Causal Inference I

Mixtape Session

Roadmap

- Directed Acyclic Graphs
  - Backdoor criterion
  - Collider bias
- Unconfoundedness and Ignorable Treatment Assignment
  - Motivating estimation with an example
  - Aggregate target parameters
  - Assumptions
- Estimators
  - Subclassification
  - Regression
  - DML
- Concluding remarks

Double De-biased Machine Larning

- This is an important paper. It's by Chernozhukov, et al. (2018) and has been very influential
- But the elements of it are found in ideas that I think we have to cover because otherwise the technical mechanics of it will overshadow important parts that I think you can't skip
- DML, as I'll call it, ultimately is a machine learning based method for handling a large number of variables (features) to reconstruct counterfactuals
- But these variables still cannot just be haphazardly selected and we need to see why – there are no magic wands in causal inference sadly

Control for X

- The easiest way to motivate the core identifying assumptions for DML is to start with graphs
- Then move into regression and explain the concepts of curse of dimensionality and why contemporary data sources swamp our matching estimators
- We rely therefore on extrapolation using penalized regressions and other ML methods, but these are still just ”control for X”

Adjusting for variables

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

Judea Pearl and DAGs

- As you maybe know, Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology
- We will only focus on one idea called the backdoor criterion and link that to unconfoundedness, a potential outcomes framework that underlies the DML estimator
- I will be emphasizing things that others don't emphasize when introducing DML, but that is just my style

Further reading

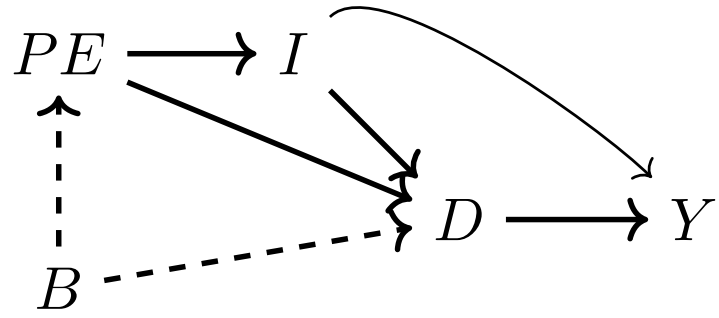
1. Molak (2023), Causal Inference and Discovery in Python (good)
2. Facure (2023),
   Causal Inference in Python: Applying Causal Inference in the Tech Indu
   (great)
3. Facure (2022) Causal Inference for the Brave and the Bold (also fantastic)
4. Brigham Frandsen's online workshops which I'll share for free

Design vs. Model

- DAGs are way too tempting in my opinion, and I think they may take us backwards a little, but they are core parts of modern causal inference mapping onto ML, and in particular tech
- They are extremely helpful for thinking through the differences between good and bad controls, which DML will not be able to differentiate

Causal model

- The causal model is sometimes called the structural model, but for us, I prefer the former as it's less alienating
- It's the system of equations describing the relevant aspects of the world
- It necessarily is filled with causal effects associated with some particular comparative statics
- Consider the following diagram representing the returns to education with simplified confounders



- B is a **parent** of PE and D
- PE and D are **descendants** of B
- There is a **direct (causal) path** from D to Y
- There is a **mediated (causal) path** from B to Y through D
- There are six **paths** from PE to Y but none are direct, but some of them are different in other ways

Where do DAGs come from?

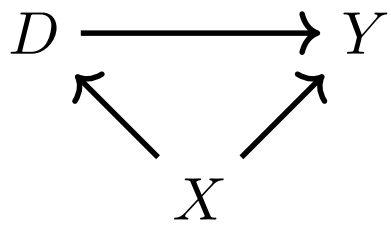
- DAGs are meant to represent "contemporary agreement among experts" – if you aren't willing to present your DAG before a room of experts, it's likely you shouldn't use it at all
- But more than that, the DAG is a local graph – it's the modeling of the treatment assignment mechanism in your local context, not merely some model of monopolies or whateer
- Your DAG should be a reasonable approximation of D and Y parents (confounders) and direct and indirect effects of D on Y
- We get ideas for DAGs from theory, models, observation, experience, prior studies, intuition, as well as conversations with domain experts

Unconfoundedness and the backdoor criterion

- DAGs help us understand the source of problems in our observational (non-experimental) data that make inferring causality hard
- But it also can help us see a way out in some situations
- We will focus today on the unconfoundedness research design, which is best described in causal graphs with the concept of the **backdoor criterion**
- As we will see, the DAG helps you solve the problem of choosing covariates for a model to resolve selection bias, but to do so requires confidence in your DAG

Confounding

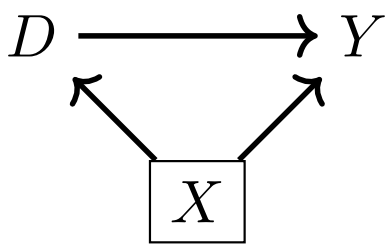
- Confounding occurs when when the treatment and the outcomes have a common parent node as that creates spurious correlation between D and Y



- Confounders are causing selection bias $E[Y^0|D=1,X] \neq E[Y^0|D=0,X]$ due to differences in X which are shaping Y^0 differentially by treatment status

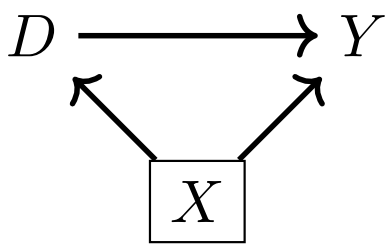
Backdoor Paths

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



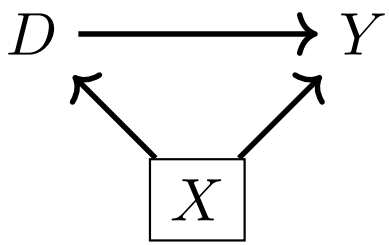
Backdoor Paths

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



Backdoor Paths

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



Backdoor criterion

Backdoor criterion

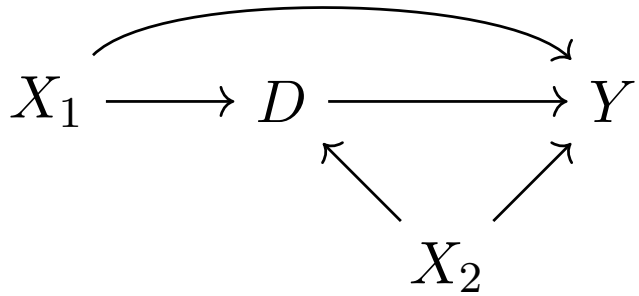
Conditioning on X satisfies the backdoor criterion with respect to (D, Y) directed path if:

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

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

What control strategy meets the backdoor criterion?

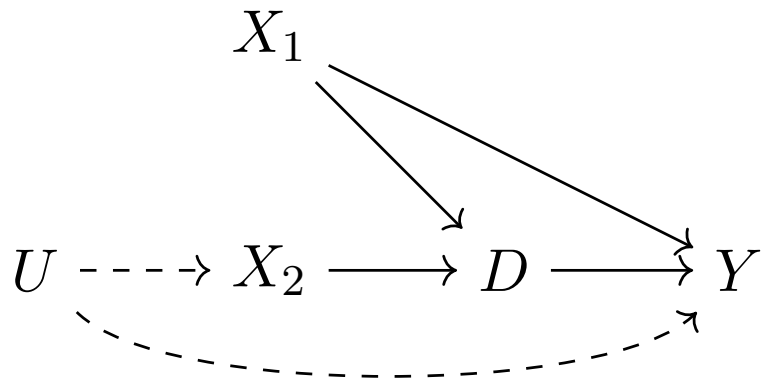
- List all backdoor paths from D to Y. I'll wait.



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

What if you have an unobservable?

- List all the backdoor paths from D to Y.



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

Collider bias

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

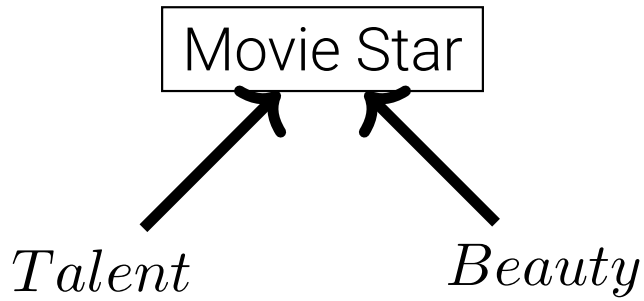
Example 1: Movie stars

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

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

Movie star DAG

Imagine casting directors pick movie stars based on talent and beauty



Talent and beauty can become correlated even though they are independent

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

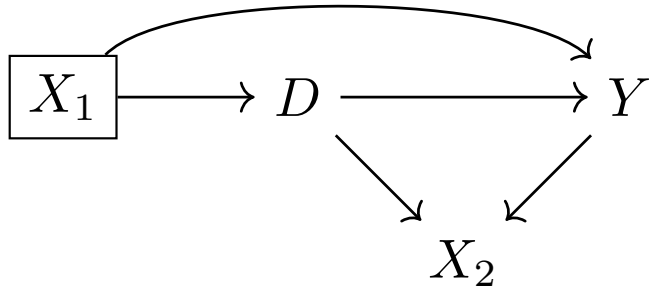
Sample selection?

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

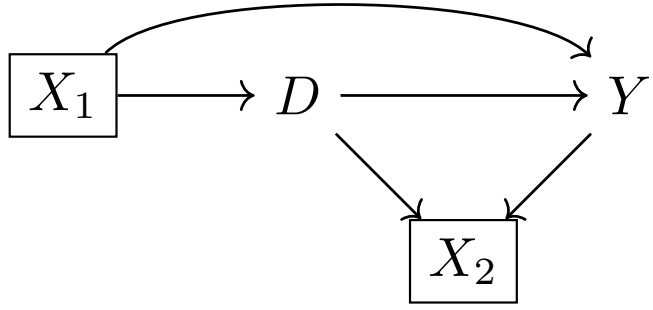
$$beauty_i = \alpha + \delta talent_i + \beta star_i + \varepsilon_i$$

- It’s not just sample selection

- **Colliders can be outcomes (and often those are the ones)**
  - → There is only one backdoor path from D to Y

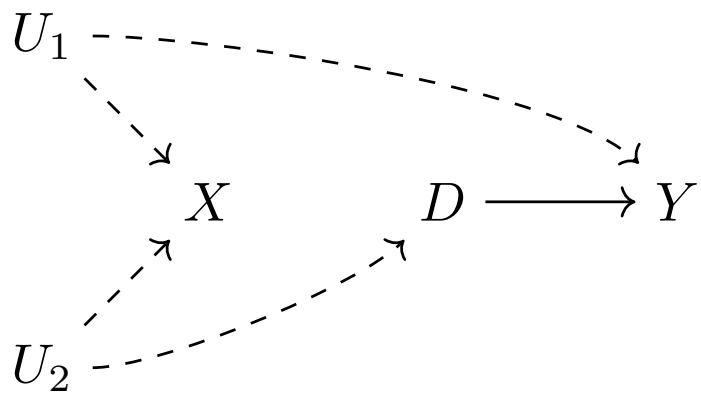


  - → Conditioning on X_1 blocks the backdoor path
  - → But what if we also condition on X_2?

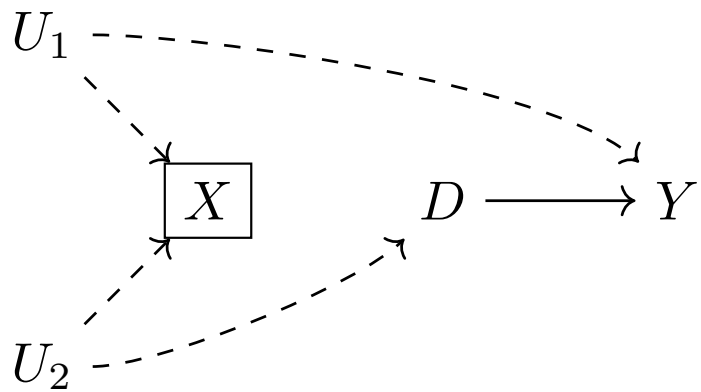


  - → Conditioning on X_2 opens up a new path, creating new spurious correlations between D and Y

- **Colliders could be pre-treatment covariates (called M-bias because it looks like an M)**
  - → Name the backdoor paths. Is it open or closed?

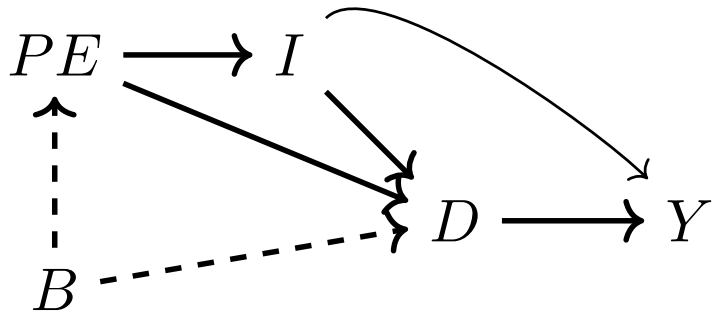


  - → But what if we condition on X?



Testing the Validity of the DAG

- The DAG makes testable predictions, and there is also new work on "causal discovery" which is outside my area of expertise (see Molak and the DoWhy package in python)
- Conditional on D and I, parental education (PE) should no longer be correlated with Y
- Can be hard to figure this out by hand, but software can help (e.g., Daggity.net is browser based, Causal Fusion is more advanced)
- Causal algorithms tend to be DAG based and are becoming popular in industry



Covariate selection without DAGs

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


BEWARE
OF DOG
YAPPP SUPER BANAY!

Falsifications as a test

- 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 or average age of the state
- Imbens and Rubin (2015) suggested using the lagged outcome (pre-treatment) as a way of checking, as those have similar confounder structures
- But just be careful because without some understanding of the treatment assignment mechanism, you could make serious mistakes

Roadmap

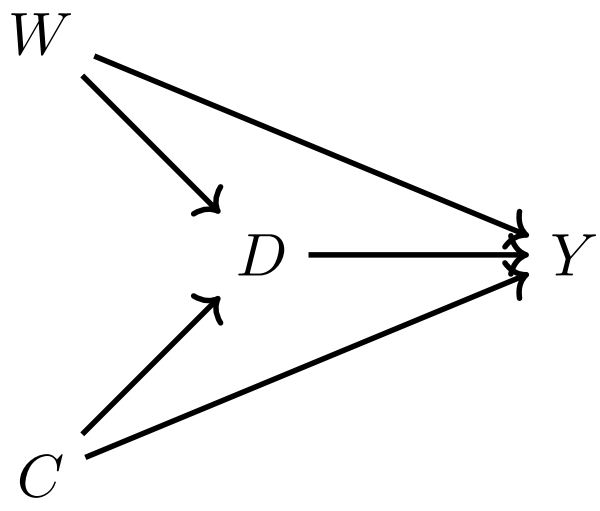
- Directed Acyclic Graphs
  - Backdoor criterion
  - Collider bias
- Unconfoundedness and Ignorable Treatment Assignment
  - Motivating estimation with an example
  - Aggregate target parameters
  - Assumptions
- Estimators
  - Subclassification
  - Regression
  - DML
- Concluding remarks

Titanic example and a simple DAG

- What if we wanted to know the causal effect of being seated in first class (D) on survival (Y) in the sinking of the Titanic cruiser in the early 20th century?
- Domain knowledge: as the Titanic sank, the captain called for women (W) and children (D) to go first

Titanic DAG

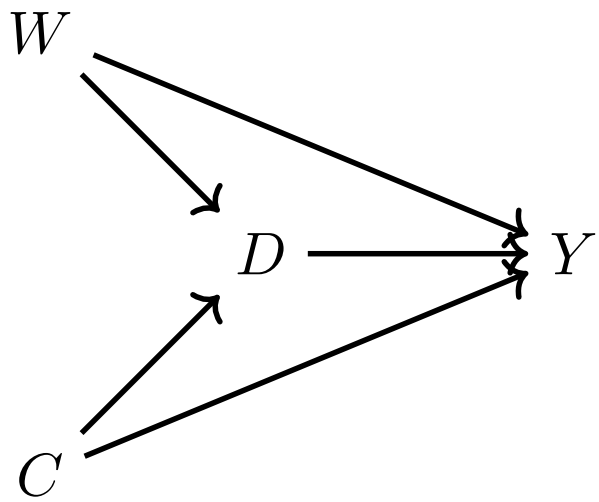
Figure: Titanic sinking as a DAG



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

Simple DAG

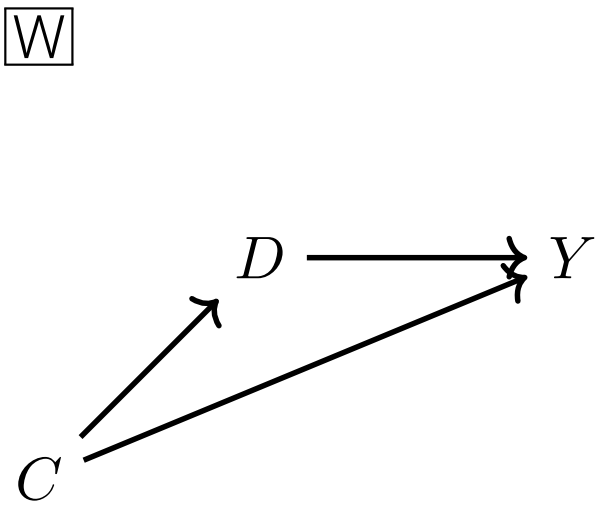
Figure: A simple DAG illustrating selection on observables.



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

Simple DAG

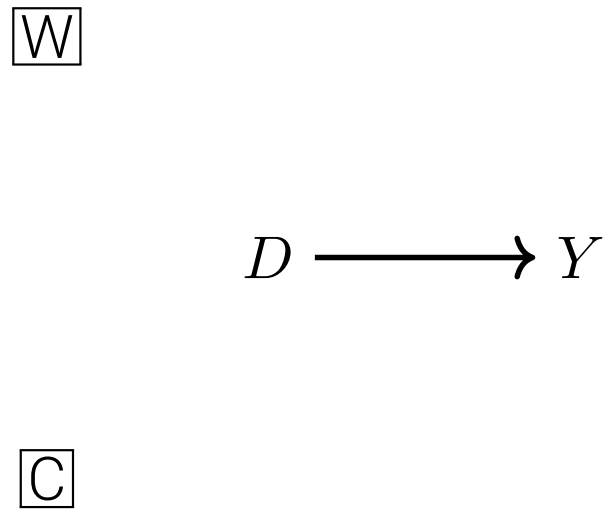
Figure: The same simple DAG illustrating selection on observables only with the direct edge from D to Y deleted and backdoor W blocked.



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

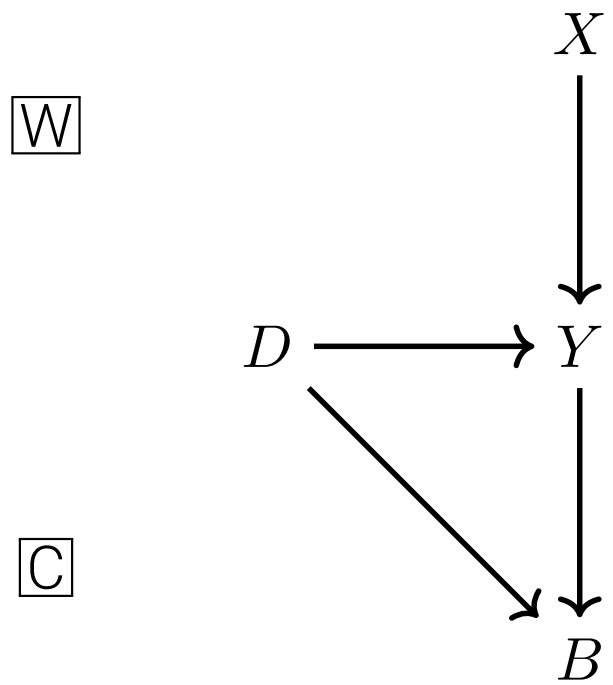
- Confounders may or may not be observed, but they must be known if they are confounders as confounders create backdoor paths from D to Y
- Visually, solid lines means they are “quantified” (i.e., in the data), whereas dashed lines mean they are either not defined correctly or not in the dataset (“unobserved”)
- 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
- Let us now call C and W "known and quantified confounders" because the model said these were necessary, they were observed (no dashed line) and they were confounders
- Let's add two more variables – one we discussed already, but one we haven't

Modification of the original DAG

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



4. Conditional on C and W, the collider path $D \rightarrow B \leftarrow Y$ is closed by the collider B
5. And X never appears on any of our backdoor paths, so it is irrelevant,

Covariate

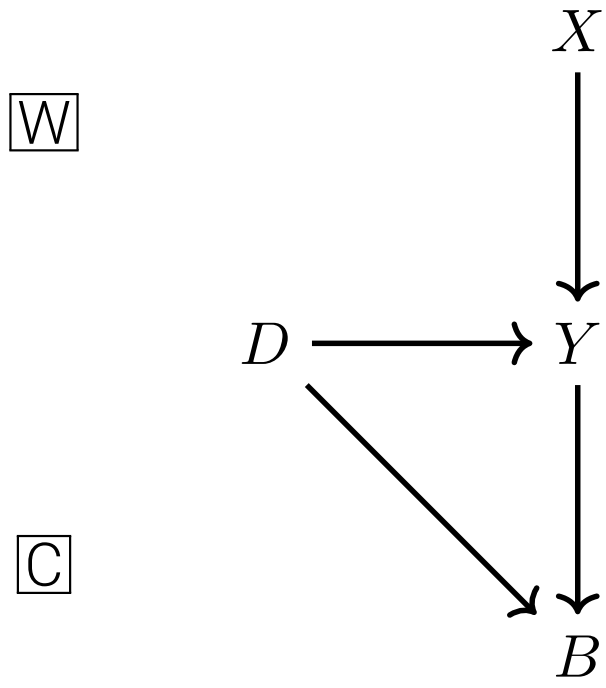
Definition of a Covariate

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

- Think of it as explaining the outcome, but not correlated with the treatment variable (therefore it's in the error term of a regression model)
- 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
- For us today, I am going to try to distinguish between "confounders" which are needed to estimate causal effects and "covariates" which help with precision

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”)

Two useful readings

1. Cunningham (2023), “Which variables do I need to control for?” Substack post, https://causalinf.substack.com/p/which-variables-do-i-need-to-control
2. Cinelli, Forney, and Pearl (2022), “A Crash Course in Good and Bad Controls”, forthcoming in *Journal Sociological Methods and Research* (previously technical report R-493), available in our readings directory

Which variables do I need to control for?

Five types of variables

SCOTT CUNNINGHAM
FEB 25, 2023

42 22 2 Share

Which Variables Do I Need?

I started this substack off wanting to discuss inexact matching, but I just think it would be helpful that before we get into the nuts and bolts of inexact matching, a clear explanation of how to achieve unconfoundedness may be useful. It comes up a lot, and I think sometimes there isn't enough clear exposition about it out there, so I figured why not just do this now, and next week I'll wrap up my inexact matching substack.

Most of us learned causal inference for the very first time in a stats class where we learned about "running regressions" and "controlling for covariates". If the class was somewhat advanced, that introduction to the ordinary least squares formulas might even be followed by learning the Frisch-Waugh-Lovell theorem where we learned OLS was "partialing out" those extra variables effects so that we could focus just on the partial relationship between the covariate of interest and the outcome. And then if we went even further, we might then learn about the theoretical properties of OLS whereby if all the confounders were included in the model and measured well, and the data was reasonably smooth and treatment effects homogenous and additive, then under certain OLS specifications, we could obtain estimates of the ATE. Without those covariates in the model, we learned that the model suffered from "omitted variable bias". What did that mean? It meant that had we not controlled for the confounders, then our OLS estimates wouldn't be causal. Those are fun moments in everybody's life, I think — realizing that even outside the experiment, I might actually get an estimate of a causal effect by "running a regression".

A Crash Course in Good and Bad Controls

Carlos Cinelli* Andrew Forney† Judea Pearl ‡

March 21, 2022

Abstract

Many students of statistics and econometrics express frustration with the way a problem known as "bad control" is treated in the traditional literature. The issue arises when the addition of a variable to a regression equation produces an unintended discrepancy between the regression coefficient and the effect that the coefficient is intended to represent. Avoiding such discrepancies presents a challenge to all analysts in the data intensive sciences. This note describes graphical tools for understanding, visualizing, and resolving the problem through a series of illustrative examples. By making this "crash course" accessible to instructors and practitioners, we hope to avail these tools to a broader community of scientists concerned with the causal interpretation of regression models.

Introduction

Students, data analysts, and empirical social scientists have likely encountered the problem of "bad controls" (Angrist and Pischke, 2009, 2014). The problem arises when an analyst needs to decide whether or not the addition of a variable to a regression equation helps getting estimates closer to the parameter of interest. Analysts have long known that some variables, when added to the regression equation, can produce unintended discrepancies between the regression coefficient and the effect that the coefficient is expected to represent. Such variables have become known as "bad controls," to be distinguished from "good controls" (also known as "confounders" or "deconfounders") which are variables that must be added to the regression equation to eliminate what came to be known as "omitted variable bias" (Angrist and Pischke, 2009; Steiner and Kim, 2016; Cinelli and Hazlett, 2020a,b).

*Department of Statistics, University of Washington, Seattle. Email: cinelli@uw.edu

†Department of Computer Science, Loyola Marymount University, Los Angeles. Email: Andrew.Forney@lmu.edu

‡Department of Computer Science, University of California, Los Angeles. Email: judea@cs.ucla.edu.
This research was supported in parts by grants from the National Science Foundation [#IIS-2106908], Office of Naval Research [#N00014-17-S-12091 and #N00014-21-1-2351], and Toyota Research Institute of North America [#PO000897].

Defining the target parameter comes first

- So you want to estimate causal effects using “controls” – remember our steps, though – step 1 is to define the parameter (e.g., ATE, ATT, ATU)
- Also something called the $CATE$ – conditional ATE which is an ATE for certain values of X (e.g., ATE for men) – which is quite common in ML because of the focus on estimating more targeted ”individual treatment effects” defined by observables

Defining the target parameter comes first

- Covariate adjustment strategies like regression or matching can identify these, but they aren't the same in non-experimental data if there's heterogenous treatment effects (i.e., δ_i differs for different i people)
- So it is imperative up front you decide which aggregate causal parameter you're going to be trying to estimate as each one has different assumptions and different methodologies (as well as interpretations)

Aggregate causal parameters have missing potential outcomes

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

Assumptions

- When attempting to estimate aggregate causal parameters using covariate adjustment, there are two assumptions
  1. **Unconfoundedness**: fairly controversial to some because of what it implies about human behavior
  2. **Common support**: since most people stop at assumption 1, there's much less attention to common support
- Today we will assume you satisfy unconfoundedness, but because we are solely focused on DML, I will be mostly moving around common support to go directly into regression based methods

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 of X, this assumption states they choose treatments by flipping coins (not because they think the treatment helped them)
- Eliminating all backdoor paths on a DAG through blocking satisfies unconfoundedness; also called ignorability

Identifying assumption I: Unconfoundedness

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

Economic meaning of backdoor criterion

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

Why do economists so dislike "conditional independence"?

A humble theory from an economist

SCOTT CUNNINGHAM
JAN 4, 2023

25 11 1 Share

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

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

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 imo (I seem to be alone in describing strong unconfoundedness this way)

But maybe you actually want the ATT, we can go with strictly weaker assumptions – weak unconfoundedness and weak overlap – and therefore slightly more rationality

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

Maybe You Want the ATT

- ATE – it's the summarizing of all treatment effects at the unit level. So think of it as a population parameter and if you are confused, ask yourself "Do I want to know the causal effect for every possible person?"
- ATT – it's summarizing the treatment effects only for the treatment group. Think of it as a subpopulation parameter. Do you want to know the causal effect but only for the treatment group?
- Efficacy of vaccines (ATE) versus efficacy of ventilators (ATT)
- Popularity of DiD and synth suggest that the ATT is a desirable parameter as they only identify the ATT, not the ATE, under parallel trends

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.

You could also flip out independence with respect to Y^1 and discuss common support for the control group and get back to the ATU too

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

- Directed Acyclic Graphs
  - Backdoor criterion
  - Collider bias
- Unconfoundedness and Ignorable Treatment Assignment
  - Motivating estimation with an example
  - Aggregate target parameters
  - Assumptions
- Estimators
  - Subclassification
  - Regression
  - DML
- Concluding remarks

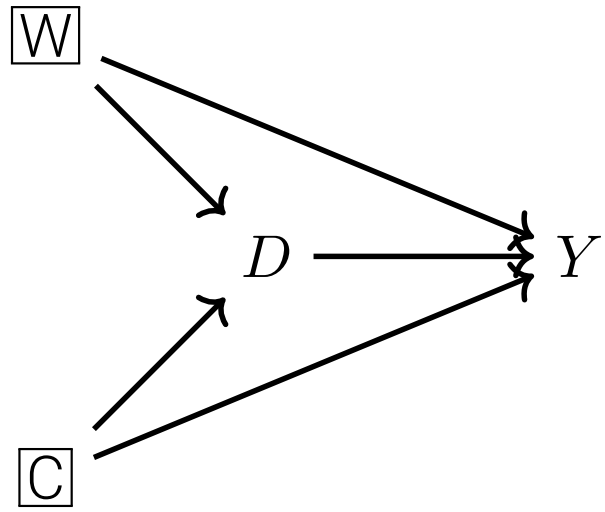
- Multivariate regression goes back to Yule (19th century)
- Weighting goes back to Cochran (late 60s)
- Matching seems to be from the 1970s
- Novel insights about regression methods are based on Oaxaca-Blinder in the 1970s, and more recently Kline and Wooldridge
- DML as we said is Chernozhukov, et al. (2016)

Subclassification method

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

Exercise: Titanic DAG

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



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

Titanic exercise

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

Table 1: Standard balance table

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"

Curse of Dimensionality

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

Empty cells

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

Moving into regression

- If we have unconfoundedness and common support – complete common support – then an unbiased estimate of these aggregate causal parameters is possible using nonparametric matching
- But when we have a large number of covariates, we do not have common support because of the curse of dimensionality, and regression tends to be used instead (and therefore ML too)
- Let's look at a proof by Imbens and Rubin explaining the additional assumptions needed by regression generally

Core OLS assumptions

- Return to OLS model with additive covariates controls:

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

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

OLS Assumptions

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

OLS Assumptions

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

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

for some values of the three unknown parameters, 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)

Great quote

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

Prediction vs causal inference and OLS

- OLS is the “best linear predictor” given the covariates because it minimizes the sum of all the residuals (squared)
- So it’s like it finds a *line* that goes through the data in such a way that the prediction error is on average as small as possible
- But out of sample is another matter – that line *fits the data* and ML has been about trying to improve it for prediction *out of the data*
- But to Tymon’s point, that still isn’t causal inference because in causal inference we are predicting fictional counterfactuals, not realized outcomes

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):

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

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

Constant Treatment Effects and Linearity

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

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

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

Simulation

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

Heterogenous Treatment Effects wrt X

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

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

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

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

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

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

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

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

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

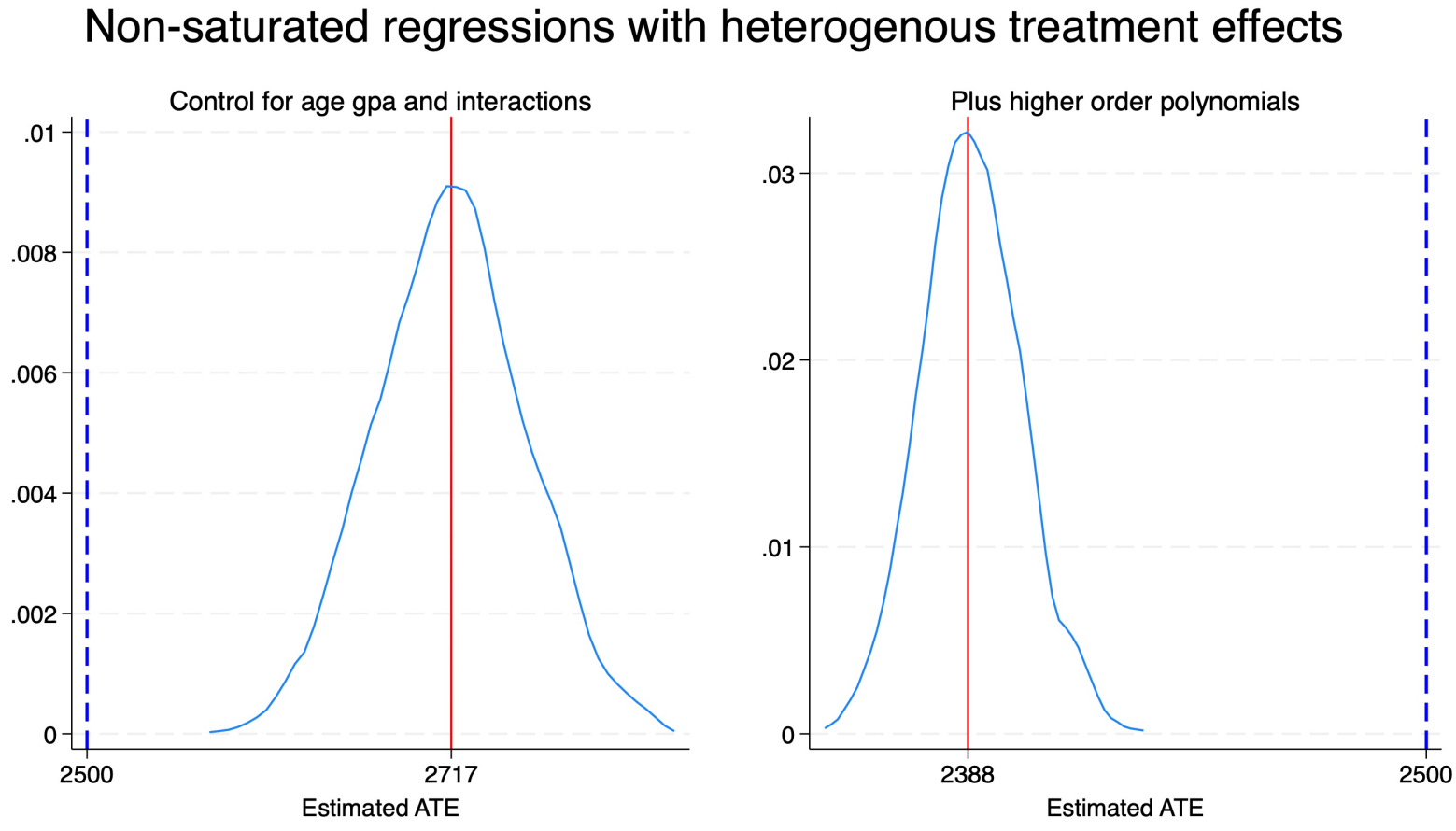
Parameters

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

Constant Treatment Effects and Linearity

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

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

Coefficient on Treatment Dummy is Wrong



ATE is 2500 and ATT is 1980

Commentary

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

Heterogenous treatment effects

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

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

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 effects

Take conditional expectations of the *potential* outcomes:

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

Average treatment effect is:

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

Regression adjustment

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

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

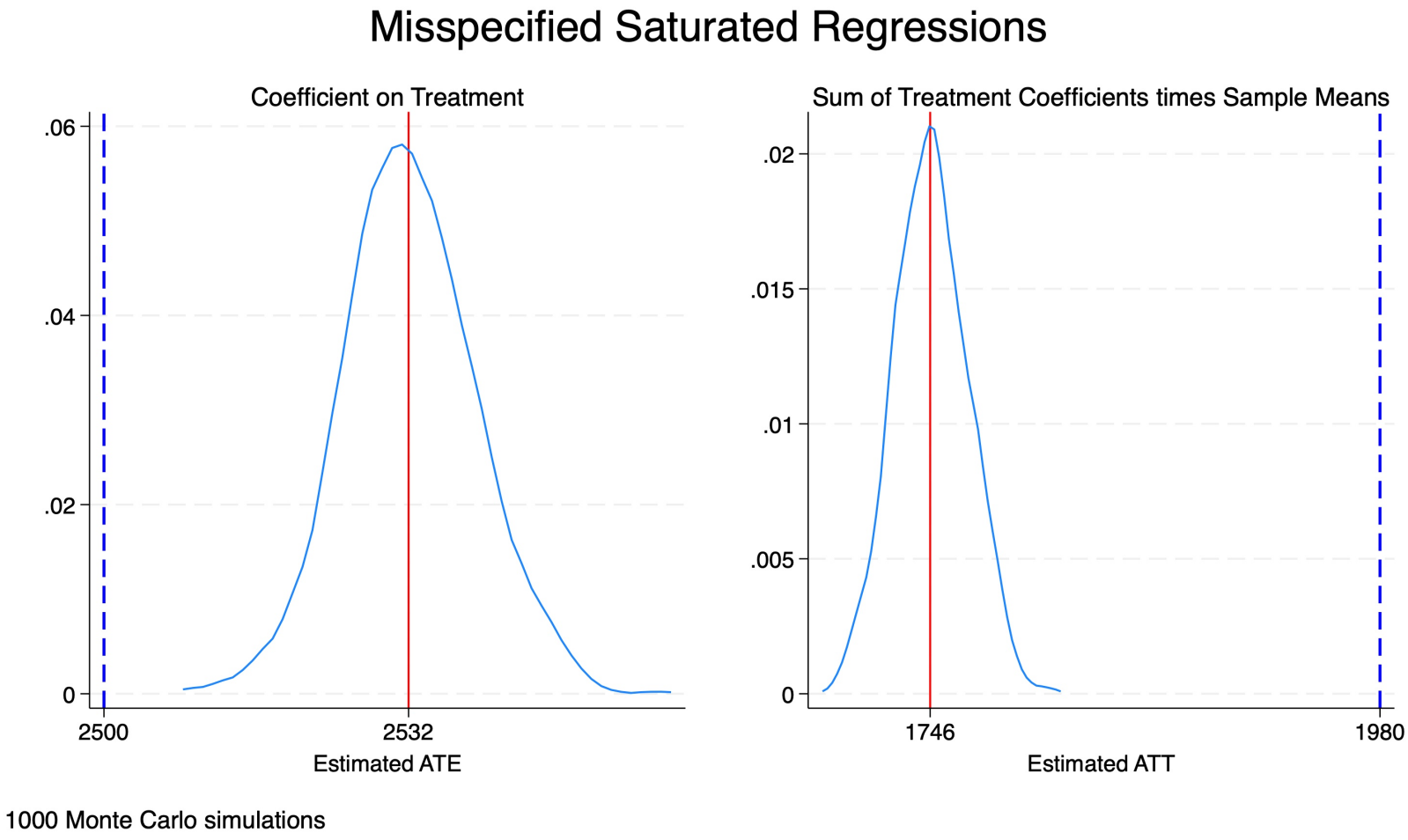
$\widehat{\delta}$ is the ATE but the ATT is equal to $\widehat{\delta} + \widehat{\beta_{11}} E[X_i | D_i = 1]$ where $E[X_i | D_i = 1]$ is the sample average of X_i for the treatment group

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

Misspecified Saturated OLS Regression

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

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

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

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

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

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

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

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

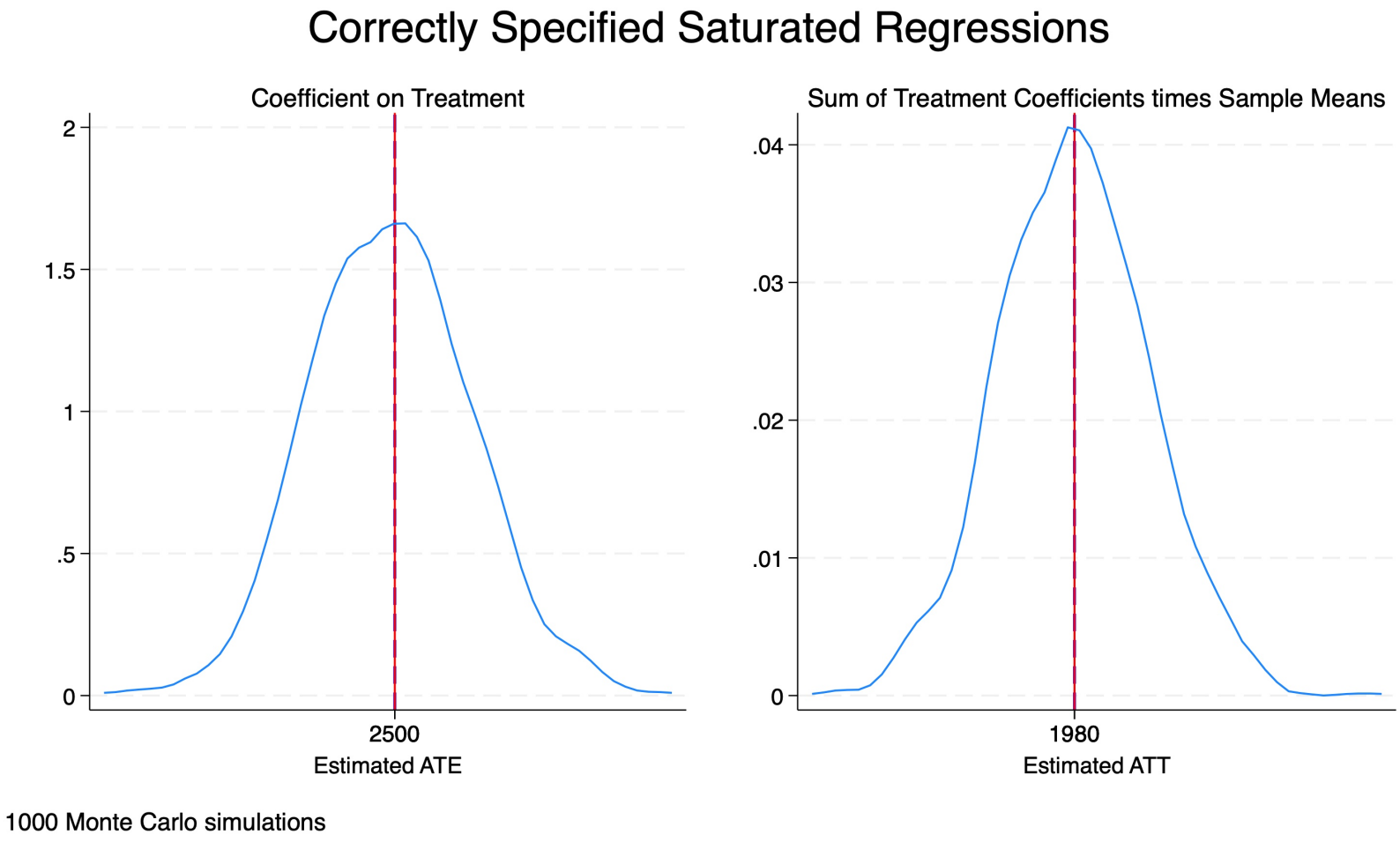
Misspecified Saturated OLS Regression



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`

Curse of dimensionality and DML

So recall our conversation then that we needed both unconfoundedness and common support, but you won't hear much about common support with DML because it extrapolates – but there's some other reasons too

1. Machine Learning Techniques: DML incorporates machine learning methods that are robust to high-dimensional data. Methods like Lasso, Random Forests, or Gradient Boosting are designed to handle large numbers of features, potentially mitigating the curse of dimensionality to some extent.

Curse of dimensionality and DML

2. Double-Residualization: The double-residualization step in DML can help reduce dimensionality by focusing on the component of the outcome and treatment variable that cannot be explained by the high-dimensional controls (we'll see this in a second with the Frisch-Waugh-Lovell theorem)

Curse of dimensionality and DML

3. Sparsity: In many empirical applications, the true underlying model may be sparse, even if it's high-dimensional. Many machine learning techniques incorporated within DML are good at picking up this sparsity. But what is sparsity in this context?

Sparsity in DML

- **What is Sparsity?**: In high-dimensional models, sparsity refers to the property where only a small number of variables (features) have a significant impact on the outcome, even though there could be a large number of variables involved.

Sparsity in DML

- **Why is it Important?**
  - → *Efficiency*: A sparse model reduces computational burden.
  - → *Interpretability*: Easier to focus on key variables, making the model more understandable.
  - → *Generalizability*: Sparse models are less likely to overfit, improving their performance on unseen data.

Sparsity in DML

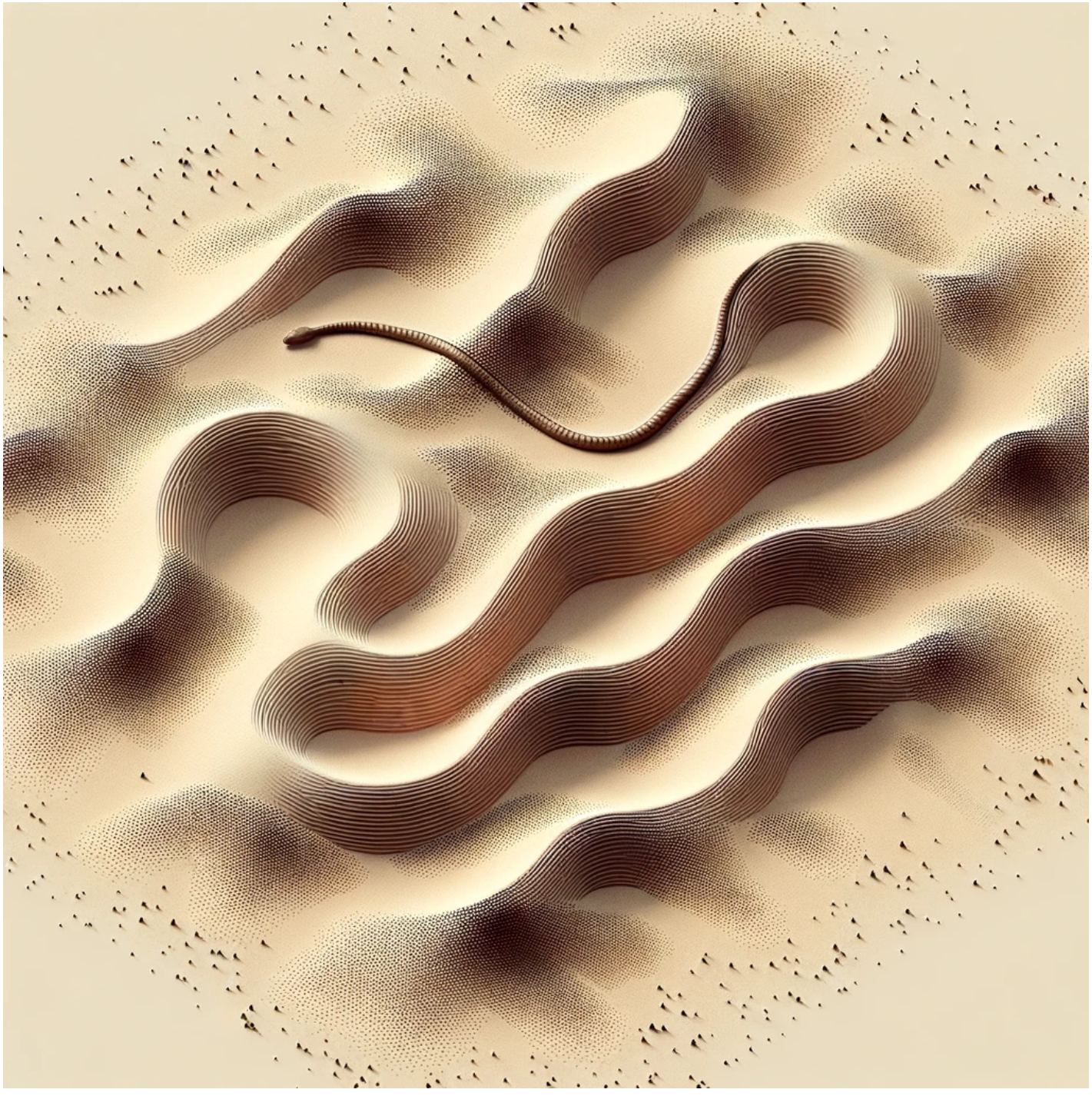
- **DML and Sparsity**: Many machine learning techniques used within DML are adept at identifying these significant, non-zero parameters. Algorithms like LASSO, Random Forests, and certain neural networks can efficiently pick up this sparsity.

Curse of dimensionality and DML

4. Regularization: Regularization methods used in machine learning can help to avoid overfitting, which is closely related to the curse of dimensionality. By penalizing the complexity of the model, regularization can help make the estimation problem more manageable, even in high dimensions.
5. Cross-Fitting: This approach ensures that the model used to predict each observation is fitted on a separate sample, which can help to alleviate overfitting and, consequently, some issues related to high dimensionality.

Overfitting - The ’Overeager Student’

- An overfit model performs exceedingly well on the data it’s trained on, but poorly on new, unseen data.
- Imagine a student who memorizes every question in the textbook but fails to understand the underlying concepts.
- It’s like fitting a curve that passes through every data point, capturing noise as well as the trend.



Underfitting - The ’Lazy Student’

- Imagine a student who only studies the chapter titles, missing all the details.
- An underfit model is too simple to capture the complexities in the data.
- It performs poorly both on the data it was trained on and on new, unseen data.

The ’Goldilocks Zone’ - Best Fitting

- The ideal model is in the ’Goldilocks Zone’: Not too simple, not too complex.
- It performs well on both the training data and new, unseen data.
- Balancing bias (systematic error) and variance (random error) is key to finding this sweet spot.

Balancing Bias and Variance

- **Bias**: Tendency of the model to consistently learn the wrong thing. High bias leads to underfitting.
- **Variance**: Tendency of the model to be overly sensitive to fluctuations in the training data. High variance leads to overfitting.
- The goal is to find a balance: Low bias and low variance, which equates to high accuracy on unseen data.

Flexibility and Overfitting in ML

Credits: Facure's Online Book

- The power you gain with ML is flexibility.
- ML can capture complicated functional forms in nuisance relationships.
- But this flexibility leads to the risk of overfitting.

Consequences of Overfitting

- Overfitting causes residuals to be smaller than they should be.
- Overfitting captures not only the relationship between Y and X but also D and Y, biasing the residual regression.

Overfitting and Treatment Variance

- Overfitting can also affect the treatment variable.
- Reduced variance in treatment leads to higher variance in the final estimator.
- Violation of positivity can also have similar effects.

Cross Prediction and Out-of-Fold Residuals

- Solution: Use cross prediction and out-of-fold residuals.
- Data is divided into K parts.
- Each part is estimated using the remaining $K-1$ parts to avoid driving residuals to zero.

Curse of dimensionality and DML

6. Interpretable Parameters: DML is often used to estimate a low-dimensional target (the treatment effect) while controlling for high-dimensional covariates. The curse of dimensionality is less of an issue when the parameter of interest is low-dimensional.

Summary and Next Steps

- We've covered the challenges and solutions in using ML for DML.
- Next, let's dive into a step-by-step example with more technical discussion

Frisch-Waugh-Lovell

- Old theorem showing that you can decompose a multivariate regression into a set of steps and get numerically the same number
- Used by DML so we should review it, though it is more often called “orthogonalized”

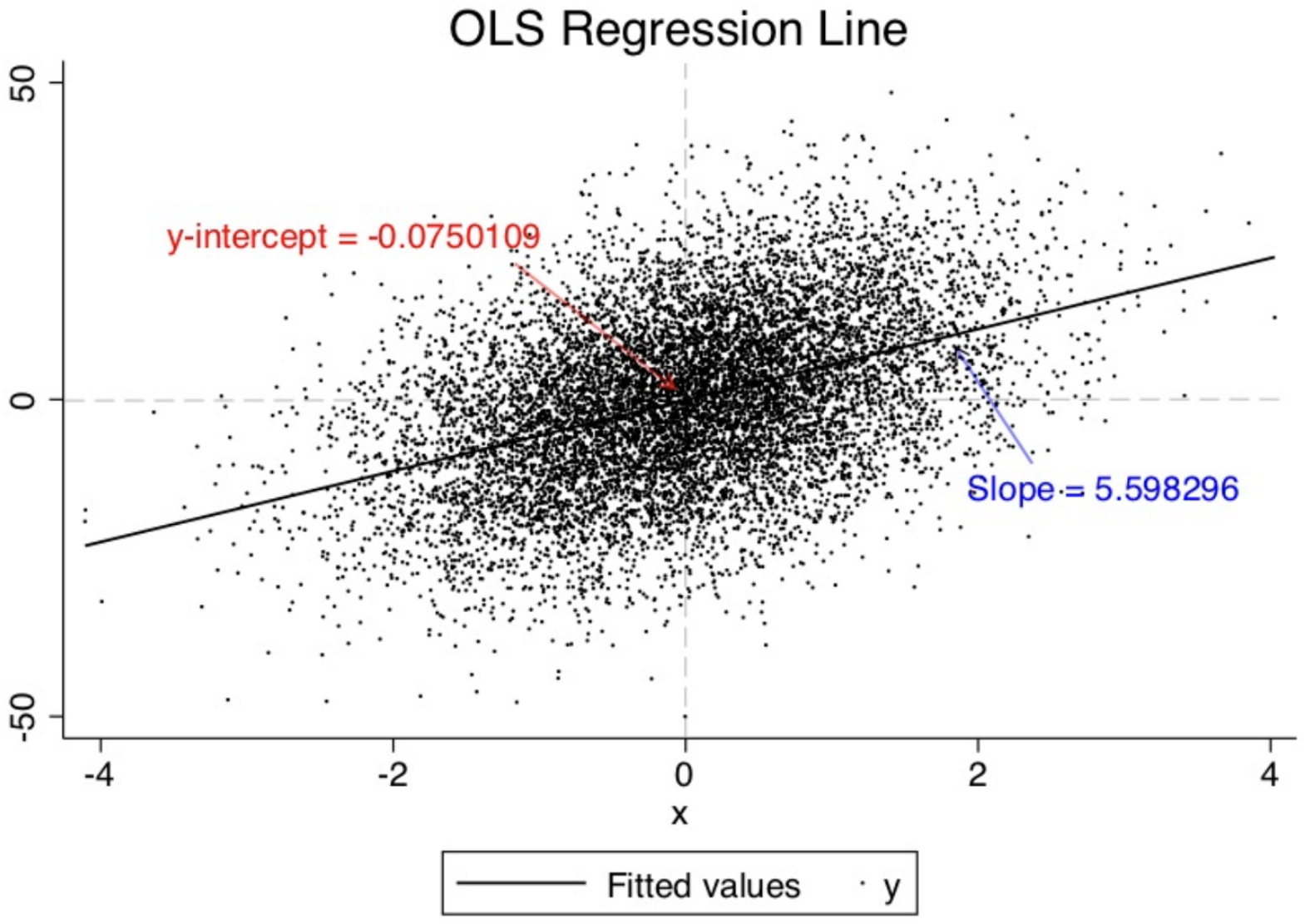
OLS Simulated Code

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 is written out as a scaled covariance:

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

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

Applying FWL theorem

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

Applying FWL theorem

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

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

- If family size is random, then we can interpret $\widehat{\beta_1}$ as the ATE
- But how do we interpret $\widehat{\beta_1}$ if `family size` is only conditionally random?

Applying FWL theorem

- Assume the longer model is:

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

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

FWL Theorem

FWL Theorem

Assume your main multiple regression model of interest:

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

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

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

and $\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}$ 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

Flexible Regression Model

Consider the flexible regression model:

$$Y_i = \tau D_i + g(X_i) + \epsilon_i$$

Our goal is to estimate the treatment effect τ.

DML Steps

1. Predict Y_i using X_i with Machine Learning and compute the residuals: $\tilde{Y} = Y - \hat{Y}_{\text{DML}}$
2. Predict D_i using X_i with Machine Learning and compute the residuals: $\tilde{D} = D - \hat{D}_{\text{DML}}$
3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$

Note: $\hat{Y}_{\text{DML}}$ and $\hat{D}_{\text{DML}}$ should be predictions generated by a machine learning model trained on a set of observations that does not include i. We accomplish this via cross-fitting.

DML and FWL

- Frisch-Waugh-Lovell (FWL) theorem is often viewed as a precursor to the debiasing steps in DML.
- Essentially, FWL allows you to estimate partial effects by taking residuals.
- You first remove the effect of control variables, and then run a simple univariate regression on these ”purified” residuals to get unbiased estimates.
- This is fundamentally what DML is doing but in a more flexible, nonparametric setting.

Frisch-Waugh-Lovell (FWL) Theorem

Consider the multiple linear regression model:

$$Y = X\beta + D\gamma + \epsilon$$

Recall from earlier – the FWL theorem tells us that we can obtain γ (effect of D) by:

1. Regressing D on X to get residuals $\tilde{D}$
2. Regressing Y on $\tilde{D}$ to get γ

Connection to DML

Now consider the model

$$Y_i = \tau D_i + g(X_i) + \epsilon_i$$

The steps in DML closely resemble FWL but are designed for a flexible function $g(x)$:

1. Regress Y_i on X_i using ML to get residuals $\tilde{Y}_i$
2. Regress D_i on X_i using ML to get residuals $\tilde{D}_i$
3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$ to estimate τ

This extends FWL to a machine learning setting.

Summary

- Both FWL and DML focus on ”purifying” the variables of interest by removing the influence of control variables.
- DML extends the logic of FWL to a flexible, nonparametric setting using machine learning methods.
- The core idea in both approaches is to focus on residuals, effectively controlling for confounders.

Cross-Fitting in DML

1. Divide the sample into K folds.
2. For $k = 1, \ldots, K$
   a. Train a model to predict Y given X, leaving out observations i in fold k: $\hat{Y}_{-k}(x)$
   b. Train a model to predict D given X, leaving out observations i in fold k: $\hat{D}_{-k}(x)$
   c. Form residuals $\tilde{Y}_i = Y_i - \hat{Y}_{-k}(X_i)$ and $\tilde{D}_i = D_i - \hat{D}_{-k}(X_i)$
3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$

Caveats and considerations:

- Use cross-validation to choose tuning parameters.
- For inference, use robust standard errors from the last step.

Role of Tuning Parameters in ML Models

- Tuning parameters, also known as hyperparameters, control the complexity of the machine learning models used.
- For example, in Lasso, the regularization term λ is a tuning parameter.
- These parameters are not learned from the data but must be set prior to the learning process.
- Too high or too low values can lead to overfitting or underfitting.

Cross-Validation for Tuning Parameter Selection

- Cross-validation helps us to choose the optimal tuning parameters.
- By splitting the data into training and validation sets multiple times, we can find the parameters that yield the best out-of-sample performance.
- This makes the DML approach more robust.

Why Out-of-Fold Prediction?

- Using the same data for training and prediction can introduce bias.
- Out-of-fold prediction ensures that each observation's residual is computed using a model that did not train on that observation.
- This step is crucial for unbiased treatment effect estimation.

Stability and Generalization in DML

- Cross-validation ensures the model's findings are stable across different subsamples.
- This adds an extra layer of validation and makes the findings more likely to generalize.

What is K-Folding?

- The dataset is randomly partitioned into K equally (or nearly equally) sized folds.
- One fold is used as the test set, and the remaining $K-1$ folds are used as the training set.
- This process is repeated K times, each time using a different fold as the test set.

Cross-Fitting in DML

- In DML, cross-fitting helps to ensure that the model is not using the same observations for both training and testing.
- This is essential for unbiased estimation.

Cross-Fitting Illustrated

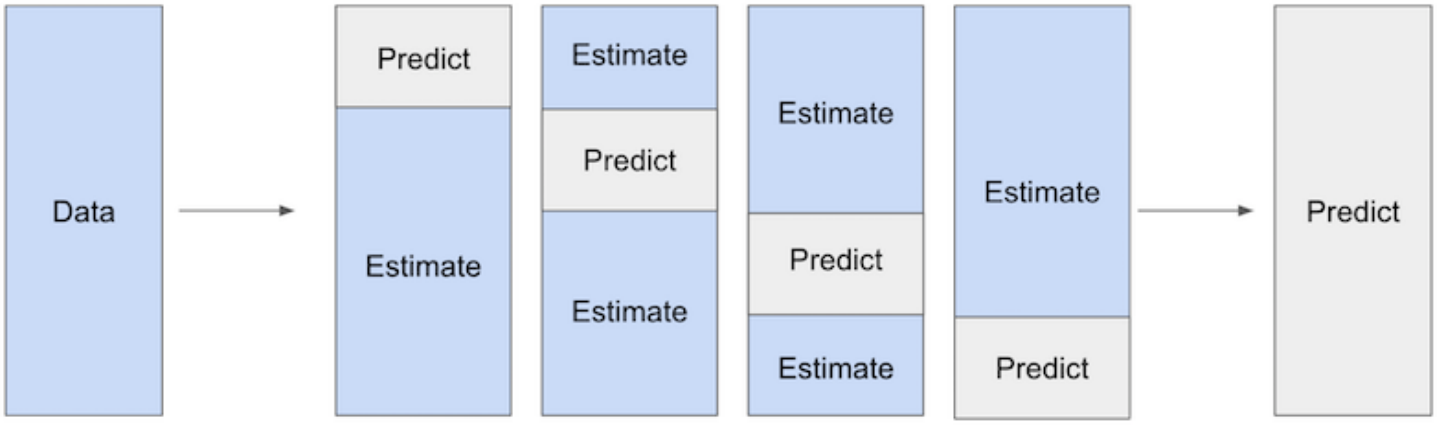


Figure: An illustration of cross-fitting in DML

How Cross-Fitting Works in DML

1. For each fold k:
   - → Train the model on $K-1$ folds to predict Y and D based on X.
   - → Use this trained model to predict the values in the left-out fold.
   - → Compute residuals $\tilde{Y}$ and $\tilde{D}$ for the left-out fold.
2. After looping through all K folds, you'll have residuals $\tilde{Y}$ and $\tilde{D}$ for all observations.
3. Regress $\tilde{Y}$ on $\tilde{D}$ to obtain the treatment effect.

Why Not Use All Observations?

- Using the same data for both training and testing could lead to overfitting.
- Overfitting would bias our estimates and make them unreliable.
- Cross-fitting via k-folding helps to mitigate this by ensuring each observation is ”out-of-sample” in one of the K iterations.

Role of Debiasing in DML

Credits: Facure's Online Book

- Role of D model: Debiasing the treatment.
- Residuals $\tilde{D}$ are treatment values where confounding bias has been removed.
- $\tilde{D}$ is orthogonal to X.

Example: Ice Cream and Temperature

- In Facure's example with Ice Cream and Temperature:
- Residuals remove the influence of weekends on higher prices.
- This results in an unbiased representation of the treatment (price).

Role of Denoising in DML

- Role of Y model: Denoising the outcome.
- It creates a version of the outcome where variance due to X has been explained.
- This makes causal estimation easier as the noise is reduced.

Summary

- We’ve delved into the roles of debiasing and denoising in DML.
- These steps are crucial for unbiased and accurate causal estimation.

Conditional ATE in DML

Credits: Facure's Online Book

- Double/Debiased ML is not limited to ATE.
- Can also be used to estimate Conditional Average Treatment Effect (CATE).
- Causal parameter changes based on unit's covariates.

Estimating CATE

- Use the same residualized version of treatment and outcome.
- Interact these residuals with other covariates.
- Fit a linear CATE model.

Final Linear Model for CATE

- Randomized test set used for CATE predictions.
- Final model is linear, enabling mechanical computation of CATE.

Generalization: R-Learner

- This procedure is a general one.
- Known as R-Learner, inspired by the work of Robinson (1988).
- Emphasizes the role of residualization.
- New loss function defined, which can be minimized in various ways.

Summary

- DML provides a flexible framework to estimate both ATE and CATE.
- CATE allows for nuanced treatment effect estimates based on covariates.

Roadmap

- Directed Acyclic Graphs
  - Backdoor criterion
  - Collider bias
- Unconfoundedness and Ignorable Treatment Assignment
  - Motivating estimation with an example
  - Aggregate target parameters
  - Assumptions
- Estimators
  - Subclassification
  - Regression
  - DML
- Concluding remarks

Buyer beware

- Difficult paper to read; encourage you to read online material
- Also remember the core assumptions remain: unconfoundedness
- That means avoiding colliders and having all known and quantified confounders
- Machine learning, including DML, can be “naive” too if the large dimension of features includes unknowingly colliders

Double Machine Learning and Automated Model Selection: A Cautionary Tale

Paul Hünermund[†] Beyers Louw[‡] Itamar Caspi[*]

[†] *Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK.*
E-mail: phu.si@cbs.dk
[‡] *Maastricht University, Tongersestraat 53, 6211 LM Maastricht, NL.*
E-mail: jb.louw@maastrichtuniversity.nl
[*] *Bank of Israel, P.O.Box 780, 91007, Jerusalem, IL*
E-mail: itamar.caspi@boi.org.il

This version: May 25, 2023
First version: August 26, 2021

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

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

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