Causal Inference

Mixtape Session

MIXTAPE SESSIONS

Roadmap

Introduction to course

Foundations of causal inference
- Princeton Industrial Relations and the Nobel
- Design versus Model
- Potential outcomes
- Selection bias
- Independence

Directed Acyclic Graphs
- Graph notation
- Backdoor criterion
- Collider bias
- Front door criterion
- Concluding remarks

Welcome to Mixtape Sessions!

- Mixtape Sessions is a new educational platform designed to "democratize causal inference" at all levels
- Causal inference, in my mind, is an *applied* field as much as it is a *technical* field and so learning more about it is to also learn about a range of topics not normally covered in an econometrics course
- These include econometric estimation, detailed exposition of research design elements, but also coding practices, handling of data, more detailed dives on specific topics and even advice on publishing and communicating results,

5-day Causal Inference Workshop

- Our workshop together is 5-days, 8am to 5pm CST, with 15 min breaks on the hour and a 1-hour lunch break at noon CST
- It will mix exposition, discussion of papers, coding exercises and discussion as best as I can
- It’s essentially a semester’s worth of material

Github repo

- We will communicate with one another regularly in the Discord channel and I will always be monitoring it
- I will be distributing things to you, like code and slides, via the github repo: `https://github.com/Mixtape-Sessions/Causal-Inference.git`
- Each lecture will be recorded and then uploaded to Vimeo as a password protected file that you'll have access to into perpetutity

What to expect

1. **Confidence**: You will feel like you have a good understanding of design-based causal inference by the end such that it doesn't feel so mysterious or intimidating
2. **Comprehension**: You will have learned a lot both conceptually but also in various specifics, particularly with regards to issues around identification and estimation
3. **Competency**: you will have had some experience working together implementing these methods using code in Stata and R, syntax, possession of programs, knowledge of packages

Topics

1. Foundations: Day 1
2. RDD: Day 1-2
3. IV: Day 2-3
4. DiD: Day 4
5. Synthetic control: Day 5
6. Selection on observables: Day 5

Different types of prediction

Prediction machines

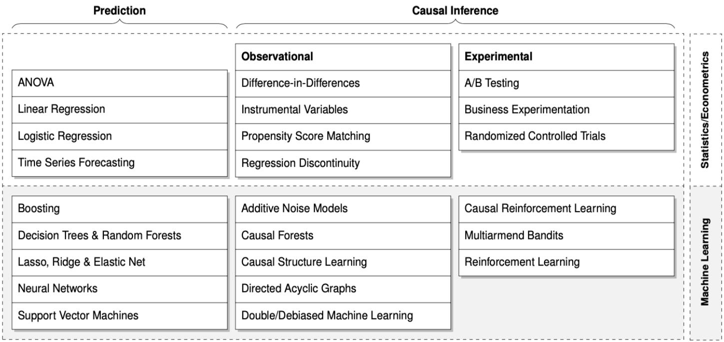
- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- “Does this person have heart disease?”, “How many books will I sell?”
- It is not predictions of what effect a choice will have, though

Causal inference

- Causal inference is also a type of prediction, but it’s a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

Identification problem

Figure 1: Examples of popular data analysis algorithms in statistics and econometrics, as well as machine learning and artificial intelligence, classified according to prediction and causal inference methods. Causal inference methods are further differentiated according to observational (based on ex-post observed data) and experimental approaches.



- Causal inference is the right columns; prediction the left columns
- This course is about the middle top column only, but that is only because it's a focused course on a particular type of causal inference methodology

Roadmap

Introduction to course

Foundations of causal inference

- Princeton Industrial Relations and the Nobel
- Design versus Model
- Potential outcomes
- Selection bias
- Independence

Directed Acyclic Graphs

- Graph notation
- Backdoor criterion
- Collider bias
- Front door criterion
- Concluding remarks

Princeton and credibility

- October 2021's Nobel Prize in economics went to Card, Angrist and Imbens
- But it's arguably as much to Princeton's mid 1980s Industrial Relations group as it's ground zero for the credibility revolution
- Starts with Orley Ashenfelter, advised by Albert Rees, who had been working on job trainings programs

Princeton and credibility

- Panel models were not satisfactory for recovering causal effects in Ashenfelter dip situations
- Increasingly less focus on modeling the outcome and more focus on treatment variation
- Key players are Orley Ashenfelter, David Card (Orley's student), Josh Angrist (Card and Orley's student), Alan Krueger (hired by Orley), Bob Lalonde (Card and Orley's student) and then a generation of students (Levine, Currie, Pischke)

Collaborations

- Angrist writes a dissertation on Vietnam draft, focusing on randomized instruments – broadly there's a growing interest in what randomization buys you
- Angrist goes to Harvard, overlaps with Imbens for a year, they are mentored by Gary Chamberlain, begin working with Don Rubin
- Chamberlain recommends potential outcomes framework over a different one that had been used at that time (latent index)
- This course is about these people and those ideas and their subsequent development – what I call the "design based causal inference"

Theoretical Models

All models are wrong but some are useful – George Box

- Models reduce an infinitely complex world into something so simple that it borders on being false, and yet that's a feature not a bug
- But the connection between the theoretical model (e.g., utility maximization and labor supply functions) and the empirical micro models have moved through phases
- Card (2014; 2021) gave this interpretation of labor in a speech he gave at Michigan and then the Nobel acceptance speech

Causality and the model

Empirical micro was divided along two almost philosophical approaches to causal inference over the years (and still is)

- **Model**: Causality is model-based. It only exists within the framework of a theory that says “X causes Y” (e.g., Heckman)
- **Design**: Causality is design-based. No causality without *physical* manipulation of a treatment X (e.g., Rubin, Holland)

Approximating models

1. **Approximating models**: Consumer demand, labor supply models (e.g., Mincer 1958; 1974)
   - → Theory implies $y_i = f_i(x_i)$ with restrictions on f_i (e.g., concavity)
   - → Researcher estimates a simpler version

$$y_i = \alpha + x_i\beta + \varepsilon_i$$

Exact models

2. **Exact models**: Models gives us all causes ("complete DGP")
   - → More structural approach to identification, less focused on physical assignment of treatments
   - → Estimate model parameters and distribution of heterogeneity
   - → Functional form, useful for welfare analysis

Working model

3. **Working model**: Program evaluation (e.g., Princeton)
   - → Focus is on physical assignment of treatments (putting it in Fisher tradition on the RCT)
   - → Model formulates questions, intuition, but does not necessarily assist with identification

Confidence differs

Schools of thoughts use their models in very different ways

- **Design**: Focus tends to be falsifiable predictions from theory (e.g., immigration reduces domestic employment)
- **Model**: Stipulate complete models with an accompanying goal of estimating parameters, do welfare analysis, counterfactual estimation from within the model

Topics

Dependence on the model vs freed from the model for causal inference increases topics

- **Design**: Anything goes, “economics is what economists study”, happiness, fringe stuff (e.g., sex work) (opening up topics)
- **Model**: Neoclassical topics due to needing agreed upon models (limiting topics)

Design vs Model *within* Design

Confusingly, within the broadly design tradition we will often now hear about “design vs. model” identification. What?

- **Design-design**: emphasizes randomization for identification which is inside RCTs, IV, even matching (i.e., conditional independence)
- **Design-model**: restricts the “behavior” of unobserved potential outcomes through appeals to parallel trends (DiD), smoothness (RDD), factor models (synthetic control)

Design lasting impact

- Potential outcomes, counterfactuals and causality, focus on treatment assignment, randomization, credible instruments (Orley: “I am old enough to remember when people wouldn’t even say what instruments they used”)
- Substantive specification tests: balance across covariates, pre-treatment comparisons, event studies, falsification – maybe even the growing page length!
- Focus on data quality, replication, data warehouses, journals requiring authors submit programs, pre-registration of RCTs, and increasingly, the lab model

Introduction to Counterfactuals

- Aliens come and orbit earth, see sick people in hospitals and conclude “doctors are hurting people”
- They kill the doctors, unplug patients from machines, throw open the doors – many patients inexplicably die
- Ridiculous to us but only because we know what hospitals are – they don’t
- Consider this: aren’t we the aliens in our research?
- Three types of errors

#1: Correlation and causality are different

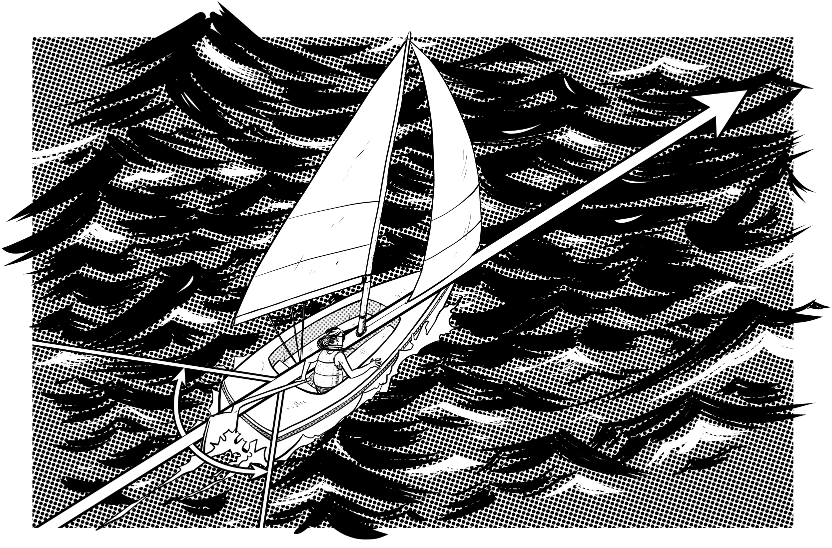
Causal is one unit, correlation is many units

- Causal question: “If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?”
- Correlation question:

$$\frac{Cov(D,Y)}{\sqrt{Var_D}\sqrt{Var_Y}}$$

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- What if cat killed the rooster?
- *Post hoc ergo propter hoc*: “after this, therefore, because of this”



#3: No correlation does not mean no causality!

- A sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- The same aliens observe from space and say “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken.” So they send her a new rudder
- They’re wrong but why are they wrong? There is, after all, no correlation
- Example: Fed and open market operations

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if hospitalized at time } t \\ 0 \text{ if not hospitalized at time } t \end{cases}$$

  where i indexes an individual observation, such as a person

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1 \text{ health if hospitalized at time } t \\ 0 \text{ health if not hospitalized at time } t \end{cases}$$

  where j indexes a counterfactual state of the world

Moving between worlds

- I'll drop t subscript, but note – these are potential outcomes for the same person at the exact same moment in time
- A potential outcome Y^1 is not the historical outcome Y either conceptually or notationally
- Potential outcomes are hypothetical states of the world but historical outcomes are ex post realizations
- Major philosophical move here: go from the potential worlds to the actual (historical) world based on your treatment assignment

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, equals $Y_i^1 - Y_i^0$

Definition 3: Switching equation

An individual's observed health outcomes, Y, is determined by treatment assignment, D_i, and corresponding potential outcomes:

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\
Y_i &= \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}
\end{aligned}
$$

Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$
\begin{aligned}
E[\delta_i] &= E[Y_i^1 - Y_i^0] \\
&= E[Y_i^1] - E[Y_i^0]
\end{aligned}
$$

So what's the problem?

Definition 4: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

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

Conditional Average Treatment Effects

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

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

Any collection of treatment effects

- Notice how in all three of these, all we did was take the defined treatment effect at the individual and aggregate
- We will see this again with IV when we introduce the "local" average treatment effect
- Just keep in mind – these parameters can be defined, but they cannot be calculated due to the switching equation

Good and bad variation

- Naive use of statistical models will often find and take advantage of all types of variation for the purpose of prediction
- But causal inference is much more cautious because it only uses *some* of the variation
- This is better seen with a story and a decomposition

Causality and comparisons

- Two groups of people infected with COVID have different mortality rates – one group is on vents, the other isn't
- People on vents have higher mortality than those who aren't. Is the vent assignment harming or helping them?
- If we compare average health outcomes for a group of COVID infected individuals on vents (treatment) versus in the population (control), what exactly are we measuring?

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by the sample averages:

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

I tend to use expectation operators $E[.]$ but note we are using samples $E_N(.)$

SDO

- Simple difference in mean outcomes is our first estimator
- Notice that we switched from potential outcomes to observed outcomes
- This means that because the SDO is based on the switching equation, it uses data
- So when is the SDO causal and when is it not?

Potentially biased comparisons

Decomposition of the SDO

The SDO can be decomposed into the sum of three parts:

$$\begin{aligned} E[Y^1|D=1] - E[Y^0|D=0] &= ATE \\ &+ E[Y^0|D=1] - E[Y^0|D=0] \\ &+ (1-\pi)(ATT - ATU) \end{aligned}$$

Seeing is believing so let's work through this identity!

Use LIE to decompose ATE into the sum of four conditional average expectations

$$\begin{aligned}
\text{ATE} &= E[Y^1] - E[Y^0] \\
&= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\
&\quad -\{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\}
\end{aligned}$$

Substitute letters for expectations

$$\begin{aligned}
E[Y^1|D=1] &= a \\
E[Y^1|D=0] &= b \\
E[Y^0|D=1] &= c \\
E[Y^0|D=0] &= d \\
\text{ATE} &= e
\end{aligned}$$

Rewrite ATE

$$e = \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\}$$

Move SDO terms to LHS

$$
\begin{aligned}
e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\
e &= \pi a + b - \pi b - \pi c - d + \pi d \\
e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\
0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)
\end{aligned}
$$

Rewrite from previous slide

$$\mathbf{a} - \mathbf{d} \quad = \quad e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)$$

Substitute conditional means

$$\begin{aligned}
E[Y^1|D=1] - E[Y^0|D=0] \quad &= \quad \text{ATE} \\
&+(E[Y^0|D=1] - E[Y^0|D=0]) \\
&+(1-\pi)(\{E[Y^1|D=1] - E[Y^0|D=1]\} \\
&-(1-\pi)\{E[Y^1|D=0] - E[Y^0|D=0]\}) \\
E[Y^1|D=1] - E[Y^0|D=0] \quad &= \quad ATE \\
&+(E[Y^0|D=1] - E[Y^0|D=0]) \\
&+(1-\pi)(ATT - ATU)
\end{aligned}$$

Decomposition of difference in means

$$
\begin{aligned}
\underbrace{E_N[y_i|d_i=1]-E_N[y_i|d_i=0]}_{\text{SDO}} \quad = \quad & \underbrace{E[Y^1]-E[Y^0]}_{\text{Average Treatment Effect}} \\
& + \underbrace{E[Y^0|D=1]-E[Y^0|D=0]}_{\text{Selection bias}} \\
& + \underbrace{(1-\pi)(ATT-ATU)}_{\text{Heterogenous treatment effect bias}}
\end{aligned}
$$

Using the switching equation, we get $E_N[Y|D=1] \to E[Y^1|D=1]$, $E_N[Y|D=0] \to E[Y^0|D=0]$ and $(1-\pi)$ is the share of the population in the control group.

Independence

Independence assumption

Treatment is assigned to a population independent of that population's potential outcomes

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

This is random or quasi-random assignment and ensures mean potential outcomes for the treatment group and control group are the same. Also ensures other variables are distributed the same for a large sample.

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

Random Assignment Solves the Selection Problem

$$\underbrace{E_N[y_i|d_i=1] - E_N[y_i|d_i=0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}}$$

$$+ \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}}$$

$$+ \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

- If treatment is independent of potential outcomes, then swap out equations and **selection bias** zeroes out:

$$E[Y^0|D=1] - E[Y^0|D=0] = 0$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D=1] - E[Y^0|D=1]$$
$$\text{ATU} = E[Y^1|D=0] - E[Y^0|D=0]$$

- Rewrite the third row bias after $1-\pi$:

$$\begin{aligned} ATT - ATU &= \mathbf{E[Y^1 \mid D=1]} - E[Y^0|D=1] \\ &\quad -\mathbf{E[Y^1 \mid D=0]} + E[Y^0|D=0] \\ &= 0 \end{aligned}$$

- If treatment is independent of potential outcomes, then:

$$\begin{aligned} E_N[y_i|d_i=1] - E_N[y_i|d_i=0] &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

SUTVA

- Potential outcomes model places a limit on what we can measure: the "stable unit-treatment value assumption" . Horrible acronym.
    1. **S**: ***s****table*
    2. **U**: across all ***u****nits*, or the population
    3. **TV**: ***t****reatment-value* ("treatment effect", "causal effect")
    4. **A**: ***a****ssumption*
- As this is a bit of a pregnant concept, let's go slow

SUTVA: Unit-level assignment only

- Most people, if they know of SUTVA, tend to associate with one of its elements not its core definition
- It's core definition is actually the switching equation:

$$Y_{i,t} = D_{i,t}Y^1_{i,t} + (1 - D_{i,t})Y^0_{i,t}$$

- Notice now the i and t subscripts; think of what that means

SUTVA: No Anticipation

- A particular unit i at some point in time t assigns potential outcome for unit i at time t to outcome based on *its* contemporanous treatment assignment for the same i unit at the same t time
- Outcomes are **not** someone else's (spillovers), nor on future assignment (anticipation)
- Example: I increase spending based on a future raise I haven't yet gotten

SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable”
- If we are estimating the effect of vents on covid symptoms, we assume everyone is getting the same kinds of vents more or less.
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*

SUTVA: No spillovers to other units

- What if putting someone on a ventilator causes someone else to be more or less likely to develop severe covid symptoms?
- Have to think hard about externalities, particularly with transmissible diseases
- SUTVA means that you don't have a problem like this.
- If there are no externalities from treatment, then δ_i is stable for each i unit regardless of whether someone else receives the treatment too, but herd immunity must be considered when it comes to cures

SUTVA: Partial equilibrium only

Easier to imagine this with a different example.

- Let’s say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?
- Scaling up a policy can be challenging to predict if there are rising costs of production
- What if expansion requires hiring lower quality teachers just to make classes?
- That’s a general equilibrium effect; we only estimated a partial equilibrium effect (external versus internal validity)

Demand for Learning HIV Status

- Rebecca Thornton implemented an RCT in rural Malawi for her job market paper at Harvard in mid-2000s
- At the time, it was an article of faith that you could fight the HIV epidemic in Africa by encouraging people to get tested; but Thornton wanted to see if this was true
- She randomly assigned cash incentives to people to incentivize learning their HIV status
- Also examined whether learning changed sexual behavior.

Experimental design

- Respondents were offered a free door-to-door HIV test
- Treatment is randomized vouchers worth between zero and three dollars
- These vouchers were redeemable once they visited a nearby voluntary counseling and testing center (VCT)
- Estimates her models using OLS with controls

Why Include Control Variables?

To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the SDO, we might estimate:

$$Y_i = \alpha + \delta D_i + \gamma X_i + \eta_i$$

Why Control Variables?

- There are 2 main reasons for including additional controls in the regression models:
    1. Conditional random assignment. Sometimes randomization is done *conditional* on some observable (e.g., gender, school, districts)
    2. Exogenous controls increase precision. Although control variables X_i are uncorrelated with D_i, they may have substantial explanatory power for Y_i. Including controls thus reduces variance in the residuals which lowers the standard errors of the regression estimates.
- Ongoing work by econometricians is investigating this more carefully

Table: Impact of Monetary Incentives and Distance on Learning HIV Results

	1	**2**	**3**	**4**	**5**
Any incentive	0.431***	0.309***	0.219***	0.220***	0.219 ***
	(0.023)	(0.026)	(0.029)	(0.029)	(0.029)
Amount of incentive		0.091***	0.274***	0.274***	0.273***
		(0.012)	(0.036)	(0.035)	(0.036)
Amount of incentive^2			-0.063***	-0.063***	-0.063***
			(0.011)	(0.011)	(0.011)
HIV	-0.055*	-0.052	-0.05	-0.058*	-0.055*
	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Distance (km)				-0.076***	
				(0.027)	
Distance^2				0.010**	
				(0.005)	
Controls	Yes	Yes	Yes	Yes	Yes
Sample size	2,812	2,812	2,812	2,812	2,812
Average attendance	0.69	0.69	0.69	0.69	0.69

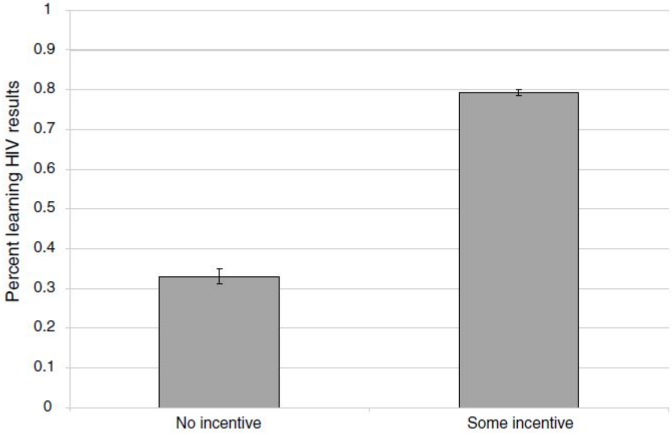


Figure: Visual representation of cash transfers on learning HIV test results.

Results

- Even small incentives were effective
- Any incentive increases learning HIV status by 43% compared to the control (mean 34%)
- Next she looks at the effect that learning HIV status has on risky sexual behavior

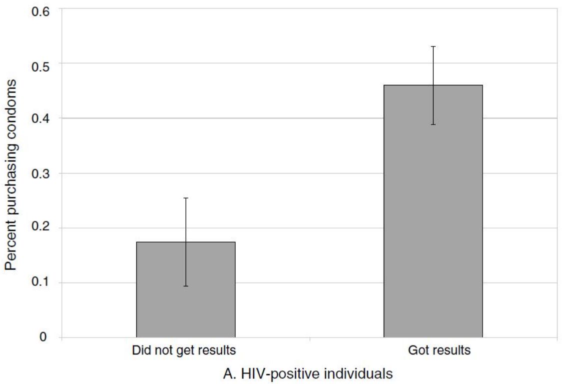


Figure: Visual representation of cash transfers on condom purchases for HIV positive individuals.

Table: Reactions to Learning HIV Results among Sexually Active at Baseline

Dependent variables:	**Bought condoms**		**Number of condoms bought**	
	OLS	**IV**	**OLS**	**IV**
Got results	-0.022	-0.069	-0.193	-0.303
	(0.025)	(0.062)	(0.148)	(0.285)
Got results $\times$ HIV	0.418***	0.248	1.778***	1.689**
	(0.143)	(0.169)	(0.564)	(0.784)
HIV	-0.175**	-0.073	-0.873	-0.831
	(0.085)	(0.123)	(0.275)	(0.375)
Controls	Yes	Yes	Yes	Yes
Sample size	1,008	1,008	1,008	1,008
Mean	0.26	0.26	0.95	0.95

Results

- For those who were HIV+ and got their test results, 42% more likely to buy condoms (but shrinks and becomes insignificant at conventional levels with IV).
- Number of condoms bought – very small. HIV+ respondents who learned their status bought 2 more condoms

Discussion

- What's in your field a causal question you find interesting that you wish you could answer?
- Describe the way you would conduct the RCT by explaining the following:
  - → What's the treatment? Express it as a binary variable.
  - → How will you assign this so that SUTVA holds and independence is achieved?
  - → What is the outcome you are interested in?
- Describe the steps you would take to do this if you had all the money in the world

Roadmap

- Introduction to course
- Foundations of causal inference
  - Princeton Industrial Relations and the Nobel
  - Design versus Model
  - Potential outcomes
  - Selection bias
  - Independence
- Directed Acyclic Graphs
  - Graph notation
  - Backdoor criterion
  - Collider bias
  - Front door criterion
  - Concluding remarks

Judea Pearl and DAGs

- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology
- They 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

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

Further reading

1. Pearl (2018) The Book of Why: The New Science of Cause and Effect, Basic Books (*popular*)
2. Morgan and Winship (2014) Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press, 2nd edition (*excellent*)
3. Pearl, Glymour and Jewell (2016) Causal Inference In Statistics: A Primer, Wiley Books (*accessible*)
4. Pearl (2009) Causality: Models, Reasoning and Inference, Cambridge, 2nd edition (*difficult*)
5. Cunningham (2021) Causal Inference: The Mixtape, Yale, 1st edition (*best choice, no question*)

Design vs. Model

- DAGs tend to be focused more on the theory of treatment assignment in the world
- As such it's compatible with design-based approaches
- But assumptions in design based approaches tend to emphasize selection into treatment which is not exactly what is meant here

Causal model

- The causal model is sometimes called the structural model, but for us, I prefer the former as it's less alienating
- Think of this as more connected to the model-based approach discussed earlier
- 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
- To illustrate, I will assume a Beckerian human capital model

Human capital model: statements *and* graphs

Let's describe my simplified Beckerian human capital model.

- Individuals maximize utility by choosing consumption and schooling (D) subject to multi-period budget constraint
- Education has current costs but longterm returns
- But people choose different levels of schooling based on a number of things we will call "background" (B) which won't be in the dataset ("unobserved")
- And own-schooling will also be because of parental schooling (PE)
- Finally, wages (Y) are a function of parental schooling

Becker's human capital causal model

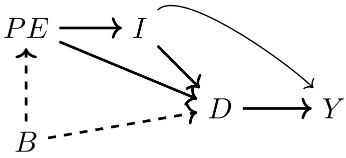
We can represent that causal model visually

$$\begin{tikzcd}
PE \arrow[r] \arrow[drr] & I \arrow[dr] \arrow[drr, bend left] & & \\
& & D \arrow[r] & Y \\
B \arrow[uu, dashed] \arrow[urr, dashed] & & &
\end{tikzcd}$$

PE is parental education, B is "unobserved background factors (i.e., "ability")", I is family income, D is college education and Y is log wages. The DAG is an approximation of Becker's underlying (causal) human capital model.

Arrows, but also *missing arrows*

Before we dive into all this notation, couple of things

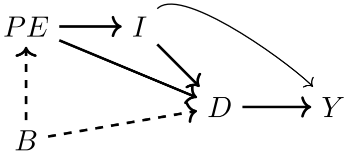


PE and D are caused by B. But why doesn't B cause Y?? Do you believe this? Why/why not? We can dispute this, but notice – we can see the assumption, which is transparent and communicates the author's beliefs, as well as the needed assumptions in their forthcoming *empirical* model. Every empirical strategy makes assumptions, but oftentimes they are not as transparent to us as this is.

$$
\begin{tikzcd}
PE \arrow[r] \arrow[drr] & I \arrow[dr] \arrow[drr, bend left] & & \\
& & D \arrow[r] & Y \\
B \arrow[uu, dashed] \arrow[urr, dashed] & & &
\end{tikzcd}
$$

- 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 four **paths** from PE to Y but none are direct, and one is unlike the others

Colliders



Notice anything different with this DAG? Look closely.

- D is a **collider** along the path $B \rightarrow D \leftarrow I$ (i.e., “colliding” at D)
- D is a **noncollider** along the path $B \rightarrow D \rightarrow Y$

Summarizing Value of DAGs imo

1. Facilitates the task of designing identification strategy for estimating average causal effects
2. Facilitates the task of testing compatibility of the model with your data
3. Visualizes the identifying assumptions which opens up the model to critical scrutiny

Creating DAGs

- The DAG is a *relevant* causal relationships describing the relationship between D and Y
- It will include:
  - $\rightarrow$ All direct causal effects among the *relevant* variables in the graph
  - $\rightarrow$ All common causes of any pair of *relevant* variables in the graph
- No need to model a dinosaur stepping on a bug causing in a million years some evolved created that impacted your decision to go to college
- We get ideas for DAGs from theory, models, observation, experience, prior studies, intuition
- Sometimes called the data generating process.

Confounding

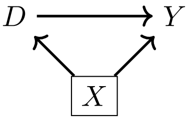
- Omitted variable bias has a name in DAGs: “confounding”
- Confounding occurs when when the treatment and the outcomes have a common cause or parent which creates spurious correlation between D and Y

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

The *correlation* between D and Y no longer reflects the causal effect of D on Y

Backdoor Paths

- Confounding creates **backdoor paths** between treatment and outcome ($D \leftarrow X \rightarrow Y$) – i.e., spurious correlations
- Not the same as mediation ($D \rightarrow X \rightarrow Y$)
- We can "block" backdoor paths by conditioning on the common cause X
- Once we condition on X, the correlation between D and Y estimates the causal effect of D on Y
- Conditioning means calculating $E[Y|D=1,X] - E[Y|D=0,X]$ for each value of X then combining (e.g., integrating)



Blocked backdoor paths

A backdoor path is blocked if and only if:

- It contains a noncollider that has been conditioned on
- Or it contains a collider that has not been conditioned on

Examples of blocked paths

Examples:

1. Conditioning on a noncollider blocks a path:

$$\begin{tikzcd} \boxed{X} \arrow[r] \arrow[rr, bend left] & Z & Y \end{tikzcd}$$

2. Conditioning on a collider opens a path (i.e., creates spurious correlations):

$$\begin{tikzcd} Z \arrow[r] & \boxed{X} & Y \arrow[l] \end{tikzcd}$$

3. *Not* conditioning on a collider blocks a path:

$$Z \longrightarrow X \longleftarrow Y$$

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

In words: If X satisfies the backdoor criterion with respect to (D, Y), then controlling for or matching on X identifies the causal effect of D on Y

What control strategy meets the backdoor criterion?

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

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

- 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

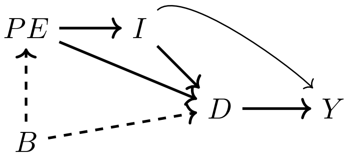
$$
\begin{tikzcd}
\boxed{X_1} \arrow[dr] \arrow[drrr, bend left] & & & \\
& X_3 \arrow[rr] & & D \arrow[r] & Y \\
\boxed{X_2} \arrow[ur] \arrow[urrrr, bend right] & & & &
\end{tikzcd}
$$

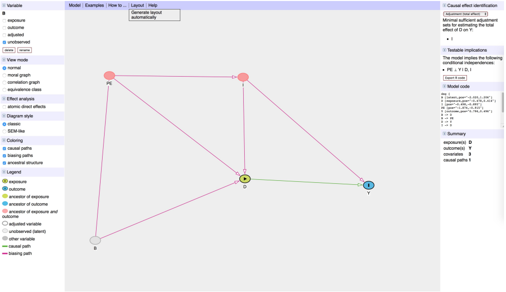
$$
\begin{tikzcd}
X_1 \arrow[dr] \arrow[drrrr, bend left] & & & & \\
& \boxed{X_3} \arrow[rr] & & D \arrow[r] & Y \\
X_2 \arrow[ur] \arrow[urrrr, bend right] & & & &
\end{tikzcd}
$$

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

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




Variable
B
exposure
outcome
adjusted
unobserved
delete
rename
View mode
normal
moral graph
correlation graph
equivalence class
Effect analysis
atomic direct effects
Diagram style
classic
SEM-like
Coloring
causal paths
biasing paths
ancestral structure
Legend
exposure
outcome
ancestor of exposure
ancestor of outcome
ancestor of exposure and outcome
adjusted variable
unobserved (latent)
other variable
causal path
biasing path
Model
Examples
How to ...
Layout
Help
Generate layout automatically
PE
I
D
Y
B
Causal effect identification
Adjustment (total effect)
Minimal sufficient adjustment sets for estimating the total effect of D on Y:
• I
Testable implications
The model implies the following conditional independences:
• PE ⊥ Y | D, I
Export R code
Model code
dag {
B [latent,pos="-2.020,1.206"]
D [exposure,pos="-0.478,0.414"]
I [pos="-0.498,-0.895"]
PE [pos="-1.874,-0.915"]
Y [outcome,pos="0.796,0.496"]
B -> D
B -> PE
D -> Y
I -> D
Summary
exposure(s) D
outcome(s) Y
covariates 3
causal paths 1

Collider bias

- **Conditioning on a collider introduces spurious correlations; can even mask causal directions**
  - → There is only one backdoor path from D to Y

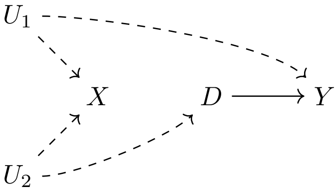
$$
\begin{tikzcd}
\boxed{X_1} \arrow[r] \arrow[rrr, bend left] & D \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& & X_2 &
\end{tikzcd}
$$

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

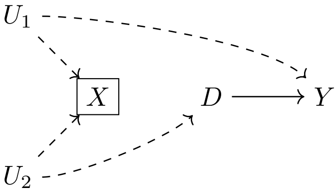
$$
\begin{tikzcd}
\boxed{X_1} \arrow[r] \arrow[rrr, bend left] & D \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& & \boxed{X_2} &
\end{tikzcd}
$$

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

- **Even controlling for pretreatment covariates can create bias**
  - → Name the backdoor paths. Is it open or closed?



  - → But what if we condition on X?



Living in reality - he doesn't love you

- **Fact #1**: We can't know if we have a collider bias (confounder) problem without making assumptions about the causal model (i.e. not in the codebook)
- **Fact # 2**: You can't just haphazardly throw in a bunch of controls on the RHS (i.e., "the kitchen sink") bc you may inadvertently be conditioning on a collider which can lead to massive biases
- **Fact # 3**: You have no choice but to leverage economic theory, intuition, intimate familiarity with institutional details and background knowledge for research designs.
- **Fact #4**: You can only estimate causal effects with **data** and **assumptions**.

Examples of collider bias

Bad controls

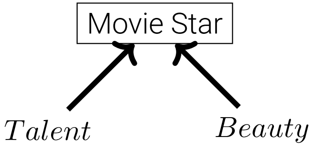
- Angrist and Pischke in MHE talk about a specific type of danger associated with controlling for an outcome – “bad controls”
- The problem is not controlling for an outcome;
- The problem is controlling for a collider and don’t correct for *that*
- This has implications for when you work with non-random administrative data, too

Sample selection example of collider bias

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)?

- What if the sample consists *only* of movie stars?



Stata code

```
clear all
set seed 3444

* 2500 independent draws from standard normal distribution
set obs 2500
generate beauty=rnormal()
generate talent=rnormal()

* Creating the collider variable (star)
gen score=(beauty+talent)
egen c85=pctile(score), p(85)
gen star=(score>=c85)
label variable star "Movie star"

* Conditioning on the top 15%
twoway (scatter beauty talent, mcolor(black) msize(small) msymbol(smx)),
ytitle(Beauty) xtitle(Talent) subtitle(Aspiring actors and actresses) by(star,
total)
```

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.

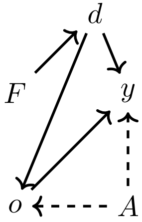
Stata

- Run Stata file star.do

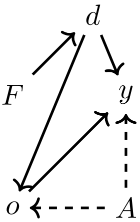
Occupational sorting and discrimination example of collider bias

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

DAG



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



Mediation and Backdoor paths

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

Stata model (Erin Hengel)

- Erin Hengel (`www.erinhengel.com`) and I worked out this code and she gave me permission to put in my Mixtape
- Let's look at collider_discrimination.do or collider_discrimination.R together

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 $d \rightarrow o \rightarrow y$ and $d \rightarrow y$
- Because it seems intuitive to control for occupation, notice column 2 - the sign flips!
- We are only able to isolate the direct causal effect by conditioning on ability and occupation, but ability is unobserved

Administrative data

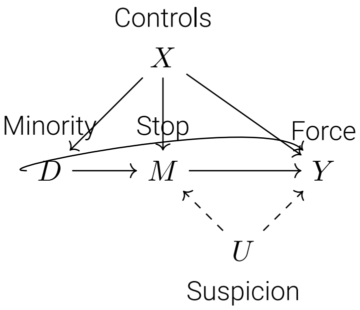
- Admin data has become extremely common, if not absolutely necessary
- But naive use of admin data can be dangerous if the drawing of the sample is itself a collider problem (Heckman 1979; Elwert and Winship 2014)
- Let's look at a new paper by Fryer (2019) and a critique by Knox, et al. (2019)

Collider bias and police use of force

- Claims of excessive and discriminator use of police force against minorities (e.g., Black Lives Matter, Trayvon Martin, Michael Brown, Eric Garner)
- Challenging to identify
  - → Police-citizen interactions are conditional on interactions having already been triggered
  - → That initial interaction is unobserved
- Fryer (2019) is a monumental study for its data collection and analysis: Stop and Frisk, Police-Public Contact Survey, and admin data from two jurisdictions
- Codes up almost 300 variables from arrest narratives which range from 2-100 pages in length – shoeleather!

Initial interaction

- Fryer finds that blacks and Hispanics were more than 50% more likely to have an interaction with the policy in NYC Stop and Frisk as well as Police-Public Contact survey
- It survives extensive controls – magnitudes fall, but still very large (21%)
- Moves to admin data
- Conditional on police interaction, *no* racial differences in officer-related shootings
- Fryer calls it one of the most surprising findings in his career
- Lots of eyes on this study as a result of the counter intuitive results; published in JPE
- Knox, et al (202) claim his data is itself a collider. What?



Fryer told us $D \to M$ exists from both Stop and Frisk and Police-Public. But note: admin data is instances of M stops, which is itself a collider. If this DAG is true, then spurious correlations enter between M and Y which may dilute our ability to estimate causal effects.

Knox, et al (2020)

- Move from DAG to more contemporary potential outcomes notation to design relevant parameters
- Use potential outcomes and bounds
- Even with lower bound estimates of the incidence of police violence against civilians is more than 5x higher than what Fryer (2019) finds
- Heckman (1979) – we *cannot* afford to ignore sample selection

Mechanisms

- Rarely does an intervention operate directly on an outcome
  - → Parental substance abuse causes foster care removals not because foster care witness substance abuse, but because parents abuse and neglect their children when they abuse drugs
- The presence of mechanisms, it turns out, is valuable because of their policy relevance, but also because we can use them *sometimes* for identification

$$
\begin{tikzcd}
D \arrow[r] & M \arrow[r] & Y \\
& U \arrow[ul, dashed] \arrow[ur, dashed] &
\end{tikzcd}
$$

- D is confounded by U; therefore we cannot identify the causal effect of D on Y using the backdoor criterion bc $D \leftarrow U \rightarrow Y$ cannot be blocked
- Pearl (2009) showed that this DAG actually does allow us to recover the effect of D on Y, though – just not via the backdoor criterion
- We'll now look at a lesser known method of identification called the frontdoor criterion

Front door criterion

If one or more unblocked back door paths connect a causal variable to an outcome variable, the causal effect is identified by conditioning on a set of observed variables M that make up the identifying mechanism if and only if: 1) the variables in M intercept all directed paths from the causal variable to the outcome ("exhaustiveness"); 2) No unblocked back-door paths connecting the causal variable to the variables in the set M and all back door paths from the variables in M to the outcome can be blocked by conditioning on D ("isolation")

Exhaustiveness

- Exhaustiveness means the variables M are the only paths through which D impacts Y.
- In other words, rules out direct effects that bypass M altogether
- “only through M” in place of exhaustiveness and you get the idea

Isolation

- Mechanism itself is not confounded with respect to Y
- There does not exist some additional unobservable creating a back door path between M and Y
- It's a truly closed system, and as such, you're going to be making a strong argument so good luck

Three step method

1. Estimate the effect of D on M. Consider a regression of M on D or simple difference in mean D with respect to M

$$D = \alpha_0 + \beta M + \epsilon$$

   - → M is isolated, so it is not confounded
   - → $D \leftarrow U \rightarrow Y \leftarrow M$ which is blocked bc Y is a collider
   - → Therefore $\widehat{\beta}$ identifies β

2. Estimate the effect of M on Y conditional on X
   - → Gets you an unbiased estimate of M effect on Y bc only backdoor path from M to Y is $M \leftarrow D \leftarrow U \rightarrow Y$
   - → So long as we condition on D this path is blocked

$$Y = \alpha_1 + \gamma M + \psi D + \varepsilon$$

3. Multiply $\widehat{\gamma} \times \widehat{\beta}$ and you get the causal effect of D on Y

Examples have been elusive

- Pearl has suggested smoking as an example of this
- Smoking causes tar build-up, tar build-up causes lung cancer, smoking is endogenous to confounders
- Requires smoking to not have a direct effect on lung cancer, which is incorrect
- But a new paper by Bellemare, et al. (2021) provides a plausible example involving tipping and Uber

DGP

We will define a few error terms. Let U be our unobserved confounder variable drawn from the standard normal distribution. And let Z be drawn from the uniform distribution from 0 to 1. And let the following three error terms, ε_D, ε_M and ε_Y be normally distributed as well. Then let our variables of interest come from the following linear system:

$$\begin{aligned} D_i &= 0.5U_i + \varepsilon_D \\ M_i &= Z_i D_i + \varepsilon_M \\ Y_i &= 0.5M_i + 0.5U_i + \varepsilon_Y \end{aligned}$$

Code is available (I may run it for you now), but let's look at the results

Table: Simulation results Bellemare, et al. (2021)

Variables:	**Benchmark**	**Naive**	**Front Door**		**Direct effect**
	Y	**Y**	**M**	**Y**	**Y**
Treatment D	0.257***	0.451***	0.505***	0.198***	0.005
	(0.004)	(0.003)	(0.001)	(0.004)	(0.003)
Mechanism M				0.500***	0.500***
				(0.003)	(0.003)
Confounder U	0.492***				0.491***
	(0.004)				(0.004)
Estimated causal effect ($\widehat{\delta}$)	0.257***	0.451***		0.252***	–
	(0.004)	(0.003)		(0.002)	
N	100,000	100,000	100,000	100,000	100,000

Real world data

- Harrington (2019) nots shared rides typically result in lower tips for Uber drives: “on average, about 17% of rideshares end up with the driver getting tipped. For strips where a shared trip was authorized, that number is halved to a measly 8.6%.”
- Drivers experiencing such declines probably think it’s caused by sharing rides (e.g., bystander effects, freeriding, etc.)
- But maybe it’s selection – cheapskates share rides
- Let’s use the front door criterion to check

Assumed Uber Tipping DAG

- Let D here be authorizing a shared ride (regardless of whether a shared ride occurred), M be a dummy measuring one if sharing did occur, Y be the amount the passengers tipped and U be the unobserved covariates.
- Use the front door criterion, conditional on a series of geographic and time fixed effects using data on over 95 million Uber and Lyft rides in Chicago in 2019.
- Estimate the effect of authorization on both whether a passenger tips as well as how much, what they call the extensive and intensive margin of tipping, respectively.
- These data come from a data portal maintained by Chicago's Department of Business aAffairs and Consumer Protection's Transportation Network Providers and is freely available for download from the City of Chicago's website.

Assumptions

- It’s the same DAG as before so I won’t redraw it
- Key to this DAG is to consider that once the authorization to share a ride is initiated (the treatment), then when the ride is shared (the mechanism), the authors argue that their extensive set of fixed effects will yield plausible conditions for isolation and exhaustiveness are guaranteed.
- This means there is no direct effect of authorization on tipping, nor does there exist an unblocked backdoor path from sharing a ride and tipping itself.

Estimation

- Using the logic of the front door criterion, the authors estimate the same two step procedure as shown in the previous simulation with the caveat that they include extensive fixed effects so as to create conditional conditions for isolation and exhaustiveness.
- For illustrative purposes, I will only focus on the effect at the extensive margin (i.e., on whether a passenger tipped at all).

Table: Estimation results for tipping at the extensive margin from **?**

Variables:	Naive	Front Door	
	Tipped	**Shared Trip**	**Tipped**
Sharing authorized D	-0.0628***	0.6769***	-0.0550***
	(0.0001)	(0.0002)	(0.0002)
Shared trip M			-0.0115***
			(0.0002)
Full fare	0.0050***	-0.0064***	0.0049***
	(0.00001)	(0.00001)	(0.00003)
Estimated causal effect ($\widehat{\delta}$)	-0.0628***		-0.0078**
	(0.0001)		(0.0001)
N	95,670,449	95,670,449	95,670,449

Interpretation

- Column 1: naive regression simply compares tipping between authorized and non-authorized sharing (6.3pp reduction in tipping)
- Front door criterion: 1pp reduction
- Not surprising drivers don't want ride shares, but authors argue it's caused by selection (i.e., the people using ride shares) not ride share itself
- Unclear if you banned it whether it would increase driver earnings in other words
- Cool paper but it took me forever to download 95 million observations

Summarizing all of this

- Your dataset will not come with a codebook flagging some variables as “confounders” and other variables as “colliders” because those terms are always context specific
- Except for some unique situations that aren’t generally applicable, you also don’t always know statistically you have an omitted variable bias problem; but both of these are fatal for any application
- You only know to do what you’re doing based on *knowledge about data generating process*.
- All identification must be guided by theory, experience, observation, common sense and knowledge of institutions
- DAGs absorb that information and can be then used to write out the explicit identifying model

DAGs are not panacea

- DAGs cannot handle, though, reverse causality or simultaneity
- So there are limitations. “All models are wrong but some are useful”
- They are also not popular (see Twitter ongoing debates which have descended into light hearted jokes as well as aggressive debates)
- But I think they are helpful and while not *necessary*, showcase what is necessary – assumptions
- Heckman (1979) can maybe provide some justification at times