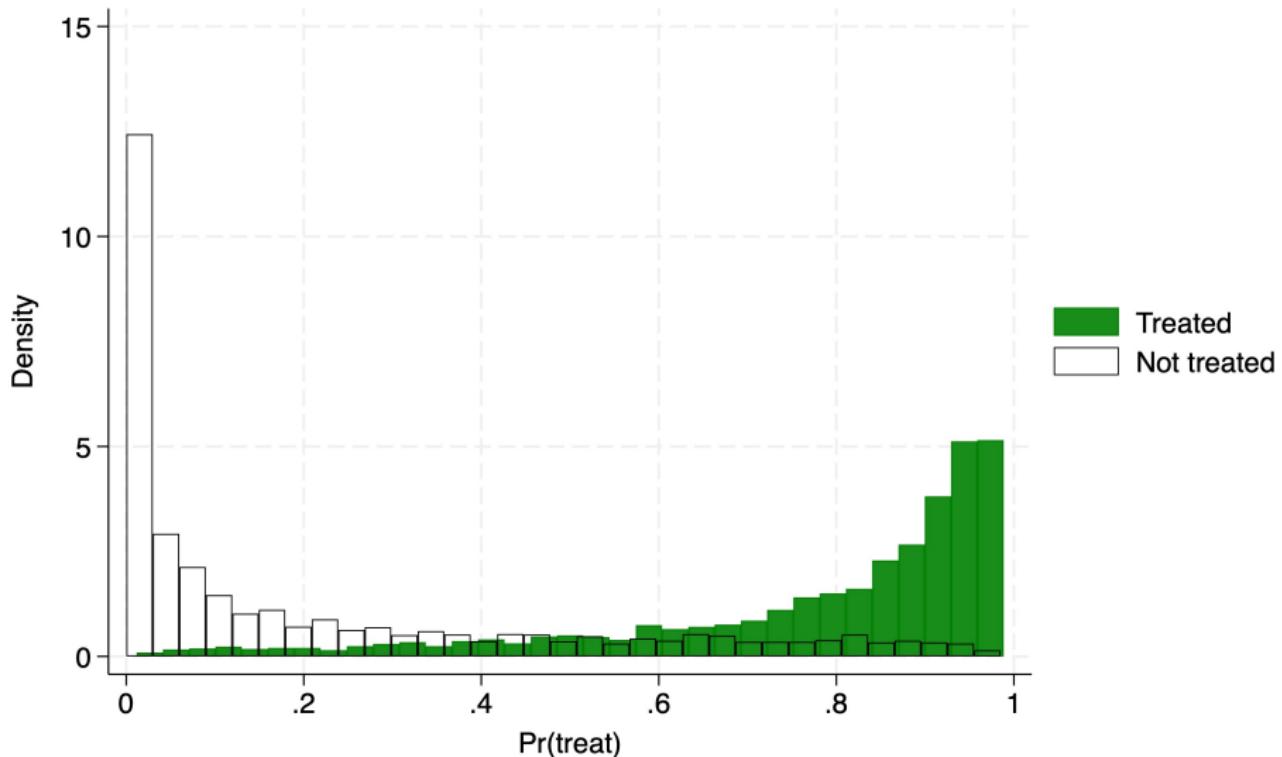
Standard errors can be constructed a few different ways:

- We need to adjust the standard errors for first-step estimation of  $\rho(X)$ 
  - Parameteric first step: Newey and McFadden (1994)
  - Non-parametric first step: Newey (1994)
- IPW is a smooth estimator which means the bootstrap is valid for inference (Adudumilli 2018 and Bodory et al. 2020)

## Check for common support using histograms

- Assessing whether there are units in both groups for whichever parameter you're focused on is simple with propensity score as shown earlier using histograms of the propensity score for treated and control
- Crump, et al. (2009) suggest keeping propensity scores within the interval [0.1,0.9] ("trimming") but any trimming will drop units and dropping units means moving away from the parameter
- Let's look at a picture again just to remind ourselves

# Assessing overlap



# Estimating ATE with IPW

- Any of the parameters including the ATE are possible with propensity scores
- Assumptions are strongest for ATE: full unconfoundedness and full support versus weak versions of both
- You should only focus on the ATE if you care about the entire population's treatment effects, but remember that requires more assumptions than if you just focused on the ATT

# Estimating ATE with IPW

## Estimating ATE with IPW

Given  $Y^1, Y^0 \perp\!\!\!\perp D|X$  and common support, then

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - \rho(X)}{\rho(X) \cdot (1 - \rho(X))}\right]\end{aligned}$$

Notice that since treated units are missing counterfactuals, but so are controls, all of the data is weighted for the ATE (not just the controls for ATT)

# Inverse Probability Weighting

Proof.

$$\begin{aligned} E \left[ Y \cdot \frac{D - \rho(X)}{\rho(X)(1 - \rho(X))} \middle| X \right] &= E \left[ \frac{Y}{\rho(X)} \middle| X, D = 1 \right] \rho(X) \\ &\quad + E \left[ \frac{-Y}{1 - \rho(X)} \middle| X, D = 0 \right] (1 - \rho(X)) \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

and the results follow from integrating over  $P(X)$  and  $P(X|D = 1)$ . □

## Other comments about propensity scores

- Only other comments to make is that you can get outlier weights
- If the propensity score is very high for a control group in the ATT, the weight can explode
- Certain normalizations of the way the IPW is constructed to try and minimize those influences
- Stata's `teffects` or R's `ipw` both let you estimate parameters using IPW and get standard errors
- But you can do it manually using OLS and use the propensity scores as analytical weights

## Double robust estimators

- You can have the right covariates but the wrong model and unbiasedness requires the correct model
- What if you had a way to control for the covariates using propensity scores and something else like regression?
- Buys you some insurance against model misspecification if such a thing existed

# Double robust estimators

- Question: Is it possible to combine the virtues of regression and propensity score?
- Answer: Double robust estimators
- DR address the model misspecification problem by combining propensity scores with other methods
- Basic idea in all of them was to control for covariates twice *at the same time* without paying for it (two for one)

# Double robust estimators

- We say that estimators combining regression with IPW are double robust so long as
- The regression for the outcome is properly specified, or
- The propensity score is properly specified
- We give ourselves two chances to get it right (either/or not both/and) but if neither is properly specified, then you didn't really gain much
- Three strikes in baseball

# Doubly Robust Estimator

Easy to verify: with true  $m_1(X)$  (an outcome regression or OR model) and  $\rho(X)$

$$\begin{aligned} ATE &= \mathbb{E} \left[ \frac{DY}{\rho(X)} - \frac{D - \rho(X)}{\rho(X)} m_1(X) \right] - \mathbb{E} \left[ \frac{(1 - D)Y}{1 - \rho(X)} + \frac{D - \rho(X)}{1 - \rho(X)} m_0(X) \right] \\ &= \mathbb{E} \left[ m_1(X_i) + \frac{D_i \{Y_i - m_1(X_i)\}}{\rho(X_i)} \right] - \mathbb{E} \left[ m_0(X_i) + \frac{(1 - D_i) \{Y_i - m_0(X_i)\}}{1 - \rho(X_i)} \right] \\ &= \mu_1 - \mu_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \end{aligned}$$

## Doubly Robust Estimator

In the previous formula, replace the true PS  $\rho(X)$  and outcome  $m_1(X)$  by the estimated ones from postulated models  $\hat{\rho}(X)$  and  $\hat{m}_1(X)$ , we obtain two augmented estimators:

$$\widehat{\tau}_{dr} = \hat{\mu}_{1,dr} - \hat{\mu}_{0,dr}$$

$$= \frac{1}{N} \sum_{i=1}^N \left[ \frac{D_i Y_i - D_i - \hat{\rho}(X_i)}{\hat{\rho}(X_i)} \hat{m}_1(X_i) \right] - \frac{1}{N} \sum_{i=1}^N \left[ \frac{(1 - D_i) Y_i - D_i - \hat{\rho}(X_i)}{1 - \hat{\rho}(X_i)} \hat{m}_0(X_i) \right]$$

# Double Robust Estimator

- The two estimators are mathematically the same, but they have different statistical implications:
  - the first estimator augments an IPW estimator by outcome regression (OR);
  - the second augments an OR estimator by IPW.
- The first estimator is usually referred to as the “doubly-robust (DR) estimator” and we saw it in Sant’Anna and Zhou (2020) actually for diff-in-diff

# Doubly Robust Estimator

Now focus on the DR estimator

Notice that  $\hat{\mu}_{1,dr}$  and  $\hat{\mu}_{0,dr}$  have the same structure

$$\hat{\mu}_{1,dr} = N^{-1} \sum_{i=1}^N \left[ \frac{D_i Y_i}{\hat{\rho}_i} - \frac{(D_i - \hat{\rho}_i \hat{\rho}_i) \hat{m}_1}{\hat{\rho}_i} \right]$$

$$\hat{\mu}_{0,dr} = N^{-1} \sum_{i=1}^N \left[ \frac{(1 - D_i) Y_i}{1 - \hat{\rho}_i} + \frac{(D_i - \hat{\rho}_i) \hat{m}_0}{1 - \hat{\rho}_i} \right]$$

We will study the properties of  $\hat{\mu}_{1,dr}$  as an estimator for  $\mu_1 = \mathbb{E}\{Y(1)\}$ ;  
 $\hat{\mu}_{0,dr}$  is symmetric

# Doubly Robust Estimator

What does  $\hat{\mu}_{1,dr}$  estimate? Simple algebra shows that  $\mu_{1,dr}$  converges to

$$\begin{aligned} & \mathbb{E} \left[ \frac{DY}{\rho(X)} - \frac{D - \rho(X)}{\rho(X)} m(X) \right] \\ &= \mathbb{E}\{Y(1)\} + \mathbb{E} \left[ \frac{D - \rho(X)}{\rho(X)} \{Y(1) - m(X)\} \right] \end{aligned}$$

- Thus, in order for  $\hat{\mu}_{1,dr}$  to estimate  $\mathbb{E}\{Y(1)\}$ , the second term in (4) must be 0
- Regression estimator augmented by weighted residuals
- When does the second term = 0?

# Doubly Robust Estimator

When does  $R = \mathbb{E} \left[ \frac{D - \rho(X)}{\rho(X)} \{Y(1) - m_1(X)\} \right] = 0?$

- **Scenario 1:** Postulated propensity score model  $\rho(X; \beta)$  is incorrect, but postulated regression model  $m_1(X; \alpha)$  is correct

$$R = \mathbb{E} \left[ \mathbb{E} \left[ \frac{D - \rho(X)}{\rho(X)} \{Y(1) - m_1(X)\} \middle| X \right] \right]$$

$$= \mathbb{E} \left[ \frac{\rho_{true}(X) - \rho(X)}{\rho(X)} \mathbb{E}\{Y(1) - m_1(X)\} \middle| X \right]$$

- Easy to see that  $\mathbb{E}\{Y(1) - m_1(X)|X\} = \mathbb{E}\{Y(1)|X\} - m_1(X) = m_{1,true}(X) - m_1(X) = 0$

# Doubly Robust Estimator

When does  $\mathbb{E} \left[ \frac{D - \rho(X)}{\rho(X)} \{Y(1) - m_1(X)\} \right] = 0$ ?

- **Scenario 2:** Postulated propensity score model  $\rho(X; \beta)$  is correct, but postulated regression model  $m_1(X; \alpha)$  is incorrect

$$\begin{aligned} R &= \mathbb{E} \left[ \mathbb{E} \left[ \frac{D - \rho(X)}{\rho(X)} \{Y(1) - m_1(X)\} \middle| X \right] \right] \\ &= \mathbb{E} \left[ \frac{\hat{e}_{true}(X) - \rho(X)}{\rho(X)} \mathbb{E}\{Y(1) - m_1(X)\} \middle| X \right] \end{aligned}$$

- Equals to zero because  $\hat{e}_{true}(X) - \rho(X) = 0$

## Doubly Robust Estimator

- In both cases, the second term goes to 0 in large samples, and thus  $\hat{\mu}_{1,dr}$  is consistent (asymptotically unbiased) for  $E(Y(1))$ .
- Similarly,  $\hat{\mu}_{0,dr}$  is consistent for  $E(Y(0))$ , and hence  $\hat{\tau}_{dr}$  is consistent for the ATE.
- Obviously, if both models are correct,  $\hat{\tau}_{dr}$  is consistent for estimating the ATE.

# Doubly Robust Estimator

**Double Robustness:**  $\hat{\tau}_{dr}$  is a consistent estimator of the ATE if either the propensity score model or the potential outcome model is, but not necessarily both are, correctly specified

- DR is a large sample property
- Offers protection against model mis-specification: gives you two chances to get it right (and wrong)!
- If  $\rho(X)$  and  $m_1(X)$  are modeled correctly,  $\hat{\tau}_{dr}$  will have smaller variance than the IPW estimator (in large samples)
- If the outcome model  $m_1(X)$  is correct,  $\hat{\tau}_{dr}$  has larger variance (in large samples) than the direct regression estimator
- ... but gives protection in the event it is not

## Propensity score matching

- Matching, or “imputation”, is another way that utilizes the  $\hat{p}_i(X_i)$
- Matching estimation based on the propensity score has the same first step as IPW, but not the second and third steps
- Common support starts to be more complex with imputation methods because you will need to decide how far away from a unit’s own propensity score is a tolerable distance to be considered a “neighbor”

## Standard matching strategy

- Pair each treatment unit  $i$  with one or more *comparable* control group unit  $j$ , where comparability is in terms of proximity, or distance, to the estimated propensity score
- Impute the unit's missing counterfactual outcome  $Y_{i(j)}$  based on the unit or units chosen in the previous step
- If more than one are “nearest neighbors”, then use the neighbors’ weighted outcomes

$$Y_{i(j)} = \sum_{j \in C(i)} w_{ij} Y_j$$

where  $C(i)$  is the set of neighbors with  $W = 0$  of the treatment unit  $i$  and  $w_{ij}$  is the weight of control group units  $j$  with  $\sum_{j \in C(i)} w_{ij} = 1$

# Imputing the counterfactuals

Let the ATT be our parameter of interest:

$$E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1]$$

We estimate it as follows

$$\widehat{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} \left[ Y_i - Y_{i(j)} \right]$$

where  $N_T$  is the number of matched treatment units in the sample.

Note the difference between *imputation* and IPW – the only weight here is  $\frac{1}{N_T}$

# Matching methods

- The probability of observing two units with exactly the same propensity score is in principle zero if  $Pr(X = x)$  is continuous
- Several matching methods have been proposed in the literature, but the most widely used are:
  - Stratification matching
  - Nearest-neighbor matching (with or without caliper)
  - Radius matching
  - Kernel matching
- Typically, one treatment unit  $i$  is matched to several control units  $j$ , but sometimes one-to-one matching is used

# Stratification

- Stratification based on the propensity score is a multi step process that bears resemblance to the stratification/subclassification method proposed by Cochran (1968)
- Method uses brute force to achieve the balancing property discussed earlier, which is then used with weighted differences in means within propensity score “strata”
- Dehejia and Wahba (2002) used stratification matching in their seminal paper

# Stratification: Achieving Balance

First create “propensity score strata” inside which you have balanced covariates

1. Sort the data by propensity score and divide into groups of observations with similar propensity scores (e.g., percentiles)
2. Within each strata, test (e.g., t-test) whether the means of the  $k$  covariates are equal between treatment and control
3. If so, then stop. If not, it means the covariates aren’t balanced *within that propensity score strata* so then divide that strata in half and repeat step 2
4. If a particular covariate is unbalanced for multiple groups, modify the initial logit or probit equation by including higher order terms and/or interactions with that covariate and repeat

## Propensity score matching

- Next we review explicit imputation based on the propensity score or what is sometimes called propensity score matching
- King and Nielsen (2019) is a critique of using propensity scores *for matching* (i.e., imputation)
- But not a critique of the propensity score itself or to stratification, regression adjustment, or IPW
- Issues raised have to do with forced balance through trimming and a myriad of other common choices made by the researcher

## Ad hoc user choices introduce bias

*"[The] more balanced the data, or the more balance it becomes by [trimming] some of the observations through matching, the more likely propensity score matching will degrade inferences." – King and Nielsen (2019)*

# Nearest Neighbor

Pretty similar to covariate matching. Formula is

$$\widehat{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} \left[ Y_i - \sum_{j \in C(i)_M} w_{ij} Y_j \right]$$

- $N_T$  is the number of treated units  $i$  and  $N_C$  is number of control units  $j$
- $w_{ij}$  is equal to  $\frac{1}{N_C}$  if  $j$  is a control unit and zero otherwise
- And unit  $j$  is chosen as a control for  $i$  if it's propensity score is nearest to that of  $i$

## NN Matching: Bias vs. Variance

How far away on the propensity score will you use is what makes some of the different types of matching proposed differ

- Matching just one nearest neighbor minimizes bias at the cost of larger variance
- Matching using additional nearest neighbors increases the bias but decreases the variance

## NN Matching: Bias vs. Variance

Matching with or without replacement

- with replacement keeps bias low at the cost of larger variance
- without replacement keeps variance low but at the cost of potential bias

## Distance between treatment and control units

- What was historically done was limiting “distance” through various *ad hoc* choices
- Imagine these choices as creating like a cowboy rope lasso that matches to everything inside that circle
- There were two common ways for creating the circle – caliper matching and radius matching.

## Caliper matching

- Caliper matching is a variation on NN matching that tries to build brakes into the algorithm as to avoid “bad neighbors” by imposing a tolerable maximum distance (e.g., 0.2 units in the propensity score away from a treatment unit  $i$ ’s propensity score)
- Note – this is a one-to-one imputation, and if there doesn’t exist anybody in the control group unit  $j$  within that “caliper”, then treatment unit  $i$  is discarded which as with all trimming changes the parameter we are estimating
- It’s difficult to know what this caliper should be *ex ante*, hence why I said it is somewhat *ad hoc*

## Radius matching

- Each treatment unit  $i$  is matched with the control group units whose propensity score are in a “predefined neighborhood” of the propensity score of the treatment unit.
- **All** the control units with  $\hat{\rho}_j(X_j)$  falling within a radius  $r$  from  $\hat{\rho}_i(X_i)$  are matched to the treatment unit  $i$  – this is what distinguishes it from calipers, and makes it more similar to covariate matching (Abadie and Imbens 2006, 2008)
- The smaller the radius, the better the quality of the matches, but the higher the possibility some treatment units are not matched because the neighborhood does not contain control group units  $j$

## Software

- You can use `-teffects`, `psmatch`- to get at these two nearest neighbor approaches by setting the number of matches
- You can use `-pscore2`- for stratification
- You can use the `MatchIt` package in R

## Failure of econometric estimators (LaLonde 1986)

- Evaluation of the Job Trainings Program (NSW) has a rich history in causal inference
- Bob LaLonde (passed away November 2015) was a Card and Ashenfelter student at Princeton whose job market paper evaluated, not NSW itself, but econometric methods one would use in something like NSW
- Dehejia and Wahba (1999; 2002) used LaLonde's data with propensity score matching and found they could recover known effects
- Critiques by Petra Todd, Jeff Smith and others followed which I'll summarize

## Summarizing LaLonde (1986)

- Very clever study that combined experimental and non-experimental data to ascertain whether popular econometric methods could recover unbiased effects when those effects were already known
- Damning conclusion – 1986 AER (it was LaLonde's JMP) found econometric methods failed to get the number right, and worse, failed to get the sign right
- Was a critical paper in the emerging “credibility crisis” within labor and helped fuel the type of work we now broadly consider to be design based causal inference

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of \$900

## LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
  - His estimates were way off and usually the wrong sign
  - Conclusion was influential in policy circles and led to greater push for more experimental evaluations

# Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
  - **Treatment group**: received all the benefits of NSW program
  - **Control group**: left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

# NSW Program

- Treatment group members were:
  - guaranteed a job for 9-18 months depending on the target group and site
  - divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
  - paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

# NSW Program

- Other details about the NSW program:
  - Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
  - Post-treatment: after their term expired, they were forced to find regular employment
  - Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

# NSW Data

- NSW data collection:
  - MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
  - Conducted up to 4 post-baseline interviews
  - Different sample sizes from study to study can be confusing, but has simple explanations

# NSW Data

- Estimation:
  - NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$SDO = \frac{1}{N_t} \sum_{D_i=1} Y_i - \frac{1}{N_c} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e.,  $(Y^0, Y^1) \perp\!\!\!\perp D$ .

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful –  
 $\approx \$900$  (LaLonde 1986) to  $\$1,800$  (Dehejia and Wahba 2002)  
depending on the sample used

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW  
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA<sup>a,b</sup>

Name of Comparison Group <sup>d</sup>	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted <sup>c</sup> (3)	Unadjusted (4)	Adjusted <sup>c</sup> (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted <sup>c</sup> (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

<sup>a</sup> The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

<sup>b</sup> Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

<sup>c</sup> The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

<sup>d</sup> See Table 3 for definitions of the comparison groups.

## Switching out the control group

- Think of \$800 to \$900 as the “ground truth” since row 1 was using the RCT
- LaLonde “drops” the experimental controls (which satisfied independence) and “replaces” it with six different draws from two nationally representative surveys (PSID and CPS)
- Now the dataset contains a negatively selected treatment group compared to a nationally representative control group
- Will selection on observable methods “work”?

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW  
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA<sup>a,b</sup>

Name of Comparison Group <sup>d</sup>	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted <sup>c</sup> (3)	Unadjusted (4)	Adjusted <sup>c</sup> (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted <sup>c</sup> (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	(\$455 (704))	(\$697 (760))	(\$509 (967))	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

<sup>a</sup> The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

<sup>b</sup> Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

<sup>c</sup> The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

<sup>d</sup> See Table 3 for definitions of the comparison groups.

# Imbalanced covariates for experimental and non-experimental samples

covariate	All		CPS	NSW		
	mean	(s.d.)	Controls	Trainees	N <sub>t</sub> = 297	t-stat
			N <sub>c</sub> = 15,992			
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

## Dehejia and Wahba (1999)

- Dehejia and Wahba (DW) update LaLonde's original study using propensity score matching
  1. Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". Journal of the American Statistical Association, vol. 94(448): 1053-1062 (pdf)
- Can propensity score matching improve over the estimators that LaLonde examined?

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
<b>NSW/Lalonde:<sup>a</sup></b>									
Treated	297	24.63 (.32)	10.38 (.09)	.80 (.02)	.09 (.01)	.73 (.02)	.17 (.02)	3,066 (236)	
Control	425	24.45 (.32)	10.19 (.08)	.80 (.02)	.11 (.02)	.81 (.02)	.16 (.02)	3,026 (252)	
<b>RE74 subset:<sup>b</sup></b>									
Treated	185	25.81 (.35)	10.35 (.10)	.84 (.02)	.059 (.01)	.71 (.02)	.19 (.02)	2,096 (237)	1,532 (156)
Control	260	25.05 (.34)	10.09 (.08)	.83 (.02)	.1 (.02)	.83 (.02)	.15 (.02)	2,107 (276)	1,267 (151)
<b>Comparison groups:<sup>c</sup></b>									
PSID-1	2,490	34.85 [.78]	12.11 [.23]	.25 [.03]	.032 [.01]	.31 [.04]	.87 [.03]	19,429 [991]	19,063 [1,002]
PSID-2	253	36.10 [1.00]	10.77 [.27]	.39 [.04]	.067 [.02]	.49 [.05]	.74 [.04]	11,027 [853]	7,569 [695]
PSID-3	128	38.25 [1.17]	10.30 [.29]	.45 [.05]	.18 [.03]	.51 [.05]	.70 [.05]	5,566 [686]	2,611 [499]
CPS-1	15,992	33.22 [.81]	12.02 [.21]	.07 [.02]	.07 [.02]	.29 [.03]	.71 [.03]	14,016 [705]	13,650 [682]
CPS-2	2,369	28.25 [.87]	11.24 [.19]	.11 [.02]	.08 [.02]	.45 [.04]	.46 [.04]	8,728 [667]	7,397 [600]
CPS-3	429	28.03 [.87]	10.23 [.23]	.21 [.03]	.14 [.03]	.60 [.04]	.51 [.04]	5,619 [552]	2,467 [288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE74 = earnings in calendar year 19x.

<sup>a</sup> NSW sample as constructed by Lalonde (1986).

<sup>b</sup> The subset of the Lalonde sample for which RE74 is available.

<sup>c</sup> Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

PSID-1 and CPS-1 are identical to those used by Lalonde. CPS2-3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

Table 2. Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparison Groups From the PSID and the CPS<sup>a</sup>

A. Lalonde's original sample										B. RE74 subsample (results do not use RE74)								C. RE74 subsample (results use RE74)												
Comparison group	NSW				NSW				NSW				NSW				NSW				NSW									
	treatment	Unrestricted differences	treatment	Unrestricted differences in	group	Quasi-difference	group	Quasi-difference	earnings less comparison	group	Quasi-difference	group	Quasi-difference	group	Quasi-difference	earnings less comparison	group	Quasi-difference	group	Quasi-difference	group	Quasi-difference	group	Quasi-difference	earnings less comparison					
	1978	1975–1978							1978							1978					1978				1975–1978					
	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Controlling for all variables <sup>f</sup>						
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)				
NSW	886	798	879	802	820	1,794	1,572	1,750	1,631	1,612	1,794	1,688	1,750	1,672	1,668	1,640	1,632	1,638	1,632	1,638	1,632	1,638	1,632	1,638	1,632	1,638				
	(472)	(472)	(467)	(468)	(468)	(633)	(637)	(632)	(637)	(633)	(633)	(636)	(636)	(632)	(632)	(636)	(636)	(632)	(632)	(636)	(632)	(636)	(632)	(632)	(636)	(632)	(636)			
PSID-1	-15,578	-8,057	-2,380	-2,119	-1,844	-15,205	-7,741	-582	-265	186	-15,205	-879	-582	218	731															
	(913)	(990)	(680)	(746)	(762)	(1155)	(1175)	(841)	(861)	(901)	(1155)	(931)	(841)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)	(866)			
PSID-2	-4,020	-3,482	-1,364	-1,694	-1,876	-3,647	-2,810	721	298	111	-3,647	94	721	907	688															
	(781)	(935)	(729)	(878)	(885)	(960)	(1082)	(886)	(1004)	(1032)	(960)	(1042)	(886)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)	(1004)			
PSID-3	697	-509	629	-552	-576	1,070	35	1,370	243	298	1,070	821	1,370	822	825															
	(760)	(967)	(757)	(967)	(968)	(900)	(1101)	(897)	(1101)	(1105)	(900)	(1100)	(897)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)	(1101)			
CPS-1	-8,870	-4,416	-1,543	-1,102	-987	-8,498	-4,417	-78	525	709	-8,498	-8	78	739	972															
	(562)	(577)	(426)	(450)	(452)	(712)	(714)	(537)	(567)	(560)	(712)	(572)	(537)	(547)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)	(550)		
CPS-2	-4,195	-2,341	-1,649	-1,129	-1,149	-3,822	-2,208	-263	371	305	-3,822	615	371	263	790															
	(533)	(620)	(459)	(551)	(551)	(671)	(746)	(574)	(662)	(666)	(671)	(672)	(574)	(654)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)	(658)		
CPS-3	-1,008	-1	-1,204	-263	-234	-635	375	-91	844	875	-635	1,270	844	875	91	91	91	91	91	1,326	1,326	1,326	1,326	1,326	1,326	1,326	1,326	1,326		
	(539)	(681)	(532)	(677)	(675)	(657)	(821)	(641)	(808)	(810)	(657)	(798)	(641)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)	(798)		

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID-1, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly recreate his subset. Column (5) differs because the data file that we obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

<sup>a</sup> Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

<sup>b</sup> The estimates based on the NSW control group are unbiased estimates of the treatment impacts for the original sample (\$886) and for the RE74 sample (\$1,794).

<sup>c</sup> The exogenous variables used in the regressions-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

<sup>d</sup> Regresses RE78 on a treatment indicator and RE75.

<sup>e</sup> The same as (d), but controls for the additional variables listed under (c).

<sup>f</sup> Controls for all pretreatment covariates.

## Covariate imbalance

- Conditional on the propensity score, the covariates are independent of the treatment, suggesting that the distribution of covariate values should be the same for both treatment and control groups
- This can be checked as we have data on all three once we've estimated the propensity score
- DW note that the two samples have severe imbalance on *observables* – a huge number of non-experimental controls have propensity scores almost exactly equal to 0
- Their analysis will “trim” (which will ultimately have implications for interpretation)

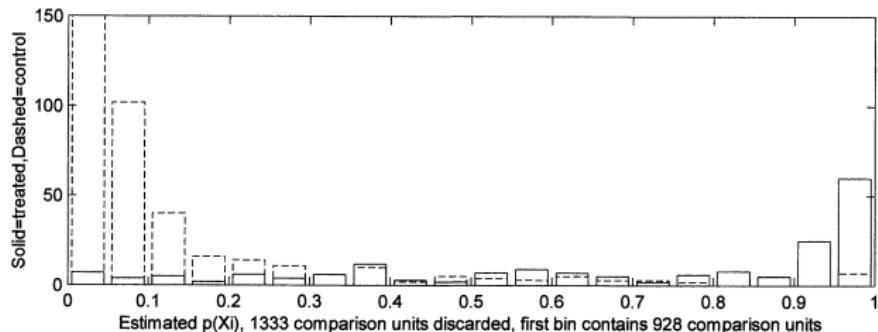


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8-.85, .85-.9, and .9-.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

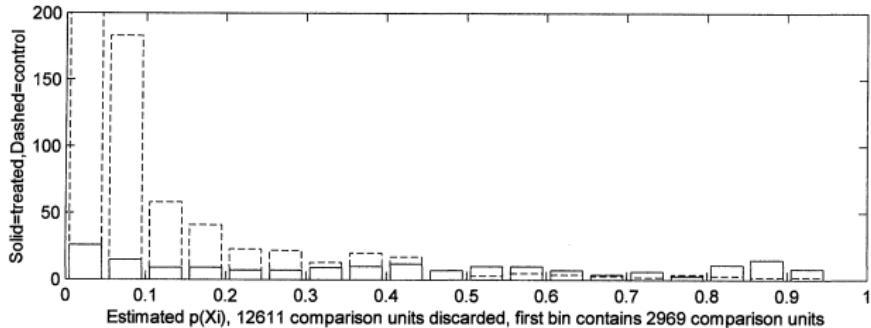


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45-.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score <sup>b</sup>	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted <sup>a</sup>	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations <sup>c</sup>	(7) Unadjusted	(8) Adjusted <sup>d</sup>
NSW	1,794 (633)	1,672 (638)						
PSID-1 <sup>e</sup>	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 <sup>f</sup>	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 <sup>f</sup>	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 <sup>g</sup>	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 <sup>g</sup>	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 <sup>g</sup>	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

<sup>a</sup> Least squares regression: RE78 on a constant, a treatment indicator, age, age<sup>2</sup>, education, no degree, black, Hispanic, RE74, RE75.<sup>b</sup> Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).<sup>c</sup> Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.<sup>d</sup> Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)].

Propensity scores are estimated using the logistic model, with specifications as follows:

<sup>e</sup> PSID-1: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, married, no degree, black, Hispanic, RE74, RE75, RE74<sup>2</sup>, RE75<sup>2</sup>, u74\*black).<sup>f</sup> PSID-2 and PSID-3: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, no degree, married, black, Hispanic, RE74, RE74<sup>2</sup>, RE75, RE75<sup>2</sup>, u74, u75).<sup>g</sup> CPS-1, CPS-2, and CPS-3: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education\*RE74, age<sup>3</sup>).

*Table 4. Sample Means of Characteristics for Matched Control Samples*

<i>Matched samples</i>	<i>No. of observations</i>	<i>Age</i>	<i>Education</i>	<i>Black</i>	<i>Hispanic</i>	<i>No degree</i>	<i>Married</i>	<i>RE74 (U.S. \$)</i>	<i>RE75 (U.S. \$)</i>
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.

MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

## Replies by econometricians to DW

- Heckman, Smith and Todd concluded from their own work that in order for matching estimators to have low bias, you need the following:
  1. A rich set of variables related to program participation and predictive of  $Y^0$  labor market outcomes,
  2. Nonexperimental comparison group be drawn from the same local labor markets as the participants and
  3. Dependent variable (e.g., earnings) be measured in the same way for participants and nonparticipants
- All three of these conditions fail to hold in DW (1999, 2002) according to Smith and Todd (2005)
- DW also note the importance of conditioning on pre-treatment lagged outcomes (e.g., real earnings in  $t - 1, t - 2$ , etc.) as well as *trimming*

## Smith and Todd, diff-in-diff, doubly robust

- Difference-in-differences with propensity scores tended to work well in Smith and Todd (2005) though the effect sizes are much larger
- In my Causal Inference II workshop, we use Sant'anna And Zhao's double robust DiD and get nearly the exact same parameter estimate as the experimental finding

## Coding together

- Let's spend some time together trying to replicate the Lalonde study ourselves
- I use replicate lightly – our goal is syntax only
- It's in the Lalonde lab on github for this workshop

# Interpreting OLS coefficients

- Most common causal model is OLS with covariate controls

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- OLS is the best linear predictor (Angrist and Pischke 2009)
- But the best linear predictor of the *realized* outcome is not the same as being the best linear predictor of the *missing potential outcome*

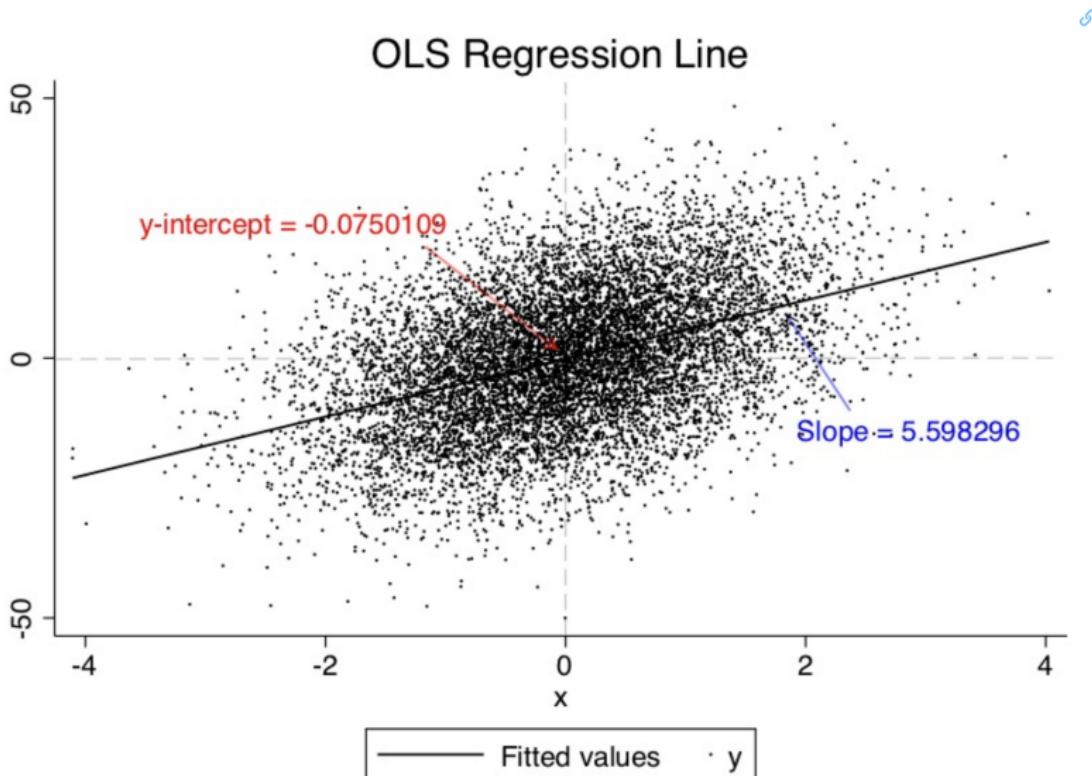
# OLS Simulated Code

[Stata Code](#)[R Code](#)[Python Code](#)

## ▼ Code

```
set seed 1
clear
set obs 10000
gen x = rnormal()
gen u  = rnormal()
gen y  = 5.5*x + 12*u
reg y x
predict yhat1
gen yhat2 = -0.0750109 + 5.598296*x // Compare yhat1 and yhat2
sum yhat*
predict uhat1, residual
gen uhat2=y-yhat2
sum uhat*
twoway (lfit y x, lcolor(black) lwidth(medium)) (scatter y x, mcolor(black) msymbol(point)), title(OLS Regression Line)
rvfplot, yline(0)
```

# OLS Extrapolates Using Functional Form (e.g., Lines)



## Bivariate vs multivariate regressions

- Interpreting the OLS coefficient in bivariate regression is written out as a scaled covariance:

$$\hat{\beta}_1 = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

- But when we are looking at a multivariate regression, what is it?

## Applying FWL theorem

- Frisch-Waugh-Lovell also helps us interpret  $\hat{\beta}_1$  when there are covariates
- Use the FWL theorem to turn the multivariate regression coefficient into the simple bivariate one
- Angrist and Pischke (2009) call FWL the “regression anatomy theorem” and Filoso (2013) has an excellent and original proof

# Applying FWL theorem

- Can we estimate the causal effect of family size on labor supply by regressing labor supply ( $Y$ ) on family size ( $X$ )?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- If family size is random, then we can interpret  $\hat{\beta}_1$  as the ATE
- But how do we interpret  $\hat{\beta}_1$  if `family size` is only conditionally random?

## Applying FWL theorem

- Assume the longer model is:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_1 \text{White}_i + \gamma_2 \text{Married}_i \\ + \gamma_3 \text{Age}_i + \gamma_4 \text{Employed}_i + u_i$$

- Assume unconfoundedness and constant treatment effects so that family size is independent conditional on all controls
- FWL shows that in a multivariate regression, any one coefficient fitted can be reconceived as a simple scaled covariance of the outcome and residualized treatment variable

# FWL Theorem

## FWL Theorem

Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

and an auxiliary regression in which the variable  $x_{1i}$  is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i$$

and  $\tilde{x}_{1i} = x_{1i} - \hat{x}_{1i}$  being the residual from the auxiliary regression. The parameter  $\beta_1$  can be rewritten as:

$$\beta_1 = \frac{Cov(y_i, \tilde{x}_{1i})}{Var(\tilde{x}_{1i})}$$

## FWL Proof (Proof by Filoso 2013)

To prove the theorem, note  $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\hat{x}_{ki}] = E[f_i]$ , and plug  $y_i$  and residual  $\tilde{x}_{ki}$  from  $x_{ki}$  auxiliary regression into the covariance  $cov(y_i, \tilde{x}_{ki})$

$$\begin{aligned}\beta_k &= \frac{cov(y_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\ &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\ &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{var(f_i)}\end{aligned}$$

1. Since by construction  $E[f_i] = 0$ , it follows that the term  $\beta_0 E[f_i] = 0$ .
2. Since  $f_i$  is a linear combination of all the independent variables with the exception of  $x_{ki}$ , it must be that

$$\beta_1 E[f_i x_{1i}] = \cdots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \cdots = \beta_K E[f_i x_{Ki}] = 0$$

3. Consider now the term  $E[e_i f_i]$ . This can be written as:

$$\begin{aligned}
 E[e_i f_i] &= E[e_i f_i] \\
 &= E[e_i \tilde{x}_{ki}] \\
 &= E[e_i(x_{ki} - \hat{x}_{ki})] \\
 &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}]
 \end{aligned}$$

Since  $e_i$  is uncorrelated with any independent variable, it is also uncorrelated with  $x_{ki}$ : accordingly, we have  $E[e_i x_{ki}] = 0$ . With regard to the second term of the subtraction, substituting the predicted value from the  $x_{ki}$  auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\hat{\gamma}_0 + \hat{\gamma}_1 x_{1i} + \cdots + \hat{\gamma}_{k-1} x_{k-1i} + \hat{\gamma}_{k+1} x_{k+1i} + \cdots + \hat{\gamma}_K x_{Ki})]$$

Once again, since  $e_i$  is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows  $E[e_i f_i] = 0$ .

4. The only remaining term is  $E[\beta_k x_{ki} f_i]$  which equals  $E[\beta_k x_{ki} \tilde{x}_{ki}]$  since  $f_i = \tilde{x}_{ki}$ . The term  $x_{ki}$  can be substituted using a rewriting of the auxiliary regression model,  $x_{ki}$ , such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= E[\beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})]] \\ &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}] \tilde{x}_{ki})]\} \\ &= \beta_k var(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between  $E[x_{ki}|X_{-k}]$  and  $\tilde{x}_{ki}$ . From previous derivations we finally get

$$cov(y_i, \tilde{x}_{ki}) = \beta_k var(\tilde{x}_{ki})$$

which completes the proof.

## FWL

- Let's review the FWL "partialing out" interpretation of OLS estimated  $\hat{\beta}_1$  in code
- We will use the fwl.do at /Labs/Matching for this

## Core OLS assumptions

- Return to OLS model with additive covariates controls:

$$Y_i = \alpha + \delta D_i + \beta X_i + \gamma Z_i + \varepsilon_i$$

- Which average treatment effect parameter does  $\hat{\delta}$  estimate? What assumptions are imposed by exogeneity?
- This hopefully is where I'll be able to convince you of how important it is that you identify ahead of time *which* causal effect
- This is **not** a criticism of OLS but rather of that specification above

# OLS Assumptions

- Typically we assume that the mean error is zero conditional on all covariates, called exogeneity
- This is a pregnant assumption as it turns out
- Imbens and Rubin discussed it in their book on causal inference from 2015
- I'll pull out their quote and proof but we will then discuss some other materials

# OLS Assumptions

*"In many empirical studies in social sciences, causal effects are estimated through linear regression, where, typically it is implicitly assumed that in the super-population,*

$$E[Y_i^D | X_i] = \alpha + \delta_{sp} \cdot D + X_i \beta$$

*for some values of the three unknown parameters,  $\alpha$ ,  $\delta_{sp}$  and  $\beta$  where  $\delta_{sp} = E_{sp}[Y_i^1 - Y_i^0]$ ."*

## What about OLS? (Imbens and Rubin 2015)

"Defining  $\varepsilon_i = Y_i - \delta_{sp} \cdot D_i - X_i\beta$  so that we can write

$$Y_i = \alpha + \delta_{sp} \cdot D_i + X_i\beta + \varepsilon_i$$

it is then assumed that

$$\varepsilon_i \perp\!\!\!\perp D_i, X_i$$

This assumption is often referred to as **exogeneity** of the treatment (and the pre-treatment variables) in the econometrics literature."

## OLS Assumptions

*"The regression function is interpreted as a causal relation, in our sense of the term "causal", namely that if we manipulate the treatment  $D_i$ , then the outcome would change in expectation by an amount  $\delta_{sp}$ . Hence in the potential outcomes formulation, we have*

$$Y_i^0 = \alpha + X_i\beta + \varepsilon_i$$

$$Y_i^1 = Y_i^0 + \delta_{sp}$$

## OLS Assumptions

"Then, because  $\varepsilon_i$  is a function of  $Y_i^0$  and  $X_i$  given the parameters,

$$Pr(D_i = 1|Y_i^0, Y_i^1 X_i) = Pr(D_i|\varepsilon_i, X_i),$$

and by exogeneity of the treatment indicator, we have

$$Pr(D_i|\varepsilon_i, X_i) = Pr(D_i|X_i)$$

and thus [conditional independence] holds."

# OLS Assumptions

*"However, the exogeneity assumption combines unconfoundedness with functional form and constant treatment effect assumptions that are quite strong, and arguably unnecessary." – Imbens and Rubin (2015)*

# Regression with correct functional form



Imbens and Rubin are saying regression can identify an aggregate causal parameter under unconfoundedness because it assumes constant treatment effects and the correct functional form which is like an army without a bridge shooting arrows and using catapults to the target. If the functional form is right, then they'll hit the target everytime.

# Constant Treatment Effects and Linearity

Most commonly used method is OLS where the outcome is an additive model of the observed outcome,  $Y$ , on the treatment,  $D$ , and covariates,  $X$  like:

$$Y_i = \alpha + \delta D_i + \beta_1 X_i + \varepsilon_i$$

Take conditional expectations

$$E[Y_i | D_i = 1, X_i] = \alpha + \delta E[D_i | D_i = 1, X_i] + \beta_1 E[X_i | D_i = 1, X_i]$$

$$E[Y_i | D_i = 0, X_i] = \alpha + \delta E[D_i | D_i = 0, X_i] + \beta_1 E[X_i | D_i = 0, X_i]$$

# Constant Treatment Effects and Linearity

Replace realized variables with potential notation (both outcomes and covariates):

$$E[Y_i^1 | D_i = 1, X_i] = \alpha + \delta + \beta_{11} E[X_i^1 | D_i = 1, X_i]$$

$$E[Y_i^0 | D_i = 0, X_i] = \alpha + \beta_{01} E[X_i^0 | D_i = 0, X_i]$$

$X^1$  is the effect of  $X$  on  $Y^1$  when treated and  $X^0$  is effect on  $Y^0$  when not treated; note if treatment changes  $X$  then  $X^1 \neq X^0$  and it will be a problem

# Constant Treatment Effects and Linearity

With a binary treatment variable, the OLS estimator is equivalent to simple difference in conditional means:

$$\begin{aligned}\hat{\delta} &= E[Y_i^1 | D_i = 1, X_i] - E[Y_i^0 | D_i = 0, X_i] \\ &= \left( \alpha + \delta E[D_i | D_i = 1, X_i] + \beta_{11} E[X_i^1 | D_i = 1, X_i] \right) \\ &\quad - \left( \alpha + \delta E[D_i | D_i = 0, X_i] + \beta_{01} E[X_i^0 | D_i = 0, X_i] \right) \\ &= \delta + \beta_{11} E[X_i^1 | D_i = 1, X_i] - \beta_{01} E[X_i^0 | D_i = 0, X_i]\end{aligned}$$

OLS model requires: (1) linearity, (2) treatment cannot change covariate values (e.g., bad controls), (3)  $\beta_{11} = \beta_{01}$  (homogenous treatment effects with respect to  $X$ ).

# Simulation

- Simulation will have linear DGP, heterogeneity with respect to covariates and common support violation (1000 trials)
- Will show a variety of estimators and specifications so that we see how to recover causal parameters with regression and matching
- Focus is on estimated ATE and estimated ATT under a variety of specifications

# Heterogenous Treatment Effects wrt $X$

```
* Simulation with heterogenous treatment effects, unconfoundedness and OLS estimation
clear all
program define het_te, rclass
version 14.2
syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]

clear
drop _all
set obs 5000
gen treat = 0
replace treat = 1 in 2501/5000

* Poor pre-treatment fit
gen age = rnormal(25,2.5)      if treat==1
replace age = rnormal(30,3)       if treat==0
gen gpa = rnormal(2.3,0.75)     if treat==0
replace gpa = rnormal(1.76,0.5)   if treat==1

su age
replace age = age - `r(mean)'

su gpa
replace gpa = gpa - `r(mean)'

gen age_sq = age^2
gen gpa_sq = gpa^2
gen interaction=gpa*age

gen y0 = 15000 + 10.25*age + -10.5*age_sq + 1000*gpa + -10.5*gpa_sq + 500*
interaction + rnormal(0,5)
gen y1 = y0 + 2500 + 100 * age + 1000*gpa
gen delta = y1 - y0

su delta // ATE = 2500
su delta if treat==1 // ATT = 1980
local att = r(mean)
scalar att = `att'
gen att = `att'

gen earnings = treat*y1 + (1-treat)*y0
```

## Parameters

- We have two parameters: the ATE is \$2500 but the ATT is \$1980
- What is the specification for each of them?
- Let's look at what people usually do
- 1,000 simulations of DGP with regression estimates plotting coefficient on treatment dummy: first with just age and GPA, second with the precise model used for  $Y^0$  (but not  $Y^1$ )

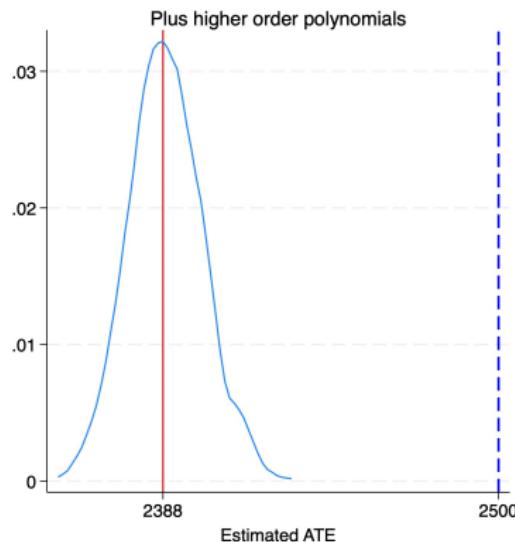
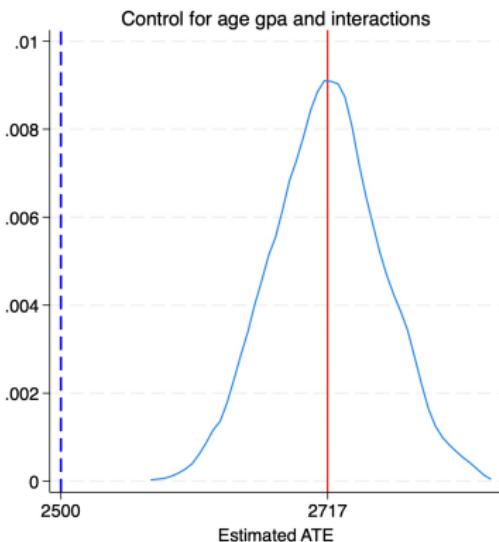
# Constant Treatment Effects and Linearity

```
* Regression 1: constant treatment effects, no quadratics
reg earnings treat age gpa, robust
local treat1=_b[treat]
scalar treat1 = `treat1'
gen treat1=`treat1'

* Regression 2: constant treatment effects, quadratics and interaction
reg earnings treat age age_sq gpa gpa_sq c.gpa#c.age, robust
local treat2=_b[treat]
scalar treat2 = `treat2'
gen treat2=`treat2'
```

# Coefficient on Treatment Dummy is Wrong

Non-saturated regressions with heterogenous treatment effects



ATE is 2500 and ATT is 1980

# Commentary

- Three ifs and a then:
  - If unconfoundedness held, and
  - if the potential outcome model was linear, and
  - if the treatment effect had been homogenous with respect to age and GPA,
  - then the coefficient on the treatment variable would have been the ATE
- But it wasn't because homogeneity with respect to  $X$  was not true (recall  $Y(0)$  coefficients were not the same as  $Y(1)$  coefficients)
- Why were these models biased? Didn't we have the correct specification in the second graph?

## Heterogenous treatment effects

Write down a simplified version of the DGP from the code:

$$Y_i^0 = \alpha + \beta_{01}X_i + \varepsilon_i$$

$$Y_i^1 = \alpha + \beta_{01}X_i + \delta D_i + \beta_{11}X_i \times D_i + \varepsilon_i$$

Notice that the setup before,  $X_i$  had a different effect on  $Y^0$  than it did on  $Y_i^1$  – that's because of heterogenous treatment effects with respect to conditioning set.

# Heterogenous treatment effects

Take conditional expectations of the *potential* outcomes:

$$E[Y_i^0 | D_i = 1, X_i] = \alpha + \beta_{01} E[X_i | D_i = 1, X_i]$$

$$E[Y_i^1 | D_i = 1, X_i] = \alpha + \beta_{01} E[X_i | D_i = 1, X_i]$$

$$+ \delta + \beta_{11} E[X_i | D_i = 1, X_i]$$

Average treatment effect is:

$$\begin{aligned} E[Y_i^1 | D_i, X_i^1] - E[Y_i^0 | D_i, X_i] &= \left( \alpha + \beta_{01} E[X_i] + \delta + \beta_{11} E[X_i \times D_i | D_i = 1, X_i] \right) \\ &\quad - \left( \alpha + \beta_{01} E[X_i] \right) \\ &= \delta + \beta_{11} E[X_i \times D_i | D_i = 1] \end{aligned}$$

## Regression adjustment

This implies an interacted OLS model or what Wooldridge (2010) calls regression adjustment:

$$Y_i = \alpha + \delta D_i + \beta_{01} X_i + \beta_{11} D_i \times X_i + \varepsilon_i$$

$\widehat{\delta}$  is the ATE but the ATT is equal to  $\widehat{\delta} + \widehat{\beta}_{11} E[X_i | D_i = 1]$  where  $E[X_i | D_i = 1]$  is the sample average of  $X_i$  for the treatment group

We will estimate two models: (1) once with simplified but incorrectly specified saturated and (2) another with the correctly specified saturated model – warning, it's a huge pain and you can easily mess it up even with just a few variables

# Misspecified Saturated OLS Regression

```
* Regression 3: Heterogenous treatment effects, partial saturation
regress earnings i.treat##c.age##c.gpa, robust
local ate1=_b[1.treat]
scalar ate1 = `ate1'
gen ate1='ate1'

* Obtain the coefficients
local treat_coef = _b[1.treat]
local age_treat_coef = _b[1.treat#c.age]
local gpa_treat_coef = _b[1.treat#c.gpa]
local age_gpa_treat_coef = _b[1.treat#c.age#c.gpa]

* Save the coefficients as scalars and generate variables
scalar treat_coef = `treat_coef'
gen treat_coef_var = `treat_coef'

scalar age_treat_coef = `age_treat_coef'
gen age_treat_coef_var = `age_treat_coef'

scalar gpa_treat_coef = `gpa_treat_coef'
gen gpa_treat_coef_var = `gpa_treat_coef'

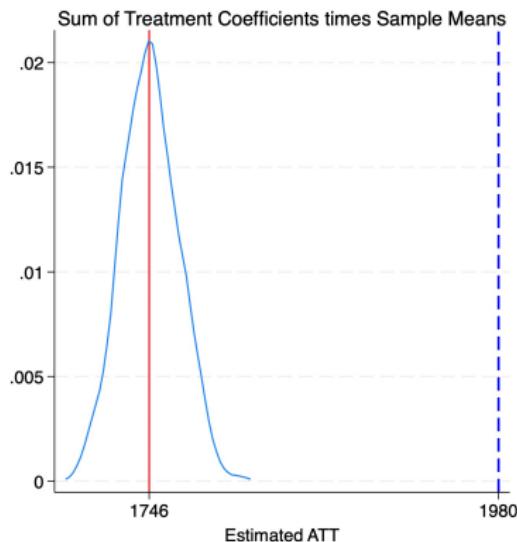
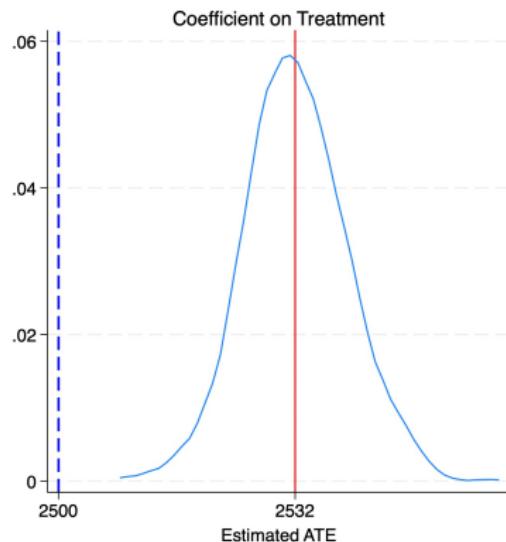
scalar age_gpa_treat_coef = `age_gpa_treat_coef'
gen age_gpa_treat_coef_var = `age_gpa_treat_coef'

* Calculate the mean of the covariates
egen mean_age = mean(age), by(treat)
egen mean_gpa = mean(gpa), by(treat)

* Calculate the ATT
gen treat3 = treat_coef_var + ///
    age_treat_coef_var * mean_age + ///
    gpa_treat_coef_var * mean_gpa + ///
    age_gpa_treat_coef_var * mean_age * mean_gpa if treat == 1
```

# Misspecified Saturated OLS Regression

## Misspecified Saturated Regressions



1000 Monte Carlo simulations

# Comically Long Saturated OLS Regression

```
* Regression 4: Fully saturated regression model
#delimit ;
regress earnings i.treat##c.age
           i.treat##c.age_sq
           i.treat##c.gpa
           i.treat##c.gpa_sq
           i.treat##c.age##c.gpa;
#delimit cr

local ate2=_b[i.treat]
scalar ate2= `ate2'
gen ate2= `ate2'

* Obtain the coefficients
local treat_coeff = _b[1.treat] // 0
local age_coeff = _b[2.treat] // 1
local agesq_coeff = _b[3.treat] // 2
local gpa_coeff = _b[4.treat] // 3
local gpasq_coeff = _b[5.treat] // 4
local age_gpa_coeff = _b[6.treat] // 5

* Save the coefficients as scalars and generate variables
scalar treat_coeff = `treat_coeff'
gen treat_coeff_var = `treat_coeff' // 0
scalar age_treat_coeff = `age_coeff'
gen age_treat_coeff_var = `age_coeff' // 1
scalar agesq_treat_coeff = `agesq_coeff'
gen agesq_treat_coeff_var = `agesq_coeff' // 2
scalar gpa_treat_coeff = `gpa_coeff'
gen gpa_treat_coeff_var = `gpa_coeff' // 3
scalar gpasq_treat_coeff = `gpasq_coeff'
gen gpasq_treat_coeff_var = `gpasq_coeff' // 4
scalar age_gpa_coeff = `age_gpa_coeff'
gen age_gpa_coeff_var = `age_gpa_coeff' // 5

* Calculate the mean of the covariates
su age if treat==1
local mean_age = `r(mean)'
gen mean_age = `mean_age'

su age_sq if treat==1
local mean_agesq = `r(mean)'
gen mean_agesq = `mean_agesq'

su gpa if treat==1
local mean_gpa = `r(mean)'
gen mean_gpa = `mean_gpa'

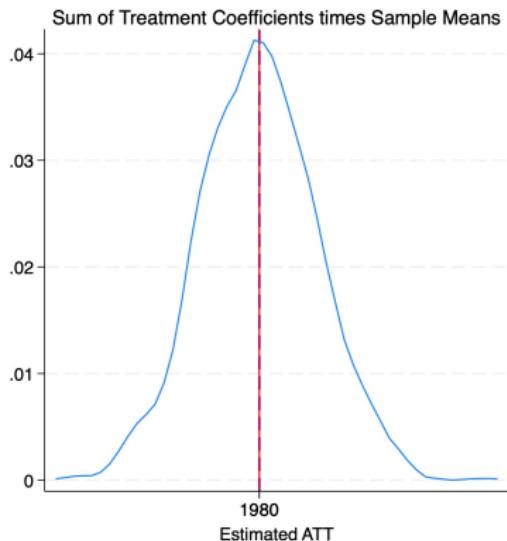
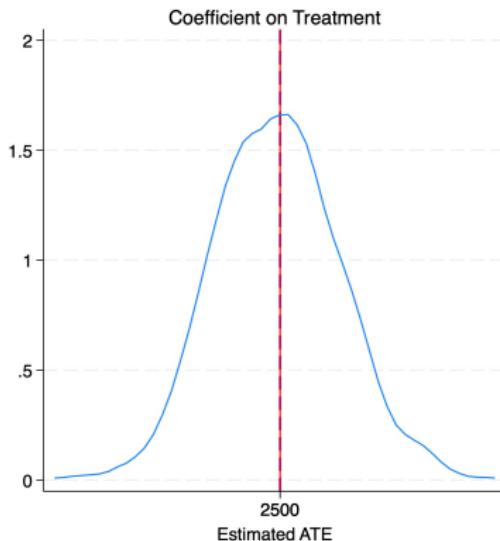
su gpasq if treat==1
local mean_gpasq = `r(mean)'
gen mean_gpasq = `mean_gpasq'

su agegpa if treat==1
local mean_agegpa = `r(mean)'
gen mean_agegpa = `mean_agegpa'

* Calculate the ATT
gen treat4 = `treat_coeff_var' + // 0
           `age_treat_coeff_var' * mean_age + // 1
           `agesq_treat_coeff_var' * mean_agesq + // 2
           `gpa_treat_coeff_var' * mean_gpa + // 3
           `gpasq_treat_coeff_var' * mean_gpasq + // 4
           `age_gpa_coeff_var' * mean_agegpa
```

# Correctly Saturated OLS Regression

## Correctly Specified Saturated Regressions



1000 Monte Carlo simulations

## Regression adjustment

- Notice how the same regression (fully interacted) led to *both* the ATE and the ATT?
- That means if you don't state ahead of time which parameter you want, how are you going to know how to set it up, and how will you know how to recover it?
- Thankfully software exists that does this for you called "regression adjustment" by Wooldridge (2010) or Oaxaca-Blinder (Kline 2011; Graham and Pinto 2022)
- In Stata teffects, you can just the `ra` to get it – see `comparison.do`

## Great quote

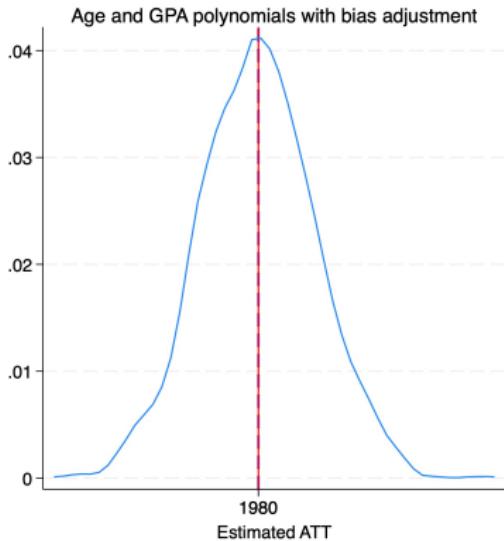
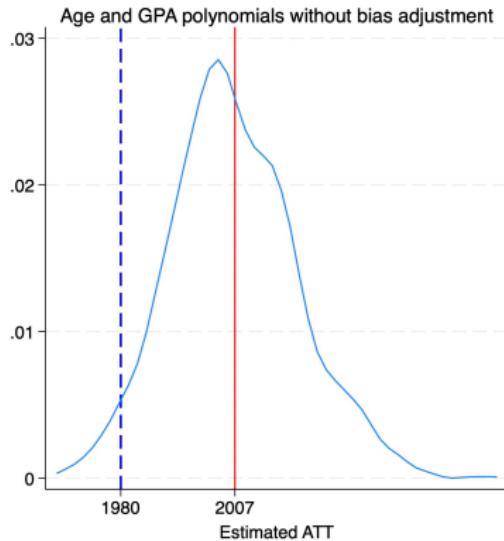
*"OLS is ... the linear projection of Y on D and X [and] provides the best linear predictor of Y given D and X (Angrist and Pishcke 2009). However, if our goal is to conduct causal inference, then this is not, in fact, a good reason to use this method. OLS is 'best' in predicting actual outcomes, but causal inference is about predicting [fictional potential] outcomes. ... In other words, OLS ... is optimal for predicting 'what is'. Instead we are interested in predicting 'what would be' if treatment were assigned differently."* - Tymon Słoczyński, Restat 2022

# Matching

- Now let's estimate the ATT (\$1980) using nearest neighbor matching by minimizing Mahalanobis distance on age, GPA, polynomials and interaction
- One line in Stata using `teffects` and only 1 match (variance is simple to estimate until we use matches multiple times, then the variance grows)
- In R, the package is `Matching`, not sure in python

# Matching Estimation

## Nearest Neighbor Matching with Minimized Maha Distance



Estimated ATT from 1000 simulations using nearest neighbor matching

## Commentary

- Technically I was only “fully interacting” – full saturation would be to interact the treatment dummy with every value of the covariates yielding a huge number of parameters likely that cannot be estimated
- Regression adjustment within -teffects- will do this for you
- But note, with heterogeneity you have to use the fully saturated model (I just hadn’t dummed every value of the covariates) to get both the ATE and the ATT
- Otherwise you are imposing strong and unnecessary assumptions on the data that the treatment effects are the same for all values of  $X$  and constant treatment effects so that the single coefficient is the ATE and the ATT

# Misspecified functional form regression visual



Using our medieval castle example, if you have the wrong functional form, then regression won't hit.

## Nonlinear DGP

- We've seen what happens if there's heterogenous treatment effects with respect to covariates
- But this was resolvable with "full interaction" or "regression adjustment"
- But Imbens and Rubin (2015) noted that exogeneity also implied functional forms, namely linearity
- Let's examine this now using a very unusual DGP using `nonlinear_matching.do`

# Nonlinear DGP

```
clear
drop _all
set obs 5000
gen treat = 0
replace treat = 1 in 2501/5000
gen age = rnormal(35,2.5)           if treat==0
replace age = rnormal(30,2)          if treat==1
gen gpa = rnormal(2.3,0.75)         if treat==0
replace gpa = rnormal(1.76,0.25)    if treat==1

su age
replace age = age - `r(mean)'

su gpa
replace gpa = gpa - `r(mean)'

su age, detail
replace age = age - `r(mean)'

gen gpa_age = age*gpa

su age, detail

* Admittedly weird nonlinear data generating process: Y0
gen y0 = rnormal()
replace y0 = 200 + rnormal() if age < `r(p25)'
replace y0 = 0 + runiform() if age>= `r(p25)' & age<=`r(p75)'
replace y0 = 150 + rnormal() if age>`r(p75)'

su gpa_age, detail

* Admittedly weird nonlinear data generating process: Y1
gen y1 = 0
replace y1 = y0 + 2000 * (0.25)*gpa_age + rnormal(1,25) if gpa_age >= `r(p5)' & gpa_age < `r(p25)'
replace y1 = -10*y0 + 25*age + (0.1)*gpa if age >= `r(p25)' & gpa_age < `r(p50)'
replace y1 = 10*y0 + (5)*gpa_age + gpa + rnormal(5,5) if gpa_age >= `r(p50)' & gpa_age<`r(p75)'
replace y1 = y0 + (0.05) * gpa_age + 2*age if gpa_age>= `r(p75)'

gen delta = y1-y0

su delta // ATE = approximately 15
local ate = r(mean)
scalar ate = `ate'
gen ate = `ate'

su delta if treat==1 // ATT = approximately 272
local att = r(mean)
scalar att = `att'
```

# Nonlinear DGP, OLS with RA

```
* Regression: Heterogenous treatment effects with age
regress earnings i.treat##c.age i.treat##c.gpa i.treat##c.gpa_age, robust

** ATE
local reg_ate2=_b[1.treat]
scalar reg_ate2 = `reg_ate2'
gen reg_ate2=`reg_ate2'

** ATT
* Obtain the coefficients
local treat_coef = _b[1.treat]
local age_treat_coef = _b[1.treat#c.age]
local gpa_treat_coef = _b[1.treat#c.gpa]
local gpaage_treat_coef = _b[1.treat#c.gpa_age]

* Save the coefficients as scalars and generate variables
scalar treat_coef = `treat_coef'
gen treat_coef_var = `treat_coef'

scalar age_treat_coef = `age_treat_coef'
gen age_treat_coef_var = `age_treat_coef'

scalar gpa_treat_coef = `gpa_treat_coef'
gen gpa_treat_coef_var = `gpa_treat_coef'

scalar gpaage_treat_coef = `gpaage_treat_coef'
gen gpaage_treat_coef_var = `gpaage_treat_coef'

* Calculate the mean of the age covariate for treatment group only
egen mean_age = mean(age) if treat==1
egen max_age = max(mean_age)
replace mean_age = max_age if treat==0

egen mean_gpa = mean(gpa) if treat==1
egen max_gpa = max(mean_gpa)
replace mean_gpa = max_gpa if treat==0

egen mean_gpaage = mean(gpaage) if treat==1
egen max_gpaage = max(mean_gpaage)
replace mean_gpaage = max_gpaage if treat==0

* Calculate the ATT
gen reg_att = treat_coef_var + /// 0
                age_treat_coef_var * mean_age + /// 1
                gpa_treat_coef_var * mean_gpa + /// 2
                gpaage_treat_coef_var * mean_gpaage
```

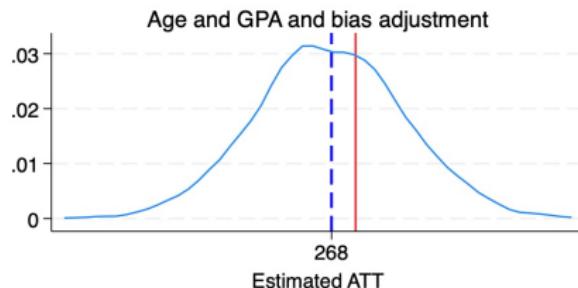
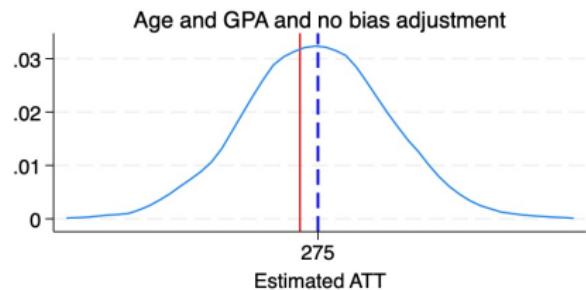
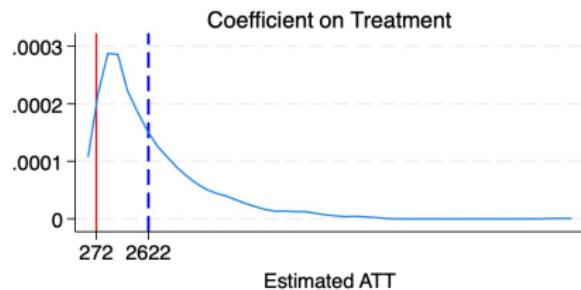
# Nonlinear DGP and Matching

```
** Distance minimization matching method
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(maha)
mat b=e(b)
local nn_att1 = b[1,1]
scalar nn_att1=`nn_att1'
gen nn_att1=`nn_att1'

** Distance minimization matching method with bias adjustment
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(maha) biasadj(age gpa)
mat b=e(b)
local nn_att2 = b[1,1]
scalar nn_att2=`nn_att2'
gen nn_att2=`nn_att2'
end
```

# Nonlinear DGP, OLS and Matching

## Saturated Regression and Nearest Neighbor Matching



The dashed blue line is estimated ATT from 1000 simulations using saturated OLS and nearest neighbor matching.

# Understanding the Standard OLS Model

- Tymon S loczyński's 2022 Restat paper decomposes the following OLS model estimate of  $\tau$  into weighted average of other parameters of interest

$$y = \alpha + \tau d + X\beta + u, \quad (4)$$

- Recall this is the model we found to have observed bias in earlier simulations due to the assumption of constant treatment effects (Imbens and Rubin 2015).
- But Tymon's paper addresses the interpretation of  $\tau$ , the OLS estimand, when treatment effects are heterogeneous.

## Weights in Heterogeneous Treatment Effects

- With heterogeneous treatment effects,  $\hat{\tau}$  is both biased and pulled towards the smallest group's aggregate causal parameter (i.e., ATT vs ATU)
- In other words, the weight OLS places on the average effect for each group is inversely related to the group's size.
- As the number of treated units increases (as a fraction of the population), the weight on ATT decreases
- This suggests OLS estimates based on this additive controls specification will be biased towards the smallest group's average treatment effect making it inappropriate when treatment effects vary across subjects

# A Pragmatic View of Heterogeneous Treatment Effects

- Tymon's paper provides a pragmatic perspective on the main result and derives corollaries suggesting diagnostic methods for when:
  1. The treatment is binary.
  2. OLS is used.
  3. The researcher does not want to assume  $ATT = ATU$ .
- Diagnostics help detect when OLS weights deviate from what is needed for consistent estimation of ATE or ATT.

# Diagnostics for OLS with Heterogeneous Treatment Effects

- The diagnostics:
  - Are bounded between 0 and 1 in absolute value.
  - Reflect the proportion of bias from the difference between ATU and ATT.
- A diagnostic close to 0 implies OLS may be suitable, while a value far from 0 suggests the need for alternative methods.

# Simple Diagnostics for ATT and ATE

- Special case diagnostics provide simple rules-of-thumb:
  - For ATT estimation: Diagnostic is equal to the proportion of treated units,  $P(d = 1)$ .
  - For ATE estimation: Diagnostic is  $2 \times P(d = 1) - 1$ , or twice the deviation from 50%.
- OLS approximates ATE well if the size of treated and untreated groups is similar.
- For ATT, a small proportion of treated units is necessary.

## Main Result: Algebra of OLS and Descriptive Estimands

- Focuses on the algebra of OLS and descriptive estimands.
- Potential outcomes and conditions for causal interpretation discussed in Section IIC.
- Main result does not require causal interpretation assumptions.

## Linear projection

- Linear projection is a concept from the theory of linear regression describing the expected value of the dependent variable as a linear function of independent variables
- It does not necessarily incorporate randomness or an error term and so should be considered distinct from the “data generating process” which does
- It posits a relationship that says, “If we knew the true parameters, this is how we would expect  $Y$  to change, on average, with changes in  $X$  and  $D$ ”

# Linear Projection

- Let  $L(\cdot|\cdot)$  denote the linear projection.
- Interested in the interpretation of  $\tau$  in the linear projection of  $y$  on  $d$  and  $X$ :

$$L(y|1, d, X) = \alpha + \tau d + X\beta \quad (2)$$

- This projection is distinct from the structural conditional mean.

# Probability of Treatment and Propensity Score

- The unconditional probability of treatment:

$$\rho = P(d = 1) \quad (3)$$

- Propensity score from the linear probability model:

$$p(X) = L(d|1, X) = \alpha_p + X\beta_p \quad (4)$$

- $p(X)$  is the best linear approximation to the true propensity score.
- Equations (2) and (4) provide the specification.

## Flexibility of the Linear Projection

- Equation (2) can be seen as partially linear, including powers and cross-products of control variables.
- The propensity score approximation can be made very accurate.

# Linear Projections with Propensity Score

- Separate linear projections of  $y$  on  $p(X)$  for  $d = 1$  and  $d = 0$ :

$$L[y|1, p(X), d = 1] = \alpha_1 + \gamma_1 \times p(X) \quad (5)$$

$$L[y|1, p(X), d = 0] = \alpha_0 + \gamma_0 \times p(X) \quad (6)$$

- Definitions are based on the variability of the propensity score within treatment groups.

## Assumptions for Linear Projections

- Assumption 1: Existence and uniqueness of linear projections in (2) and (4).
- Assumption 2: Existence and uniqueness of linear projections in (5) and (6).
- These assumptions guarantee the linear projections are well-defined and unique.
- Both of these assumptions (fn. 3) are generally innocuous although the second one rules out a small number of interesting applications like regression adjustments in completely randomized experiments – but in these cases OLS will estimate the ATE consistently (Imbens and Rubin 2015)

## Defining Average Partial Linear Effects

- The average partial linear effect of  $d$ :

$$\tau_{\text{APLE}} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \times E[p(X)] \quad (7)$$

- The average partial linear effect of  $d$  on group  $j$  (where  $j = 0, 1$ ):

$$\tau_{\text{APLE},j} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \times E[p(X)|d = j] \quad (8)$$

- These estimands are well-defined under Assumptions 1 and 2.

## Causal Interpretation and Theorem 1

- Causal interpretation requires additional assumptions beyond 1 and 2.
- Theorem 1 is more general, relying only on Assumptions 1 and 2.
- It applies to the algebraic properties of OLS estimators.

## Weighted Average Interpretation of OLS (Theorem 1)

Under assumptions (1) and (2)

$$\tau = \omega_1 \times \tau_{APLE,1} + \omega_0 \times \tau_{APLE,0}$$

where the weights are equal to:

$$\omega_1 = \frac{(1 - \rho) \times V[p(X)|d = 0]}{\rho \times V[p(X)|d = 1] + (1 - \rho) \times V[p(X)|d = 0]}$$

and  $\omega_0 = 1 - \omega_1$

## Theorem 1: Interpretation of the OLS Estimand

- The OLS estimand  $\tau$  is a convex combination of  $\tau_{\text{APLE},1}$  and  $\tau_{\text{APLE},0}$ .
- Reflects the weighted average outcome of a three-step procedure:
  1. Obtain the propensity score  $p(X)$ .
  2. Calculate  $\tau_{\text{APLE},1}$  and  $\tau_{\text{APLE},0}$  for  $d = 1$  and  $d = 0$  via linear projections.
  3. Compute the weighted average of  $\tau_{\text{APLE},1}$  and  $\tau_{\text{APLE},0}$ .

## Weighting in the Estimation of $\tau$

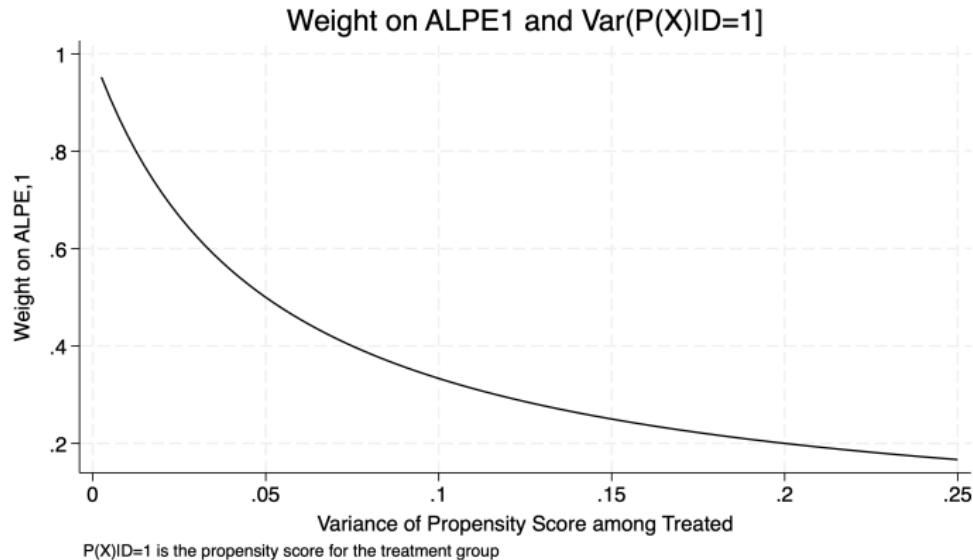
- The weight on  $\tau_{\text{APLE},1}$ ,  $\omega_1$ , decreases with  $V[p(X)|d = 1]$  and the probability of treatment  $\rho$ .
- Conversely, the weight on  $\tau_{\text{APLE},0}$ ,  $\omega_0$ , increases with  $V[p(X)|d = 1]$  and  $\rho$ .
- The weighting scheme implies that larger groups receive less weight on their respective  $\tau_{\text{APLE},j}$ .

## Assumptions for an Illustrative Example

The following graph shows how the weight on  $\tau_{APLE,1}$  decreases as the variance of the propensity score among the treated increases using the following simplifications for illustrative purposes:

- The variance among the non-treated  $V[p(X)|d = 0]$  is constant at 0.1.
- The probability of treatment  $\rho$  is 0.5, meaning that half of the population is treated.
- The variance among the treated  $V[p(X)|d = 1]$  varies from 0.1 to 0.9 for the purpose of the illustration.

# Weight on $\tau_{\text{APLE},1}$ with Varying Variance among Treated



*Figure:* The relationship between the weight  $\omega_1$  on  $\tau_{\text{APLE},1}$  and the variance of the propensity score among the treated  $V[p(X)|d = 1]$ . As the variance among the treated increases, the weight on  $\tau_{\text{APLE},1}$  decreases.

# Factors Causing the Variance of the Propensity Score to Rise

## **Understanding Variance:**

- The variance is not simply  $p(1 - p)$  but the spread of the estimated propensity scores  $p(X)$  within the treatment group.
- It reflects the degree to which the likelihood of treatment (given covariates) varies among those actually treated.

# Factors Causing the Variance of the Propensity Score to Rise

**The variance of the propensity score among the treated,  
 $V[p(X)|d = 1]$ , can increase due to several factors:**

- **Heterogeneity in Covariates:** A diverse range of covariates within the treated group can lead to a wide array of propensity scores, increasing variance.
- **Model Specification:** The choice of model and its ability to accurately capture the relationship between covariates and treatment affects variance. Mis-specification can inflate variance.
- **Overlap with Control Group:** Significant overlap in propensity scores between treated and untreated groups can raise variance, reflecting less distinction based on covariates.

## Stata exercises

We will now look at two simulations to help you understand both of these

## Intuition Behind the Weighting Scheme

- This counterintuitive result may be understood through the lens of finite sample properties.
- OLS places more weight on treatment effects that are more precisely estimated.
- Further intuition and alternative proof are provided in Online Appendix B2.
- The framework aligns with the insights from Angrist (1998) and Angrist and Pischke (2009) regarding the weighting of treatment effects in OLS.

# Causal Interpretation of OLS

Theorem 1 ("Weighted Average Interpretation of OLS") is general due to its reliance solely on the existence and uniqueness of linear projections; however, for a causal interpretation of OLS, we must introduce:

- Potential outcomes  $y(1)$  and  $y(0)$  instead of realized outcomes
- Defined causal parameters: ATE, ATT, and ATU
- Introduce assumptions that will ensure a causal interpretation of  $\tau$

## Assumption 3 for Causal Interpretation

Unconfoundedness in Mean Potential Outcomes or "Ignorability in Means"

1.  $E[y(1)|X, d] = E[y(1)|X]$
2.  $E[y(0)|X, d] = E[y(0)|X]$

Assumptions 3 is our standard unconfoundedness assumption but limited to means (slightly weaker but it's all we need)

# Assumption 4 for Causal Interpretation

## Linearity in Potential Outcomes

Assumption 4 posits that the conditional expectations of potential outcomes are linear in the propensity score:

1.  $E[y(1)|X] = \alpha_1 + \gamma_1 \times p(X)$
2.  $E[y(0)|X] = \alpha_0 + \gamma_0 \times p(X)$

## Historical and Practical Context of Assumption 4

While not a common assumption, it is not necessarily strong due to:

- The flexibility in the specification of  $X$ , which can include nonlinear terms.
- The automatic satisfaction of this assumption in saturated models.

The linearity assumption has historical roots and practical implications:

- Historically aligned with Rosenbaum and Rubin (1983).
- Assumed linearity of  $E[d|X]$  in several notable econometric works.
- It is restrictive but can be mitigated by model specification choices.

## Corollary 1: Causal Interpretation of OLS

Under assumptions the first two linear projection related assumptions and these two new causal assumptions, then

$$\tau = \omega_1 \times \tau_{ATT} + \omega_0 \times \tau_{ATU}$$

In words, this states that under those four assumptions, the OLS weights from theorem 1 apply to the causal parameters we care about, the ATT and the ATU, and so therefore  $\tau$  has a causal interpretation.

But there's a catch: the greater the proportion of treated units, the smaller the OLS weight on the ATT, which is a problem given  $ATE = \rho \times ATT + (1 - \rho)ATU$

## Proving Corollary 1 with Assumptions 3 and 4

- Assumption 3 (Unconfoundedness) allows us to estimate the ATE using realized outcomes, as the expected potential outcomes are equal to the observed outcomes conditional on covariates  $X$ :

$$E[y(1) - y(0)|X] = E[y|X, d = 1] - E[y|X, d = 0]$$

- Assumption 4 specifies the functional form of the expected potential outcomes as linear in the propensity score:

$$E[y(1)|X] = \alpha_1 + \gamma_1 \times p(X), \quad E[y(0)|X] = \alpha_0 + \gamma_0 \times p(X)$$

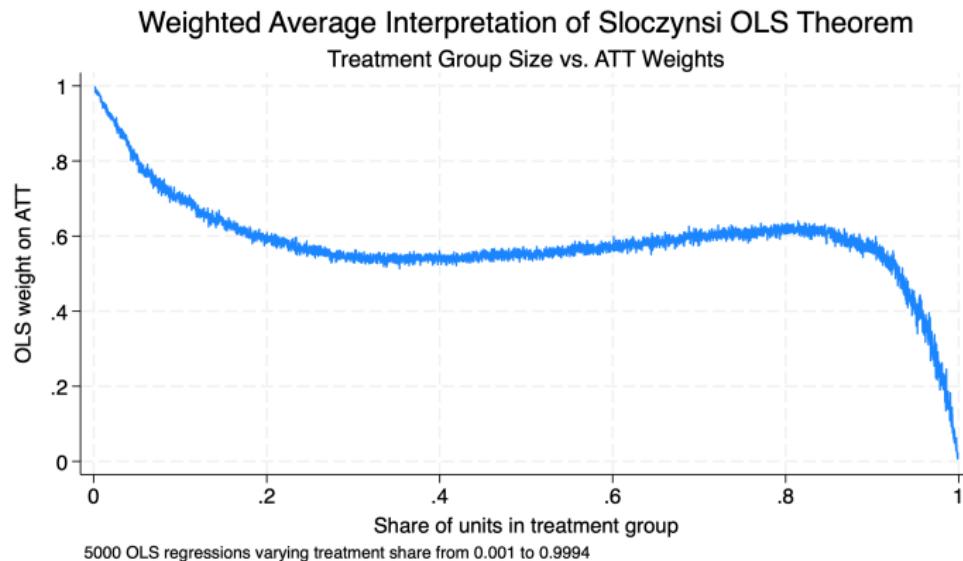
- Together, they imply that the OLS estimand for the treated ( $\tau_{ATT}$ ) and untreated ( $\tau_{ATU}$ ) equates to the Average Partial Linear Effects (APLE):

$$\tau_{ATT} = \tau_{APLE,1}, \quad \tau_{ATU} = \tau_{APLE,0}$$

# OLS Weights and Causal Parameters

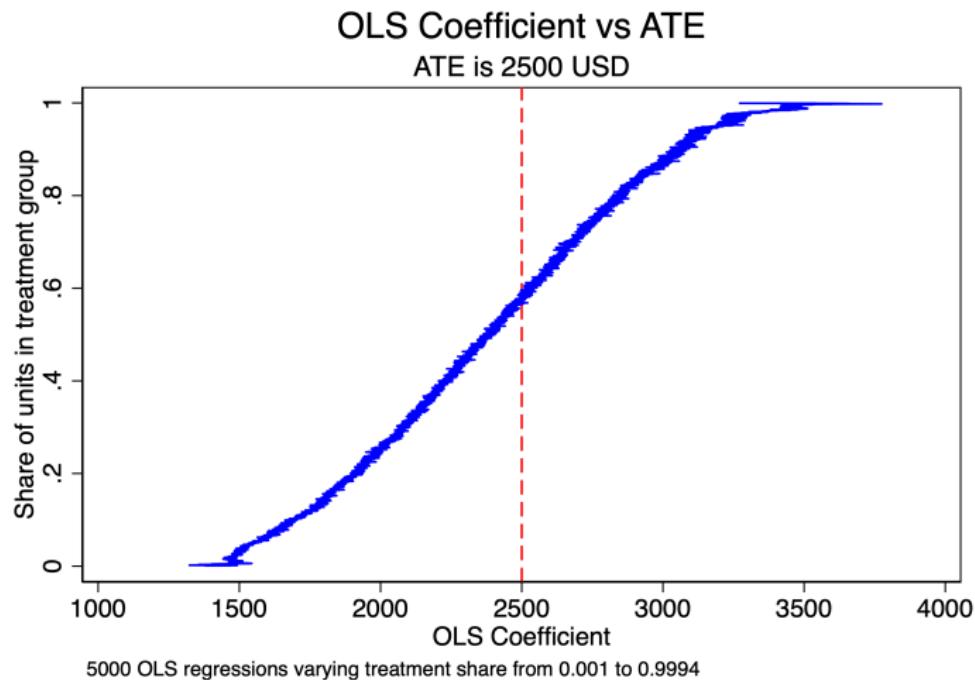
- Note earlier I said "there's a catch" – the OLS weights are inversely related to the size of the treatment group shares
- The greater the proportion of treated units, the smaller the weight on  $\tau_{ATT}$  in the OLS estimand.
- The OLS approach is optimal for predicting actual outcomes, not necessarily for causal inference which aims at predicting counterfactuals.

# Visualizing the problem one way



*Figure:* As more units are placed in treatment (blue line), the weight OLS places on the ATT.

## Visualizing the problem another way



*Figure:* As more units are placed in treatment (top right), OLS is weighted towards the ATU (\$3,037) and when few are in the treatment (bottom left), it is

## Predicting Outcomes vs. Predicting Counterfactuals

- OLS prioritizes weights based on the size of the group when predicting actual outcomes.
- For causal inference (predicting counterfactuals), the emphasis should be on the coefficients predicting outcomes for the smaller group.

# Adjustment for Predicting Counterfactuals

$$\tau_{ATE} = [E(y|d=1) - E(y|d=0)] - [(1-\rho)\beta_1 + \rho\beta_0] \times [E(X|d=1) - E(X|d=0)]$$

- OLS and ATE share structural similarities but differ in weight assignment.
- In ATE, larger groups receive smaller weights on their specific coefficients, opposite to OLS weighting.

# Bias of coefficients

**Corollary 2.** *Under assumptions 1 and 2,*

$$\begin{aligned}\tau - \tau_{ATE} = & \underbrace{w_0 \times (\tau_{APLE,0} - \tau_{ATU}) + w_1 \times (\tau_{APLE,1} - \tau_{ATT})}_{\text{bias from nonlinearity}} \\ & + \underbrace{\delta \times (\tau_{ATU} - \tau_{ATT})}_{\text{bias from heterogeneity}},\end{aligned}$$

## Bias of coefficients

where  $\delta = \rho - w_1 = \frac{\rho^2 \times V[p(X)|d=1] - (1-\rho)^2 \times V[p(X)|d=0]}{\rho \times V[p(X)|d=1] + (1-\rho) \times V[p(X)|d=0]}$ .  
under assumptions 1, 2, 3, and 4,

$$\tau - \tau_{ATE} = \delta \times (\tau_{ATU} - \tau_{ATT}).$$

# Bias of coefficients

**Corollary 3.** *Under assumptions 1 and 2,*

$$\begin{aligned}\tau - \tau_{ATT} = & \underbrace{w_0 \times (\tau_{APLE,0} - \tau_{ATU}) + w_1 \times (\tau_{APLE,1} - \tau_{ATT})}_{\text{bias from nonlinearity}} \\ & + \underbrace{w_0 \times (\tau_{ATU} - \tau_{ATT})}_{\text{bias from heterogeneity}}.\end{aligned}$$

*Also, under assumptions 1, 2, 3, and 4,*

$$\tau - \tau_{ATT} = w_0 \times (\tau_{ATU} - \tau_{ATT}).$$

# Prevalence of OLS Bias in Evaluating Labor Market Programs

- OLS may frequently show substantial bias for ATE and ATT in published studies
- In studies like Card, Kluve, and Weber (2018), with a mean treatment rate of 17.7%, biases are common:
  - The expected bias in OLS for ATE is 64.6% of the ATU-ATT difference.
  - For ATT, the expected OLS bias is 17.7% of the ATU-ATT difference.
- These findings indicate that biases in OLS estimates could often be significant.
  - LaLonde's (1986) training program effects aligning with ATT.
  - Aizer et al.'s (2016) cash transfer effects resembling ATU.
- Implementation of results is available in R and Stata through the `hettreatreg` packages.

## Issues with OLS Weights Illustrated by the NSW Program

- NSW program provides a classic example of OLS weights problem.
- LaLonde (1986) and Angrist and Pischke (2009) assessments using the Dehejia and Wahba (2002) subsample (see Stata code together)
  - OLS close to experimental benchmark but driven by a small treated sample.
  - Economic disadvantage of treated vs. CPS comparison group suggests large ATT and ATU differences.
- Treated sample's size (1.1%) implies that OLS will be skewed towards the ATT

# Tymon's replication of Angrist and Pischke

TABLE 1.—THE EFFECTS OF A TRAINING PROGRAM ON EARNINGS

	(1)	(2)	(3)	(4)
Original estimates				
OLS	-3,437*** (612)	-78 (596)	623 (610)	794 (619)
Diagnostics				
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\hat{w}_0^* = \hat{\beta}$	0.011	0.011	0.011	0.011
$\hat{\delta}$	-0.970	-0.987	-0.971	-0.971
$\hat{\delta}^* = 2\hat{\beta} - 1$	-0.977	-0.977	-0.977	-0.977
Decomposition				
$\widehat{ATT}$	-3,373*** (620)	-69 (595)	754 (619)	928 (630)
$\hat{w}_1$	0.981	0.999	0.983	0.983
$\widehat{ATU}$	-6,753*** (1,219)	-6,289** (2,807)	-6,840*** (1,294)	-6,840*** (1,319)
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\widehat{ATE}$	-6,714*** (1,206)	-6,218** (2,777)	-6,754*** (1,281)	-6,751*** (1,305)
Demographic controls	✓		✓	✓
Earnings in 1974				✓
Earnings in 1975		✓	✓	✓
$\hat{\beta} = \hat{P}(d = 1)$	0.011	0.011	0.011	0.011
Observations	16,177	16,177	16,177	16,177

The estimates in the top panel correspond to column 2 in table 3.3 in Angrist and Pischke (2009, p. 89). The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, Black, and Hispanic. For treated individuals, earnings in 1974 correspond to real earnings in months 13 to 24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. Formulas for  $w_0$ ,  $w_1$ , and  $\delta$  are given in theorem 1 and corollary 2. Following these results, OLS =  $\hat{w}_0 \times \widehat{ATT} + \hat{w}_1 \times \widehat{ATU}$ . Estimates of ATE, ATT, and ATU are sample analogs of  $\widehat{TATE}$ ,  $\widehat{TAPLE}_1$ , and  $\widehat{TAPLE}_0$ , respectively. Also,  $\widehat{ATE} = \hat{\beta} \times \widehat{ATT} + (1 - \hat{\beta}) \times \widehat{ATU}$ . Huber–White standard errors (OLS) and bootstrap standard errors ( $\widehat{ATE}$ ,  $\widehat{ATT}$ , and  $\widehat{ATU}$ ) are in parentheses. Statistically significant at \*10%, \*\*5%, and \*\*\*1%.

## Stata simulation

Let's do this ourselves using `lalonde.do` in the `labs/matching` on github repo

# Replications of Lalonde data using Dehejia and Wahba sample with CPS control

		Dependent variable: Real earnings in 1978				
	(1)	(2)	(3)	(4)	(5)	
Controls:	No controls	Demographics	RE75 Only	Demo and RE75	Demo, RE74 and RE75	
Treatment	-8497.516*** (581.916)	-3436.795*** (612.056)	-77.705 (596.215)	622.547 (609.582)	<b>793.587</b> (618.609)	

Each column corresponds to a set of observable controls below:

- Demographics: Age, Age-squared, Years of schooling, Marriage dummy, High school dropout dummy, Black and Hispanic dummies
- Real earnings in 1975 (RE75)
- Real earnings in 1974 (RE74)

Our focus will be on the last estimate. Question is what is this?

## Regression Adjustment Application and Results

- Estimate the regression adjustment method interacting treatment with all covariates to recover estimates of all parameters:
  - $\widehat{ATE} = -\$4,930$
  - $\widehat{ATT} = \$796$
  - $\widehat{ATU} = -\$4,996$
- OLS estimate \$794 shows a weight on ATT nearly 100%, reflecting the treated sample's size of 1.1%
- Illustrates "weight reversal" principle: OLS weights do not reflect expected proportions.

## General Implications of the OLS Weight Anomaly

- The NSW example generalizes beyond this specific case.
- OLS weights can misrepresent effects, influenced by treatment group proportion.
- OLS estimand does not always align with intuitive weight distributions for ATT and ATU.

## Concluding remarks

- Last years have shown a *particular* OLS specification yields biased estimates of all causal parameters when there exists heterogeneity in the treatment effects
  - Goodman-Bacon (2021) for instance showed this with the common two way fixed effects specification and dynamic treatment effects
- Turns out it was not a diff-in-diff problem only – it's an OLS problem more generally
  - We saw it also in a paper by Goldsmith-Pinkham, Hull and Kolsar on contamination bias in linear models with multiple treatment effects and heterogeneity
- Here we see the perversity of the effects under unconfoundedness and heterogeneity which is the OLS coefficient counterintuitively favors the smaller groups and therefore their causal parameters

## Concluding remarks

- We mentioned work by Anna Aizer more likely is estimating coefficients that are more like the ATU than either the ATE or the ATT because of how large the treatment group is
- This has implications for analysis in which the proportion of groups differ in size due to population differences or over-representation of one group over another (i.e., Asians in criminal justice)
- The smaller the group, the larger their influence on the OLS coefficient
- Lesson here is that regression adjustment as well as nearest neighbor matching with bias adjustment may need to be your core models, not the additive in covariates OLS specification we are accustomed to

# Roadmap

## Selecting Covariates Using Directed Acyclic Graphs

- Graph notation

- Backdoor criterion

- Collider bias

- What if you don't know the DAG?

## Unconfoundedness and Ignorable Treatment Assignment

- Aggregate target parameters

- Assumptions

## Estimators

- Subclassification

- Exact and Inexact Matching

- Propensity scores

- Regressions

## Concluding remarks

## Comments

- Unconfoundedness means that the confounders are known and quantified, meaning they're in your data and well measured
- Also means that within the dimensions of those covariates is an RCT ("independence")
- Common support is also needed with matching, but for OLS you rely on extrapolation and functional forms
- Without a prior behavioral model guiding you, it's very hard to defend unconfoundedness (identification by convenience)

# When not to use unconfoundedness methods

- Individual sorting based on rationality is strong and subtle
  - May be easier to defend in some situations though – we are studying the effect of a prison assignment where if you get a suicide risk score, a team of trained inmates go meet you
  - Only happened in 16 prisons – I have 84 others and I know each inmate score
  - They are picking their score to a degree, by picking their suicidality, but you wouldn't say it was done *rationally*
- College attendance, major, marriage, children, divorce – very hard to imagine that for people with identical covariate values they all flipped coins
- Unconfoundedness is a strong assumption, and the weaker ones (like with respect to  $Y^0$ ) may be easier to defend which gets you to the ATT

# Common Support

- Unconfoundedness says on average,  $E[Y^0|D = 1, X] = E[Y^0|X]$  – that is, it doesn't depend on  $D$  so you can just switch the known for unknown ones
- But even if unconfoundedness holds doesn't mean your dataset is large enough that the one to one matches can happen – that's failure to have enough matches because the dimensions are too large
- So much of the literature is how to handle failed common support
- Bias adjustment methods are one way to address that but even they can't work miracles

## My opinions: Parameter first

- You have three average causal effects associated with three populations
- You will need up front to choose which one, and justify why that is the one you want
- The conditional ATEs like ATT or ATU have fewer assumptions
- ATT is often the one you want anyway (e.g., effect of discrimination on blacks not everyone)

# My opinions: Regression

- Simple regression model

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon_i$$

- It requires extremely strong assumptions like constant treatment effects and has weird weighting properties as we discussed
- Functional form must be right also (linearity)
- It is simply unnecessary to estimate this any longer

## Remember Imbens and Rubin

- Additive OLS model with exogeneity imposes strong assumptions on the DGP: constant treatment effects, unconfoundedness and functional form assumptions
- Matching allows for heterogeneous treatment effects but requires common support; OLS uses extrapolation in its place
- You can address heterogeneous treatment effects with fully interaction called regression adjustment (itself extremely uncommon in practice), but that does not address nonlinear DGP

## Quote that great Tymon Słoczyński quote (Restat 2022)

*"An important motivation for using  $Y = \alpha + \delta D + \beta X + \varepsilon$  and OLS is that the linear project of  $Y$  on  $D$  and  $X$  provides the best linear predictor of  $Y$  given  $D$  and  $X$  (Angrist and Pischke 2009). However if our goal is to conduct causal inference, then this is not, in fact, a good reason to use this method. Ordinary least squares is "best" in predicting actual outcomes, but causal inference is about predicting missing [fictional] outcomes. In other words, the OLS weights are optimal for predicting "what is". Instead, we are interested in predicting "what would be" if treatment were assigned differently."* (Tymon Słoczyński, Restat 2022)

## My opinions: Regression adjustment

- If you want to estimate something with a regression, then assuming unconfoundedness and linearity you can use regression adjustment (RA)
- Recall: regression adjustment allows you to estimate both the ATE and the ATT but will require a saturated model
- You'll want to use software for this as it's too easy to make a mistake

## Nearest neighbor matching with bias adjustment

- Consider Abadie and Imbens (2006; 2008; 2011) nearest neighbor and minimize the Mahalanobis distance metric
- This only requires unconfoundedness and common support, not constant treatment effects and not parametric functional form like regressions
- But most likely you will not have common support, so consider an outcome regression adjustment called bias adjustment

# Inverse probability weighting

- If you use propensity scores, then consider using the IPW as it's got fewer ad hoc choices and does not require knowing the true propensity score (for efficiency)
- You may want to also use outcome regression or double robust to correct for any model misspecification (consistency)
- Always plot the propensity score histograms and assess overlap for the parameter you want