

Causal Inference II

MIXTAPE SESSION



Roadmap

In Pursuit of the ATT

Potential Outcomes

Independence and Selection Bias

Unconfoundedness and Ignorable Treatment Assignment

Exact and Inexact Matching

Saturated Regressions

Synthetic control

Interpolation with non-negative weighting

Extrapolation with Conservative Negative Weighting

Difference-in-differences

Four averages and three subtractions

Covariates

Model Misspecification

Alternatives to TWFE

Conclusion

Welcome!

- Scott Cunningham, professor of economics at Baylor University, author of Causal Inference: the Mixtape
- Identifying the causal parameter, ATT, using unconfoundedness and synthetic control (maybe diff-in-diff)
- Deep dive into the potential outcomes framework

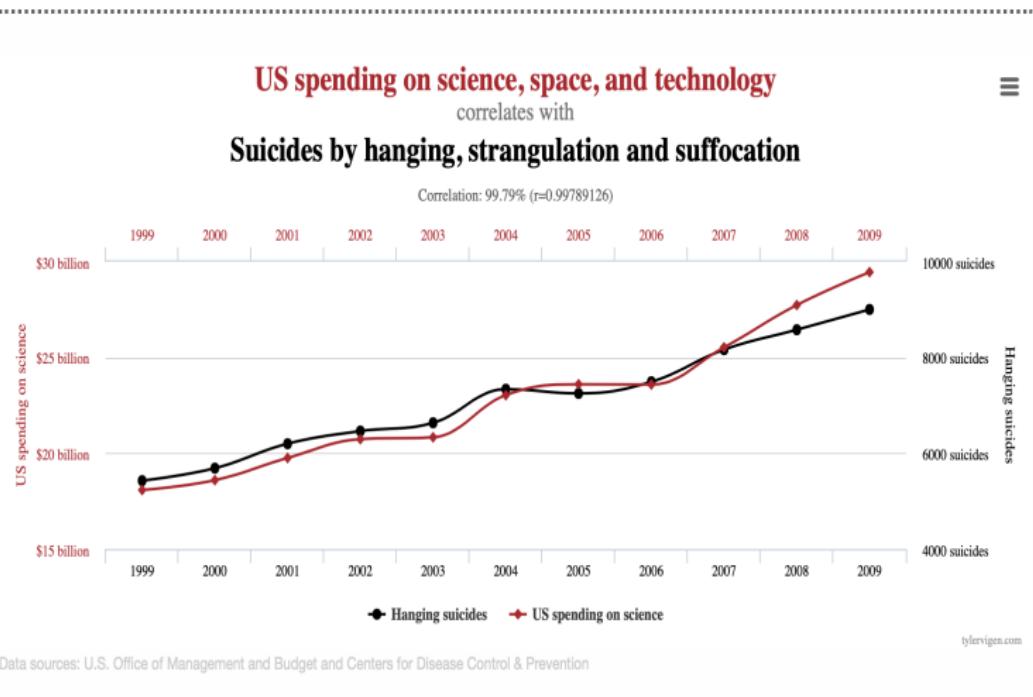
What my pedagogy is like

- Long day that don't feel long because it's high energy, with regular breaks including lunch
- Move between the econometrics, history of thought, videos, applications, code, spreadsheets, exercises
- Ask questions at any point; I'll do my best to answer them

Outline

1. Potential outcomes and the ATT
2. Selection Bias, Treatment Assignment and Randomization
3. Unconfoundedness, Regression, Matching and Propensity Score Weighting
4. Synthetic control with and without negative weighting
5. Difference-in-differences, covariates and staggered roll out (if time, but likely won't be)

Mandatory Spurious Correlation Slide



What is Causality?

- Causal inference has many mothers and fathers
- Aristotle was different than Hume, Mill, and Lewis – I am not comfortable saying to Aristotle he was wrong
- What I can do is explain the potential outcomes framework of causal inference and use it to discuss methods and tools that estimate parameters consistent with that view
- “All models are wrong but some are useful” – George Box

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

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

#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? And which causal effect is it? And why should we believe it is? 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 one

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 Nearly Killed Causal Inference

Mill's counterfactuals were immensely valuable though for the clarity of the definition of causality, but it came at a huge price – it made the knowledge impossible because of missing information

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

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:

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

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

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*.

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
- The aggregate causal parameters are *definitions* of summaries but cannot be calculated directly bc of missing data problem
- But they can be estimated, which is probably a distinction in epistemology as it's knowledge but of a different type ("warranted belief")

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

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

Note: this is a *written* formula for the definition of the ATE and so is *always* true. Also, notice that 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

- Eliminating selection bias requires understanding the selection mechanism – why did units end up treated but not others?
- Sorting into treatment based on potential outcomes always implies selection bias
 1. I chose to get a PhD because I thought I would be less happy without it – i.e., Y^0 maybe was lower for me than others
 2. I chose to get a PhD because I thought it would make me happier – i.e., Y^1 maybe was higher for me than others
 3. I chose to get a PhD because treatment effects were positive –
 $\delta = Y^1 - Y^0$
- More rational, more efficient, our decision making processes, the worse the bias gets!

Illustrating selection bias with spreadsheets

- Patients come to the Perfect Doctor who knows each person's treatment effects and assigns treatment based on it
- 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:

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

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 $ATT = ATU$.

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)

Roadmap

In Pursuit of the ATT

- Potential Outcomes

- Independence and Selection Bias

Unconfoundedness and Ignorable Treatment Assignment

- Exact and Inexact Matching

- Saturated Regressions

Synthetic control

- Interpolation with non-negative weighting

- Extrapolation with Conservative Negative Weighting

Difference-in-differences

- Four averages and three subtractions

- Covariates

- Model Misspecification

- Alternatives to TWFE

Conclusion

Self selection based on gains

- Rational actors almost by definition are thought to “self-select into treatment” making non-designed comparisons potentially misleading
- RCTs can overcome the selection bias that emerges when independence is violated, but are there other ways too?
- What parameter will we pursue? Consider the ATT

ATE vs ATT

- The RCT will identify the ATE, but under randomization, the $ATE=ATT=ATU$
- Outside the RCT, to identify the ATE requires a stronger set of assumptions, but it's not just about the assumptions
- Questions is which parameter is it you actually do want to know for decision making?

ATE vs ATT

- Pfizer wants to vaccinate the world – since everyone will get the treatment, they want to know the ATE
- If the goal ultimately is to implement something for every person, usually the ATE is the parameter you want to know (e.g., vaccines)
- But there is also the ATT – the returns to the program for those people on the program (e.g., ventilators will only be given to compromised people not all people)

Adjusting for variables

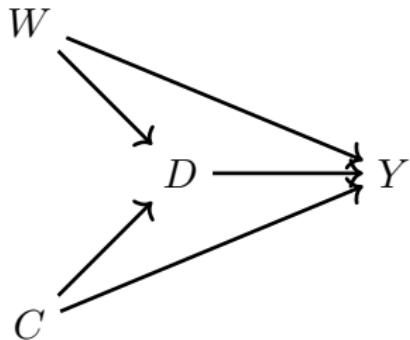
- One of the first things you learn in a methods course is multivariate regression “controlling for X ”
- What is this? Why do we do this? What should X be? What causal parameter does it help identify?
- Unconfoundedness, selection on observables, ignorable treatment assignment are different terms describing the same thing – the RCT is still occurring, only within the dimensions of a conditioning set of confounders and covariates

Which covariates?

- One of the values of causal directed acyclic graphs (DAG) is it allows you to formally select variables needed for covariate adjustment
- One such approach is the backdoor criterion which states that if you can condition on X such that all backdoor paths close, then you can identify some aggregate causal parameter
- But this requires a model, and I don't mean a theoretical model that you might learn as some abstract theory about education
- It's a model of treatment assignment, which is local in nature, and when it occurs outside the RCT requires expert knowledge

Simple DAG

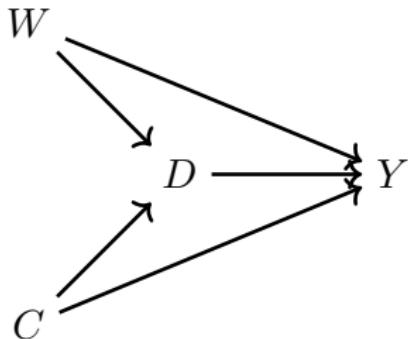
Figure: A simple DAG illustrating selection on observables.



Write down all paths, both direct from D to Y and indirect or “backdoor paths”

Simple DAG

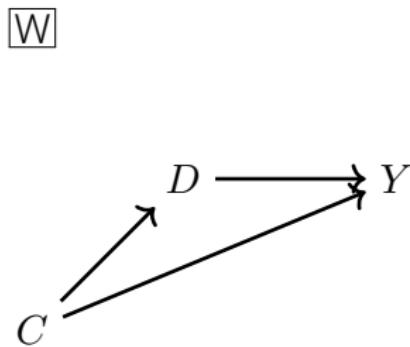
Figure: A simple DAG illustrating selection on observables.



1. $D \rightarrow Y$, the direct edge representing a causal effect with associated causal parameter like the ATE, ATT, etc.

Simple DAG

Figure: The same simple DAG illustrating selection on observables only with the direct edge from D to Y deleted and backdoor W blocked.



2. $D \leftarrow \boxed{W} \rightarrow Y$ is a backdoor from D to Y through W . **Block it**

Remaining variation after blocking

Figure: Visualization of Backdoor Criterion

[W]

$D \longrightarrow Y$

[C]

2. $D \leftarrow [W] \rightarrow Y$ is a backdoor from D to Y through W . **Block it**
3. $D \leftarrow [C] \rightarrow Y$ is a backdoor from D to Y through C . **Block it**

Definition of Known and Quantified Confounders

Definition of a Known and Quantified Confounder

Variable C is a *known* and *quantified confounders* if the researcher believes it causes units to select into treatment ($C \rightarrow D$) and also independently determine outcome Y , or $C \rightarrow Y$. Confounders are always known, which requires prior knowledge. And to be quantified, they must be correctly measured in your dataset.

Known and Quantified Confounder

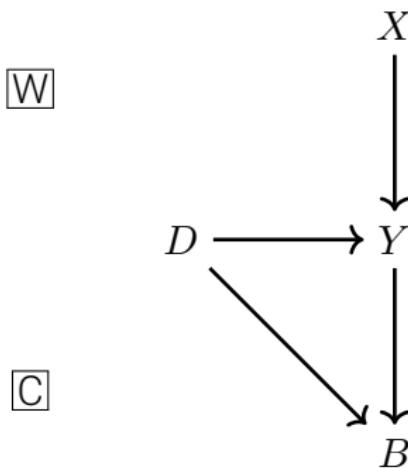
- Confounders may or may not be observed, but they must be known if they are confounders as confounders create backdoor paths from D to Y
- Visually, solid lines means they are “quantified” (i.e., in the data), whereas dashed lines mean they are either not defined correctly or not in the dataset (“unobserved”)
- Backdoor criterion is appropriate only for known and quantified confounders – if either known or quantified is missing, this material today is not to be used

DAG tells us what we need to condition on

- If we “block” on C and W , then the *only* explanation of why D and Y are then correlated is causal
- Depending on the model we estimate, and explicit assumptions made about potential outcomes, then we are able to identify an aggregate causal parameter
- We call C and W the “known and quantified confounders” because the model said these were necessary, they were observed (no dashed line) and they were confounders
- So what’s a collider, and what’s a covariate? Let’s now add those into the simple DAG

Modification of the original DAG

Figure: A DAG illustrating confounders (W and C) versus colliders (B) versus exogenous covariates (X).



4. You cannot get from D to Y via X so it is not a backdoor path

Covariate

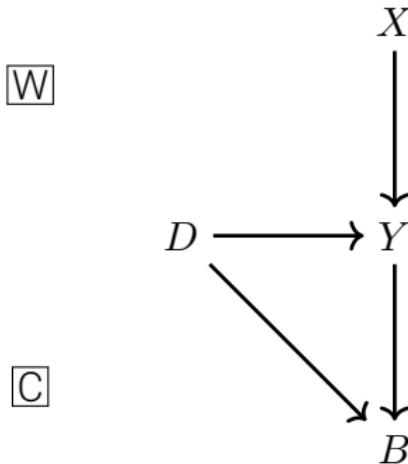
Definition of a Covariate

Variable X is a covariate if it causes Y but does not cause the treatment status D .

- Think of it as in the error term, but not correlated with the treatment variable
- Including X in a model can increase precision of estimates of D on Y simply by reducing residual variance, but should have no effect on point estimates
- Keep “confounder” and “covariate” distinct
- Covariates can be time invariant or change over the time – that’s not relevant

Modification of the original DAG

Figure: A DAG illustrating confounders (W and C) versus colliders (B) versus exogenous covariates (X).



5. You cannot get from D to Y via B so it is a collider, but if you control for it, that path opens up and introduces selection bias ("bad controls")

Colliders

Definition of a Collider

Variable B is a collider if there exists $D \rightarrow B \leftarrow Y$ along the path from D to Y .

- Colliders block backdoor paths so long as they are not blocked
- If you block on a collider, then the backdoor path opens, unless there exists a non-collider that you block to close it
- Conditioning on a collider introduces selection bias and depending on the magnitudes of $D \rightarrow B$ and $B \leftarrow Y$ relative to $D \rightarrow Y$, the distortion of estimated effect of D on Y may be extreme

Summarizing “which variables”

- Comparing treatment and control group of exactly the same values of known and quantified confounders will allow you to estimate aggregate causal parameters of interest
- Covariates can improve precision but do not reduce bias
- Colliders must be left alone, otherwise they introduce bias unless another non-collider can block them

Contrast this with ordinary practices

- Person attempts to “control for omitted variable bias” by including as many “controls” as possible
- Person does not even attempt to think about treatment assignment mechanism and therefore has no idea what variables are colliders, covariates or confounders
- Big data approaches to covariate adjustment is *very dangerous* – colliders introduce bias and without a model, there is no way you know what those are

Ad Hoc

- Short of an outright DAG, then the thing to be thinking about is this:
 - What set of covariates are highly predictive of Y^0 ?
 - What set of covariates are highly predictive of D ?
 - Are these covariates distributed enough across both treatment and control?
- This is more of a hunch approach, but at least it's based on reasoning through the treatment assignment mechanism as opposed to "kitchen sink regressions"

Identifying assumption I: Unconfoundedness

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. There exists a set X of known and quantified confounders such that after adjusting for them, treatment assignment is *independent of potential outcomes*.

- Conditional on X , treatment assignment is **random**
- For a large group of people within the same strata, they flipped coins as opposed to sought treatments that helped them
- Eliminating all backdoor paths on a DAG through blocking satisfies unconfoundedness; also called ignorability

Identifying assumption I: Unconfoundedness

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. There exists a set X of known and quantified confounders such that after adjusting for them, treatment assignment is *independent of potential outcomes*.

$$\begin{aligned} E[Y^0 | D = 1, X = x] &= E[Y^0 | D = 0, X = x] \\ E[Y^1 | D = 1, X = x] &= E[Y^1 | D = 0, X = x] \end{aligned}$$

Unconfoundedness justifies substituting units in treatment for control based on $X = x$ – but only if there are exact matches (next slide)

Identifying assumption II: Common support

For ranges of X , there is a positive probability of being both treated and untreated

- There exists units in treatment and control with same values of X – you can't make the substitutions otherwise
- Dimension k means every specific combination of the conditioning set (e.g., not males and old, but adult males, adult females, youth male, youth female)
- Testable because common support is observable unlike unconfoundedness, but as you can imagine if the dimensions of X gets large (and with a continuous covariate it's infinite!) then it won't hold in any finite sample!

Assumptions combined

But if we have them both (represented below), we can even outside of an RCT estimate the ATE through nonparametric matching

1. $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (strong unconfoundedness)
2. $0 < Pr(D = 1|X) < 1$ with probability one (common support)

Comparing groups of individuals who have the same values of X , treatment is no longer based gains, δ .

The second term implies we have people in treatment and control for every strata of X

Implications of assumptions

- Assumption 1 lets you plug Y for Y^j with the switching equation

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

- Assumption 2 lets you weight over the covariate distribution

$$\begin{aligned} \delta_{ATE} &= E[Y^1 - Y^0] = E\left[E[Y^1 - Y^0 | X]\right] \\ &= \int E[Y^1 - Y^0 | X, D = 1] dPr(X) \\ &= \int (E[Y | X, D = 1] - E[Y | X, D = 0]) dPr(X) \end{aligned}$$

In Defense of the ATT

If we want the ATE, we need strong unconfoundedness and strong overlap

But if we want the ATT, we can go with strictly weaker assumptions

ATT requires weak unconfoundedness and a weaker common support condition

ATT Identification

We can modify those assumptions and weaken both which helps a lot

1. $Y^0 \perp\!\!\!\perp D|X$ (weak unconfoundedness)
2. $Pr(D = 1|X) < 1$ (with $Pr(D = 1) > 0$) (weak support)

We don't need full common support because we don't need to find counterfactuals for the control group – we only need units in the control group that match with our treatment group

Selection is weaker too, like I said – they are not entirely irrational, but who knows if it helps you

Summarizing

Weighted averages under both assumptions:

$$\delta_{ATT} = \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X|D = 1)$$

We match units in treatment and control because under weak unconfoundedness they're substitutable, and we use weak common support so that we can actually do it, then we take weighted averages over the differences.

Exact matching



Quote from Imbens and Rubin 2015

"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others."

Stratification weighting doesn't exactly always help you see the imputation – where you literally “fill in” missing counterfactuals either through estimating them or simply plug-in, but matching does

Matching will match a treated unit to a comparison unit that is similar (or in exact matching identical) on a known and quantified confounder

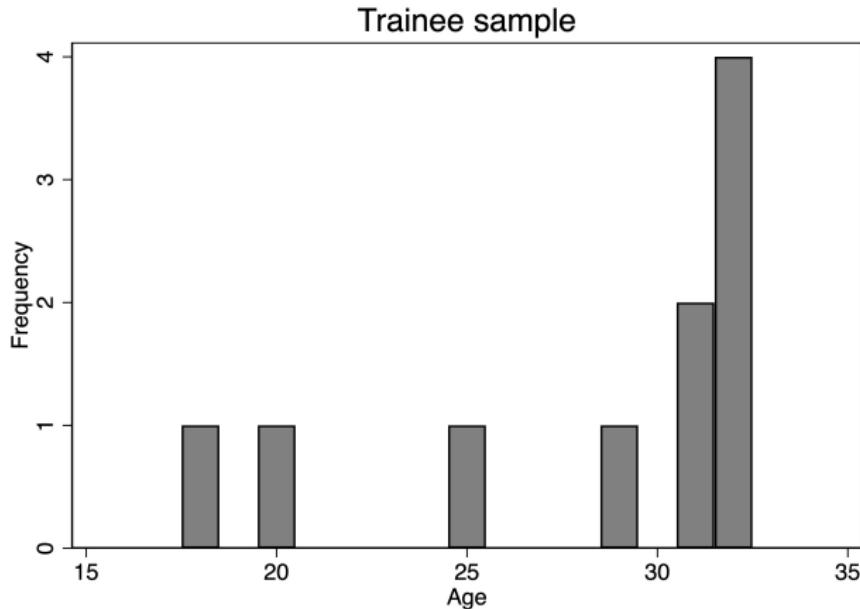
Training example (unmatched)

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	31	\$ 26,629	1	29	\$ 23,178
2	31	\$ 26,633	2	39	\$ 33,817
3	18	\$ 15,324	3	33	\$ 27,061
4	32	\$ 27,717	4	46	\$ 43,109
5	32	\$ 27,725	5	32	\$ 26,040
6	25	\$ 20,762	6	39	\$ 33,815
7	32	\$ 27,716	7	31	\$ 25,052
8	32	\$ 27,719	8	33	\$ 27,060
9	20	\$ 16,723	9	25	\$ 19,787
10	29	\$ 24,552	10	29	\$ 23,173
			11	27	21,416
			12	32	26,040
			13	20	16,246
			14	41	36,316
			15	18	15,046
			16	29	23,178
			17	49	47,559
			18	32	26,040
			19	27	21,418
			20	46	43,109
Mean	28.2	\$24,150	Mean	32.85	\$27,923

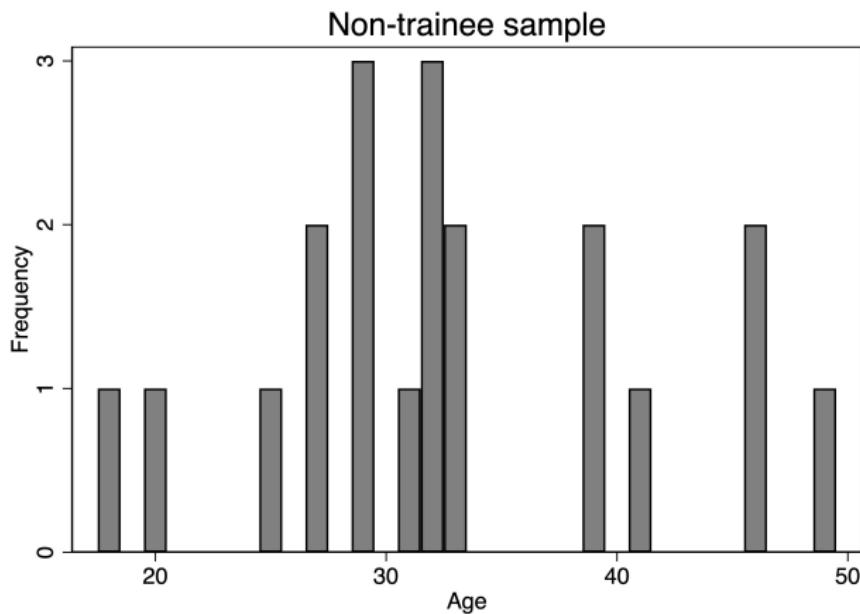
$$SDO = \$24,150 - 27,923 = -\$3,773$$

Age Imbalance

Figure: Age distribution of a job training program's trainees (figure a) versus a sample of workers who were not enrolled in the trainee program (figure b).



Age Imbalance



Exact matching

- Exact matching finds a person in the control group whose value of X_j is exactly equal to each person in the treatment group i
- Will not work if the conditioning set includes a continuous variable
- Will also not work if K gets large (curse of dimensionality)

ATT estimator

We will focus on the ATT for the rest of today and the equation is:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) \quad (1)$$

where $Y_{j(i)}$ is the j^{th} unit matched to the i^{th} unit based on the j^{th} being "exactly equal to" the i^{th} unit with respect to the X conditioning set

Number of matches

What if I find two or more M units with the identical X value? Then what?

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right) \quad (2)$$

Notice that we are only dealing with Y_i^0 by matching; The Y_i^1 is fine as is.

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, find all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.
2. For each unit i , replace its missing potential outcome, Y_i^0 , with the matched j units' realized outcomes, $\frac{1}{M} \sum Y_{j(i)}$, from Step 1. Do this for all i units in the treatment group.
3. For each unit i , calculate the difference between realized earnings and matched earnings, $\hat{\delta}_i = Y_i - \frac{1}{M} \sum Y_{j(i)}$.
4. Finally, estimate the sample ATT by averaging over all i differences in earnings from Step 3 as $\frac{1}{N_T} \sum \hat{\delta}_i$, where N_T is the number of treatment units.

Matched sample

Table: Training example with matched sample using exact matching

Trainees			Matched Sample		
Unit	Age	Earnings	Matched Unit	Age	Earnings
1	31	\$26,693	2	31	\$25,052
2	31	\$26,691	2	31	\$25,052
3	18	\$15,392	18	18	\$15,046
4	32	\$27,776	5	32	\$26,045
5	32	\$27,779	5	32	\$26,045
6	25	\$20,821	4	25	\$19,787
7	32	\$27,778	5	32	\$26,045
8	32	\$27,780	5	32	\$26,045
9	20	\$16,781	8	20	\$16,246
10	29	\$24,610	6	29	\$23,178
Mean	28.2	\$24,210	Mean	28.2	\$22,854

$$\widehat{ATT} = \$24,210 - \$22,854 = \$1,356$$

Estimated ATT using Exact Matching

Weak unconfoundedness of Y^0 with respect to age justified substituting one group for another

But matching bias still exists if you fail common support – unconfoundedness is necessary but not sufficient

Even weak support is rare due to the curse of dimensionality

Inexact matching



Curse of Dimensionality

- Assume we have k covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low, medium, high, etc.)
- The number of strata is 3^k . For $k = 10$, then it's $3^{10} = 59,049$
- Common support problems isn't about the covariates – you probably will find white people in treatment and control, women in both groups, college educated in both groups
- Common support is about white women with a college degree in both groups – it's about the dimensions which is why the curse grows fast

Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both, and that violates our second assumption
- We can always use “finer” classifications, but finer cells worsens the dimensional problem, so we don’t gain much from that. ex: using 10 variables and 5 categories for each, we get $5^{10} = 9,765,625$.
- Matching methods really force us to see these curses; they’re often hidden from OLS because OLS doesn’t tell us it is just doing various extrapolations
- Simple weighting methods is also a problem if the cells are “too coarse”

Propensity scores

- Propensity scores were developed by Rosenbaum and Rubin (1983) as a way of reducing the dimension of the conditioning set of X
- But it still requires common support – propensity scores don't solve the curse of dimensionality problem from the perspective of bias
- If you don't have exact matches in the dimensions of X then you'll still be matching units with similar propensity scores but it won't overcome the bias
- There are methods that will adjust the propensity score estimation which I'll briefly mention at the end

To Look Like Someone Else

- When we can make synthetic xerox copies of ourselves, that's exact matching
- But what if we can only make similar copies of ourselves, like fraternal, but not identical, twins? That's nearest neighbor matching – a form of "inexact matching", sort of like fraternal twins
- Introduces bias bc of inexact matching, but the magnitude of the bias depends on the severity of the discrepancy
- We can improve on nearest neighbor matching using bias adjustment (Abadie and Imbens 2011)

Nearest Neighbor Matching

- Estimate $\hat{\delta}_{ATT}$ by *imputing* the missing potential outcome of each treatment unit i using the observed outcome from that outcome's "nearest" neighbor j in the control set using X for the matching

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where $Y_{j(i)}$ is the observed outcome of a control unit such that $X_{j(i)}$ is the **closest** value to X_i among all of the control observations (eg match on X)

Matching

- We could also use the average observed outcome over M closest matches:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

- Works well when we can find good matches for each treatment group unit, so M is usually defined to be small (i.e., $M = 1$ or $M = 2$)

Matching example with single covariate

i	Y_i^1	Y_i^0	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Matching example with single covariate

i	Y_i^1	Y_i^0	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Match and plug in!

Matching example with single covariate

i	Y_i^1	Y_i^0	D_I	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

$$\widehat{\delta}_{ATT} = \frac{1}{3} \cdot (6 - 9) + \frac{1}{3} \cdot (1 - 0) + \frac{1}{3} \cdot (0 - 9) = -3.7$$

Measuring the matching discrepancy

- What does it mean to be close when I am working with a large number of covariates?
- What if we had a way of measuring a match in terms of how “close” each unit’s X_i value was to the matched X_j
- Let’s do that and use the square root of the sum of all squared differences in each unit’s $X_i - X_{j(i)}$ as a measure of how bad the match is
- This is called the Euclidean distance

Euclidean distance

Definition: Euclidean distance

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

Let's do this together – sometimes it helps to manually calculate this

https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing

Inexact matching: Random match 1

Table 32: Matching on two covariates at random (first attempt)

Trainee sample				Non-Trainees				Matched sample #1			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	4	39	1.76	12775
2	29	2.80	12250	2	27	1.78	10075	20	48	1.87	14800
3	24	3.92	11000	3	21	1.84	8725	12	36	1.70	12100
4	27	2.29	11750	4	39	1.76	12775	8	33	1.97	11425
5	33	2.50	13250	5	38	1.61	12550	1	20	1.89	8500
6	22	1.34	10500	6	29	1.74	10525	15	43	1.45	13675
7	19	1.66	9750	7	39	1.57	12775	18	30	1.86	9000
8	20	2.60	10000	8	33	1.97	11425	7	39	1.57	12775
9	21	1.94	10250	9	24	1.81	9400	3	21	1.84	8725
10	30	3.37	12500	10	30	2.02	10750	11	33	1.64	11425
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	34.2	1.76	\$11,520

Euclidean distance: 45.8.

Estimated ATT equals \$11,075 - \$11,520 = -\$445.

Inexact matching: Random match 2

Table 33: Matching on two covariates at random (second attempt)

Trainee sample				Non-Trainees				Matched sample #2			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	13	22	1.66	8950
2	29	2.80	12250	2	27	1.78	10075	5	38	1.61	12550
3	24	3.92	11000	3	21	1.84	8725	1	20	1.89	8500
4	27	2.29	11750	4	39	1.76	12775	20	48	1.87	14800
5	33	2.50	13250	5	38	1.61	12550	15	43	1.45	13675
6	22	1.34	10500	6	29	1.74	10525	9	24	1.81	9400
7	19	1.66	9750	7	39	1.57	12775	6	29	1.74	10525
8	20	2.60	10000	8	33	1.97	11425	17	19	1.86	8275
9	21	1.94	10250	9	24	1.81	9400	5	38	1.61	12550
10	30	3.37	12500	10	30	2.02	10750	18	30	1.86	9000
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	31	1.74	\$10,822.50

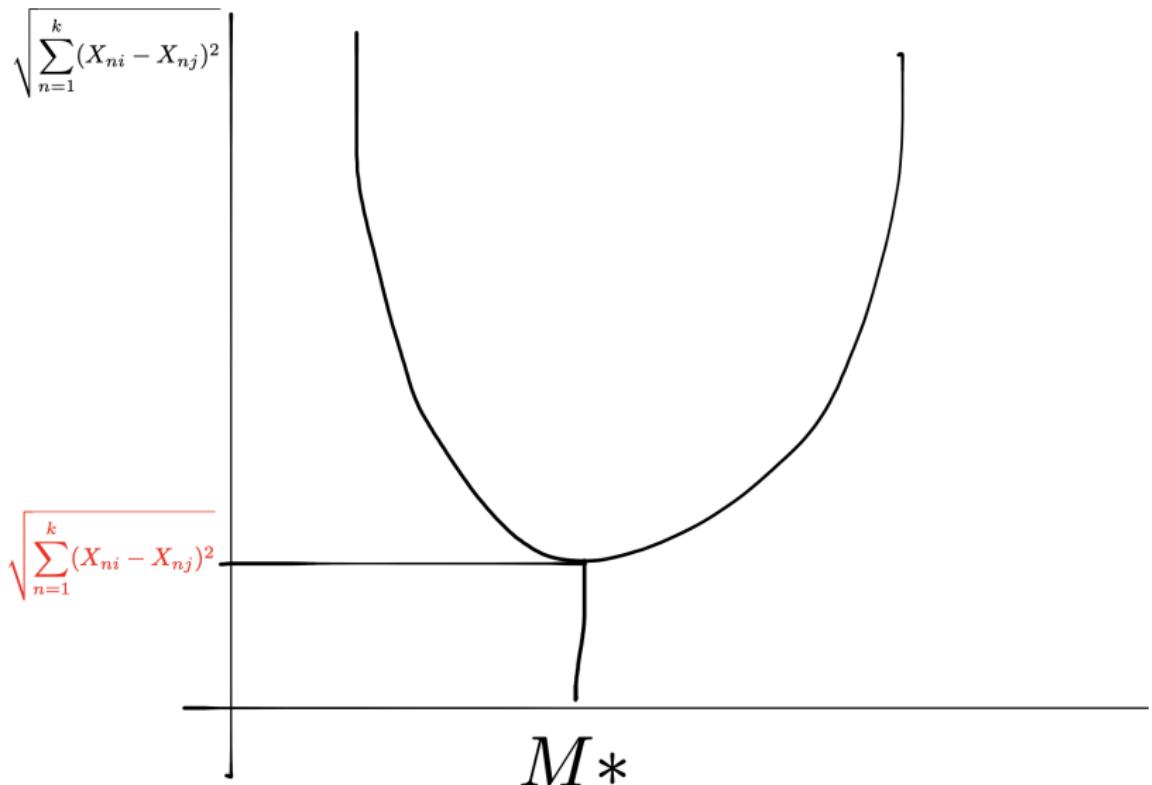
Euclidean distance: 32.53.

Estimated ATT equals $\$11,075 - \$10,822.50 = \$252.50$.

Minimizing the Euclidean distance

- Abadie and Imbens (2006) show that there exists a unique solution to the matching problem that minimizes a given distance metric
- **Matching** in R and **teffects** in Stata (not sure in python)
- But the idea here is that any other match will always have a higher Euclidean distance so I've drawn a picture!

Visualization of Optimal Match



Inexact matching by minimizing the Euclidean distance

Table 34: Matching on two covariates with minimized Euclidian distance

Trainee sample				Non-Trainees				Optimal Match			
Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings	Unit	Age	GPA	Earnings
1	18	1.28	9500	1	20	1.89	8500	14	18	1.89	8050
2	29	2.80	12250	2	27	1.78	10075	6	29	1.74	10525
3	24	3.92	11000	3	21	1.84	8725	9	24	1.81	9400
4	27	2.29	11750	4	39	1.76	12775	2	27	1.78	10075
5	33	2.50	13250	5	38	1.61	12550	8	33	1.97	11425
6	22	1.34	10500	6	29	1.74	10525	13	22	1.66	8950
7	19	1.66	9750	7	39	1.57	12775	17	19	1.86	8275
8	20	2.60	10000	8	33	1.97	11425	1	20	1.89	8500
9	21	1.94	10250	9	24	1.81	9400	3	21	1.84	8725
10	30	3.37	12500	10	30	2.02	10750	10	30	2.02	10750
				11	33	1.64	11425				
				12	36	1.70	12100				
				13	22	1.66	8950				
				14	18	1.89	8050				
				15	43	1.45	13675				
				16	39	1.88	12775				
				17	19	1.86	8275				
				18	30	1.86	9000				
				19	51	1.96	15475				
				20	48	1.87	14800				
Mean	24.3	2.37	\$11,075					Mean	24.3	1.85	\$9457.50

Minimized Euclidean distance: 3.00.

Estimated ATT* equals \$11,075 - \$9457.50 = \$1,607.50.

Other distance metrics

- Our example treated a one unit difference in age and one unit difference in GPA as the same, but those scales are different and matter a lot
- The Euclidean distance is not invariant to changes in the scale of the X 's.
- Alternative distance metrics that are invariant to changes in scale are more commonly used
- Normalized Euclidean distance and Mahalanobis distance both try to normalize it so that scale doesn't matter

Normalized Euclidean distance

Definition: Normalized Euclidean distance

A commonly used distance is the normalized Euclidean distance:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

where

$$\hat{V}^{-1} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \dots, \hat{\sigma}_k^2)$$

Normalized Euclidean distance

- Notice that the normalized Euclidean distance is equal to:

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}_n^2}}$$

- Thus, if there are changes in the scale of X_{ni} , these changes also affect $\hat{\sigma}_n^2$, and the normalized Euclidean distance does not change

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \widehat{\Sigma}_X^{-1} (X_i - X_j)}$$

where $\widehat{\Sigma}_X$ is the sample variance-covariance matrix of X .

Matching and the Curse of Dimensionality

- The larger the dimensions of the conditioning set, the less likely common support holds, and you can't not do it because you need these covariate dimensions to satisfy weak unconfoundedness!
- This problem is caused by the finite dataset, and it introduces a particular type of selection bias
- Curses are only overcome with new spells
- Abadie and Imbens (2011) derived a way to reduce the bias (bias adjustment or bias correction)

Deriving the matching bias

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where each i and $j(i)$ units are matched, $X_i \approx X_{j(i)}$ and $D_{j(i)} = 0$.

Define potential outcomes and switching eq.

$$\mu^0(x) = E[Y|X = x, D = 0] = E[Y^0|X = x],$$

$$\mu^1(x) = E[Y|X = x, D = 1] = E[Y^1|X = x],$$

$$Y_i = \mu^{D_i}(X_i) + \varepsilon_i$$

Deriving the matching bias

Substitute and distribute terms

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) \\ &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Deriving the matching bias

Difference between sample estimate and population parameter is:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT}) \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Algebraic manipulation and simplification:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_i) - \delta_{ATT}) \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} (\mu^0(X_i) - \mu^0(X_{j(i)})) .\end{aligned}$$

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$.

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\hat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\hat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\hat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D=1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\hat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D=1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly

Deriving the matching bias

Note $\hat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\hat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$ may not converge to zero and can be very large!

Deriving the matching bias

Note $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$ may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$ may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Deriving the matching bias

Note $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$ as $N \rightarrow \infty$. However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$ may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$ may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Bias is often an issue when we match in many dimensions

Solutions to matching bias problem

The bias of the matching estimator is caused by large matching discrepancies $\|X_i - X_{j(i)}\|$ which is virtually guaranteed by the curse of dimensionality. However:

1. But the matching discrepancies are observed. We can always check in the data how well we're matching the covariates.
2. For $\widehat{\delta}_{ATT}$ we can sometimes make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making N_C large).
3. If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques

Matching with bias correction

- Each treated observation contributes

$$\mu^0(X_i) - \mu^0(X_{j(i)})$$

to the bias.

- Bias-corrected (BC) matching:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[(Y_i - Y_{j(i)}) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})) \right]$$

where $\widehat{\mu^0}(X)$ is an estimate of $E[Y|X = x, D = 0]$. For example using OLS but other maybe too (neural nets?).

- Under some conditions, the bias correction eliminates the bias of the matching estimator without affecting the estimator's variance.

Steps

1. Regress Y on X with OLS except only use the control sample:

$$Y_j = \alpha + \beta X_j + \varepsilon_j$$

where j are the units for which $D_j = 0$.

Steps

2. Use the fitted values $\hat{\alpha}$ and $\hat{\beta}$ to predict $\hat{\mu}^0(X)$ for both the i and the matched $j(i)$ units:

$$\begin{aligned}\hat{\mu}_i^0 &= \hat{\alpha} + \hat{\beta}X_i \\ \hat{\mu}_{j(i)}^0 &= \hat{\alpha} + \hat{\beta}X_{j(i)}\end{aligned}$$

Steps

3. Subtract $\hat{\mu}_i^0(X_i) - \hat{\mu}_{j(i)}^0(X_{j(i)})$, our estimate of the selection bias caused by matching discrepancies, from the sample estimate of the *ATT*:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[(Y_i - Y_{j(i)}) - (\hat{\mu}^0(X_i) - \hat{\mu}^0(X_{j(i)})) \right]$$

Steps

4. Estimate Abadie-Imbens robust standard error (Abadie and Imbens 2006; 2008; 2011)

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{(10 - 8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4 - 1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3} \\ &+ \frac{(10 - 9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33\end{aligned}$$

Matching bias: Implications for practice

Matching bias arises because of the effect of large matching discrepancies on $\mu^0(X_i) - \mu^0(X_{j(i)})$ due to a lack of common support. To minimize matching discrepancies:

1. Use a small M (e.g., $M = 1$). Larger values of M produce large matching discrepancies.
2. Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement.
3. Try to match covariates with a large effect on $\mu^0(\cdot)$ particularly well.

Large sample distribution for matching estimators

- Cannot use the bootstrap, so Abadie and Imbens derived the variance (Abadie and Imbens 2008)
- Matching estimators have a Normal distribution in large samples (provided the bias is small):

$$\sqrt{N_T}(\widehat{\delta}_{ATT} - \delta_{ATT}) \xrightarrow{d} N(0, \sigma_{ATT}^2)$$

- For matching without replacement, the “usual” variance estimator:

$$\widehat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2,$$

is valid.

Large sample distribution for matching estimators

- For matching with replacement:

$$\begin{aligned}\widehat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2 \\ &+ \frac{1}{N_T} \sum_{D_i=0} \left(\frac{K_i(K_i - 1)}{M^2} \right) \widehat{var}(\varepsilon | X_i, D_i = 0)\end{aligned}$$

where K_i is the number of times observation i is used as a match.

- $\widehat{var}(Y_i | X_i, D_i = 0)$ can be estimated also by matching. For example, take two observations with $D_i = D_j = 0$ and $X_i \approx X_j$, then

$$\widehat{var}(Y_i | X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of $\widehat{var}(\varepsilon_i | X_i, D_i = 0)$

Heterogeneity and OLS

- Most common causal model is OLS with covariates (“run regressions with controls”) but under heterogeneity, it can break down
- Under constant treatment effects, then given exogeneity of the error with respect to covariates, OLS is unbiased estimator of the constant causal effect
- It is also best of all linear unbiased estimators (BLUE)
- But as Imbens and Rubin (2015) note here, complex functional forms and heterogeneous treatment effects create some challenges

What about OLS? (Imbens and Rubin 2015)

"In many empirical studies in social sciences, causal effects are estimated through linear regression, where, typically it is implicitly assumed that in the super-population,

$$E[Y_i^D | X_i] = \alpha + \delta_{sp} \cdot D + X_i \beta$$

for some values of the three unknown parameters, α , δ_{sp} and β where $\delta_{sp} = E_{sp}[Y_i^1 - Y_i^0]$."

What about OLS? (Imbens and Rubin 2015)

"Defining $\varepsilon_i = Y_i - \delta_{sp} \cdot D_i - X_i\beta$ so that we can write

$$Y_i = \alpha + \delta_{sp} \cdot D_i + X_i\beta + \varepsilon_i$$

it is then assumed that

$$\varepsilon_i \perp D_i, X_i$$

This assumption is often referred to as **exogeneity** of the treatment (and the pre-treatment variables) in the econometrics literature."

What about OLS? (Imbens and Rubin 2015)

"The regression function is interpreted as a causal relation, in our sense of the term "causal", namely that if we manipulate the treatment D_i , then the outcome would change in expectation by an amount δ_{sp} . Hence in the potential outcomes formulation, we have

$$\begin{aligned} Y_i^0 &= \alpha + X_i\beta + \varepsilon_i \\ Y_i^1 &= Y_i^0 + \delta_{sp} \end{aligned}$$

What about OLS? (Imbens and Rubin 2015)

"Then, because ε_i is a function of Y_i^0 and X_i given the parameters,

$$Pr(D_i = 1|Y_i^0, Y_i^1 X_i) = Pr(D_i|\varepsilon_i, X_i),$$

and by exogeneity of the treatment indicator, we have

$$Pr(D_i|\varepsilon_i, X_i) = Pr(D_i|X_i)$$

and thus [conditional independence] holds."

What about OLS? (Imbens and Rubin 2015)

"However, the exogeneity assumption combines unconfoundedness with functional form and constant treatment effect assumptions that are quite strong, and arguably unnecessary."

Constant Treatment Effects and Linearity

Most commonly used method is OLS where the outcome is an additive model of the observed outcome, Y , on the treatment, D , and covariates, X like:

$$Y_i = \alpha + \delta D_i + \beta_1 X_i + \varepsilon_i$$

Take conditional expectations

$$\begin{aligned} E[Y_i | D_i = 1, X_i] &= \alpha + \delta E[D_i | D_i = 1, X_i] + \beta_1 E[X_i | D_i = 1, X_i] \\ E[Y_i | D_i = 0, X_i] &= \alpha + \delta E[D_i | D_i = 0, X_i] + \beta_1 E[X_i | D_i = 0, X_i] \end{aligned}$$

Constant Treatment Effects and Linearity

Replace realized variables with potential notation (both outcomes and covariates):

$$\begin{aligned} E[Y_i^1 | D_i = 1, X_i] &= \alpha + \delta + \beta_{11} E[X_i^1 | D_i = 1, X_i] \\ E[Y_i^0 | D_i = 0, X_i] &= \alpha + \beta_{01} E[X_i^0 | D_i = 0, X_i] \end{aligned}$$

May seem somewhat unorthodox to also let X have potential status, but you'll see why in a minute

Constant Treatment Effects and Linearity

OLS Estimator is a simple difference in conditional means:

$$\begin{aligned}\hat{\delta} &= E[Y_i^1 | D_i = 1] - E[Y_i^0 | D_i = 0] \\ \hat{\delta} &= \left(\alpha + \delta E[D_i | D_i = 1] + \beta_{11} E[X_i^1 | D_i = 1] \right) \\ &\quad - \left(\alpha + \delta E[D_i | D_i = 0] + \beta_{01} E[X_i^0 | D_i = 0] \right) \\ &= \delta + \beta_{11} E[X_i^1 | D_i = 1] - \beta_{01} E[X_i^0 | D_i = 0]\end{aligned}$$

OLS model requires three things: (1) linearity, (2) covariates to be independent of treatment status (i.e., treatment cannot cause covariates to change), (3) $\beta_{11} = \beta_{01}$ (homogenous treatment effects with respect to X).

Simulation

- Following code will maintain linearity but have common support violation to show OLS does not require common support, but does require linearity (it extrapolates based on functional form which is quite spectacular with correct model)
- I will also create heterogenous treatment effects
- But I will also violate the previous requirement that $\beta_{11} = \beta_{01}$ so that you can see the bias that forms on average across 1,000 simulations
- Will show a variety of estimators and specifications so that we see how to recover causal parameters with regression and matching

Heterogenous Treatment Effects wrt X

```
* Simulation with heterogenous treatment effects, unconfoundedness and OLS estimation
clear all
program define het_te, rclass
version 14.2
syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]

    clear
    drop _all
    set obs 5000
    gen treat = 0
    replace treat = 1 in 2501/5000

    * Poor pre-treatment fit
    gen age = rnormal(25,2.5)      if treat==1
    replace age = rnormal(30,3)      if treat==0
    gen gpa = rnormal(2.3,0.75)    if treat==0
    replace gpa = rnormal(1.76,0.5) if treat==1

    su age
    replace age = age - `r(mean)'

    su gpa
    replace gpa = gpa - `r(mean)'

    gen age_sq = age^2
    gen gpa_sq = gpa^2
    gen interaction=gpa*age

    gen y0 = 15000 + 10.25*age + -10.5*age_sq + 1000*gpa + -10.5*gpa_sq + 500*
interaction + rnormal(0,5)
    gen y1 = y0 + 2500 + 100 * age + 1000*gpa
    gen delta = y1 - y0

    su delta // ATE = 2500
    su delta if treat==1 // ATT = 1980
    local att = r(mean)
    scalar att = `att'
    gen att = `att'

    gen earnings = treat*y1 + (1-treat)*y0
```

Parameters

- Ordinarily we look at the coefficient on the treatment dummy to obtain an estimate
- But we have two parameters: the ATE is \$2500 but the ATT is \$1980
- How do we get both of them? Let's look at what people usually do
- 1,000 simulations of DGP with regression estimates plotting coefficient on treatment dummy: first with just age and GPA, second with the precise model used for Y^0 (but not Y^1)

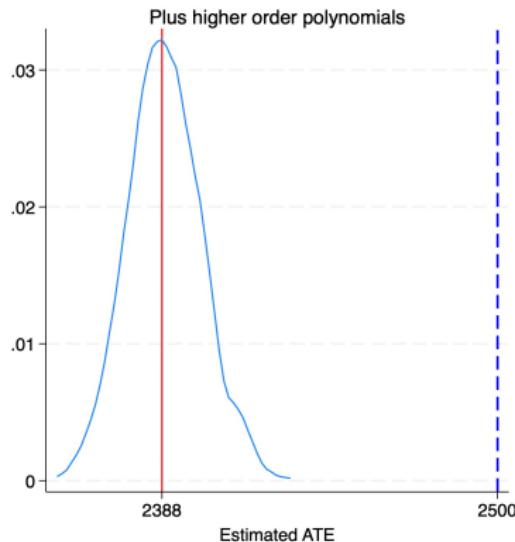
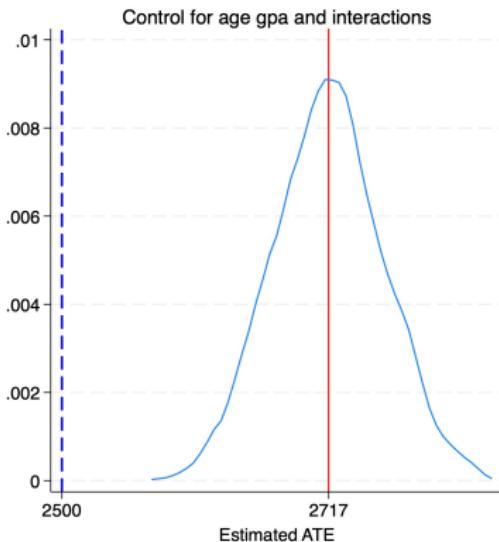
Constant Treatment Effects and Linearity

```
* Regression 1: constant treatment effects, no quadratics
reg earnings treat age gpa, robust
local treat1=_b[treat]
scalar treat1 = `treat1'
gen treat1=`treat1'

* Regression 2: constant treatment effects, quadratics and interaction
reg earnings treat age age_sq gpa gpa_sq c.gpa#c.age, robust
local treat2=_b[treat]
scalar treat2 = `treat2'
gen treat2=`treat2'
```

Coefficient on Treatment Dummy is Wrong

Non-saturated regressions with heterogenous treatment effects



ATE is 2500 and ATT is 1980

Commentary

- Three ifs and a then:
 - If unconfoundedness held, and
 - if the potential outcome model was linear, and
 - if the treatment effect had been homogenous with respect to age and GPA,
 - then the coefficient on the treatment variable would have been the ATE
- But it wasn't because homogeneity with respect to X was not true (recall $Y(0)$ coefficients were not the same as $Y(1)$ coefficients)
- So what had we done? Bear with me but this will pay off

Heterogenous treatment effects

Write down a simplified version of the DGP from the code:

$$Y_i^0 = \alpha + \beta_{01} X_i^0 + \varepsilon_i$$

$$Y_i^1 = \alpha + \beta_{01} X_i^0 + \delta D_i + \beta_{11} X_i^1 \times D_i + \varepsilon_i$$

Notice that the setup before, X_i had a different effect on Y^0 than it did on Y_i^1 – that's because of heterogenous treatment effects with respect to conditioning set.

Heterogenous treatment x'effects

Take conditional expectations of the *potential* outcomes:

$$E[Y_i^0|D_i = 1, X_i] = \alpha + \beta_{01}E[X_i^0]$$

$$E[Y_i^1|D_i = 1, X_i] = \alpha + \beta_{01}E[X_i^0] + \delta + \beta_{11}E[X_i^1 \times D_i|D_i = 1, X_i^1]$$

Average treatment effect is:

$$\begin{aligned} E[Y_i^1|D_i, X_i^1] - E[Y_i^0|D_i, X_i^0] &= \left(\alpha + \beta_{01}E[X_i^0] + \delta + \beta_{11}E[X_i^1 \times D_i|D_i = 1, X_i^1] \right) \\ &\quad - \left(\alpha + \beta_{01}E[X_i^0] \right) \\ &= \delta + \beta_{11}E[X_i^1|D_i = 1] \end{aligned}$$

assuming $X_i^1 = X_i^0$. OLS model accounting for heterogeneity must be "fully saturated".

Estimation

Our saturated OLS model is:

$$Y_i = \alpha + \delta D_i + \beta_{01} X_i + \beta_{11} D_i \times X_i + \varepsilon_i$$

$\widehat{\delta}$ is the ATE but the ATT is equal to $\widehat{\delta} + \beta_{11} E[X_i | D_i = 1]$ where $E[X_i | D_i = 1]$ is the sample average of X_i for the treatment group

We will estimate two models: (1) once with simplified but incorrectly specified saturated and (2) another with the correctly specified saturated model – warning, it's a huge pain and you can easily mess it up even with just a few variables

Misspecified Saturated OLS Regression

```
* Regression 3: Heterogenous treatment effects, partial saturation
regress earnings i.treat##c.age##c.gpa, robust
local ate1=_b[1.treat]
scalar ate1 = `ate1'
gen ate1=`ate1'

* Obtain the coefficients
local treat_coef = _b[1.treat]
local age_treat_coef = _b[1.treat#c.age]
local gpa_treat_coef = _b[1.treat#c.gpa]
local age_gpa_treat_coef = _b[1.treat#c.age#c.gpa]

* Save the coefficients as scalars and generate variables
scalar treat_coef = `treat_coef'
gen treat_coef_var = `treat_coef'

scalar age_treat_coef = `age_treat_coef'
gen age_treat_coef_var = `age_treat_coef'

scalar gpa_treat_coef = `gpa_treat_coef'
gen gpa_treat_coef_var = `gpa_treat_coef'

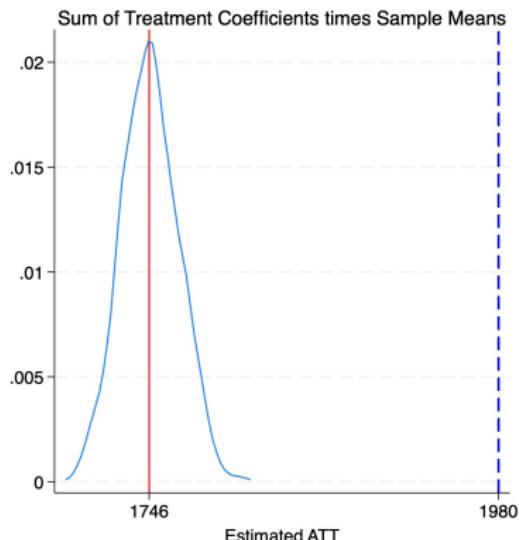
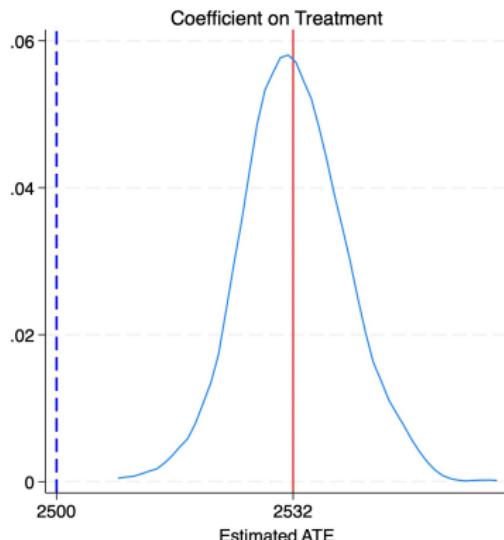
scalar age_gpa_treat_coef = `age_gpa_treat_coef'
gen age_gpa_treat_coef_var = `age_gpa_treat_coef'

* Calculate the mean of the covariates
egen mean_age = mean(age), by(treat)
egen mean_gpa = mean(gpa), by(treat)

* Calculate the ATT
gen treat3 = treat_coef_var + ///
    age_treat_coef_var * mean_age + ///
    gpa_treat_coef_var * mean_gpa + ///
    age_gpa_treat_coef_var * mean_age * mean_gpa if treat == 1
```

Misspecified Saturated OLS Regression

Misspecified Saturated Regressions



1000 Monte Carlo simulations

Comically Long Saturated OLS Regression

```
+ Regression 4: Fully saturated regression model
#delimit ;
regress earnings i.treat##c.age
           i.treat##c.age_sq
           i.treat##c.gpa
           i.treat##c.gpa_sq
           i.treat##c.age##c.gpa;
#delimit cr

local ate2= b[1,treat]
scalar ate2= `ate2'
gen ate2= `ate2'

* Obtain the coefficients
local treat_coeff = `b[1,treat]' // 0
local age_treat_coeff = `b[1,treat##c.age]' // 1
local agesq_treat_coeff = `b[1,treat##c.age_sq]' // 2
local gpa_treat_coeff = `b[1,treat##c.gpa]' // 3
local gpasq_treat_coeff = `b[1,treat##c.gpa_sq]' // 4
local age_gpa_coeff = `b[1,treat##c.gpa]' // 5

* Save the coefficients as scalars and generate variables
scalar treat_coeff = `treat_coeff'
gen treat_coeff_var = `treat_coeff' // 0
scalar age_treat_coeff = `age_treat_coeff'
gen age_treat_coeff_var = `age_treat_coeff' // 1
scalar agesq_treat_coeff = `agesq_treat_coeff'
gen agesq_treat_coeff_var = `agesq_treat_coeff' // 2
scalar gpa_treat_coeff = `gpa_treat_coeff'
gen gpa_treat_coeff_var = `gpa_treat_coeff' // 3
scalar gpasq_treat_coeff = `gpasq_treat_coeff'
gen gpasq_treat_coeff_var = `gpasq_treat_coeff' // 4
scalar age_gpa_coeff = `age_gpa_coeff'
gen age_gpa_coeff_var = `age_gpa_coeff' // 5

* Calculate the mean of the covariates
su age if treat==1
local mean_age = `r(mean)'
gen mean_age = `mean_age'

su age_sq if treat==1
local mean_agesq = `r(mean)'
gen mean_agesq = `mean_agesq'

su gpa if treat==1
local mean_gpa = `r(mean)'
gen mean_gpa = `mean_gpa'

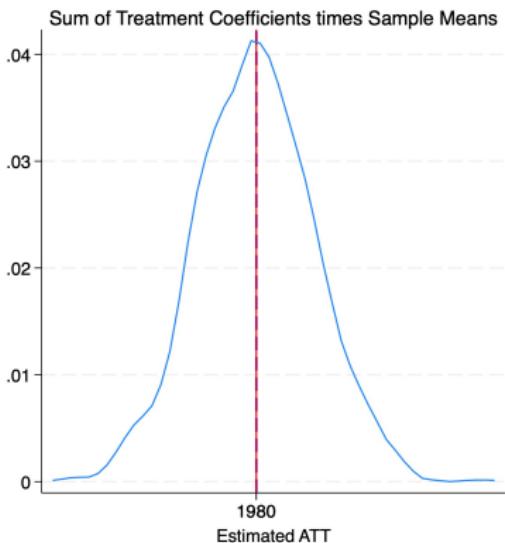
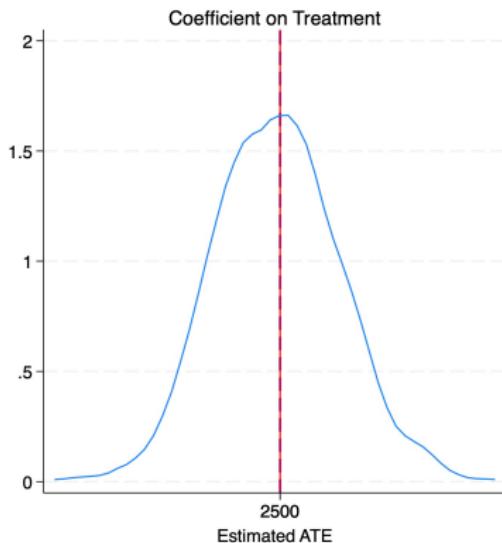
su gpasq if treat==1
local mean_gpasq = `r(mean)'
gen mean_gpasq = `mean_gpasq'

su agegpa if treat==1
local mean_agegpa = `r(mean)'
gen mean_agegpa = `mean_agegpa'

* Calculate the ATT
gen treat4 =
    treat_coeff_var + /// 0
    age_treat_coeff_var * mean_age + /// 1
    agesq_treat_coeff_var * mean_agesq + /// 2
    gpa_treat_coeff_var * mean_gpa + /// 3
    gpasq_treat_coeff_var * mean_gpasq + /// 4
    age_gpa_coeff_var * mean_agegpa
```

Correctly Saturated OLS Regression

Correctly Specified Saturated Regressions



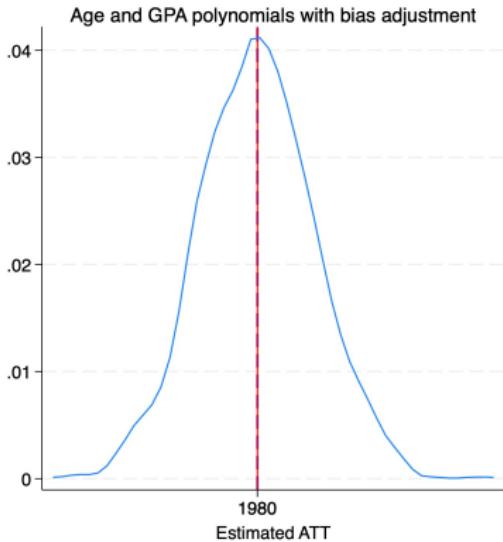
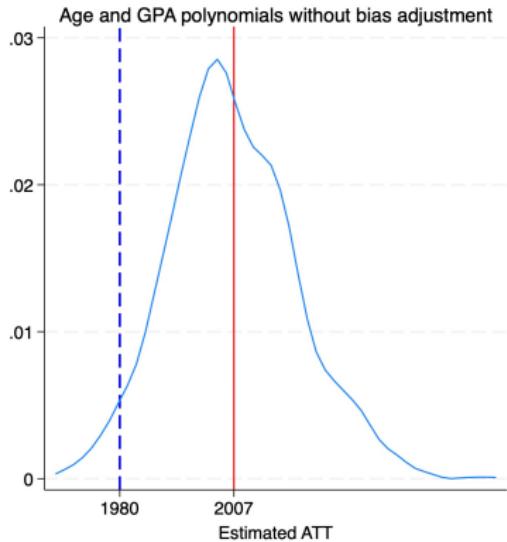
1000 Monte Carlo simulations

Matching

- Now let's estimate the ATT (\$1980) using nearest neighbor matching by minimizing Mahalanobis distance on age, GPA, polynomials and interaction
- One line in Stata using `teffects` and only 1 match (variance is simple to estimate until we use matches multiple times, then the variance grows)
- In R, the package is `Matching`, not sure in python

Matching Estimation

Nearest Neighbor Matching with Minimized Maha Distance



Estimated ATT from 1000 simulations using nearest neighbor matching

Commentary

- Full saturation tips – remember to center the covariates at the start, be sure to specify ahead of time which parameter (ATT or ATE) and note that the covariates must not be based on the treatment (“exogenous” and confounders only)
- Other methods, more sophisticated, have emerged such as outcome regressions where you impute missing counterfactuals, $\widehat{E[Y^0|D=1, X]}$ using the comparison group only and Y^0 as a function of X
- But note the saturation is not merely so you can examine heterogeneity at the margin – you need it even to get the correct ATE as we saw which is distinct from the ATT

Commentary

- Unconfoundedness requires that you *know* and *include* all confounders to adjust comparisons when estimating treatment effects
- Without a prior behavioral model guiding you, it's very hard to defend conditional independence (borderline disingenuous)
- If you are unwilling to use DAGs, you may want to ask yourself why you are comfortable running regressions with covariates?

Avoiding dimensionality problems

- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method which is a dimension reduction method of reducing K covariates used for adjusting into a single scalar called the propensity score
- Propensity score is simply the frequentist share of units in the treatment group with values of X compared to all units with values of X
- Can be then used to weight conditional mean outcomes of the comparison group to get the ATT as well as match, but I'm going to focus on weighting

Basic idea behind propensity scores

- Earlier we matched on X 's to compare units "near" one another based on some distance but matching discrepancies and sparseness created problems
- Propensity scores summarize covariate information about treatment selection into a single number bounded between 0 and 1 (i.e., a probability)
- Rather than compare units with similar values of X , we compare units with similar **estimated conditional probabilities of treatment**
- Important theorem shows that once we adjust comparisons using the propensity score, we do not need to adjust for X

Formal Definition

Definition of Propensity score

A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables: $p(X) = Pr(D = 1|X)$

For the ATT, we need weak unconfoundedness and weak overlap (i.e., only need overlap for the treatment group)

Propensity score theorem

Propensity score theorem

If $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (unconfoundedness), then $(Y^1, Y^0) \perp\!\!\!\perp D|\rho(X)$
where $\rho(X) = Pr(D = 1|X)$, the propensity score

- Conditioning on the propensity score is enough to have independence between D and (Y^1, Y^0) (Rosenbaum and Rubin 1983)
- Valuable theorem because of dimension reduction and convergence rate issues which can introduce biases

Step 1: Estimate the propensity score

- Estimate the conditional probability of treatment using probit or logit model (or more sophisticated)

$$Pr(D_i = 1|X_i) = F(\beta X_i)$$

- Use the estimated coefficients to calculate the propensity score for each unit i

$$\hat{\rho}_i(X_i) = \hat{\beta} X_i$$

- Note that each unit i now has a predicted probability of treatment given the values of their covariates relative to everyone else's
- Frequentist probability – you've basically just obtained the likelihood someone who "looks like you" would be treated (regardless of whether you were in fact treated)

Step 2: Estimation of ATT with IPW

- IPW uses the estimated propensity score to reweight the outcomes (e.g., Robins and Rotnitzky 1995, Imbens 2000, Hirano and Imbens 2001)
- IPW is non-parametric – you are just taking averages and multiplying by weights
- There are also fewer implementation choices – you aren't choosing how many neighbors to include, how far away a neighbor can be – but you still have to closely examine common support
- There are bias adjustment methods called double robust where you combine imputing counterfactuals with weighting by the propensity score

Step 2: Estimation of ATT with IPW

Estimating ATT with IPW

Given $Y^0 \perp\!\!\!\perp D|X$ and common support, then

$$\begin{aligned}\delta_{ATT} &= E[Y^1 - Y^0|D = 1] \\ &= \frac{1}{Pr(D = 1)} \cdot E \left[Y \cdot \frac{D - \rho(X)}{1 - \rho(X)} \right]\end{aligned}$$

Notice that when $D = 1$, the outcome is not weighted, but when $D = 0$ it is. You're missing the Y^0 for the treatment, not Y^1 so you weight the treatment group Y values alone and weight "up" or "down" the comparison groups by their propensity scores

Step 3: Standard Errors

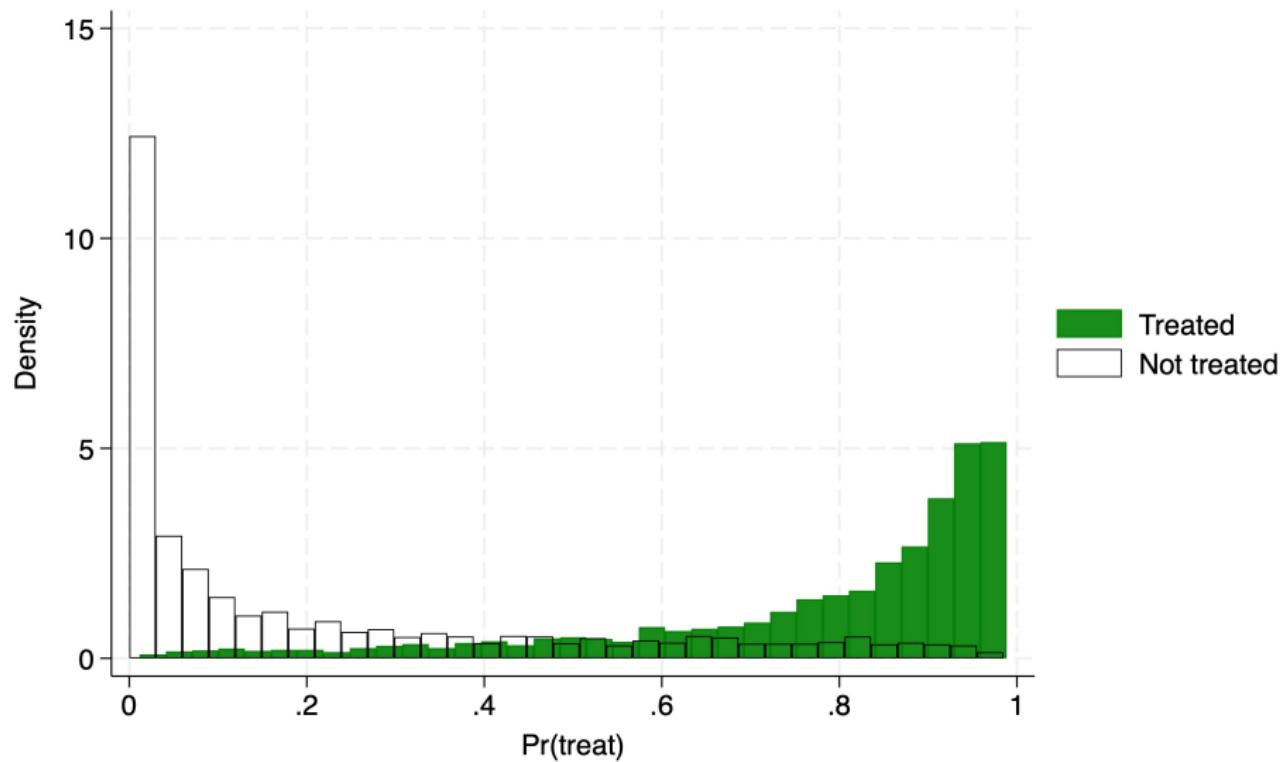
Standard errors can be constructed a few different ways:

- We need to adjust the standard errors for first-step estimation of $\rho(X)$
 - Parameteric first step: Newey and McFadden (1994)
 - Non-parametric first step: Newey (1994)
- IPW is a smooth estimator which means the bootstrap is valid for inference (Adudumilli 2018 and Bodory et al. 2020)

Practical uses

- One of the things I find immensely valuable is that when you can collapse the conditioning set into a single scalar, then assessing overlap is simple
- Just create histograms of the propensity score distribution for treatment and control
- Crump, et al. (2009) suggest keeping propensity scores within the interval [0.1,0.9] (“trimming”) but note any time you drop a unit, you are moving away from the ATT

Assessing overlap



Concluding Remarks

- Unconfoundedness implies two things: that there are no unknown confounders and that for groups of units with the same values of all confounders, treatment is assigned to units independent of either both potential outcomes (for the ATE) or just Y^0 (for the ATT)
- Second assumption is also strong but testable – common support. Can be checked with propensity score
- If there's heterogeneous treatment effects with respect to X , then saturated regressions as well as other methods can recover the ATT, but OLS requires linearity (but not common support), but matching and weighting require common support (but not linearity) – the former extrapolates based on outcome modeling, but the latter interpolates

Roadmap

In Pursuit of the ATT

- Potential Outcomes

- Independence and Selection Bias

Unconfoundedness and Ignorable Treatment Assignment

- Exact and Inexact Matching

- Saturated Regressions

Synthetic control

- Interpolation with non-negative weighting

- Extrapolation with Conservative Negative Weighting

Difference-in-differences

- Four averages and three subtractions

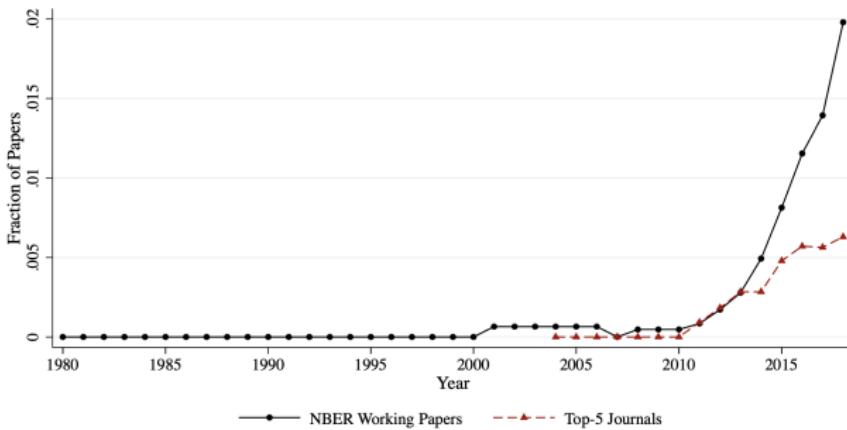
- Covariates

- Model Misspecification

- Alternatives to TWFE

Conclusion

D: Synthetic Control



What is synthetic control

- Synthetic control has been called the most important innovation in causal inference of the last two decades (Athey and Imbens 2017)
- Originally designed for comparative case studies, but newer developments have extended it to multiple treated units as well as differential timing
- Continues to also be methodologically a frontier for applied econometrics, so consider this talk a starting point for you

What is a comparative case study

- Comparative case studies compare a single unit to another unit to make causal inference
- Single treated unit is usually a country, state, firm, etc.
- Social scientists traditionally tackled them either qualitatively and quantitatively (more traditional economic approach)

Qualitative comparative case studies

- In qualitative comparative case studies, the goal might be to reason *inductively* the causal effects of events or characteristics of a single unit on some outcome, oftentimes through logic and historical analysis.
 - Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America (but he is regularly comparing the US to France)
- Sometimes there may not be an explicit counterfactual, or if there is, it's not principled (subjective researcher decision)
- Quantitative claims about causal effects are unlikely – de Toqueville's won't claim GDP per capita fell \$500 when compared against France

Traditional quantitative comparative case studies

- Traditional quantitative comparative case studies are explicitly causal designs in that there is a treatment and control, usually involving natural experiment on a single aggregate unit
- Comparison focuses on the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same *ad hoc* aggregate control group (Card 1990; Card and Krueger 1994)
- It'll essentially be diff-in-diff, but it may not use the event study, and the point is the choice of controls is a subset of all possible controls

Pros and cons

- Pros:
 - Takes advantage of policy interventions that take place at an aggregate level (which is common and so this is useful)
 - Aggregate/macro data are often available (which may be all we have)
- Cons:
 - Selection of control group is *ad hoc* – opens up researcher biases, even unconscious
 - Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

Description of the Mariel Boatlift

- In 1980, Fidel Castro allowed anyone to leave Cuba so long as they did in the fall from the Mariel boat dock.
- The Mariel Boatlift brought 100,000 Cubans to Miami which increased the Miami labor force by 7%
- Card (1990) uses the Mariel Boatlift as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- His question was how do inflows of immigrants affect the wages and employment of natives in local US labor markets?
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and comparison cities







Selecting control groups

- His treatment group was low skill workers in Miami since that's where Cubans went
- But which control group?
- He chose Atlanta, Los Angeles, Houston, Tampa-St. Petersburg

Why these four?

Tables 3 and 4 present simple averages of wage rates and unemployment rates for whites, blacks, Cubans, and other Hispanics in the Miami labor market between 1979 and 1985. For comparative purposes, I have assembled similar data for whites, blacks, and Hispanics in four other cities: Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. These four cities were selected both because they had relatively large populations of blacks and Hispanics and because they exhibited a pattern of economic growth similar to that in Miami over the late 1970s and early 1980s. A comparison of employment growth rates (based on establishment-level data) suggests that economic conditions were very similar in Miami and the average of the four comparison cities between 1976 and 1984.

Diff-in-diff

Differences-in-differences estimates of the effect of immigration on unemployment^a

Group	Year			
	1979 (1)	1981 (2)	1981–1979 (3)	
Whites				
(1)	Miami	5.1 (1.1)	3.9 (0.9)	- 1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	- 0.1 (0.4)
(3)	Difference Miami-comparison	0.7 (1.1)	- 0.4 (0.95)	- 1.1 (1.5)
Blacks				
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Difference Miami-comparison	- 2.0 (1.9)	- 3.0 (2.0)	- 1.0 (2.8)

^a Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Parallel trends

- His estimate is unbiased if the change in Y^0 for the comparison cities correctly approximates the unobserved ΔY^0 for the treatment group
- But Card largely focused on covariates, and in a relatively casual way (“similar growth”) and does not report much
- Black result would have been positive, too, were it not that the comparison cities growth was smaller – uncertainty about null result being from no effect or arbitrary control group

Synthetic Control

- Abadie and Gardeazabal (2003) introduced synthetic control in the AER in a study of a terrorist attack in Spain (Basque) on GDP
- Revisited again in a 2010 JASA with Diamond and Hainmueller, two political scientists who were PhD students at Harvard (more proofs and inference)
- Basic idea is to use a combination of comparison units as counterfactual for a treated unit where the units are chosen according to a data driven procedure

Researcher's objectives

- Our goal here is to reproduce the counterfactual of a treated unit by finding the combination of untreated units that best resembles the treated unit *before* the intervention in terms of the values of k relevant covariates (predictors of the outcome of interest)
- Method selects *weighted average of all potential comparison units* that best resembles the characteristics of the treated unit(s) - called the "synthetic control"

Synthetic control method: advantages

- “Convex hull” means synth is a weighted average of units which means the counterfactual is a collection of comparison units that on average track the treatment group over time.
- Constraints on the model use non-negative weights which does not allow for extrapolation
- Makes explicit the contribution of each comparison unit to the counterfactual
- Formalizing the way comparison units are chosen has direct implications for inference

Notation and setup

Suppose that we observe $J + 1$ units in periods $1, 2, \dots, T$

- Unit “one” is exposed to the intervention of interest (that is, “treated” during periods $T_0 + 1, \dots, T$)
- The remaining J are an untreated reservoir of potential controls (a “donor pool”)

Group-time ATT with only one treated group

Using same potential outcomes notation as we've been using, define the ATT parameter as a dynamic group-time

$$\begin{aligned}\delta_{1t} &= Y_{1t}^1 - Y_{1t}^0 \\ &= Y_{1t} - Y_{1t}^0\end{aligned}$$

for each post-treatment period, $t > T_0$ and Y_{1t} is the outcome for unit one at time t . We will estimate Y_{1t}^0 using the J units in the donor pool

Estimating W weights

- Let $W = (w_2, \dots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \dots, J + 1$ and $w_2 + \dots + w_{J+1} = 1$. Each value of W represents a potential synthetic control
- Let X_1 be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^* = (w_2^*, \dots, w_{J+1}^*)'$ is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints

Donor weights and characteristic weights

Abadie, et al. consider

$$\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where X_{jm} is the value of the m -th covariates for unit j and V is some $(k \times k)$ symmetric and positive semidefinite matrix

More on the V matrix

Typically, V is diagonal with main diagonal v_1, \dots, v_k . Then, the synthetic control weights w_2^*, \dots, w_{J+1}^* minimize:

$$\sum_{m=1}^k v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

where v_m is a weight that reflects the relative importance that we assign to the m -th variable when we measure the discrepancy between the treated unit and the synthetic controls

How this works

This method of “minimizing pre-treatment characteristics” is very similar to nearest neighbor matching from Abadie and Imbens (2006)

Let's look at it together; it should help you understand, too, the idea of the V matrix being crucial

https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing

Donor weights and characteristic weights

- If all pre-treatment characteristics are treated equally, then there exists a solution to the minimization, W^*
- But if pre-treatment characteristics were emphasized differently in the minimization (say one was given a weight of 0.00001), then there is a different solution, W^* '.
- Making a choice about V (the weight on pre-treatment characteristics' importance) is unavoidable – a non-choice is a choice as you'll just weight all characteristics the same

Keeping in mind our goal - create a good synth

- The synthetic control $W^*(V^*)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment
- Therefore, the V^* weights directly shape W^*

Estimating the V matrix

Choice of v_1, \dots, v_k can be based on

- Assess the predictive power of the covariates using regression
- Subjectively assess the predictive power of each of the covariates, or calibration inspecting how different values for v_1, \dots, v_k affect the discrepancies between the treated unit and the synthetic control
- Minimize mean square prediction error (MSPE) for the pre-treatment period (default):

$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^J w_j^*(V^*) Y_{jt} \right)^2$$

Choosing pre-treatment characteristics

- Original papers provided no guidance, and it's unclear how these characteristics connect to the factor model
- Goal is over a long stretch pre-treatment to create the convex hull that contains the treatment group
- Lots of early variation in what was done (including my papers), and some effort has now been made to be more principled to avoid p-hacking

Cross validation

Abadie suggests model selection using specifications guided by sample splitting into training and validation:

- Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
- For any given V , calculate $W^*(V)$ in the training period.
- Minimize the MSPE of $W^*(V)$ in the validation period

Avoiding Cherry Picking Synth

Ferman, Pinto and Possbaum (2020) note that there are opportunities for p-hacking in model selection and recommend reporting a variety of specifications rather than just one

1. Use all pre-treatment lagged outcomes as your X characteristics
2. Use first three-fourths
3. Use first half
4. Use all odd years
5. Use all even years
6. Use pre-treatment outcome mean
7. Use three outcome values

With and without covariates, report p-values (explained later) and event study plot

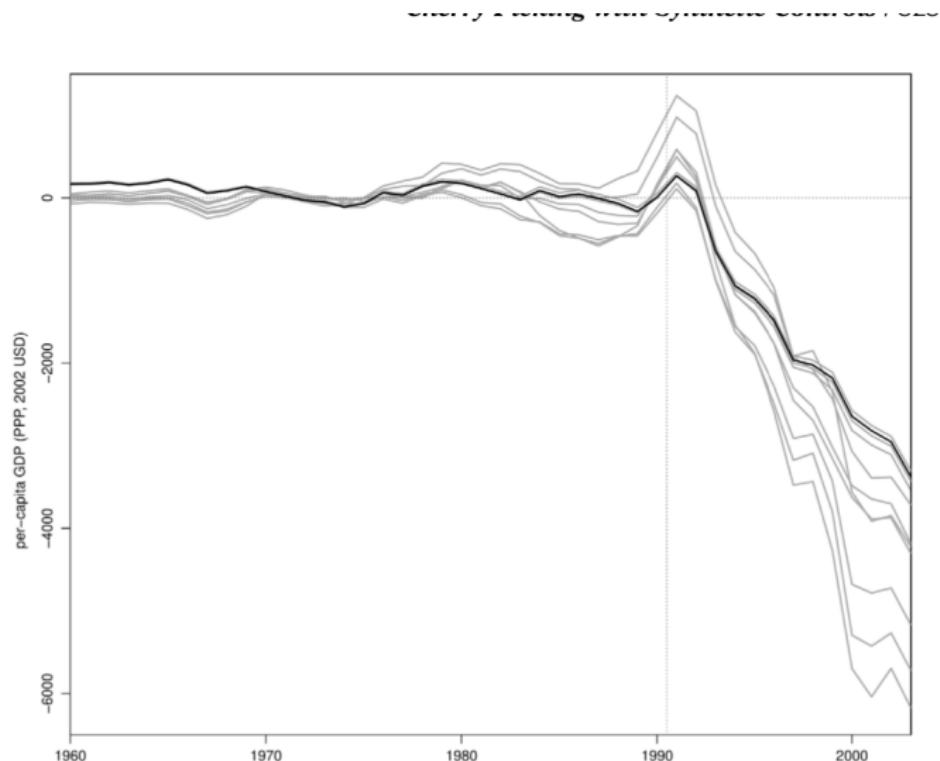
14 specification p-values

Table 3. Specification searching—database from Abadie et al. (2015).

Specification	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
p-value	0.059	0.059	0.059	0.118	0.118	0.059	0.059	0.059
Specification	(5a)	(5b)	(6a)	(6b)	(7a)	(7b)		
p-value	0.118	0.059	0.588	0.059	0.353	0.059		

Notes: We analyze 14 different specifications. The number of the specifications refers to: (1) all pre-treatment outcome values, (2) the first three-fourths of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) pre-treatment outcome mean (original specification by Abadie, Diamond, & Hainmueller, 2010), and (7) three outcome values. Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates trade openness, inflation rate, industry share, schooling levels, and investment rate.

14 specification event studies



Notes: The solid black line is the original specification by Abadie, Diamond, and Hainmueller (2015) and gray lines are specifications 1 through 5. The vertical line denotes the beginning of the post-treatment period.

Synth identification

- In diff-in-diff, the core assumption was parallel trends
- Only needed one pretreatment period for identification but needed parallel trends starting at that period
- Synthetic control does not use parallel trends for identification; it uses a factor model of Y^0
- But you need a long pre-treatment series, not just for the event study, but for modeling the heterogeneity adequately

Assumption: Y^0 is determined by factor model

What about unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors?

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it}$$

where α_t is an unknown common factor with constant factor loadings across units, and λ_t is a vector of unobserved common factors

With some manipulation

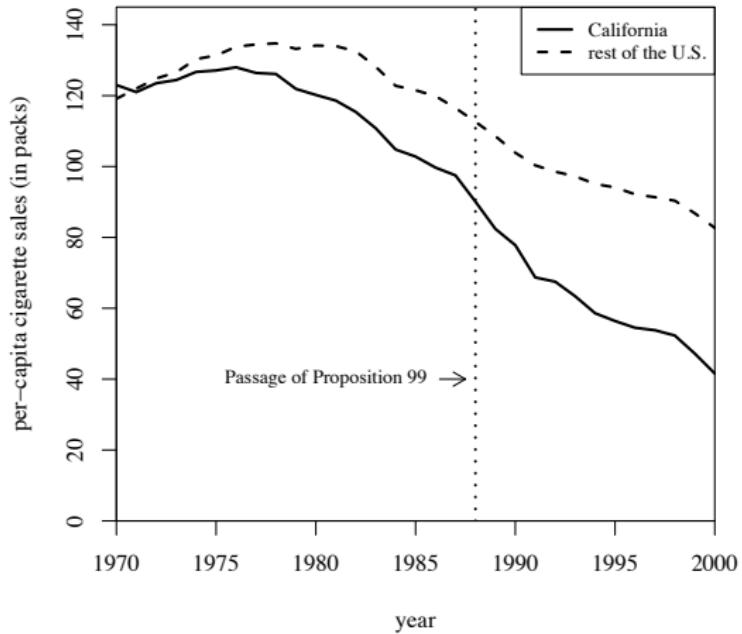
$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

- If $\sum_{t=1}^{T_0} \lambda_t' \lambda_t$ is nonsingular, then RHS will be close to zero if number of preintervention periods is “large” relative to size of transitory shocks
- Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time
- Main takeaway: you need a long pre-treatment time period to use this and the fit must be excellent

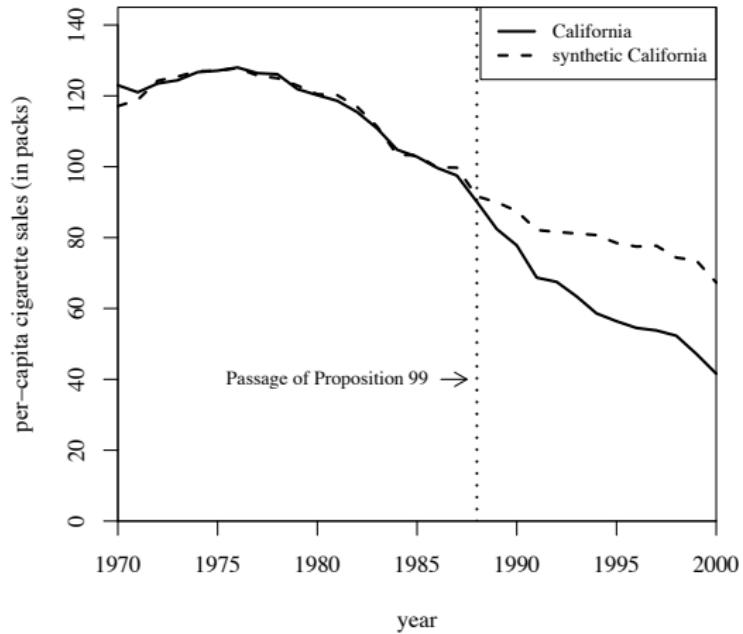
Example: California's Proposition 99

- In 1988, California first passed comprehensive tobacco control legislation:
 - increased cigarette tax by 25 cents/pack
 - earmarked tax revenues to health and anti-smoking budgets
 - funded anti-smoking media campaigns
 - spurred clean-air ordinances throughout the state
 - produced more than \$100 million per year in anti-tobacco projects
- Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)

Cigarette Consumption: CA and the Rest of the US



Cigarette Consumption: CA and synthetic CA

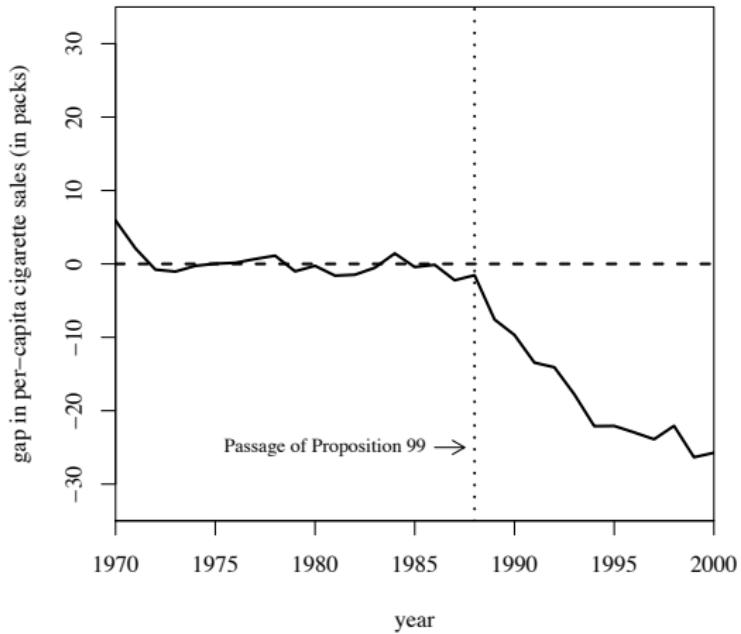


Predictor Means: Actual vs. Synthetic California

Variables	Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Smoking Gap between CA and synthetic CA



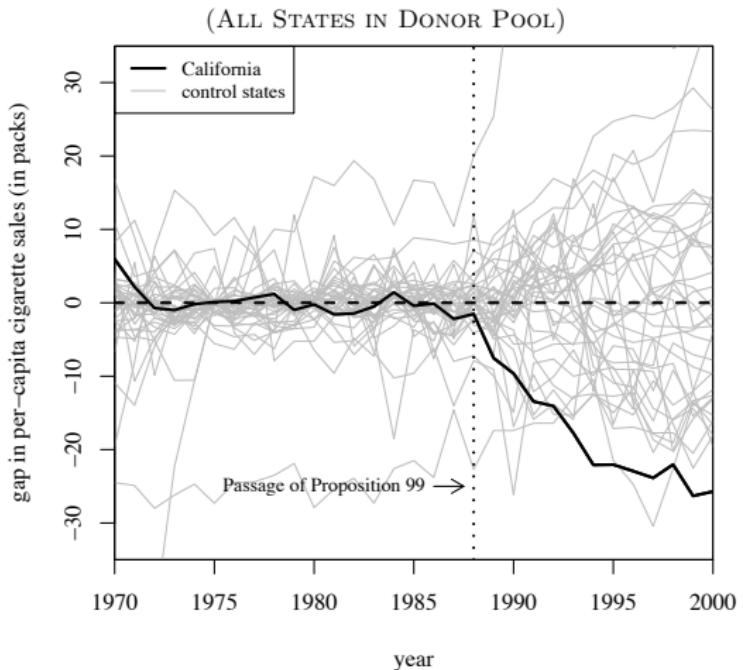
Inference

- To assess significance, we calculate exact p-values under Fisher's sharp null using a test statistic equal to after to before ratio of RMSPE
- Exact p-value method
 - Iteratively apply the synthetic method to each country/state in the donor pool and obtain a distribution of placebo effects
 - Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example the post-Prop. 99 RMSPE is:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^T \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

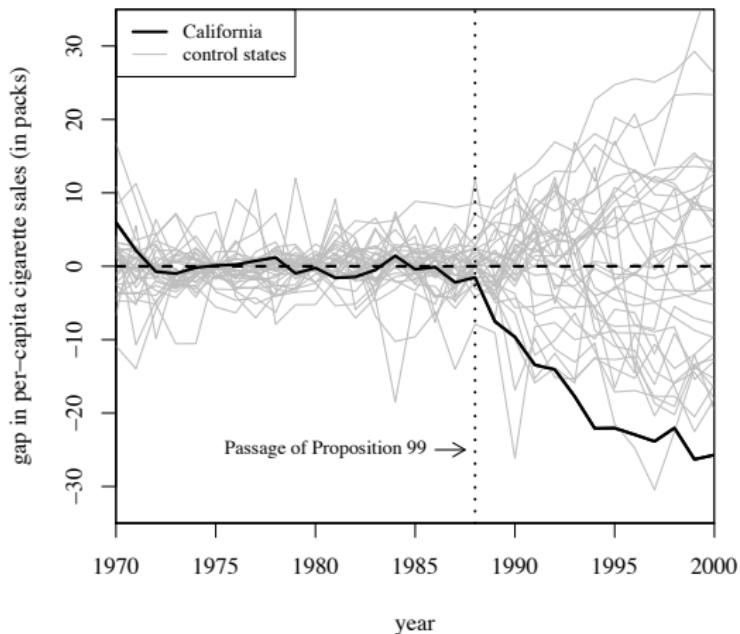
and the exact p-value is the treatment unit rank divided by J

Smoking Gap for CA and 38 control states



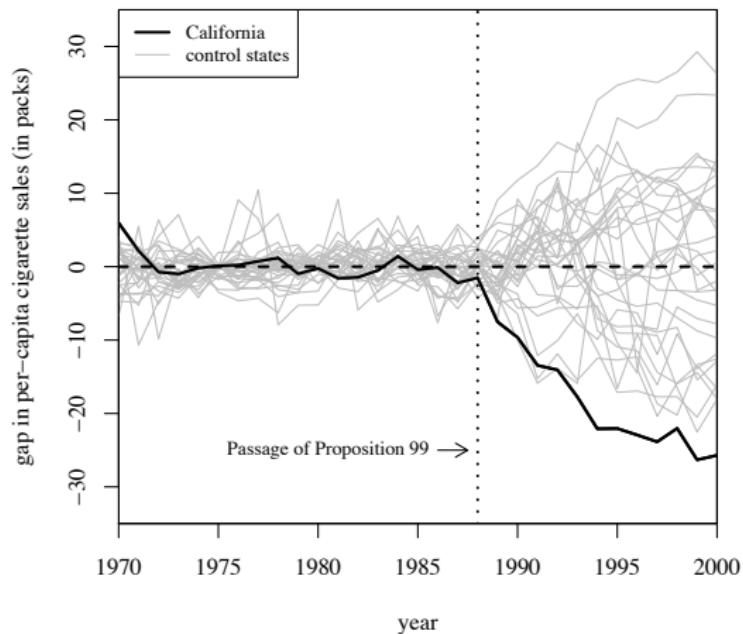
Smoking Gap for CA and 34 control states

(PRE-PROP. 99 MSPE \leq 20 TIMES PRE-PROP. 99 MSPE FOR CA)



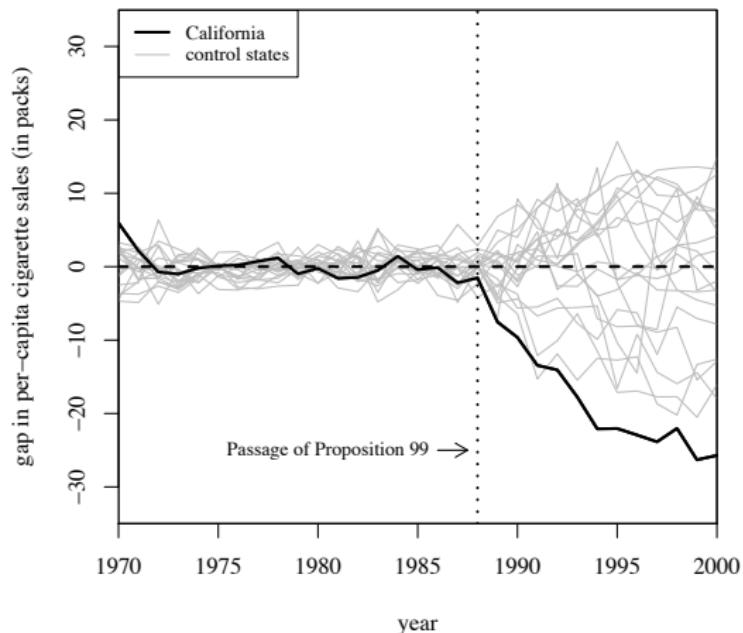
Smoking Gap for CA and 29 control states

(PRE-PROP. 99 MSPE \leq 5 TIMES PRE-PROP. 99 MSPE FOR CA)

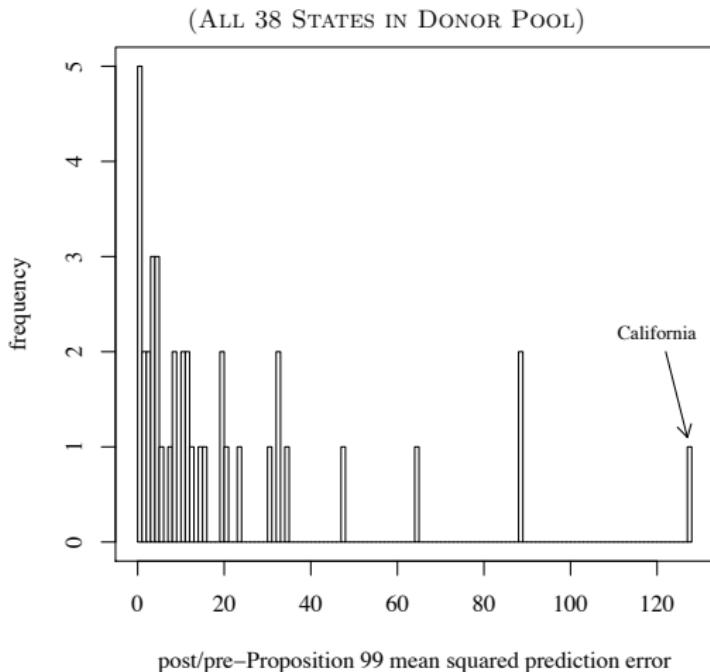


Smoking Gap for CA and 19 control states

(PRE-PROP. 99 MSPE \leq 2 TIMES PRE-PROP. 99 MSPE FOR CA)



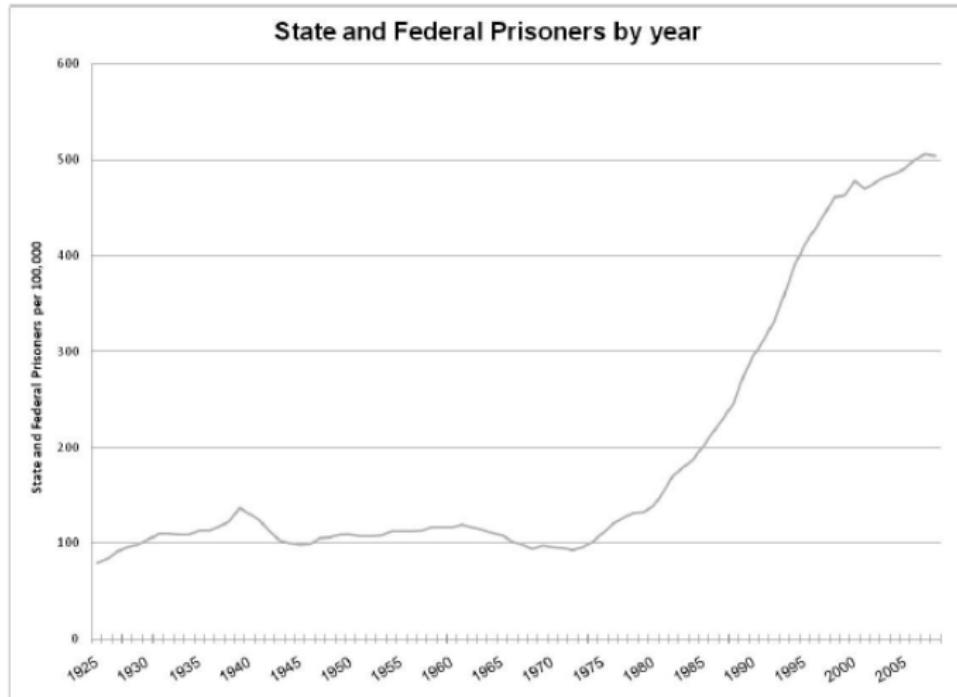
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Coding exercise

- The US has the highest prison population of any OECD country in the world
- 2.1 million are currently incarcerated in US federal and state prisons and county jails
- Another 4.75 million are on parole
- From the early 1970s to the present, incarceration and prison admission rates quintupled in size

Figure 1
History of the imprisonment rate, 1925 - 2008



Source: www.albany.edu/sourcebook/tost_6.html

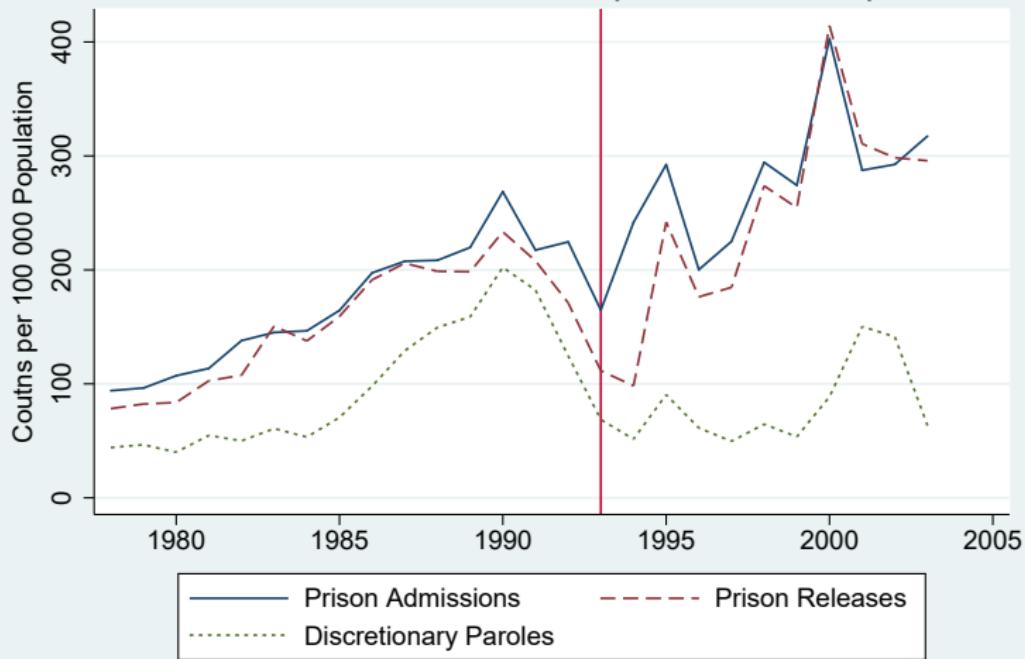
Prison constraints

- Prisons are and have been at capacity for a long time so growth in imprisonment would bite on state corrections
- Managing increased flows can only be solved by the following:
 - Prison construction
 - Overcrowding
 - Paroles
- Texas chooses overcrowding

Ruiz v. Estelle 1980

- Class action lawsuit against TX Dept of Corrections (Estelle, warden).
- TDC lost. Lengthy period of appeals and legal decrees.
- Lengthy period of time relying on paroles to manage flows

Texas Prison Flows Measures per 100 000 Population

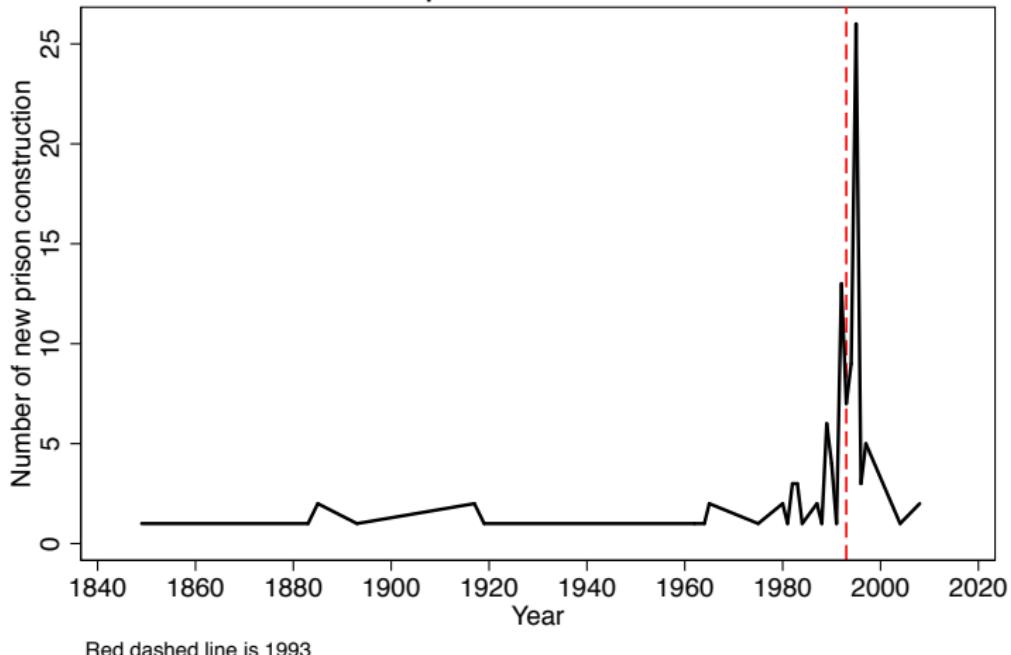


Texas prison boom

Governor Ann Richards (D) 1991-1995

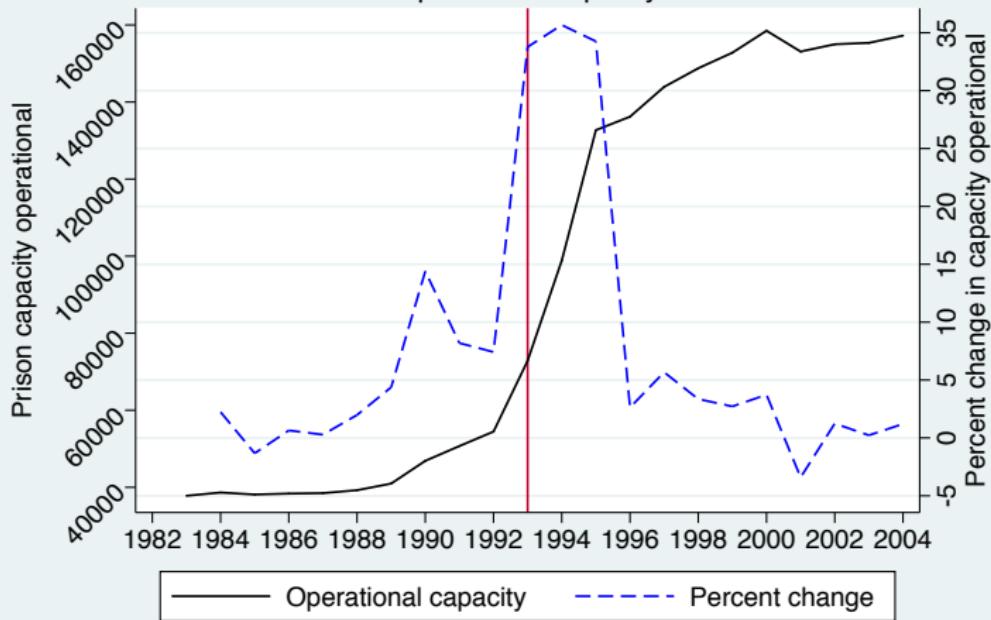
- Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
- Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
- Building of new prisons (private and public)

New prison construction

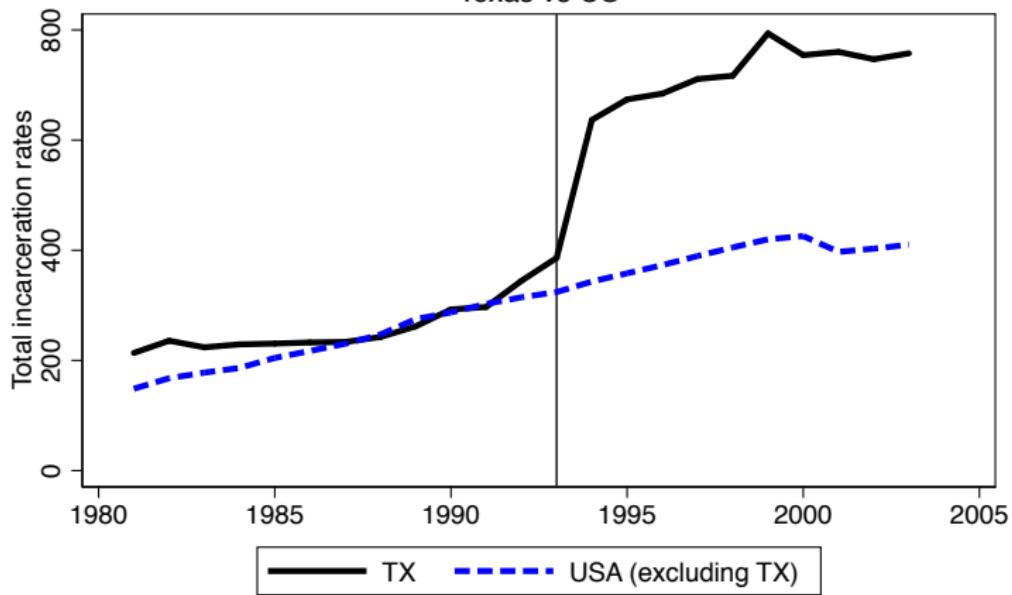


Texas prison growth

Operational capacity



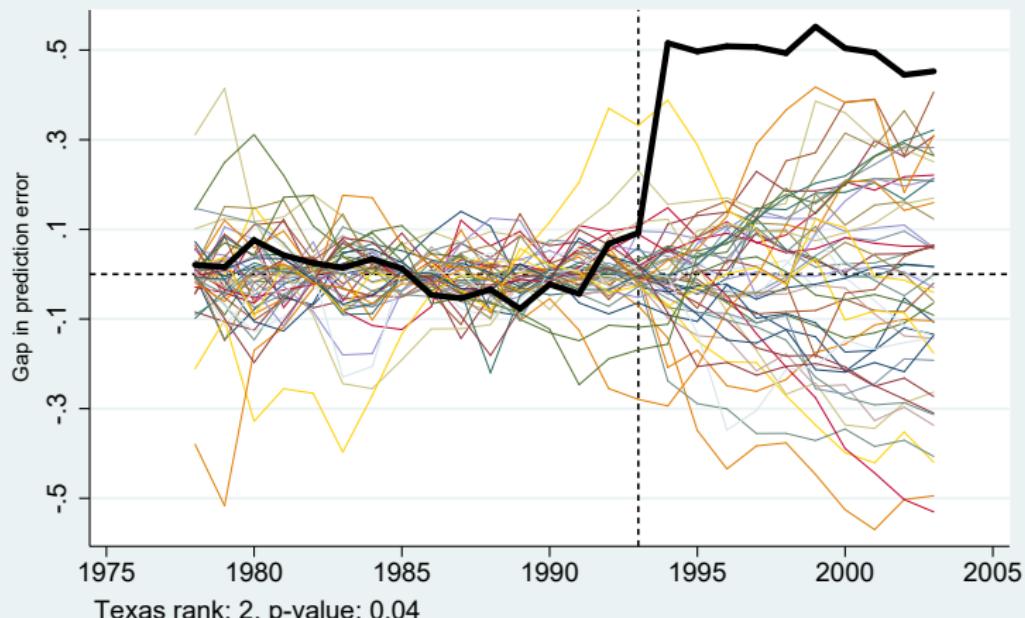
Total incarceration per 100 000 Texas vs US



1993 starts the prison expansion

Incarcerated persons per 100,000

1993 Treatment



Coding together

- Let's go to Mixtape Sessions repository now into /Labs/Texas
- I'll walk us through the Stata and R code so you understand the syntax and underlying logic
- But then I have us a practice assignment

Synth Opponents

- Synthetic control has opponents
- One criticism is people will say the weights are implausible
- But that's because they have their own "mental synth" model that tells them what weights should or shouldn't be
- Tell story about conference where Abadie presented Germany paper
- Synth tells you the weights; OLS you have to calculate them and no one does

Regression vs synth weights

TABLE 1 Synthetic and Regression Weights for West Germany

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.09	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	United Kingdom	0	0.06
Japan	0.16	0.19	United States	0.22	0.13

Notes: The synthetic weight is the country weight assigned by the synthetic control method. The regression weight is the weight assigned by linear regression. See text for details.

Comment about Negative Weights

- Matching principles favor non-negative weighting because similar groups are best comparisons for a treated group, not dissimilar ones
- But what if you're Michael Jordan and wanting to know his synthetic control – can you find a positively weighted average group of players that approximate him over his career if he is an outlier?
- You cannot with non-negative weights, but you could with negative weights

Comment about Negative Weights

- Negative weights would mean finding a truly horrible player and then negatively weighting him
- But if you are never willing to negatively weight, some questions you cannot answer
- Non-negative weights by Abadie are not the best fit – they are the best fit given you can't negatively weight
- Perhaps there is a compromise – the least worst negative weighting, or “augmented synthetic control”

Non-negative weighting and poor fit

"The applicability of the [ADH2010] method requires a sizable number of pre-intervention periods. The reason is that the credibility of a synthetic control depends upon how well it tracks the treated unit's characteristics and outcomes over an extended period of time prior to the treatment. **We do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small.** A sizable number of post-intervention periods may also be required in cases when the effect of the intervention emerges gradually after the intervention or changes over time." (my emphasis, Abadie, et al. 2015)

Introducing Augmented Synthetic Control

- Synthetic control has built in constraints forcing weights to be non-negative
- Convex hull constraint ensures that synth is a feasible counterfactual in that it is formed by a combination of control units similar on pre-intervention characteristics
- Improves the validity of the estimated effect as there exists interpolated comparison group; similar to common support concept
- But, the convex hull constraint reduces extrapolation bias from comparing dissimilar units, but at the cost of failing to find matches at all

What is augmented synthetic control?

- Eli Ben-Michael, Avi Feller and Jesse Rothstein present a modification to ADH in which they allow for negative weights, but only minimally so
- This model will “augment” the original synthetic control model by adjusting for pre-treatment imbalance using doubly robust bias adjustment
- Augmentation is conservative; it uses **penalized ridge regression** but with constraints such that the negative weighting is only to the convex hull, not to the center of the convex hull

Optimal weights

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

$Y'_0 w^*$ (i.e., optimally weighted donor pool) is the unit 1 “synthetic control”. We are hoping that \widehat{Y}_1^0 with $Y'_0 w^*$ based on “perfect fit” pre-treatment

Slight change in synth notation

- Assume that our outcome, Y_{jt} , follows a factor model where $m(\cdot)$ are pre-treatment outcomes:

$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

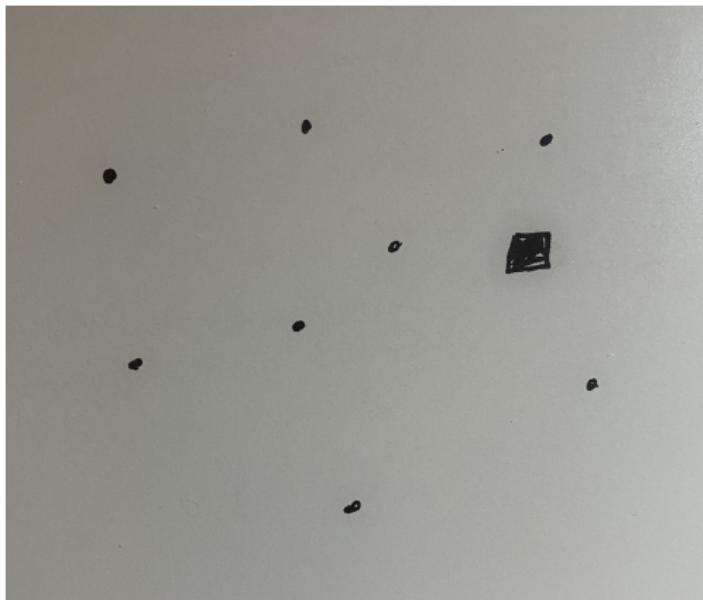
- Since $\widehat{m}(\cdot)$ estimates the post-treatment outcome, let's view it as estimated bias, analogous to bias correction for inexact matching from earlier (Abadie and Imbens 2011)

Bias correction

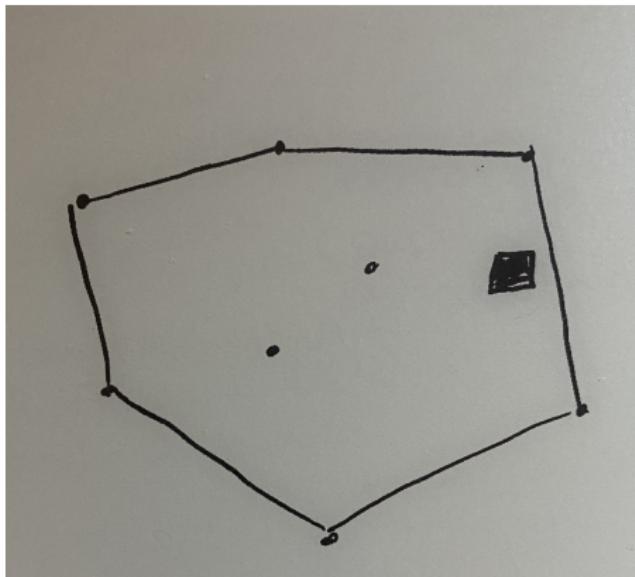
$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

- When the weights achieve exact balance, the bias of synthetic control decreases with T
- The intuition is that for a large T (T not transitory shocks), you achieve balance by balancing the latent parameter on the unobserved heterogeneity in our factor model

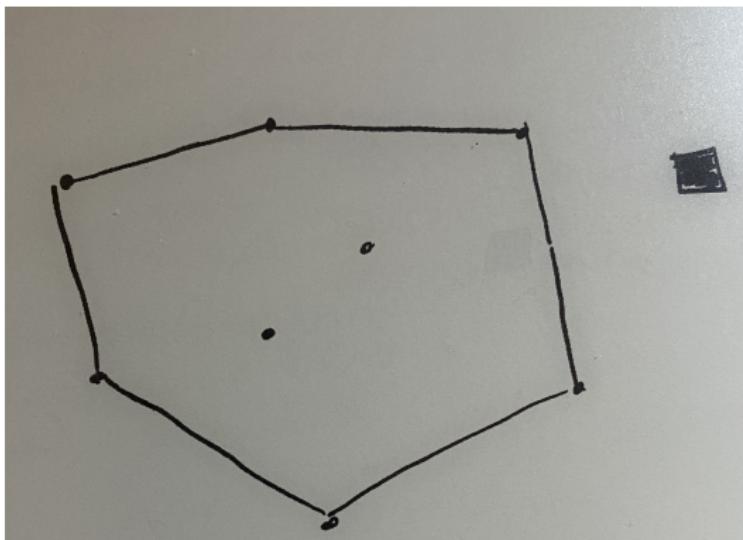
Treatment and control units



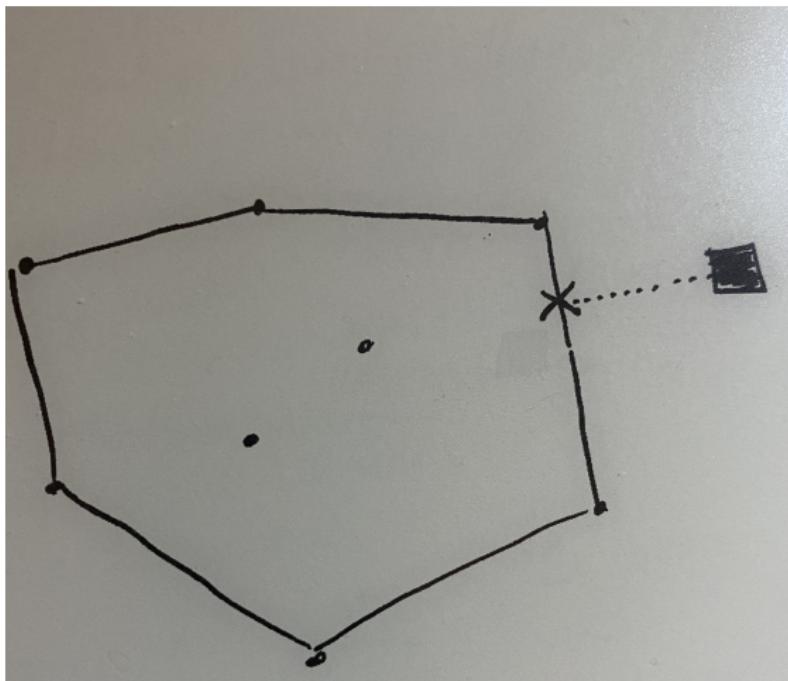
Convex hull – ideal for synth



Outside the convex hull bc of dimensionality



Outside the convex hull bc of dimensionality



Reweighting Original Synth

- Adjust the synthetic control approach to adjust for poor fit pre-treatment.
- We don't estimate new donor pool weights – we estimate the bias and then reweight the original weights, which introduces the negative weighting
- The augmented synthetic control estimator for Y_{jt}^0 is on the next slide, but the bias adjustment is very similar to Abadie and Imbens (2011) earlier (i.e., it subtracts out linearly)

Setup of the estimator

Let's adjust synthetic control for this bias. First we'll apply the **bias correction**. Then we'll do the doubly robust augmented **inverse probability weighting**. Let $Y_1^{aug,0}$ be the augmented potential outcome

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_j + \hat{m}(X_1) - \sum_{D_j=0} \hat{w}_j \hat{m}(X_j) \\ &= \hat{m}(X_1) + \sum_{D_j=0} \hat{w}_j (Y_j - \hat{m}(X_j)) \end{aligned}$$

Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left(\hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) Note how in the first line the traditional synthetic control weighted outcomes are corrected by the imbalance in a particular function of the pre-treatment outcomes \hat{m} .

Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left(\hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) Since \hat{m} estimates the post-treatment outcome, we can view this as an estimate of the bias due to imbalance, which is similar to how you address imbalance in matching with a bias correction formula (Abadie and Imbens 2011).

Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left(\hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) So if the bias is small, then synthetic control and augmented synthetic control will be similar because that interior term will be zero.

Interpreting line 2

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left(\hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (2) The second equation is equivalent to a double robust estimation which begins with an outcome model but then re-weights it to balance residuals.

Interpreting line 2

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left(\hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (2) The second equation has a connection to inverse probability weighting (they show this in an appendix)

Ridge Augmented SCM

$$\arg \min_{\eta_0, \eta} \frac{1}{2} \sum_{D_j=0} (Y_j - (\eta_0 + X'_j \eta))^2 + \lambda^{ridge} \|\eta\|_2^2$$

Here we estimate $\hat{m}(X_j)$ with ridge regularized linear model and penalty hyper parameter λ^{ridge} . We then adjust for imbalance using the $\hat{\eta}^{ridge}$ parameter as a weight on the outcome model itself.

Go back to that weighting but use the ridge parameters

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_j + \left(X_1 - \sum_{D_j=0} \hat{w}_j^{synth} X_j \right) \hat{\eta}^{ridge} \\ &= \sum_{D_j=0} \hat{w}_j^{aug} Y_j \end{aligned}$$

What you're trying to do is adjust with the \hat{w}_j^{aug} weights to improve balance.

The ridge weights are key to the augmentation

$$\hat{w}_j^{aug} = \hat{w}_j^{synth} + (X_j - X_0' \hat{w}_j^{synth})' (X_0' X_0 + \lambda I_{T_0})^{-1} X_i$$

The second term is adjusting the original synthetic control weights, w_j^{synth} for better balance. Again remember – we are trying to address the bias due to imbalance. You can achieve better balance, but at higher variance and can introduce negative weights.

Ridge will allow negative weights via extrapolation

$$\hat{w}_j^{aug} = \hat{w}_j^{synth} + (X_j - X_0' \hat{w}_j^{synth})' (X_0' X_0 + \lambda I_{T_0})^{-1} X_i$$

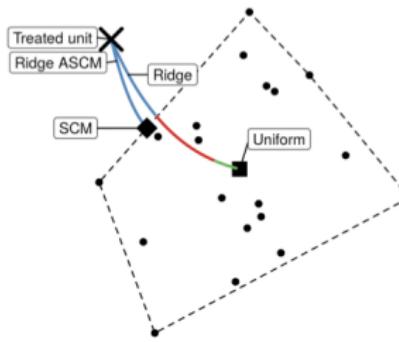
Relaxing the constraint from synth that weights be non-negative, as non-negative weights prohibit extrapolation. But we don't have synthetic control on the simplex, so we *must* extrapolate, otherwise synth will be biased.

Summarizing and some comments

- When the treated unit lies in the convex hull of the control units so that the synth weights exactly balance lagged outcomes, then SCM and Ridge ASCM are the same
- When synth weights do not achieve exact balance, Ridge ASCM will use negative weights to extrapolate from the convex hull to the control units
- The amount of extrapolation will be determined by how much imbalance we're talking about and the estimated hyperparameter $\hat{\lambda}^{ridge}$
- When synth has good pre-treatment fit or when λ^{ridge} is large, then adjustment will be small and the augmented weights will be close to the SCM weights

Intuition

Ridge begins at the center of control units, while Ridge ASCM begins at the synth solution. Both move towards an exact fit solution as the hyperparameter is reduced. It is possible to achieve the same level of balance with non-negative weights. Both ridge and Ridge ASCM extrapolate from the support of the data to improve pre-treatment fit relative to synth alone. Let's look at a picture!



- In convex hull
- Out of convex hull
- Weights in simplex

(a) Treated and control units with the convex hull marked as a dashed line. Ridge and Ridge ASCM estimates in solid.

Conformal Inference

Whereas Abadie, et al. used Fisher sharp null and randomization inference, newer synth models will conduct inference based on conformal inference (Chernozhukov et al. 2019)

We can get 95% point-wide confidence intervals, or use a jackknife method by Barber et al (2019)

Comparing four synth models

- Non-negative weighted synth vs three augmented synth models (no covariates, covariates, residualized covariates)
- Augmenting synth with a ridge outcome regression reduces bias relative to synth alone in application (but note at the cost of negative weights)
- This underscores the importance of the recommendation Abadie, et al. (2015) make which is that synth should be used in settings with excellent pre-treatment fit

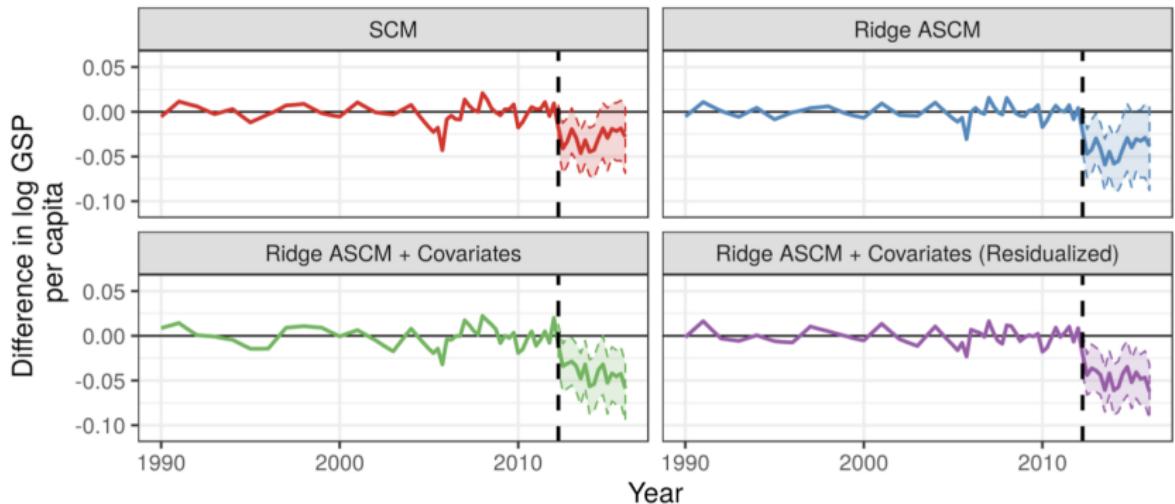
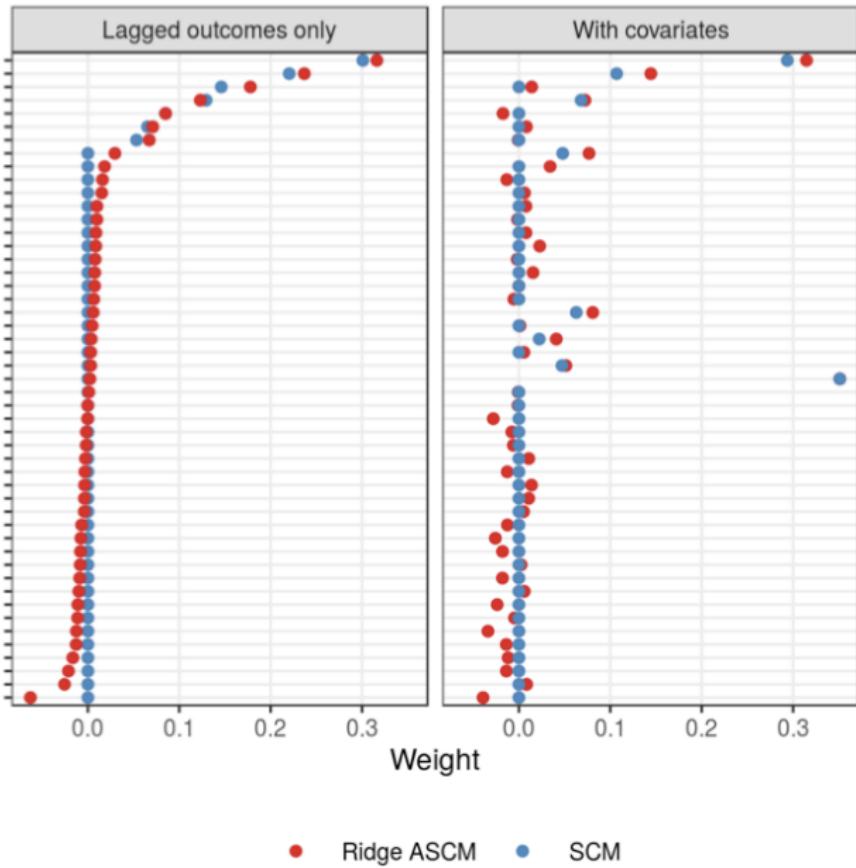


Figure 6: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.



Couple of minor points

- Hyper parameter chosen using cross validation
- This can be extended to auxiliary covariates as opposed to just lagged outcomes (section 6)

Concluding remarks

- Non-negative weights preclude extrapolation forcing us to use similar groups as synthetic controls, but it may also make the project impossible if our treatment group is too “unique”
- Ridge regression augmentation is like the “least worst” way to do negative weighting – it’s bias adjustment with weights estimated using ridge that shifts the convex hull the minimum distance needed to balance through negative weights
- Augmented synth will dominate synth in those instances by extrapolating outside the convex hull

More synthetic control

- Synthetic control remains an active area in econometrics and Athey and Imbens have contributed two new ones (matrix completion with nuclear norm regularization and synthetic difference-in-differences)
- MCNN can handle panel data and a variety of designs including staggered adoption and uses imputation based on vertical and horizontal regressions in pre-treatment period
- Synthetic control continues to evolve in such a way that it appears to be encroaching on difference-in-differences, but it's important to know that synth needs a longer pre-treatment time series than diff-in-diff which we will look at now

Roadmap

In Pursuit of the ATT

- Potential Outcomes

- Independence and Selection Bias

Unconfoundedness and Ignorable Treatment Assignment

- Exact and Inexact Matching

- Saturated Regressions

Synthetic control

- Interpolation with non-negative weighting

- Extrapolation with Conservative Negative Weighting

Difference-in-differences

- Four averages and three subtractions

- Covariates

- Model Misspecification

- Alternatives to TWFE

Conclusion

Types of evidence

Difference-in-differences does not rely on covariates or randomization, so the types of evidence needed shifts and can be somewhat more involved – the underlying estimator is itself simple, but the evidence is not

- You are building a case, the prosecutor before a judge and jury, always in battle with the defense attorney
- You are producing evidence, not proofs, and that evidence has particular broadly defined forms that can help you on the front end
- Your goal in my humble opinion should be reasonable and theoretically relevant falsifications with particular kinds of data visualization, starting with the event study

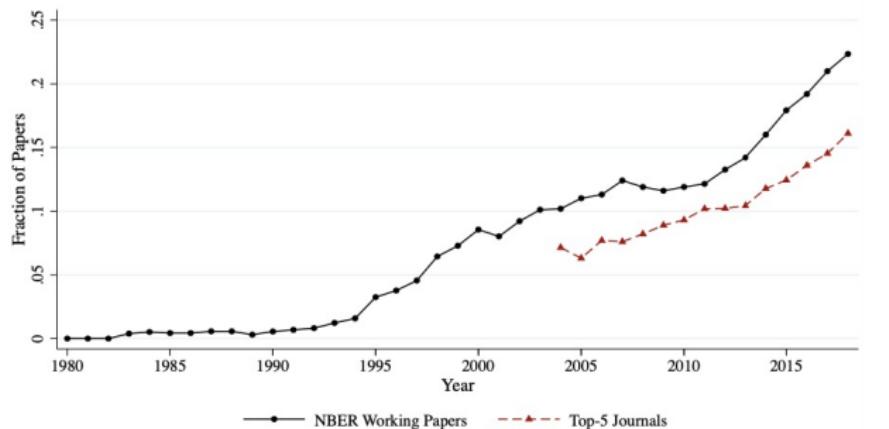
What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Mostly associated with “big shocks” happening in space over time

*“A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us.” – Daniel McFadden
(Nobel Laureate recipient with Heckman 2002)*

Figure: Currie, et al. (2020)

A: Difference-in-Differences



Difference-in-differences

Table: Treatment and Control, Before and After

Companies	Time	Outcome	D_1	D_2
Treatment	Before	$Y = L$	$T_L + D$	D
	After	$Y = L + T_L + D$		
				D
Comparison	Before	$Y = SV$	T_{SV}	
	After	$Y = SV + T_{SV}$		

$$\hat{\delta}_{did} = D + (T_L - T_{SV})$$

This simple method yields an unbiased estimate of D if $T_{SV} = \textcolor{red}{T}_L$

Relationship to OLS

- Orley Ashenfelter graduated from Princeton in the 1970s, takes a job in Washington DC and begins studying “job trainings programs”
- Empirical crisis in empirical macro and empirical labor back in the 1970s – Orley, David Card, Bob Lalonde, Alan Krueger at Princeton all helped bring attention to it and began pushing for solutions, one of which was RCTs in labor but also diff-in-diff as well as better instruments
- Listen to Orley explain the connection he made between two way fixed effects and difference-in-differences; it was born out of a need to explain OLS to an American bureaucrat

<https://youtu.be/WnB3EJ8K71g?t=126>

Equivalence

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist}$$

$$\hat{\delta} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

- Orley claims that the OLS estimator of δ and the “four averages and three subtractions” are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

DiD equation

Orley's "four averages and three subtractions"

$$\hat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

k are the people in the job training program, U are the untreated people not in the program, $Post$ is after the trainees took the class, Pre is the period just before they took the class, and $E[y]$ is mean earnings.

Does $\hat{\delta}$ equal the ATT? If so when? If not why not?

Potential outcomes and the switching equation

$$\hat{\delta} = \underbrace{\left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)}_{\text{Switching equation}} + \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\hat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\left[E[Y_k^0|Post] - E[Y_k^0|Pre] \right] - \left[E[Y_U^0|Post] - E[Y_U^0|Pre] \right]}_{\text{Non-parallel trends bias in 2x2 case}}$$

Identification through parallel trends

Parallel trends

Assume two groups, treated and comparison group, then we define parallel trends as:

$$E(\Delta Y_k^0) = E(\Delta Y_U^0)$$

In words: “The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees”.

It's in red because parallel trends is untestable and critically important to estimation of the ATT using any method, OLS or “four averages and three subtractions”

What is parallel trends

- Parallel trends assumes away the selection bias associated with comparisons
- The assumption is thought to be more plausible than simply assuming simple comparisons held equal
$$E[Y^0|D = 0] = E[Y^0|D = 1]$$
- But it is still a strong assumption, and differs from the assumptions have in the RCT which though also untestable, is nearly guaranteed by randomization
- Most of the hard part of the work involves the old fashioned detective work and the work of making good arguments with good exhibits (tables and figures)

Understanding parallel trends through worksheets

Before we move into regression, let's go through a simple exercise to really pin down these core ideas with simple calculations

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

OLS Specification

- Simple DiD equation will identify ATT under parallel trends
- But so will a particular OLS specification (two groups and no covariates)
- OLS was historically preferred because
 - OLS estimates the ATT under parallel trends
 - Easy to calculate the standard errors
 - Easy to include multiple periods
- People liked it also because of differential timing, continuous treatments and covariates, but those are more complex so we address them later

Minimum wages

- Card and Krueger (1994) have a famous study estimating causal effect (ATT) of minimum wages on employment
- Exploited a policy change in New Jersey between February and November in mid-1990s where minimum wage was increased, but neighbor PA did not
- Using DiD, they do not find a negative effect of the minimum wage on employment which is part of its legacy today, but I mainly present it to illustrate the history and the design principles



Binyamin Appelbaum



@BCAppelbaum



Replies to @BCAppelbaum

The Nobel laureate James Buchanan wrote in the Wall Street Journal that Card and Krueger were undermining the credibility of economics as a discipline. He called them and their allies "a bevy of camp-following whores."

3:49 PM · Mar 18, 2019



179



Reply



Share

[Read 18 replies](#)

Card on that study

"I've subsequently stayed away from the minimum wage literature for a number of reasons. First, it cost me a lot of friends. People that I had known for many years, for instance, some of the ones I met at my first job at the University of Chicago, became very angry or disappointed. They thought that in publishing our work we were being traitors to the cause of economics as a whole."

But let's listen to Orley's opinion about the paper's controversy at the time. <https://youtu.be/M0tbuRX4eyQ?t=1882>

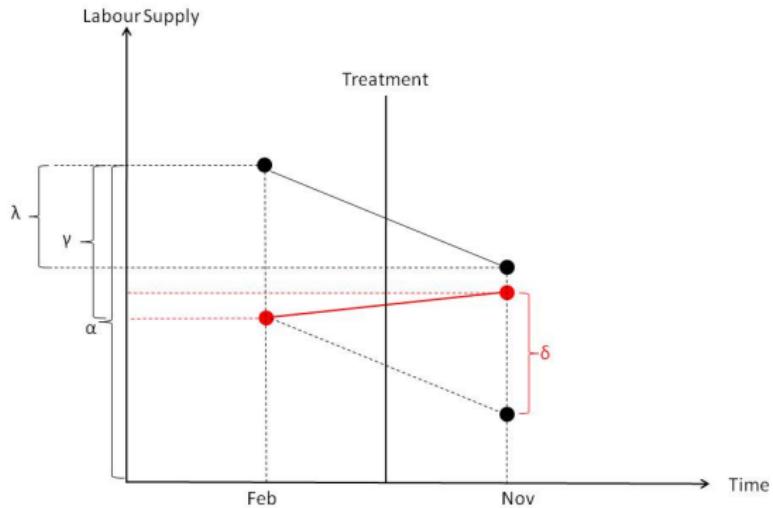
OLS specification of the DiD equation

- The correctly specified OLS regression is an interaction with time and group fixed effects:

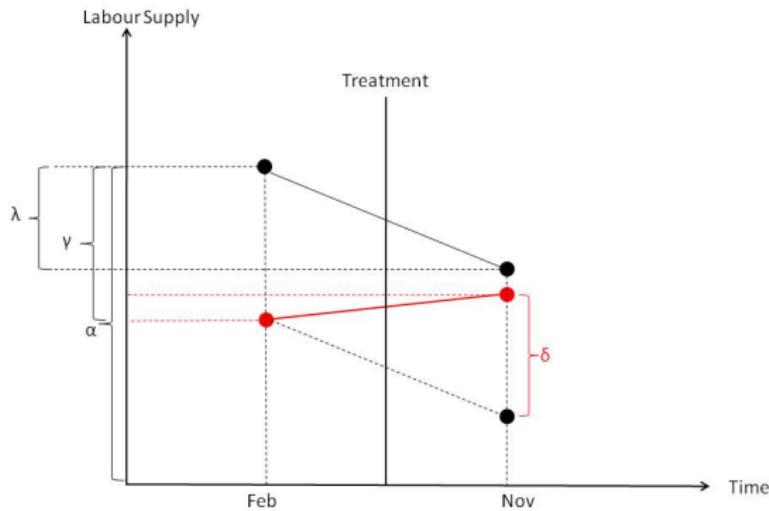
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DiD equation: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$

$$Y_{ist} = \alpha + \gamma N J_s + \lambda d_t + \delta (N J \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Notice how OLS is “imputing” $E[Y^0|D = 1, Post]$ for the treatment group in the post period? It is only “correct”, though, if parallel trends is a good approximation

Inference

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely underestimate the standard deviation of the estimators
- Standard errors are biased downward (i.e., too small, over reject)
- They proposed three solutions, but most only use one of them (clustering)

Inference

- 1 Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
- 2 Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)

Most people will simply cluster, but there are issues if you have too few clusters. They mention a third way but it's only a curiosity.

Main DiD assumptions

There are actually three DiD assumptions in the basic design, but you usually only hear about the first:

1. Parallel trends – concerns changes in Y^0 , one of which is a fictional change because the post treatment Y^0 doesn't exist for the treated
2. No anticipation (next slide)
3. SUTVA (slide after next)

No Anticipation

- No anticipation means that the treatment effect happens only at the time that the treatment occurs or after, but not before
 - **Example 1:** Tomorrow I win the lottery, but don't get paid yet. I decide to buy a new house today. That violates NA
 - **Example 2:** Next year, a state lets you drive without a driver license and you know it. But you can't drive without a driver license today. This satisfies NA.
- Requires that baseline in pre-period is untreated $Y = Y^0$ for treated group
- Crucial for DiD equation collapsing to ATT plus PT bias

SUTVA

- Stable Unit Treatment Value Assumption (Imbens and Rubin 2015) focuses on what happens when in our analysis we are combining units (versus defining treatment effects)
 1. **No Interference:** a treated unit cannot impact a control unit such that their potential outcomes change (unstable treatment value)
 2. **No hidden variation in treatment:** When units are indexed to receive a treatment, their dose is the same as someone else with that same index
 3. **Scale:** If scaling causes interference or changes inputs in production process, then #1 or #2 are violated
- Shifts from defining treatment effects to estimating them, which means being careful about who is the control group, how you define treatments and what questions can and cannot be answered with this method
- Again, we need our control group to be untreated ($Y = Y^0$) otherwise DiD doesn't equal ATT + PT

Violating parallel trends exercise

- Parallel trends are needed so we can impute the missing $E[Y^0|D = 1]$ with $E[Y^0|D = 0]$ either explicitly or implicitly
- Which means if parallel trends isn't true, then the imputation isn't correct and therefore estimates are biased
- To illustrate this, let's go through the document again – this time to tab 2

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

Event studies and pre-trends

- Parallel trends involves, Y^0 , specifically
 $\Delta E[Y^0|D = 1] = \Delta E[Y^0|D = 0]$
→ Notice that parallel trends is about Y^0 in other words, not Y^1
- We cannot verify the red term, because the change is post-treatment and thus counterfactual (fictional)
- But there are other non-red $\Delta E[Y^0|D = 1]$ that aren't fictional which we can investigate, but where?

"Pre-trends" are also $\Delta E[Y^0|D = 1]$, just non-fictional in nature

Testing for parallel pre-trends is a type of falsification for selection bias

Intuition behind event studies

- Checking pre-trends is **not** a test for parallel trends as there is no formal test for parallel trends
- It's akin to finding a smoking gun – maybe someone planted it, but dismiss it is irresponsible
- Do not overweight nor underweight parallel pre-trends
- Even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias is a parallel trends violation)

Event study regression

- Event studies have a simple OLS specification with only one treatment group and one never-treated group

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

- where D is an interaction of the treatment dummy with the calendar year
- Treatment occurs in year 0, no anticipation, drop baseline $t - 1$
- All “four averages and three differences” calculations will use $t - 1$ as “pre” which is why it must be untreated (no anticipation)
- Includes q leads or anticipatory effects and m lags or post treatment effects

Event study regression

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

Typically you'll plot the coefficients and 95% CI on all leads and lags
(binned or not, trimmed or not)

Under no anticipation, then you expect $\hat{\mu}$ coefficients to be zero, which gives you confidence that parallel trends holds (but is not a guarantee, and there are still specification issues – see Jon Roth's work)

Under parallel trends, $\hat{\delta}$ are estimates of the ATT at points in time

Medicaid and Affordable Care Act example



Volume 136, Issue 3
August 2021

< Previous Next >

Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data [Get access >](#)

Sarah Miller, Norman Johnson, Laura R Wherry

The Quarterly Journal of Economics, Volume 136, Issue 3, August 2021, Pages 1783–1829,

<https://doi.org/10.1093/qje/qjab004>

Published: 30 January 2021

[Cite](#) [Permissions](#) [Share ▾](#)

Abstract

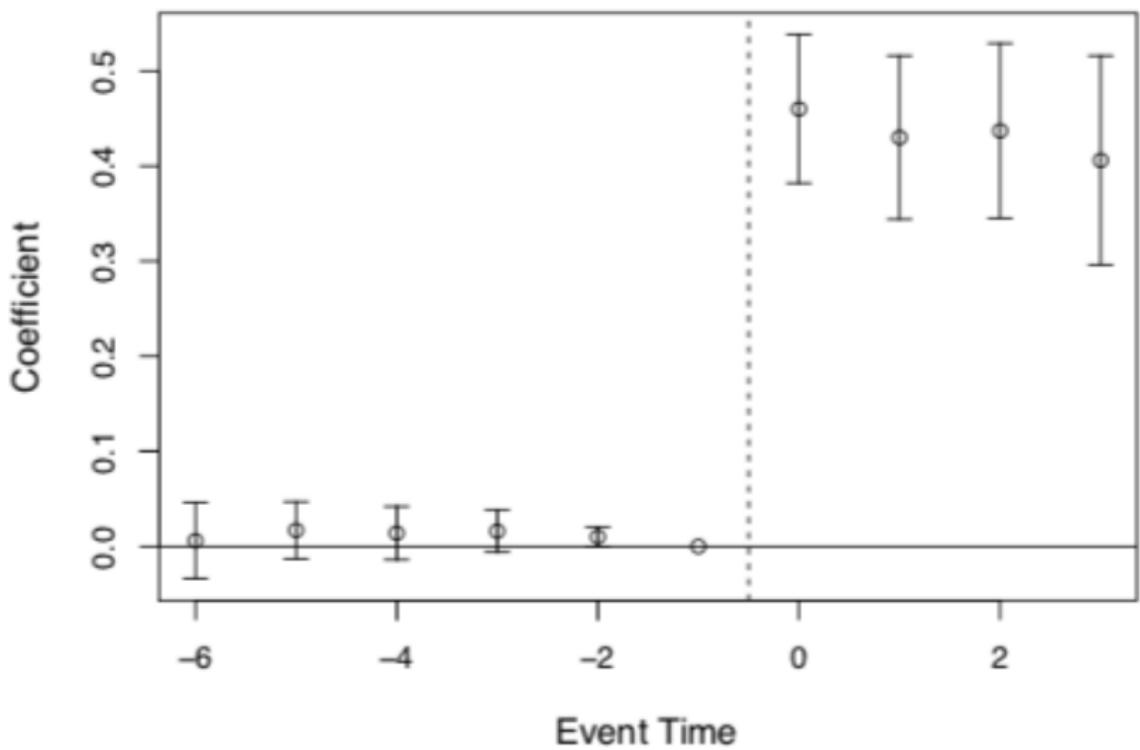
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults most likely to benefit using survey information on socioeconomic status, citizenship status, and public program participation. We find that prior to the ACA expansions, mortality rates across expansion and nonexpansion states trended similarly, but beginning in the first year of the policy, there were significant reductions in mortality in states that opted to expand relative to nonexpander states. Individuals in expansion states experienced a 0.132 percentage point decline in annual mortality, a 9.4% reduction over the sample mean, as a result of the Medicaid expansions. The effect is driven by a reduction in disease-related deaths and grows over time. A variety of alternative specifications, methods of inference, placebo tests, and sample definitions confirm our main result.

JEL: H75 - State and Local Government: Health; Education; Welfare; Public Pensions, I13 - Health Insurance, Public and Private, I18 - Government Policy; Regulation; Public Health

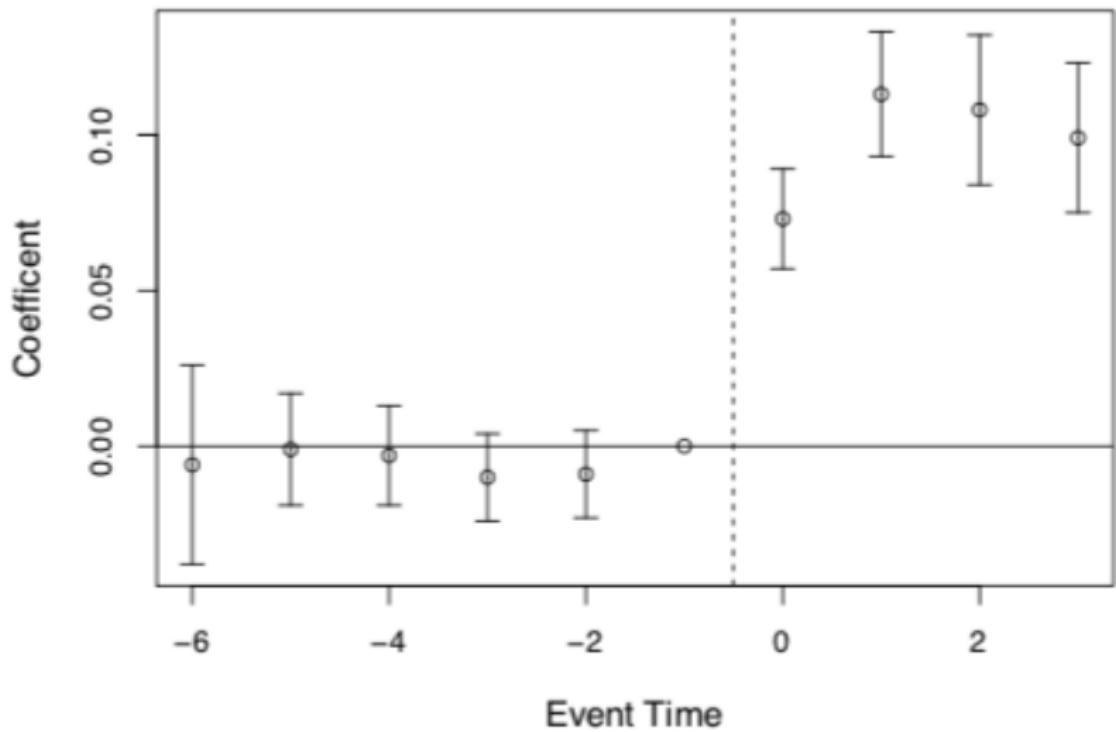
Issue Section: Article

Types of evidence

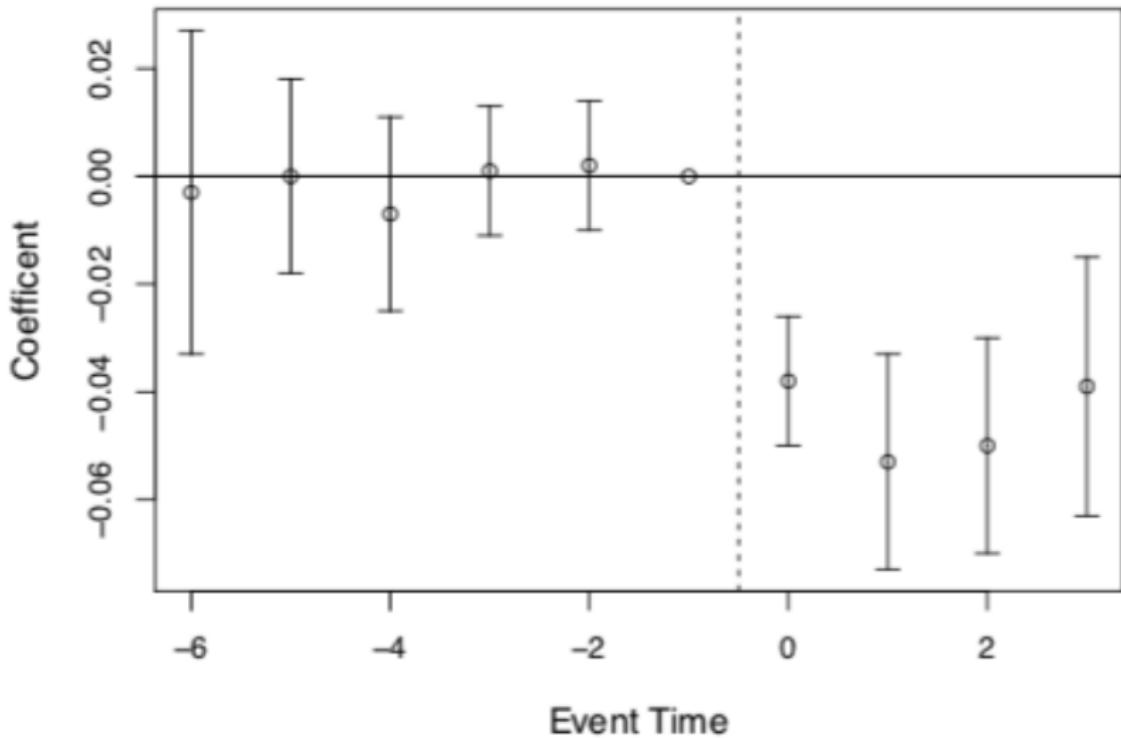
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Main Results** – Show your main results (the point of the paper)
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Mechanisms** – Find some reason explaining why the treatment affects the outcome via some “mechanism”
- **Event study** – Show leads and lags on mortality



(a) Medicaid Eligibility



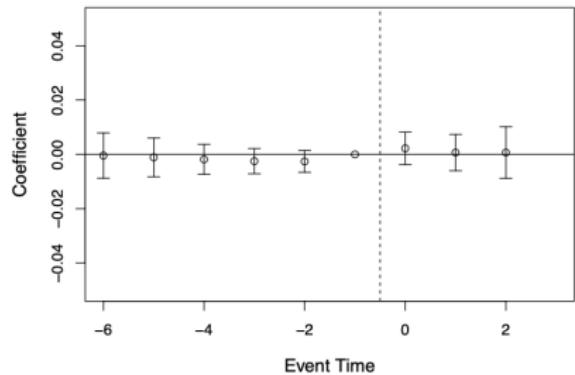
(b) Medicaid Coverage



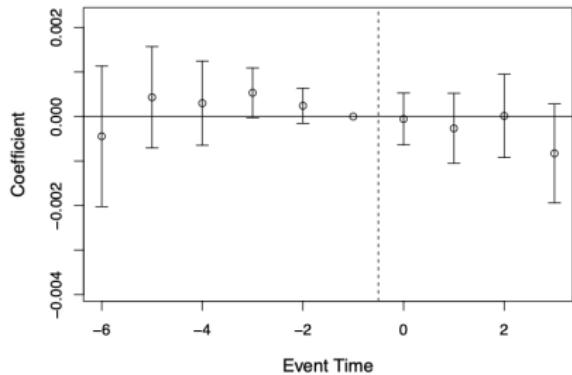
(c) Uninsured

Falsifications on elderly

Age 65+ in 2014



(c) Medicaid Coverage



(d) Annual Mortality

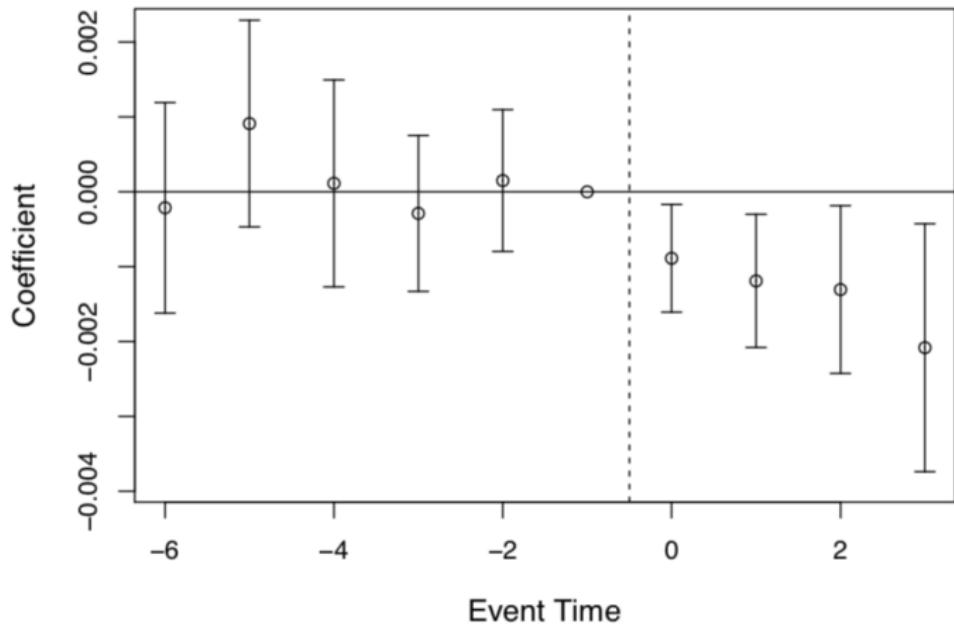


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on annual mortality

Why covariates?

- The inclusion of covariates in diff-in-diff models is not about trying to find random variation in the treatment within values of the dimension of X_n as we discussed earlier
- It is *only* to re-establish parallel trends
- This is itself different than how covariates will be used in synthetic control, too – probably the least understood element of diff-in-diff

Correcting the missingness problem

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

We were always missing Y^0 values for the treatment group units, but parallel trends allowed us to impute it using the change in $[Y^0]|D = 0$ as a guide

But if that trend is not a good guide, then we cannot.

Conditional parallel trends

The DiD equation yields:

$$\begin{aligned}\hat{\delta} &= \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right) \\ &= \text{ATT} + \text{Non-parallel trends bias}\end{aligned}$$

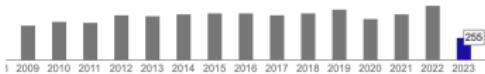
If we believe that conditional on covariates, parallel trends holds, but only within values of X , then there are methods we can use that incorporate covariates into the DiD equation and unbiasedness returns

The inclusion of covariates has particular regression specifications, plus there are alternative methods too, and we will review them

Outcome Regression Paper

Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme

Authors James J Heckman, Hidehiko Ichimura, Petra E Todd
Publication date 1997/10/1
Journal The review of economic studies
Volume 64
Issue 4
Pages 605-654
Publisher Wiley-Blackwell
Description This paper considers whether it is possible to devise a nonexperimental procedure for evaluating a prototypical job training programme. Using rich nonexperimental data, we examine the performance of a two-stage evaluation methodology that (a) estimates the probability that a person participates in a programme and (b) uses the estimated probability in extensions of the classical method of matching. We decompose the conventional measure of programme evaluation bias into several components and find that bias due to selection on unobservables, commonly called selection bias in econometrics, is empirically less important than other components, although it is still a sizeable fraction of the estimated programme impact. Matching methods applied to comparison groups located in the same labour markets as participants and administered the same questionnaire eliminate much of the bias as conventionally ...
Total citations Cited by 9508



Heckman, Ichimura and Todd (1997) is Petra and Hide's most cited paper and Heckman's second most cited!

Doubly Robust Paper

Doubly robust difference-in-differences estimators

Authors Pedro HC Sant'Anna, Jun Zhao

Publication date 2020/11/1

Journal Journal of Econometrics

Volume 219

Issue 1

Pages 101-122

Publisher North-Holland

Description This article proposes doubly robust estimators for the average treatment effect on the treated (ATT) in difference-in-differences (DID) research designs. In contrast to alternative DID estimators, the proposed estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified. We also derive the semiparametric efficiency bound for the ATT in DID designs when either panel or repeated cross-section data are available, and show that our proposed estimators attain the semiparametric efficiency bound when the working models are correctly specified. Furthermore, we quantify the potential efficiency gains of having access to panel data instead of repeated cross-section data. Finally, by paying particular attention to the estimation method used to estimate the nuisance parameters, we show that one can sometimes construct doubly robust DID ...

Total citations Cited by 398



Sant'Anna and Zhao (2020) is Pedro's second most cited paper

Doubly Robust Difference-in-differences

- DR models control for covariates twice – once using the propensity score, once using outcomes adjusted by regression – and are unbiased so long as:
 - The regression specification for the outcome is correctly specified
 - The propensity score specification is correctly specified
- Sant'Anna and Zhao (2020) incorporated DR into DiD by combining inverse probability weighting and outcome regression into a single DiD model
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so it merits close study

Identification assumptions I: Data

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is possible but assumes stationarity which is a kind of stability assumption, but I'll use panel representation.

Cross-sections will be potentially violated with changing sample compositions (e.g., the Napster example).

Identification assumptions II: Modification to parallel trends

Assumption 2: Conditional parallel trends

Counterfactual trends for the treatment group are the same as the control group for all values of X

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Identification assumptions III: Common support

Assumption 3: Common support

For some $e > 0$, the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on X is $\leq 1 - e$.

Heckman, et al doesn't use the propensity score so we need a more general expression of support

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997 (will require correct model too)
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005) (will require correct model too)

Outcome regression

This is the Heckman, et al. (1997) approach where the potential outcome evolution for the treatment group is imputed with a regression based only on X_b for the control group *only*

$$\hat{\delta}^{OR} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i)) \right]$$

where \bar{Y} is the sample average of Y among units in the treatment group at time t and $\hat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$.

Outcome regression

$$\hat{\delta}^{OR} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i)) \right]$$

1. Regress changes ΔY on X among untreated groups using baseline covariates only
2. Get fitted values of the regression using all X from $D = 1$ only.
Average those
3. Calculate change in this fitted Y among treated with the average fitted values

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\hat{\delta}^{ipw} = \frac{1}{E_N[D]} E \left[\frac{D - \hat{p}(X)}{1 - \hat{p}(X)} (Y_1 - Y_0) \right]$$

where $\hat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

These models cannot be ranked

- Outcome regression needs $\hat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\hat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified

TWFE

Consider our earlier TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose three *more* assumptions

Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

$$E[Y_1^1 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

Conditional parallel trends implies

$$E[Y_1^0 - Y_0^0 | D = 1, X] = E[Y_1^0 - Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] - E[Y_0^0 | D = 1, X] = E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0^0 | D = 1, X] + E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0 | D = 1, X] + E[Y_1 | D = 0, X] - E[Y_0 | D = 0, X]$$

Switching equation substitution

Last line from the switching equation. This gives us:

$$E[Y_1^0 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Now compare this with our earlier Y^1 expression

$$E[Y_1^1 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

We can define our target parameter, the ATT, now in terms of the fixed effects representation

Collecting terms

TWFE representation of our conditional expectations of the potential outcomes

$$E[Y_1^1|D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$

$$E[Y_1^0|D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

Substitute these into our target parameter

$$\begin{aligned} ATT &= E[Y_1^1|D = 1, X] - E[Y_1^0|D = 1, X] \\ &= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X) \\ &= \delta + (\theta_1 X - \theta_2 X) \end{aligned}$$

What if $\theta_1 X \neq \theta_2 X$?

Assumption 4: Homogeneous treatment effects in X

TWFE requires homogenous treatment effects in X (i.e., the treatment effect is the same for all X)

If X is sex, then effects are the same for males and females.

If X is continuous, like income, then the effect is the same whether someone makes \$1 or \$1 million.

X-specific trends

TWFE also places restrictions on covariate trends for the two groups too. Take conditional expectations of our TWFE equation.

$$E[Y_1|D = 1] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}$$

$$E[Y_0|D = 1] = \alpha_1 + \alpha_3 + \theta X_{10}$$

$$E[Y_1|D = 0] = \alpha_1 + \alpha_2 + \theta X_{01}$$

$$E[Y_0|D = 0] = \alpha_1 + \theta X_{00}$$

X-specific trends

Now take the DiD formula:

$$\delta^{DD} = \left((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10}) \right) - \left((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00}) \right)$$

Eliminating terms, we get:

$$\delta^{DD} = \delta + (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})$$

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

We need “no X -specific trends” for the treatment group (assumption 5) and comparison group (assumption 6)

Intuition: No X -specific trends means the evolution of potential outcome Y^0 is the same regardless of X . This would mean you cannot allow rich people to be on a different trend than poor people, for instance.

Without these six, in general TWFE will not identify ATT.

Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
 1. Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
 2. Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
 3. TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one

Double Robust DiD

$$\delta^{dr} = E \left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E \left[\frac{p(X)(1-D)}{(1-p(X))} \right]} \right) (\Delta Y - \mu_{0,\Delta}(X)) \right]$$

$p(x)$: propensity score model

$$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}(X)$, where $\mu(X)$ is a model for

$$m_{d,t} = E[Y_t | D = d, X = x]$$

So that means $\mu_{0,\Delta}$ is just the control group's change in average Y for each $X = x$

Double Robust DiD

$$\delta^{dr} = E \left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E \left[\frac{p(X)(1-D)}{(1-p(X))} \right]} \right) (\Delta Y - \mu_{0,\Delta}(X)) \right]$$

Notice how the model controls for X : you're weighting the adjusted outcomes using the propensity score

The reason you control for X twice is because you don't know which model is right. DR DiD frees you from making a choice without making you pay too much for it

Efficiency

- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DiD estimator is also DR for inference
- Let's skip to Monte Carlos

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

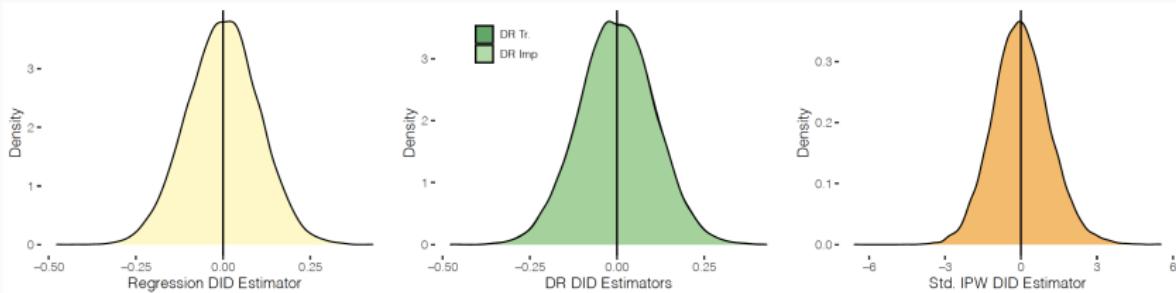
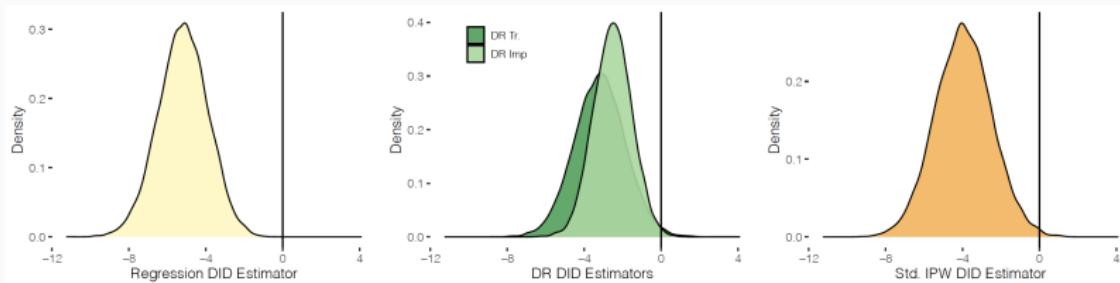


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



Two-way fixed effects

- When working with panel data, the so-called “two-way fixed effects” (TWFE) estimator was the workhorse estimator
- And from the start, it was used with diff-in-diff
- But at the start, it wasn’t staggered adoption – it was a much simpler design in which a group was treated in one year, and a comparison group wasn’t

Two OLS Models

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist} \quad (3)$$

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist} \quad (4)$$

First equation is used for simple designs when everyone is treated at once; second equation was used when different groups were treated at different times ("differential timing")

First equation works; second one only sometimes works

Discussion of estimate

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- If you estimate the above with OLS and there are more than one treatment “groups” or cohorts, then what does $\hat{\delta}$ estimate?
- TWFE estimates a coefficient that is a weighted average over all difference-in-differences equations (“four averages and three subtractions”)
- But some of them are “illegal” and introduce bias and therefore incorrect specifications must be become salient and avoided
- How many of these diff-in-diff equations are there when you have more than one treatment group?

K^2 distinct DDs

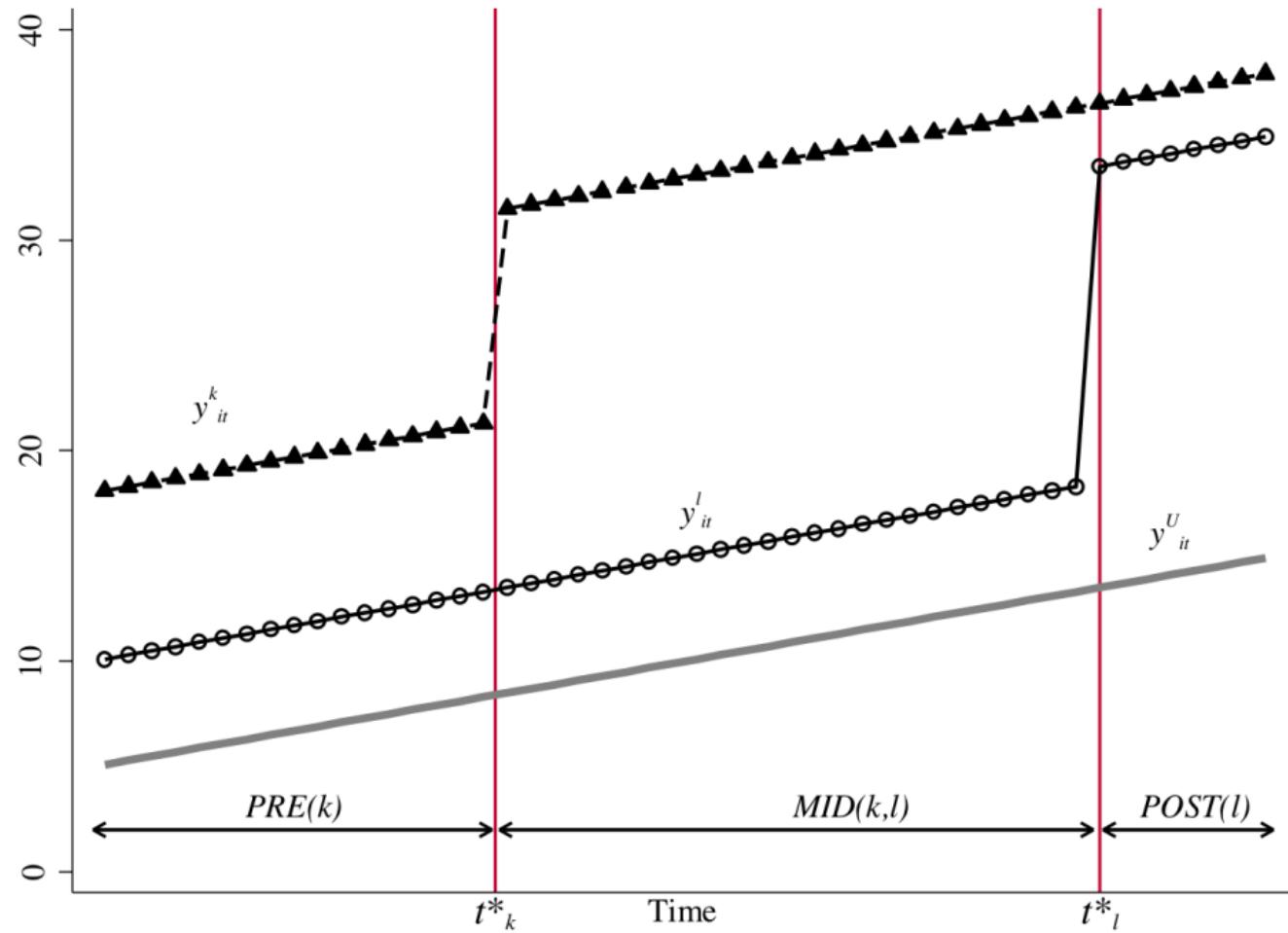
Let's look at 3 timing groups (a, b and c) and one untreated group (U).
With 3 timing groups, there are 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

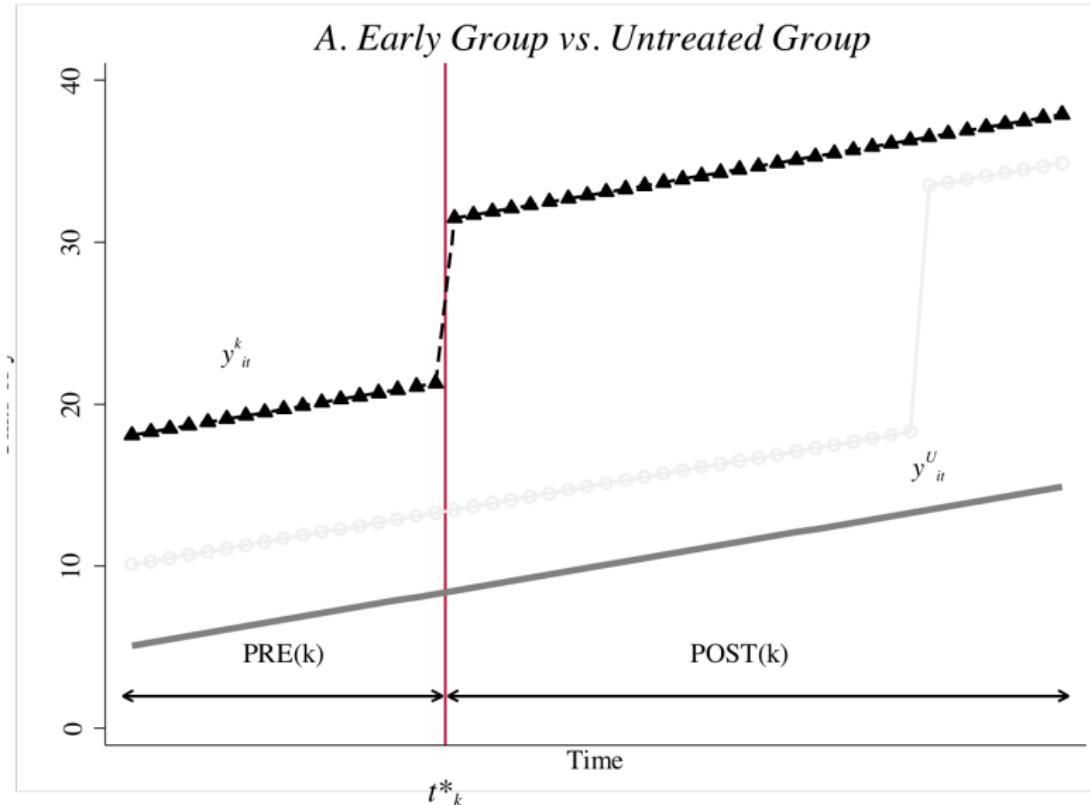
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

- Let there be two treatment groups (k, l) and one untreated group (U)
- k, l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote \bar{D}_k as the share of time each group spends in treatment status
- Denote $\hat{\delta}_{jb}^{2x2}$ as the canonical 2×2 DD estimator for groups j and b where j is the treatment group and b is the comparison group

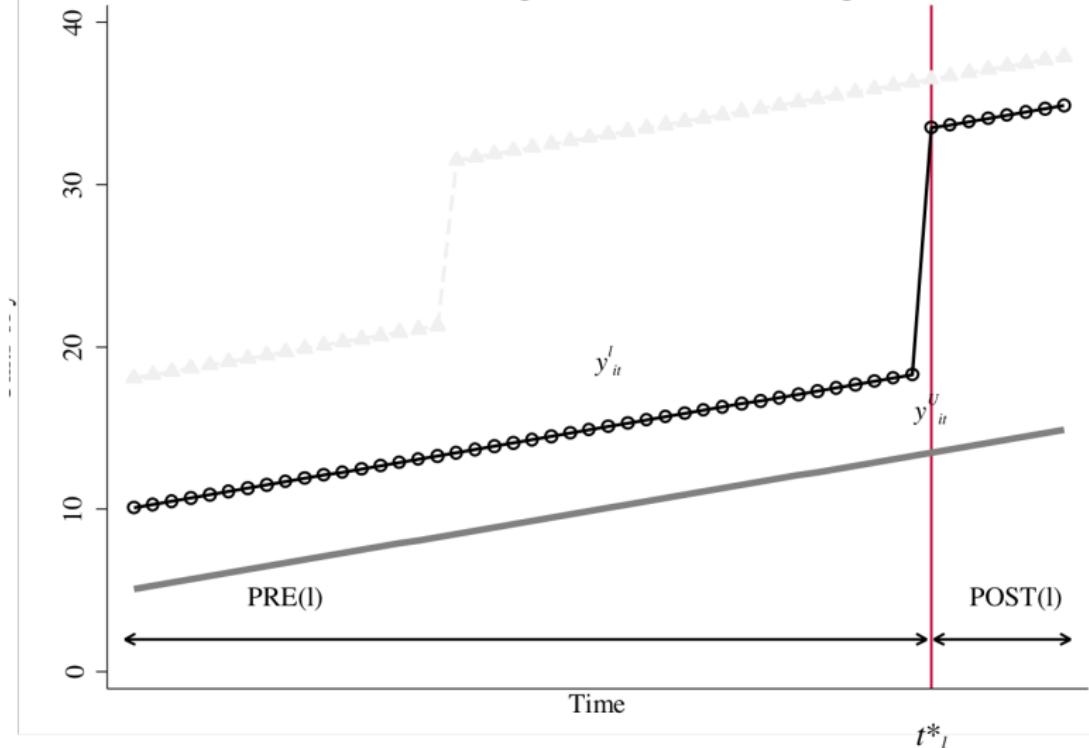


$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$

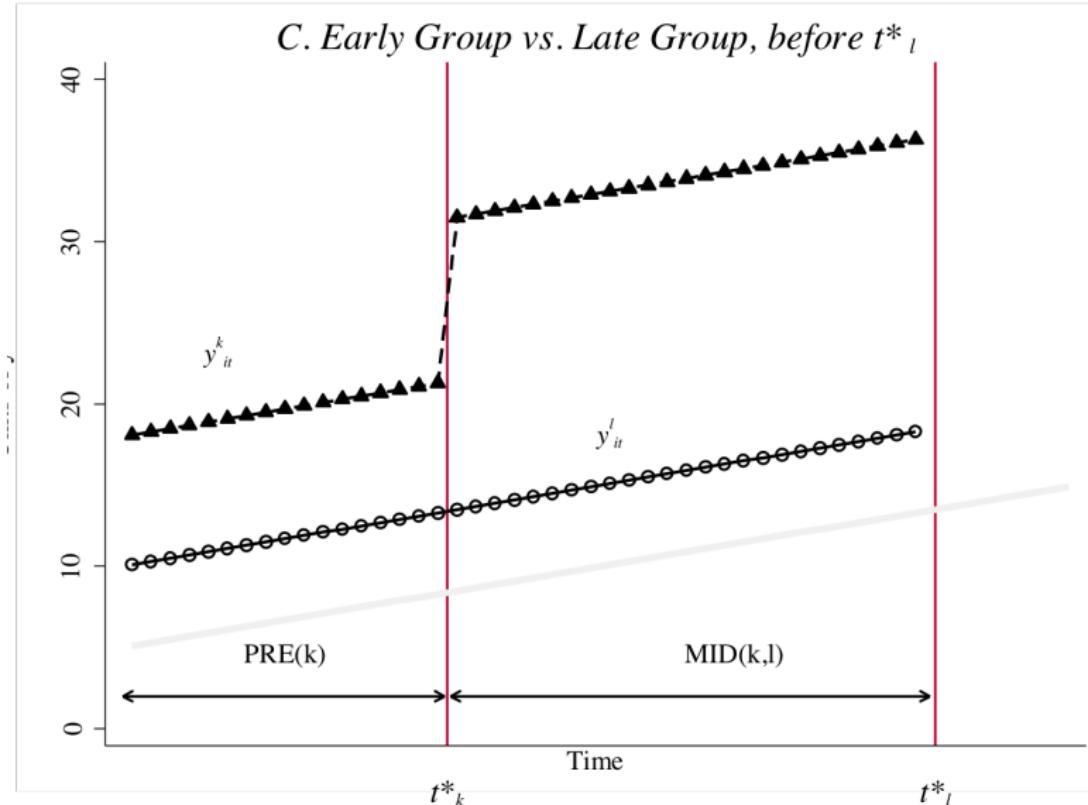


$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_l^{post(l)} - \overline{y}_l^{pre(l)} \right) - \left(\overline{y}_U^{post(l)} - \overline{y}_U^{pre(l)} \right)$$

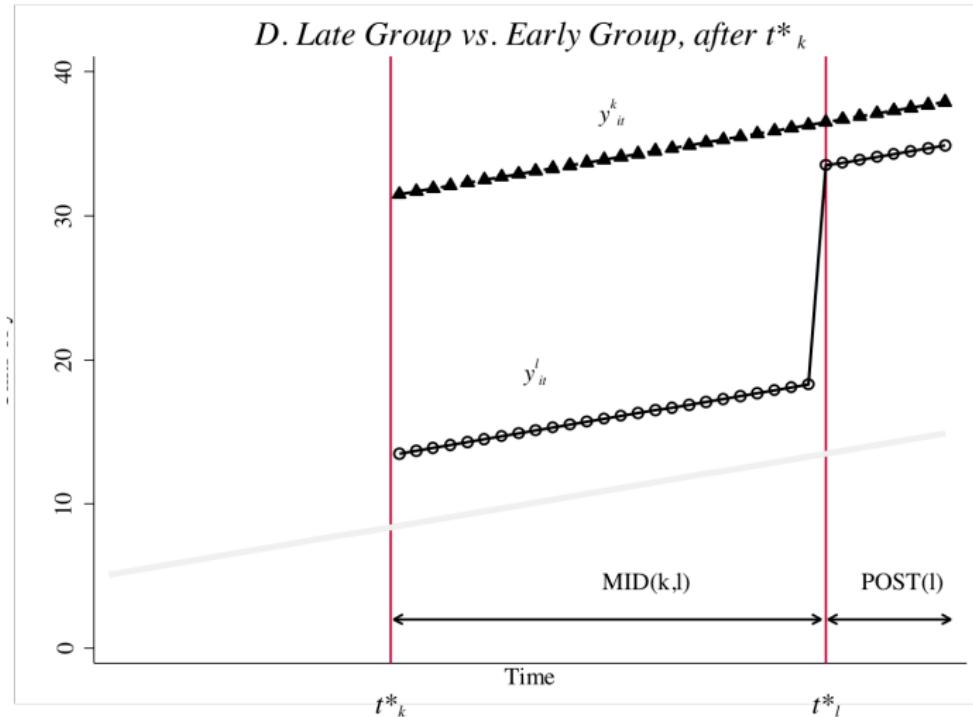
B. Late Group vs. Untreated Group



$$\delta_{kl}^{2x2,k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{Pre(k,l)} \right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k,l)} \right)$$



$$\delta_{lk}^{2x2,l} = \left(\bar{y}_l^{POST(k,l)} - \bar{y}_l^{MID(k,l)} \right) - \left(\bar{y}_k^{POST(k,l)} - \bar{y}_k^{MID(k,l)} \right)$$



Bacon decomposition

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

TWFE estimate of $\widehat{\delta}$ is equal to a weighted average over all group 2x2
(of which there are 4 in this example)

$$\widehat{\delta}^{TWFE} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \widehat{\delta}_{lk}^{2x2,l} \right]$$

where that first 2x2 combines the k compared to U and the l to U
(combined to make the equation shorter)

Third, the Weights

$$\begin{aligned}s_{ku} &= \frac{n_k n_u \bar{D}_k (1 - \bar{D}_k)}{\widehat{Var}(\tilde{D}_{it})} \\ s_{kl} &= \frac{n_k n_l (\bar{D}_k - \bar{D}_l) (1 - (\bar{D}_k - \bar{D}_l))}{\widehat{Var}(\tilde{D}_{it})} \\ \mu_{kl} &= \frac{1 - \bar{D}_k}{1 - (\bar{D}_k - \bar{D}_l)}\end{aligned}$$

where n refer to sample sizes, $\bar{D}_k(1 - \bar{D}_k)$ ($\bar{D}_k - \bar{D}_l$) $(1 - (\bar{D}_k - \bar{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.

Weights discussion

- Two things to note:
 - More units in a group, the bigger its 2x2 weight is
 - Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
 - $\bar{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
 - $\bar{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
 - $\bar{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
 - $\bar{D} = 0.6$. Then $0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the “treated on treated” weights (i.e., $\bar{D}_k - \bar{D}_l$)
- Same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\bar{D}_k - \bar{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

Back to TWFE

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- So we know that the estimate is a weighted average over all “four averages and three subtractions” but is that good or bad?
- It’s good if it’s unbiased; it’s bad if it isn’t, and the decomposition doesn’t tell us which unless we replace realized outcomes with potential outcomes
- Bacon shows that TWFE estimate of δ needs two assumptions for unbiasedness:
 1. variance weighted parallel trends are zero and
 2. no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs (not just weighted average of DiDs)

Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

$$\begin{aligned}\hat{\delta}_{kU}^{2x2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\ \hat{\delta}_{kl}^{2x2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)\end{aligned}$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated yet).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions we get:

$$\widehat{\delta}_{lk}^{2x2} = ATT_{l,Post(l)} + \underbrace{\Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \widehat{\delta}_{kl}^{2x2,l} \right]$$

where we will make these substitutions

$$\begin{aligned}\widehat{\delta}_{kU}^{2x2} &= ATT_k(Post) + \Delta Y_l^0(Post, Pre) - \Delta Y_U^0(Post, Pre) \\ \widehat{\delta}_{kl}^{2x2,k} &= ATT_k(Mid) + \Delta Y_l^0(Mid, Pre) - \Delta Y_l^0(Mid, Pre) \\ \widehat{\delta}_{lk}^{2x2,l} &= ATT_l Post(l) + \Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

$$p \lim_{n \rightarrow \infty} \hat{\delta}_{n \rightarrow \infty}^{TWFE} = VWATT + VWPT - \Delta ATT$$

- Notice the number of assumptions needed even to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Model can flip signs (does not satisfy a “no sign flip property”)

Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- I'll impose "unit level parallel trends", which is much stronger than we need (we only need average parallel trends)
- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects
- Two types of situations: constant versus dynamic treatment effects

[https://docs.google.com/spreadsheets/d/
1dI67eNNE2zrX4KrkoFvej-cKxqHkM8yJdMpD-0uE4q8/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1dI67eNNE2zrX4KrkoFvej-cKxqHkM8yJdMpD-0uE4q8/edit?usp=sharing)

Constant vs Dynamic Treatment Effects

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	10	0	0	0
1988	10	0	0	0
1989	10	0	0	0
1990	10	0	0	0
1991	10	0	0	0
1992	10	8	0	0
1993	10	8	0	0
1994	10	8	0	0
1995	10	8	0	0
1996	10	8	0	0
1997	10	8	0	0
1998	10	8	6	0
1999	10	8	6	0
2000	10	8	6	0
2001	10	8	6	0
2002	10	8	6	0

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

- Heterogenous treatment effects across time and across groups
- Cells are called “group-time ATT” (Callaway and Sant’anna 2020) or “cohort ATT” (Sun and Abraham 2020)
- ATT is weighted average of all cells and +82 with uniform weights 1/60

Estimation

Estimate the following equation using OLS:

$$Y_{ist} = \alpha_i + \gamma_t + \delta D_{it} + \varepsilon_{ist}$$

Table: Estimating ATT with different models

Truth	(TWFE)	(CS)	(SA)	(BJS)
\widehat{ATT}	82	-6.69***		

The sign flipped. Why? Because of extreme dynamics (i.e., $-\Delta ATT$)

Bacon decomposition

Table: Bacon Decomposition (TWFE = -6.69)

DD Comparison	Weight	Avg DD Est
Earlier T vs. Later C	0.500	51.800
Later T vs. Earlier C	0.500	-65.180

T = Treatment; C= Comparison

$$(0.5 * 51.8) + (0.5 * -65.180) = -6.69$$

While large weight on the “late to early 2x2” is suggestive of an issue, these would appear even if we had constant treatment effects

Callaway and Sant'Anna 2020

CS is a DiD estimator used for estimating and then summarizing smaller ATT parameters under differential timing and conditional parallel trends into more policy relevant ATT parameters (either dynamic or static)

Along with Goodman-Bacon and Sun and Abraham, CS won paper of the year for Journal of Econometrics; coauthor was Pedro Sant'Anna, former Microsoft economist

When is CS used

Just some examples of when you'd want to consider it:

1. When treatment effects differ depending on when it was adopted
2. When treatment effects change over time
3. When shortrun treatment effects are different than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

CS estimates the ATT by identifying smaller causal effects and aggregating them using non-negative weights

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

Each cell contains that group's ATT(g,t)

$$ATT(g, t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

CS identifies all feasible ATT(g,t)

Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Group-time ATT estimates are simple (weighted) differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Allows us ways to choose our aggregations
- Inference is the bootstrap

Notation

- T periods going from $t = 1, \dots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t .
- C is also binary and indicates a control group unit equalling one if “never treated” (can be relaxed though to “not yet treated”) → Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = \Pr(G_g = 1 | X, G_g + C = 1)$$

Assumptions

Assumption 1: Sampling is iid (panel data, but repeated cross-sections are possible)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

CS Estimator (the IPW version)

$$ATT(g, t) = E \left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E \left[\frac{\hat{p}(X)C}{1-\hat{p}(X)} \right]} \right) (Y_t - Y_{g-1}) \right]$$

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. CS uses the never-treated or the not-yet-treated as controls but never the already-treated

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

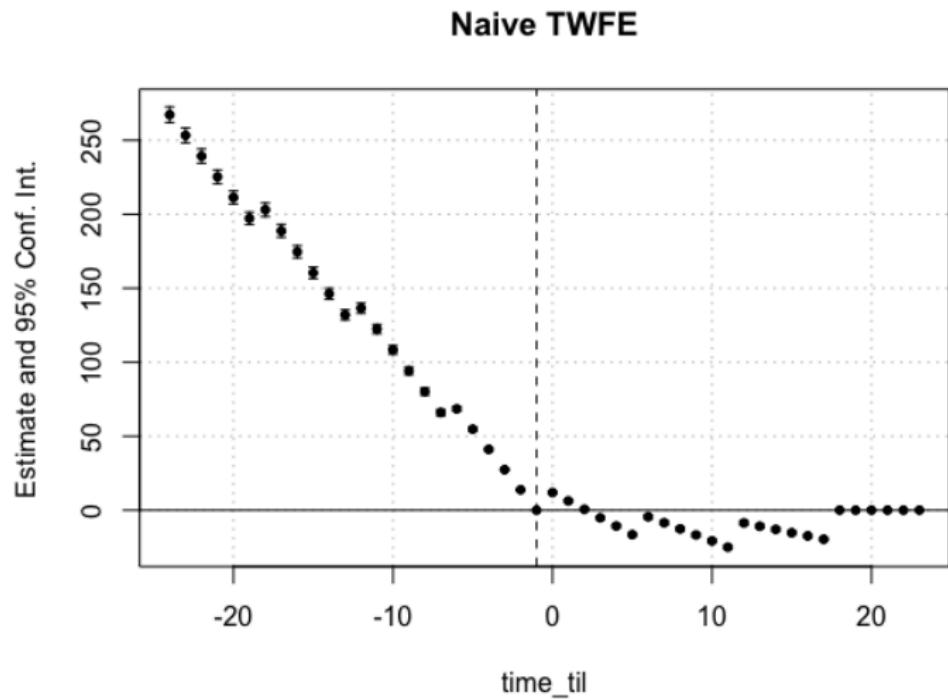
Group-time ATT

Truth					CS estimates				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)	Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0	1981	-0.0548	0.0191	0.0578	0
1986	10	0	0	0	1986	10.0258	-0.0128	-0.0382	0
1987	20	0	0	0	1987	20.0439	0.0349	-0.0105	0
1988	30	0	0	0	1988	30.0028	-0.0516	-0.0055	0
1989	40	0	0	0	1989	40.0201	0.0257	0.0313	0
1990	50	0	0	0	1990	50.0249	0.0285	-0.0284	0
1991	60	0	0	0	1991	60.0172	-0.0395	0.0335	0
1992	70	8	0	0	1992	69.9961	8.013	0	0
1993	80	16	0	0	1993	80.0155	16.0117	0.0105	0
1994	90	24	0	0	1994	89.9912	24.0149	0.0185	0
1995	100	32	0	0	1995	99.9757	32.0219	-0.0505	0
1996	110	40	0	0	1996	110.0465	40.0186	0.0344	0
1997	120	48	0	0	1997	120.0222	48.0338	-0.0101	0
1998	130	56	6	0	1998	129.9164	56.0051	6.027	0
1999	140	64	12	0	1999	139.9235	63.9884	11.969	0
2000	150	72	18	0	2000	150.0087	71.9924	18.0152	0
2001	160	80	24	0	2001	159.9702	80.0152	23.9656	0
2002	170	88	30	0	2002	169.9857	88.0745	29.9757	0
2003	180	96	36	0	2003	179.981	96.0161	36.013	0
2004	190	104	42	4	2004				
2005	200	112	48	8	2005				
2006	210	120	54	12	2006				
2007	220	128	60	16	2007				
2008	230	136	66	20	2008				
2009	240	144	72	24	2009				
ATT	82				Total ATT	n/a			
Feasible ATT	68.3333333				Feasible ATT	68.33718056			

Biased event studies

- Sun and Abraham decompose the leads and lags from standard TWFE event study models
- Each lead and lag is equal to the sum of three terms:
 - Target parameter with positive weights summing to one if binning or one if not binning
 - All other treatment effect terms associated with every other lead and lag with weights summing to zero (i.e. negative weights)
 - Parameters associated with dropped coefficients
- Standard TWFE event studies require all treatment groups have the same dynamics ("homogenous treatment profiles") which our data didn't have

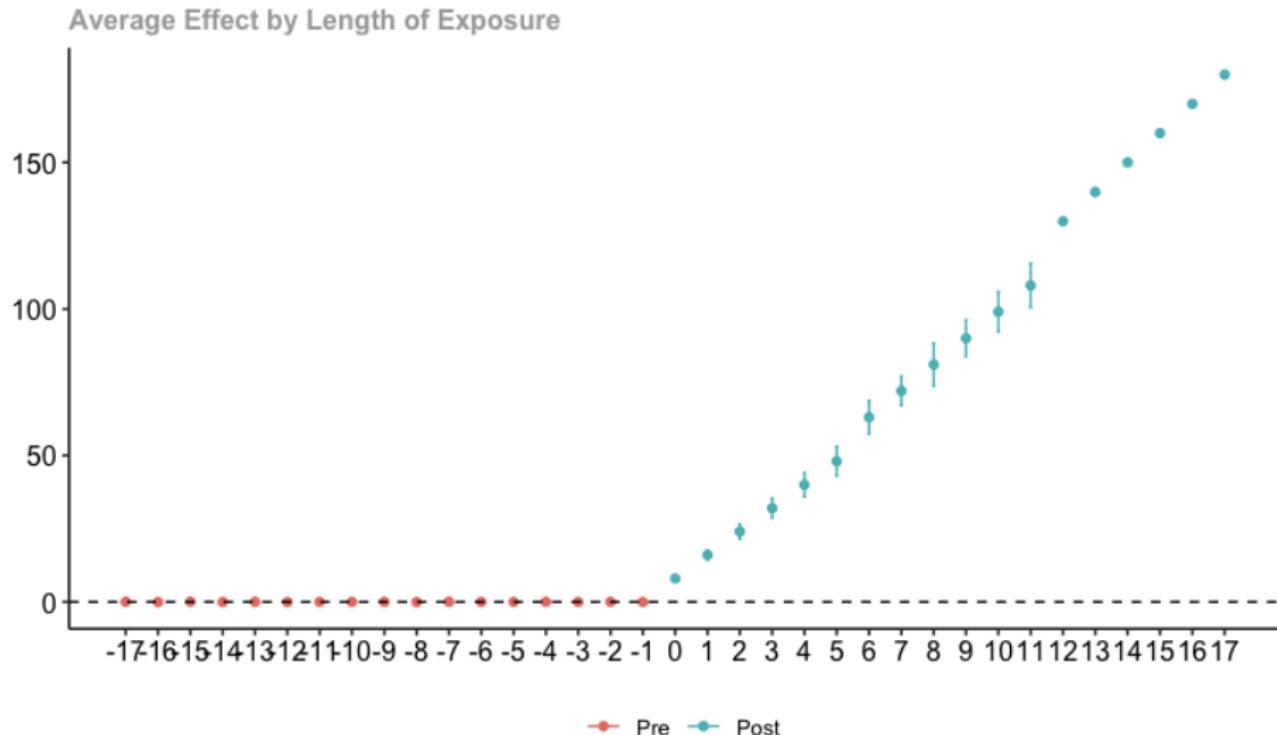
Misspecified TWFE Event Studies



Alternative Event Study Models

- Use CS or equivalent and estimate building block $ATT(g,t)$ parameters
- Aggregate into an event study by averaging over relative time periods
- CS uses bootstrapping for confidence intervals; Sun and Abraham derive the variance

Event studies



Roadmap

In Pursuit of the ATT

Potential Outcomes

Independence and Selection Bias

Unconfoundedness and Ignorable Treatment Assignment

Exact and Inexact Matching

Saturated Regressions

Synthetic control

Interpolation with non-negative weighting

Extrapolation with Conservative Negative Weighting

Difference-in-differences

Four averages and three subtractions

Covariates

Model Misspecification

Alternatives to TWFE

Conclusion

Concluding remarks

- Important elements of the diff-in-diff design
 - DiD equation is four averages and three differences
 - ATT equation is difference in two averages, one of which is counterfactual
 - DiD equals ATT if parallel trends hold and the comparison group is untreated
- Including *time-varying* covariates in the canonical OLS specification requires additional assumptions
- Staggered adoption requires either a different TWFE specification (full saturation, Wooldridge 2023) or using a different robust estimator like CS

Weights and Imputation

- Lots of mention of negative weights
 - OLS in synth negatively weighted *donor pool units*
 - TWFE in staggered adoption diff-in-diff negatively weighted *treatment effects*
- Causal inference itself is a series of weighted averages or imputations which are justified under strict assumptions
- Remember: parameters first
- Thank you so much!