

# Causal Inference I

MIXTAPE SESSION

---



# Roadmap

Introduction to course

Managing expectations

History of causal inference

Potential outcomes

Naive correlation

Potential outcomes notation

Selection bias

Independence

Example of physical experimentation: eBay advertising

# Welcome!

- Scott Cunningham, professor of economics at Baylor University, author of Causal Inference: the Mixtape, applied microeconomist
- Causal inference is a sub-category within econometrics most often associated with “treatment effects”, as opposed to economic theory
- Workshops can be helpful ways to plug into one’s methodological training as the field has become very impactful (arguably won the Nobel Prize last October)

# Workshop Goals: Potential Outcomes and Synthetic Control

1. **Confidence:** You will feel like more comfortable with the **potential outcomes model** and **synthetic control estimator**
2. **Comprehension:** You will have learned more about the unique features of each of these, such as imputation of counterfactuals using selected control groups
3. **Competency:** You will know have code to tinker with and study

# 4-hour Workshop

- Goals are modest: cover potential outcomes, cover synthetic control
- Discuss possible extensions using machine learning methods

# Credibility Problem Within Empirical Economics

- Some history can help sometimes understand where all this comes from
- Tremendous turmoil in applied economics in late 1970s and early 1980s
- Several high profile papers found serious credibility problems in many applied studies, particularly those using macro data
- This criticism eventually turned to labor economics (Lalonde 1985; Ashenfelter and Card 1985)

## Lewis 1986 book

*"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. This criticism was particularly salient because many of the new methods had been tested initially on the union wage effect question (e.g., Lee 1978)." – Card (2022)*

## Replication problems

- Deward, Thursby and Anderson (1986) suggest that applied estimates can't be replicated partly bc many authors won't share their data or their programs
- They tried to replicate the papers at the JMCB journal between 1980 and 1984 (there was a data sharing agreement in 1982 and 1/4 authors still won't respond to request; before that 2/3 failed to respond)
- Only 2 of 9 papers could be reproduced exactly; 5 had substantial errors

## Take the con out of econometrics

- Famous paper by Ed Leamer – “Take the con out of econometrics”
- He says in Hendry, Leamer and Poirier (1990), “we don’t take empirical work seriously in economics. It’s not the source by which economists accumulate their opinions, by and large.”
- Hits hard at Princeton’s Industrial Relations Section where Card, Ashenfelter, Krueger, Angrist, Lalonde, and more are. They seek to fix this

## Princeton IRS Model

- Ashenfelter 1970s work in government focuses on extreme self-selection into job trainings programs ("Ashenfelter dip"); Ashenfelter and Card (1985) note longitudinal studies are not designed for this
- Princeton Industrial Relations Section Model: more transparency, more credible "research design" that make explicit sources of identification and work hard to verify the legitimacy of that source
- Krueger notes that he would read NEJM and they'd often include a short description of the study's "research design"

# Natural experiments

- Richard Freeman had been pushing for natural experiments for years
  - “big shocks” like federal minimum wages in Puerto Rico
- Card really pushes this early on
  - Immigration labor supply shifts from the Mariel Boatlift paper (Card 1990);
  - Minimum wage increases from neighboring state comparisons (Card and Krueger 1994)
- Steven Levitt seems instrumental in the late 1990s for bringing natural experiments and this labor economics approach to crime, more or less cleaning house; also very good abt sharing data and programs

# Today

- World is very different – naive correlational analysis is more quickly identified and methods have been developed moving past it
- Strong focus on the idea of the counterfactual being approximated with “good comparisons”
- Good comparisons are replacements for theoretical counterfactuals which are impossible to observe
- Key point: causal inference estimates counterfactuals which is a kind of prediction problem, but a very different type of one

# Roadmap

Introduction to course

Managing expectations

History of causal inference

Potential outcomes

Naive correlation

Potential outcomes notation

Selection bias

Independence

Example of physical experimentation: eBay advertising

# Introduction to Counterfactuals

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

# #1: Correlation and causality are different

Causal is one unit, correlation is many units

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

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

## #2: Coming first may not mean causality!

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



### #3: No correlation does not mean no causality!

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

# What is causality?

*"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 (philosopher)*

Key idea → counterfactual. Counterfactuals are neither past nor future. They are alternative histories created by thought experiments but we use them as framing devices to decipher causality in our timeline

Causal inference is also fundamentally practical because if we know causality, then we can also know not only the past, but also the future (policy)

# Different types of 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

# Identification problem

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

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

# Potential outcomes

- Conceptual framework for design based causal inference is counterfactual reasoning
- Counterfactual modeling becomes linked to potential outcomes models with Don Rubin 1970s work on propensity scores
- Potential outcomes notation was created by Jerzy Neyman (1923) and led Ronald Fisher (1925) to suggest RCTs
- Huge push by Donald Rubin, a Neyman “grandson”, in the 1970s and 1980s with Paul Rosenbaum on the propensity score
- Guido Imbens, a Rubin coauthor, notes that it was crucial to his work on IV with Angrist, but it may even have made that work more appealing *outside economics*

<https://youtu.be/cm8V65AS5iU?t=1097>

## Potential outcomes notation

- Let the treatment be a binary variable:

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

where  $i$  indexes an individual observation, such as a person

## Potential outcomes notation

- Potential outcomes:

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

where  $j$  indexes a counterfactual state of the world

## Realized vs potential outcomes

- Potential outcomes  $Y^1$  are not the realized outcomes  $Y$  either conceptually or notationally
- Potential outcomes are hypothetical states of the world but realized outcomes are *ex post* the observed outcomes we have in our data due to treatment assignment
- This distinction is subtle, creates challenges at the introductory stage, it isn't how econometrics was historically taught, except for at the very beginning; again Imbens on this

<https://youtu.be/cm8V65AS5iU?t=1175>

# Important definitions

## Definition 1: Individual treatment effect

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

## Definition 3: Fundamental problem of causal inference

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

## Definition 2: Switching equation

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

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

# Missing data problem

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

# Average Treatment Effects

## Definition 4: Average treatment effect (ATE)

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

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

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

# Conditional Average Treatment Effects

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

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

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

Cannot be calculated because  $Y_i^1$  and  $Y_i^0$  do not exist *for the same unit i* due to switching equation. This is the parameter we will focus on today because synthetic control *only* can be used to estimate this one.

## Good and bad variation

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

## Definition 7: Simple difference in mean outcomes (SDO)

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

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

I tend to use expectation operators  $E[\cdot]$  but note we are using samples  $E_N[\cdot]$

# SDO

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

# Potentially biased comparisons

## Decomposition of the SDO

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

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

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

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

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

Substitute letters for expectations

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

Rewrite ATE

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

## Move SDO terms to LHS

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

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

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

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

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

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

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

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

Rewrite from previous slide

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

Substitute conditional means

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

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

## Decomposition of difference in means

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

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

## Selection bias

- Notice this term “selection bias”

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

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

## Goal of causal inference

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

This imputation process happens sometimes explicitly (nearest neighbor matching) and sometimes implicitly (RCTs)

Let's look what happens in an RCT *and why* this addresses selection bias term  $E[Y^0|D = 1]$  and  $E[Y^0|D = 0]$

# 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:

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

$$SDO = ATE$$

# SUTVA

- Potential outcomes model places a limit on what we can measure:  
the “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 spillovers, poorly defined treatments and scale

## SUTVA: No spillovers to other units

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

## SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable” (i.e., notice that the super scripts contain either 0 or 1, not 0.55, 0.27)
- If we are estimating the effect of aspirin on headaches, we assume treatment is 200mg per person in the treatment
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*; you may have to think of it as multiple arms

# SUTVA: Scale can affect stability of treatment effects

Easier to imagine this with a different example.

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

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

BY THOMAS BLAKE, CHRIS NOSKO, AND STEVEN TADELIS<sup>1</sup>

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.<sup>2</sup>

# 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

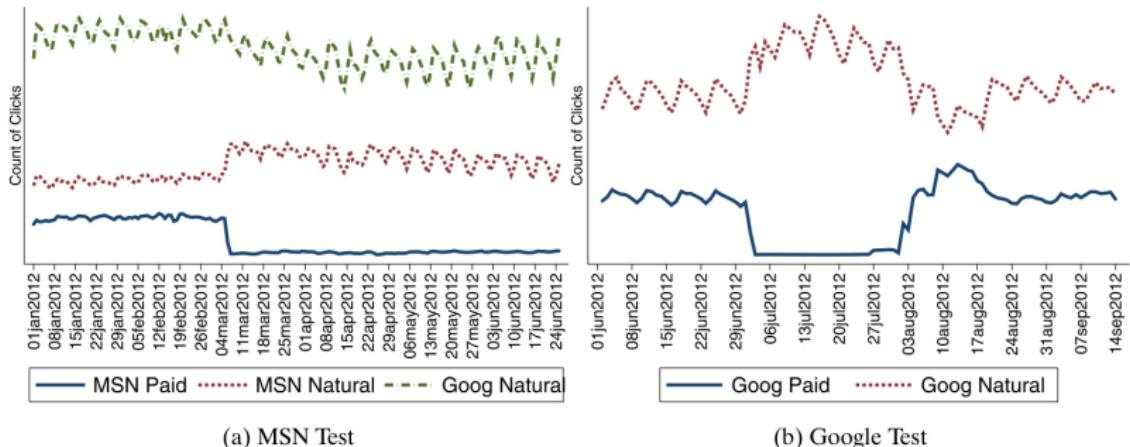


FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for ‘ebay’ terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

## 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."*

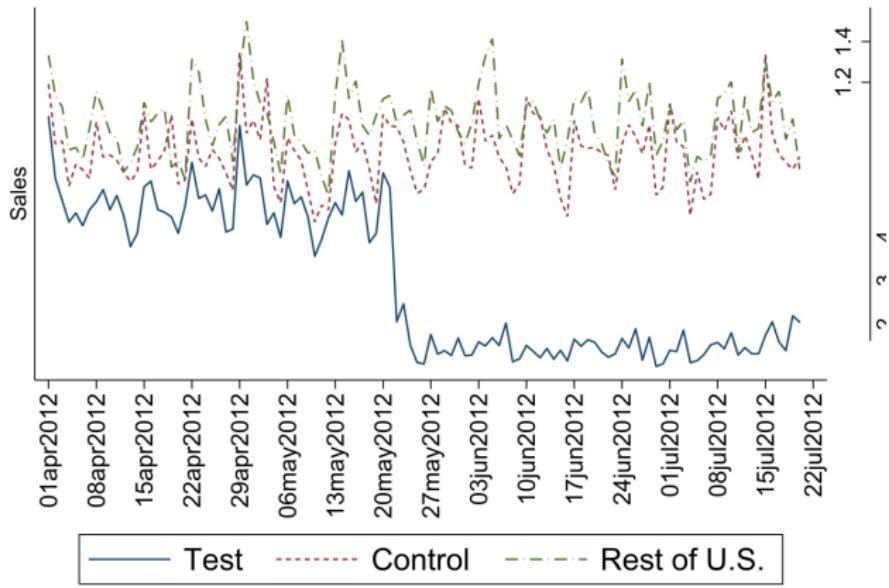
## RCT

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

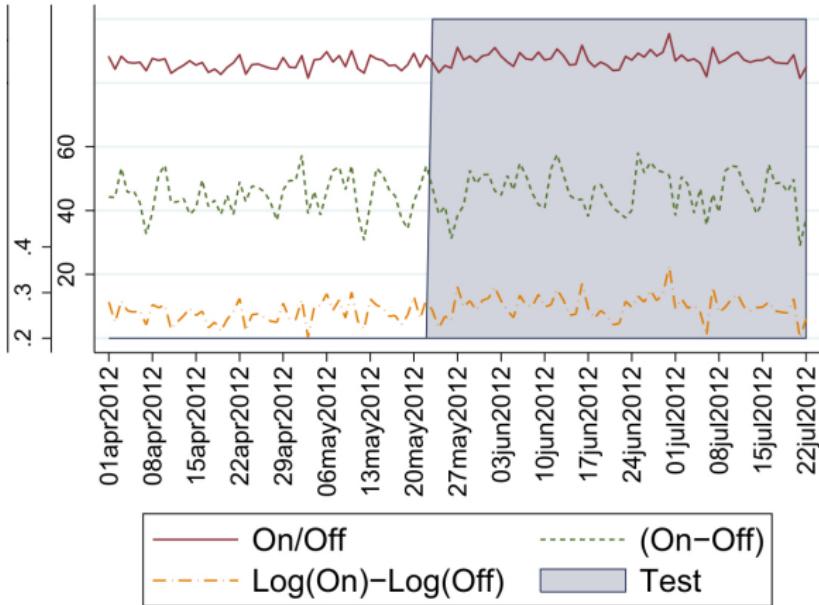
## Design of the experiment

- 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
$\Delta \ln(Rev)$ ( $\beta$ )	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%

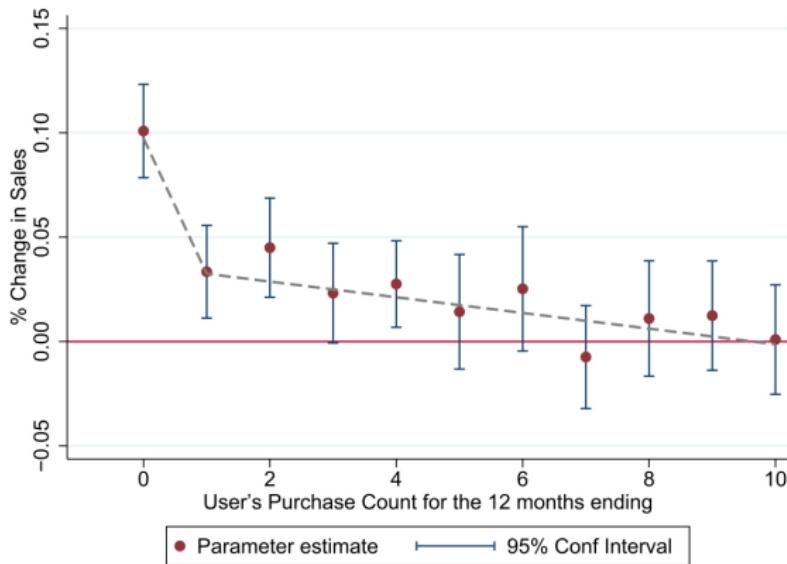
*Figure:* Spending effect on revenue using OLS but not the randomization.  
 Effects are gigantic.

	(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
$\Delta \ln(Rev)$ ( $\beta$ )	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.

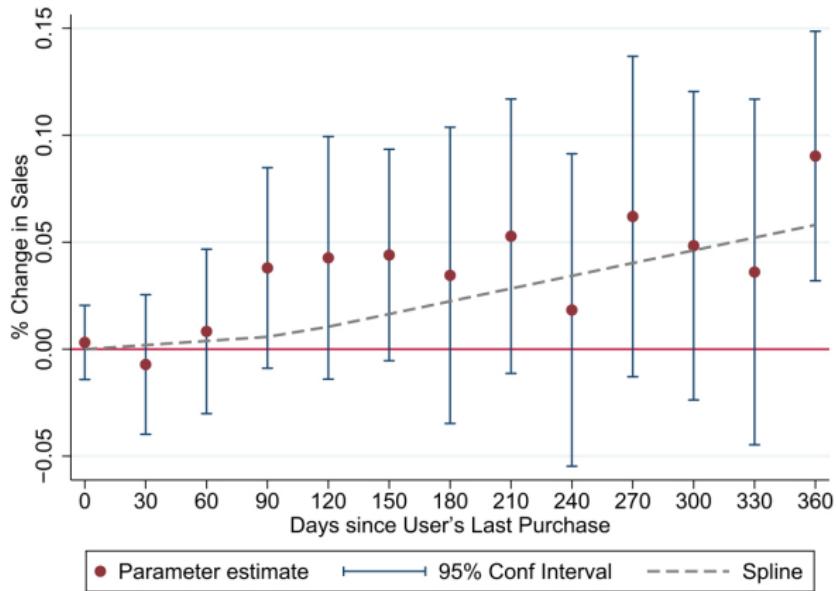
## Heterogenous treatment effects

- 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

- OLS was biased, not because OLS was bad because the comparisons were bad (under randomization, OLS is fine)
- 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

## Next: Synthetic control

- Very hard to imagine a situation where randomization wouldn't help us solve this identification problem, if not completely solve it altogether
- But unfortunately, many problems we simply cannot overcome with randomization – not because it's scientifically undesirable, but *because no one will let us randomize!*
- Next we will look at a particular method that attempts to address the problem of causal inference by finding a set of peers that approximate the counterfactual – the synthetic control