Causal Inference I

Mixtape Session

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

Welcome!

- I'm Scott Cunningham, Ben H. Williams professor of economics at Baylor University, author of Causal Inference: the Mixtape
- I enjoy learning about economics, econometrics and particularly causal inference and run workshops on causal inference in order to help people make the material more accessible
- I work on a variety of eclectic topics – sex work, methamphetamine drug policy, abortion policy, and more recently, detecting and treating self harm in prisons and jails
- Causal inference is an old field but which has increasingly drawn people to it (Nobel Prize two years ago maybe helped)

Expectations

- This workshop will cover potential outcomes, randomization, unconfoundedness, instrumental variables and regression discontinuity design
  - → Causal Inference 2 will cover diff-in-diff
  - → Causal Inference 3 will cover synthetic control
- My teaching style is to try and not leave any one behind so I emphasize a lot of basics, applications, simple things repeatedly and then extend it in more advanced material

Class goals

1. **Curiosity**: You will be intrigued and interested enough in this material that you want to learn more of it by the end
2. **Confidence**: You will feel like you have a good understanding of causal inference so that by the end it doesn't feel all that mysterious or intimidating
3. **Comprehension**: You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation
4. **Competency**: You will have more knowledge of programming syntax in Stata and R (and python!) so that later you can apply this in your own work

Github repo

- We will communicate with one another regularly in the Discord channel and I will always be monitoring it
- Encourage you to talk to each other there, help one another, network with one another, coauthor with one another!
- I will be distributing things to you, like code and slides, via a github repo at my platform Mixtape Sessions
- Each lecture will be recorded and then uploaded to Vimeo, and we will be experimenting with various AI tools as well

What is Causal Inference?

- Causality as a subject when studied by philosophers is divided between metaphysics (a theory of what is) and epistemology (what is knowledge)
- Causal inference will cover both of these – how will we define causal effects? What rules will we use to make a deduction that a causal *belief* (inference) is warranted?
- It's a rich field, like a mighty river, with many rivers and streams that feed into it and I will try my best to point to them along the way
- But the majority of this comes from when two approaches formed a new paradigm which is largely what modern causal inference means

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

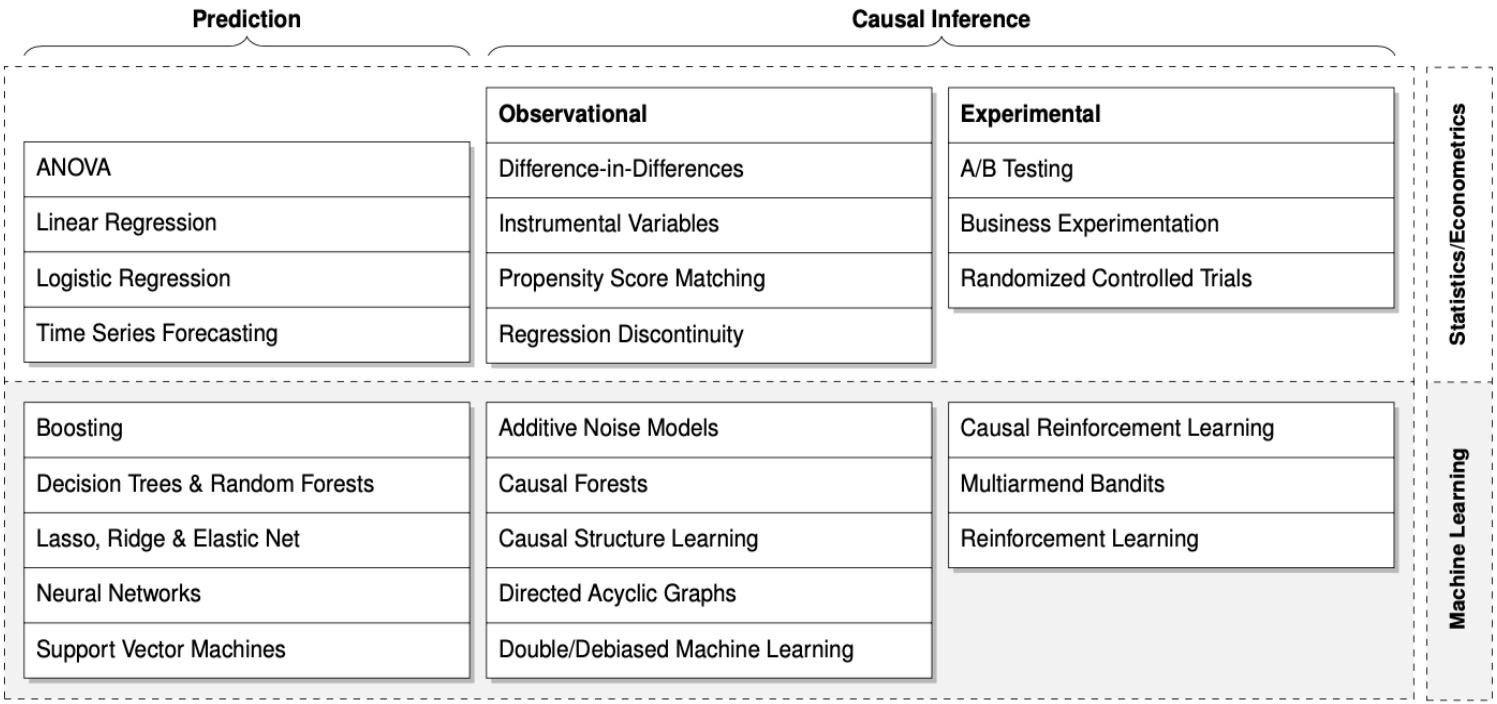
Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

Causal Inference vs Prediction

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 vs Prediction

Traditional prediction

- 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

Naive causal inference

- Aliens estimate a model showing a systematic correlation between COVID deaths and ventilators
- They conclude doctors are killing patients with ventilators so they come to earth to liberate the patients, but it only makes things worse
- Their error was they confused correlation with causality, but deeper than that, they didn't understand how the world worked
- *We are the aliens in our research*

#1: Correlation and causality are different concepts

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:

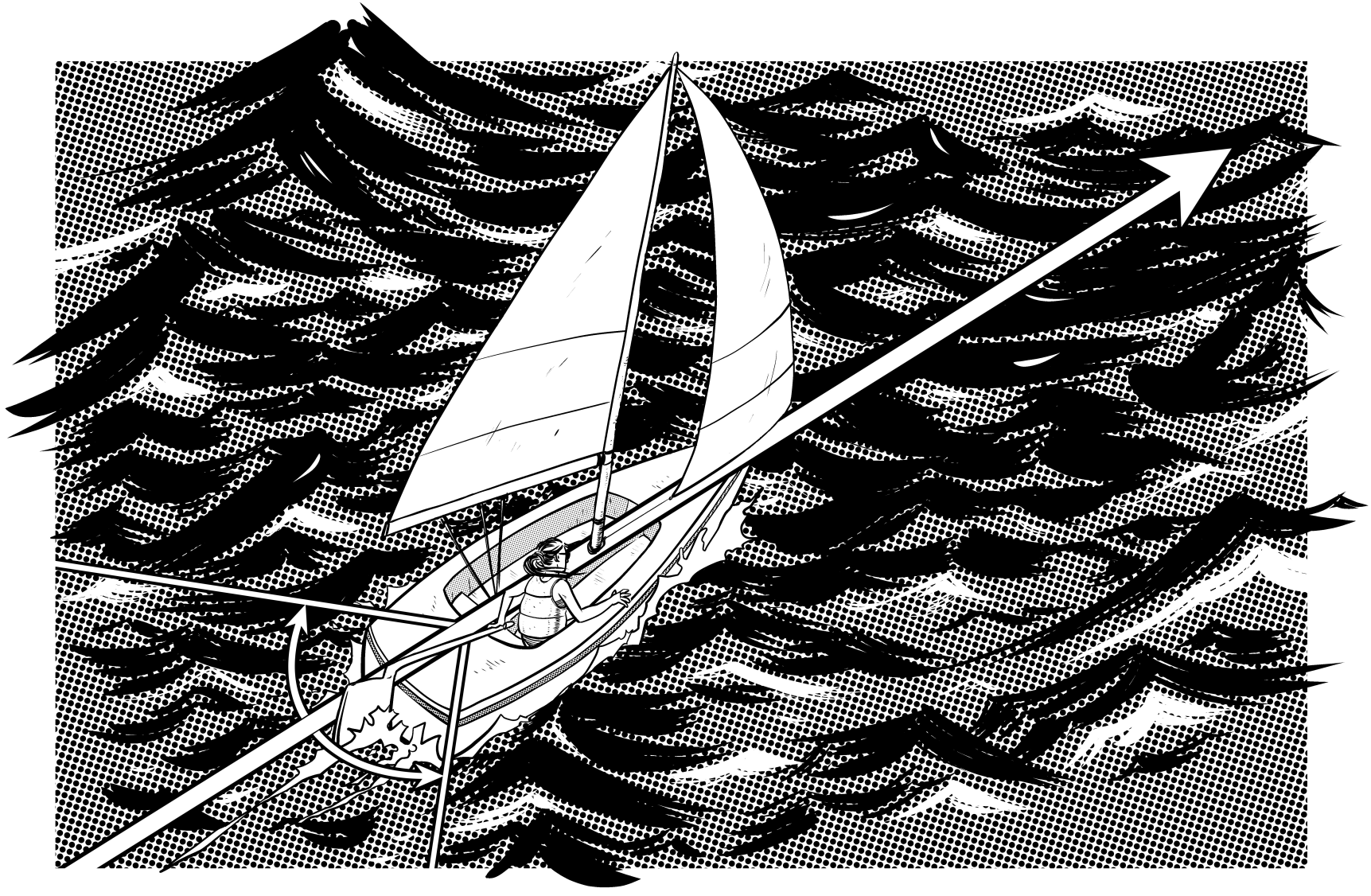
$$\frac{Cov(D,Y)}{\sqrt{Var_D}\sqrt{Var_Y}}$$

- Error extends to predictive modeling that isn’t based on causal frameworks

#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: Causality may mask correlations!



Modeling is Not the First Step

Most of us simply estimate models and cross our fingers that that coefficient is causal, but is it? When is it? Why is it? And which causal effect is it? And when is it reasonable to believe it?

We have to introduce concepts and notation first otherwise we will extend the correlation fallacy

Definition and Identification Come First

1. Turn the research question (“what is the causal effect of an advertising campaign on sales?”) into a specific aggregate causal parameter
2. Describe the narrow set of beliefs that make that parameter obtainable with data
3. Build a model that uses the data and the beliefs to estimate the causal parameter?

Most of us skip (1) and many skip (2) and go straight to (3) but hopefully today I’ll convince you that that’s how errors are introduced, even after one understands that causal inference is not merely correlational

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

Counterfactuals by Philosophers

“If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death.” – John Stuart Mill (19th century moral philosopher and economist)

“Causation is something that makes a difference, and the difference it makes must be a difference from what would have happened without it.” – David Lewis (20th century philosopher)

Statisticians

“Yet, although the seeds of the idea that [causal effects are comparisons of potential outcomes] can be traced back at least to the 18th century [most likely he means David Hume], the formal notation for potential outcomes was not introduced until 1923 by Neyman.” – Don Rubin (1990)

Jerzy Neyman's Notation

- Jerzy Neyman's 1923 masters thesis describes a field experiment with differing plots of land (imagine hundreds of square gardens) and many different "varieties" of fertilizer that farmers could apply to the land
- "U_{ik} is the yield of the ith variety on the kth plot..." (Neyman 1923)
- He calls U_{ik} "potential yield", as opposed to the realized yield because i (the fertilizer type) described all possible fertilizers that could be assigned to each k square garden
- Though only one fertilizer will be assigned to the land, many possible fertilizer assignments were possible beforehand, each with their own outcome

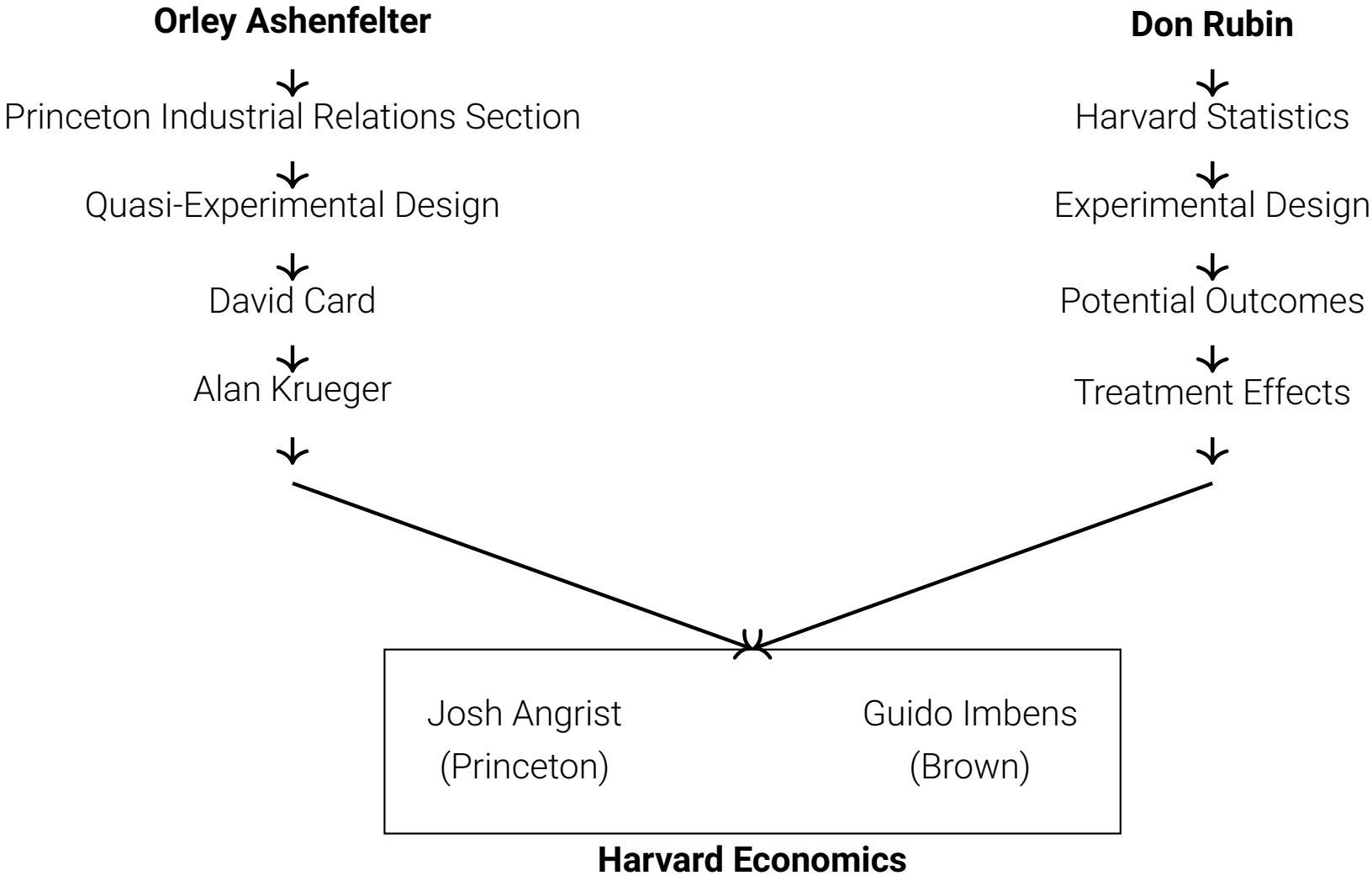
Jerzy Neyman's Notation

- For each fertilizer there is an associated "potential yield" that he collapses into U which he considers to be "a priori fixed but unknown" (Rubin 1990)
- Farmers draw fertilizer from an urn, like a bingo ball from a bingo ball machine, with replacement and apply it to each square garden
- Fertilizer assignment moves us from "all possible outcomes" to "realized outcome" terminology
- Neyman's urn model was a classic thought experiment, but it was also stochastically identical to the completely randomized experiment
- His arch-rival, Ronald Fisher, realizes this and publishes a book two years later calling for *randomizization* as the basis for causal inference

Treatment assignment mechanism

"Before the 20th century, there appears to have been only limited awareness of the concept of the assignment mechanism. Although by the 1930s, randomized experiments were firmly established in some areas of scientific investigation, notably in agricultural experiments, there was no formal statement for a general assignment mechanism and, moreover, not even formal arguments in favor of randomization until Fisher (1925)." (Imbens and Rubin 2015)

Two rivers into causal inference



Background I: Harvard Stats and Potential Outcomes

- Don Rubin, former chair of Harvard stats, is the main source of potential outcomes, building on Jerzy Neyman's 1923 work.
- Rubin's influential 1970s papers advocated for causal inference using contrasts of $Y(1)$ and $Y(0)$.
- Neyman's notation, initially in Polish, was translated into English in 1990, likely due to Rubin.
- Rubin expanded Neyman's ideas from experiments to observational studies, leading to developments like propensity score methods.
- Economics was slow to adopt these methods initially.

Background II: Princeton Industrial Relations Section

- Late 1970s and early 1980s: little ”credibility” in empirical labor studies.
- Princeton Industrial Relations Section: older than the economics dept, rigorous, non-partisan focus on US ”manpower”, highly empirical.
- Key faculty are Orley Ashenfelter, David Card, Alan Krueger.
- Key students include Bob Lalonde, Josh Angrist, Steve Pischke, John Dinardo, Janet Currie, Anne Case, and many more.
- Listen to David Card: https://youtu.be/1soLdywFb_Q?si=BCVqYeRz6jYiwHTQ&t=1580

Background II: Princeton Industrial Relations Section

Example of Princeton Paradigm Emerging

- Lalonde (1986) was a groundbreaking study, recently reviewed by Guido Imbens and Yiqing Xu (2024)
- Lalonde, a student of Card and Orley, analyzed an RCT on a job training program, finding an average treatment effect of +$800.
- He then replaced the experimental control group with survey data, reran econometric methods, and couldn't replicate the results.
- Orley and Card emphasized randomization in their 1985 Restat article, advocating for its exploitation in studies.

IV is Especially Bad

- Instrumental variables are like finding a four leaf clover – you should take it as a good omen
- It must satisfy several stringent conditions so it's hard to do well
- But in the old days, people used counterfeit instruments due to low standards which contributed to the empirical crisis
- Listen to Orley Ashenfelter:

  https://www.youtube.com/watch?v=GG627T4GbqU

H. Gregg Lewis on IV

"After reviewing virtually every study since 1963, Lewis reached the awkward conclusion that simple OLS of union wage effects were more useful and reliable than those based on IV or endogenous selection approaches. The problem, in his view, was that researchers used arbitrary and unsupported assumptions to identify their models with little or no concern for the validity of their assumptions or the implications of their findings." – Card (2022)

Angrist goes to Harvard

- Josh Angrist writes a paper evaluating the effect of military service on career earnings using IV
- Josh is Card and Orley's student (and Card and Josh win the Nobel prize in 2021 with Imbens)
- But his instrument is decidedly different – he uses "actually randomized" instruments. He uses the draft lottery
- Finds military services *causes* reductions in career earnings, goes on the Restud tour, gets a job at Harvard

Imbens goes to Harvard

- Guido Imbens (wins the Nobel with Card and Angrist) is a Dutch econometrics student in a masters program in the UK who follows a professor to Brown for his PhD
- Very interesting personality – unusually open minded, non-dogmatic, very intellectually curious
- https://youtu.be/cm8V65AS5iU?si=FiqecXsHOwjIk9Qc&t=540
- He goes to Harvard and overlaps with Angrist for one year
- Much of what we cover ends up being both parts of this tradition forming a paradigm as well as introducing econometric tools

Progress is made and progress is not made

- Part of the fusion between Princeton and Cambridge that I noted was to recast outcomes as potential outcomes, a technical term
- We need to make a distinction between now the idea of data (“realized outcomes”) and these hypothetical concepts represented by Neyman’s notation (“potential outcomes”)
- Let’s introduce potential outcomes and assume an intervention from the pandemic –
  - → Patient is either placed on a ventilator or not (”treatment” is the ventilator)
  - → What effect will it have on their health (e.g., mortality)?
  - → What is an ”effect” and how will we define it?

Potential outcomes notation

Let the treatment be a binary variable:

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

where i indexes an individual observation, such as a person

Potential outcomes notation

Potential outcomes:

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

where j indexes a potential treatment status for the same i person at the same t point in time

Potential outcomes

- Prior to being placed on the ventilator, there are multiple possibilities that could happen
  - → They could be placed on the ventilator, call it Y^1, and something could happen, or
  - → They could not be placed on the ventilator, call it Y^0, and something could happen
- There are only two things that could happen because the treatment only has two sides to it – you either are on it or you aren't
- Neyman, on the other hand, had an urn consisting of balls representing a large number of treatments (fertilizer types) and so had many potential outcomes

Potential outcomes vs realized outcomes

- Experimentalists thought of potential outcomes as ”real” in that either of those beforehand could happen – Rubin ”a priori fixed by unknown”
- But once the decision to put the patient on the ventilator happens, it eliminates one of them, and that’s the difference between potential outcomes and realized outcomes
- Potential outcomes are more or less describing what would happen under different treatments, but realized outcomes are describing what did in fact happen for a given treatment

Treatment Assignment Mechanism

- The second part of causal inference in the design tradition has to do with the reason that a person gets on a ventilator
- Did a doctor choose to do it? Did the patient choose to do it? Did an algorithm choose to do it?
- But it's more than that – *why* was the choice made, not *who* made it
- For instance, a doctor might always make it, but it matters if she flipped a coin vs read the chart and made the "best decision"

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, associated with a ventilator is equal to $Y_i^1 - Y_i^0$.

Important definitions

Definition 2: Switching equation

An individual's realized health outcome, Y_i, is determined by treatment assignment, D_i which selects one of the potential outcomes:

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

Not the same as treatment assignment mechanism. Treatment assignment mechanism describes why one patient but not another was put on the ventilator

Missing data problem

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*.

This is my reason from saying Mill's counterfactual framework derailed the quest for causal effects given counterfactuals are fictional

Missing data problem

- Fundamental problem of causal inference is deep and impossible to overcome – not even with more data (you will always with more data be missing one of the potential outcomes)
- Causal inference is a missing data problem
- All of causal inference involves imputing missing counterfactuals and not all imputations are equal

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

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

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

Aggregate parameters based on individual treatment effects are *summaries* of individual treatment effects

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

Average Treatment Effects are Simple Summaries

- Notice how in all three of these, all we did was take the defined treatment effect at the individual and aggregate
- Because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Missing data in this context isn't missing your car keys – it's missing unicorns and fire breathing dragons (fictional vs real data)
- While we cannot measure average causal effects, we can estimate them, but only in situations and we review one – randomization

Simple Comparisons

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by comparing the sample average outcome for the treatment group ($D = 1$) with a comparison group ($D = 0$)

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

SDO is not a causal parameter because it's comparing Y^1 and Y^0 for different units, not the same units, so what is it measuring?

Decomposition of the SDO

Decomposition of the SDO

The SDO is made up of three things:

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

where π is the share of units in the treatment group. Now let's see how we get here.

Two ways to rewrite the ATE

- Before we get started, let's look closely at the definition of the ATE
  - → We can express the ATE as the weighted average of the ATT and the ATU, and ...
  - → We can express the ATE as the sum of four conditional means multiplied by corresponding weights (Law of iterated expectations)
- They are in fact the exact same formula once you write down the definition of the ATT and the ATU
- Let's do it together before we get started:
  https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing

Begin with ATE definition

Rewrite the definition of the ATE

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

Let's make this easier to read by replacing the last row with letters

Change notation

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 definition

Rewrite ATE

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

Simple manipulation of ATE definition

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

Carry forward from previous slide

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

Replace letters with original terms

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

Purple terms are based on missing counterfactuals and therefore cannot be calculated.

This is an *identity*

Decomposition of the SDO

Decomposition of the SDO

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

Although we started with π (the share of units in treatment), note we have weighted the heterogeneity bias term by $1-\pi$ (the share of units in control)

Estimate SDO with sample averages

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

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

Illustrating selection bias with spreadsheets

- Eliminating selection bias requires understanding the selection mechanism – why did units end up treated but not others?
- Perfect Doctor exercise: assume a doctor sees patients, knows each person's treatment effects, despite counterfactuals, and assigns treatment based on whether gains are positive or not
- Illustrate decomposition together using numerical example: https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq3Oa6EQ/edit?usp=sharing

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

Design vs model based approaches to selection bias

- Historically two ways that this selection bias was addressed: modeling it directly (Heckman, and others) and by design (Rubin, Princeton)
    1. Design-based methods. Experimental design and randomization
    2. Model-based methods. Model the selection bias and then remove it mechanically
- Both have been highly influential, but constitute different approaches, and our workshop largely focuses on the former not the latter

Selection bias and Design

- Selection bias in causal inference is when one or both mean potential outcome differ by treatment status
- Source of the bias is caused by the why people get treated, or what's called the "treatment assignment *mechanism*"
- But as we will see, if the treatment was assigned randomly (exogeneity), the implications differ than if they were assigned "optimally" (endogeneity)

Treatment assignment mechanisms

- Two extreme examples of a treatment assignment mechanism:
  1. randomization (i.e., taking the medicine because a coin flip told you to take it) versus
  2. sorting on one or both potential outcome (i.e., taking the medicine because it'll help you) which I call the "Perfect Doctor" but which Heckman and others call a "Roy model"
- Bias comes from how treatment is assigned and that mechanism dictates the direction we have to take

Three forms of selection bias

- In causal inference, selection bias is caused by different mean potential outcomes by treatment status, of which there are three possibilities:
    1. Selection on Y^0: You chose the treatment because of what will happen if you didn't
    2. Selection on Y^1: You chose the treatment because of what will happen if you do
    3. Selection on gains, δ: You chose the treatment because the net benefits were positive
- All three cause biased estimates of the ATE, though the degree to which it fully distorts the estimates depends on those different reasons for sorting into treatment

Humans cause selection bias, not models

- An unusual paradox in causal inference is that our successful programs are the hardest to study:
  - → The better our programs are, the better humans get at assigning treatments to people, the more interventions target people correctly based on treatment gains, the worse the selection bias is
  - → But the less efficient our programs, the more erratic the assignment mechanism is, the more mistakes there are in targeting people with their best options, the less the selection bias is
- The reason for the "natural experiment" drive in causal inference was based on what we could deduce about selection bias from designs, which was more like the latter
- But solely depending on the latter would be limiting if our aim was to study our most successful programs (e.g., food stamps or STAMP)

Summarizing the goals of causal inference

Our goal in causal inference is to estimate aggregate causal parameters with data using treatment assignment mechanisms that plausibly eliminate selection bias

Depending on the treatment assignment mechanism, certain procedures are allowed and others are prohibited

Let's look what happens in an RCT *and why* this addresses selection bias term $E[Y^0|D=1]$ and $E[Y^0|D=0]$ to see why Fisher (1925) recommended it

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:

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

Identification with Full Independence

$$\underbrace{E_N[Y_i|D_i=1] - E_N[Y_i|D_i=0]}_{\text{Estimate of SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{0}_{\text{Selection bias}} + \underbrace{0}_{\text{Heterogenous treatment effect bias}}$$

SDO is unbiased estimate of ATE with randomized treatment assignment because it sets selection bias to zero and $ATT = ATU$.

What about partial independence?

- For homework, write down the independence terms and substitute it into the decomposition if selection into treatment is independent of Y^0 but not Y^1
- What if it is independent of Y^0 but not Y^1?
- Can you describe formally the bias caused by partial independence? What terms are eliminated? What terms remain?
- We will review it tomorrow

Interference when aggregating units

- While treatment effects are defined at individual level, aggregate parameters combine units
- This therefore means that for the aggregate parameters to be stable, one unit's treatment choice cannot "interfere" with another unit's potential outcomes
- Placing limits on those possibilities creates challenges for definitions and estimation that are probably huge headaches, even in the RCT
- Violations are an active area of scholarship and important for social networks, peer effects and various platforms (e.g., Twitter)

SUTVA

- SUTVA stands for “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*
- Means that a potential outcome has nothing to do with others’ treatment assignment – only one’s own, and only today

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
- Can be mitigated with careful delineation of treatment and control units so that interference is impossible, may even require aggregation (e.g., classroom becomes the unit, not students)

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

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)

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

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

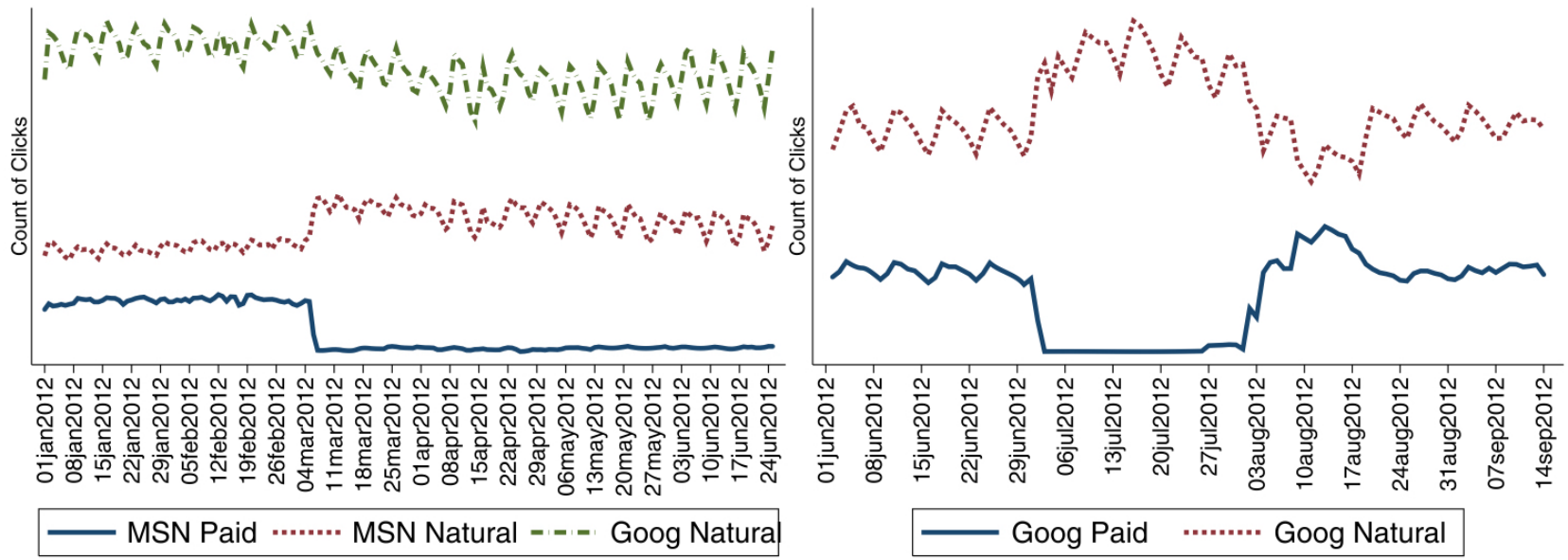
- Treatment was targeted ads at particular people conducting particular types of keyword search
- Consumers who choose to click on ads are loyal and already informed about products with high likelihood to buy already
- Problem is ads are targeting people at the end of their search, so the question is 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.”*

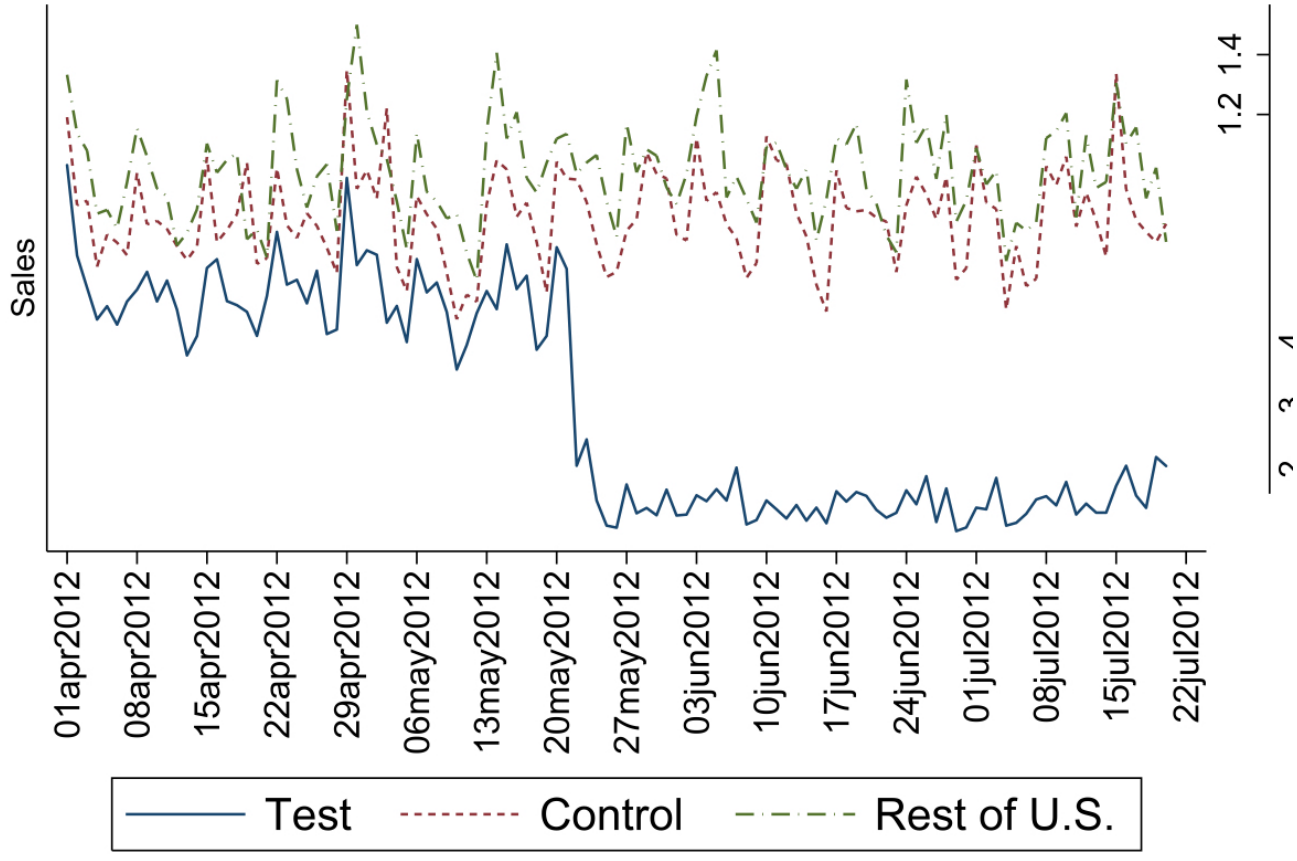
RCT

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

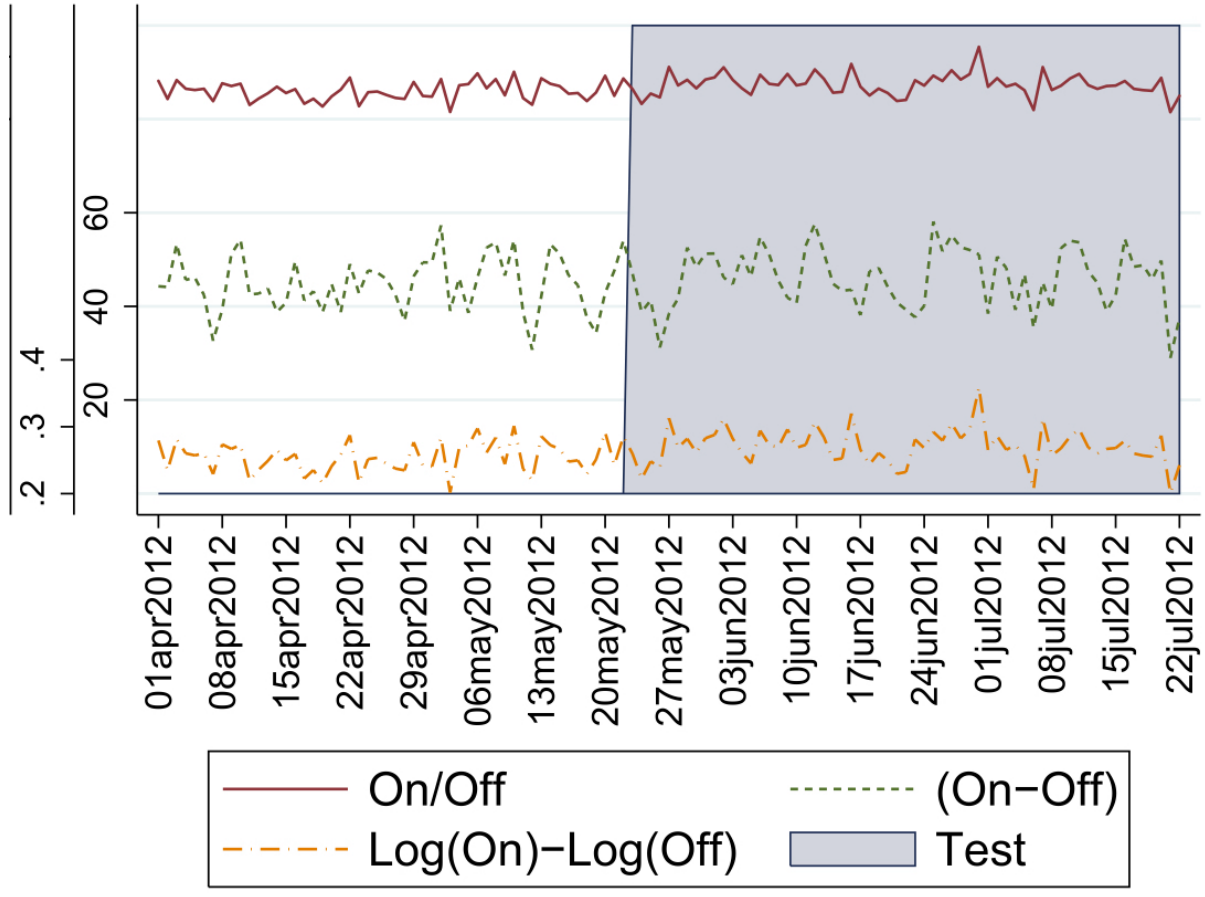
Design of the experiment

- 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%

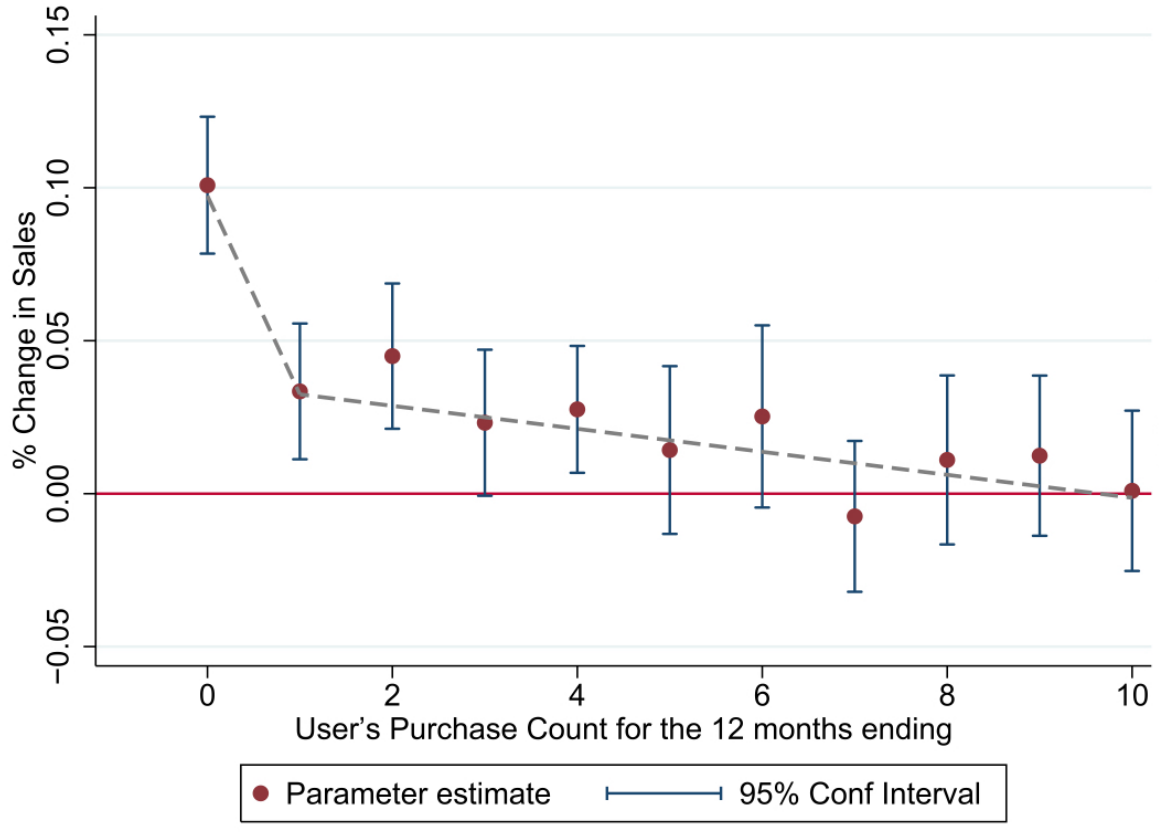
Figure: Spending effect on revenue using OLS but not the randomization. Effects are gigantic.

	(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.

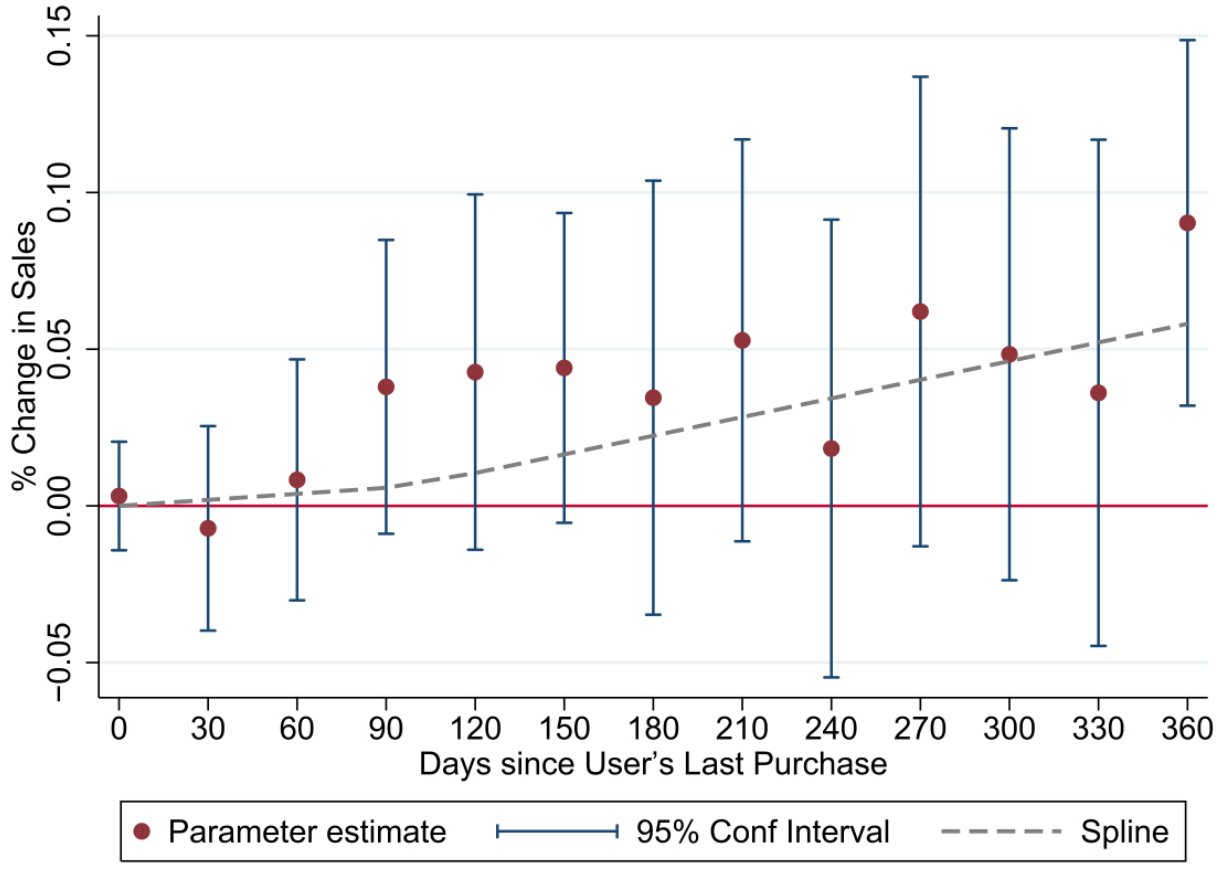
Heterogenous treatment effects

- 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

Roadmap

What is Mixtape Sessions?

Foundational ideas

Potential Outcomes

- History of potential outcomes
- Experimental Design Fuses with Quasi-Experimental Design

Independence and Selection Bias

Industry example of RCT: eBay advertising

Tennessee's small class size RCT: STAR

Example 1: Krueger (1999)

- Krueger (1999) econometrically re-analyzes a randomized experiment to determine the causal effect of class size on student achievement
- The project is the Tennessee Student/Teacher Achievement Ratio (STAR) experiment from the 1980s
- 11,600 students and their teachers were *randomly* assigned to one of the following three groups:
    1. Small class of 13-17 students
    2. Regular class of 22-25 students
    3. Regular class of 22-25 students with a full-time teacher's aide
- After the assignment, the design called for students to remain in the same class type for four years
- Randomization occurred within schools

Regression analysis of experiments

- With randomization one could simply calculate SDO (simple difference in mean outcomes) for the treatment and control group and know that SDO=ATE because of independence
- Nonetheless, it is often useful to analyze experimental data with regression analysis (see MW section 3.2.2; MHE ch. 2)
- Assume that treatment effects are constant – i.e., $Y_i^1 - Y_i^0 = \delta \; \forall i$
- Substitute into a rearranged switching equation (Definition 2):

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\
Y_i &= Y_i^0 + (Y_i^1 - Y_i^0) D_i \\
Y_i &= Y_i^0 + \delta D_i \\
Y_i &= E[Y_i^0] + \delta D_i + Y_i^0 - E[Y_i^0] \\
Y_i &= \alpha + \delta D_i + \eta_i
\end{aligned}
$$

  where η_i is the random part of Y_i^0
- This is a regression equation that could be used to estimate the causal effect of D on Y

Regression analysis of experiments (cont.)

- The conditional expectation, $E[Y_i|D_i]$, with treatment status switched on and off gives:

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

- Subtracting the latter from the former, we get:

$$\underbrace{E[Y_i|D_i=1] - E[Y_i|D_i=0]}_{\text{SDO}} = \underbrace{\delta}_{\text{Treatment Effect}} + \underbrace{E[\eta_i|D_i=1] - E[\eta_i|D_i=0]}_{\text{Selection bias}}$$

- We can estimate SDO using least squares but there's other options as well
- In the STAR experiment, D_i, equalled one if the student was enrolled in a small class and had been *randomly* assigned
- Recall that randomization implies that treatment is independent of potential outcomes, and therefore the selection bias vanishes

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 prior equation, we might estimate:

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

- 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. (here that's the school). We'll discuss "conditional independence assumption" when we cover matching.
  2. Additional 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.

Regression in Krueger (1999)

Krueger estimates the following econometric model

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 \text{REG/A}_{cs} + \alpha_s + \varepsilon_{ics}$$

- i indexes a student, c indexes a class, and s indexes a school
- Y_{ics} is the student's percentile score
- $SMALL_{cs}$ is a dummy equalling 1 if she is assigned to a small class.
- REG/A_{cs} is a dummy equalling 1 if she was assigned to a regular class with an aide
- α is a "school fixed effect" which is a vector of school-specific dummy variables. He conditions on school fixed effects because randomized classroom assignment occurred *within* schools.

Regression results Kindergarten

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			A. Kindergarten	
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master's degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Regression results 1st grade

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			B. First grade	
Small class	8.57	8.43	7.91	7.40
	(1.97)	(1.21)	(1.17)	(1.18)
Regular/aide class	3.44	2.22	2.23	1.78
	(2.05)	(1.00)	(0.98)	(0.98)
White/Asian (1 = yes)	—	—	6.97	6.97
			(1.18)	(1.19)
Girl (1 = yes)	—	—	3.80	3.85
			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61
			(.87)	(.87)
White teacher	—	—	—	−4.28
				(1.96)
Male teacher	—	—	—	11.82
				(3.33)
Teacher experience	—	—	—	.05
				(0.06)
Master's degree	—	—	—	.48
				(1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

Problem 1: Attrition

A common problem in randomized experiments on humans is *attrition* – i.e., people leaving the experiment

- If attrition is random, then attrition affects the treatment and control groups in the same way and our estimates remain unbiased
- But in this application, attrition is probably non-random: especially good students from large classes may have enrolled in private schools creating a selection bias problem
- Krueger addresses this concern by imputing the test scores (from their earlier test scores) for all children who leave the sample and then re-estimates the model including students with imputed test scores.

Problem 1: Attrition

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

	Actual test data		Actual and imputed test data	
Grade	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32 (.76)	5900	5.32 (.76)	5900
1	6.95 (.74)	6632	6.30 (.68)	8328
2	5.59 (.76)	6282	5.64 (.65)	9773
3	5.58 (.79)	6339	5.49 (.63)	10919

Estimates of reduced-form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects; a dummy variable for student gender; and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

- Non-random attrition hardly biases the results.

Problem 2: Switch Classrooms after Random Assignment

"It is virtually impossible to prevent some students from switching between class types over time." (Krueger 1999, p. 506)

B. First grade to second grade

	Second grade			
First grade	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

- Interpreting Krueger's"transition matrix" (above)
  - → If students remained in their same class type over time, all the off-diagonal elements would be zero
  - → Interpretation: Of the 1,482 first graders assigned to small classrooms, 1,435 remained in small classes; 23 and 24 switched into regular and regular with aide classes in the second grade
- If students with stronger expected academic potential were more likely to move into the small classes, then these transitions would bias a simple comparison of outcomes across class types.

Problem 2: Switch Classrooms after Random Assignment

- Subjects moved between treatment and control groups. How to address this?
- A common solution to this problem is to use initial classroom assignment (i.e., small or regular classes) as an *instrument* for actual assignment. We will discuss instrumental variables later, so I will hold off on that now.
- Krueger reports regression results where instead of the student's actual status as the treatment variable, he regresses performance against their *randomly assigned class size*. This is called the "reduced form" model, and we learn more about this when we cover IV.
- In Kindergarten, OLS and reduced form are the same because students remained in their initial class for at least one year.
- From grade 1 onwards, OLS and reduced form results differ.

Problem 2: Switch Classrooms after Random Assignment

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
B. First grade								
Small class	8.57	8.43	7.91	7.40	7.54	7.17	6.79	6.37
	(1.97)	(1.21)	(1.17)	(1.18)	(1.76)	(1.14)	(1.10)	(1.11)
Regular/aide class	3.44	2.22	2.23	1.78	1.92	1.69	1.64	1.48
	(2.05)	(1.00)	(0.98)	(0.98)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian (1 = yes)	—	—	6.97	6.97	—	—	6.86	6.85
			(1.18)	(1.19)			(1.18)	(1.18)
Girl (1 = yes)	—	—	3.80	3.85	—	—	3.76	3.82
			(.56)	(.56)			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61	—	—	−13.65	−13.77
			(.87)	(.87)			(.88)	(.87)
White teacher	—	—	—	−4.28	—	—	—	−4.40
				(1.96)				(1.97)
Male teacher	—	—	—	11.82	—	—	—	13.06
				(3.33)				(3.38)
Teacher experience	—	—	—	.05	—	—	—	.06
				(0.06)				(.06)
Master's degree	—	—	—	.48	—	—	—	.63
				(1.07)				(1.09)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

Comments

- Natural experiments are valuable, but they don't always have the same certainty the way an RCT does
- We use natural experiments when people won't let us run the RCTs we want to run!
- Findings from natural experiments often push others to run RCTs – like at eBay

Going forward

- Now let's move into a set of tools that will help us in two of the areas we cover: DAGs
- Matching/regression and instrumental variables both depend critically on knowing something about the data generating process
- We'll be learning one way to assist you