Causal Inference

Mixtape Session

Roadmap

- Introduction to course
- Foundations of causal inference
  - Princeton Industrial Relations Section
  - Design versus Model
  - Potential outcomes
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
- Directed Acyclic Graphs
  - Graph notation
  - Backdoor criterion
  - Collider bias
  - Front door criterion
  - Concluding remarks

Welcome to Mixtape Sessions!

- 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
- Mixtape Sessions is an educational platform designed to "democratize causal inference" at all levels by helping bridging people with teachers

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

Class goals

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. IV: Day 2
3. RDD: Day 3
4. DiD: Day 4
5. Synthetic control and remaining: Day 5

Different types of prediction

Prediction machines

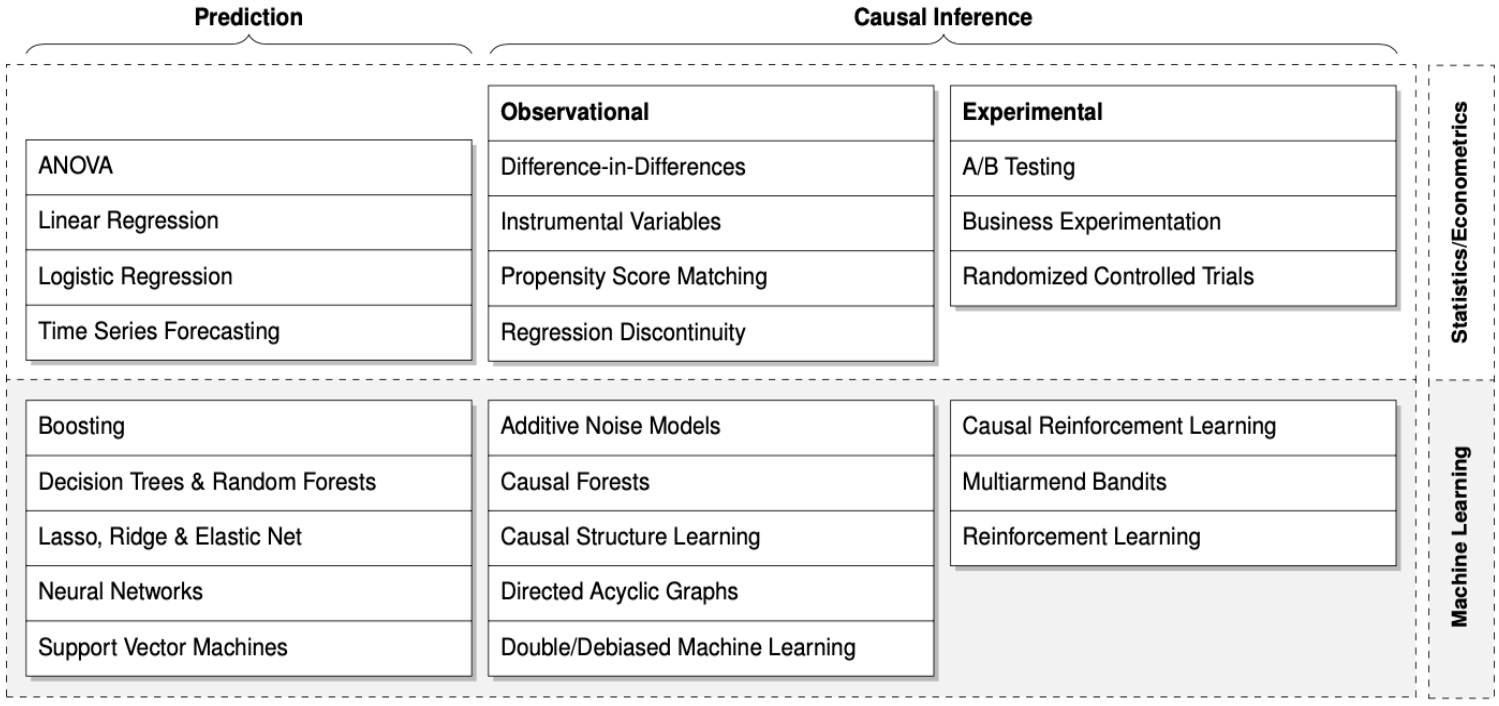
- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- “Does this person have heart disease?”, “How many books will I sell?”
- It is not predictions of what effect a choice will have, though

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.



Roadmap

- Introduction to course
- Foundations of causal inference
  - Princeton Industrial Relations Section
  - Design versus Model
  - Potential outcomes
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
- Directed Acyclic Graphs
  - Graph notation
  - Backdoor criterion
  - Collider bias
  - Front door criterion
  - Concluding remarks

Princeton Industrial Relations Section

- October 2021's Nobel Prize in economics went to Card, Angrist and Imbens
- Princeton's mid 1980s Industrial Relations Section: Orley Ashenfelter, who advises Card, Angrist, LaLonde, hires Krueger, etc.

Princeton and credibility

- Panel models were not satisfactory for recovering causal effects in Ashenfelter dip situations
- Princeton IRS economists focus less on modeling the outcome and more on the treatment variation
- LaLonde JMP (AER 1986) as well as Ashenfelter and Card (RESTAT 1985) cast doubt on econometric evaluators, increase demand for explicit randomization

Angrist and Imbens and the 1990s

- Angrist writes a dissertation using randomized instruments (Vietnam draft), goes to Harvard, overlaps with Imbens for a year, they are mentored by Gary Chamberlain, work with Don Rubin, write their famous LATE paper
- Chamberlain recommends potential outcomes framework over a different one that had been used at that time (latent index) and that seems to make the work more generally attractive (like to Rubin)
- Let’s spend twenty minutes listening to them

Angrist, Imbens and Harvard

Josh Angrist on the negative results at the time (10 min)

https://youtu.be/ApNtXe-JDfA?t=1885

Guido Imbens on the reception of their work (10 min)

https://youtu.be/cm8V65AS5iU?t=799

Causality and the model

Empirical labor economics and the economic model (Card speech 2014)

- **Model**: Causality exists within the framework of a theory that says “D causes Y” (e.g., Heckman)
- **Design**: Causality is design-based and can be discerned with *physical* manipulation of a treatment D (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

Empirical labor seems to shift towards more of the “working model” approach with exceptions and this is largely connected with an approach that is somewhat agnostic about the underlying model

Topics broaden

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's distinctiveness

- Main idea: focus is on the assignment of units to treatment (e.g., physical randomization, instruments, running variables, propensity score)
- Favors deep institutional knowledge and domain knowledge (as opposed to black box style modeling)
- Can help answer causal questions when RCTs are difficult to implement (e.g., too expensive, not realistic, not ethical)
- Potential outcomes model is really core in all of this

Introduction to Counterfactuals

- Aliens come and orbit earth, see people dying in hospitals and conclude “doctors are hurting people”
- They kill the doctors, unplug patients from machines, throw open the doors – many patients inexplicably die
- *We are the aliens in our research*

#1: Correlation and causality are different

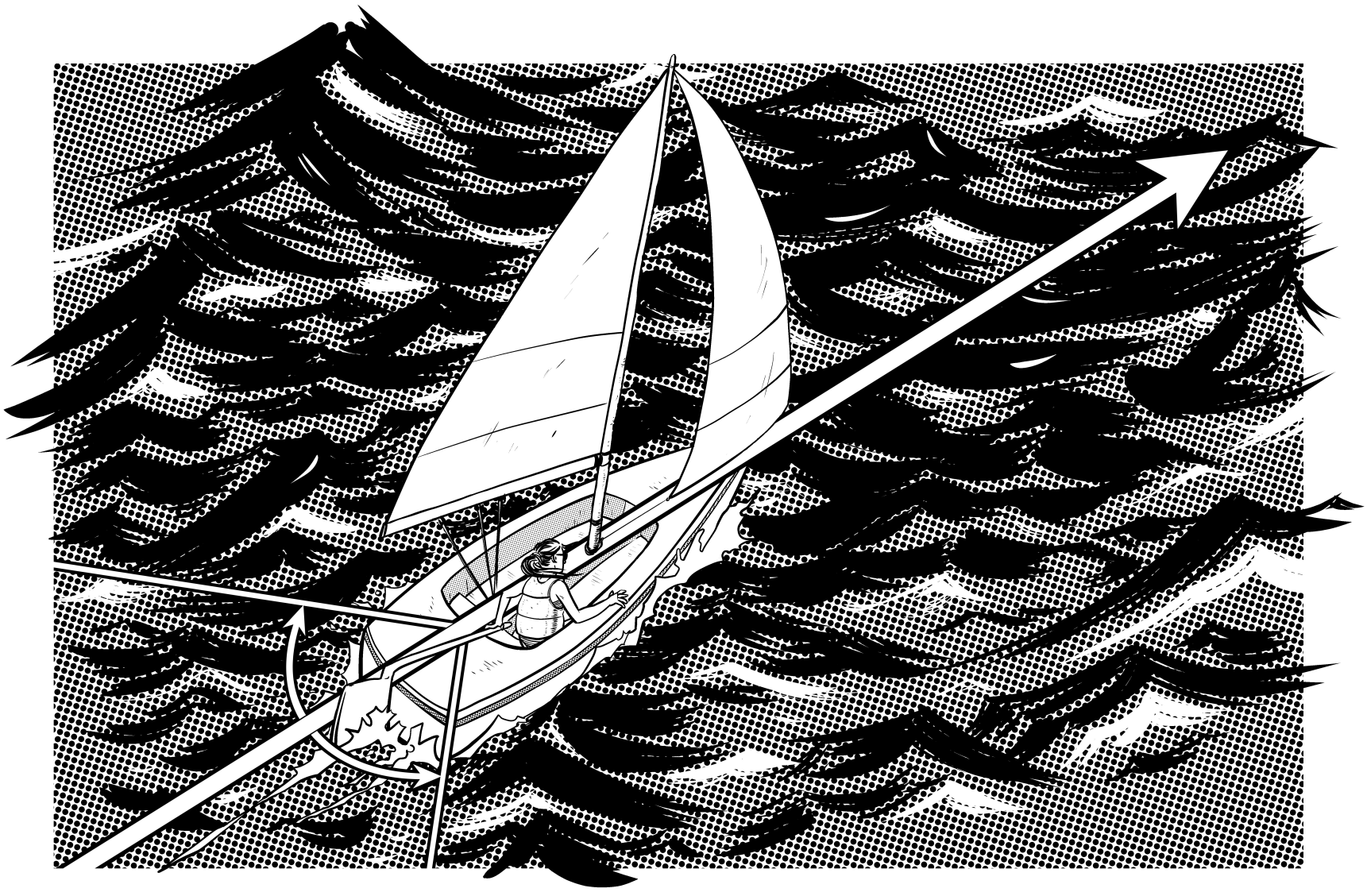
Causal is one unit, correlation is many units

- 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

History of potential outcomes

- The conceptual framework for design based causal inference is the *potential outcomes* model which roots causality in counterfactual reasoning
- Jerzy Neyman (1923) and Ronald Fisher (1925) linked causal inference to randomized physical experiments
- It continues into the present through Donald Rubin in the 1970s and 1980s with Paul Rosenbaum on the propensity score
- Continues through Rubin's collaborations with Josh Angrist and Guido Imbens (this year's Nobel Prize winners) on instrumental variables

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:

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

  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 observed outcome Y either conceptually or notationally
- Potential outcomes are hypothetical states of the world but observed outcomes are ex post realizations

Important definitions

Definition 1: Individual treatment effect

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

Definition 3: 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*.

Definition 2: Switching equation

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

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}$$

Missing data problem

- Causal inference is fundamentally a missing data problem requiring prediction, not of the present or the future, but of a missing past – sometimes explicitly (nearest neighbor, synthetic control), sometimes implicitly (RDD, IV)
- Aggregate parameters based on individual treatment effects are descriptions of causal effects
- Fundamental problem of causal inference holds because of the switching equation *even with big data*

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

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

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

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

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:

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

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

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:

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

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

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

- Epistemology: what beliefs are warranted and what beliefs are not
- Without counterfactuals, we do not *know* treatment effects, but with groups of data we can sometimes obtain *estimates*
- We do this by making comparisons of groups treated and not treated
- But not all comparisons are equal – selection bias (e.g., aliens making unwarranted conclusions about causality because of failing to use design)
- We will decompose a simple estimator so we can see what *selection bias* is

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:

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

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

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

Decomposition of difference in means

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

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

Selection bias

- Notice this term “selection bias”

$$E[Y^0|D=1] \neq E[Y^0|D=0]$$

- Selection bias was the problem that the aliens failed to overcome – without treatment Y^0, COVID patients on vents $(D=1)$ would likely have been different from those with COVID not on vents $(D=0)$

What is selection bias?

Let's put this into words.

1. Put $E[Y^0|D=1] \neq E[Y^0|D=0]$ to your best friend who hasn't taken this course
2. If people choose a treatment, $D=1$, or control, $D=0$, because they expect it benefits them, $Y^1 - Y^0 > 0$, or doesn't $Y^1 - Y^0 \leq 0$ then what do you suspect is true about the mean value of Y^0 for the treatment and control groups?

Perfect doctor exercise (and breakout)

- Chronic PTSD has historically been treated with cognitive behavior therapies like mindfulness, but recent work shows therapist assisted MDMA (street name: ecstasy), are effective too
- The Perfect Doctor can perfectly assess whether mindfulness practices or MDMA is more beneficial for treating a patient's chronic PTSD ($Y^1 - Y^0$ is positive or negative), and makes treatment assignments ($D = 1$ or 0) depending on its impact
- We will go through an exercise together (go along on your end with your google sheet) analyzing the implications of the perfect doctor's choices on a range of statistics, followed by a breakout session where you do it alone

Goal of causal inference

Our goal in all of causal inference is to estimate aggregate causal parameters by modeling treatment assignment

We seek to *impute* (either explicitly or implicitly) missing counterfactuals using various techniques such that selection bias is eliminated

Let's look what happens in an A/B test *and why* this addresses the term $E[Y^0|D=1]$ and $E[Y^0|D=0]$

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 $1-\pi$:

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

- 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”
  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*
- Largely about spillovers, poorly defined treatments and scale

SUTVA: No spillovers to other units

- What if we impose a treatment at one neighborhood but not a contiguous one?
- Treatment may spill over causing $Y = Y^1$ even for the control units because of spillovers from treatment group
- Informs the design stage

SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable” (i.e., notice that the super scripts contain either 0 or 1, not 0.55, 0.27)
- If we are estimating the effect of aspirin on headaches, we assume treatment is 200mg per person in the treatment
- 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*; you may have to think of it as multiple arms

SUTVA: Scale can affect stability of treatment effects

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)

Econometrica, Vol. 83, No. 1 (January, 2015), 155–174

CONSUMER HETEROGENEITY AND PAID SEARCH EFFECTIVENESS: A LARGE-SCALE FIELD EXPERIMENT

BY THOMAS BLAKE, CHRIS NOSKO, AND STEVEN TADELIS[1]

Internet advertising has been the fastest growing advertising channel in recent years, with paid search ads comprising the bulk of this revenue. We present results from a series of large-scale field experiments done at eBay that were designed to measure the causal effectiveness of paid search ads. Because search clicks and purchase intent are correlated, we show that returns from paid search are a fraction of non-experimental estimates. As an extreme case, we show that brand keyword ads have no measurable short-term benefits. For non-brand keywords, we find that new and infrequent users are positively influenced by ads but that more frequent users whose purchasing behavior is not influenced by ads account for most of the advertising expenses, resulting in average returns that are negative.

KEYWORDS: Advertising, field experiments, causal inference, electronic commerce, return on investment, information.

1. INTRODUCTION

ADVERTISING EXPENSES ACCOUNT for a sizable portion of costs for many companies across the globe. In recent years, the Internet advertising industry has grown disproportionately, with revenues in the United States alone totaling $36.6 billion for 2012, up 15.2 percent from 2011. Of the different forms of Internet advertising, paid search advertising, also known in industry as "search engine marketing" (SEM), remains the largest advertising format by revenue, accounting for 46.3 percent of 2012 revenues, or $16.9 billion, up 14.5 percent from $14.8 billion in 2010. Google Inc., the leading SEM provider, registered $46 billion in global revenues in 2012, of which $43.7 billion, or 95 percent, were attributed to advertising.[2]

Internet advertising facts

- In 2012, revenues from Internet advertising was $36.6 billion and has only grown since
- Paid search (“search engine marketing”) is the largest format by revenue (46.3% of 2012 revenues, or $16.9 billion)
- Google is leading provider (registered $46 billion in global revenues in 2012 of which 95% was attributed to advertising)

Selection bias

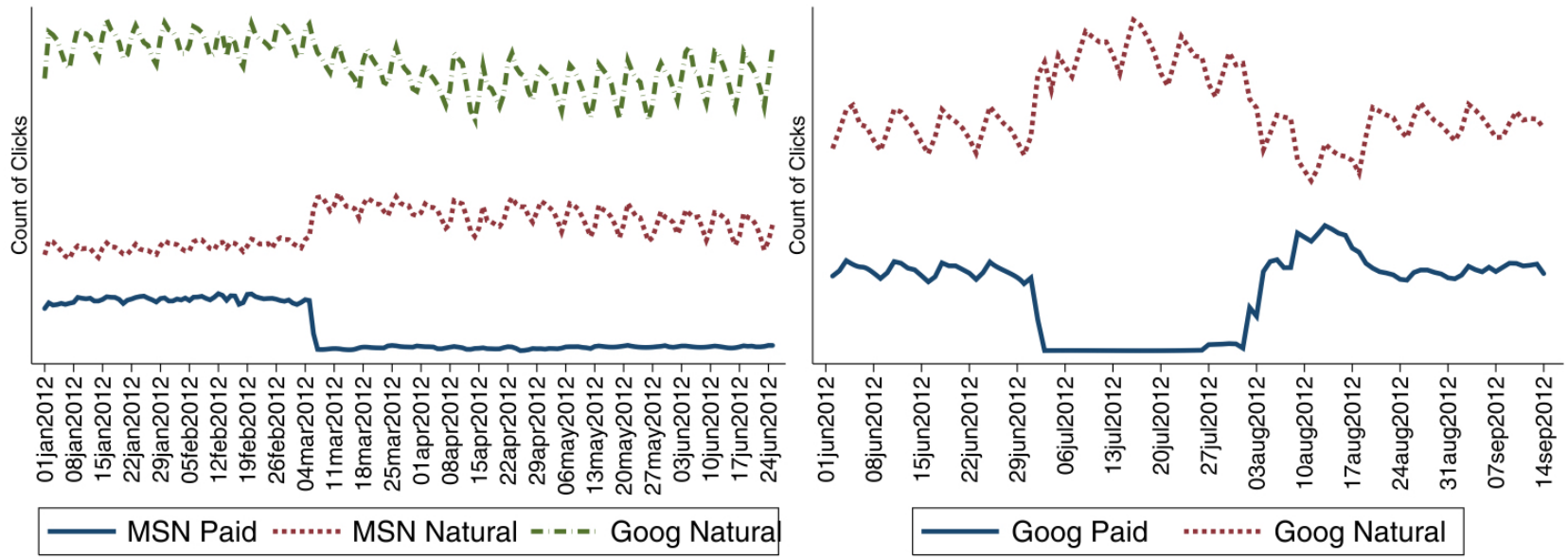
- Treatment was targeted ads at particular people conducting particular types of keyword search
- Consumers who choose to click on ads are loyal and already informed about products with high likelihood to buy already
- Problem is ads are targeting people at the end of their search, so the question is whether they would've found it already (i.e., $E[Y^0|D=1] \neq E[Y^0|D=0]$)

Selection bias

- Estimated return on investment using OLS found ROI of over 1600%
- Compared this to experimental methods and found ROI of -63% with a 95% CI of $[-124\%, -3\%]$, rejecting the hypothesis that the channel yielded short-run positive returns
- Think back to perfect doctor – Even without the treatment (Y^0), the treated group observationally would've still found a way

Natural experiment

- Study began with a naturally occurring and somewhat fortuitous event at eBay
- eBay halted SEM queries for brand words (i.e., queries that included the term eBay) on Yahoo! and Microsoft but continued to pay for these terms on Google
- Blake, Nosky and Tadelis (2015) showed almost all of the foregone click traffic and attributed sales were captured by natural search
- Substitution between paid and unpaid traffic was nearly one to one complete



(a) MSN Test

(b) Google Test

FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for 'ebay' terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

Interpretation of natural experiment

"The evidence strongly supports the intuitive notion that for brand keywords, natural search is close to a perfect substitute for paid search, making brand keyword SEM ineffective for short-term sales. After all, the users who type the brand keyword in the search query intend to reach the company's website, and most likely will execute on their intent regardless of the appearance of a paid search ad."

Selection bias

Observational data masked causal effect (recall the decomposition of the any non-designed estimation strategy)

> *“Advertising may appear to attract these consumers, when in reality they would have found other channels to visit the company’s web-site. We overcome this endogeneity challenge with our controlled experiments.”*

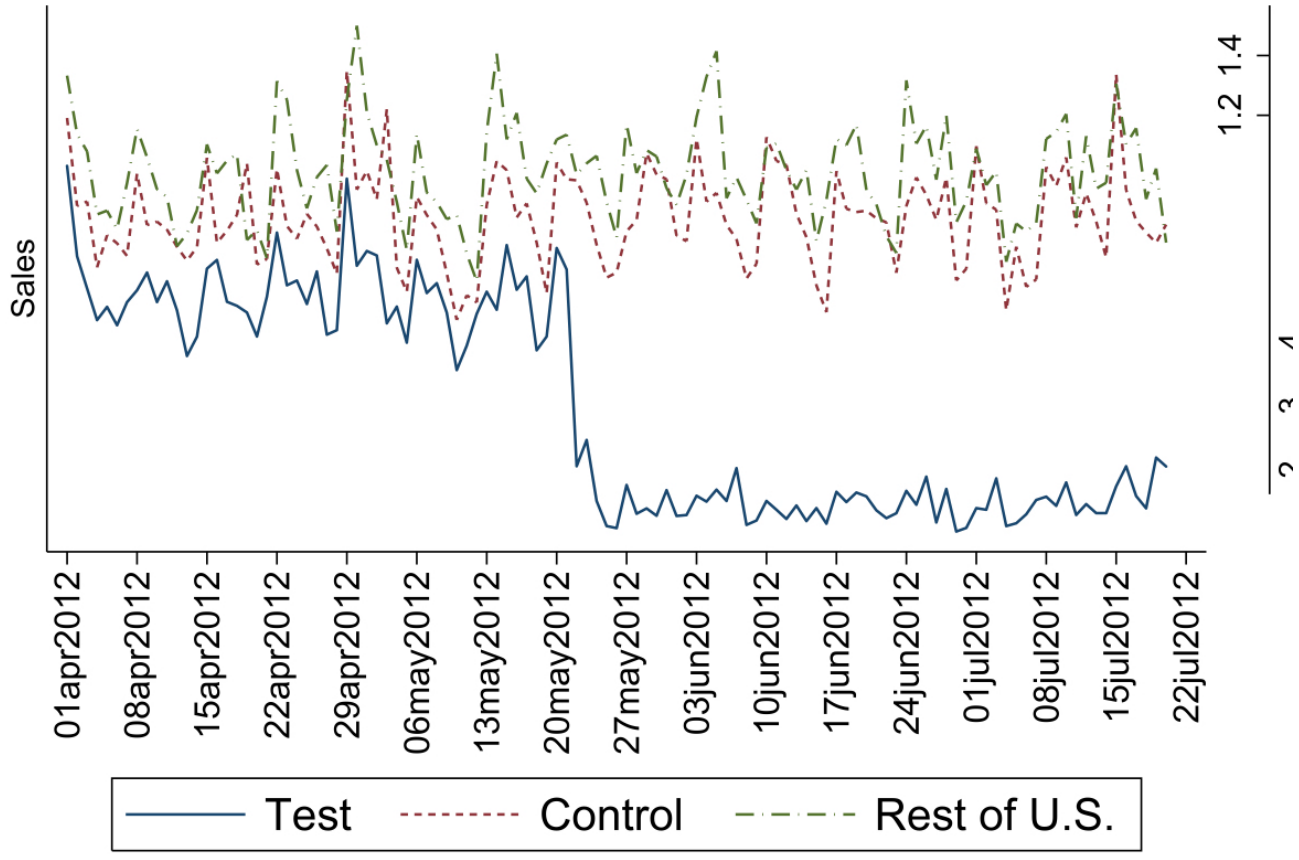
RCT

Natural experiment was valuable, but eBay could run a large scale RCT.

Use this finding of a nearly one-to-one substitution once paid search was dropped to convince eBay to field a large scale RCT discontinuing non-band key words

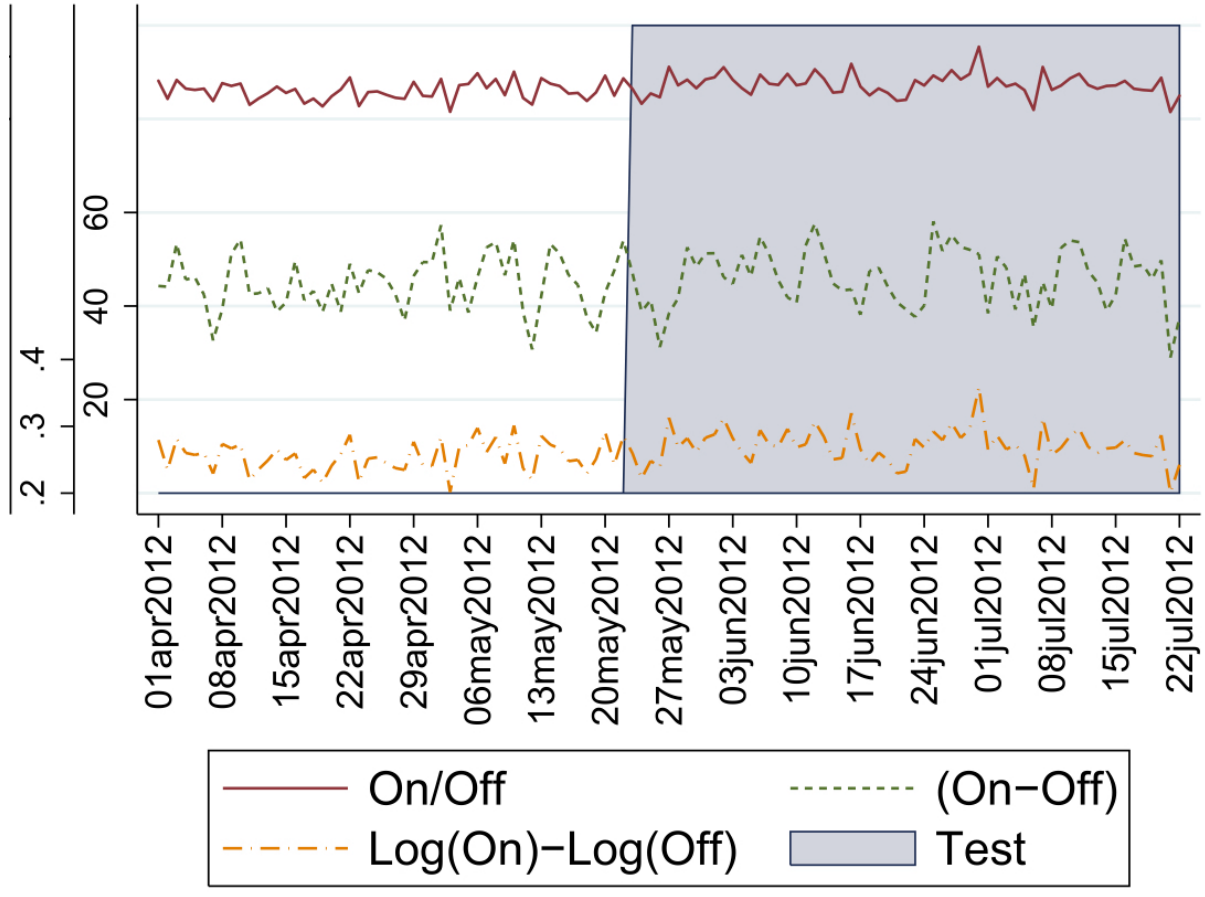
Design of the experiment

- Randomly assigned 30 percent of eBay's US traffic to stop all bidding for all non-brand keywords for 60 days
- Some random group of users, in other words, were exposed to ads; a control group did not see the ads
- Used Google's geographic bid feature that can accurately identify geographic market of the user conducting the search
- Ads were suspended in 30 percent of markets to reduce the scope of the test and minimize the potential cost and impact to the business



(a) Attributed Sales by Region

Figure: Attributed sales due to clicking on a Google link (treatment group)



(b) Differences in Total Sales

Figure: Differences in total sales by market (treatment to control)

	OLS	
	(1)	(2)
Estimated Coefficient	0.88500	0.12600
(Std Err)	(0.0143)	(0.0404)
DMA Fixed Effects		Yes
Date Fixed Effects		Yes
N	10,500	10,500
$\Delta \ln(Spend)$ Adjustment	3.51	3.51
$\Delta \ln(Rev)$ (β)	3.10635	0.44226
Spend (Millions of $)	$51.00	$51.00
Gross Revenue (R′)	2,880.64	2,880.64
ROI	4,173%	1,632%
ROI Lower Bound	4,139%	697%
ROI Upper Bound	4,205%	2,265%

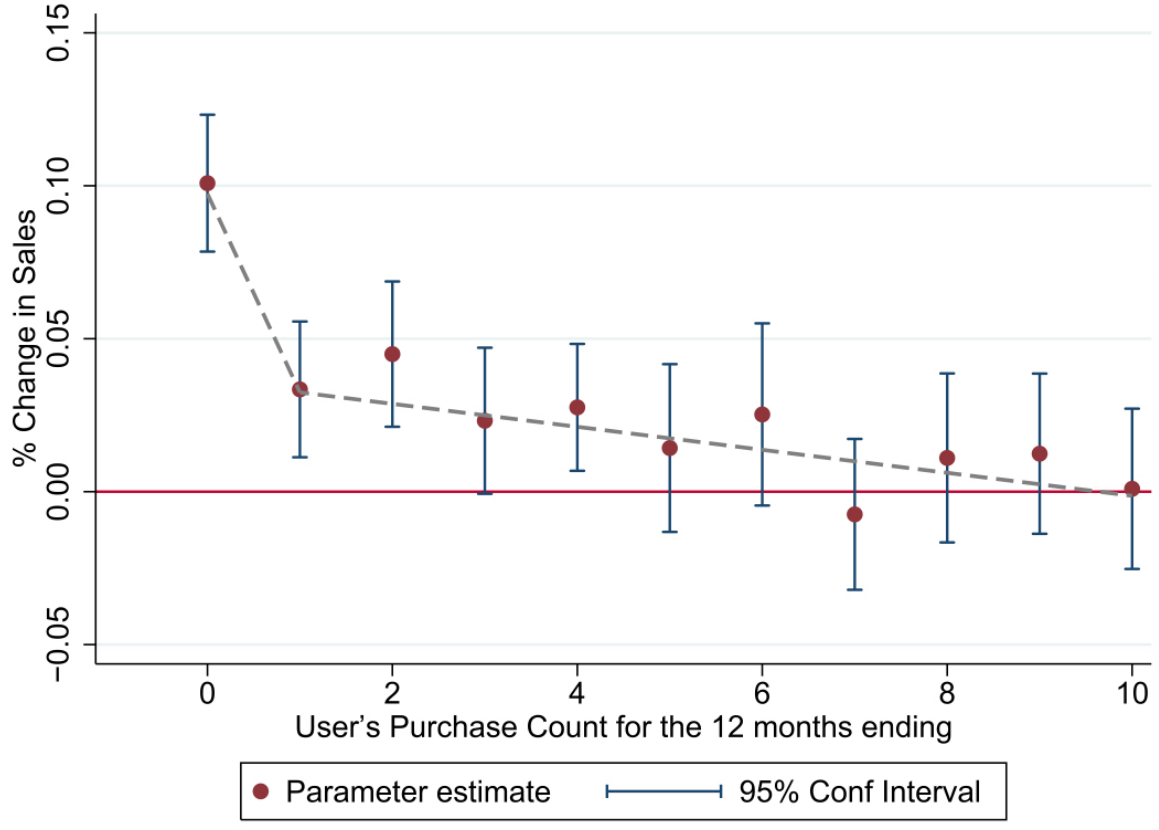
Figure: Spending effect on revenue using OLS but not the randomization. Effects are gigantic.

	(5)
Estimated Coefficient	0.00659
(Std Err)	(0.0056)
DMA Fixed Effects	Yes
Date Fixed Effects	Yes
N	23,730
$\Delta \ln(Spend)$ Adjustment	1
$\Delta \ln(Rev)$ (β)	0.00659
Spend (Millions of $)	$51.00
Gross Revenue (R′)	2,880.64
ROI	−63%
ROI Lower Bound	−124%
ROI Upper Bound	−3%

Figure: Spending effect on revenue using the randomization. Effects are negative.

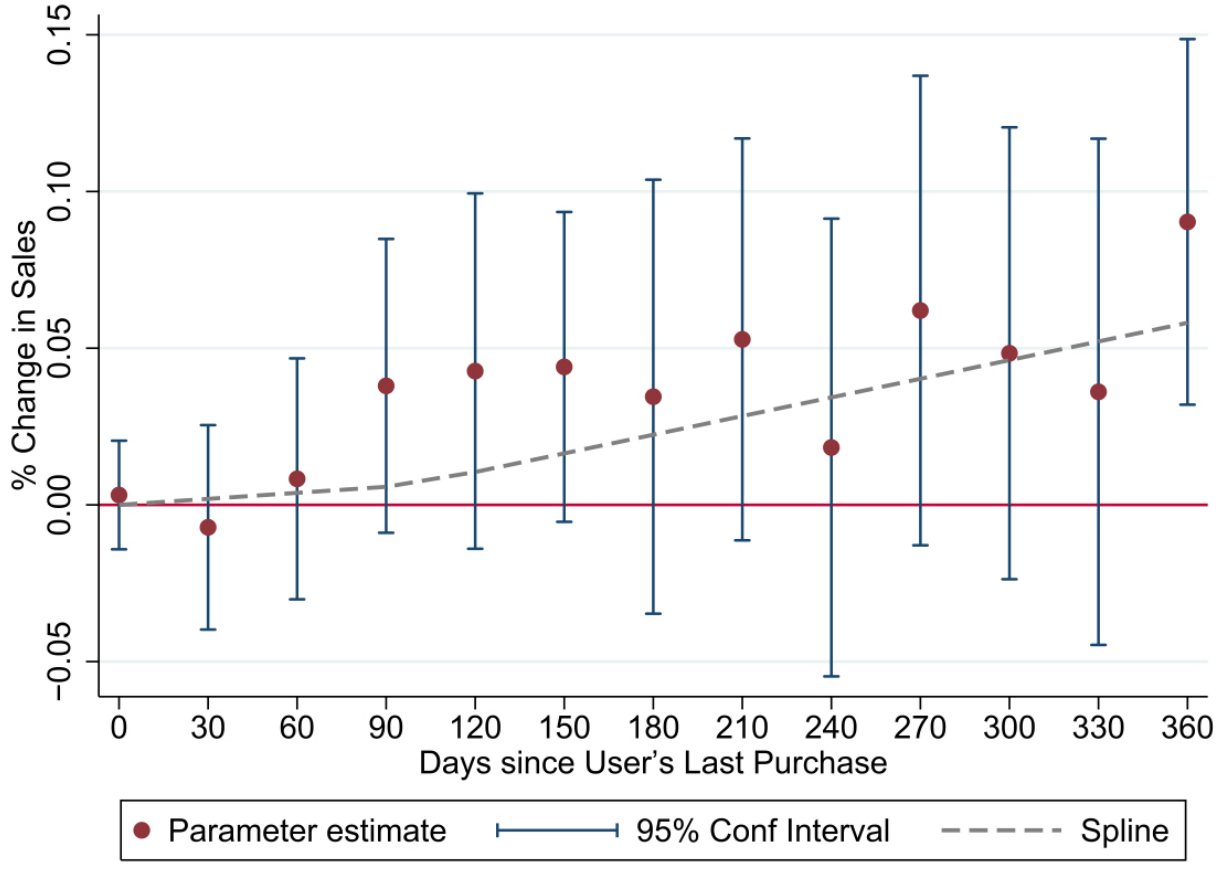
Heterogenous treatment effects

- Recall how the potential outcomes model explicitly models individual treatment effects could be unique and that the perfect doctor showed selection on gains masked treatment effects, perhaps even reversing sign
- Search advertising in this RCT only worked if the consumer had no idea that the company had the desired product
- Large firms like eBay with powerful brands will see little benefit from paid search advertising because most consumers already know that they exist, as well as what they have to offer



(a) User Frequency

Figure: Effects on new users are positive and large, but not others.



(b) User Recency

Figure: Effects are largest for "least active" customers.

Why are causal effects small?

- They suggest that the brand query tests found small causal returns because users simply substituted from the paid search clicks to the natural search clicks
- If that's the case, then it's explicitly a selection bias story

$$E[Y^0|D=1] \neq E[Y^0|D=0]$$

where D is being shown the branded advertisement based on search (i.e., they were already going there)

- They weren't using branded search for information; they were using to *navigate*

Self selection based on gains

- Potential outcomes is the foundation of the physical experiment because the physical experiment assigns units to treatments *independent* of potential outcomes, Y^0, Y^1
- This is important because outside of the physical experiment, we expect people select those important treatments based on whether, subjectively, they think $Y^1 > Y^0$ or $Y^1 \leq Y^0$.
- Rational actors almost by definition are thought to “self-select into treatment” making non-designed comparisons potentially misleading – sometimes by a little, sometimes by a lot

Natural experiments

- Recall how Blake, et al. (2015) used that natural even in which eBay stopped paid search on two search engines but not Google?
- While the phrase “natural experiment” can be a misnomer, as these are not real experiments, it is nonetheless a helpful hook to hang your hat on as we go forward
- Natural experiments are important if the RCT we want to run may not be one we can run for any number of reasons
- Our course runs down alternative strategies, but each of them must replace independence with something else
- Separate in your minds identification from estimation – it’ll help as we progress

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

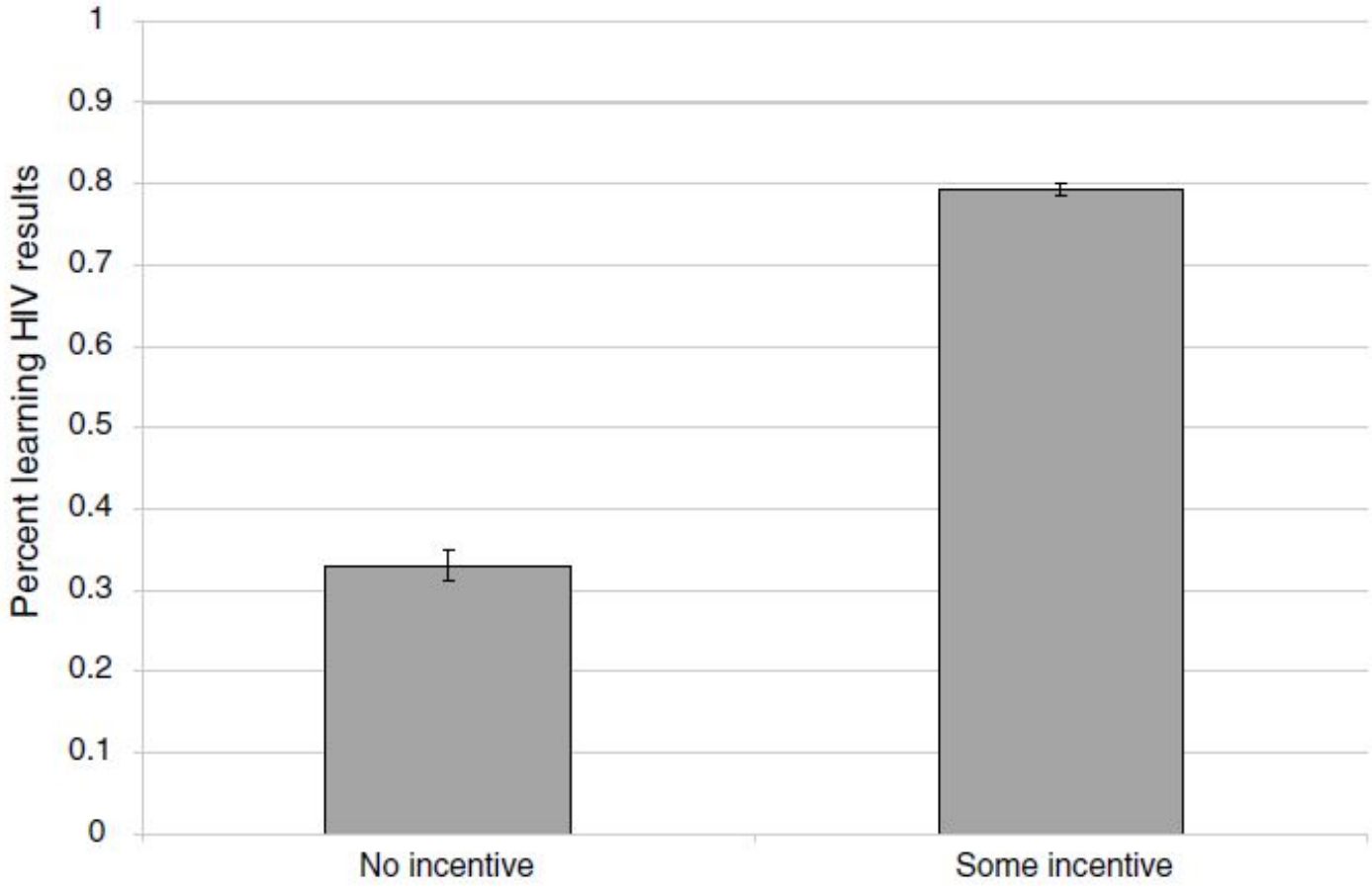


Figure: Visual representation of cash transfers on learning HIV test results.

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

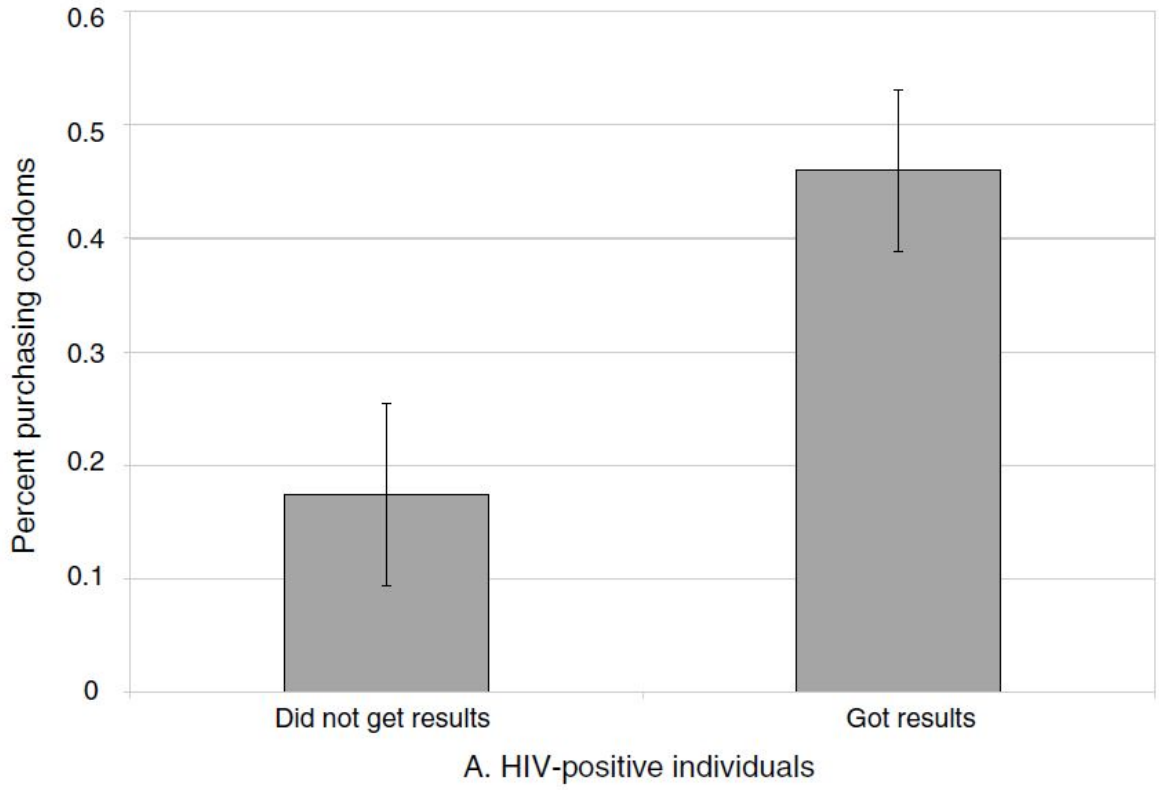


Figure: Visual representation of cash transfers on condom purchases for HIV positive individuals.

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 × 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 Section
  - Design versus Model
  - Potential outcomes
  - Selection bias
  - Independence
  - Example of physical experimentation: eBay advertising
- Directed Acyclic Graphs
  - Graph notation
  - Backdoor criterion
  - Collider bias
  - Front door criterion
  - Concluding remarks

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 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

Further reading

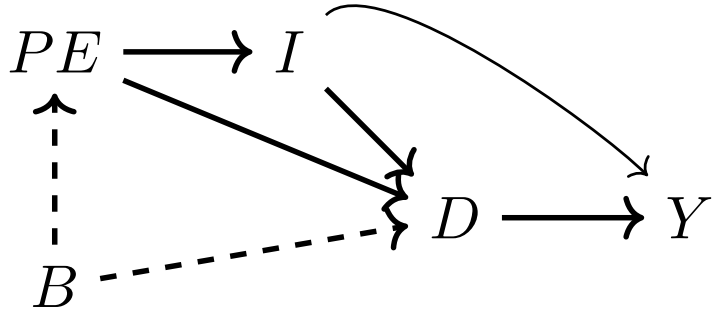
1. Pearl (2018) The Book of Why: The New Science of Cause and Effect, Basic Books (*popular*)
2. Morgan and Winship (2014) Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press, 2nd edition (*excellent*)
3. Pearl, Glymour and Jewell (2016) Causal 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

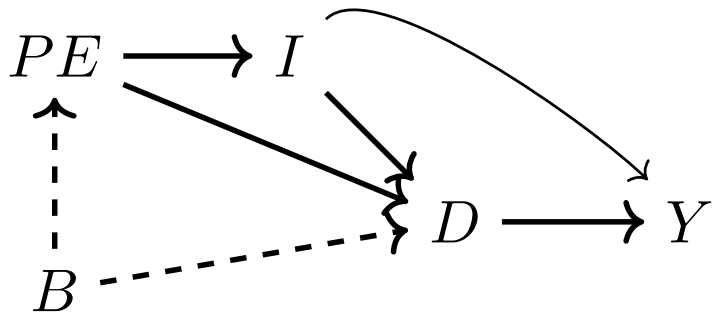
Causal model

- 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



- 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

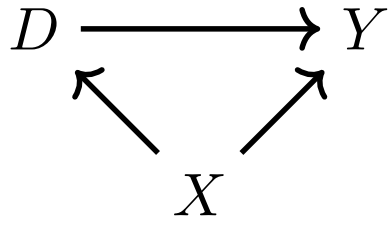
1. Facilitates the task of designing identification strategy for estimating average causal effects
2. Facilitates the task of testing compatibility of the model with your data
3. Visualizes the identifying assumptions which opens up the model to critical scrutiny

Creating DAGs

- The DAG is a *relevant* causal relationships describing the relationship between D and Y
- It will include:
  - → All direct causal effects among the *relevant* variables in the graph
  - → 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

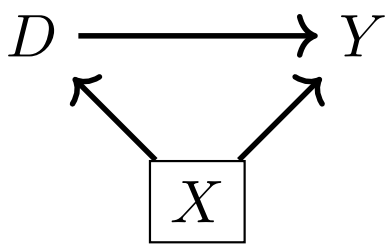
- Omitted variable bias has a name in DAGs: “confounding”
- Confounding occurs when when the treatment and the outcomes have a common cause or parent which creates spurious correlation between D and Y



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

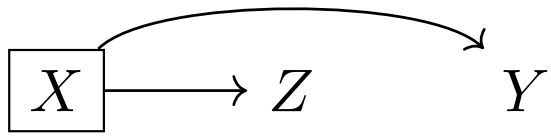
A backdoor path is blocked if and only if:

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

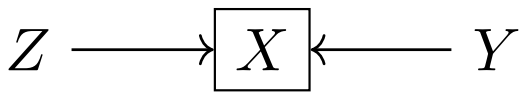
Examples of blocked paths

Examples:

1. Conditioning on a noncollider blocks a path:



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



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

Backdoor criterion

Backdoor criterion

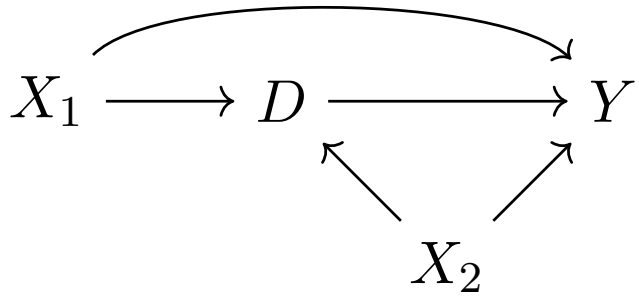
Conditioning on X satisfies the backdoor criterion with respect to (D, Y) directed path if:

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?

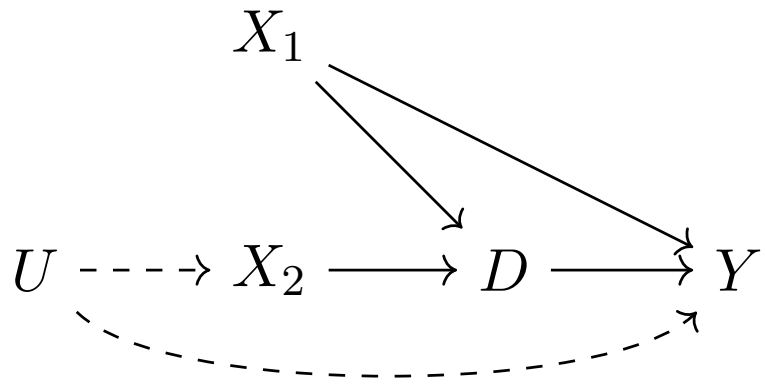
- List all backdoor paths from D to Y. I'll wait.



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

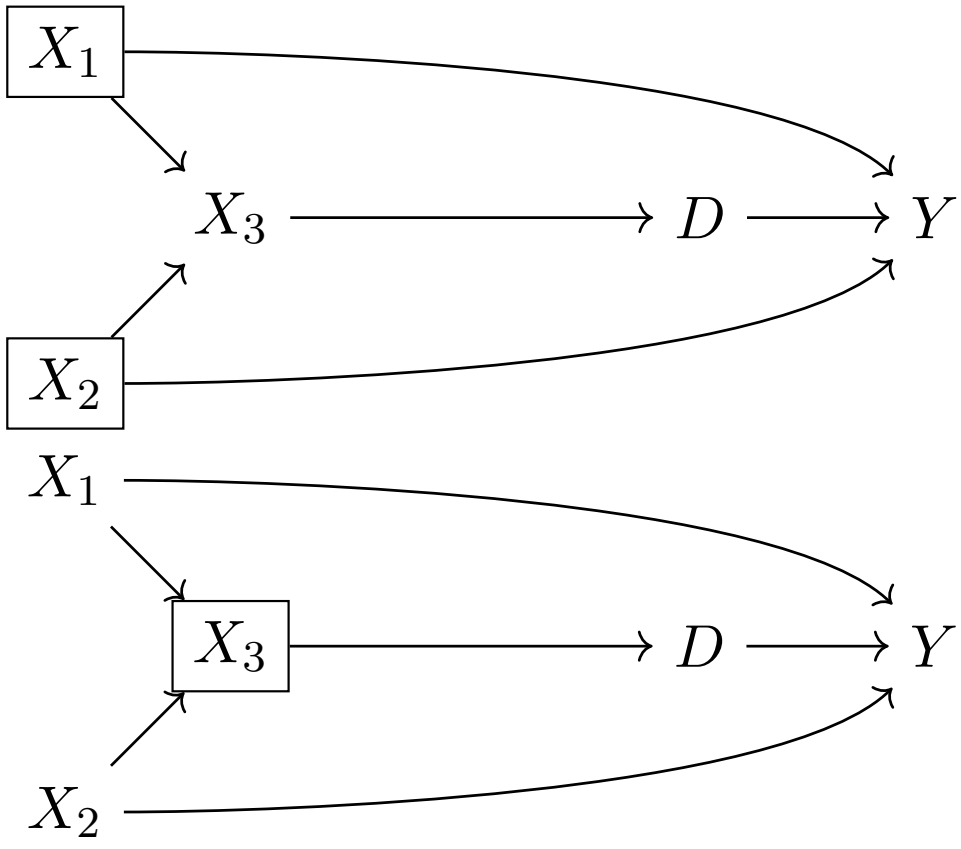
What if you have an unobservable?

- List all the backdoor paths from D to Y.



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

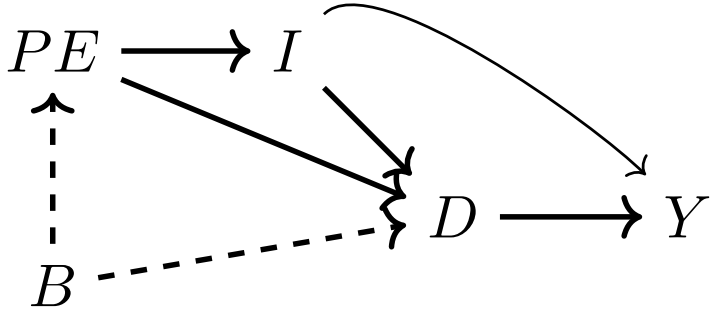
Multiple strategies



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

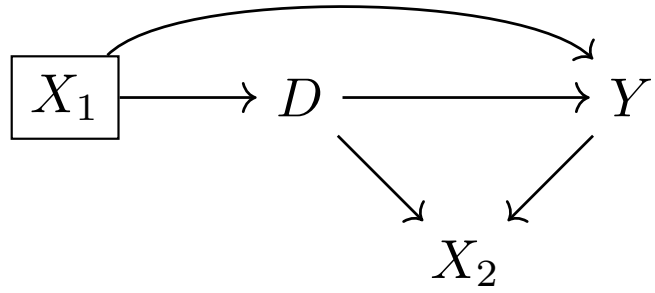
Testing the Validity of the DAG

- 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

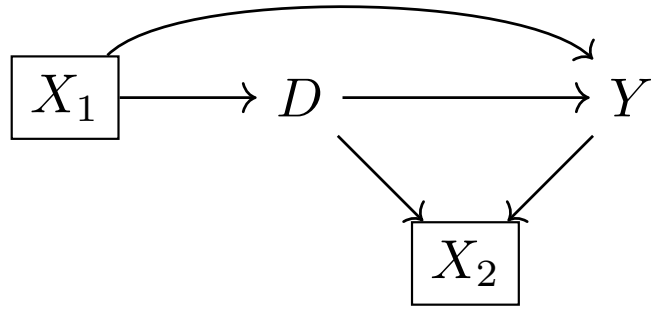


Collider bias

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

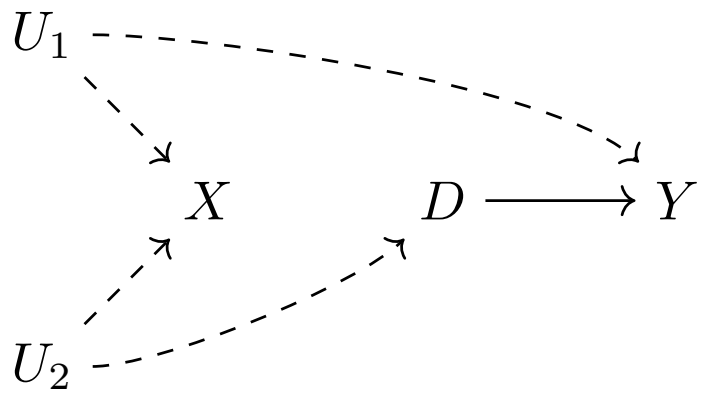


  - → Conditioning on X_1 blocks the backdoor path
  - → But what if we also condition on X_2?

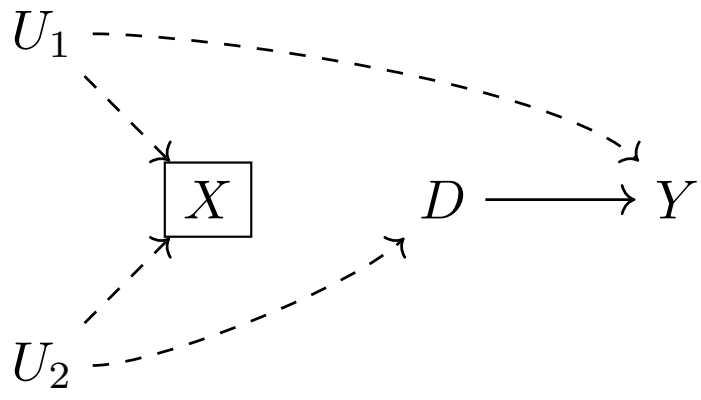


  - → Conditioning on X_2 opens up a new path, creating new spurious correlations between D and Y

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



  - → But what if we condition on X?



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?
- Look at python code

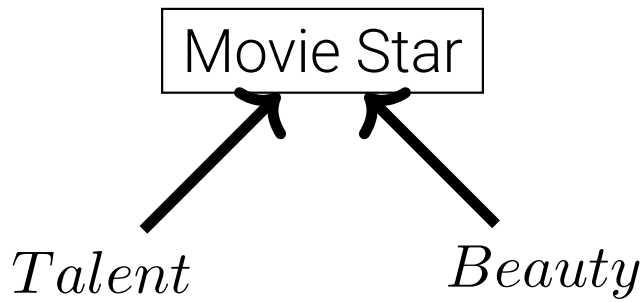
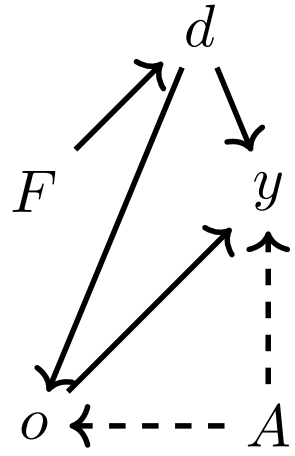


Figure: Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent.

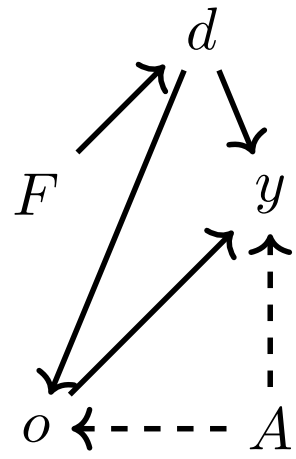
Occupational sorting and discrimination example of collider bias

- 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. $d \rightarrow o \leftarrow A \rightarrow y$

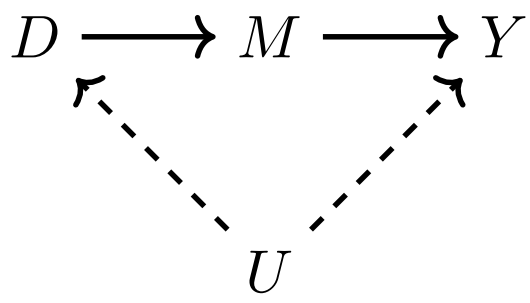
Table: Regressions illustrating collider bias with simulated gender disparity

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

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



- 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

Uber and tipping

- 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

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

Breakout sessions

- DAG survey response bias example
- DAG front door criterion example
- DAG backdoor criterion example

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 everywhere 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