Causal Inference I

Mixtape Session

MIXTAPE SESSIONS

Roadmap

- Selecting Covariates Using Directed Acyclic Graphs
  - Graph notation
  - Backdoor criterion
  - Collider bias
- Unconfoundedness and Ignorable Treatment Assignment
  - Motivating estimation with an example

Graphs

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

Adjusting for variables

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

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

Judea Pearl and DAGs

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

Design vs. Model

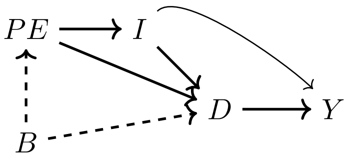
- DAGs require prior theory of treatment assignment in the world, making it more structural than design oriented methods which exploit randomization or simple treatment assignment mechanisms
- While DAGs are compatible with design based approaches, they are less likely to follow the same estimation methodologies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

Further reading

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

Causal model

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



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

DAGs are harder than they seem

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

Creating your DAG

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

Beware of Lazy DAGs

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

Concluding caveats

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

Unconfoundedness and the backdoor criterion

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

Confounding

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

$$
\begin{tikzcd}
D \arrow[rr] & & Y \\
& X \arrow[ul] \arrow[ur] &
\end{tikzcd}
$$

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

Backdoor Paths

- Confounding creates **backdoor paths** between treatment and outcome $(D \leftarrow X \rightarrow Y)$ – i.e., spurious correlations
  - $\rightarrow$ Distinct from something called a collider path $(D \rightarrow X \leftarrow Y)$
  - $\rightarrow$ 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)

$$
\begin{tikzcd}
D \arrow[rr] & & Y \\
& \boxed{X} \arrow[ul] \arrow[ur] &
\end{tikzcd}
$$

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

$$\begin{tikzcd} D \arrow[rr] & & Y \\ & \boxed{X} \arrow[ul] \arrow[ur] & \end{tikzcd}$$

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)

$$\begin{tikzcd}
D \arrow[rr] & & Y \\
& \boxed{X} \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

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.

$$\begin{tikzcd}
X_1 \arrow[r] \arrow[rr, bend left] & D \arrow[r] & Y \\
& & X_2 \arrow[ul] \arrow[u]
\end{tikzcd}$$

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

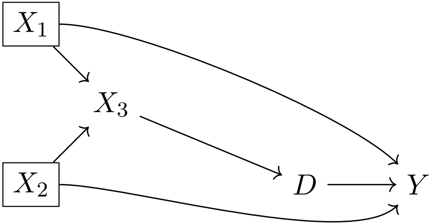
What if you have an unobservable?

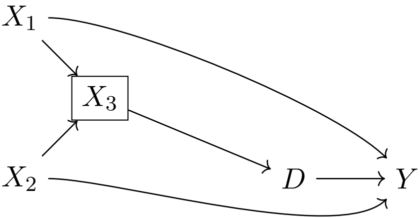
- List all the backdoor paths from D to Y.

$$
\begin{tikzcd}
 & X_1 \arrow[dr] \arrow[drr] & & \\
U \arrow[r, dashed] \arrow[rrr, dashed, bend right] & X_2 \arrow[r] & D \arrow[r] & Y
\end{tikzcd}
$$

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

Multiple strategies





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

Collider bias

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

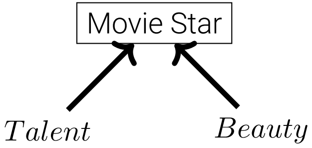
Example 1: Movie stars

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

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

Movie star DAG

Imagine casting directors pick movie stars based on talent and beauty



Talent and beauty can become correlated even though they are independent

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

Sample selection?

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

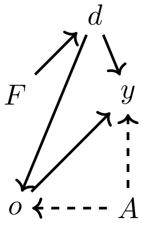
$$beauty_i = \alpha + \delta talent_i + \beta star_i + \varepsilon_i$$

- It’s not just sample selection

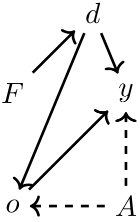
Example 2: Discrimination

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

DAG



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



Mediation and Backdoor paths

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

Table: Regressions illustrating collider bias with simulated gender disparity

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

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

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

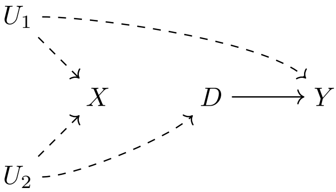
$$\begin{tikzcd}
\boxed{X_1} \arrow[r] \arrow[rrr, bend left] & D \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& & X_2 &
\end{tikzcd}$$

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

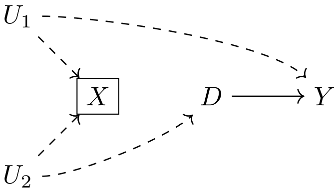
$$\begin{tikzcd}
\boxed{X_1} \arrow[r] \arrow[rrr, bend left] & D \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& & \boxed{X_2} &
\end{tikzcd}$$

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

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

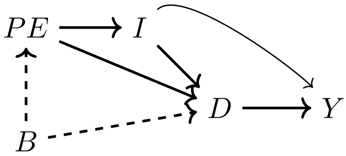


  - → But what if we condition on X?



Testing the Validity of the DAG

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



Contrast this with ordinary practices

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

Double Machine Learning and Automated Model Selection: A Cautionary Tale

Paul Hünermund[†] 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*

Covariate selection without DAGs

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

Falsifications as a test

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

Roadmap

Selecting Covariates Using Directed Acyclic Graphs

- Graph notation
- Backdoor criterion
- Collider bias

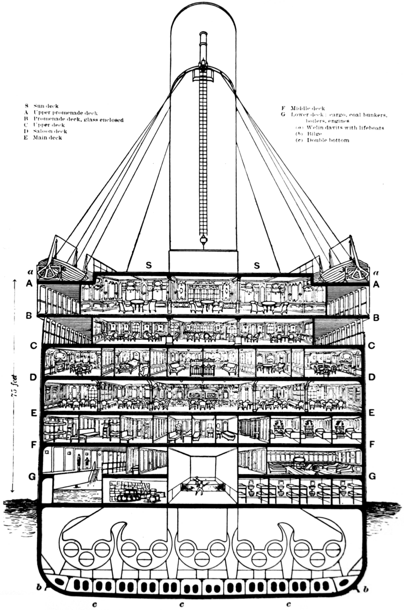
Unconfoundedness and Ignorable Treatment Assignment

- Motivating estimation with an example

Shifting gears

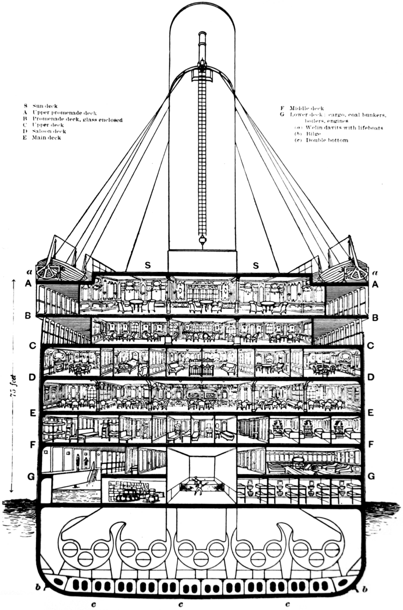
- Having reviewed the fundamentals of DAGs, we will now move into the practical realm
- You have a non-experimental study. You need to control for variables, but which ones?
- Let’s work through an example together: the sinking of the Titanic

Titanic layout



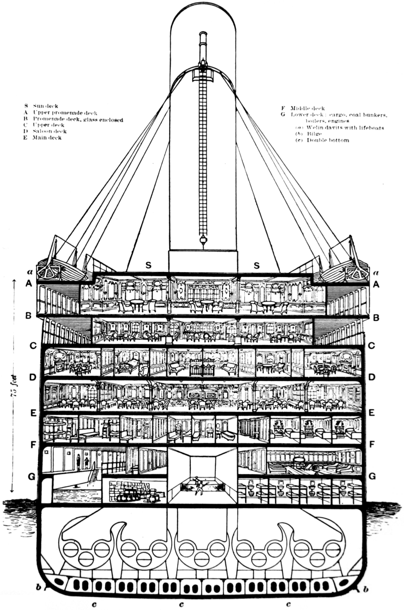
First Class: Decks B-C

Titanic layout



Second Class: Decks D-E

Titanic layout



Third Class: Decks F and G