Causal Inference I

Mixtape Session

Roadmap

- Unconfoundedness and Ignorable Treatment Assignment
  - Choosing Covariates
  - Aggregate target parameters
- Matching Estimators
  - Stratification weighting
  - Conditional Independence
  - Exact and Inexact Matching
  - Propensity scores
  - Regressions
  - Coarsened exact matching
- Concluding remarks

Adjusting for variables

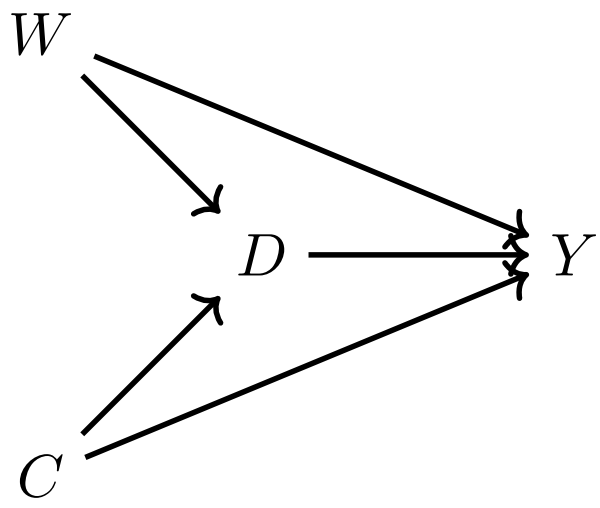
- One of the first things you learn in a methods course is multivariate regression "controlling for X"
- What is this? Why do we do this? What should X be? What causal parameter does it help identify?
- Unconfoundedness, selection on observables, ignorable treatment assignment are different terms describing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

Which covariates?

- One of the values of causal directed acyclic graphs (DAG) is it allows you to formally select variables needed for covariate adjustment
- One such approach is the backdoor criterion which states that if you can condition on X such that all backdoor paths close, then you can identify some aggregate causal parameter
- But this requires a model, and I don't mean a theoretical model that you might learn as some abstract theory about education
- It's a model of treatment assignment, which is local in nature, and when it occurs outside the RCT requires expert knowledge

Simple DAG

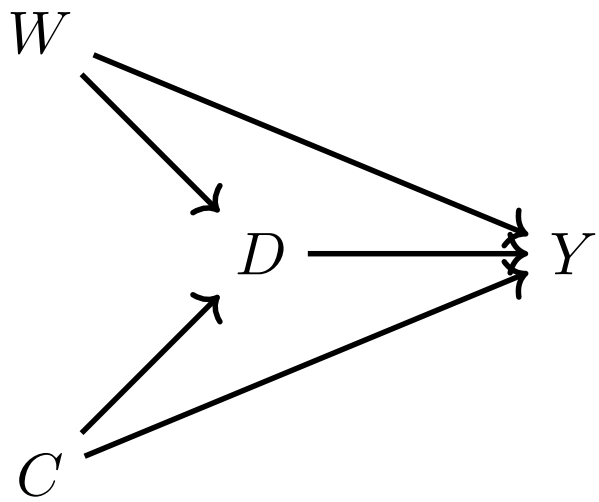
Figure: A simple DAG illustrating selection on observables.



Write down all paths, both direct from D to Y and indirect or “backdoor paths”

Simple DAG

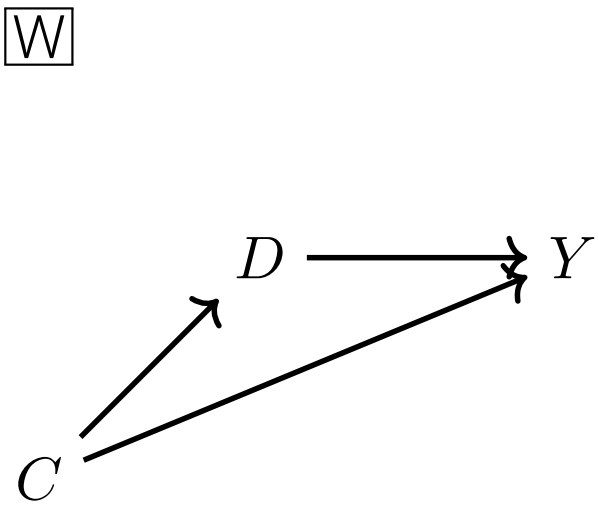
Figure: A simple DAG illustrating selection on observables.



1. $D \to Y$, the direct edge representing a causal effect with associated causal parameter like the ATE, ATT, etc.

Simple DAG

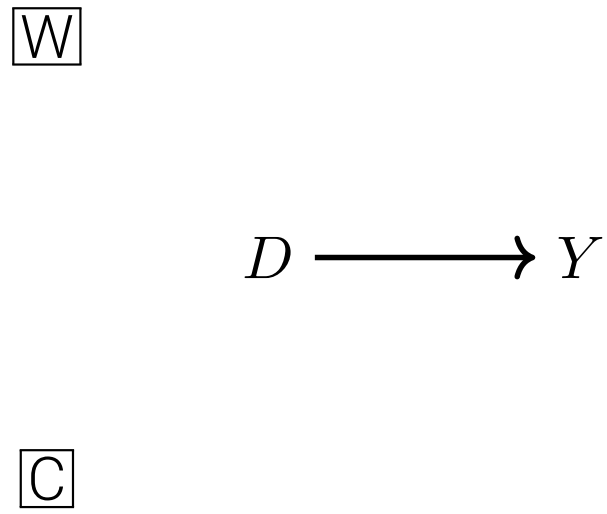
Figure: The same simple DAG illustrating selection on observables only with the direct edge from D to Y deleted and backdoor W blocked.



2. $D \leftarrow \boxed{W} \rightarrow Y$ is a backdoor from D to Y through W. Block it

Remaining variation after blocking

Figure: Visualization of Backdoor Criterion



2. $D \leftarrow \boxed{\text{W}} \rightarrow Y$ is a backdoor from D to Y through W. Block it
3. $D \leftarrow \boxed{\text{C}} \rightarrow Y$ is a backdoor from D to Y through C. Block it

Definition of Known and Quantified Confounders

Definition of a Known and Quantified Confounder

Variable C is a *known* and *quantified confounders* if the researcher believes it causes units to select into treatment ($C \rightarrow D$) and also independently determine outcome Y, or $C \rightarrow Y$. Confounders are always known, which requires prior knowledge. And to be quantified, they must be correctly measured in your dataset.

Known and Quantified Confounder

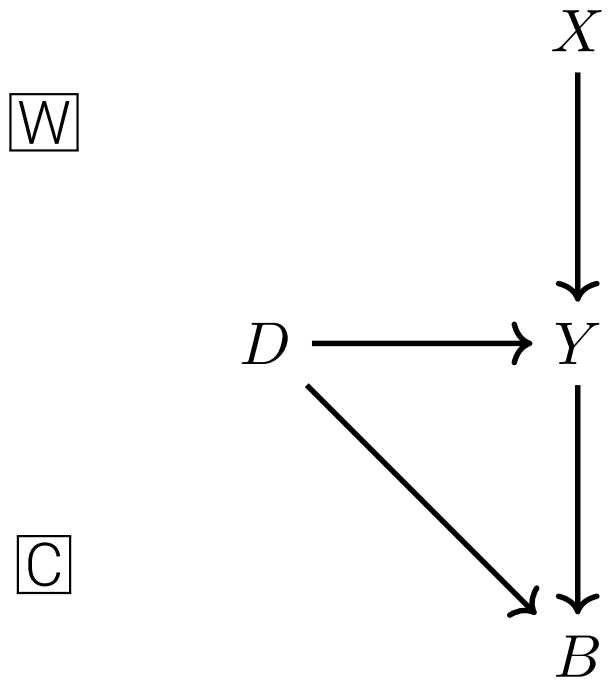
- Confounders may or may not be observed, but they must be known if they are confounders as confounders create backdoor paths from D to Y
- Visually, solid lines means they are “quantified” (i.e., in the data), whereas dashed lines mean they are either not defined correctly or not in the dataset (“unobserved”)
- Backdoor criterion is appropriate only for known and quantified confounders – if either known or quantified is missing, this material today is not to be used

DAG tells us what we need to condition on

- If we "block" on C and W, then the *only* explanation of why D and Y are then correlated is causal
- Depending on the model we estimate, and explicit assumptions made about potential outcomes, then we are able to identify an aggregate causal parameter
- We call C and W the "known and quantified confounders" because the model said these were necessary, they were observed (no dashed line) and they were confounders
- So what's a collider, and what's a covariate? Let's now add those into the simple DAG

Modification of the original DAG

Figure: A DAG illustrating confounders (W and C) versus colliders (B) versus exogenous covariates (X).



4. You cannot get from D to Y via X so it is not a backdoor path

Covariate

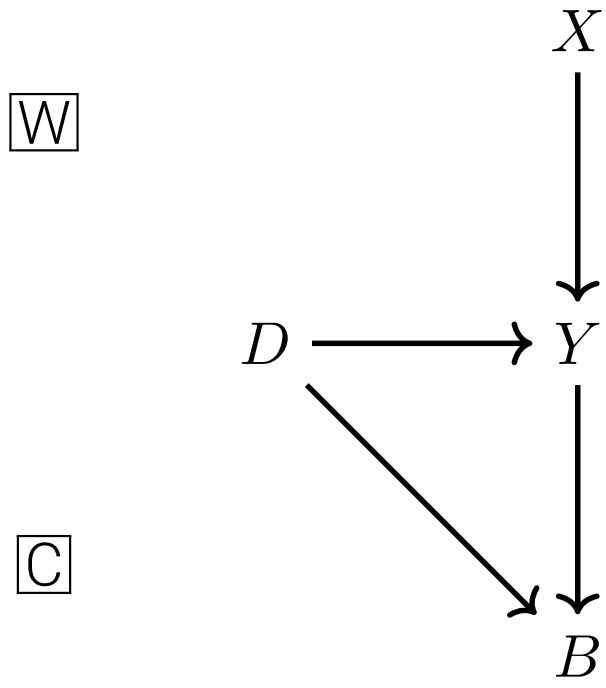
Definition of a Covariate

Variable X is a covariate if it causes Y but does not cause the treatment status D.

- Think of it as in the error term, but not correlated with the treatment variable
- Including X in a model can increase precision of estimates of D on Y simply by reducing residual variance, but should have no effect on point estimates
- Keep “confounder” and “covariate” distinct
- Covariates can be time invariant or change over the time – that’s not relevant

Modification of the original DAG

Figure: A DAG illustrating confounders (W and C) versus colliders (B) versus exogenous covariates (X).



5. You cannot get from D to Y via B so it is a collider, but if you control for it, that path opens up and introduces selection bias (“bad controls”)

Colliders

Definition of a Collider

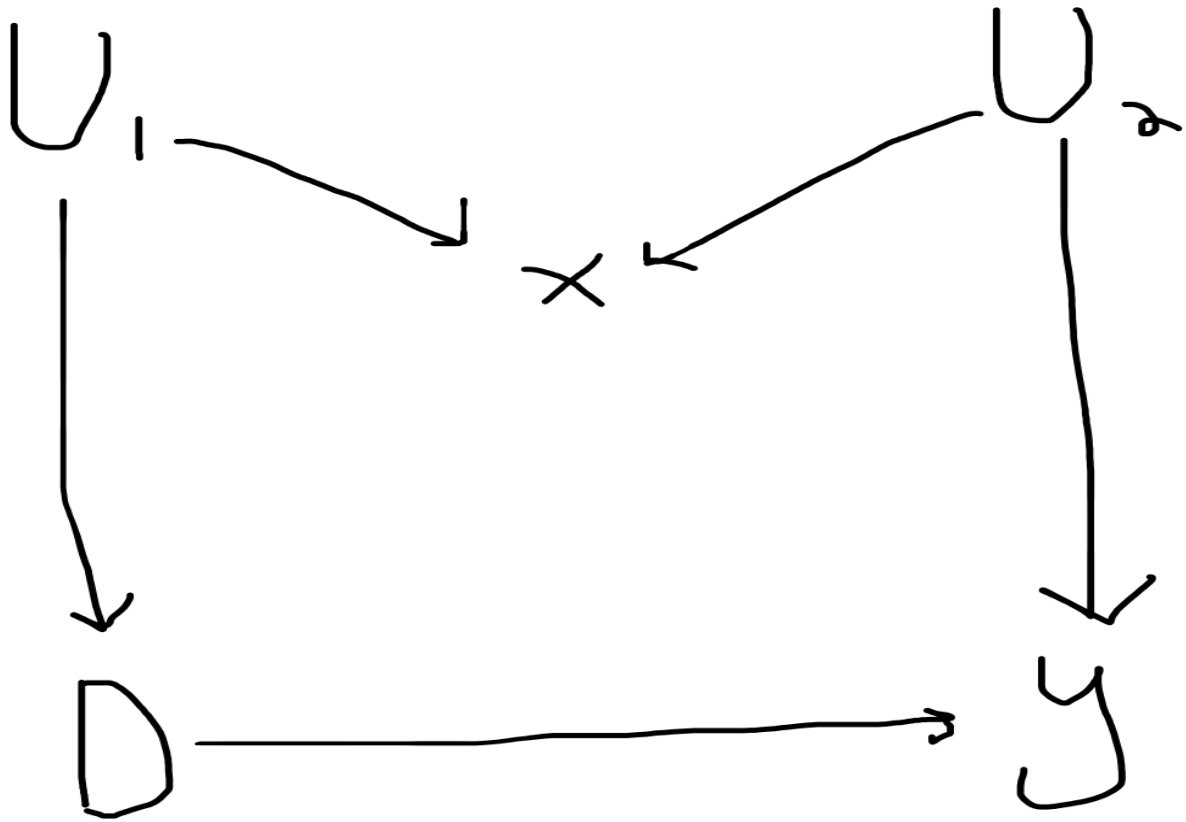
Variable B is a collider if there exists $D \rightarrow B \leftarrow Y$ along the path from D to Y.

- Colliders block backdoor paths so long as they are not blocked
- If you block on a collider, then the backdoor path opens, unless there exists a non-collider that you block to close it
- Conditioning on a collider introduces selection bias and depending on the magnitudes of $D \rightarrow B$ and $B \leftarrow Y$ relative to $D \rightarrow Y$, the distortion of estimated effect of D on Y may be extreme

Summarizing “which variables”

- Comparing treatment and control group of exactly the same values of known and quantified confounders will allow you to estimate aggregate causal parameters of interest
- Covariates can improve precision but do not reduce bias
- Colliders must be left alone, otherwise they introduce bias unless another non-collider can block them

M-bias as collider



X is pre-treatment, but if you conditioned on it, then $D \leftarrow U1 \rightarrow \text{X} \leftarrow U2 \rightarrow Y$ and X is a collider. Colliders take strange forms, so the conditioning set *must* be thoughtfully chosen based on being a strong predictor of Y^0, preferably based on expert knowledge and not a kitchen sink of all available regressors

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
- Big data approaches to covariate adjustment is *very dangerous* – colliders introduce bias and without a model, there is no way you know what those are

Covariate selection without DAGs

- Short of an outright DAG, then the thing to be thinking about is this:
  - → What set of covariates are highly predictive of Y^0?
  - → What set of covariates are highly predictive of D?
  - → Are these covariates distributed enough across both treatment and control?
- This is more of a hunch approach, but at least it's based on reasoning through the treatment assignment mechanism as opposed to "kitchen sink regressions"

Falsifications

- Covariates should not be affected by the treatment, so examining them as falsifications can help establish the credibility of unconfoundedness
- Falsification exercises are sometimes versions of this – a gun control law shouldn't affect automobile theft or petty larceny
- Imbens and Rubin (2015) suggested using the lagged outcome (pre-treatment) as a way of checking, as those have similar confounder structures
- 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)

Self selection based on gains

- Our focus on variables does not always clarify the exact nature of the matching and regression designs
- Covariates are about randomization *within* the dimensions of X
- If actors are randomly assigning themselves to treatments for identical values of X, then they are not choosing treatments optimally/rationally
- Rational actors almost by definition are thought to "self-select into treatment" making non-designed comparisons potentially misleading
- RCTs can overcome the selection bias that emerges when independence is violated, but are there other ways too?

ATE vs ATT

- The RCT will identify the ATE, but under randomization, the ATE=ATT=ATU because randomization of treatment is independent of all potential outcomes
- But outside the RCT, to identify the ATE requires a stronger set of assumptions than you need for ATT or ATU
- Questions is which parameter is it you actually do want to know for decision making?

ATE vs ATT

- Pfizer wants to vaccinate the world – since everyone will get the treatment, they want to know the ATE
- If the goal ultimately is to implement something for every person, usually the ATE is the parameter you want to know (e.g., vaccines)
- But there is also the ATT – the returns to the program for those people on the program (e.g., ventilators will only be given to compromised people not all people)

Population vs sample analogs of causal parameter

Define ATE as population mean treatment effects:

$$\delta^{ATE} = E[Y_i^1] - E[Y_i^0]$$

Define ATE sample analog as:

$$\frac{1}{N}\sum_{i}^{N}\delta_i = \frac{1}{N}\sum_{i}^{N}\left[Y_i^1 - Y_i^0\right]$$

Where N is the entire sample. This cannot be measured directly due to missing data on counterfactuals for both the treated and untreated units ("fundamental problem of causal inference").

Defined causal parameters and weights

ATE is a weighted average of ATT and ATU:

$$\delta^{ATE} = \pi \times \delta^{ATT} + (1 - \pi) \times \delta^{ATU}$$

where π is share of population in treatment group, and

$$\begin{aligned} \delta^{ATT} &= E[\delta_i | D_i = 1] \\ \delta^{ATT} &= E[\delta_i | D_i = 0] \end{aligned}$$

Show in an exercise, skip to WEIGHTS tab:

https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing

Imputation

> *"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others."*

For whichever parameter, we are always missing counterfactuals, so if we did design a method that estimated that parameter, it must be we imputed the missing conditional mean counterfactual

Some methods do this explicitly – like synthetic control – and some it's not clear, but don't ever forget that all of them are doing it

Identifying assumption I: Unconfoundedness

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D|X_i$. 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, they flipped coins as opposed to sought treatments that helped them
- Eliminating all backdoor paths on a DAG through blocking satisfies unconfoundedness; also called ignorability

Identifying assumption I: Unconfoundedness

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D|X_i$. 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 (next slide)

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, δ.

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 Y_i^0 and Y_i^1 from treatment group as j's 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 (Assumption 2) allows the match to take place as well as weight over the covariate distribution

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

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

You can get a PhD because you care about your happiness with one Y^1; you just have to have acted independent of what you gave up 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

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.

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

- Unconfoundedness and Ignorable Treatment Assignment
  - Choosing Covariates
  - Aggregate target parameters
- Matching Estimators
  - Stratification weighting
  - Conditional Independence
  - Exact and Inexact Matching
  - Propensity scores
  - Regressions
  - Coarsened exact matching
- Concluding remarks

History of stratification

- Adjusting for confounders was developed within statistics and epidemiology largely to study smoking's effect on lung cancer
- Couldn't run RCTs to examine smoking's effect, so people relied on non-experimental data, but results were not plausible to critics for a variety of reasons
- Stratification weighting was developed to adjust for the role that known and quantified confounders were playing through

Figure 1
Lung Cancer at Autopsy: Combined Results from 18 Studies

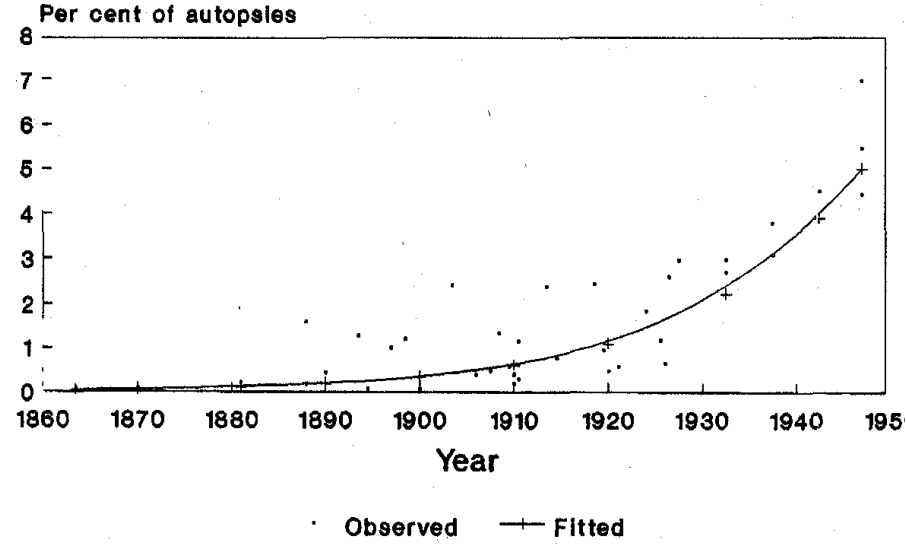


Figure 2(a)
Mortality from Cancer of the Lung in Males

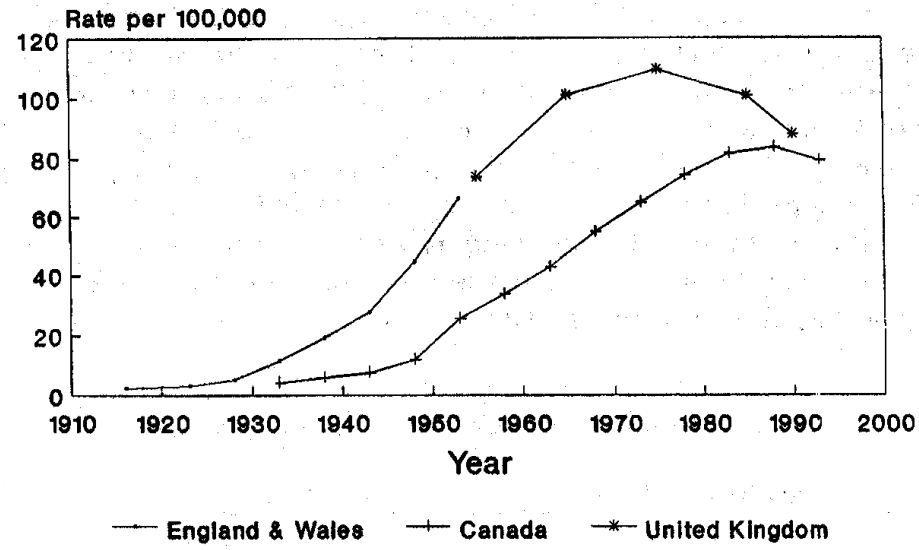
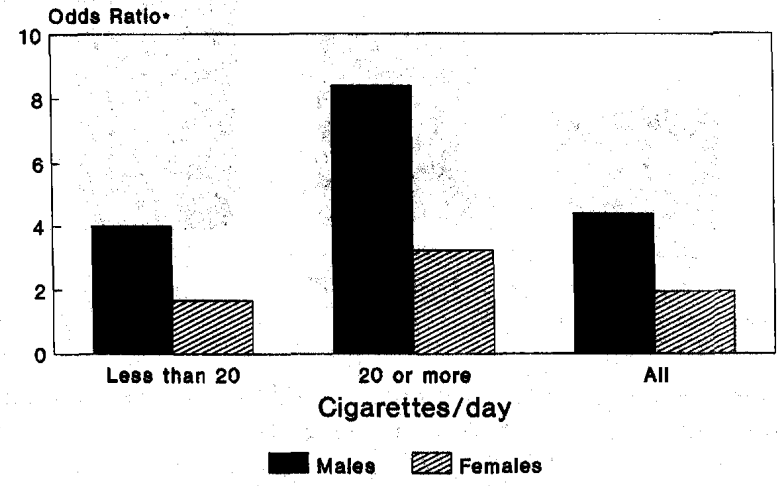
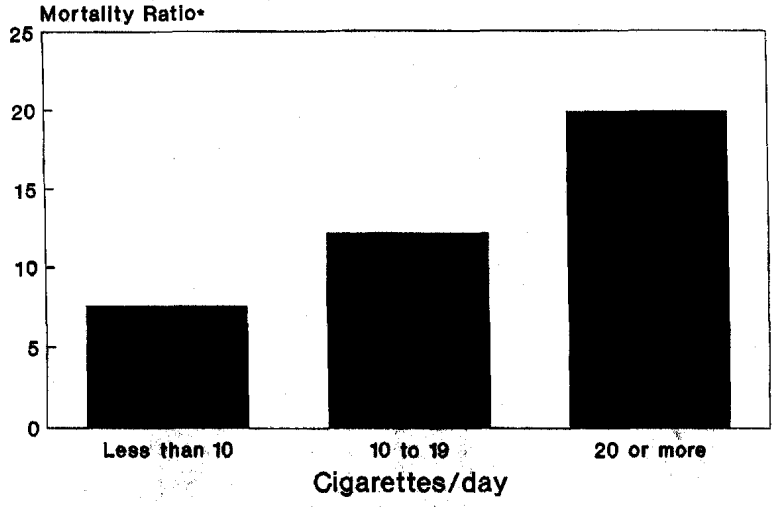


Figure 4
Smoking and Lung Cancer Case-control Studies



* Weighted mean

Figure 5
Smoking and Lung cancer Cohort Studies in Males



* weighted mean of 4 studies

Hypothesis and skepticism

Early 20th century scientists believed smoking caused lung cancer but others felt the evidence was not strong

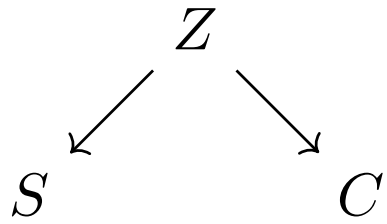
1. Sample bias due to non-random selection of subjects (only those who died)
2. Complaints about functional forms using “risk ratios” and “odds ratios”
3. Implausible magnitudes (“smell test”)
4. Killer critique: *no experimental evidence* to incriminate smoking as a cause of lung cancer

Fisher's genome confounder theory

- Ronald Fisher (recall from earlier) was a heavy smoker, died of cancer, and paid expert witness for the tobacco industry
- Famous as statistician and as a geneticist and using logic, statistics and genetic evidence proposed a contrarian theory that there existed a confounding genome, Z, which introduced selection bias into contrasts of smokers and non-smokers
- Studies showed that cigarette smokers and non-smokers were different on observables – more extraverted than non-smokers, differed in age, differed in income, differed in education, etc.

Fisher's genome confounder theory in DAG form

Smoking, S, is only *correlated* with lung cancer, C because of an unknown ant not quantified confounder Z:



Legitimate criticism, but ultimately incorrect:

> *"the [epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions." Robert Hooke (1983)*

Stratification weighting

- Simple weighting techniques were eventually introduced to adjust for these observables
- Stratification weighting goes back at least to Cochran (1968)
- We will discuss it both using his original paper, and an application
- Goal is to adjust quantities for a known and quantified confounder through weights based on the confounder's distribution in different samples

Mortality rates by country and smoking group

Table: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Mortality rates by country and smoking group

- Cigars in these data are particularly odd to me – much higher mortality rates in all three countries than non-smokers, as well as cigarette smokers.
- Another strange result – cigarette smokers in Canada and US had same mortality rates as non-smokers
- Seems weird to us, but unclear what they would've thought given the science was unsettled
- Cochran tackles *one* of the observable variables that he thinks predicts smoking and mortality – age

Non-smokers and smokers differ in age

Table: Mean ages, years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Maybe age is a confounder. It predicts smoking, but it also predicts mortality as cumulative risk of dying grows as we age

Imbalanced confounders and covariates

- Balance is a phrase often heard in causal inference – means treatment and group have different covariate/confounder means and/or distributions
- If confounders are *imbalanced*, then their means differ in each group, and will introduce selection bias (if confounders, but not if covariates)
- Weighting, matching and regression that adjust for known and quantified confounders attempt to create balance on confounders and covariates so that their effects stop

General description of stratification weighting

1. Stratify the confounder by slicing it into “strata” (e.g., age is young vs old)
2. Compare quantity of interest across each treatment status within each strata
3. Create strata specific weights
4. Weight quantity of interest by share of units across all strata

Smoking and stratification weighting

1. Stratify the smoking group into differing age groups or “bins” or “strata” (e.g., young and old)
2. Calculate mortality rates separately for young and old and treatment and control (four averages)
3. Construct “probability weights” as the proportion of each smoking group sample within a given age group
4. For treatment and control group, compute the weighted averages of the age groups mortality rates using the probability weights

This procedure will balance the observed covariate, age, between treatment and control

Smoking and stratification weighting

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: Calculate average pipe smoker death rate w/out stratification?

Smoking and stratification weighting

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: Calculate average pipe smoker death rate w/out stratification?

$$15 \cdot \left(\frac{11}{40}\right) + 35 \cdot \left(\frac{13}{40}\right) + 50 \cdot \left(\frac{16}{40}\right) = 35.5$$

Smoking and stratification weighting

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Counterfactual question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

Smoking and stratification weighting

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Counterfactual question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

$$15 \cdot \left(\frac{29}{40}\right) + 35 \cdot \left(\frac{9}{40}\right) + 50 \cdot \left(\frac{2}{40}\right) = 21.2$$

Table: Adjusted death rates using 3 age groups (Cochran 1968)

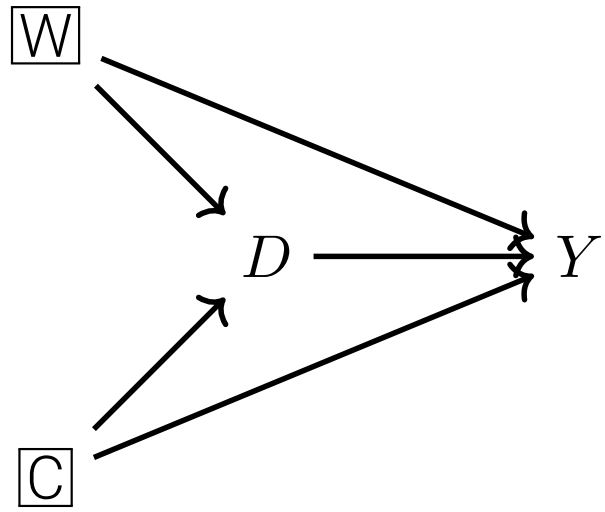
Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Exercise: Titanic sinking

- 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

Go to lab

Table 1: Simple counts

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

Variable name	**First class**		**All other classes**	
	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.

	First class		All other classes		
Strata	Obs	Mean	Obs	Mean	Total
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 parameter

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

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**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
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	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.

	First class		All other classes		
Strata	Obs	Mean	Obs	Mean	Total
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

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

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

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
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**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	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"

Potential outcomes

- Now let's introduce potential outcomes notation so that we can build more formalized models
- This helps us understand what assumptions we need to make, which is not necessarily spelled out in the DAG backdoor criterion

Recall the RCT Assumption of Independence

- Randomized treatment assignment guarantees "independence"

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

- Independence allows to estimate accurate causal effects through simple methods like differences in averages

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{by the switching equation}} \\ &= \underbrace{E[Y^1] - E[Y^0]}_{\text{by independence}} \\ &= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} \end{aligned}$$

Confounder and covariate distribution

- Just like independence implies balance on expected potential outcomes, it also implies balance on confounders and covariates which is called “common support”
- We saw this in our Thornton regressions: cash vouchers were not associated with being male, one’s age, etc.
- If we have balance on potential outcomes, that’s all we need but as that’s not observed, balance on covariates is often used to provide some evidence the randomization was done well

Violations of independence

- Problem with smoking and cancer was smoking wasn't *randomly assigned* – it was chosen by the "perfect doctor" (i.e., self selection into smoking based on factors related to potential outcomes)
- When a treatment is "dependent" on potential outcomes, it means people smoke because they expect something is better when they smoke (Y^1) than when they don't (Y^0) introducing selection bias and potentially heterogenous treatment effect bias
- Naive comparisons can be deeply misleading – covariate adjustment can resolve this if "conditional independence" happens in the data
- How do we express backdoor criterion using potential outcomes notation?

Identifying assumption I: Conditional independence

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. 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 **random**
- For a large group of people within the same strata, they flipped coins as opposed to sought treatments that helped them
- Sometimes this is called also “unconfoundedness” and worth thinking long and hard about

Identifying assumption I: Conditional independence

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

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

Allows us to write down these equalities, which means we can use comparison units as substitutes for counterfactuals so long as there exist one-to-one within $X = x$, leading to our next assumption – common support

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
- 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, and often this is where regression steps don't incorporate it

Caveat: Perfect Doctor and conditional independence

- Independence was violated if the treatment was assigned *because* we expected things to improve or not (“perfect doctor” reasoning)
- If you take an action because you think it helps and others don’t take the action because they don’t or can’t, then it is a violation of independence probably
- For a large group of people within the same strata, they flipped coins as opposed to sought treatments that helped them
- Rationality is contained in the confounders – worth reflecting a lot on

Assumptions combined

We need only two assumptions to estimate the ATE, but these are not trivial:

1. $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
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, δ.

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

Implications of assumptions

- Assumption 1 lets you plug Y for Y^j with the switching equation

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

- Assumption 2 lets you weight over the covariate distribution

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

You need fewer assumptions for ATT or ATU

Other versions of conditional independence

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

Notice how there is only one potential outcome in the independence equation. That's okay. We can still then estimate the ATT (just not the ATE).

Summarizing

Weighted averages under either assumption:

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

ATE needs independence with respect to both potential outcomes; ATT only needs it with respect to Y^0.

Review stratification weighting in light of this

- Let's now look at the way in this works within stratification if we are missing some observations
- The reason that this happens is because the dimensions K of the X conditioning set is getting larger

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

$$4 \cdot \left(\frac{13}{30} \right) + 6 \cdot \left(\frac{17}{30} \right) = 5.13$$

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{7}{10} \right) = 5.4$$

Weighting by Age and Gender ($K = 4$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Weighting by Age and Gender ($K = 4$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Not identified! What went wrong?

Weighting by Age and Gender ($K = 4$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Weighting by Age and Gender ($K = 4$)

	Death Rate			Number of	
X_k	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 5 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{4}{10} \right) = 5.1$$

Curse of Dimensionality

- Stratification methods, including OLS, may become less feasible in finite samples as the number of covariates grows (e.g., $K = 4$ was too many for this sample)
- 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 problem isn't just the number of covariates; it's the number of strata based on those covariates (you can hit the curse fast)

Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both, and that violates our second 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 doesn't tell us it is just doing various extrapolations
- Simple weighting methods is also a problem if the cells are "too coarse"

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

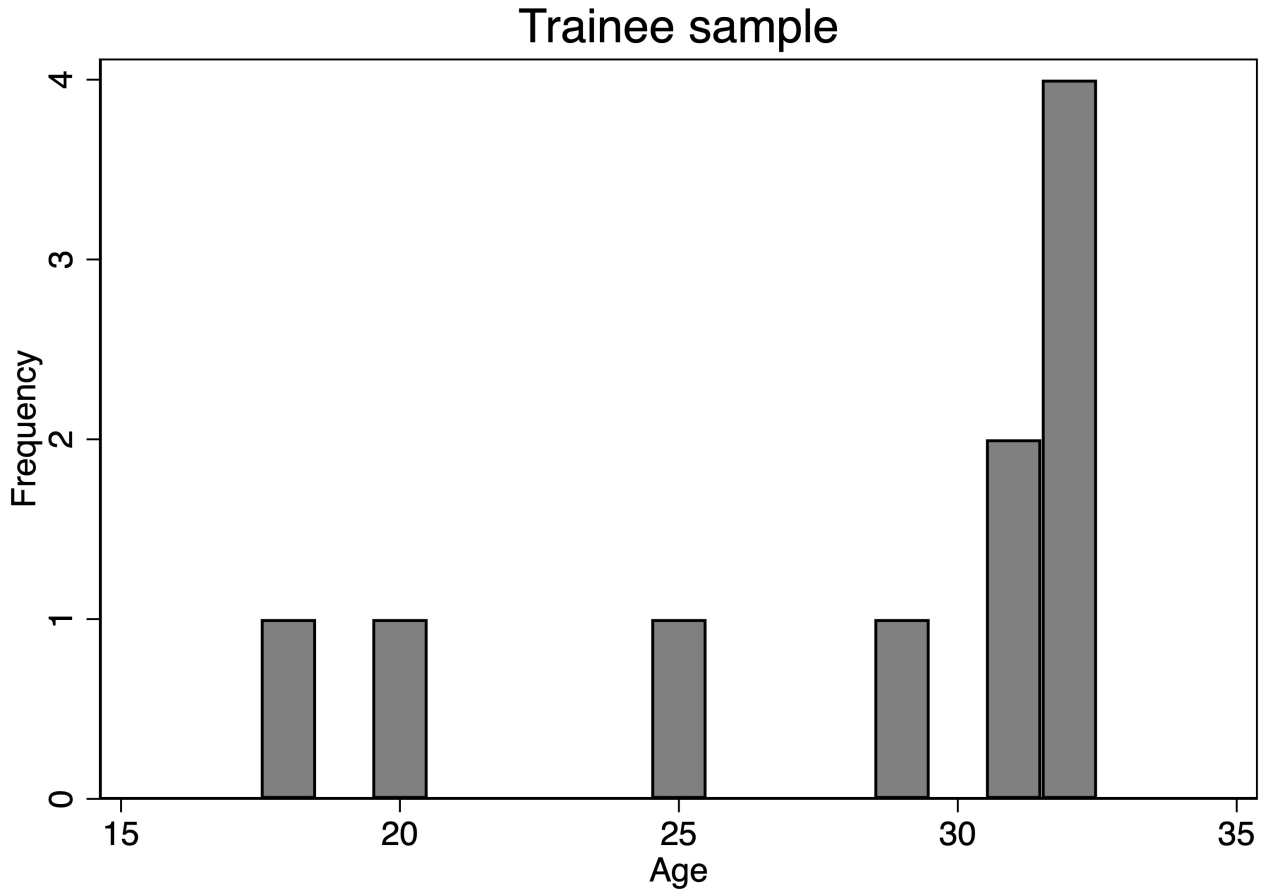
Training example (unmatched)

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

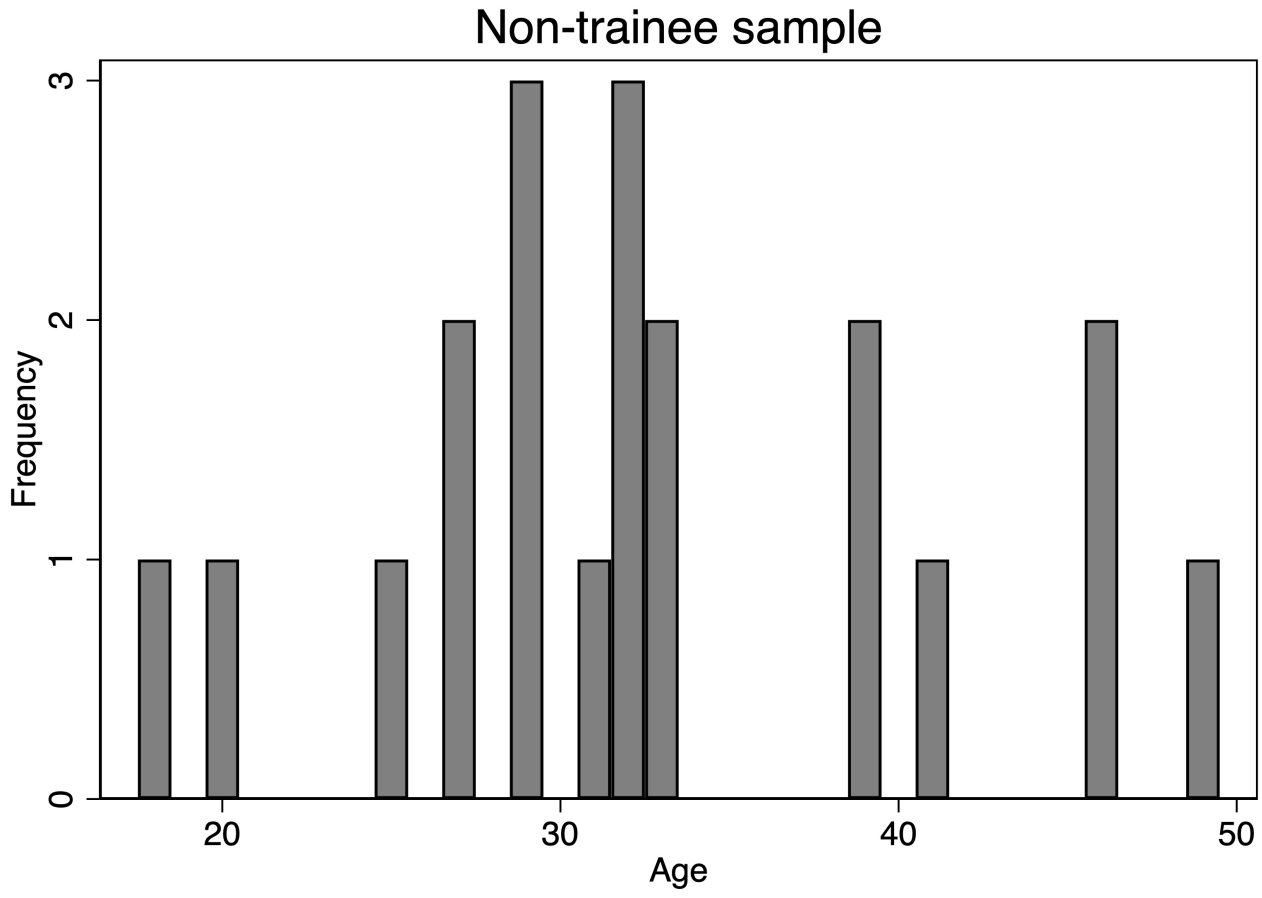
$SDO = \$24,150 - 27,923 = -\$3,773$

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:

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

where $Y_{j(i)}$ is the j^{th} unit matched to the i^{th} unit based on the j^{th} being "exactly equal to" the i^{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?

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

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, $\widehat{\delta_i} = Y_i - \frac{1}{M}\sum Y_{j(i)}$.
4. Finally, estimate the sample ATT by averaging over all i differences in earnings from Step 3 as $\frac{1}{N_T}\sum \widehat{\delta_i}$, 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

Propensity scores

- Propensity scores were developed by Rosenbaum and Rubin (1983) as a way of reducing the dimension of the conditioning set of X
- But it still requires common support – propensity scores don't solve the curse of dimensionality problem from the perspective of bias
- If you don't have exact matches in the dimensions of X then you'll still be matching units with similar propensity scores but it won't overcome the bias
- There are methods that will adjust the propensity score estimation which I'll briefly mention at the end

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:

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

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

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

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_LJVzOYvnQZWmxY8JyTcDf6YcdhkwQ/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)

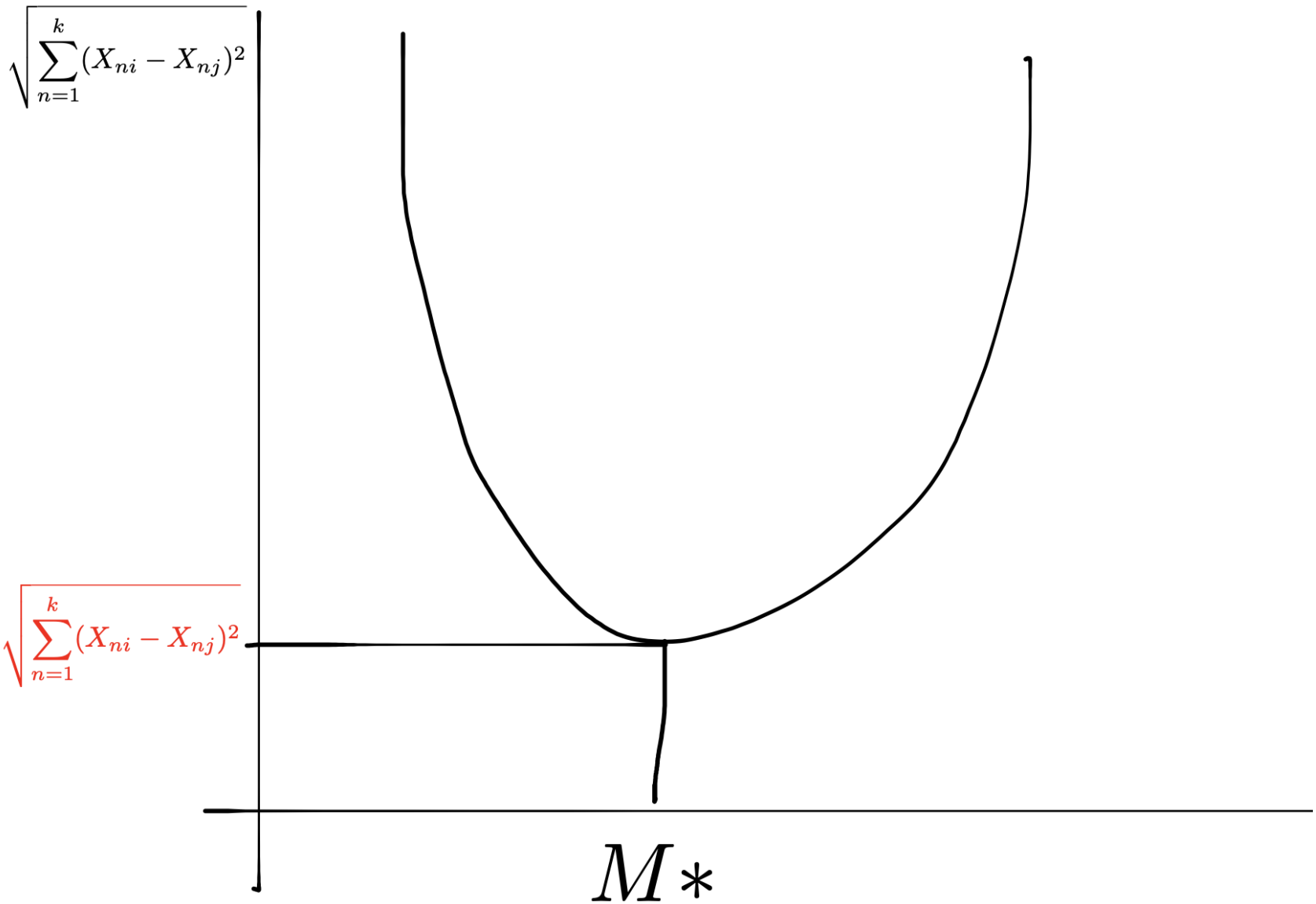
Trainee sample				Non-Trainees				Matched sample #2			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	13	22	1.66	8950
2	29	2.80	12250	2	27	1.78	10075	5	38	1.61	12550
3	24	3.92	11000	3	21	1.84	8725	1	20	1.89	8500
4	27	2.29	11750	4	39	1.76	12775	20	48	1.87	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:

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

where

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

Normalized Euclidean distance

- Notice that the normalized Euclidean distance is equal to:

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

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

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

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

where $\widehat{\Sigma}_X$ 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

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

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.

$$\begin{aligned}
\mu^0(x) &= E[Y|X = x, D = 0] = E[Y^0|X = x], \\
\mu^1(x) &= E[Y|X = x, D = 1] = E[Y^1|X = x], \\
Y_i &= \mu^{D_i}(X_i) + \varepsilon_i
\end{aligned}$$

Deriving the matching bias

Substitute and distribute terms

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

Deriving the matching bias

Difference between sample estimate and population parameter is:

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

Algebraic manipulation and simplification:

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

Deriving the matching bias

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

Deriving the matching bias

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

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

Deriving the matching bias

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

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

Now consider the implications if k is large:

Deriving the matching bias

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

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

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

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

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

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

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
- $E[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1]$ may not converge to zero and an be very large!

Deriving the matching bias

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

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

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

Deriving the matching bias

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

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

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
- $E[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1]$ may not converge to zero and an be very large!
- $E[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})]$ 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:

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

  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 $\widehat{\alpha}$ and $\widehat{\beta}$ to predict $\widehat{\mu}^0(X)$ for both the i and the matched $j(i)$ units:

$$\begin{aligned} \widehat{\mu}_i^0 &= \widehat{\alpha} + \widehat{\beta} X_i \\ \widehat{\mu}_{j(i)}^0 &= \widehat{\alpha} + \widehat{\beta} X_{j(i)} \end{aligned}$$

Steps

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

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

Steps

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

Bias adjustment in matched data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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

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

Bias adjustment in matched data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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 $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

Bias adjustment in matched data

	Potential Outcome			
unit	under Treatment	under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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 $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\widehat{\delta}_{ATT} = \frac{(10-8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4-1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3}$$

$$+ \frac{(10-9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33$$

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):

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

- For matching without replacement, the "usual" variance estimator:

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

  is valid.

Large sample distribution for matching estimators

- For matching with replacement:

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

  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 $\widehat{var}(\varepsilon_i|X_i, D_i=0))$

Propensity score as dimension reduction

- Curse of dimensionality makes matching on K covariates challenging as the dimensions grew with K (and continuous covariates had no exact matches)
- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method to reduce K covariates into a single scalar without loss of information
- Insofar as treatment is random conditional on K covariates, then it will also be random conditional on propensity score
- Variety of ways to incorporate the propensity score, but first we describe the propensity score as a dimension reduction method

What's the propensity score for?

- Once you obtain the propensity score, you typically use it to carefully explore problems with common support with respect to your selection parameter of interest
- 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 and bootstrapping
- Can even introduce "doubly robust" methods to deal with matching bias like we did with bias correction

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

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 (ATT recall needs weak unconfoundedness)

True vs Estimating Propensity Score

- In an experiment, we know the true propensity score because we controlled it – could be uniform, could be different by groups
- Outside an experiment, we don't know it
  - → We use DAGs or hunches to select covariates, careful not to include outcomes or colliders
  - → But we may still not know the functional form
- 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

Estimating with the Propensity Score

- All the propensity score will do is collapse or “reduce” the dimensions of X from K to a single scalar, but then what?
- Weight or impute – only things you can do
- We will start with weighting and then move into imputation after
- Inverse probability weighting to estimate the ATT

Step 1: Estimate the propensity score

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

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

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

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

- Note that each unit i now has a predicted probability of treatment given the values of their covariates relative to everyone else's
- Frequentist probability – you've basically just obtained the likelihood someone who "looks like you" would be treated (regardless of whether you were in fact treated)

Step 2: Estimation of ATT with IPW

- IPW uses the estimated propensity score to reweight the outcomes (e.g., Robins and Rotnitzky 1995, Imbens 2000, Hirano and Imbens 2001)
- IPW is non-parametric – you are just taking averages and multiplying by weights
- There are also fewer implementation choices – you aren't choosing how many neighbors to include, how far away a neighbor can be – but you still have to closely examine common support
- There are bias adjustment methods called double robust where you combine imputing counterfactuals with weighting by the propensity score

Step 2: 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 the treatment group Y values alone and weight "up" or "down" the comparison groups by their propensity scores

Step 3: Standard Errors

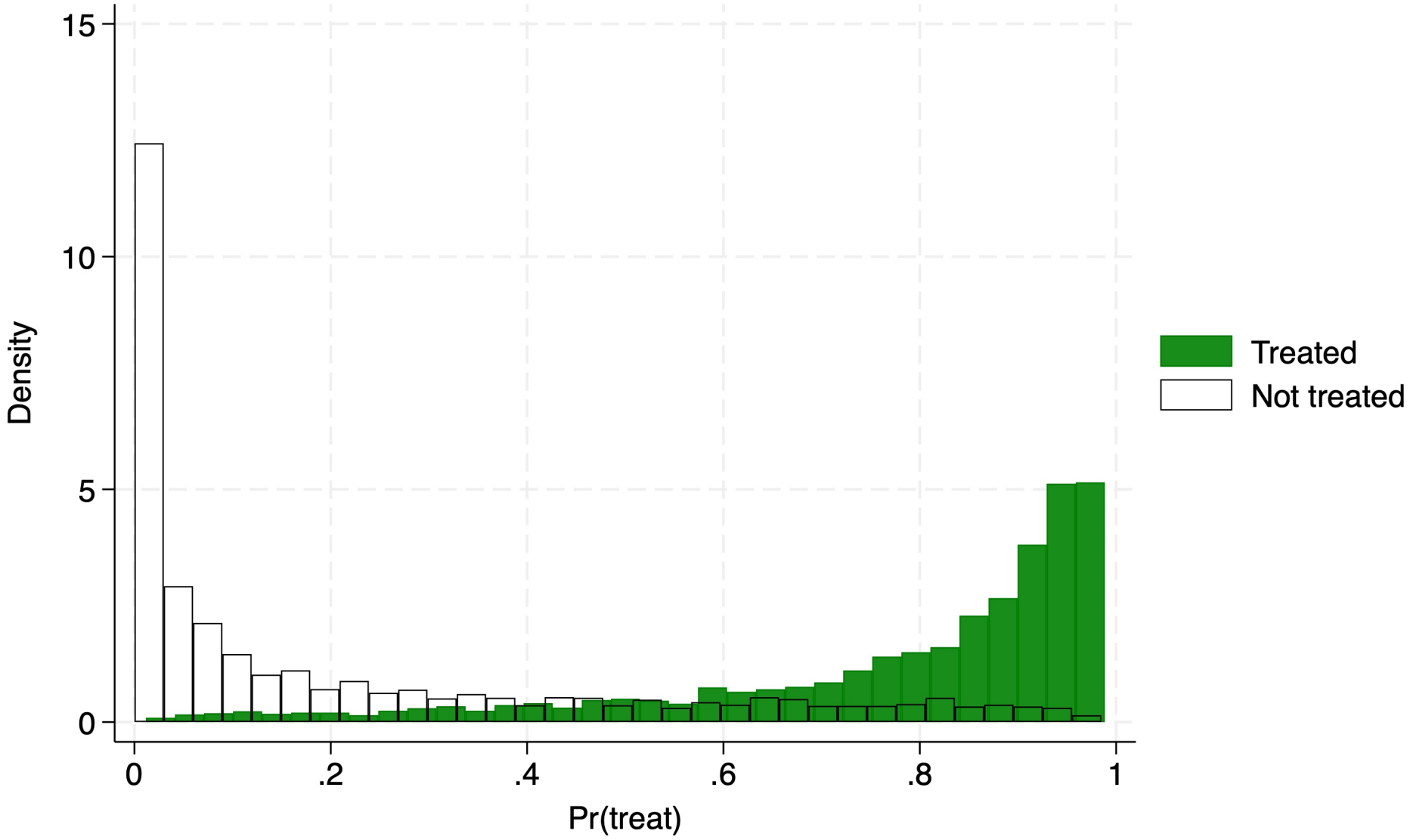
Standard errors can be constructed a few different ways:

- 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

- Recall that the matching/weighting methods all require unconfoundedness *and* common support, and whereas the first is not testable, the latter is
- Assessing whether there are units it both groups for whichever parameter you're focused on is simple with propensity score – just create histograms of the propensity score distribution for treatment 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
- "Parameters first!"

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 pursue the ATE if that is your preferred estimate, which also requires full unconfoundedness/support to be believable

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

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

- Lots of papers began to try and address the model misspecification problem by combining propensity scores with other methods called "double robust" (Robins and Rotnizky 1995; Hirano and Imbens 2001)
- Basic idea in all of them was to control for covariates twice *at the same time* without paying a price
- 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

Propensity score matching

- Matching, or “imputation”, is another way that utilizes the $\widehat{\rho_i}(X_i)$
- 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 $\widehat{\rho_j}(X_j)$ falling within a radius r from $\widehat{\rho_i}(X_i)$ 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

OLS with controls

- Most common causal model, arguably, is OLS with covariate controls used to establish mean exogeneity and eliminate omitted variable bias

$$Y_{it} = \alpha + \delta D_{it} + \beta X_{it} + \varepsilon_{it}$$

- This as it turns out is related to what we've discovered but the ways that is different matter a lot for estimation and unbiasedness

Interpreting OLS coefficients

- OLS can do a lot of things, but one thing it does it find the best linear fit of the data
- Finding the best linear fit is the same as finding the counterfactual, so to get OLS to serve that function is another task

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

- When we estimate this model, it can be easily visualized with data

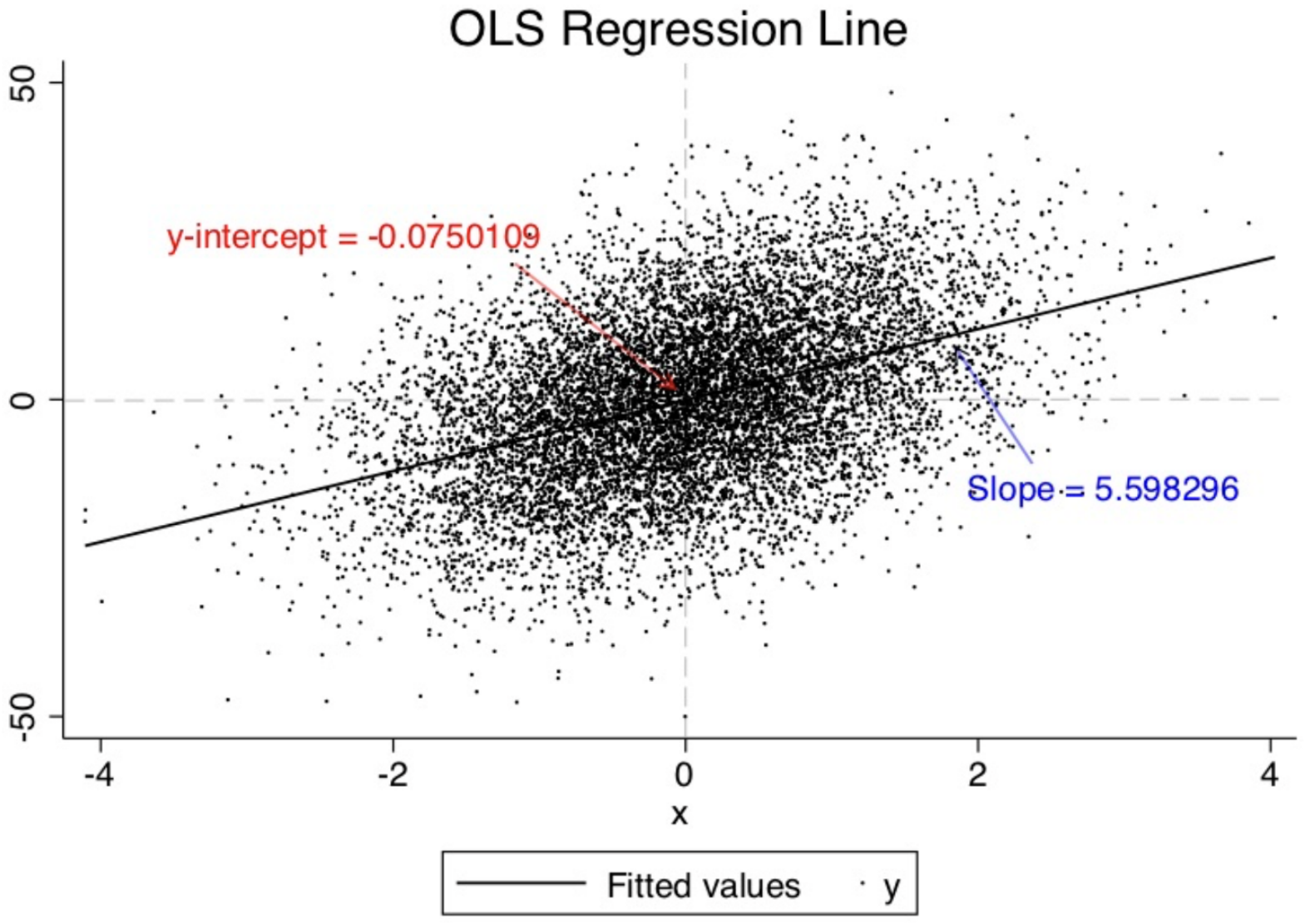
OLS Simulated Code

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(bla
msize(tiny) msymbol(point)), title(OLS Regression Line)
rvfplot, yline(0)
```

OLS Plot and Line



Bivariate vs multivariate regressions

- Interpreting the OLS coefficient in bivariate regression also has a simple express as a scaled covariance:

$$\widehat{\beta_1} = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

- But what about when we are looking at a multivariate regression – too many dimensions to visualize it
- A trick called the Frisch-Waugh-Lovell (FWL) theorem will turn the multivariate regression into the simple bivariate one
- FWL also helps us interpret $\widehat{\beta_1}$ when there are covariates

Applying FWL theorem

- Angrist and Pischke (2009) call FWL the “regression anatomy theorem”, as does Filoso (2013)
- Filoso (2013) has an excellent simplification of what is a historically complex proof

Applying FWL theorem

- FWL theorem concerns multiple linear regression and helps us interpret regression coefficients with multiple controls.
- 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 number of kids is uncorrelated with the unobserved error term, which means we can interpret $\widehat{\beta_1}$ as the ATE.
- But how do we interpret $\widehat{\beta_1}$ if `family size` is non-random?

Applying FWL theorem

- Assume that family size is random once we condition on race, age, marital status and employment. Then the model is:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \gamma_1 \texttt{White}_i + \gamma_2 \texttt{Married}_i \\ &\quad + \gamma_3 \texttt{Age}_i + \gamma_4 \texttt{Employed}_i + u_i \end{aligned}$$

- If we want to estimate average causal effect of family size on labor supply, we will need two things:
  1. a data set with all 6 variables;
  2. Number of kids (X_i) must be randomly assigned conditional on the other 4 variables
- How do we interpret $\widehat{\beta_1}$ with the additional controls?

Applying FWL theorem

- FWL shows you that in a multivariate regression, any one coefficient fitted can be reconceived as a simple scaled covariance of two variables
- It says that $\widehat{\beta_1}$ is simply a scaled covariance with the $\tilde{X_1}$ residual used instead of the actual data X

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 $\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}$ being the residual from the auxiliary regression. The parameter β_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 $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\widehat{x}_{ki}] = E[f_i]$, 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

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

FWL Proof (Proof by Filoso 2013)

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

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

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

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

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

FWL Proof (Proof by Filoso 2013)

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 orthogonoality 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 $\widehat{\beta_1}$ in code
- We will use the fwl.do at /Labs/Matching for this

Core OLS assumptions

- Most common causal model is OLS with covariates is additive in controls:

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

- Which average treatment effect parameter does $\widehat{\delta}$ estimate? What assumptions are imposed by exogeneity?
- Next we explore the OLS model and its ability (or not) to recover a given parameter of interest
- 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

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, α, δ_{sp} and β 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 tht outcome would change in expectation by an amount δ_{sp}. Hence in the potential outcomes formulation, we have

$$\begin{aligned} Y_i^0 &= \alpha + X_i\beta + \varepsilon_i \\ Y_i^1 &= Y_i^0 + \delta_{sp} \end{aligned}$$

OLS Assumptions

"Then, because ε_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)

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

$$\begin{aligned} E[Y_i|D_i = 1, X_i] &= \alpha + \delta E[D_i|D_i = 1, X_i] + \beta_1 E[X_i|D_i = 1, X_i] \\ E[Y_i|D_i = 0, X_i] &= \alpha + \delta E[D_i|D_i = 0, X_i] + \beta_1 E[X_i|D_i = 0, X_i] \end{aligned}$$

Constant Treatment Effects and Linearity

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

$$
\begin{aligned}
E[Y_i^1|D_i = 1, X_i] &= \alpha + \delta + \beta_{11} E[X_i^1|D_i = 1, X_i] \\
E[Y_i^0|D_i = 0, X_i] &= \alpha + \beta_{01} E[X_i^0|D_i = 0, X_i]
\end{aligned}
$$

May seem somewhat unorthodox to also let X have potential status, but you'll see why in a minute

Constant Treatment Effects and Linearity

OLS Estimator is a simple difference in conditional means:

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

OLS model requires three things: (1) linearity, (2) covariates to be independent of treatment status (i.e., treatment cannot cause covariates to change), (3) $\beta_{11} = \beta_{01}$ (homogenous treatment effects with respect to X).

Simulation

- Following code will maintain linearity but have common support violation to show OLS does not require common support, but does require linearity (it extrapolates based on functional form which is quite spectacular with correct model)
- I will also create heterogenous treatment effects
- But I will also violate the previous requirement that $\beta_{11} = \beta_{01}$ so that you can see the bias that forms on average across 1,000 simulations
- Will show a variety of estimators and specifications so that we see how to recover causal parameters with regression and matching

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
```

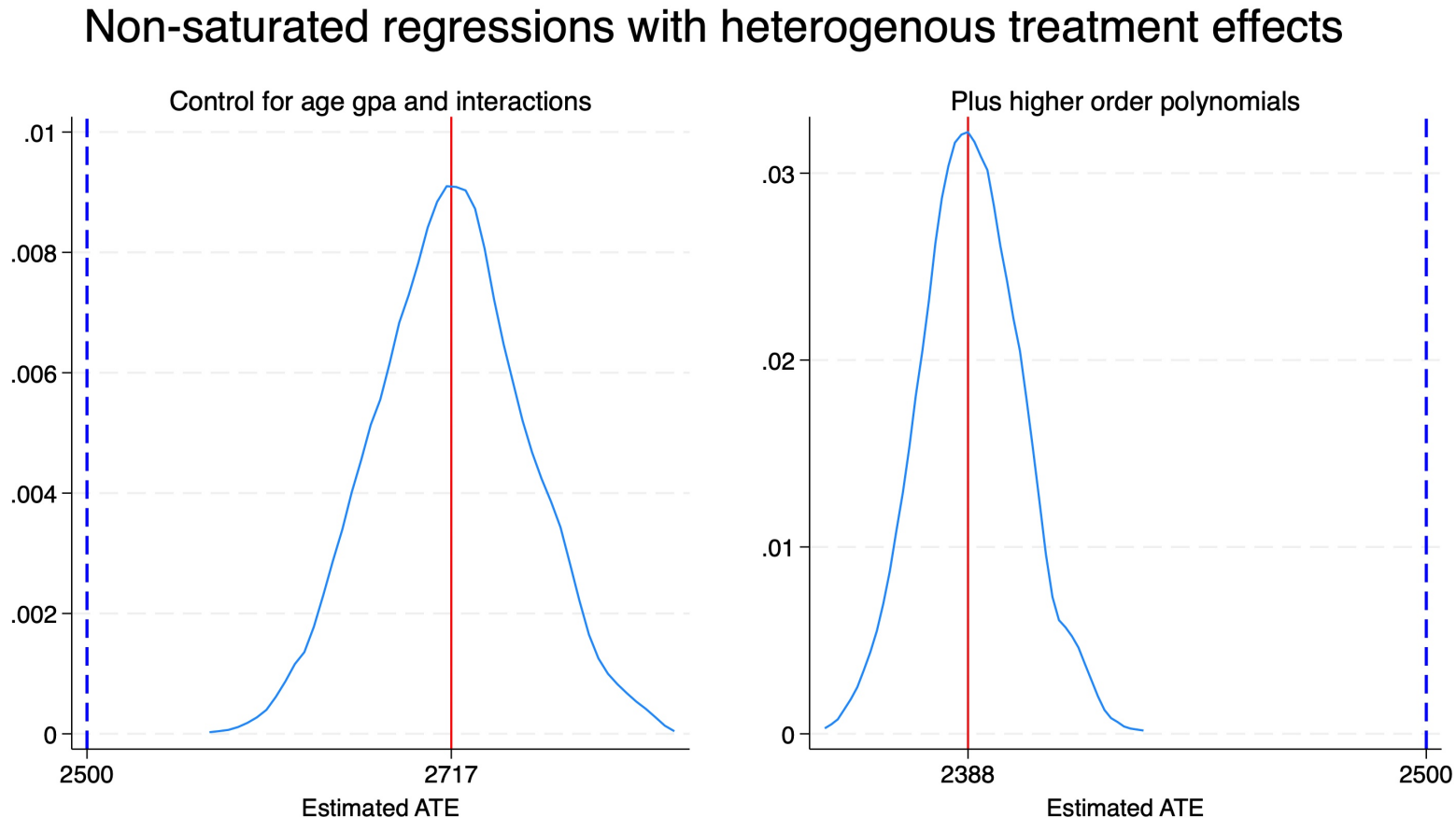
Parameters

- Ordinarily we look at the coefficient on the treatment dummy to obtain an estimate
- But we have two parameters: the ATE is $2500 but the ATT is $1980
- How do we get both 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



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)
- So what had we done? Bear with me but this will pay off

Heterogenous treatment effects

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

$$
\begin{aligned}
Y_i^0 &= \alpha + \beta_{01} X_i^0 + \varepsilon_i \\
Y_i^1 &= \alpha + \beta_{01} X_i^0 + \delta D_i + \beta_{11} X_i^1 \times D_i + \varepsilon_i
\end{aligned}
$$

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 x'effects

Take conditional expectations of the *potential* outcomes:

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

Average treatment effect is:

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

assuming $X_i^1 = X_i^0$. OLS model accounting for heterogeneity must be "fully saturated".

Estimation

Our saturated OLS model is:

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

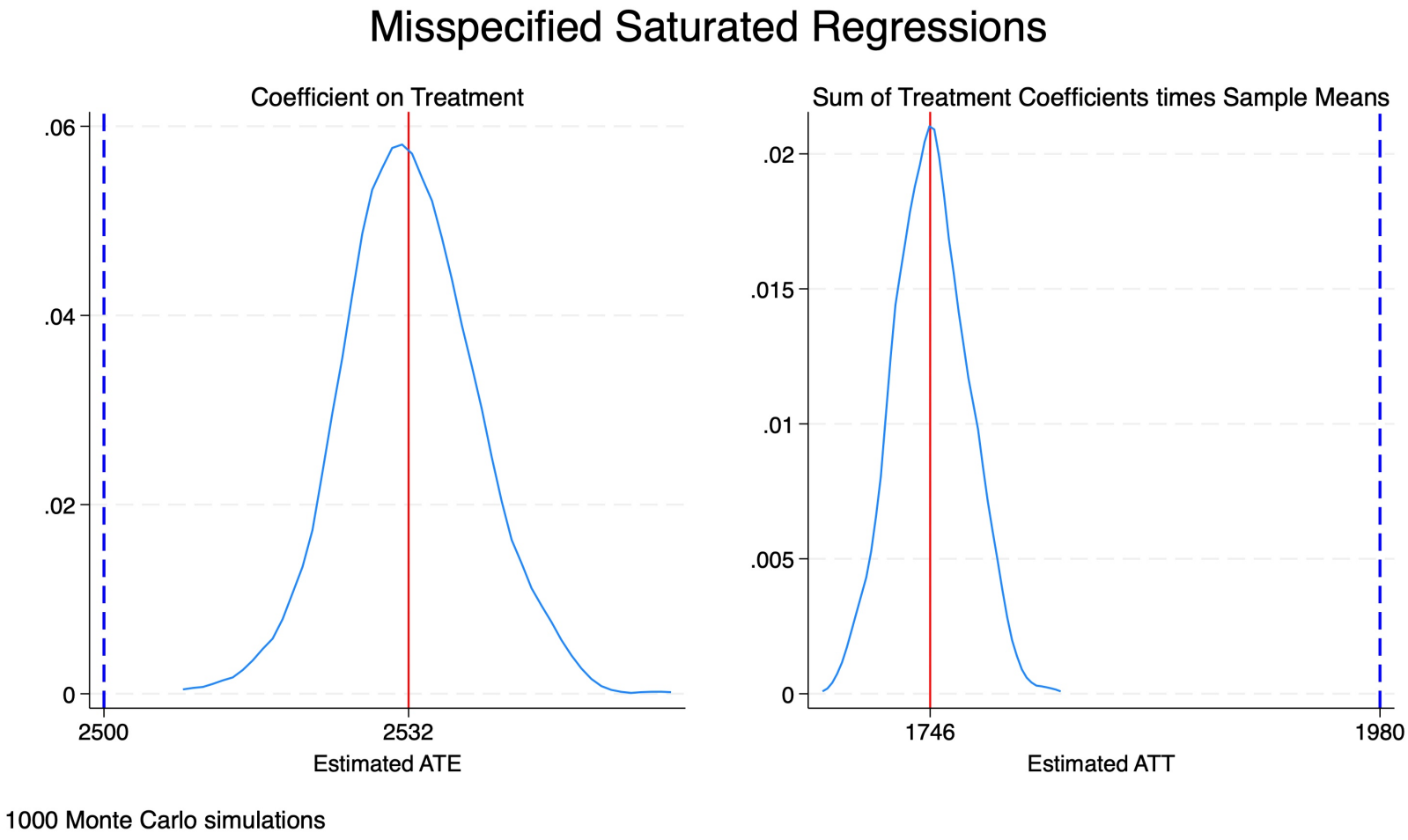
$\widehat{\delta}$ is the ATE but the ATT is equal to $\widehat{\delta} + \beta_{11} E[X_i | D_i = 1]$ where $E[X_i | D_i = 1]$ is the sample average of X_i for the treatment group

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
```

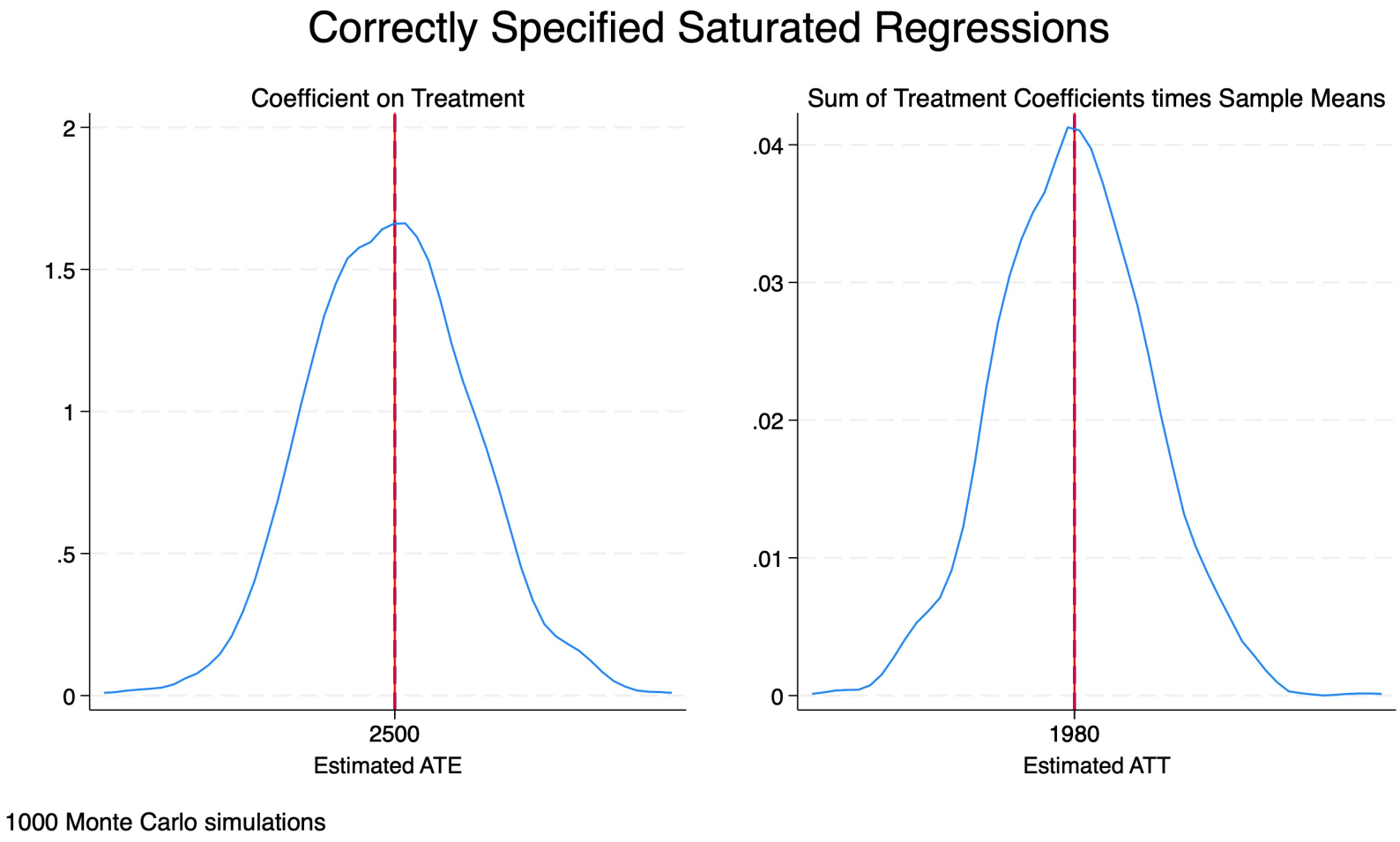
Misspecified Saturated OLS Regression



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[1.treat]
    scalar ate2 = `ate2'
    gen ate2=`ate2'

    * Obtain the coefficients
    local treat_coef        = _b[1.treat] // 0
    local age_treat_coef    = _b[1.treat#c.age] // 1
    local agesq_treat_coef  = _b[1.treat#c.age_sq] // 2
    local gpa_treat_coef    = _b[1.treat#c.gpa] // 3
    local gpasq_treat_coef  = _b[1.treat#c.gpa_sq] // 4
    local age_gpa_coef      = _b[1.treat#c.age#c.gpa] // 5


    * Save the coefficients as scalars and generate variables
    scalar  treat_coef = `treat_coef'
    gen     treat_coef_var = `treat_coef' // 0

    scalar  age_treat_coef = `age_treat_coef'
    gen     age_treat_coef_var = `age_treat_coef' // 1

    scalar  agesq_treat_coef = `agesq_treat_coef'
    gen     agesq_treat_coeff_var = `agesq_treat_coef' // 2

    scalar  gpa_treat_coef = `gpa_treat_coef'
    gen     gpa_treat_coef_var = `gpa_treat_coef' // 3

    scalar  gpasq_treat_coef = `gpasq_treat_coef'
    gen     gpasq_treat_coef_var = `gpasq_treat_coef' // 4

    scalar  age_gpa_coef = `age_gpa_coef'
    gen     age_gpa_coef_var = `age_gpa_coef' // 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 gpa_sq 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_coef_var + /// 0
                age_treat_coef_var * mean_age + /// 1
                agesq_treat_coeff_var * mean_agesq + /// 2
                gpa_treat_coef_var * mean_gpa + /// 3
                gpasq_treat_coef_var * mean_gpasq + /// 4
                age_gpa_coef_var * mean_agegpa
```

Correctly Saturated OLS Regression



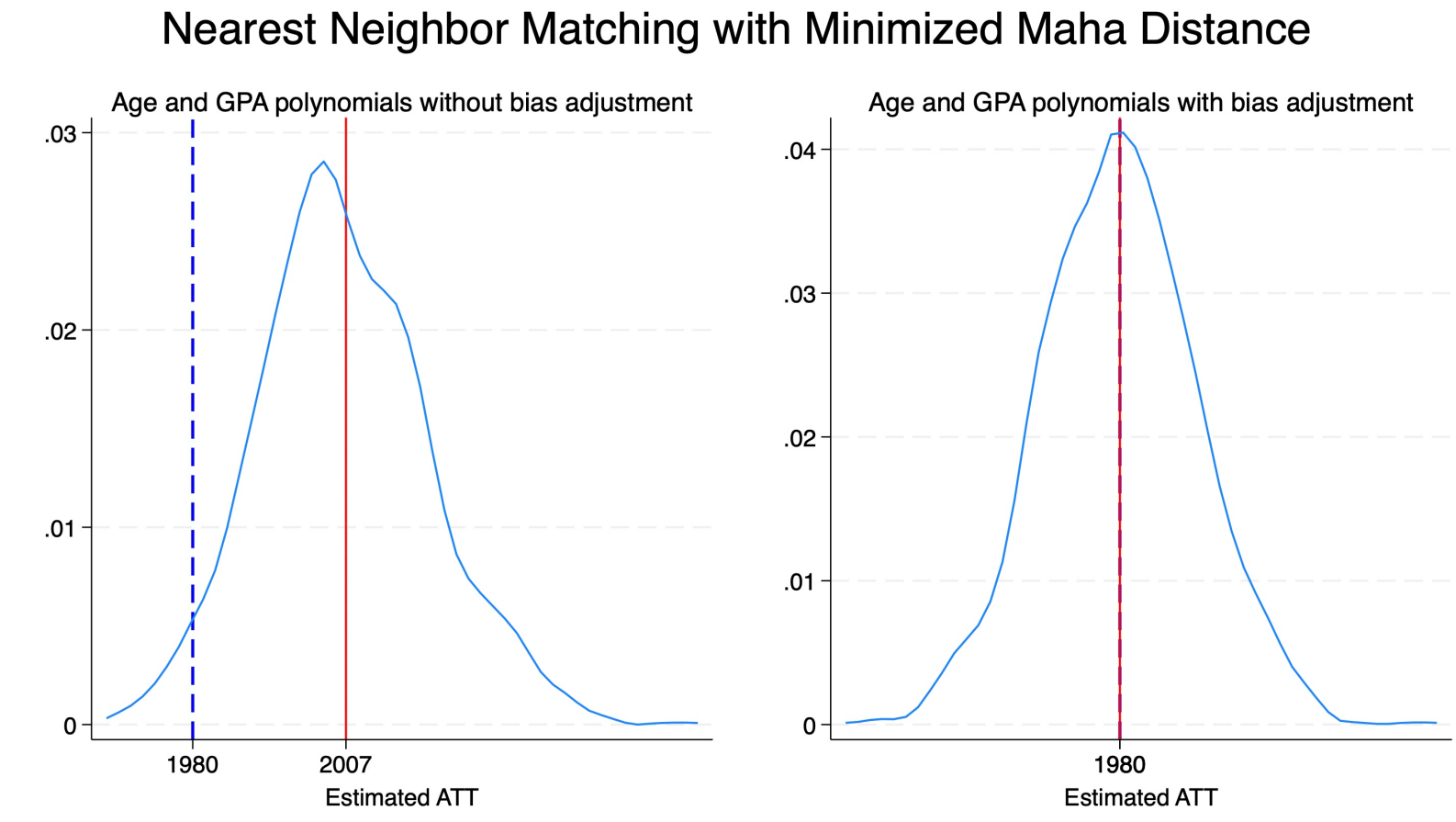
Regression adjustment

- 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`

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



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 dummied 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

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'
```

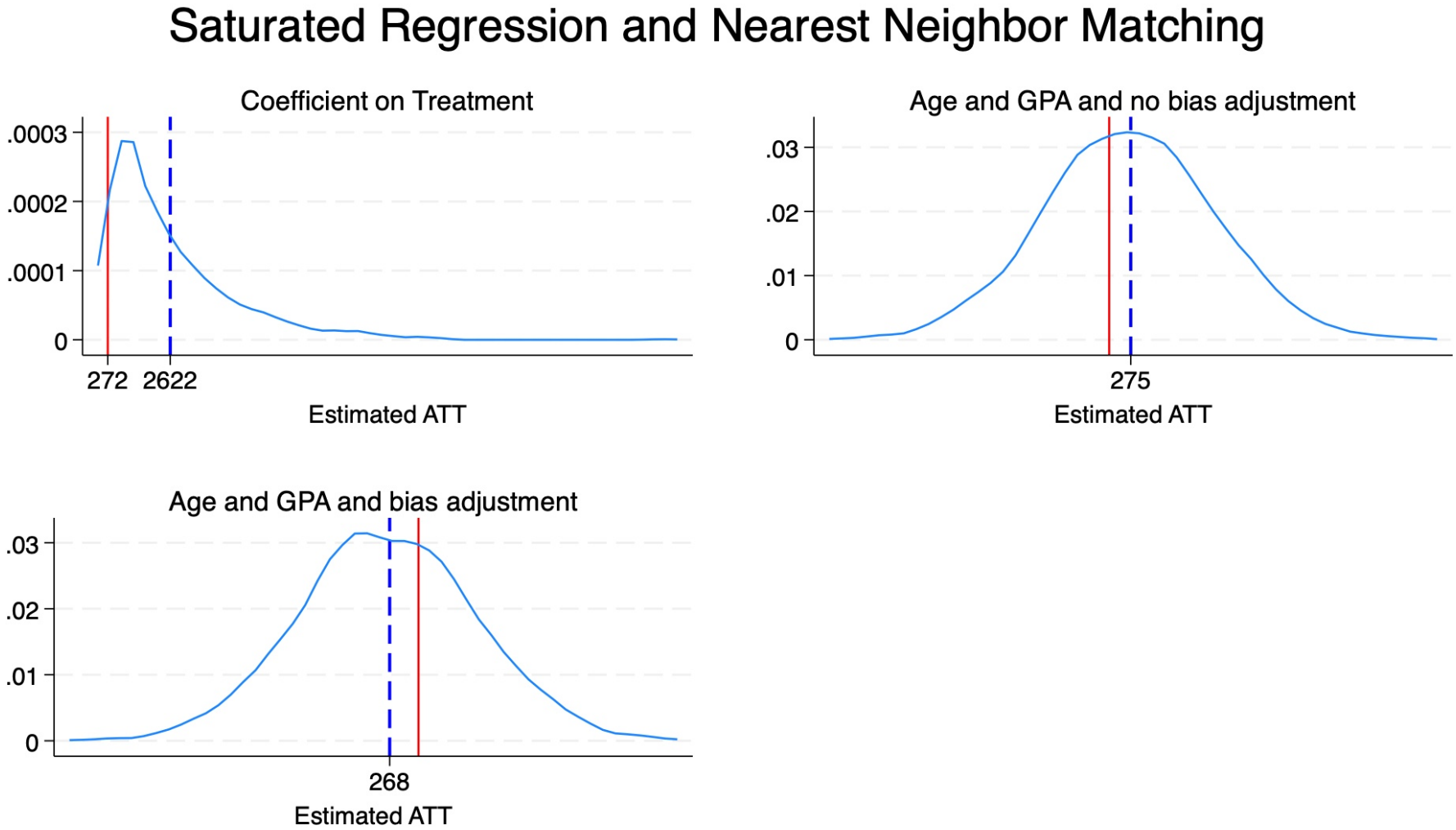
Nonlinear DGP, OLS with RA

```
    * 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



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

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*[a,b]

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

[a] 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.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

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

[d] 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*[a,b]

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

[a] 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.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

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

[d] See Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

	All		CPS Controls $N_c = 15,992$	NSW Trainees $N_t = 297$		
covariate	mean	(s.d.)	mean	mean	t-stat	diff
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

Dehija 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. $)*
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	.80	.09	.73	.17		3,066
		(.32)	(.09)	(.02)	(.01)	(.02)	(.02)		(236)
Control	425	24.45	10.19	.80	.11	.81	.16		3,026
		(.32)	(.08)	(.02)	(.02)	(.02)	(.02)		(252)
RE74 subset:[b]									
Treated	185	25.81	10.35	.84	.059	.71	.19	2,096	1,532
		(.35)	(.10)	(.02)	(.01)	(.02)	(.02)	(237)	(156)
Control	260	25.05	10.09	.83	.1	.83	.15	2,107	1,267
		(.34)	(.08)	(.02)	(.02)	(.02)	(.02)	(276)	(151)
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	.25	.032	.31	.87	19,429	19,063
		[.78]	[.23]	[.03]	[.01]	[.04]	[.03]	[991]	[1,002]
PSID-2	253	36.10	10.77	.39	.067	.49	.74	11,027	7,569
		[1.00]	[.27]	[.04]	[.02]	[.05]	[.04]	[853]	[695]
PSID-3	128	38.25	10.30	.45	.18	.51	.70	5,566	2,611
		[1.17]	[.29]	[.05]	[.03]	[.05]	[.05]	(686)	[499]
CPS-1	15,992	33.22	12.02	.07	.07	.29	.71	14,016	13,650
		[.81]	[.21]	[.02]	[.02]	[.03]	[.03]	[705]	[682]
CPS-2	2,369	28.25	11.24	.11	.08	.45	.46	8,728	7,397
		[.87]	[.19]	[.02]	[.02]	[.04]	[.04]	[667]	[600]
CPS-3	429	28.03	10.23	.21	.14	.60	.51	5,619	2,467
		[.87]	[.23]	[.03]	[.03]	[.04]	[.04]	[552]	[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; RE*x* = earnings in calendar year 19*x*.

[a] NSW sample as constructed by Lalonde (1986).

[b] The subset of the Lalonde sample for which RE74 is available.

[c] 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.

PSID1–3 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[a]

	A. Lalonde's original sample					B. RE74 subsample (results do not use RE74)					C. RE74 subsample (results use RE74)				
	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all
Comparison group	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	variables[f] (5)	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	variables[f] (5)	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	variables[f] (5)
NSW	886	798	879	802	820	1,794	1,672	1,750	1,631	1,612	1,794	1,688	1,750	1,672	1,655
	(472)	(472)	(467)	(468)	(468)	(633)	(637)	(632)	(637)	(639)	(633)	(636)	(632)	(638)	(640)
PSID-1	−15,578	−8,067	−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)	(881)	(901)	(1155)	(931)	(841)	(866)	(886)
PSID-2	−4,020	−3,482	−1,364	−1,694	−1,876	−3,647	−2,810	721	298	111	−3,647	94	721	907	683
	(781)	(935)	(729)	(878)	(885)	(960)	(1082)	(886)	(1004)	(1032)	(960)	(1042)	(886)	(1004)	(1028)
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)	(1104)
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)	(557)	(560)	(712)	(572)	(537)	(547)	(550)
CPS-2	−4,195	−2,341	−1,649	−1,129	−1,149	−3,822	−2,208	−263	371	305	−3,822	615	−263	879	790
	(533)	(620)	(459)	(551)	(551)	(671)	(746)	(574)	(662)	(666)	(671)	(672)	(574)	(654)	(658)
CPS-3	−1,008	−1	−1,204	−263	−234	−635	375	−91	844	875	−635	1,270	−91	1,326	1,326
	(539)	(681)	(532)	(677)	(675)	(657)	(821)	(641)	(808)	(810)	(657)	(798)	(641)	(796)	(798)

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID1–3, 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.

[a] Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

[b] The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample ($886) and for the RE74 sample ($1,794).

[c] 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).

[d] Regresses RE78 on a treatment indicator and RE75.

[e] The same as (d), but controls for the additional variables listed under (c).

[f] 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)

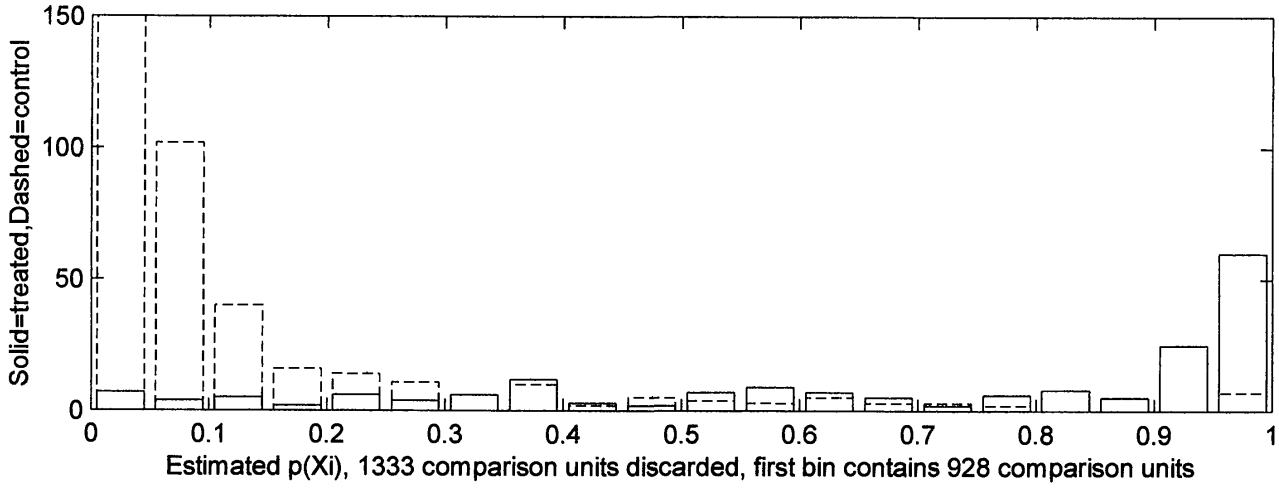


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal 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.

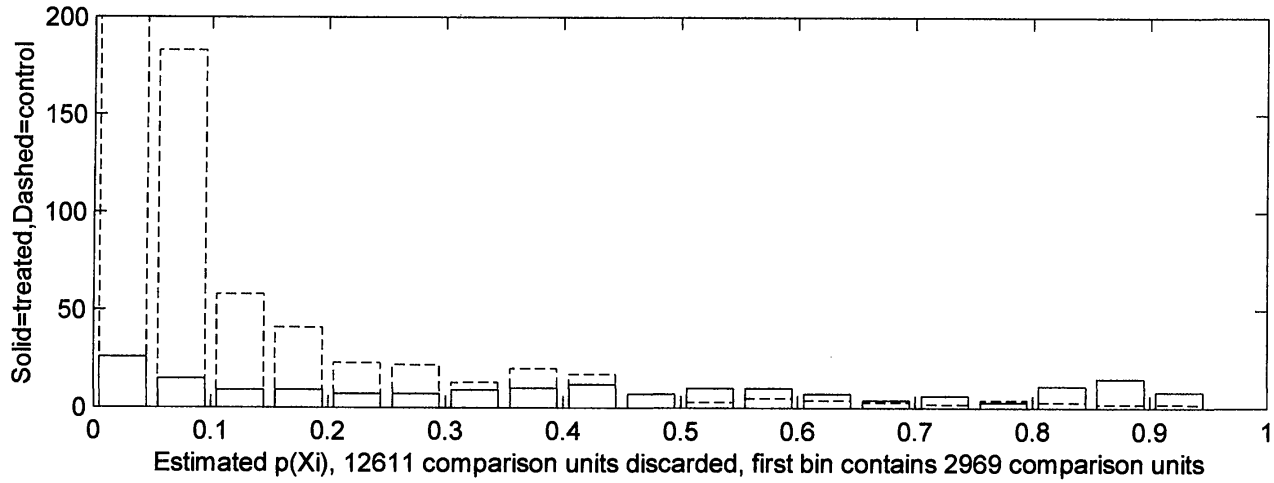


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal 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[b]	*Stratifying on the score*			*Matching on the score*	
	(1) Unadjusted	*(2) Adjusted*[a]	*(3)*	*(4) Unadjusted*	*(5) Adjusted*	*(6) Observations*[c]	*(7) Unadjusted*	*(8) Adjusted*[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] 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).

[c] 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.

[d] 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:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	*No. of observations*	*Age*	*Education*	*Black*	*Hispanic*	*No degree*	*Married*	*RE74 (U.S. $)*	*RE75 (U.S. $)*
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 githbub for this workshop

Coarsened exact matching

- There are two kinds of matching as we've said
  1. *Exact matching* matches a treated unit to all of the control units with the same covariate value. Sometimes this is impossible (e.g., continuous covariate).
  2. *Approximate matching* specifies a metric to find control units that are close to the treated unit. Requires a distance metric, such as Euclidean, Mahalanobis, or the propensity score. All of which can be implemented in Stata's `teffects`.
- Iacus, King and Porro (2011) propose another version of matching they call coarsened exact matching (CEM). Some big picture ideas

Checking imbalance

- Iacus, King and Porro (2008) say that in practice approximate matching requires setting the matching solution beforehand, then checking for imbalance after.
- Start over, repeat, until the user is exhausted by checking for imbalance.

CEM Algorithm

1. Begin with covariates X. Make a copy called $X*$
2. Coarsen $X*$ according to user-defined cutpoints or CEM's automatic binning algorithm
   - $\rightarrow$ Schooling $\rightarrow$ less than high school, high school, some college, college, post college
3. Create one stratum per unique observation of $X*$ and place each observation in a stratum
4. Assign these strata to the original data, X, and drop any observation whose stratum doesn't contain at least one treated and control unit

You then add weights for stratum size and analyze without matching.

Tradeoffs

- Larger bins mean more coarsening. This results in fewer strata.
- Fewer strata result in more diverse observations within the same strata and thus higher imbalance
- CEM prunes both treatment and control group units, which changes the parameter of interest. Be transparent about this as you're not estimating the ATE or the ATT when you start pruning

Benefits

- The key benefit of CEM is that it is in a class of matching methods called *monotonic imbalance bounding*
- MIB methods bound the maximum imbalance in some feature of the empirical distributions by an ex ante decision by the user
- In CEM, this ex ante choice is the coarsening decision
- By choosing the coarsening beforehand, users can control the amount of imbalance in the matching solution
- It’s also wicked fast.

Imbalance

- There are several ways of measuring imbalance, but here we focus on the $\mathcal{L}_1(f, g)$ measure which is

$$\mathcal{L}_1(f, g) = \frac{1}{2} \sum_{l_1 \ldots l_k} |f_{l_1 \ldots l_k} - g_{l_1 \ldots l_k}|$$

  where the f and g record the relative frequencies for the treatment and control group units.
- Perfect global imbalance is indicated by $\mathcal{L}_1 = 0$. Larger values indicate larger imbalance between the groups, with a maximum of $\mathcal{L}_1 = 1$.

Stata

- Download `cem` from Stata: `ssc install cem, replace`
- You will automatically compute the global imbalance measure, as well as several unidimensional measures of imbalance, when using `cem`
- I got a $\mathcal{L}_1 = 0.55$. What does it mean?
  - $\rightarrow$ By itself, it's meaningless. It's a reference point between matching solutions.
  - $\rightarrow$ Once we have a matching solution, we will compare its $\mathcal{L}_1$ to 0.55 and gauge the increase in balance due to the matching solution from that difference.
  - $\rightarrow$ Thus $\mathcal{L}_1$ works for imbalance as R^2 works for model fit: the absolute values mean less than comparisons between matching solutions.

More Stata

- Because `cem` bounds the imbalance ex ante, the most important information in the Stata output is the number of observations matched.
- You can also choose the coarsening as opposed to relying on the algorithm's automated binning.
- Once you have estimated the strata, you regress the outcome onto the treatment and then weight the regression by `cem_weights`. For instance,

```
regress re78 treat [iweight=cem_weight]
```

- For more on this, see Blackwell, et al. Stata journal article from 2009.

Roadmap

- Unconfoundedness and Ignorable Treatment Assignment
  - Choosing Covariates
  - Aggregate target parameters
- Matching Estimators
  - Stratification weighting
  - Conditional Independence
  - Exact and Inexact Matching
  - Propensity scores
  - Regressions
  - Coarsened exact matching
- 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

Extrapolation and functional forms

- If you do not have weak common support, then you can only estimate ATT using *extrapolation* which is model dependent on functional forms
- If you have heterogenous treatment effects, then the standard models with additive covariates is misspecified even if those are the right variables – you'll need regression adjustment for "saturation"
- Same model contains ATT and ATE, so you need to decide which one you want
- Functional form extrapolation is powerful, as it can make inference far beyond the support of the data, but it's a bit of some dark magic too

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 heterogenous treatment effects but requires common support; OLS uses extrapolation in its place
- You can address heterogenous treatment effects with fully interaction called regression adjustment (itself extremely uncommon in practice), but that does not address nonlinear DGP

Quote from Tymon Sloczynski (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 Sloczynski 2022)