Causal Inference I

Mixtape Session

Roadmap

Foundational ideas

Brief (and selected) history of causal inference

Independence and Selection Bias

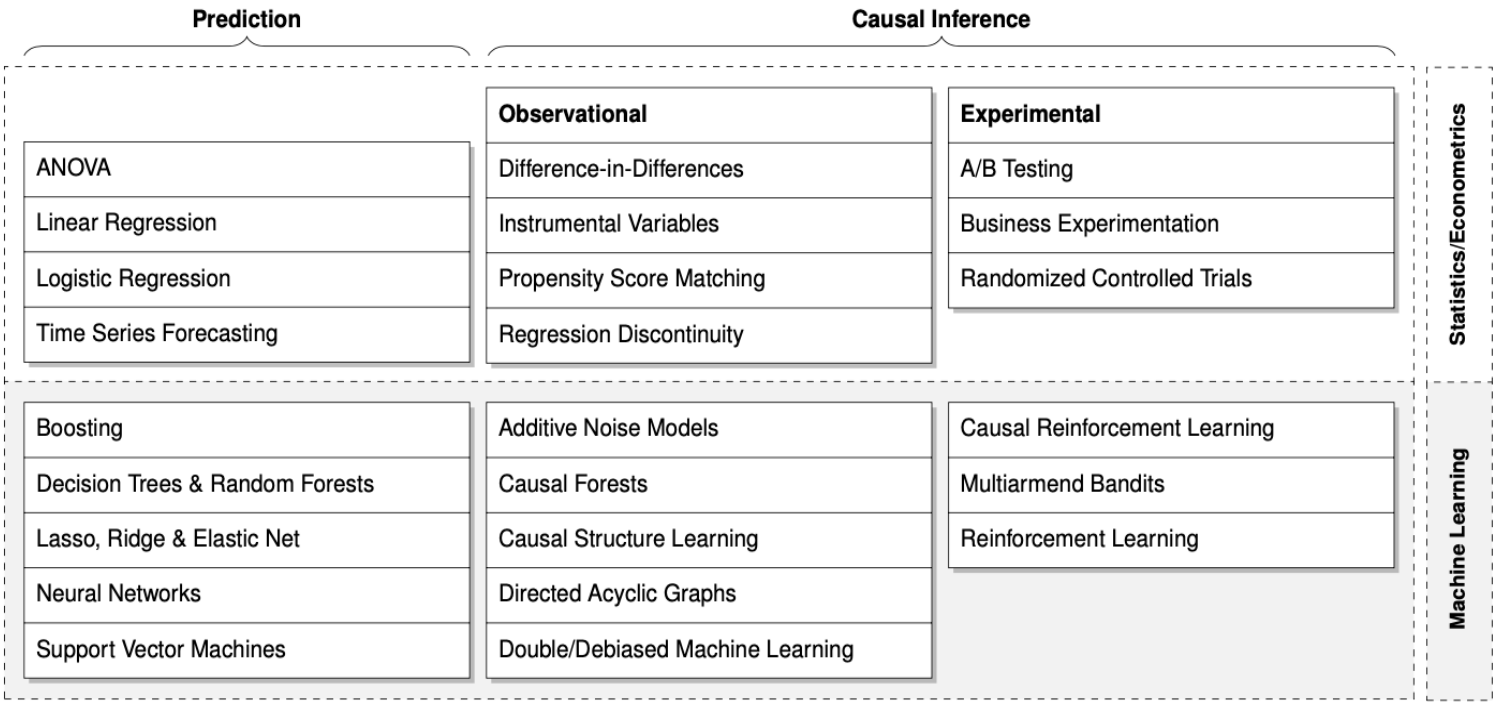
Industry example of RCT: eBay advertising

What is Causality?

- Causality is metaphysics (what is?); causal inference is epistemology (what is knowledge?) IOW, what makes a causal belief a “warranted belief”?
- Causality and causal inference have common sources (e.g., Aristotle, Hume, Mill, Lewis) but we will focus primarily on the experimental design tradition
- We use the potential outcomes framework of causal inference to build and discuss methods that estimate unbiased and meaningful **average** 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.



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:

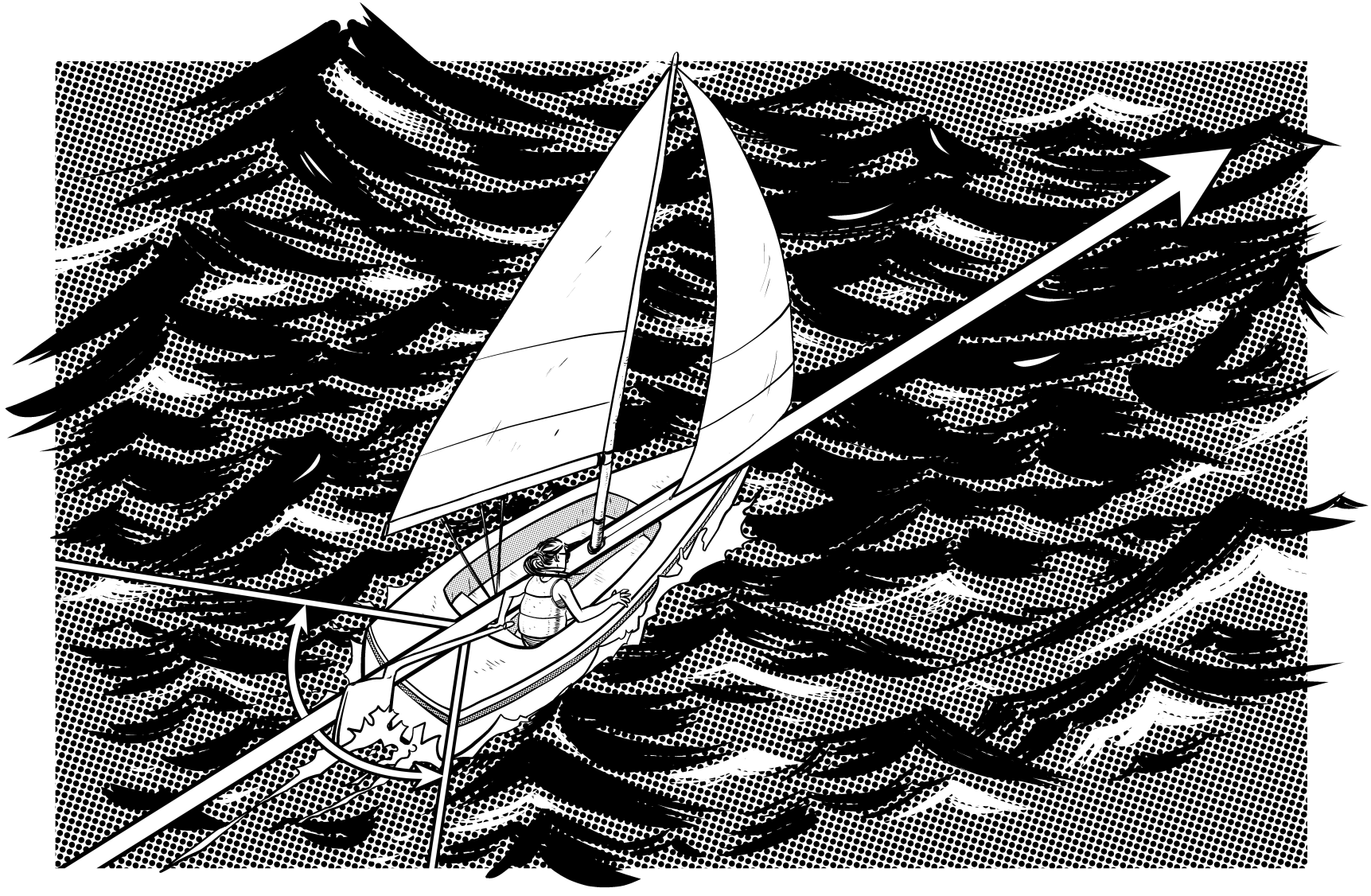
$$\frac{Cov(D,Y)}{\sqrt{Var_D}\sqrt{Var_Y}}$$

- 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

Foundational ideas

Brief (and selected) history of causal inference

Independence and Selection Bias

Industry example of RCT: eBay advertising

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)

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 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)

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:

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

Not the same as treatment assignment mechanism. Treatment assignment mechanism describes 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 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

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

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

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

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 explicitly missing counterfactuals

Decomposition of the SDO

Decomposition of the SDO

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

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

Labor by design

- Hull, Kolesar and Walters (2022) tells the economic history of causal inference which is rooted in the Princeton Industrial Relations Section in the 70s and 80s responding to an “empirical crisis”
- The core problem in causal inference is *always* selection bias due to the fact that we cannot observe each units’ counterfactual
- Historically two ways that this selection bias was addressed: modeling it directly (Heckman, and others) and by design
- Design is the experimental design rooted in Fisher and Neyman’s work in the 20s, but the Section took it as a guiding principle even outside of the lab

Selection bias

- Selection bias in surveys has to do with non-representative samples; that is not its meaning in causal inference
- Selection bias in causal inference is when the average potential outcome for treatment group and control group are different in a large sample
- Source of the bias is the treatment assignment *mechanism*
- 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"

Bias #1: Selection bias

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

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

- Interpretation
  1. When you say you chose something because you'd be unhappy if you didn't, that's selection on Y^0
  2. When you say you chose something because you'd be happy if you did it, that's selection on Y^1
  3. When you say you chose something because you'll be more happy if you do it than if you didn't do it, that's selection on $Y^1 - Y^0$
- All three of those create selection bias because all three assign treatments to units mechanically based on potential outcomes, which is always the source of selection bias

Humans cause selection bias, not statistical model

- Paradox: the better our programs get, the better that humans get at doing things, the more interventions are targeting units based on treatment gains, the less reliable correlations are
- Paradox: the less efficient our programs are, the more erratically human are at assigning interventions, the more reliable correlations are
- The former creates strong selection bias; the latter is more like a randomized experiment

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 (pass out handout)

Roadmap

Foundational ideas

Brief (and selected) history of causal inference

Independence and Selection Bias

Industry example of RCT: eBay advertising

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 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 $ATT = ATU$.

What about partial independence?

- What if selection into treatment is independent of Y^0 but not Y^1
- On a piece of paper, write down the independence terms and substitute it into the decomposition
- What is the nature of the bias? What is eliminated?

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**: ***s****table*
  2. **U**: across all ***u****nits*, or the population
  3. **TV**: ***t****reatment-value* (“treatment effect”, “causal effect”)
  4. **A**: ***a****ssumption*
- Largely about 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)

Roadmap

Foundational ideas

Brief (and selected) history of causal inference

Independence and Selection Bias

Industry example of RCT: eBay advertising

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

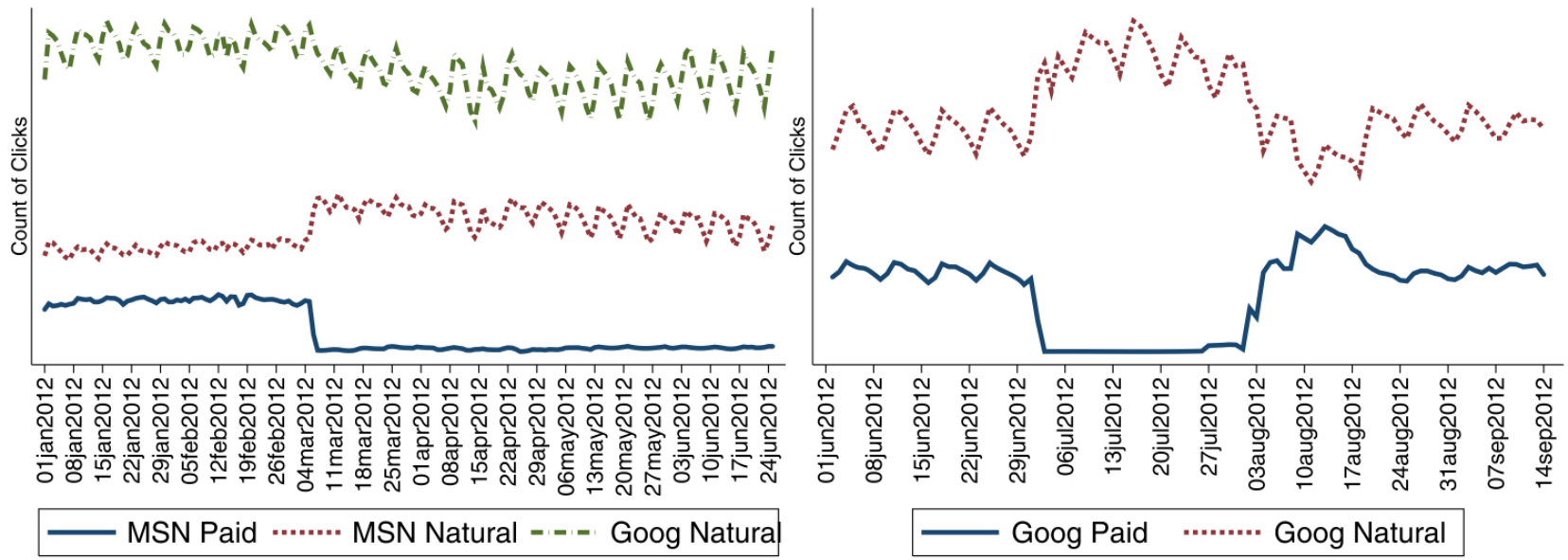
- Treatment was targeted ads at particular people conducting particular types of keyword search
- Consumers who choose to click on ads are loyal and already informed about products with high likelihood to buy already
- Problem is ads are targeting people at the end of their search, so the question is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

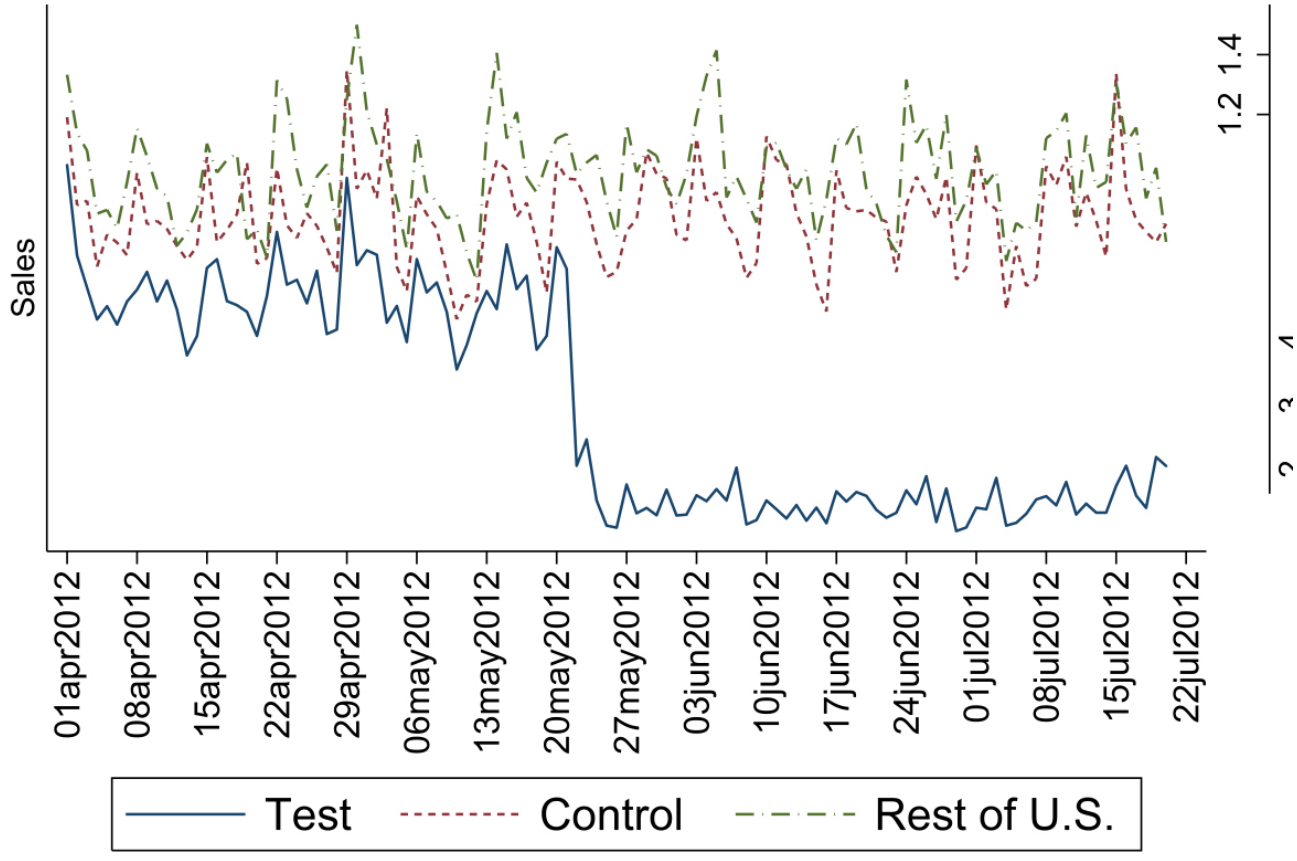
RCT

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

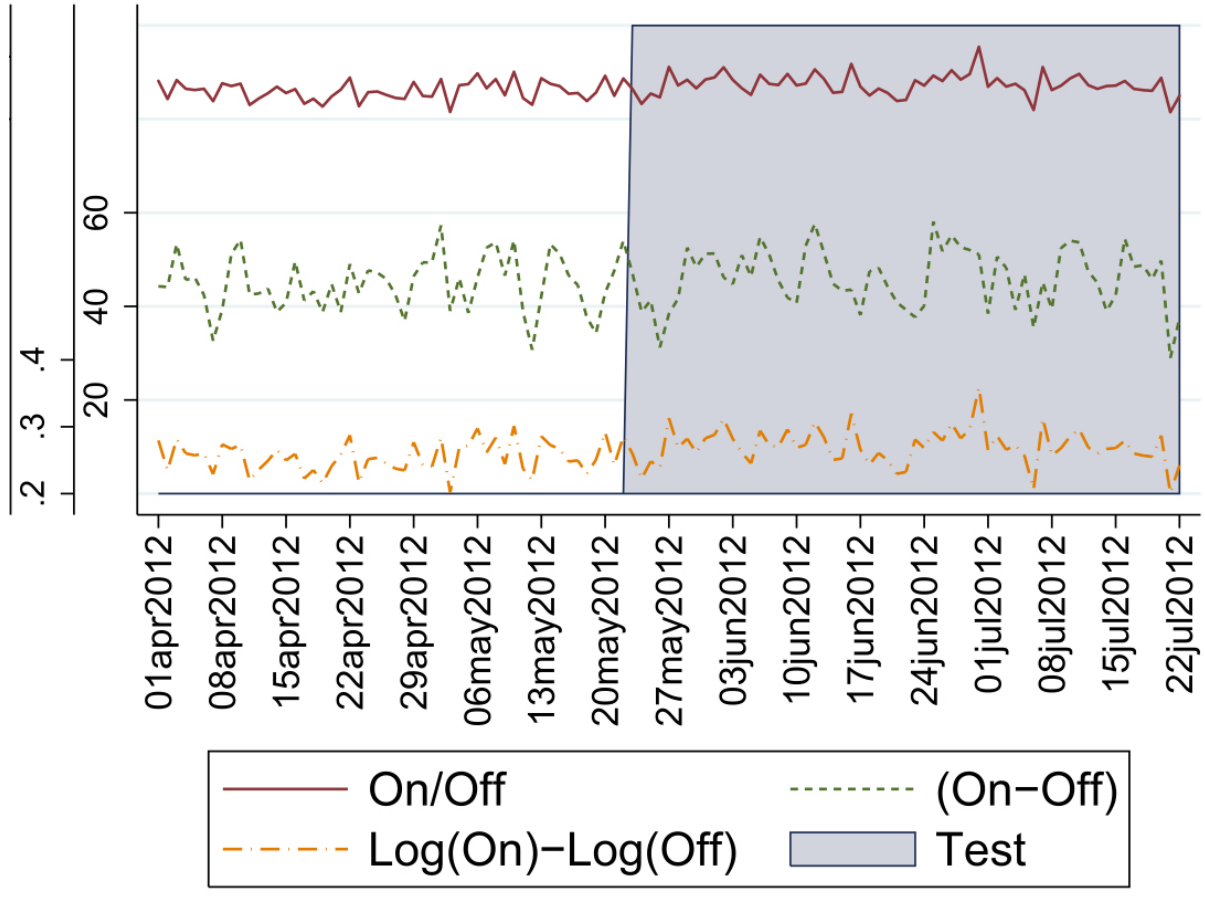
Design of the experiment

- 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%

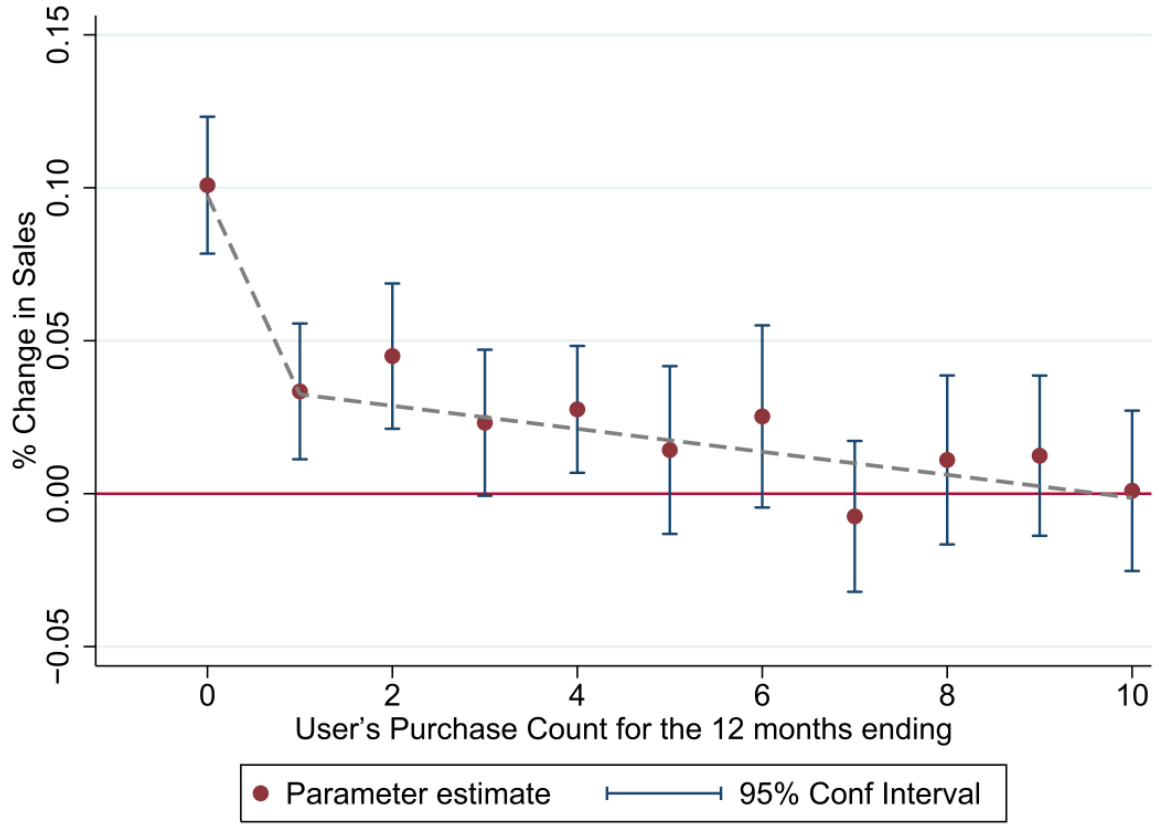
Figure: Spending effect on revenue using OLS but not the randomization. Effects are gigantic.

	(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.

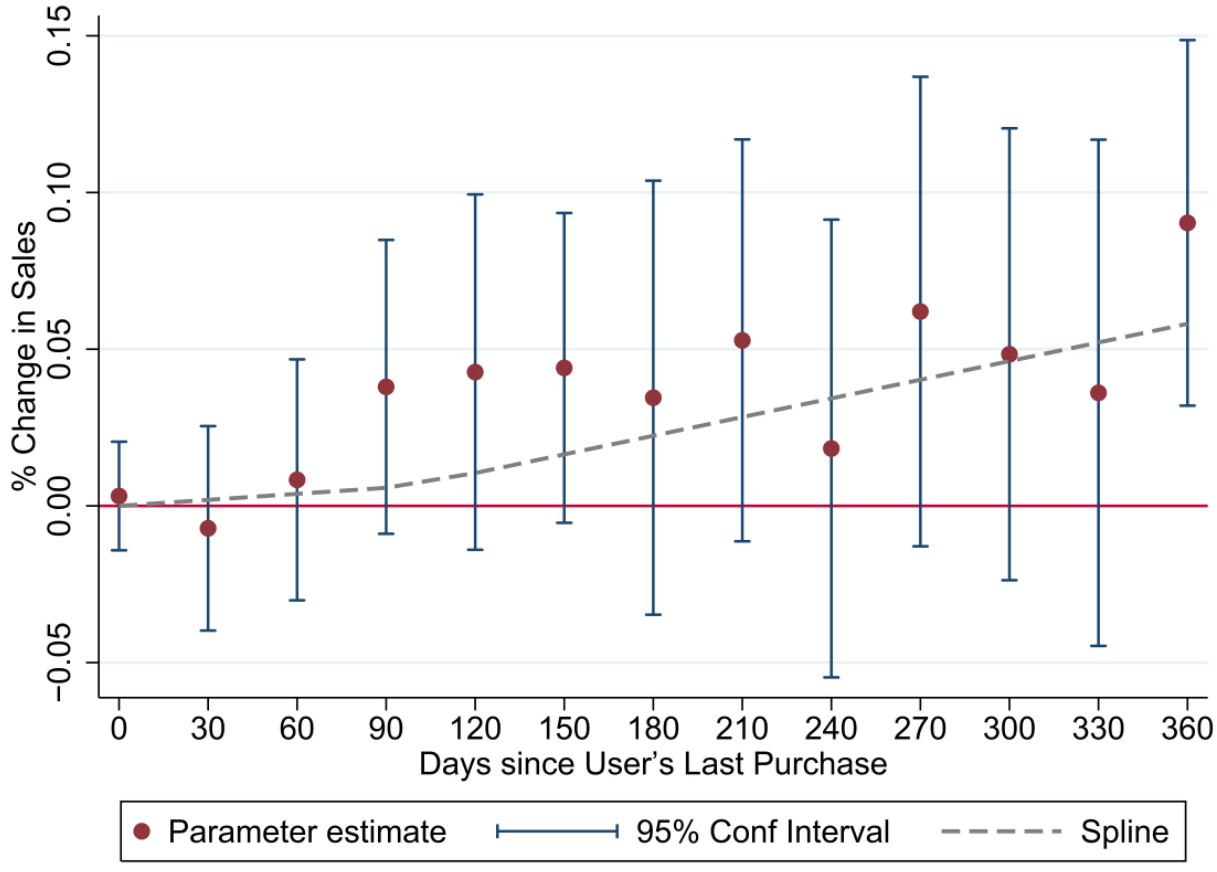
Heterogenous treatment effects

- 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

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