

Causal Inference I

MIXTAPE SESSION



Roadmap

Introduction to course

- What is Mixtape Sessions?

- Managing your expectations

- Future Tracks

Potential outcomes

- Naive causal inference

- Independence and Selection Bias

- Industry example of RCT: eBay advertising

- Policy example of RCT: HIV status

Welcome!

- Scott Cunningham, professor of economics at Baylor University, author of Causal Inference: the Mixtape
- I run workshops on causal inference all over the world, but I am not an econometrician – I'm a run of the mill applied microeconomist who studies severe mental illness who just happens to love learning and teaching
- Workshops can be helpful ways to plug into one's methodological training, and online workshops are very helpful because of the recordings, the coding together, and bunch of bells and whistles (e.g., github repositories)
- Causal inference is an old field but which has increasingly drawn people to it (Nobel Prize two years ago maybe helped)

What is Mixtape Sessions?

- Mixtape Sessions is my online platform started in November 2022 to “democratize causal inference” by helping connect people, from beginner to advanced, with material and teachers that for various reasons may not be accessible otherwise
- I became obsessed with teaching and writing about causal inference because of a strong conviction that (a) it is important and (b) not everyone had the same chances to learn it
- I wrote a 3-part sociological history of it on my substack that I encourage you to read so you can better understand the context for why I say that
- I tend to emphasize intuition, mechanics, narrow calculations, meaning, assumptions, code including actually taking time to code, advocate for data visualization – in other words the art and the science

4-day Causal Inference Workshop

- We workshop together for 4-days, 6pm to 2am CST (but 7am to 4pm GMT+8), with 15 min breaks on the hour and a 1-hour lunch break at 10:00PM CST (11am GMT+8)
- I mix exposition, discussion of papers, coding exercises and discussion as best as I can
- I'm me, and I teach how I teach, with passion, enthusiasm, deep joy, but I'm not an econometrician so sometimes I take the long way to get there when an econometrician would do it much faster

Class goals

1. **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
2. **Comprehension:** You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation
3. **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 the github repo: github.com/Mixtape-Sessions/Causal-Inference-1
- Each lecture will be recorded and then uploaded to Vimeo as a password protected file that you'll have access to into perpetuity
- Kyle Butts and I are committed to over time making the Github Repository like an open public library where the only club goods are (a) recordings, (b) Discord and (c) live lectures

Workshop (Part 1) Topics

1. Foundations and Graphs: Day 1
2. Graphs and Unconfoundedness: Day 2
3. Instrumental Variables: Day 3
4. Regression Discontinuity Design: Day 4

Today: Day 1

1. Potential outcomes
2. Randomization inference
3. Directed acyclic graphs

What's coming this fall?

- Lots of great “Mixtape Tracks” – shorter workshops taught by top professors at Brown, MIT, BYU and elsewhere
- Think of Mixtape Sessions as a bridge to more content and more people

<https://www.mixtapesessions.io/sessions/>

Three kinds of workshops

1. **The Classics:** Causal 1, 2 and 3 (new) by me
2. **The Singles:** Focusing more on specific content we covered
3. **Deep cuts:** Advanced material

The Classics

1. Causal Inference 1: Foundational material, potential outcomes, unconfoundedness, instrumental variables, regression discontinuity
2. Causal Inference 2: Difference-in-differences
3. Causal Inference 3: Synthetic control

Singles

1. Regression discontinuity design (Roci'o Titiunik, Princeton Political Science Dept)
2. Synthetic Control and Clustering (Alberto Abadie, MIT Economics *)
3. Doing Applied Research (Mark Anderson at Montana State and Dan Rees at UC3M)
4. Machine Learning and Causal Inference (Brigham Frandsen, BYU)
5. Experiments and Survey Design (Rebecca Thornton, Baylor)
6. Instrumental Variables (Peter Hull, Brown)

Deep Cuts

1. Advanced DiD (Jon Roth, Brown)
2. Advanced DiD (Brantly Callaway, UGA)
3. Machine Learning and Heterogeneous Treatment Effects (Brigham Frandsen, BYU)
4. Design (Peter Hull, Brown)
5. Shift-Share (Peter Hull, Brown)

Roadmap

Introduction to course

- What is Mixtape Sessions?

- Managing your expectations

- Future Tracks

Potential outcomes

- Naive causal inference

- Independence and Selection Bias

- Industry example of RCT: eBay advertising

- Policy example of RCT: HIV status

What is Causality?

- Causality is metaphysics; causal inference is epistemology – what makes a causal belief a “warranted belief”?
- Causal inference has many mothers and fathers
- Aristotle was different than Hume, Mill, and Lewis and as I am not comfortable saying Aristotle was wrong, I won’t
- What I can do is explain the potential outcomes framework of causal inference and use it to discuss methods and tools that estimate unbiased and meaningful causal parameters as defined by that framework

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.

Prediction		Causal Inference		Statistics/Econometrics	Machine Learning
		Observational	Experimental		
ANOVA	Linear Regression	Difference-in-Differences	A/B Testing		
Logistic Regression	Time Series Forecasting	Instrumental Variables	Business Experimentation		
		Propensity Score Matching	Randomized Controlled Trials		
		Regression Discontinuity			
Boosting	Decision Trees & Random Forests	Additive Noise Models	Causal Reinforcement Learning		
Lasso, Ridge & Elastic Net	Neural Networks	Causal Forests	Multiarmend Bandits		
	Support Vector Machines	Causal Structure Learning	Reinforcement Learning		
		Directed Acyclic Graphs			
		Double/Debiased Machine Learning			

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

Three New Ideas

1. **Counterfactual:** Philosophers come to it first and its central role in causal inference makes causality *unknowable* that the project is nearly derailed
2. **Treatment assignment mechanism:** Neyman and Fisher solve the counterfactual problem in statistics and lay the foundation of the modern randomized controlled trial (RCT) with their focus on the selection process
3. **No One Causal Effect:** There is no such thing as “the causal effect”; there’s many and your first step is to pick a parameter (not as easy as it sounds)

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

Modern Philosophers Introduce Counterfactual Comparisons

"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)

Counterfactuals Almost Derailed Causal Inference

Mill's counterfactuals were immensely valuable for the clarity of the definition as well as its intuitive validity of causality, but it came at a huge price

If I have to know what would have happened had I not eaten the dish, but I did eat the dish, then isn't it actually impossible to know the causal effect of eating the dish?

Statisticians surprisingly resolve this tension in the early 20th century with the introduction of notation and the principles of treatment assignment

Statistical origins

"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 article 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 i th variety on the k th 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 *randomization* 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)

Progress is made and progress is not made

- Econometrics traditionally modeled causality in terms of realized outcomes until recently (with some exceptions)
- 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”)
- Listen to Guido Imbens describe the transition towards modeling causality in terms of “realized outcomes”

<https://www.youtube.com/watch?v=drGkRy53bB4>

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

Realized vs potential outcomes

- Potential outcome Y^1 refers to the “a priori fixed but unknown” outcomes associated with different possible treatment assignments
- Realized outcome Y refers to the “posterior and known” outcome associated with a specific treatment assignment
- Potential outcomes become realized outcomes through treatment assignment generated by an assignment mechanism like randomization or rationality

Models vs Treatment Assignment

- Treatment assignment *mechanism* drives the entire effort to identify causal effects as some make it easy and some make it potentially *impossible*
- Put another way, the same model can be unbiased and biased depending on the treatment assignment and be utterly detectable otherwise
- Means modeling does not come first – it comes last

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:

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

Not the same as treatment assignment mechanism. Treatment assignment mechanism describes how D was assigned, not whether it was assigned.

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 have 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

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

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

Begin with ATE definition

Law of iterated expectations

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

ATE is sum of four conditional expectations (can also be rearranged as a weighted average of the ATT and the ATU)

Change notation

Substitute letters for expectations

$$E[Y^1|D = 1] = a$$

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

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

$$E[Y^0|D = 0] = d$$

$$\text{ATE} = e$$

Rewrite ATE definition

Rewrite ATE

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

Simple manipulation of ATE definition

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

$$e = \pi a + b - \pi b - \pi c - d + \pi d$$

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

$$0 = e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d}$$

$$\mathbf{a} - \mathbf{d} = e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d}$$

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

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

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

Carry forward from previous slide

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

Replace letters with original terms

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

Purple terms are explicitly missing counterfactuals

Decomposition of the SDO

Decomposition of the SDO

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

Note: this is a *rewritten* formula for the definition of the ATE and so it is an identity and thus *always* true. Also, we started with π but in the end we weight by $1 - \pi$.

Estimate SDO with sample averages

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

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.

Selection bias

- Selection bias in the potential outcomes framework is two mean potential outcomes differing for two groups,
- But one of them is fictional and the other isn't
- Source of the bias is the treatment assignment mechanism covariates

Bias #1: Selection bias

- Look very closely at the selection bias terms on their left and right hand sides

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

- Most likely, doctors “selected” units into and out of treatment based on Y^0
- Selection bias is caused by a treatment assignment mechanism that selects units into treatment based on Y^0 (also called “sorting”)

Humans cause selection bias, not statistical model

- Sorting into treatment based on potential outcomes will always create selection bias
- Following correlations between a PhD and happiness are not causal:
 1. I chose to get a PhD because I thought I would be unhappy without one – i.e., selection on Y^0
 2. I chose to get a PhD because I thought it would be happy with one – i.e., selection on Y^1
 3. I chose to get a PhD because treatment effects were positive – i.e., selection on treatment gains, $\delta = Y^1 - Y^0$
- More rational and efficient our decision making processes become, the more they are like selection on gains, the worse selection bias is

Illustrating selection bias with spreadsheets

- Eliminating selection bias requires understanding the selection mechanism – why did units end up treated but not others?
- Illustrate with Perfect Doctor exercise – doctor knows each person's treatment effects, despite counterfactuals, and assigns treatment based on whether gains are positive or not
- Illustrate decomposition using numerical example

https://docs.google.com/spreadsheets/d/10DuQqGtH_Ewea7zQoLTFYHbnvqaTVDhn2GDzq30a6EQ/edit?usp=sharing

Summarizing the goals of causal inference

Our goal in causal inference is to estimate aggregate causal parameters with data by exploiting what is known about the treatment assignment mechanism

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.

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

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

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 $1 - \pi$:

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

- If treatment is independent of potential outcomes, then:

$$E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0] = E[Y^1] - E[Y^0]$$

$$SDO = ATE$$

Identification with Randomization

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

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

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, there cannot be “interference” between one unit’s treatment choice and another unit’s potential outcome
- Creates challenges for definitions and estimation that are probably huge headaches, even in the RCT

SUTVA

- SUTVA stands for “stable unit-treatment value assumption”
 1. **S**: *stable*
 2. **U**: across all *units*, or the population
 3. **TV**: *treatment-value* (“treatment effect”, “causal effect”)
 4. **A**: *assumption*
- Largely about interference when aggregating but also 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
- 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)

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)

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

BY THOMAS BLAKE, CHRIS NOSKO, AND STEVEN TADELIS¹

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

PAID SEARCH EFFECTIVENESS

161

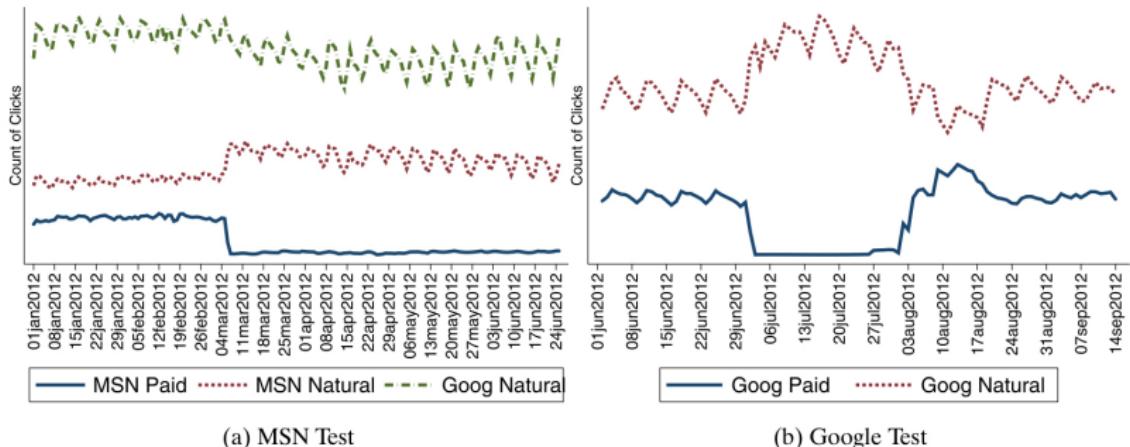


FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for ‘ebay’ terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

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 website. We overcome this endogeneity challenge with our controlled experiments."

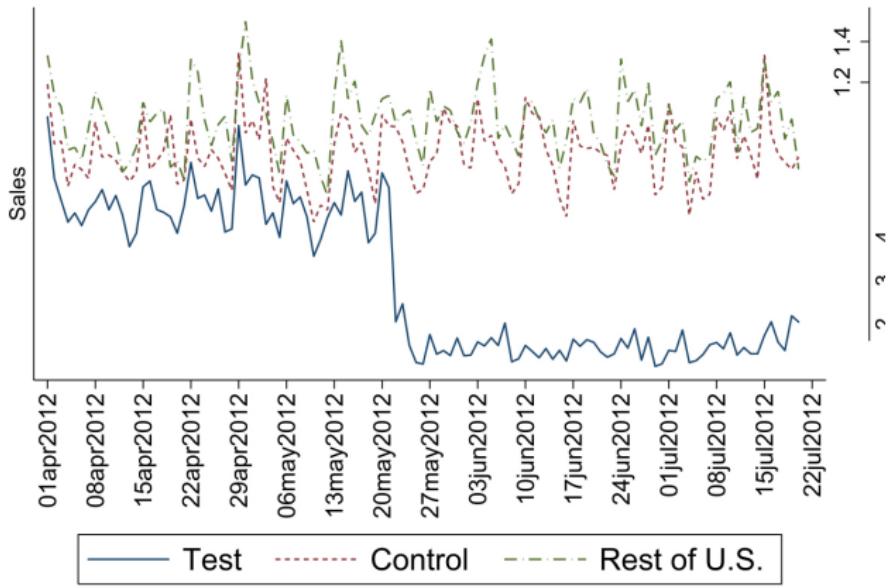
RCT

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

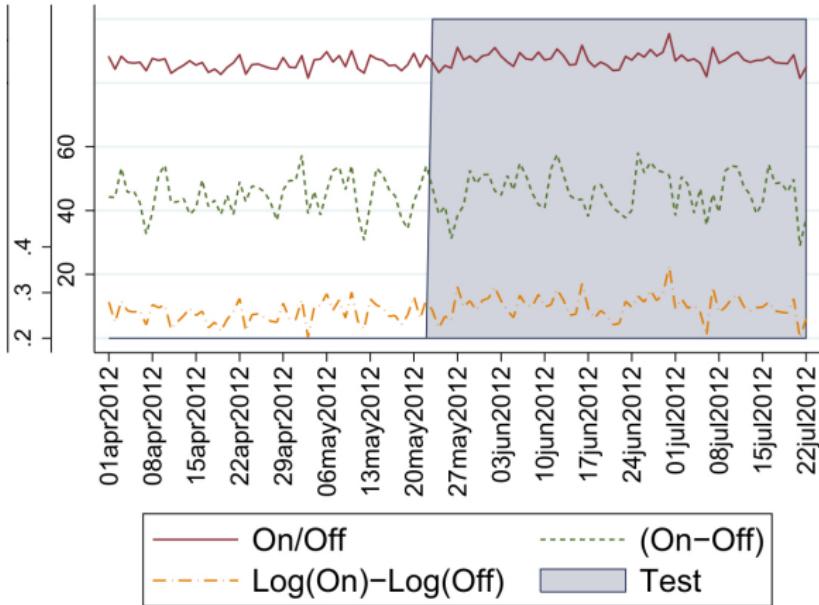
Design of the experiment

- 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
<i>N</i>	10,500	10,500
$\Delta \ln(Spend)$ Adjustment	3.51	3.51
$\Delta \ln(Rev)$ (β)	3.10635	0.44226
<i>Spend</i> (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%

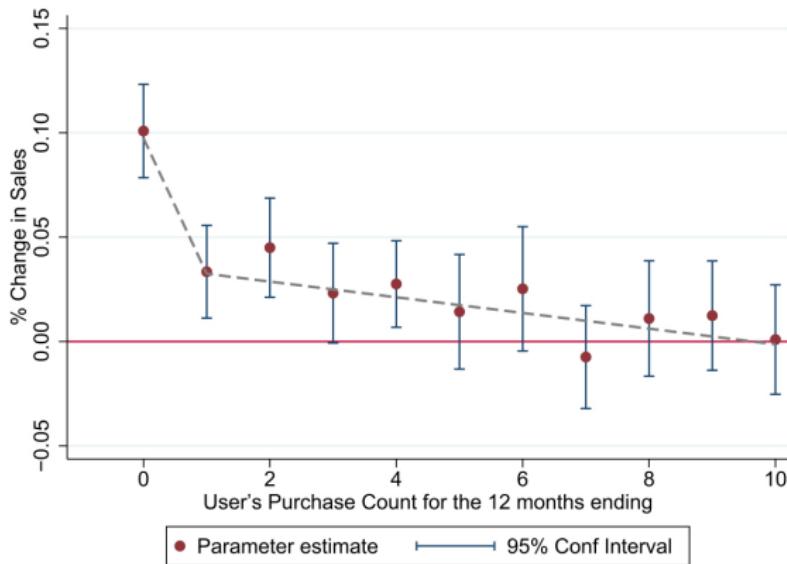
Figure: Spending effect on revenue using OLS but not the randomization.
 Effects are gigantic.

	(5)
Estimated Coefficient	0.00659
(Std Err)	(0.0056)
DMA Fixed Effects	Yes
Date Fixed Effects	Yes
<i>N</i>	23,730
$\Delta \ln(Spend)$ Adjustment	1
$\Delta \ln(Rev)$ (β)	0.00659
<i>Spend</i> (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.

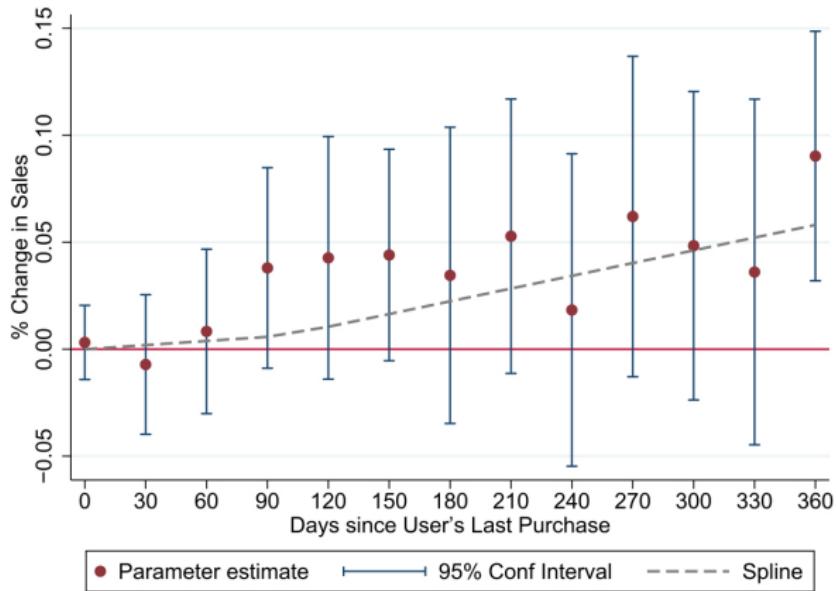
Heterogenous treatment effects

- 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

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

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, some of which we cover tomorrow

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

	1	2	3	4	5
Any incentive	0.431*** (0.023)	0.309*** (0.026)	0.219*** (0.029)	0.220*** (0.029)	0.219 *** (0.029)
Amount of incentive		0.091*** (0.012)	0.274*** (0.036)	0.274*** (0.035)	0.273*** (0.036)
Amount of incentive ²			-0.063*** (0.011)	-0.063*** (0.011)	-0.063*** (0.011)
HIV	-0.055* (0.031)	-0.052 (0.032)	-0.05 (0.032)	-0.058* (0.031)	-0.055* (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

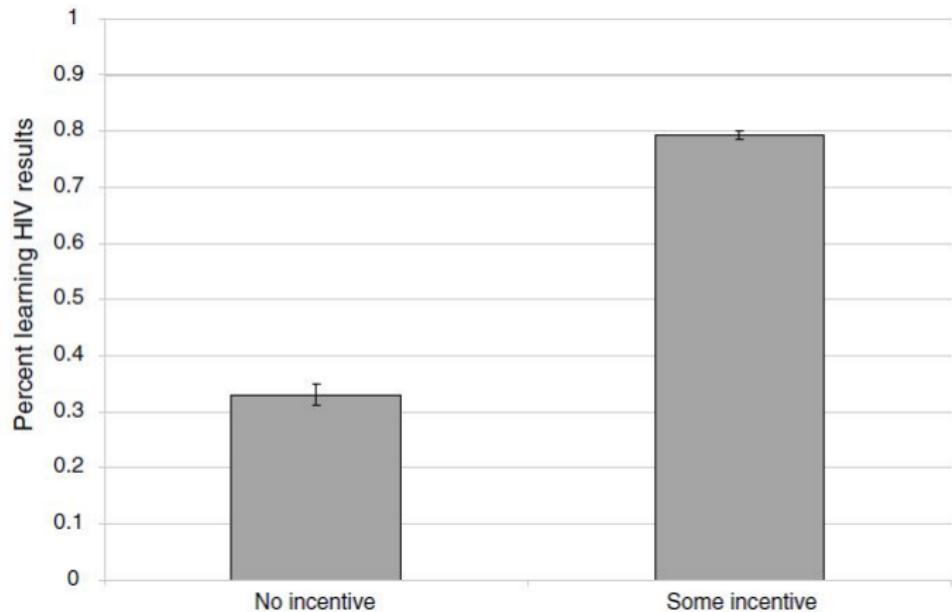


Figure: Visual representation of cash transfers on learning HIV test results.

Results

- Any incentive increases learning HIV status by 43 percentage points compared to the control (34% of controls learned HIV status)
- Next she looks at the effect that learning HIV status has on risky sexual behavior
- She had to do a lot of planning by creating two sources of randomization – the voucher and the distance to clinics – which required using instruments (we discuss it next week)

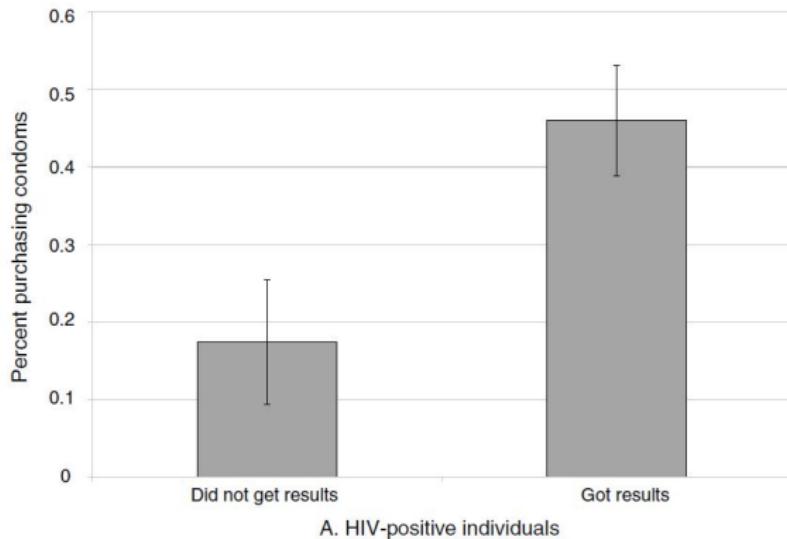


Figure: Visual representation of cash transfers on condom purchases for HIV positive individuals.

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 (0.025)	-0.069 (0.062)	-0.193 (0.148)	-0.303 (0.285)
Got results × HIV	0.418*** (0.143)	0.248 (0.169)	1.778*** (0.564)	1.689** (0.784)
HIV	-0.175** (0.085)	-0.073 (0.123)	-0.873 (0.275)	-0.831 (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

Thoughts you want to keep in mind

- Even when we don't run our own experiments, the experimental design is helpful for starting your project
- Describe the way you would conduct the RCT by explaining the following:
 - What is the research question?
 - What aggregate causal parameter?
 - What experiment am I running?
 - Who is my treatment and control group?
 - What regression will I run?
- Try saying this: "Describe the steps you would take to do this if you had all the money in the world"

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