Causal Inference II

Mixtape Session

Roadmap

Imputation DiD

- Robust efficient imputation (BJS)

Negative weighting by dCdH

Imputation Synthetic Control

- From DiD to Synth
- Abadie's non-negative weighting method
- Ben-Michael, et al's Least Negative Weighting Method
- Athey, et al's Matrix Completion with Nuclear Norm Method
- Synthetic difference-in-differences

Background

- The origin of the robust diff-in-diff papers identifying pathologies in TWFE was Borusyak and Jaravel (2016) working paper
- Both problems with static and dynamic specifications were discussed, and the identification of the “already treated” as controls
- Paper remained in working paper until 2021 when Jan Speiss was brought on; the team developed a new estimator
- We will focus primarily on the estimator, to avoid redundancies

My Outline

1. Discussion of their interpretation of “basic” DiD assumptions
2. Critique of TWFE OLS when strong assumptions don’t hold
3. Introduction of new assumptions
4. Robust efficient imputation estimator

ATT parameter

Their causal parameter target is the unit level treatment effects, but they will once they grab them aggregate to the ATT:

$$\text{ATT} = \sum_{it \in \Omega_1} w_{it} \tau_{it} = w_1' \tau$$

Note the weights.

Standard TWFE Assumptions

Their paper starts with a critique of the TWFE model by listing out the assumptions needed for unbiased estimate of the ATT:

1. Parallel trends
2. No anticipation
3. Homogenous treatment effects

Let's just discuss them each so we can the subtle differences than the others we reviewed

A1: Parallel trends

Assumption 1: Parallel trends. Expressed as an additive model of unit and time fixed effects (same exact representation as the baker simulation)
Only imposes restrictions on $Y(0)$, not treatment effects themselves.
Notice how their parallel trends uses *unit* level parallel trends (like I did with Baker.do) which is stronger than the average parallel trends of CS and SA.

A2: No anticipation

- No anticipation rules out anticipatory behavior that would cause treatment effects to materialize even before the treatment occurred:

$$Y_{it} = Y_{it}(0)$$

  for all $it \in \Omega_0$.
- Notice how as an assumption, it literally imposes $\tau = 0$ for all pre-treatment periods.

A2: No anticipation

Assumption 2: No anticipation: remember, does not mean a person is surprised by the policy. Just means that a future treatment has no treatment effect in the past.

A3: Restricted causal effects

Assumption 3 (Restricted causal effects): $B\tau 0$ for a known $M \times N_1$ matrix B of full row rank.
This is the one that places restrictions on what treatment effects can and cannot be (i.e., homogenous treatment effects). Notice the very detailed expression

If we can assume something like homogenous treatment effects, then TWFE actually is best because its ability to *correctly* extrapolate will increase efficiency. But it's when A3 is not tenable or not really ex ante justified by theory that we should be worried. There's an A3' that is a slight modification.

Critique of Common Practice

1. Under-identification in event studies
2. Negative weighting
3. Spurious identification of long-run casual effects

Underidentification problem caused by two forms of multicollinearity

Lemma 1: If there are no never-treated units, the path of [pre-treatment lead population regression coefficients] is not point identified in the fully dynamic OLS specification. In particular. adding a linear trend to this path $\{\tau_h + k(h+1)\}$ for any $k \in R$ fits the data equally well with the fixed effects coefficients appropriately modified.

This is caused by multicollinearity in calendar time when you include all year dummies, and multicollinearity in relative event time when you also include all event time dummies.

You have to drop two pre-treatment dummies with TWFE instead of one when estimating event studies (SA mention this too)

Negative weighting due to heterogenous treatment effects

Assume some simple static model with a single dummy for treatment by heterogenous treatment effects. Then they lay out a second lemma

Lemma 2: If parallel trends and no anticipation hold, then the estimand of the static OLS specification satisfies $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{OLS} \tau_{it}$ for some weights w_{it}^{OLS} that do not depend on the outcome realizations and add up to one $\sum_{it \in \Omega_1} = 1$.

The static OLS estimand cannot be interpreted as a “proper” weighted average (proper is thought to be an estimate that is a positively weighted average of all treatment effects), as some weights can be negative.

Simple illustration

Table: TWFE dynamics

$E(y_{it})$	$i = A$	$i = B$
t=1	α_A	α_B
t=2	$\alpha_A + \beta_2 + \delta_{A2}$	$\alpha_B + \beta_2$
t=3	$\alpha_A + \beta_3 + \delta_{A3}$	$\alpha_B + \beta_3 + \delta_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Static: $\delta = \delta_{A2} + \frac{1}{2}\delta_{B3} - \frac{1}{2}\delta_{A3}$.

Notice the negative weight on the furthest lag. This is what you get when A3 is not satisfied.

Short-run bias of TWFE

- TWFE OLS has a severe short-run bias
- the long-run causal effect, corresponding to the early treated unit A and the late period 3, enters with a negative weight (-1/2)
- The larger the effects in the long-run, the smaller the coefficient will be (we saw this in the Facebook graph)
- It's caused by "forbidden comparisons" (late to early treated) – we saw this with Goodman-Bacon (2021)
- Forbidden comparisons create downward bias on long-run effects with treatment effect heterogeneity, *but not with treatment effect homogeneity* – so it really is an A3 violation

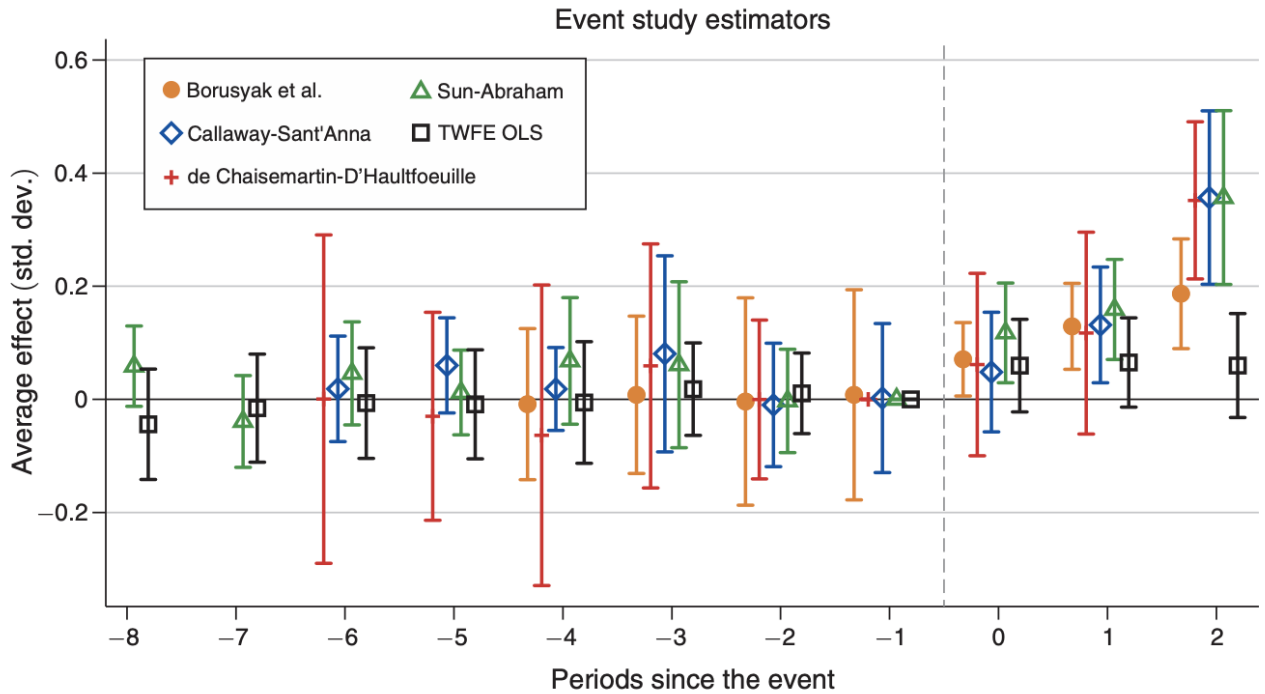


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Spurious Long-Run Causal Effects

Sometimes though you can also find long-run effects where there are none. Basically, you need to impose a lot of pre-trend restrictions to get estimates of long-run population regression coefficients. Even then you can't get them all.

OLS estimates are fully driven by unwarranted extrapolations of treatment effects across observations and may not be trusted unless strong ex ante justifications for A3 exist

Lemma 4: Suppose there are no never-treated units and let $H = max_i E_i - min_i E_i$. Then for any non-negative weights w_{it} defined over the set of observations with $K_{it} \geq \overline{H}$ (that are not identically zero), the weighted sum of causal effects $\sum_{it:K_{it} \geq \overline{H}} w_{it} \tau_{it}$ is not identified by A1 and A2.

Modifications of general model

Modification of A1 to A1':

$$Y_{it}(0) = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}$$

Assumption 4 is introduced (homoskedastic residuals). This is key, because they will be building an "efficient estimator" with BLUE like OLS properties.

Using A1' to A4, we get the "efficient estimator" which is for all linear unbiased estimates of δ_W, the unique efficient estimator $\widehat{\delta^*_W}$ can be obtained with 3 steps

Role of the untreated observations

"The idea is to estimate the model of Y_{it}^0 using the untreated observations and extrapolate it to impute Y_{it}^0 for treated observations." – Borusyak, et al. (2022)

"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others." – Imbens and Rubin (2015)

BJS Steps for Estimation of ATT

1. Estimate expected Y^0 using OLS using only the untreated observations (similar to outcome regression)
2. Then calculate $\widehat{\delta}_{it} = Y_{it}^1 - \widehat{Y}_{it}^0$
3. Then estimate target parameters as weighted sums

$$\widehat{\delta}_W = \sum_{it} w_{it} \widehat{\delta}_{it}$$

Why is this working?

- Think back to that original statement of the PT assumption – you're modeling $Y(0)_{it}$.
- That is, without treatment – so the potential outcomes do not depend on any treatment effect
- Hence where we get treatment heterogeneity
- We obtain consistent estimates of the fixed effects which are then used to extrapolate to the counterfactual units for all $Y(0)_{it \in \Omega_1}$

Pros and Cons

- Strengths: computationally fast and flexible to unit-trends, triple diff, covariates and so forth (though remember what we said about covariates)
- Since we use more pre-treatment data for this imputation, we get increased power
- But stronger parallel trends and assumptions of homoskedasticity

Comparisons to other estimators

Table 3: Efficiency and Bias of Alternative Estimators

Horizon	Estimator	Baseline simulation		More pre-periods	Heterosk. residuals	AR(1) residuals	Anticipation effects
		Variance (1)	Coverage (2)	Variance (3)	Variance (4)	Variance (5)	Bias (6)
$h = 0$	Imputation	0.0099	0.942	0.0080	0.0347	0.0072	-0.0569
	DCDH	0.0140	0.938	0.0140	0.0526	0.0070	-0.0915
	SA	0.0115	0.938	0.0115	0.0404	0.0066	-0.0753
$h = 1$	Imputation	0.0145	0.936	0.0111	0.0532	0.0143	-0.0719
	DCDH	0.0185	0.948	0.0185	0.0703	0.0151	-0.0972
	SA	0.0177	0.948	0.0177	0.0643	0.0165	-0.0812
$h = 2$	Imputation	0.0222	0.956	0.0161	0.0813	0.0240	-0.0886
	DCDH	0.0262	0.958	0.0262	0.0952	0.0257	-0.1020
	SA	0.0317	0.950	0.0317	0.1108	0.0341	-0.0850
$h = 3$	Imputation	0.0366	0.928	0.0255	0.1379	0.0394	-0.1101
	DCDH	0.0422	0.930	0.0422	0.1488	0.0446	-0.1087
	SA	0.0479	0.952	0.0479	0.1659	0.0543	-0.0932
$h = 4$	Imputation	0.0800	0.942	0.0546	0.3197	0.0773	-0.1487
	DCDH	0.0932	0.950	0.0932	0.3263	0.0903	-0.1265
	SA	0.0932	0.954	0.0932	0.3263	0.0903	-0.1265

Notes: See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

Comparisons to other estimators

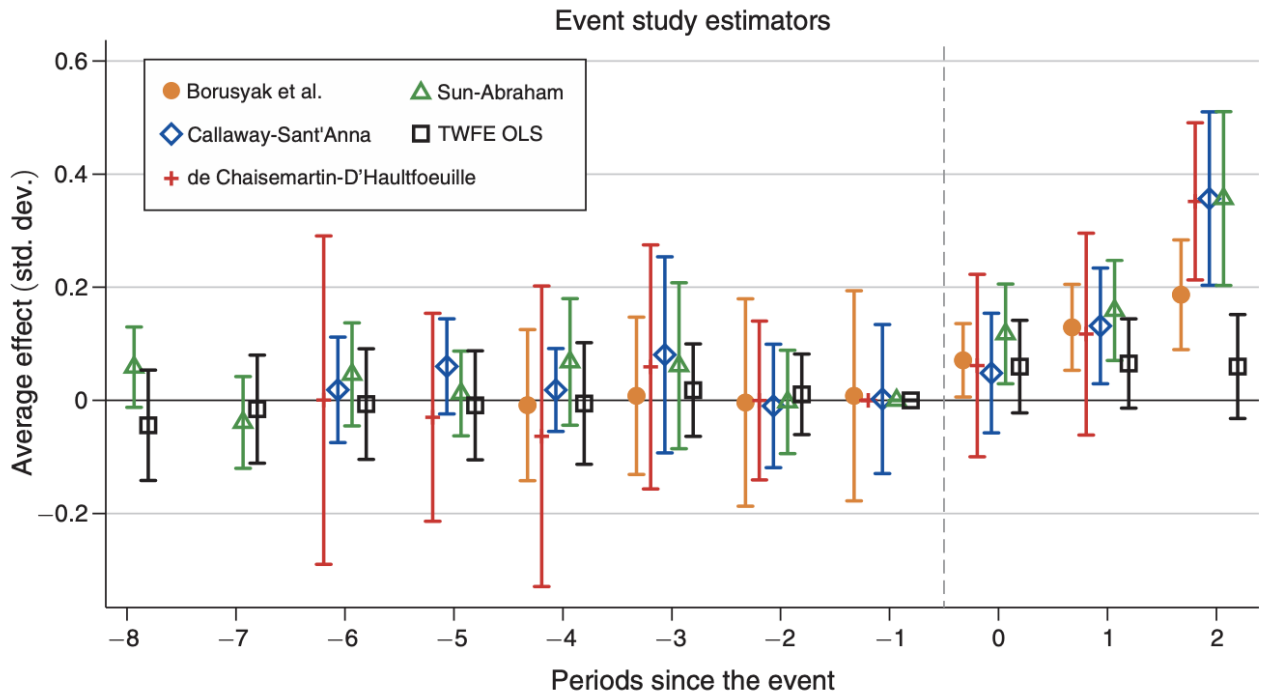


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Roadmap

Imputation DiD

- Robust efficient imputation (BJS)

Negative weighting by dCdH

Imputation Synthetic Control

- From DiD to Synth
- Abadie's non-negative weighting method
- Ben-Michael, et al's Least Negative Weighting Method
- Athey, et al's Matrix Completion with Nuclear Norm Method
- Synthetic difference-in-differences

de Chaisemartin and D'Haultfoeulle 2020

de Chaisemartin and D'Haultfouelle 2020 (dCdH) is different from the other papers in several ways

- Like SA, it's a diagnosis and a cure
- TWFE decomposition shows coefficient a weighted average of underlying treatment effects, but weights can be negative negating causal interpretation
- Propose a solution for both static and dynamic specification which does not use already treated as controls
- Treatment can turn on and off

Comment on Bacon

- Recall the Bacon decomposition – TWFE coefficients are decomposed into weighted average of all underlying 2x2s. Weights were non-negative and summed to one.
- But this decomposition was more a numerical decomposition – what exactly adds up to equal the TWFE coefficient using the data we observe?
- Bacon's decomposition is not "theoretical" – not in the way that other decompositions are. He is just explaining what OLS "does" when it calculates $\widehat{\delta}$ (positively weighted average of all 2x2 DiD equations)
- dCdH decompose TWFE coefficient into weighted average of treatment effects, not DiD

Negative weights

- dCdH impose causal assumptions and try a different decomposition strategy
- Uses as its building block the unit-specific treatment effects (not the 2x2 DiD equation)
- Their decomposition reveals that negative weights on the underlying treatment effects form with the TWFE specification we showed yesterday (similar to negative weight on dynamics with Bacon)

Negative weights (dcdh) vs positive weights (Bacon)

Bacon decomposition weights were *always* positive, because they were numerical weights on the DiD equation

dCdH are sometimes negative weights but not on the 2x2 DiD, but rather on the treatment effects themselves

dCdH notation

- Individual treatment effects (iow, not the group-time ATT):

$$\Delta_{i,t}^{g} = Y_{i,t}^{1} - Y_{i,t}^{\infty}$$

  but where the treatment is in time period g. Notice –it's not the ATT (it's i individual treatment effect)
- with defined error term as $\varepsilon_{i,t}$:

$$D_{i,t} = \alpha_i + \alpha_t + \varepsilon_{i,t}$$

- Weights:

$$w_{i,t} = \frac{\varepsilon_{i,t}}{\frac{1}{N^T}\sum_{i,t:D_{i,t}=1}\varepsilon_{i,t}}$$

Parallel trend assumption

Strong unconditional PT

Assume that for every time period t and every group g, g',

$$E[Y_t^\infty - Y_{t-1}^\infty | G = g] = E[Y_t^\infty - Y_{t-1}^\infty | G = g']$$

Assume parallel trends for every unit in every cohort in every time period.

What then does TWFE estimate with differential timing?

dCdH Theorem

Theorem – dCdH decomposition

Assuming SUTVA, no anticipation and the strong PT, then let δ be the TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \delta D_{i,t} + \varepsilon_{i,t}$$

Then it follows that

$$\delta = E\left[\sum_{i,t:D_{i,t}=1} \frac{1}{N^T} w_{i,t} \cdot \Delta^g_{i,t} \right]$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N^T} = 1$ but $w_{i,t}$ can be negative

Origins

- So once you run that specification, $\widehat{\delta}$ is going to recover a "non-convex average" over all unit level treatment effects (weights can be negative, more on this).
- Not sure who came first, because there were working papers before publications, but my understanding is dCdH was the first to prove this
- Very important theorem – established the "no sign flip property" for OLS with differential timing in the canonical static specification

Negative weights

- Very common now to hear about negative weights, and furthermore, that negative weights wipe out any causal interpretation, but why?
- Thought experiment: imagine every unit gained from the treatment, but their treatment effect when estimated was multiplied by a negative number
- It's possible it could flip the sign, but it would definitely at least pull the estimate away from the true effect
- This is dangerous – and it's caused by the forbidden contrasts (comparing treated to already treated) which is what the canonical TWFE static specification is doing (for many of us unknowingly)

Negative weights

- Doesn’t always pose a problem, but no proofs for this intuition known yet
- A large number of never-treated seems to make this less an issue
- Shrinking the spacing between treatment dates also can drive it down
- But does that mean that TWFE works, and what does it mean to work?
- TWFE still even when all the weights are positive the weighted average may not aggregate to what we think it does

Weighting

- The weights in OLS all come out of the model itself, *not the economic question*
- The economic question is “what parameter do you want? What does it look like? Who is in it?”
- And when you define the parameter up front, you’ve more or less defined the economic question you’re asking
- But OLS sort of ignores your question and just gives you what it wants

Weighting

- What makes something a good vs a bad weight?
- Not being negative is the absolute minimal requirement
- But it’s also not a good sign if you can’t really explain the weights

dCdH Solution

- dCdH propose an alternative that doesn't have the problems of TWFE – both avoiding negative weights and improving interpretability
- Recall, their model can handle reversible treatments

Roadmap

Imputation DiD

- Robust efficient imputation (BJS)

Negative weighting by dCdH

Imputation Synthetic Control

- From DiD to Synth
- Abadie's non-negative weighting method
- Ben-Michael, et al's Least Negative Weighting Method
- Athey, et al's Matrix Completion with Nuclear Norm Method
- Synthetic difference-in-differences

Moving on from DiD

- Difference-in-differences is very popular; it is often a go-to procedure for studying large policies and demand for staggered adoption will remain
- Canonical panel modeling of TWFE can with dynamic treatment effects flip the sign; will it is another matter, but it can – thankfully we have alternatives
- Treatment effects can be positive, and yet TWFE is biased down, and may even report negative effects
- Caused by non-transparent use of already treated units as controls

Summarizing

- Even if in your situation that didn't happen, how would you know without making strong prior statements about heterogeneity
- This requires *a priori* knowledge of opaque production functions that are simply not available to most researchers
- Generalizability is also an issue because the weights that TWFE uses are not the weights most likely what the policymaker needs ("variance weighted ATT")

Summarizing

- My recommendation is to use more transparent and robust methods when facing staggered adoption, and pay careful attention to how each handles covariates
- Know who your comparison group is, and how each underlying building block aggregates (i.e., weights chosen)
- Consider tables to help readers also understand, at least in the appendix

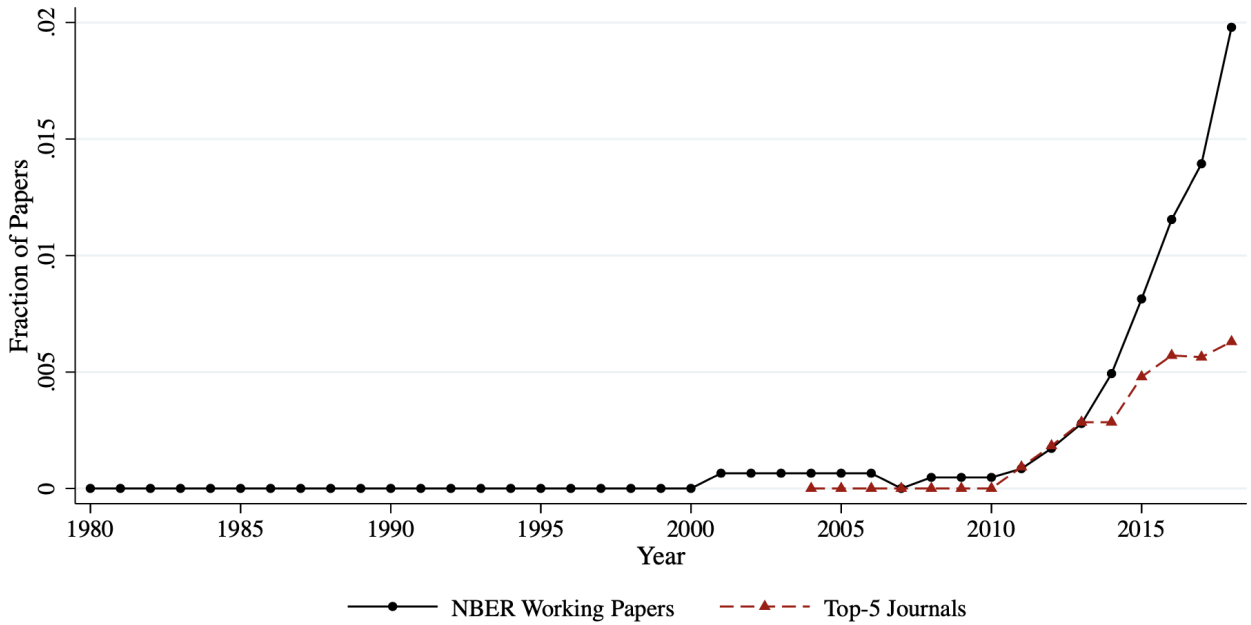
Summarizing

- Core assumptions of difference-in-differences is, apart from no anticipation and SUTVA, the parallel trends assumption
- It is not testable; people often check for it through falsifications, including the pre-treatment event study which under no anticipation leaves only *that period's* differential trend
- Even under robust diff-in-diff methods, pre-treatment trends can fail, which would simply reflect that the treatment and control group are diverging even before the intervention

Summarizing

- But what if parallel trends simply does not become believable?
- This may happen because of the event study – the *aggregate* comparison group is simply not a proxy for $E[\Delta Y^0|D=1]$, the counterfactual
- Then it may be time to shift gears and move into the synthetic control methods as they do not rely on parallel trends

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 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


EL DORADO


EL DORADO


SILVER SEA

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

Synthetic control method: disadvantages

1. Subjective researcher bias kicked down to the model selection stage
2. Significant diversity at the moment as to how to principally select models - from machine learning to modifications - as well as estimation and software
3. Part of the purpose of this procedure is to reduce subjective researcher bias
4. Ferman, Pinto and Possbaum (2020) suggest specific specifications and report all of them

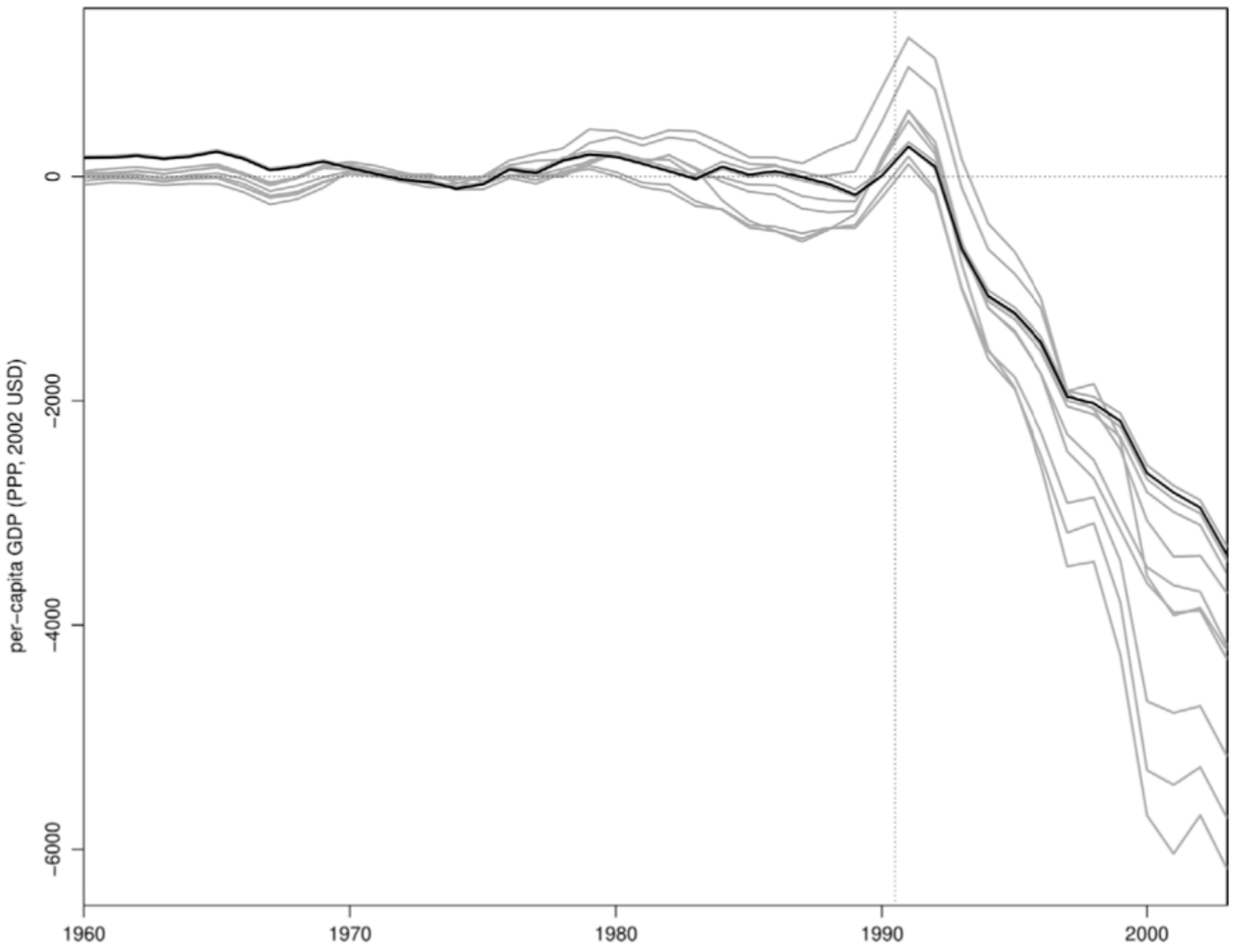
Avoiding cherry picking

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.

Avoiding cherry picking



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.

Notation and setup

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

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

Potential outcomes notation

- Let Y_{it}^0 be the outcome that would be observed for unit i at time t in the absence of the intervention
- Let Y_{it}^1 be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T.

Group-time ATT with only one treated group

Treatment effect parameter is defined as dynamic ATT where

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

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

Optimal weights

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{j+1} = 1$. 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 $W^* = (w_2^*, \ldots, w_{J+1}^*)'$ is chosen to minimize $||X_1 - X_0 W||$, subject to our weight constraints

Optimal weights differ by another weighting matrix

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 $v_1, \ldots, v_k$. Then, the synthetic control weights $w_2^*, \ldots, w_{J+1}^*$ 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

Choice of V is critical

- 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 $v_1, \ldots, v_k$ 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 $v_1, \ldots, v_k$ affect the discrepancies between the treated unit and the synthetic control
- Minimize mean square prediction error (MSPE) for the pre-treatment period (default):

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

Cross 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

Suppose Y^0 is given by a 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

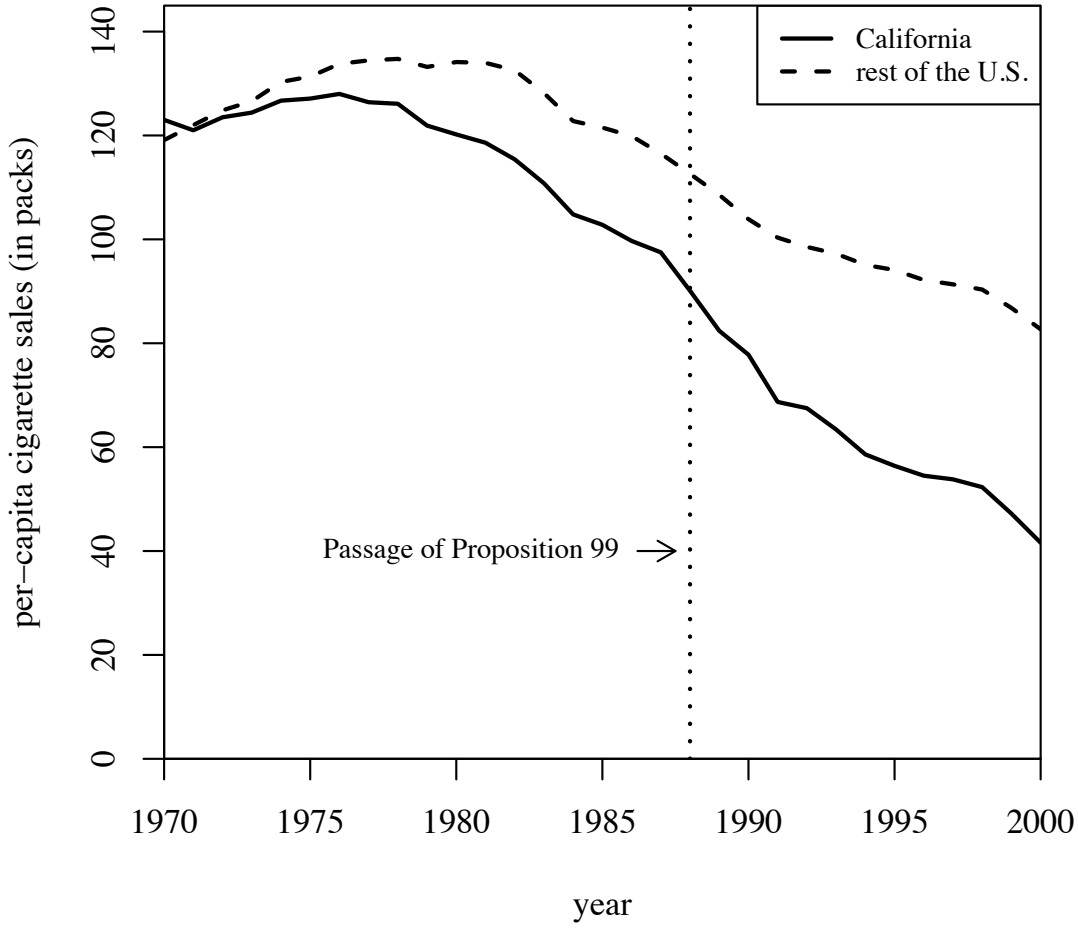
$$Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} = \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t})$$

- 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
- Proof in Appendix B of ADH (2011)

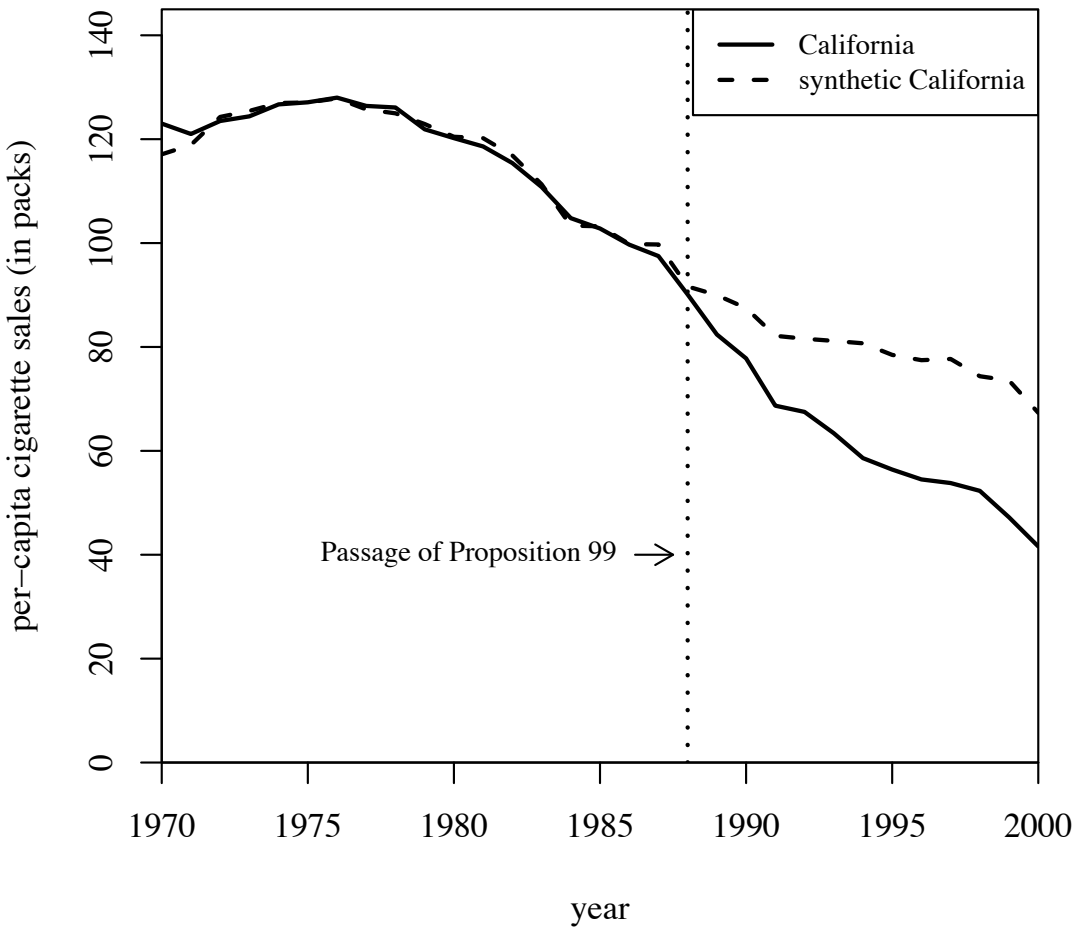
Example: California's Proposition 99

- 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

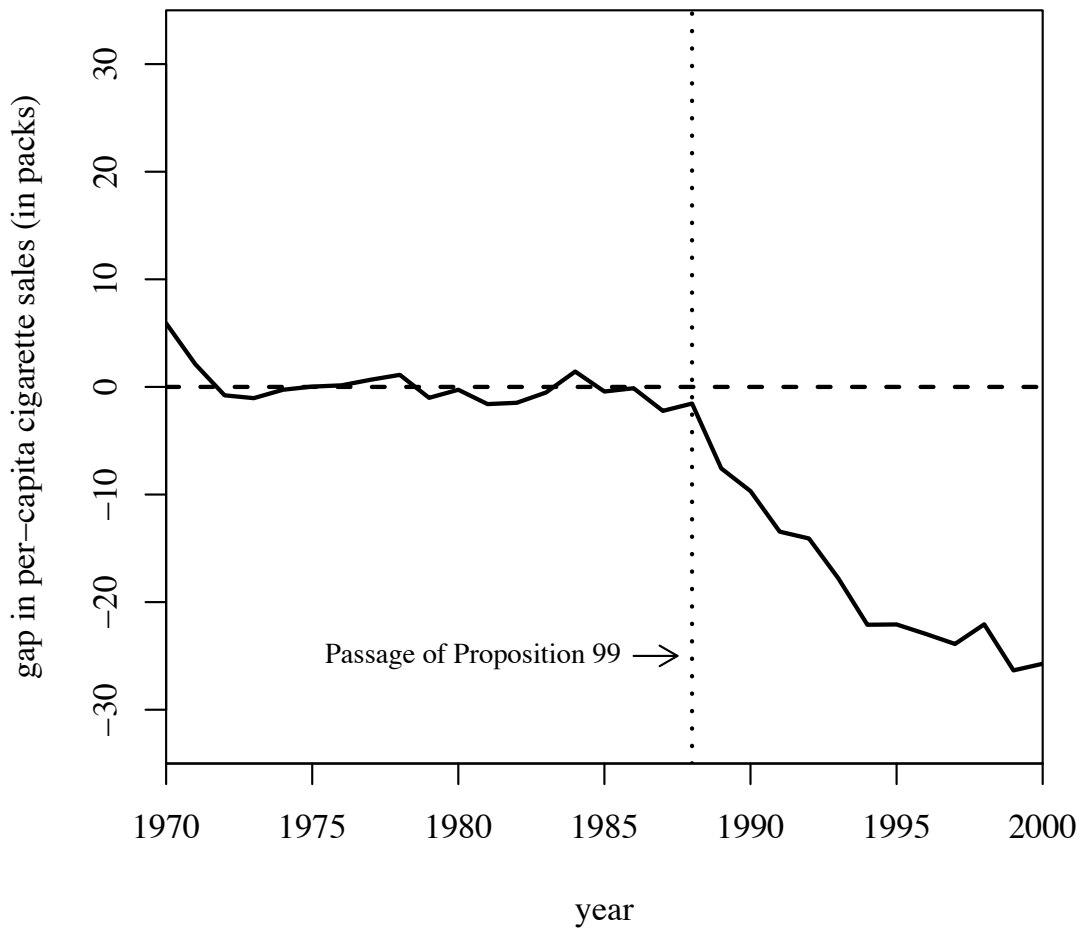


Predictor Means: Actual vs. Synthetic California

	California		Average of
Variables	Real	Synthetic	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



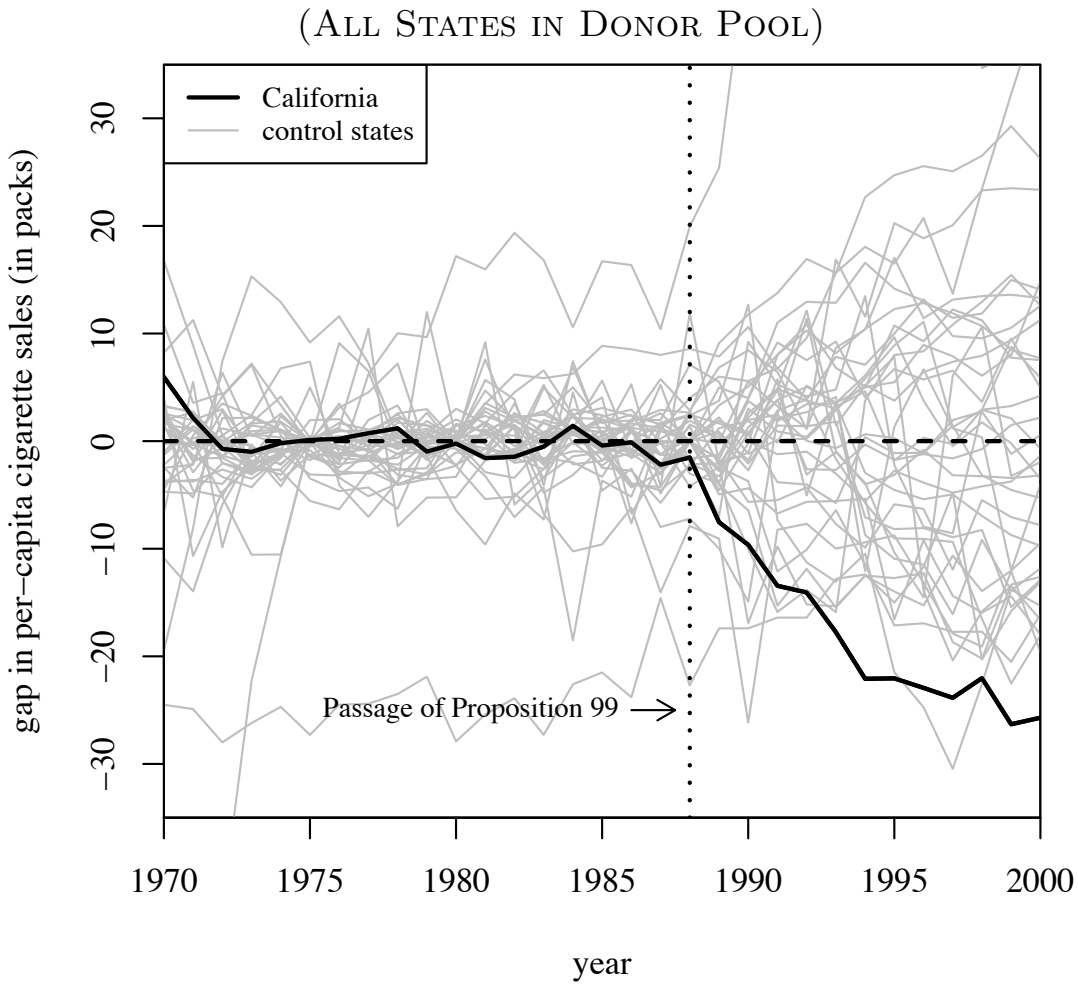
Inference

- 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:

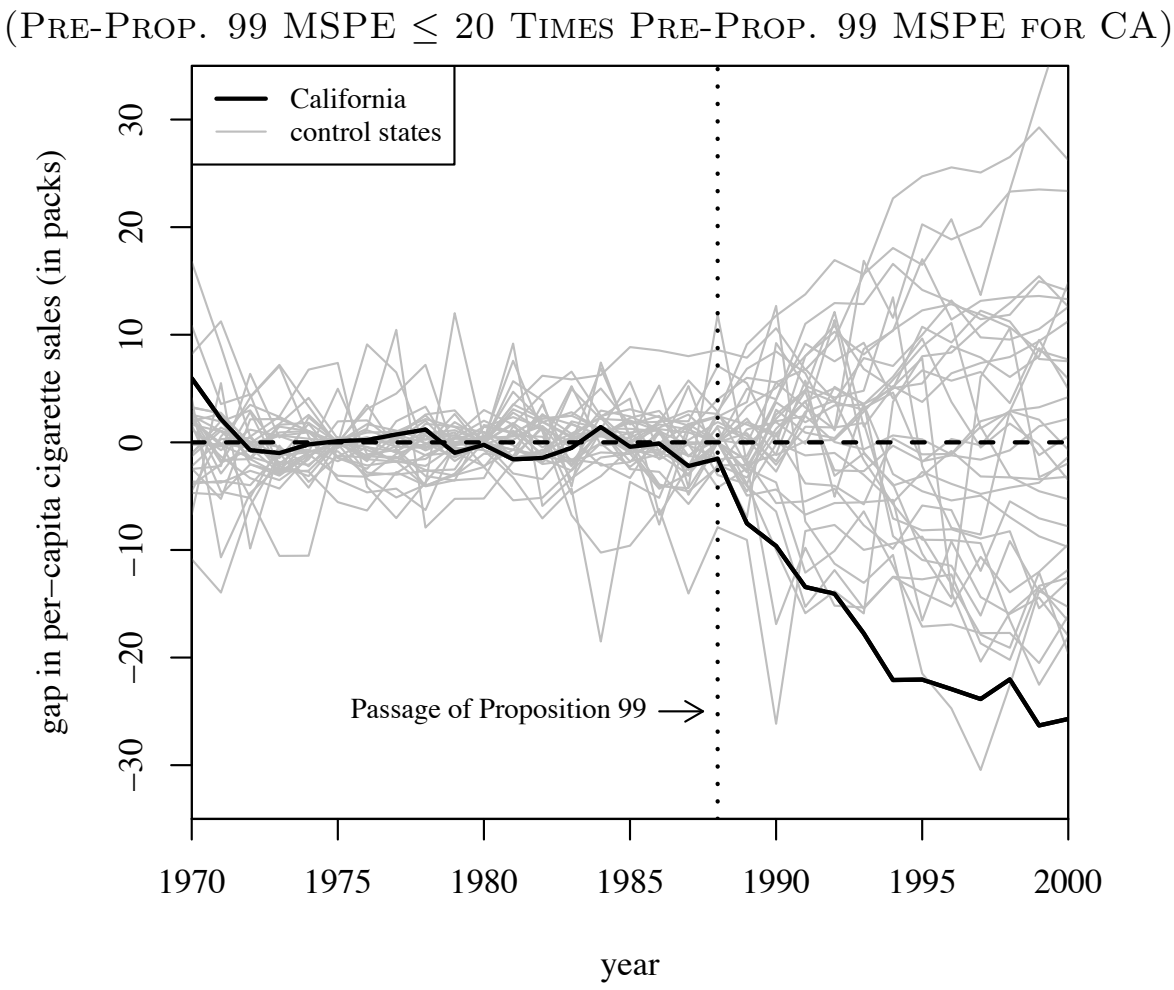
$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

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

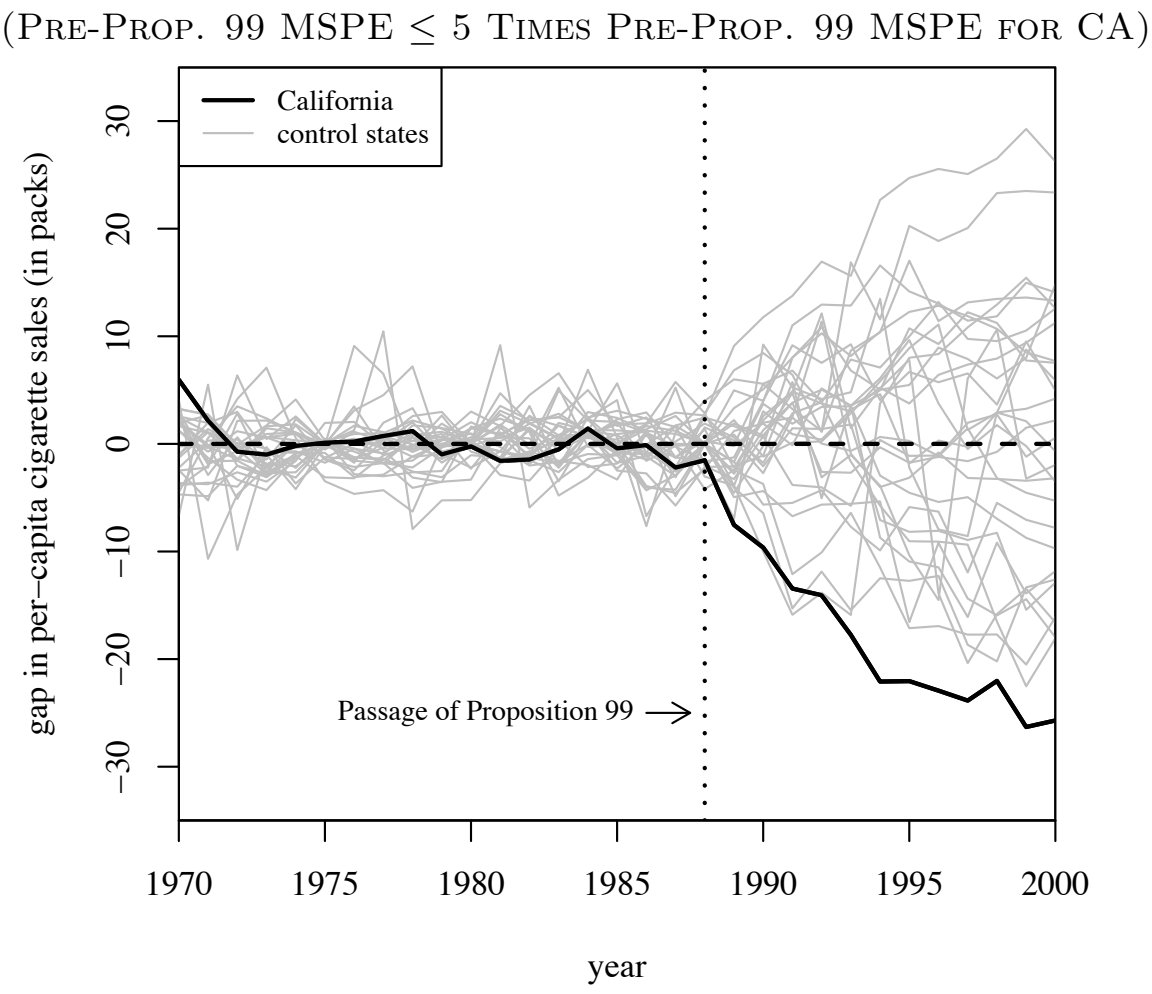
Smoking Gap for CA and 38 control states



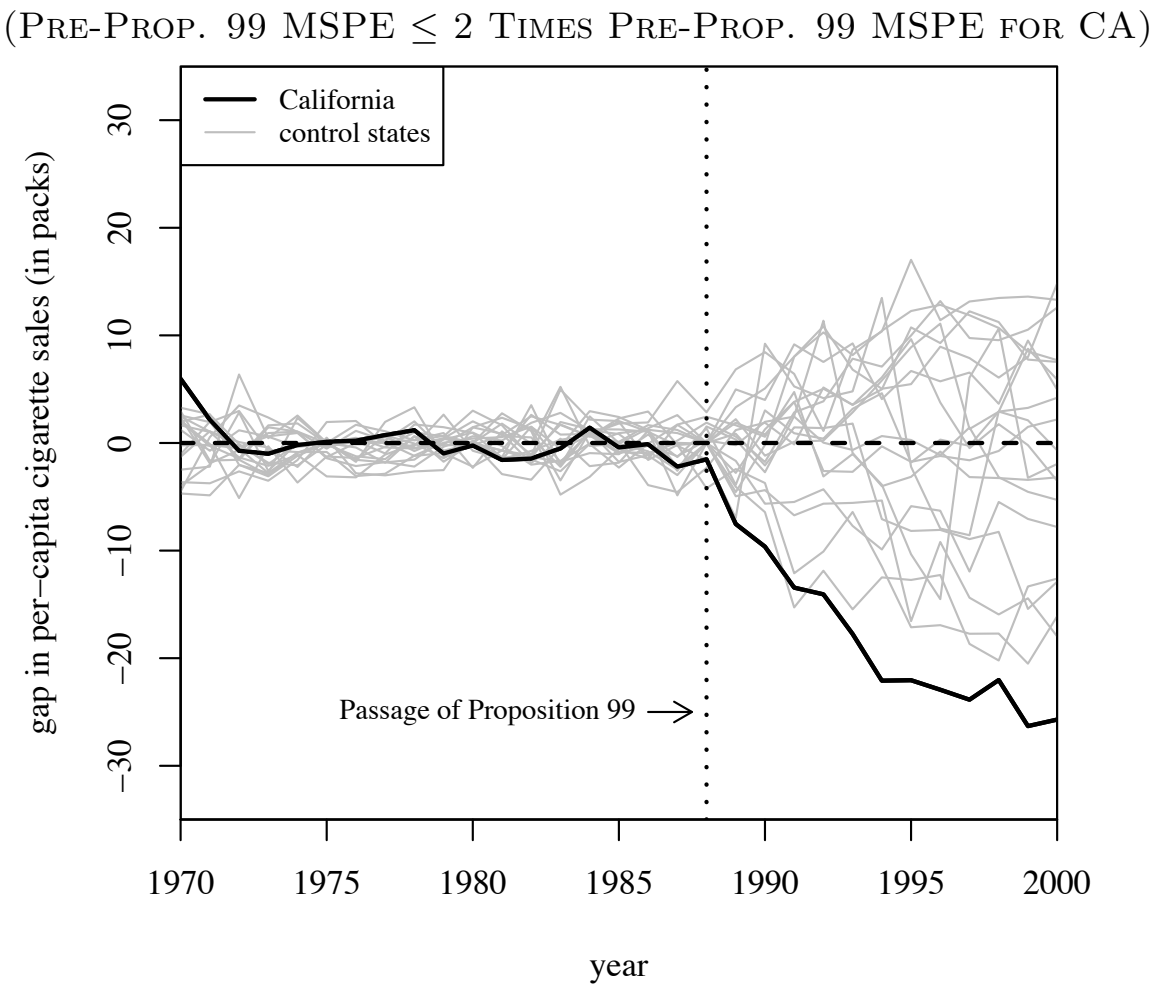
Smoking Gap for CA and 34 control states



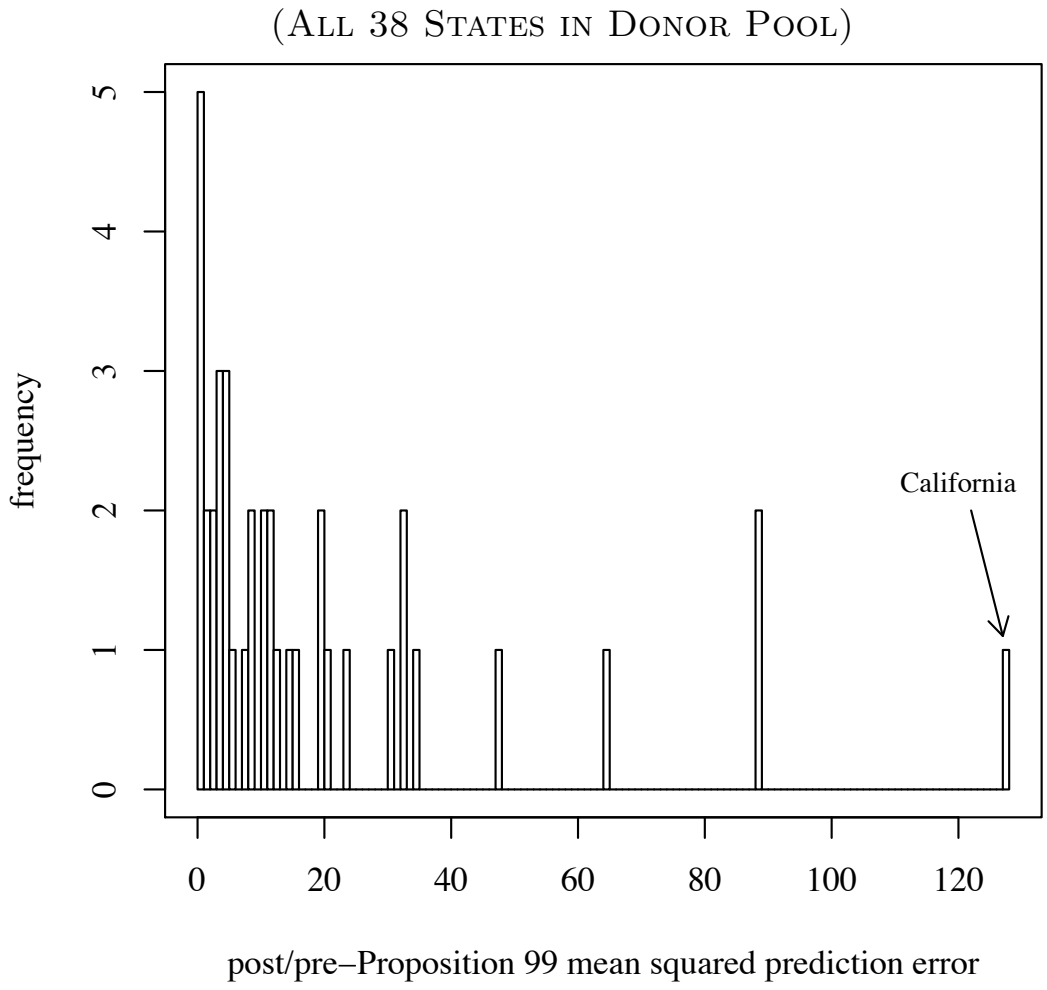
Smoking Gap for CA and 29 control states



Smoking Gap for CA and 19 control states



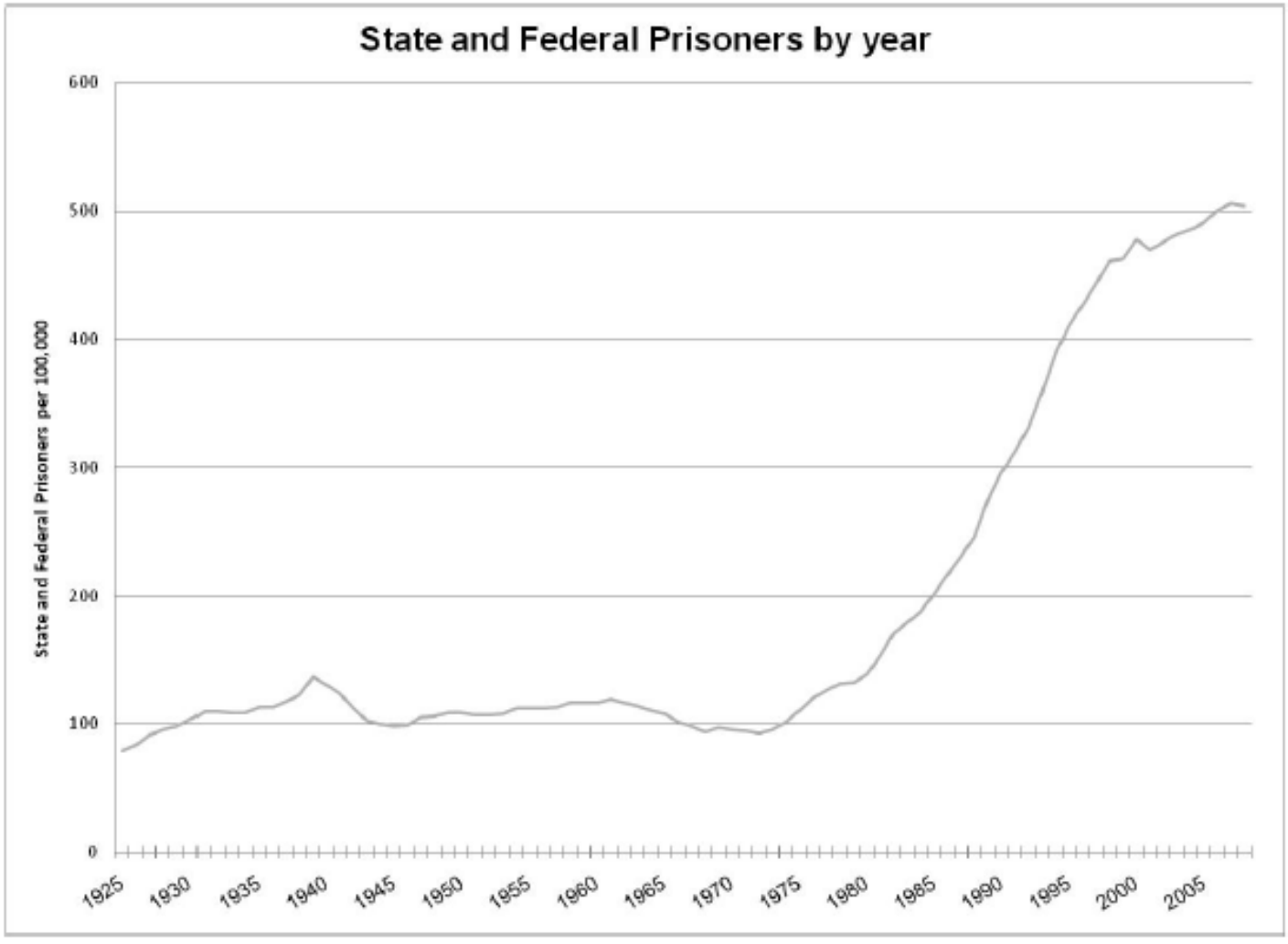
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



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



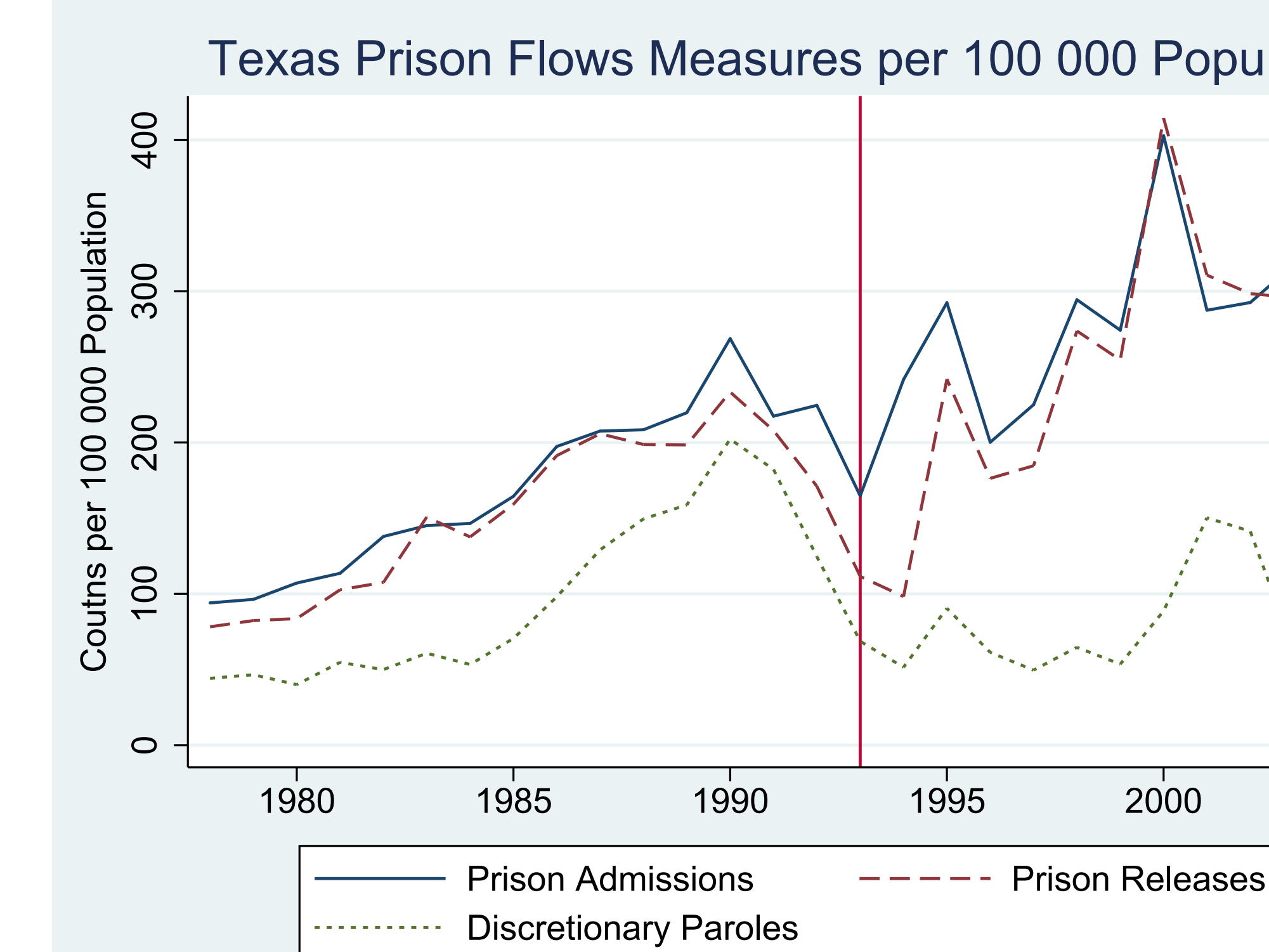
Source: www.albany.edu/sourcebook/tost_6/html

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

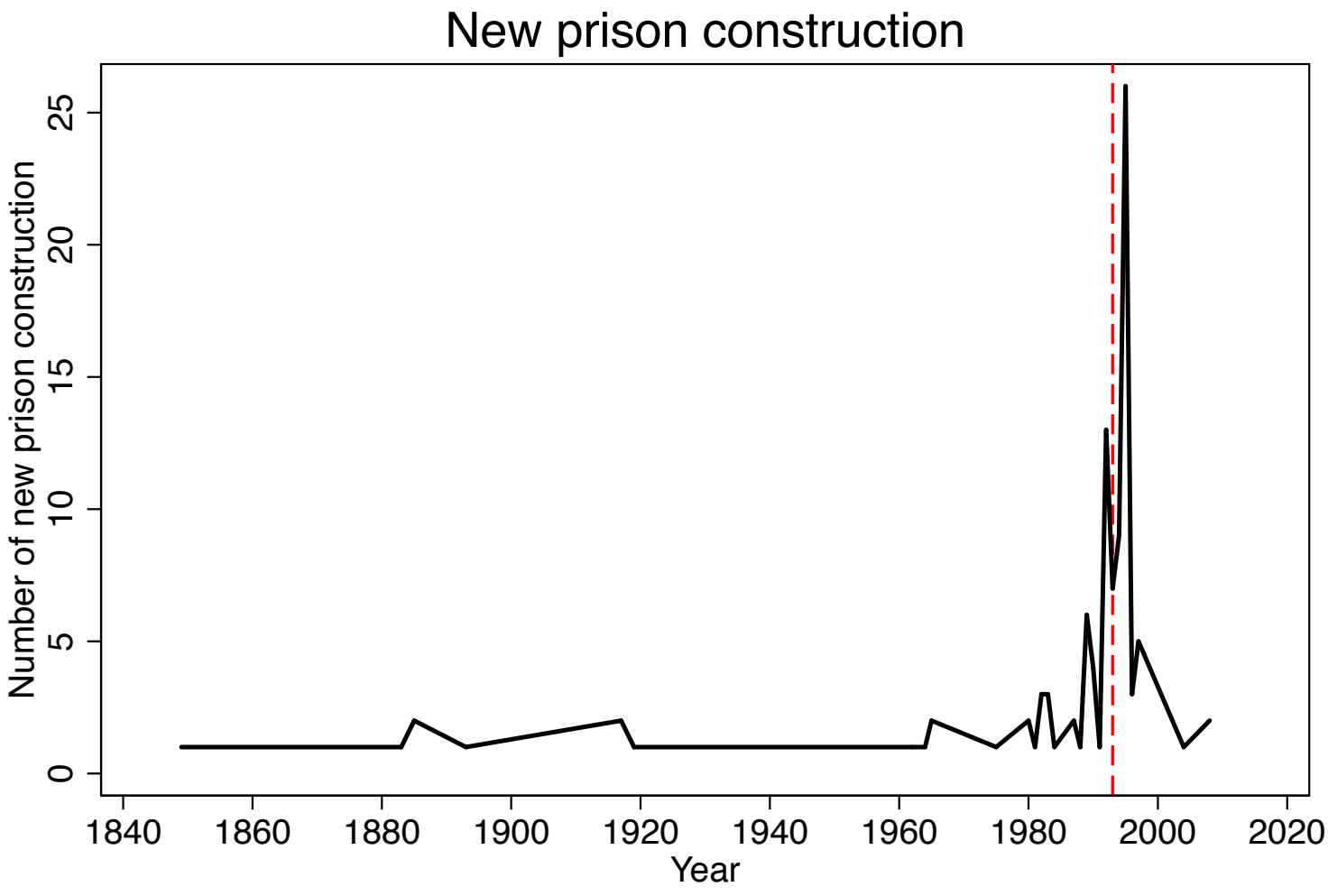
- Class action lawsuit against TX Dept of Corrections (Estelle, warden).
- TDC lost. Lengthy period of appeals and legal decrees.
- Lengthy period of time relying on paroles to manage flows


Texas Prison Flows Measures per 100 000 Popu
Coutns per 100 000 Population
0
100
200
300
400
1980
1985
1990
1995
2000
Prison Admissions
Prison Releases
Discretionary Paroles

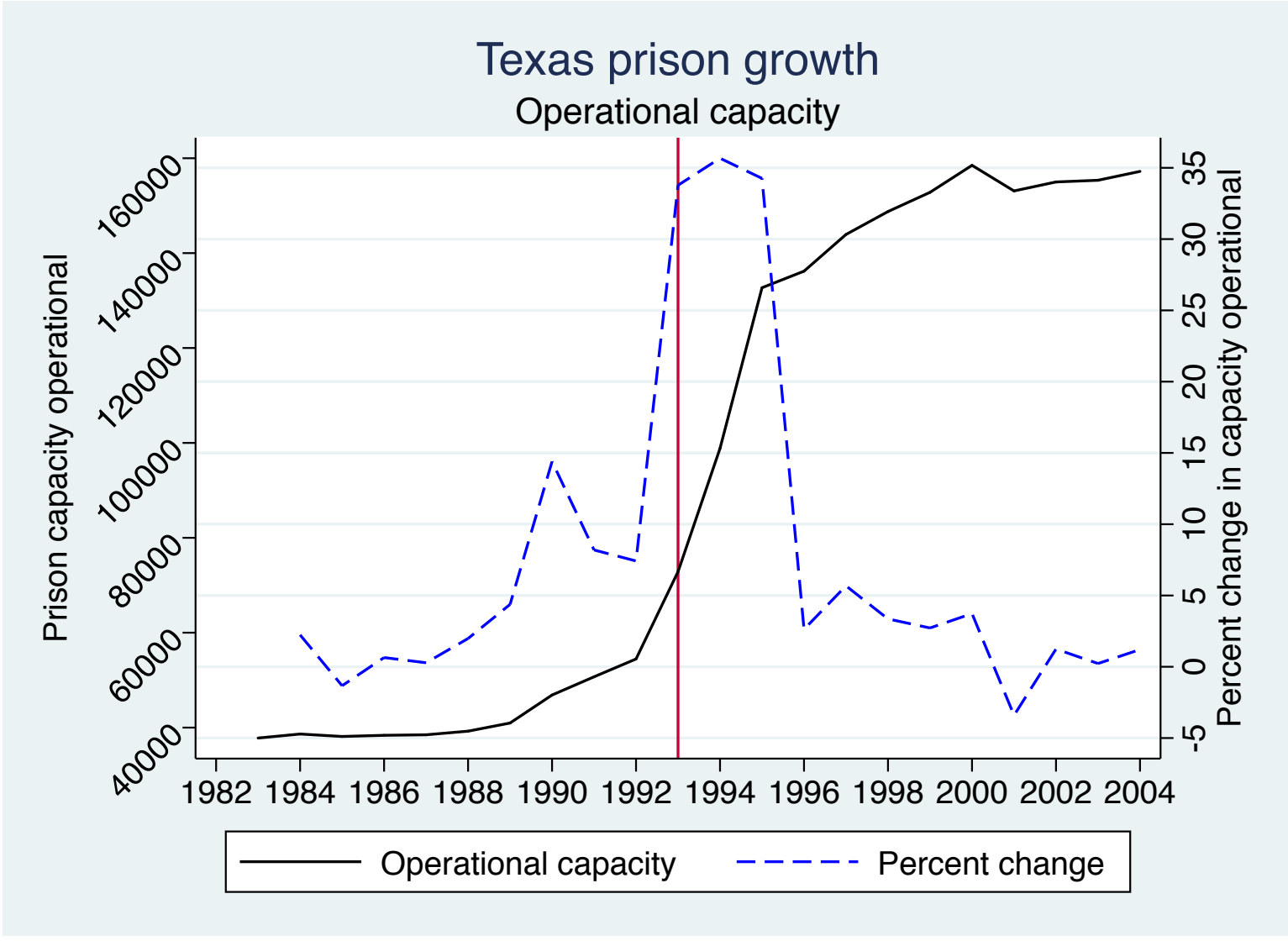
Texas prison boom

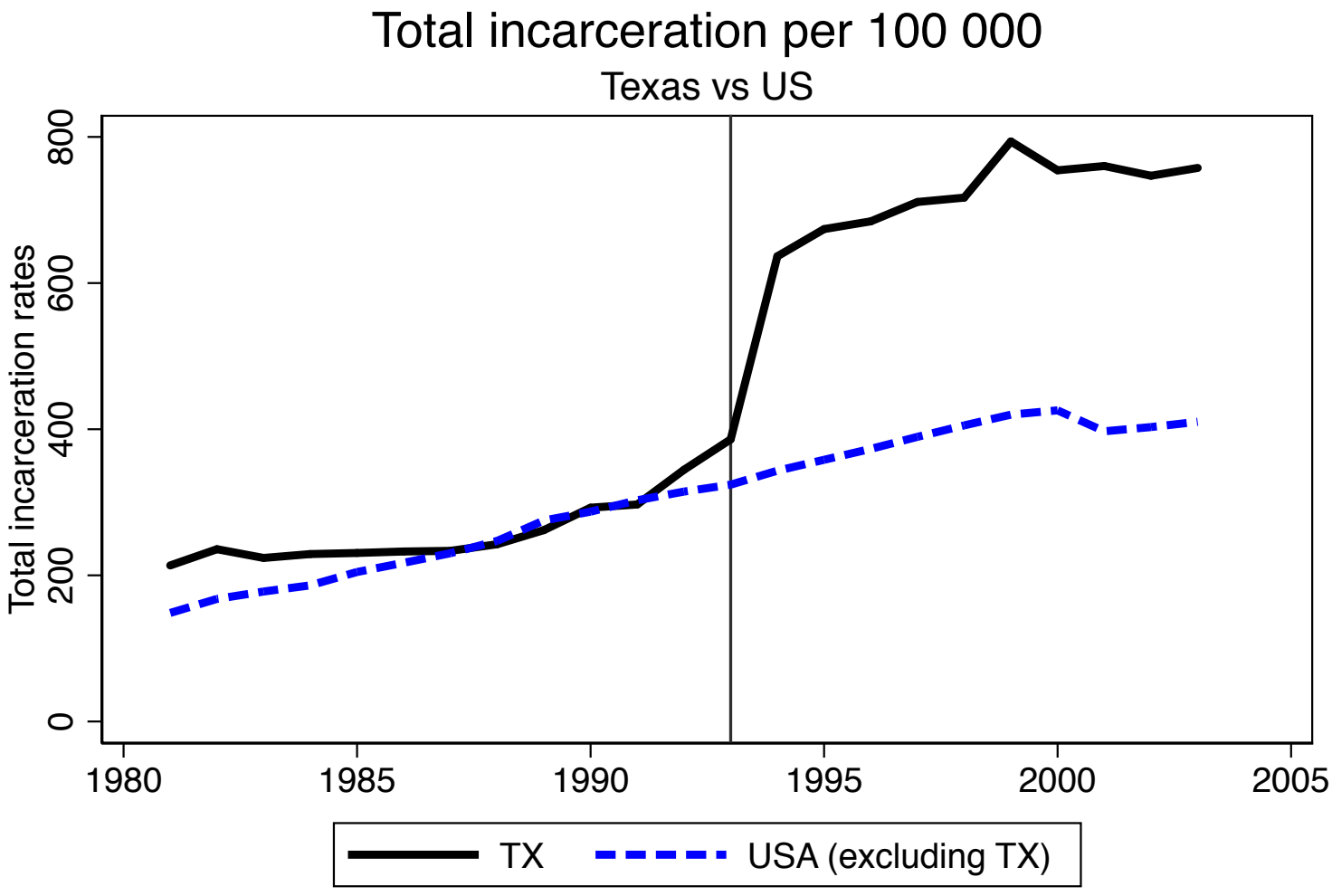
Governor Ann Richards (D) 1991-1995

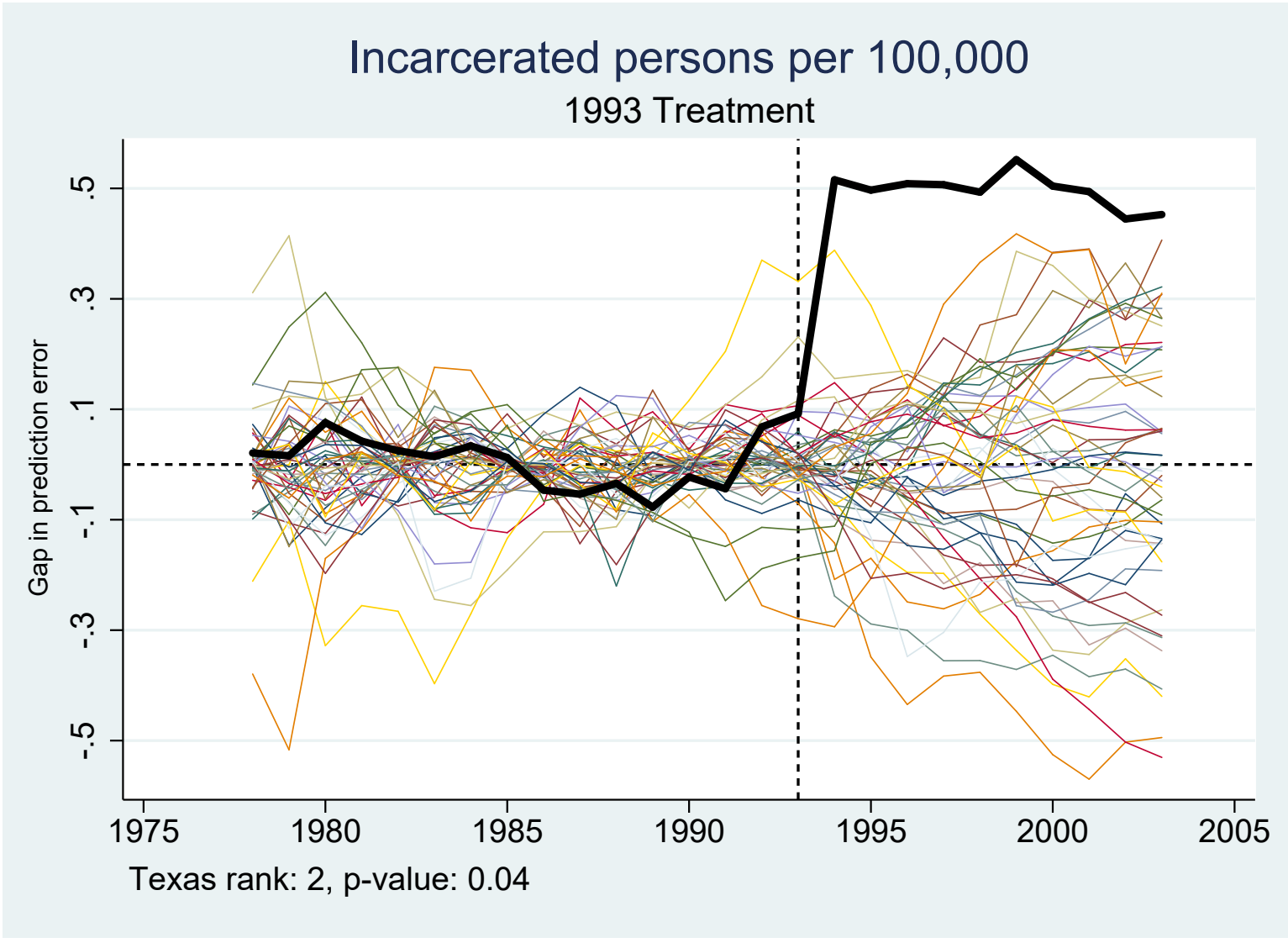
- Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
- Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
- Building of new prisons (private and public)



Red dashed line is 1993


Texas prison growth
Operational capacity
Prison capacity operational
Percent change in capacity operational
40000
60000
80000
100000
120000
140000
160000
-5
0
5
10
15
20
25
30
35
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
Operational capacity
Percent change


Total incarceration per 100 000
Texas vs US
Total incarceration rates
0
200
400
600
800
1980
1985
1990
1995
2000
2005
TX
USA (excluding TX)
1993 starts the prison expansion


Incarcerated persons per 100,000
1993 Treatment
Gap in prediction error
.5
.3
.1
-.1
-.3
-.5
1975
1980
1985
1990
1995
2000
2005
Texas rank: 2, p-value: 0.04

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

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

"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)*

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

Gist of their argument

1. ADH (“synth”) needs perfect fit and so is biased in practical settings due to the curse of dimensionality as it won’t be the case we get weights constrained to be “on the simplex”
2. Their augmentation will introduce an outcome model to estimate the bias caused by covariate imbalance
3. Introduces ridge regularization linear regression to estimate new weights to reweight synth
4. Think of it as “bias reduction” like Abadie and Imbens (2011) plus it will have doubly robust properties and be equivalent to inverse probability weighting
5. When synth is imbalanced, augmented synth will reduce bias reweighting and bias correction, and when synth is balanced, they are the same

Gist of their argument

1. Ridge regularization linear regression used to estimate weights used to reweight the original synth model
2. If synth is imbalanced, augmented synth reducees bias by reweighting and bias correction
3. When synth is balanced, the augmented and original synth are identical (but in practice, they won't be identical)
4. They argue synth DiD can be seen as a special case of augmented synth

Some topical observations

- Foregoes estimating *donor pool unit weights* (e.g., ADH, synth did, MCNN)
- Synth sequels are using penalization/regularization for estimation
- Relaxes some of the original ADH constraints, like non-negative weights (i.e., no extrapolation)
  - → This is used to address bias caused by imbalance
  - → Negative weights puts them back in the convex hull which recall we need
  - → They argue synth DiD can be seen as a special case of augmented synth

Notation

- Observe $J+1$ units over T time periods
- Unit 1 will be treated at time period $T_0 = T - 1$ (we allow for unit 1 to be an average over treated units)
- Units $j = 2$ to $J+1$ (using ADH original notation) are "never treated"
- D_j is the treatment indicator

Pre-treatment outcomes

$$\begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y^1_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y^0_{2T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{i2} & Y_{i3} & \dots & Y^0_{NT} \end{pmatrix} \equiv \begin{pmatrix} X_{11} & X_{12} & X_{13} & \dots & Y_1 \\ X_{21} & X_{22} & X_{23} & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{i2} & X_{i3} & \dots & Y_N \end{pmatrix} \equiv \begin{pmatrix} X_1 & Y_1 \\ X_0 & Y_0 \end{pmatrix}$$

This is a model of 2x2 (i.e., single last period block structure, not staggered roll out)

The last column is always post-treatment and switches from Y^1 to Y.

The last column is just showing a top row of the treated unit 1 and the bottom row of all the donor pool (i.e., we will use X_0 and Y_0 to represent all the donor pool units)

Optimal weights

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

$Y_0'w*$ (i.e., optimally weighted donor pool) is the unit 1 “synthetic control”

Predicting counterfactuals

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

We are hoping that $\widehat{Y}_1^0$ with $Y_0'w^*$ based on “perfect fit” pre-treatment

V_X matrix

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

V_x is the "importance" matrix on X_0 (Stata default chooses V_x that min pre-treatment MSE).

Penalizing the weights with ridge

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t.} \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

Modification to the original synthetic control model is the inclusion of the penalty term. “The choice of penalty is less central when weights are constrained to be on the simplex, but becomes more important when we relax this constrain.”

Convex hull

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t.} \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

These weights will be used to address imbalance, not so much the control units, bc this method is for when the weighted controls are still outside the convex hull (“simplex”)

Original ADH factor model and bias

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

Original synth factor model (with ADH notation)

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

The bias of ADH synthetic control

Perfect fit is necessary

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

Recall that the bias of ADH required “perfect fit” using their factor model (I’ll change λ factor loadings in a minute)

Perfect fit models heterogeneity

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

Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time

Remember that ADH15 quote

"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)*

Slight change in synth notation

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

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

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

Bias correction

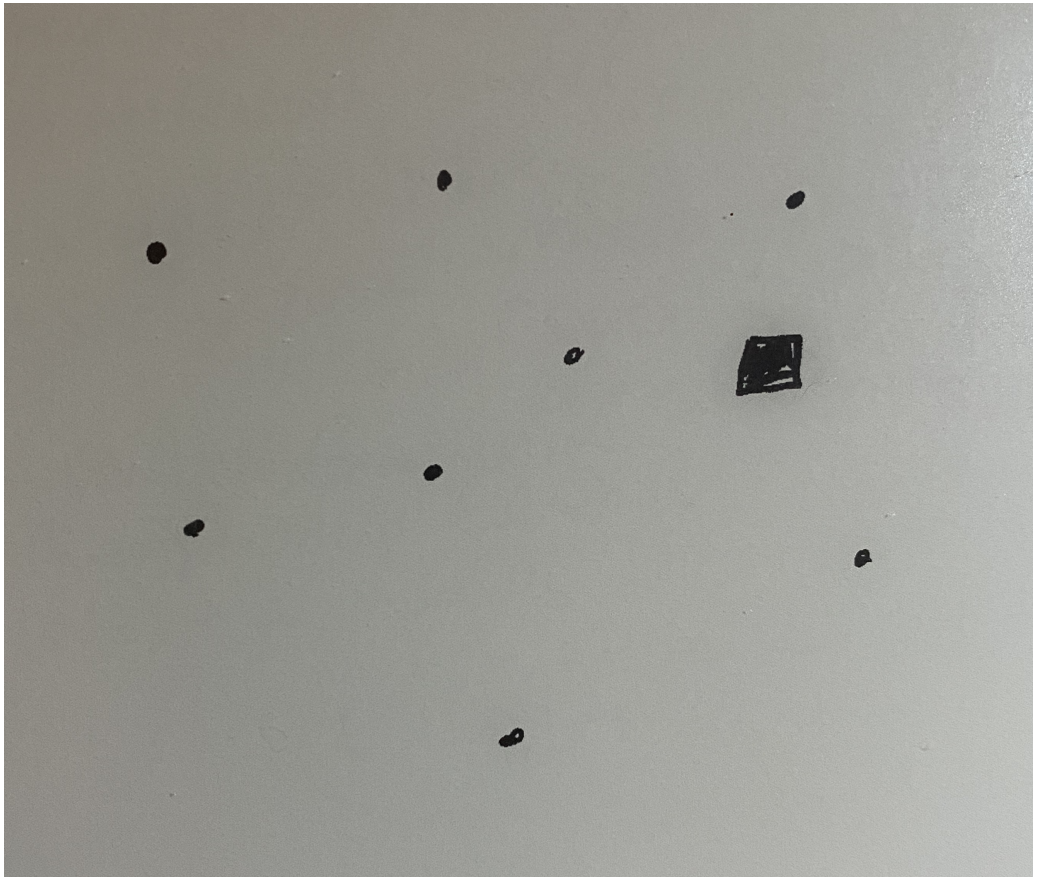
$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

- 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

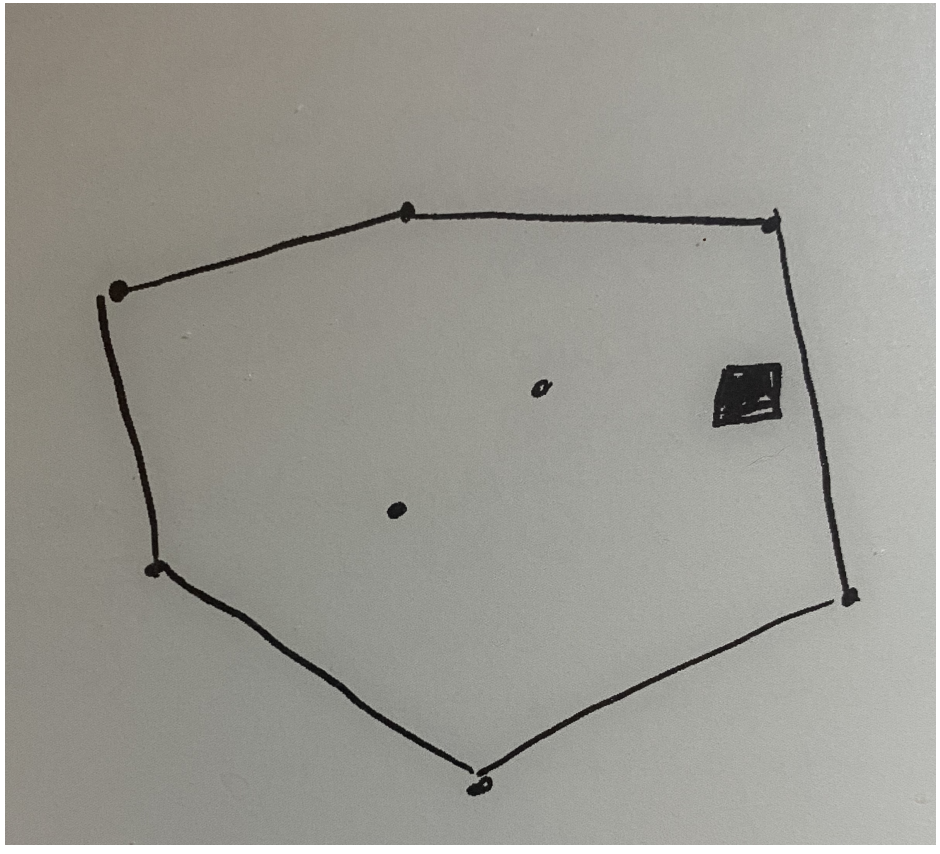
Common practice

- Usually the number of time periods isn't much larger than the number of units
- And exact balance rarely holds, which if it doesn't hold, then the unobserved heterogeneity also doesn't get deleted

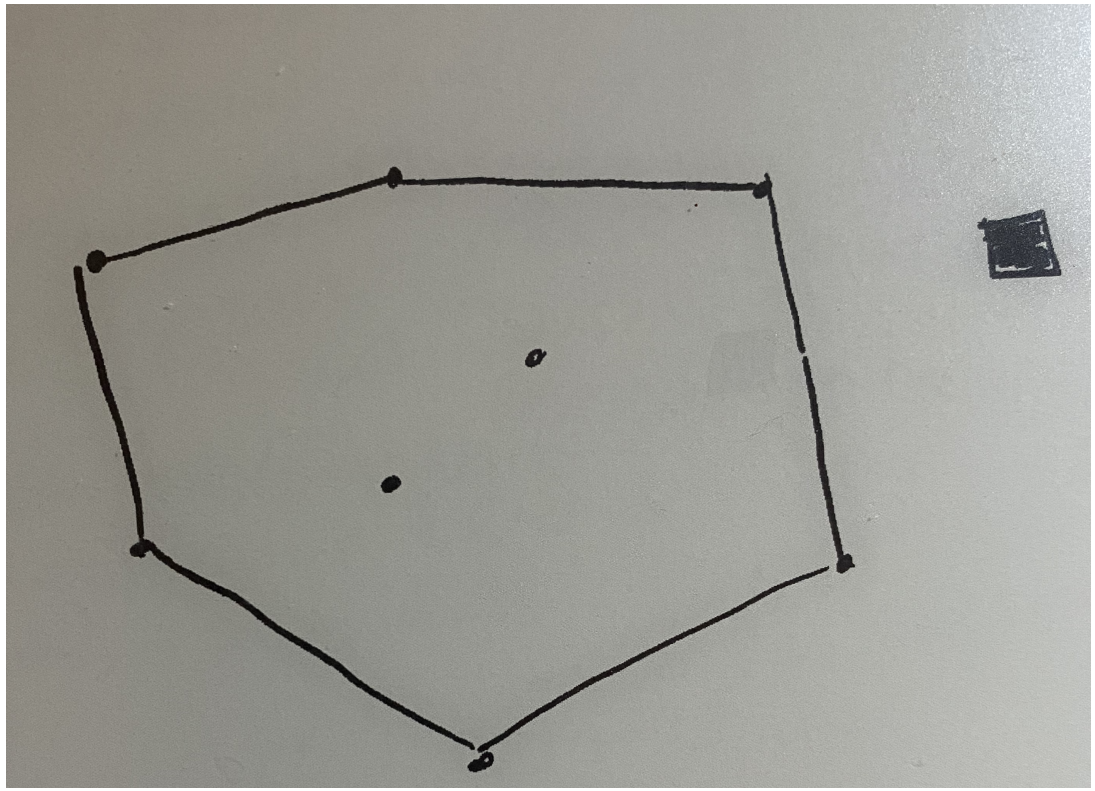
Treatment and control units



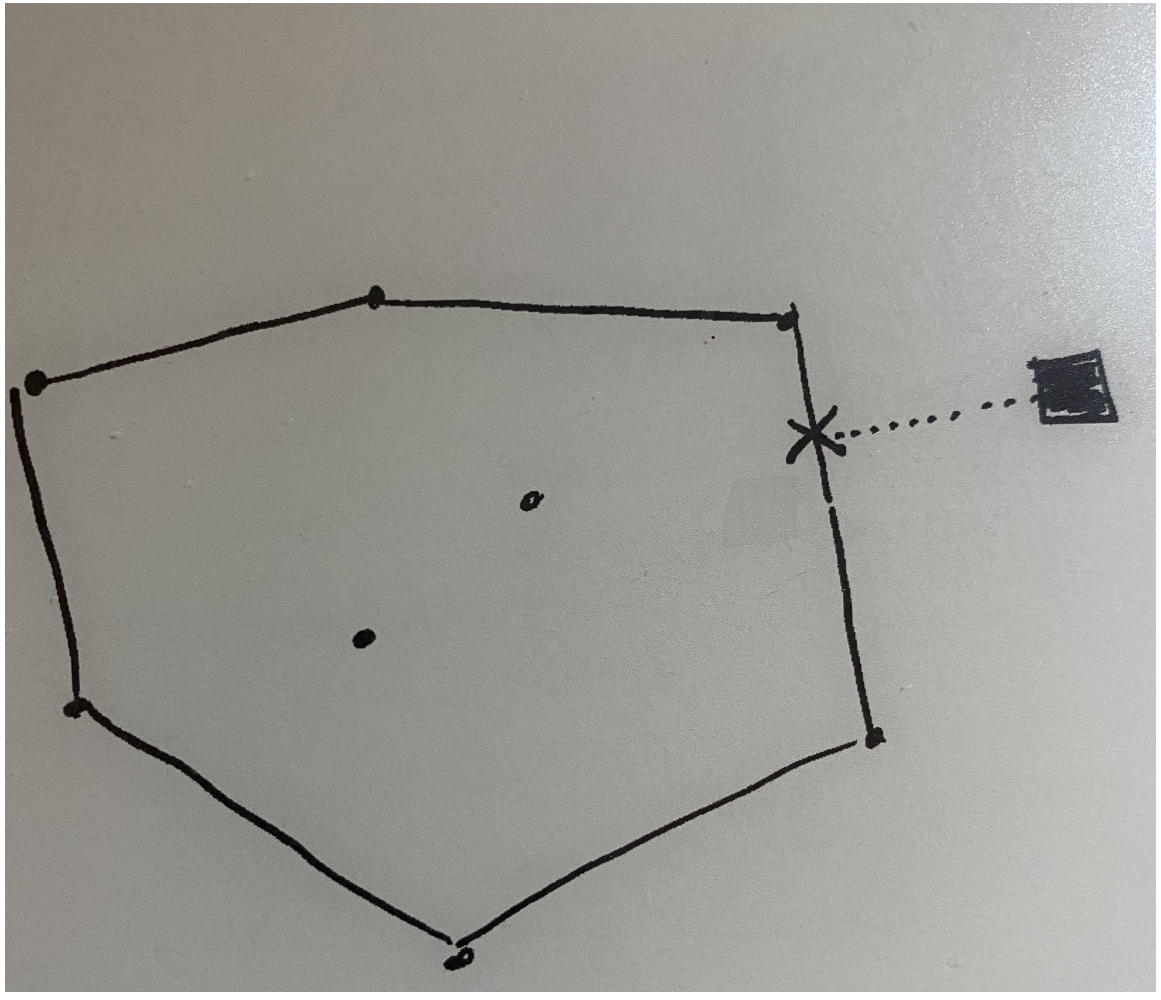
Convex hull – ideal for synth



Outside the convex hull bc of dimensionality



Outside the convex hull bc of dimensionality



Estimating the bias

- Adjust the synthetic control approach to adjust for poor fit pre-treatment.
- Recall our factor model – let $\widehat{m}_{jT}$ be an estimator for the post-treatment control potential outcome Y_{jT}^0.
- The augmented synthetic control estimator for Y_{jt}^0 is on the next slide

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

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

Interpreting line 1

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

(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 $\widehat{m}$.

Interpreting line 1

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

(1) Since $\widehat{m}$ 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

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

(1) I actually cover the bias correction of Abadie and Imbens 2011 in the mixtape! The subclassification chapter

Interpreting line 1

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

(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

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

(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

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

(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 $\widehat{m}(X_j)$ with ridge regularized linear model and penalty hyper parameter λ^{ridge}. Sorry – this is not the same λ. I didn't create this notation though! Once we have those, we adjust for imbalance using the $\widehat{\eta}^{ridge}$ parameter as a weight on the outcome model itself.

Ridge Augmented SCM

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

Once we have those, we adjust for imbalance using the $\widehat{\eta}^{ridge}$ parameter as a weight on the outcome model itself.

Go back to that weighting but use the ridge parameters

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

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

The ridge weights are key to the augmentation

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

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

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

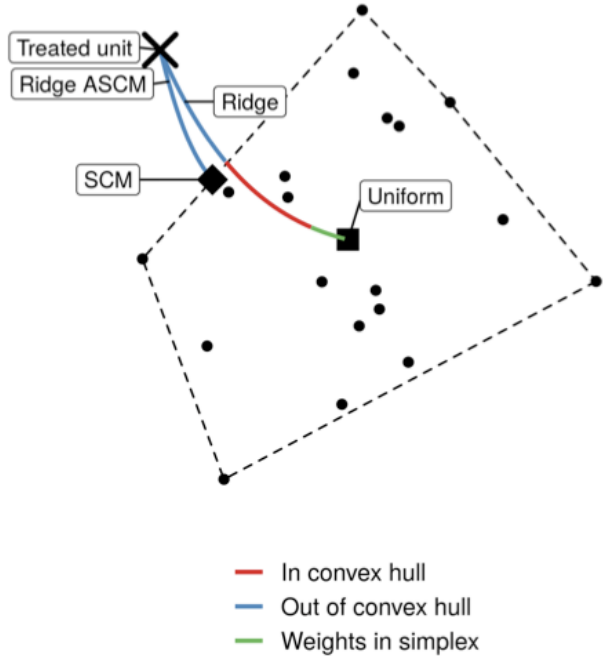
Relaxing the constraint from synth that weights be non-negative, as non-negative weights prohibit extrapolation. But we don't have synthetic control on the simplex, so we *must* extrapolate, otherwise synth will be biased.

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 $\widehat{\lambda}^{ridge}$
- 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!



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

Conformal Inference

Inference will be based on “conformal inference” method by Chernozhukov et al. (2019). We will get 95% point-wide confidence intervals. They also outline a jackknife method by Barber et al (2019).

Steps of conformal Inference

1. Choose a sharp null (i.e., no unit-level treatment effects, $\delta_0 = 0$)
   - → Enforce the null by creating an adjusted post-treatment outcome for the treated unit equal to $Y_{1T} - \delta_0$ (in other words, we get CI on the post-treatment outcomes, not the pre-treatment)
   - → Augment the original dataset to include the post-treatment time period T with the adjusted outcome and use the estimator to obtain the adjusted weights $\widehat{w(\delta_0)}$
   - → Compute a p-value by assessing whether the adjusted residual conforms with the pre-treatment residuals (see Appendix A for the exact formula)

Steps of conformal Inference

2. Compute a level α for δ by inverting the hypothesis test (see Appendix A for the exact formula)
   - $\rightarrow$ Chernozhukov et al. (2019) provide several conditions for which approximate or exact finite-sample validity of the p-values (and hence coverage of the predicted confidence intervals) can be achieved)

See Appendix A for more details

Simulations (summarized)

- They examine the performance of synth against ridge, Augmented synth with ridge regularization, demeaned synth, and fixed effects under four DGP
- Augmenting synth with a ridge outcome regression reduces bias relative to synth alone in all four simulations
- 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
- They also examine a real situation involving Kansas tax cuts in 2012

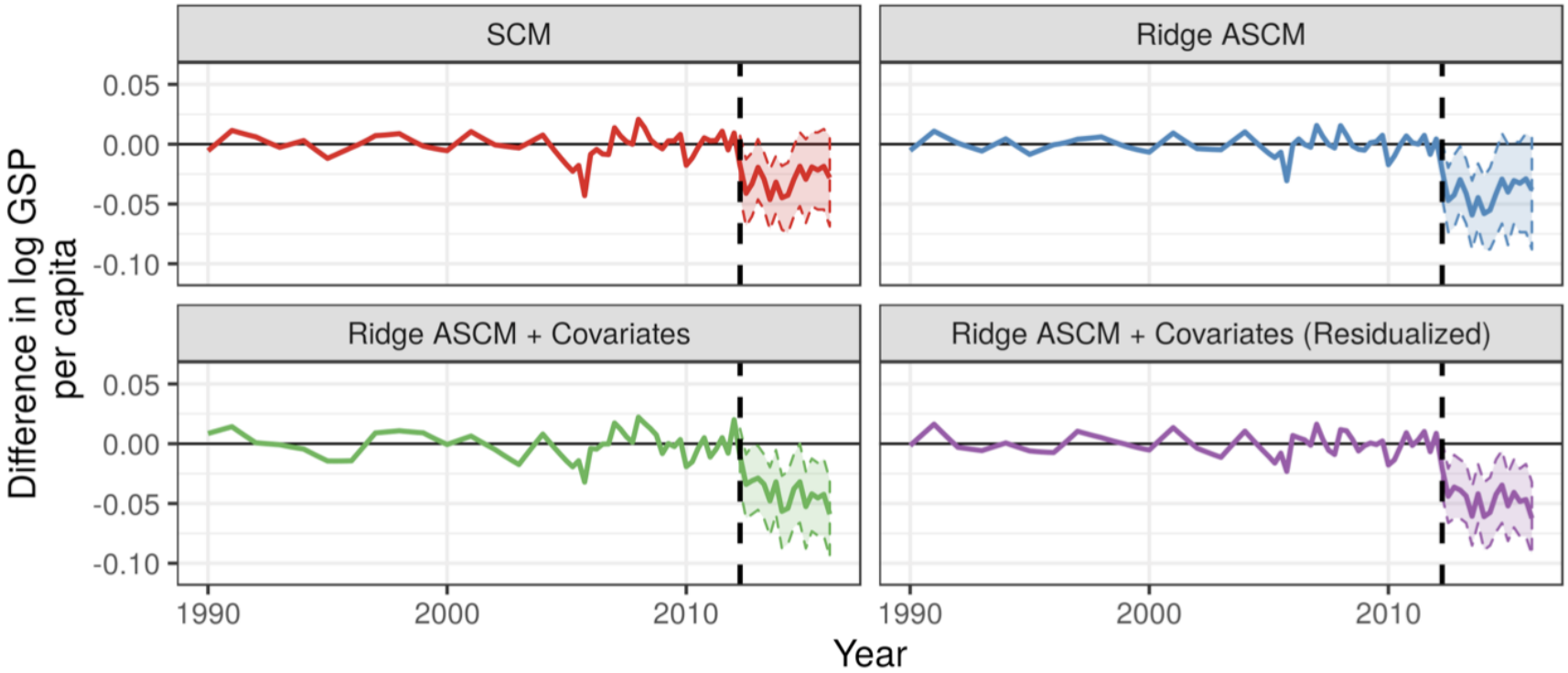
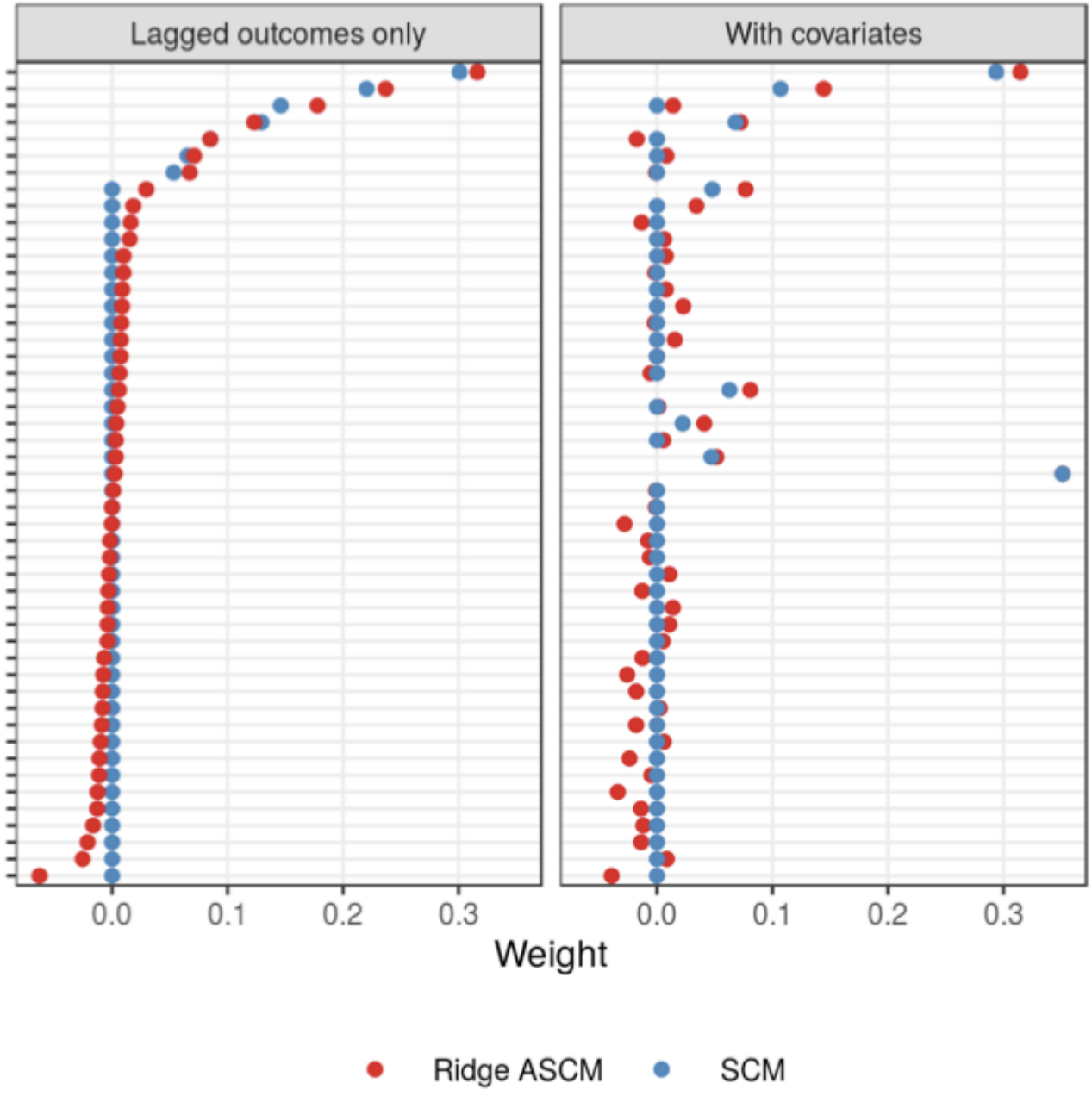


Figure 6: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.


Lagged outcomes only
With covariates
0.0
0.1
0.2
0.3
Weight
Ridge ASCM
SCM

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)

Some minor points

- We've motivated augmented synth as a kind of bias correction, but you can also think of it as correcting synth with an inverse probability weight (Appendix E)
- There's an implicit estimate of a propensity score model with ridge regularization
- Weights are odds of treatment (they're ATT weights), i.e., they're the inverse probability weighting scheme from Abadie (2005)

Augmented synth is better

- In conclusion, synthetic control is best when pre-treatment fit is excellent, otherwise it is biased
- Synthetic control avoids extrapolation by restricting weights to be non-negative and sum to one
- Ridge regression augmentation will allow for a degree of extrapolation to achieve pre-treatment balance and that creates negative weights
- Augmented synth will dominate synth in those instances by extrapolating outside the convex hull
- They also say synth DiD is a special case of their augmented synth method, which is interesting as synth DiD is also meant to nest all such modifications too (but they don't discuss augmented synth)

R code

R: https://github.com/ebenmichael/augsynth

Big idea

"The main part of the article is about the statistical problem of imputing the missing values of Y. Once these are imputed, we can estimate the causal effect of interest, δ."

"To estimate average causal effect of the treatment on the treated units, we impute the missing potential control outcomes" – Athey, et al. (2021)

Overview

- Athey, et al. (2021) unites two literatures – unconfoundedness and synthetic control
- Combines computer science with statistics to create the matrix completion with nuclear norm (MCNN) estimator
- Nuclear norm regularization is used for the imputation

What is matrix completion

- Completing a matrix means guessing at the correct values that are missing
- Hence the “completion” is just another name for “filling in” the matrix
- In causal inference, if the matrix is a matrix of potential outcomes (e.g., Y^0), then missingness is caused by treatment assignment

Here's a matrix of potential outcomes, Y^0, representing units at time t that had not been treated.

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & Y_{it}^0 \end{pmatrix}$$

Now imagine a treatment assignment, SUTVA, that flips treatment from 0 to 1 in the last period t:

$$Y = DY^1 + (1 - D)Y^0$$

Ask yourself: why are there question marks in the last column?

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Matrix completion seeks to do the following:

Matrix completion with nuclear norm will impute the last column using regularized regression:

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & \widehat{Y_{1t}^0} \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & \widehat{Y_{2t}^0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & \widehat{Y_{it}^0} \end{pmatrix}$$

And once you have those, you can calculate individual level treatment effects that can be used to aggregate to the ATT

History of matrix completion

- Open competition by Netflix in 2006 – winner would get $1m if they could improve predictive model by ten points on RMSE
- Invited a ton of competition – from MIT teams to regular everyday joes working out of their home office
- Everyone was given a database which was then tested by Netflix on a holdout dataset
- Quick progress was made followed by very slow advances
- Winner was announced in 2009

Netflix prize

- Gigantic sparsely populated matrix (100m users ranking 100k movies)
- I like Silver Linings Playbook and Lars and the Real Girl and you like Silver Linings Playbook
- Probably you'll also like Lars and the Real Girl
- So we are using correlations in the columns to "complete" missing values
- When you think about it, while it seems predictive (and it is), isn't it really a causal design?
- "If I watch Lars and the Real Girl, will I like it?"

Types of imputation

- I didn't always think of causal inference in terms of imputation because often the method was just taking existing values and manipulating them, rather than filling in missing values
- But the fundamental problem of causal inference states that causal inference is a missing data problem, so it makes sense you'd be imputing
- I tend to think therefore in terms of implicit and explicit imputation methods
- Borusyak, et al. (2021) and Athey, et al. (2021) both seem more like "explicit" imputation methods
- Callaway and Sant'Anna (2020) on the other hand is an implicit method, as is did methods more generally

Two literatures

- Lots of moving parts in this interesting paper, so my goal here is purely explainer and mostly high level at that.
- I want you to be competent and conversant in it so we also have some R code
- There's two literatures they want you to have in your mind:
  1. Unconfoundedness – $(Y^0, Y^1) \perp\!\!\!\perp D|X$ – sometimes explicitly imputes (nearest neighbor), sometimes more implicit (inverse probability weighting)
  2. Synthetic control – literally calculating a counterfactual as a weighted average over all donor pool units
- Their MCNN method will show that both are "nested" within the general framework they've developed making them actually special cases

Differences

- Conceptually different in the way they exploit patterns for causal inference
- Unconfoundedness assumes that **patterns over time**are stable *across units*
- Synth assumes **patterns across units** are stable *over time*
- Regularization nests them both
- Nuclear norm ensures a low rank matrix needed for sensible imputations

The Gist

- Factor models and interactive effects model the observed outcome as the sum of a linear function of covariates and a unobserved component that is a low rank matrix plus noise
- Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components with the rank itself estimated
- Nuclear norm regularization will be used for imputing the potential outcomes, Y^0, for all treated units
- Estimate plots and overall ATT using the estimated treatment effects

Three contributions

1. Formal results for non-random missingness when block structure allows for correlation over time. Nuclear norm is important here
2. Shows unconfoundedness and synth are in fact matrix completion methods
   - → they all have the same objective function based on the Frobenius norm for the difference between the latent matrix and the observed matrix
   - → Each approach imposes different sets of restrictions on the factors in the matrix factorization
   - → MCNN by contrast doesn't impose any restrictions – just regularization to characterize the estimator
3. Applies the method to two datasets, but I'm going to skip it though for now

Block structure

- Lots of jargon in this article – unconfoundedness, vertical and horizontal regression, fat and thin matrices.
- Unfortunately, you need to learn it all so let me try and organize it
- We define the matrix first in terms of its block structure which is describing where and when the missingness is occurring in the matrix

Unconfoundedness

- Much of the unconfoundedness literature estimates an ATE under unconfoundedness
- But it tends to focus only on a simple setup where the missingness is the last period
- Think about LaLonde (1986) – NSW treats the workers, and then you don't observe Y^0 for the treated group in the *last period*
- This is the "single-treated-period block structure" because only one *period* is missing

Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Single-treated-unit block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \cdots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \cdots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \cdots & ? \end{pmatrix}$$

Notice, this is the synthetic control design because a single unit (unit i) is missing Y^0 for the 3rd and tth periods.

Staggered adoption

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & ? & ? & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

So all of these so-called designs can be expressed in terms of missingness in the block structure, and our job therefore is to find an estimator that is general enough to manage all of them. Their MCNN will be that.

Thin and Fat matrices

- We also have to consider the relative number of panel units N and time periods T because this also shapes which regression style will be used for imputation
- Thin matrices are basically where $N >> T$, but fat matrices are ones where $T >> N$
- Approximately square ones are where T is approximately equal to N

Vertical and horizontal regression

- Two special combinations of missing data patterns and matrix shape need special attention because they are the focus of large but separate literatures
- Unconfoundedness has that single-treated period block structure with a thin matrix ($N >> T$).
- You use a large number of units and impute missing potential outcomes in the last period using controls with similar lagged outcomes
- This is the horizontal regression – imagine just running OLS on the lags and taking predicted values
- The horizontal regression holds under unconfoundedness

Vertical regression

Doudchenko and Imbens (2016) and Pinto and Furman (2019) show that Abadie, Diamond and Hainmueller (2011) can be interpreted as regressing the outcomes for the treated prior to treatment on the outcomes for controls in the same period

Fixed effects and factor models

- Both horizontal and vertical regressions exploit other patterns
- An alternative to each of them though is to consider an approach that allows for the exploitation of both stable patterns over time and stable patterns across units
- This is where their matrix completion with nearest neighbor model comes in – it does that very thing

Matrix completion with nuclear norm

- Model the $N \times T$ matrix of complete outcomes data matrix Y as:

$$Y = L* + e$$

  where $E[e|L*] = 0$
- The error term can be thought of as measurement error if you need a frame to think about it
- So you have this complete matrix, L*, and zero mean conditional independence holds

Assumption 1

Apart from the unconfoundedness assumption, we have this weird assumption!

Assumption 1

e is independent of $L*$ and the elements of e are σ-sub-Gaussian and independent of each other

Lots of matrix forms can be defined this way. But let's not get lost in the weeds – we are still just trying to estimate $L*$! That's the main storyline, not the side quest, to use Red Dead Redemption words I understand

All imputations are wrong but some are useful

- You can impute something a million different ways.
- $1 + 1 + 1 + 1 = 4$ is an imputation of the fifth unknown element and frankly just looking at it, seems wrong.
- You could minimize the sum of squared differences but if the objective function doesn't depend on $L*$, the estimator would just spit back Y and $\delta = 0$.
- They add a penalty term $||\lambda||$ to the objective function, but even then, not all of them do well.
- Turns out, it actually matters whether you regularize the fixed effects or not (just like it matters whether you regularize the constant in LASSO apparently – I decided to take their word for it)

Estimator

$$L* = \widehat{L} + \widehat{\Gamma} 1_T^T + I_N \widehat{\Delta}^T$$

where the objective function is:

$$= arg\, min_{L,\Gamma,\Delta} \left\{ \frac{1}{O} ||P_0(Y - L - \Gamma 1_T^T - 1_N \Delta^T)||_F^2 + \Lambda ||L|| \right\}$$

Fixed effects and regularization

- The penalty will likely be the nuclear norm but notice that the fixed effects are outside the penalty term. You could subsume them into L, they say, but they recommend you not doing this.
- Fraction of observations is relatively high and so the fixed effects can actually be estimated separately (apparently that is one difference between MCNN and the rest of the MC literature)
- The penalty will be chosen using cross-validation

Other norms

- One thing I thought was interesting was that the nuclear norm allowed for the construction of a low rank $L*$ matrix, but other norms actually would have weird properties
- I remember once me asking Imbens (like I had even a clue what I was talking about), “Why not use elastic net? Why are you using the nuclear norm?” He said elastic net would spit out all zeroes. I remember thinking “Why did I think I would understand what he told me?”
- One advantage of NN is its fast and convex optimization programs will do it, whereas some others won’t because of the large N or T issues
- There’s almost like a cross walk, too, between this and Borusyak, et al. (2021) but I don’t quite see it except they both leverage imputation

Conclusion

- Ultimately, this is just another model though that can be used for differential timing but at the moment, no one knows how it performs in simulations alongside Borusyak, et al. (2021), Callaway and Sant'Anna (2020) or any of the others
- So I can't really answer questions about when to use it and not to – it comes down to these very narrow assumptions
- You choose the estimator based on the problem you're studying and the assumptions – you must justify it, no one else can, but you do so by appealing to assumptions

Code

R: https://github.com/xuyiqing/gsynth

Stata: ??

New developments

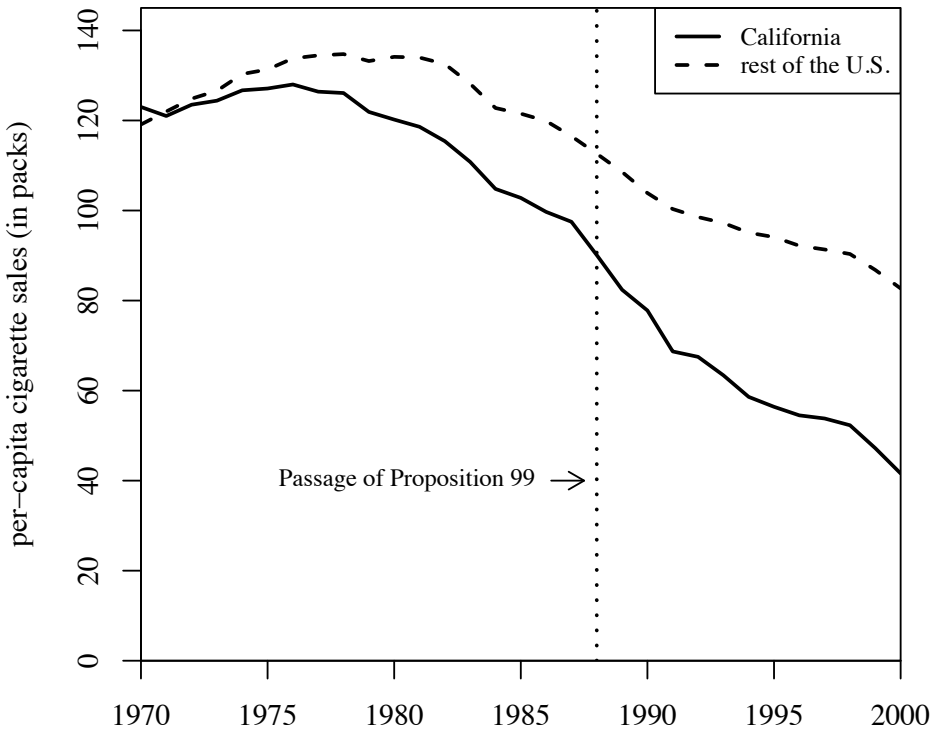
- Remember what Athey and Imbens said – “most important innovation in causal inference of the last 15 years”.
- The synthetic DiD bears some similarities to their MCNN model, but focuses on estimating weights, not the L^* matrix
- It will dominate the Abadie, Diamond and Hainmueller (2010) as they will show and addresses overfitting and other things through estimating oracle weights (which I’ll explain towards the latter half)

Imperfect fits

- Recall that ADH needs to fit a pre-treatment convex hull to model the heterogeneity
- Often, though, the fit is imperfect for various reason because weights are constrained to be non-negative and sum to one
- But this can be problematic if the treatment group can't be approximated by a weighted average of other units since the weights are fractions
- So they're going to allow for a constant level shift to "get there"

Diff-in-diff, parallel trends and pre-trends

- Recall the identifying assumption in DiD – parallel trends
- Untestable, but we often use pre-trends for an indirect test
- But in the smoking example, parallel trends didn't hold for many states
- Choice of control units matter – the average trends for many control states are roughly parallel, but not all



Weights and controls

- ADH sought a weighted average over the control units to recreate the pre-trend through a fitting exercise
- Synthetic control becomes the weighted average of controls, and then the focus is just on estimating weights
- All we ask is that the weighted average follow the same dynamic path as treatment group (a fit for each period)

Regressions up and down

- Doudchenko and Imbens (2015) note that synth weights are based on a "vertical regression" yielding coefficients on the control units (as opposed to the lags in T which is a horizontal regression)

$$Y_{1,t}^{0} = \sum_{j=2}^{J+1} \widehat{\omega_j} \times Y_{j+1,t}$$

- To the degree the fit is good pre-treatment, then the gaps post-treatment measure ATT at a point in time

Weighting across controls

Assume that the synthetic control at any period is $Y_{1,t} \approx \sum_{j=2}^{J+2} w_i \times Y_j$

- Synthetic control – weights, $\widehat{w}$, control units to get weighted average controls
  1. Use the pre-treatment data to find the optimal weights that when aggregated over control units predict treatment group outcomes ("fit")
  2. Assumes that there's a stable relationship over time, though, because this is going to be our estimated counterfactual post-treatment
- This is shown to be equivalent to a "vertical regression" where you regress units against the higher column units to get those weights
- May require regularization in the regression (if there are more units than time periods)

Weighting across time dimensions

- Forecasting – time weights, $\widehat{\lambda}$, periods to get weighted average periods
  1. Use the controls to learn an average of periods that forecast what we see post-treatment
  2. Imagine a regression, in other words, that yields coefficients on covariates, not on units, to predict future counterfactual
  3. Assumes that this relationship remains valid for the treated and we use the same average of periods to impute the Y^0 for our treatment group
- This is equivalent to a "horizontal regression" where you regress outcomes against the leads (i.e., Y_{it} against $Y_{i,t-1}$) – this is what was meant by unconfoundedness from the MCNN lecture
- Again may need regularization if there are more time periods than units

Difference-in-differences model

- They tend to equate DiD with a TWFE model

$$Y(0)_{it} = \mu + \alpha_i + \gamma_t + \varepsilon_{it}$$

  and solve for the unknown parameters
- More generally, these are the factor models

Reconciling these things

- Vertical regression (i.e., the ADH synth approach) assumes there is a stable relationship between units over time (hence why the weights accurately estimate counterfactuals post-treatment)
- Horizontal regression (i.e., the unconfoundedness approach) is similar, but assumes a stable relationship between outcomes in the treatment period and pre-treatment periods that is the same for all units
- DiD regression (TWFE): assumes an additive outcome model that captures differences between time and units

So the focus becomes about choosing between these methods

Synthetic DiD

Synthetic DID takes synth and forecasting to create a *synthetic DiD* version

- Combine these two – weighting controls using pre-treatment and weighting time using controls, then applying a type of DiD differencing – to create the synthetic DiD model
- There is a focus, just like ADH, on estimating appropriate weights
- It's doubly robust – only one has to remain valid
- Constant effects will get differenced out and the synthetic control can be *parallel* to treatment, as opposed to *identical* in pre-treatment period

Estimation of SDiD

Synthetic DiD is DiD with a synthetic control and a pre-treatment period (on the baseline, just like CS).

1. Compute the regularization parameter to match the size of a typical one-period outcome change, $\Delta_{it} = Y_{i(t+1)} - Y_{it}$, for unexposed

Estimation of SDiD

2. Estimate unit weights $\widehat{w}$ defining a synthetic control unit (just like Abadie, Diamond and Hainmueller 2010) using the pre-treatment data

$$\widehat{w}_1 + \widehat{w}^T Y_{j,pre} \approx Y_{1,pre}$$

but they allow for an intercept term so that now the weights no longer need to make the unexposed pre-trends *perfectly* match the treatment group (hence convex hull can fail to hold)

Estimation of SDiD

3. Estimate the time weights $\widehat{\lambda}$ defining a synthetic pre-treatment period using control data

$$\widehat{\lambda}_{j=1} + Y_{1,pre}\widehat{\lambda} \approx Y_{1,post}$$

Estimation

4. Compute the SDID estimator via the weighted DID regression

$$\arg\min_{\tau,\mu,\alpha,\beta} = \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \widehat{w}_i^{sdid} \widehat{\lambda}_t^{sdid} \right\}$$

Estimating the weights

Our focus then becomes about estimating $\widehat{w}$ and $\widehat{\lambda}$

5. Estimate the control weights, $\widehat{w}$, defining the control group unit via constrained least squares on the pre-treatment data. This requires weights to be non-negative and sum to one and allows for a level shift with regularization. Synthetic control is a weighted average like in ADH

Estimating the weights

6. We then estimate the time weights. $\widehat{\lambda}$, defining the synthetic pre-treatment period via constrained least squares on the control data with analogous time constraints

More formalization

Assumed data generating process – outcome is “low rank matrix” (MCNN) plus noise

$$Y = L + \tau D + E$$

where L is the systematic component and the conditional expectation of the error matrix E given the assignment matrix D and the systematic component of L is zero.

We won’t estimate L^* though, unlike MCNN

Data generating process – noise and signal

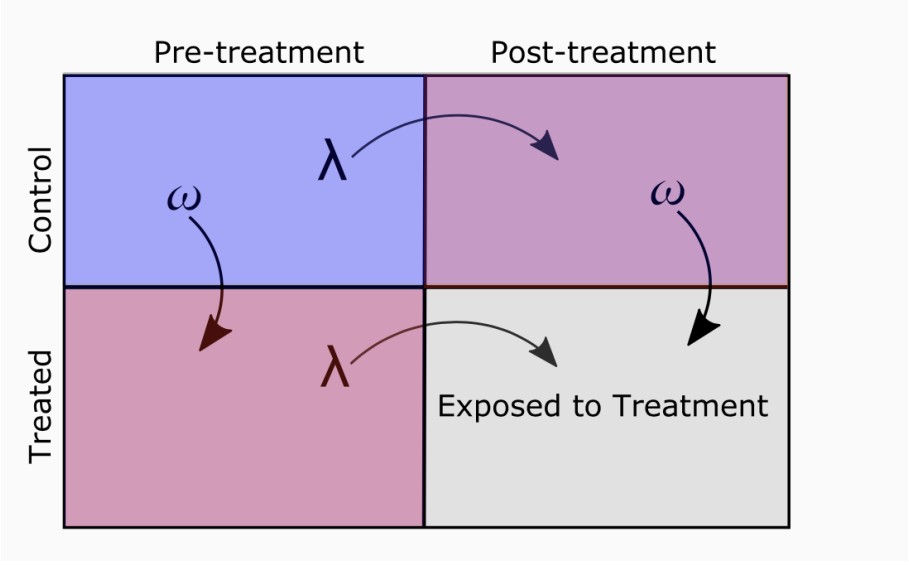
$$Y = L + \tau D + E$$

The treatment cannot depend on the error term, but may depend on the systematic elements of L (i.e., D is not randomized). Think of L as the signal, τ a matrix of treatment effects and E the noise with no autocorrelation over time or between units. The only thing random is E, our noise matrix.

Estimating the weights – high level

- Modify synthetic control weights – use penalized least squares to get a weighted average of control units with pre-trends "parallel" to the treated unit average
- But they'll allow for a constant, unlike ADH synth
- And then they'll do the same thing for the time weights, but this time they won't regularize because they want to weight more intensively the periods "just before" – ridge, they note, would "spread out the weights" over multiple time periods and they don't want that
- I'll get more into this with the oracle weights, but for now I'll just note it conceptually

Picture



(credit: David Hirshberg January 2020 slides because I can't make this picture to save my life)

Regression

- SC is weighted linear regression with no unit FEs:

$$\tau^{sc} = \text{argmin}_{\tau,\lambda} \sum_{i,t} (Y_{it} - \lambda_t - \tau D_{it})^2 \times w_i^{sc}$$

- DiD is unweighted regression with unit FEs and time FEs:

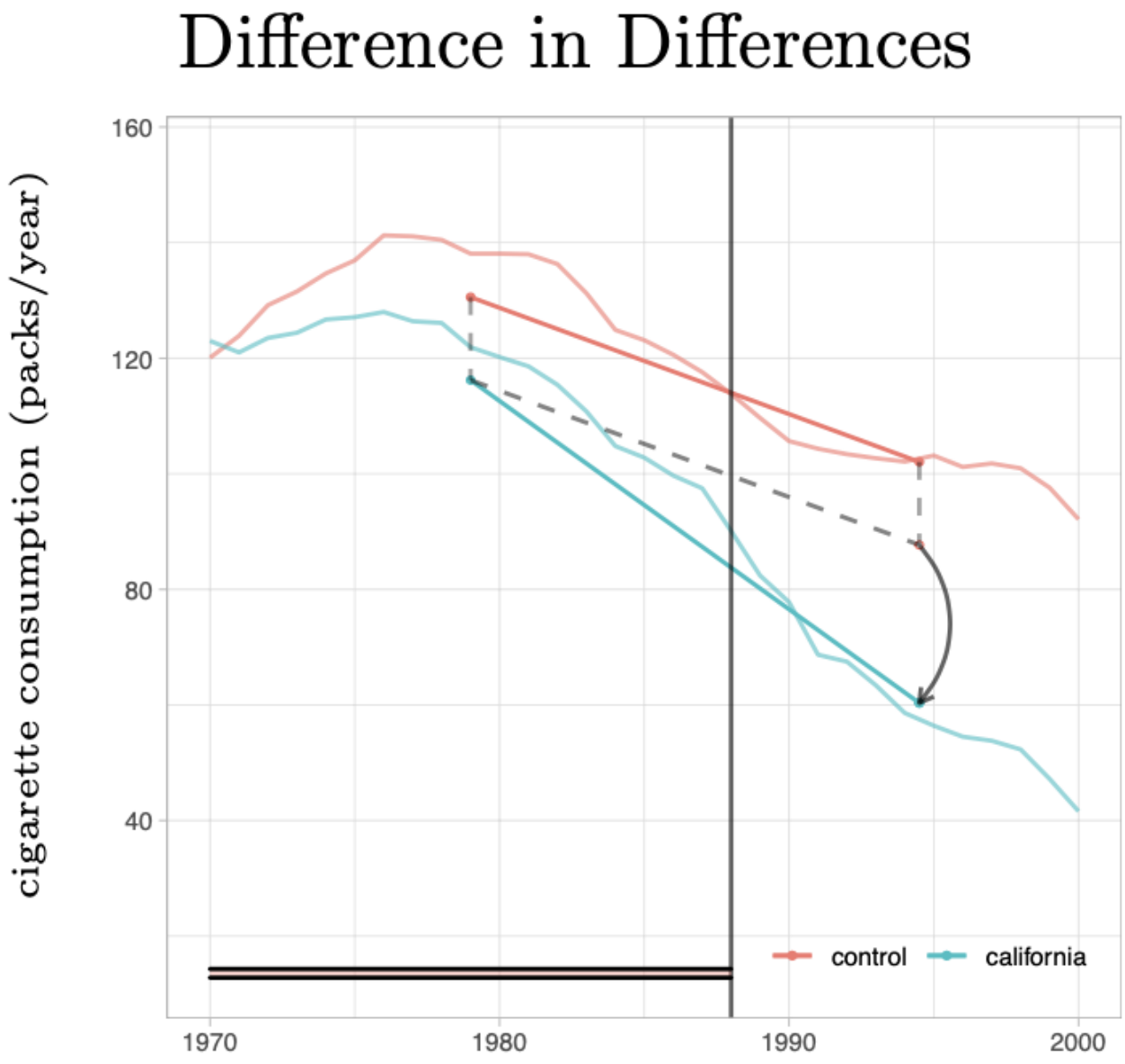
$$\text{argmin}_{\tau,\lambda,\alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2$$

- SDiD is weighted regression with unit FEs and time FEs:

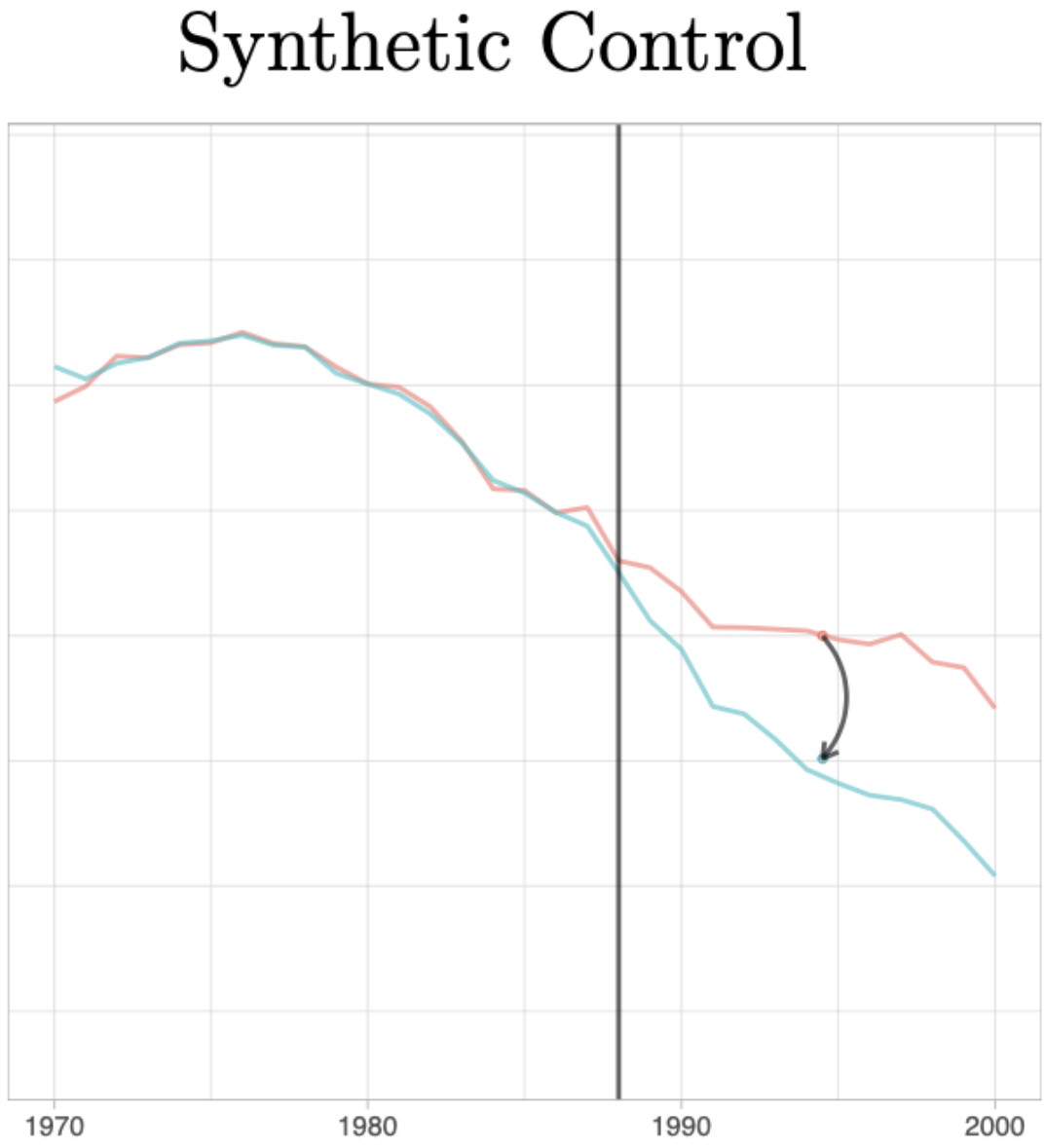
$$\text{argmin}_{\tau,\lambda,\alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2 \times w_i \times \lambda_t$$

Formal results overview

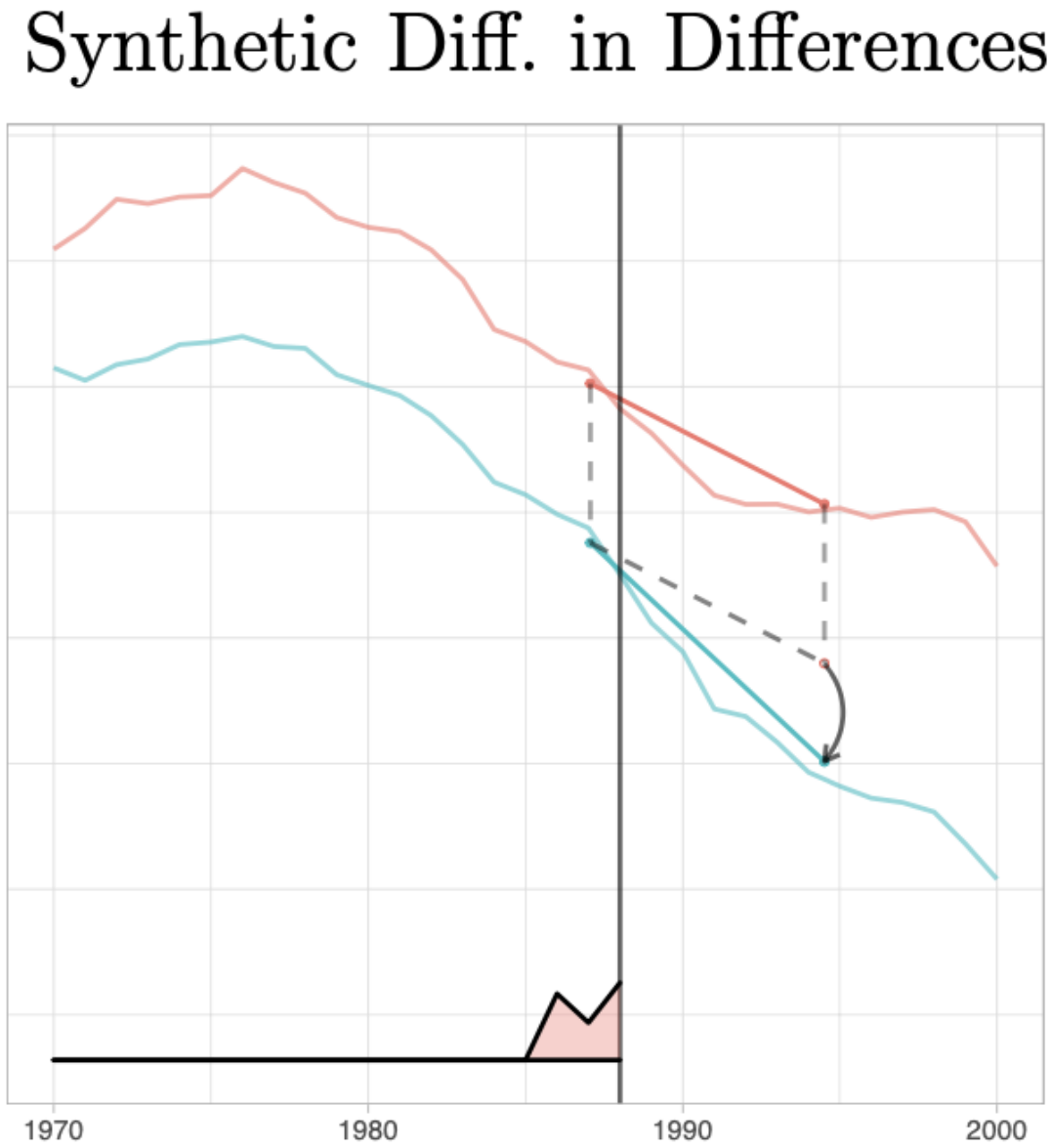
- Formal results will show SDiD is “doubly robust” (recall Sant’Anna and Zhao 2020)
- Factor model on the outcome can be a latent factor model but true model is that signal model and it’ll still be consistent
- Asymptotic normality of $\widehat{\tau}^{SDiD}$
- With oracle weights, SDiD will have “good weights”
- You can do inference through resampling like jackknife, bootstrap and randomization inference



Estimated decrease: -27.3 (17.7)



Estimated decrease: -19.6 (9.9); bad fit just prior bc weights are fitting everywhere



Estimated decrease: -15.4 (8.4). Jagged line left of 1988 is the weighting of those years

Practical problems

- Underfitting. What if I can't get a parallel synthetic control? I know because it's visible. This is an underfitting problem. We need more controls, better controls, or another method.
- Omitted variable bias. Something else happens exactly when the treatment occurs. Sorry – there isn't a solution, because you're not identified.
- Overfitting. We get a synthetic control, but it's because the plot over fit the data. This means that you've not approximated the counterfactual post-treatment. No different than in RDD when you're unable to identify the counterfactual due to functional form problems.

How to rule out overfitting: oracle weights

- Their estimator is equivalent to an “oracle estimator” which cannot overfit
- Oracle uses unit and time weights that don’t depend on the noise
- Weights minimize MSE; oracle weights minimize **expected** SE

Decomposing the bias of SDID

$$
\begin{aligned}
\widehat{\tau}^{sdid} - \tau &= \varepsilon(\widetilde{w}, \widetilde{\lambda}) + B(\widetilde{w}, \widetilde{\lambda}) + \widehat{\tau}(\widehat{w}, \widehat{\lambda}) - \widehat{\tau}(\widetilde{w}, \widetilde{\lambda}) \\
&= \text{oracle noise} + \\
&\quad \text{oracle confounding bias} + \\
&\quad \text{deviation from oracle}
\end{aligned}
$$

So they characterize these terms

Oracle noise

First term: the oracle noise

$$\varepsilon(\widetilde{w}, \widetilde{\lambda})$$

Tends to be small when the weights are small and there are a sufficient number of exposed units and time periods.

Oracle confounding bias (rows / units)

$$B(\widetilde{w}, \widetilde{\lambda})$$

Will be small when the pre-exposure oracle row (units) regression fits well and generalizes to the exposed rows :

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,pre} \approx \widetilde{w}_1^T L_{1,pre}$$

and

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,post} \approx \widetilde{w}_1^T L_{1,post}$$

Oracle confounding bias (columns / time)

$$B(\widetilde{w}, \widetilde{\lambda})$$

Will be small when the pre-exposure oracle column (time) regression fits well and generalizes to the exposed columns :

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,pre} \approx \widetilde{\lambda}_1^T L_{1,pre}$$

, and

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,post} \approx \widetilde{\lambda}_1^T L_{1,post}$$

Oracle confounding bias – neither do well

What if neither model generalizes well on its own, then there is a doubly robust property

It is sufficient for one model to predict the generalization error of the other

“The upshot is even if one of the sets of weights fails to remove the bias from the presence of L, the combination of oracle unit and time weights can compensate for such failures”

Deviation from Oracle

Core theoretical claim (All formalized in their asymptotic analysis): SDID estimator will be close to the oracle when

- The oracle time and unit weights look promising on their respective training sets

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,pre} \approx \widetilde{w}_1^T L_{1,pre}$$
$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,pre} \approx \widetilde{\lambda_1}^T L_{1,pre}$$

- and regularization is not too large for either weight

Properties

Under some assumptions, they provide then that SID:

1. SDiD is approximately unbiased and normal
2. SDiD has a variance that is optimal and estimable via clustered bootstrap

Placebo Simulation

- Big picture still – they do a simulation to evaluate bias, RMSE of estimates compared to the observed outcome, but they don't want to use randomization because that may not catch the distinct time trend
- They want the simulation to be "realistic" not "ideal" (i.e., design based identification using randomized treatment dates)
- Bertrand, et al. (2004) randomly assigned a set of states in the CPS to a placebo treatment and the rest the control and examine how well different approaches to inference for DiD covered the true effect of zero
- Only methods that were robust to serial correlation of repeated observations for a given unit (e.g., clustering by level of treatment) attained valid coverage

Treatment assignment process

- Policy: abortion laws, gun laws, minimum wages with outcome hours and unemployment rate
- Logistic regression to predict presence of regulation on four state factors from simulation outcome model M
- Goodness of fit shows that treatment assignment responds strongly to unobserved latent factors
- Assign treatment to states with probabilities from the logistic model

Some details of this placebo simulation

- They calculate average earnings over 40 years and 50 states by subtracting the overall mean and dividing by the standard deviation to get a matrix Y with $||Y||_2^2 = 1$
- They fit a rank 4 factor model M
- They then extract TWFE from there based on unit and time fixed effects F
- Extract low rank matrix as $L = M - F$
- Calculate residuals $E = Y - M$ on an AR(2) model
- Compared SDID, DiD, synthetic control and matrix completion under different baseline scenarios and SDID tends to better

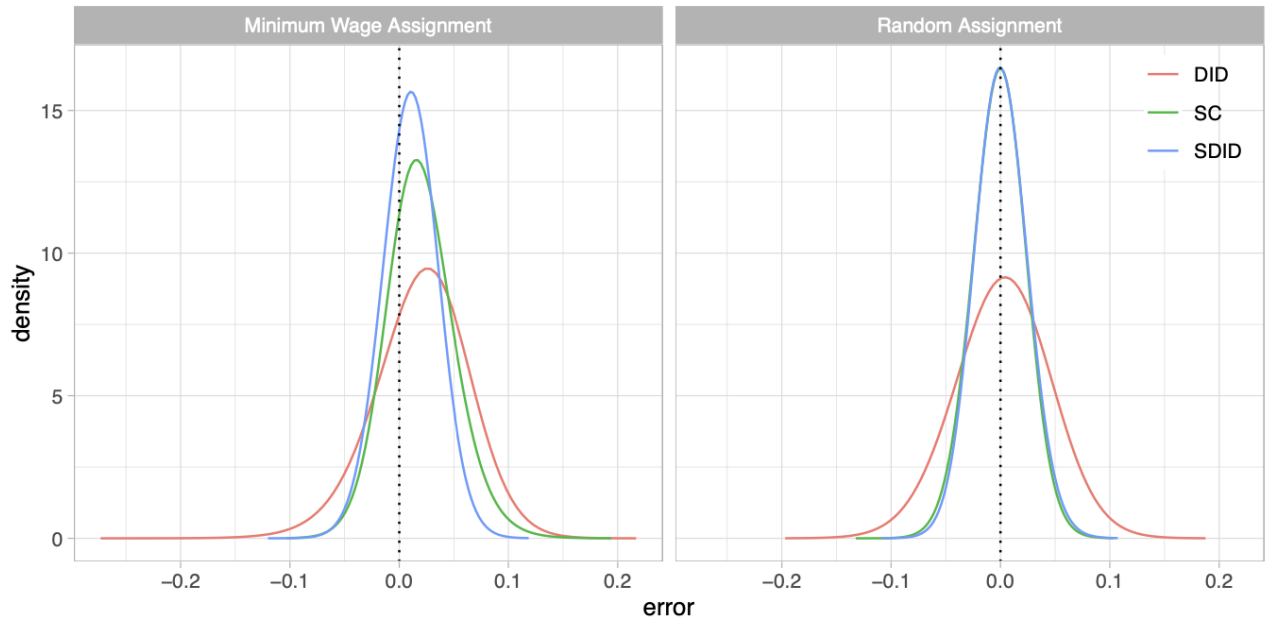


Figure 2: Distribution of the errors of SDID, SC and DID in the setting of the "baseline" (i.e., with minimum wage) and random assignment rows of Table 2.

	RMSE					Bias				
	SDID	SC	DID	MC	DIFP	SDID	SC	DID	MC	DIFP
Baseline	**0.28**	0.37	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
Outcome Model										
No Corr	**0.28**	0.38	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
No $\boldsymbol{M}$	0.16	0.18	**0.14**	**0.14**	0.16	0.01	0.04	0.01	0.01	0.01
No $\boldsymbol{F}$	0.28	**0.23**	0.49	0.35	0.32	0.10	0.04	0.21	0.15	0.07
Only Noise	0.16	**0.14**	**0.14**	**0.14**	0.16	0.01	0.01	0.01	0.01	0.01
No Noise	0.06	0.17	0.47	**0.04**	0.11	0.05	0.04	0.20	0.00	0.01
Assignment Process										
Gun Law	**0.26**	0.27	0.47	0.36	0.30	0.08	-0.03	0.15	0.15	0.09
Abortion	**0.23**	0.31	0.45	0.31	0.27	0.04	0.16	0.03	0.02	0.01
Random)	**0.24**	0.25	0.44	0.31	0.27	0.01	-0.01	0.02	0.01	-0.00
Outcome Variable										
Hours	1.90	2.03	2.06	**1.85**	1.97	1.12	-0.49	0.85	1.00	1.00
U-rate	**2.25**	2.31	3.91	2.96	2.30	1.77	1.73	3.60	2.63	1.69
Assignment Block Size										
$T_{\text{post}} = 1$	**0.50**	0.59	0.70	0.51	0.54	0.20	0.17	0.38	0.21	0.12
$N_{\text{tr}} = 1$	**0.63**	0.73	1.26	0.81	0.83	0.03	0.15	0.11	0.05	-0.02
$T_{\text{post}} = N_{\text{tr}} = 1$	1.12	1.24	1.52	**1.07**	1.16	0.14	0.24	0.33	0.16	0.11

Table 2: Simulation Results for CPS Data. The baseline case uses state minimum wage laws to simulate treatment assignment, and generates outcomes using the full data-generating process described in Section II.1.1, with $T_{\text{post}} = 10$ post-treatment periods and at most $N_{\text{tr}} = 10$ treatment states. In subsequent settings, we omit parts of the data-generating process (rows 2-6), consider different distributions for the treatment exposure variable D_i (rows 7-9), different distributions for the outcome variable (rows 10-11), and vary the number of treated cells (rows 12-14). The full dataset has $N = 50$, $T = 40$, and outcomes are normalized to have mean zero and unit variance. All results are based on 1000 simulation replications and are multiplied by 10 for readability.

Inference

This can be used to motivate practical methods for large-sample for inference. You can use conventional confidence intervals to conduct asymptotically valid inference, and they discuss three ways: jackknife, bootstrap, and placebo variance estimation.

	Bootstrap			Jackknife			Placebo		
	SDID	SC	DID	SDID	SC	DID	SDID	SC	DID
Baseline	0.96	0.93	0.89	0.93	—	0.92	0.95	0.88	0.96
Gun Law	0.97	0.96	0.92	0.94	—	0.93	0.94	0.95	0.93
Abortion	0.96	0.94	0.93	0.93	—	0.95	0.97	0.91	0.96
Random	0.96	0.96	0.92	0.93	—	0.94	0.96	0.96	0.94
Hours	0.92	0.96	0.94	0.89	—	0.95	0.91	0.90	0.96
Urate	0.78	0.74	0.38	0.71	—	0.42	0.74	0.77	0.41
$T_{\text{post}} = 1$	0.93	0.94	0.84	0.92	—	0.88	0.92	0.90	0.92
$N_{\text{tr}} = 1$	—	—	—	—	—	—	0.97	0.95	0.96
$T_{\text{post}} = N_{\text{tr}} = 1$	—	—	—	—	—	—	0.96	0.94	0.94
Resample, $N = 200$	0.94	0.96	0.92	0.95	—	0.93	0.96	0.95	0.94
Resample, $N = 400$	0.95	0.91	0.96	0.96	—	0.95	0.96	0.90	0.96
Democracy	0.93	0.96	0.55	0.94	—	0.59	0.98	0.97	0.79
Education	0.95	0.95	0.30	0.95	—	0.34	0.99	0.90	0.94
Random	0.93	0.95	0.89	0.96	—	0.91	0.95	0.94	0.91

Table 4: Coverage results for nominal 95% confidence intervals in the CPS and Penn World Table simulation setting from Tables 2 and 3. The first three columns show coverage of confidence intervals obtained via the clustered bootstrap. The second set of columns show coverage from the jackknife method. The last set of columns show coverage from the placebo method. Unless otherwise specified, all settings have $N = 50$ and $T = 40$ cells, of which at most $N_{\text{tr}} = 10$ units and $T_{\text{post}} = 10$ periods are treated. In rows 7-9, we reduce the number of treated cells. In rows 10 and 11, we artificially make the panel larger by adding rows, which makes the assumption that the number of treated units is small relative to the number of control units more accurate (we set N_{tr} to 10% of the total number of units). We do not report jackknife and bootstrap coverage rates for $N_{\text{tr}} = 1$ because the estimators are not well-defined. We do not report jackknife coverage rates for SC because, as discussed in the text, the variance estimator is not well justified in this case. All results are based on 400 simulation replications.

Some practical considerations

More treated units is worse – when we add treated units, the oracle standard deviation decreases faster leaving too little room for other sources of error to disappear in the noise

More practical considerations

Circumstances are ideal if the signal matrix L admits a good oracle synthetic control and synthetic pre-treatment period and it's too complex

- What is good? Oracle control weights distribute mass over enough control units
- Oracle time weights should distribute the rest of its mass over enough time periods

More practical considerations

Interestingly, this is an overlap assumption (like common support in matching and CS DiD):

- Many control units are like the treated ones
- Many pre-treatment periods are comparable to post-treatment ones

More practical considerations

What is “not too complex” signal matrix L? It’s one that looks different from the matrix of noise

- More about the rank of the matrix – it must be moderate rank
- Moderate means smaller than the square root of the number of control units
- A state’s behavior isn’t idiosyncratic, but characterized by a blend of industries, etc. of relatively few trends

More practical considerations

- Including more controls won't hurt you bc the set of weights is small and the error is insensitive to dimension
- Less than ideal circumstances can be problematic. The error gets worse:
  - → Signal is too complex
  - → Fit and dispersion of the oracle weights is poor

Some comments

- Conceptually, this is ADH synth combined with a simple 2x2 DiD where the weights are based on estimated time and control group weights
- Oracle weights will make improvements that don't suffer from some of the practical problems, like overfitting, that we said
- Synth DiD dominates synthetic control
- Still remains to be seen how we are going to go about choosing between these, but some things we may need to put down (ADH)

R code: synthdid

Let's look at the code together

Code: https://github.com/synth-inference/synthdid

Vignettes: https://synth-inference.github.io/synthdid/articles/more-plotting.html

Application: Melo, Neilson and Kemboi 2023

"Indoor Vaccine Mandates in US Cities, Vaccination Behavior and COVID-19 Outcomes" by Vitor Melo, Elijah Neilson and Dorothy Kemboi, 2023 working paper

Study investigates the effect of city-level vaccine mandates (implemented in US cities) on COVID-19 cases, deaths or vaccine uptake in the cities

Authors use Arkhangelsky, et al. (2021) "synthetic difference-in-differences", as well as conventional synthetic control and difference-in-differences and finds no effect of either the announcement or implementation of the mandate had any significant effect on the outcomes

Motivation

- Many policies and strategies were taken to incentivize citizens to get vaccinated and reduce COVID-19 spread
- Indoor vaccine mandates, one of the more restrictive, prevented people from entering public places (e.g., theaters, restaurants) without proof of vaccination
- Many large cities (NYC, San Francisco, LA, Seattle, Boston, Philadelphia) implemented with the stated goal to raise vaccination rates and slow spread and mortality from COVID-19

Motivation

- Vaccine viewed as crucial step toward controlling the virus and return life to normal
- Substantial number of Americans were unwilling to be immunized
- February 2021, 30% of adults say they would probably or definite not be vaccinated
- Low vaccination rates led to measures to increase uptake like mandated vaccination and weekly testing, lotteries, etc.

Mandates

- August 3, 2021, due to the Delta variant, NYC passed mandate requiring proof of vaccination to enter restaurants, concerts, stadiums and gyms
- Similar policies were adopted by other major cities soon after (see next table)
- I'll skip the prior literature for now

Timing

Table: Timing of Indoor Vaccine Mandates

City	Announced	Implemented	Repealed
NYC	8/3/21	8/16/21	3/7/22
San Francisco	8/12/21	8/20/21	3/11/22
New Orleans	8/12/21	8/16/21	3/21/22
Seattle	9/6/21	10/25/21	3/1/22
Los Angeles	11/8/21	11/29/21	3/30/22
Philadelphia	12/13/21	1/3/22	2/16/22
Boston	12/20/21	1/15/22	2/18/22
Chicago	12/21/21	1/3/22	2/28/22
DC	12/22/21	1/15/22	2/15/22

Research question

- Estimate an ATT for these cities' mandates on vaccination, cases and deaths
- Data will come from daily county level COVID-19 vaccinations, cases and deaths from the CDC aggregated to MSA by week scaled by US population estimates
- Main outcomes: Weekly measures of administered first doses of COVID-19 vaccines, cases, and deaths per 100,000 residents
- Weekly panel from December 21, 2020 to April 18, 2022 for 821 MSAs (they note various issues with data quality required dropping just under 100 MSAs) with 57,470 observations

Descriptive Statistics

Table: Descriptive Statistics

	All MSAs			Treated MSAs			Untreated MSAs		
Variable	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
First Doses per 100,000	817.47	1,344.30	458.98	1,253.50	1,237.18	827.71	812.66	1,344.65	455.01
Cases per 100,000	273.75	373.61	147.73	247.47	394.75	121.95	274.04	373.37	148.16
Deaths per 100,000	3.56	5.87	1.90	2.03	2.31	1.17	3.58	5.90	1.91
Number of observations		57,470			630			56,840	

Notes: The unit of observation is MSA week. Our sample consists of 821 MSAs, 9 of which are treated, and the period spans 70 weeks from December 21, 2020, to April 18, 2022.

Great discussion of synth DiD

"The basic idea is that the unit weights are chosen to find a convex combination of potential control states whose treatment trend in the outcome variable of interest is most parallel to that of the treated state. The inclusion of the intercept term ω_0 (made possible because of the inclusion of the unit fixed effects) is one way in which the SDID unit weights differ from those of the synthetic control weights. Instead of the weights needing to make the pre-trend control unit perfectly match that of the treated unit, as is the case with the synthetic control estimator, allowing for this intercept makes it sufficient for the weights to just make the trends parallel."

Table 3: Announcement of Indoor COVID-19 Vaccine Mandates and First-Dose Vaccine Uptake

	Dependent Variable : *Weekly First Doses per* 100,000		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
Panel A. **Boston**			
Average Effect ($\hat{\tau}$)	319.04	–140.04	72.69
95% Confidence Interval	(–1047.25, 1685.33)	(–1211.06, 930.99)	(–1160.96, 1306.33)
Panel B. **Chicago**			
Average Effect ($\hat{\tau}$)	–39.95	–28.4	–172.34
95% Confidence Interval	(–1197.23, 1117.34)	(–894.70, 837.91)	(–1188.18, 843.50)
Panel C. **Los Angeles**			
Average Effect ($\hat{\tau}$)	–143.09	–242.61	–185.31
95% Confidence Interval	(–1142.48, 856.31)	(–740.82, 255.59)	(–966.80, 596.19)
Panel D. **New Orleans**			
Average Effect ($\hat{\tau}$)	–341.38	–219.38	–209.07
95% Confidence Interval	(–1642.73, 959.97)	(–724.08, 285.32)	(–721.84, 303.70)
Panel E. **New York**			
Average Effect ($\hat{\tau}$)	–575.97	123.77	–82.59
95% Confidence Interval	(–1907.72, 755.79)	(–398.14, 645.68)	(–605.48, 440.30)
Panel F. **Philadelphia**			
Average Effect ($\hat{\tau}$)	104.16	–295.41	–303.02
95% Confidence Interval	(–1148.25, 1356.57)	(–1252.58, 661.76)	(–1401.35, 795.31)
Panel G. **San Francisco**			
Average Effect ($\hat{\tau}$)	−1197.67*	–42.89	–195.37
95% Confidence Interval	(–2504.92, 109.58)	(–566.19, 480.41)	(–726.44, 335.71)
Panel H. **Seattle**			
Average Effect ($\hat{\tau}$)	–736.58	–97.14	–207.02
95% Confidence Interval	(–1978.53, 505.38)	(–688.32, 494.03)	(–840.35, 426.32)
Panel I. **Washington DC**			
Average Effect ($\hat{\tau}$)	–253.99	18.77	–76.53
95% Confidence Interval	(–1620.28, 1112.31)	(–1059.12, 1096.67)	(–1309.86, 1156.80)

Notes: This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on first-dose vaccine uptake as measured by weekly first doses per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ($\hat{\tau}$ from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as *** p<0.01, ** p<0.05, and * p<0.1.

Table 4: Announcement of Indoor COVID-19 Vaccine Mandates and COVID-19 Cases

	Dependent Variable : *Weekly COVID*-19 Cases per 100,000		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
Panel A. **Boston**			
Average Effect ($\hat{\tau}$)	274.32	240.05	224.57
95% Confidence Interval	(–252.03, 800.67)	(–272.99, 753.09)	(–267.13, 716.27)
Panel B. **Chicago**			
Average Effect ($\hat{\tau}$)	139.6	184.48	121.14
95% Confidence Interval	(–299.65, 578.84)	(–245.53, 614.49)	(–289.63, 531.91)
Panel C. **Los Angeles**			
Average Effect ($\hat{\tau}$)	202.06	340.28***	176.49
95% Confidence Interval	(–74.98, 479.09)	(97.34, 583.22)	(–58.08, 411.05)
Panel D. **New Orleans**			
Average Effect ($\hat{\tau}$)	–22.81	6.15	–27.28
95% Confidence Interval	(–216.80, 171.19)	(–182.99, 195.28)	(–217.93, 163.36)
Panel E. **New York**			
Average Effect ($\hat{\tau}$)	–53.04	7.02	4.62
95% Confidence Interval	(–251.54, 145.46)	(–186.94, 200.98)	(–190.87, 200.12)
Panel F. **Philadelphia**			
Average Effect ($\hat{\tau}$)	110.41	290.62	114.41
95% Confidence Interval	(–368.76, 589.58)	(–180.84, 762.08)	(–329.83, 558.66)
Panel G. **San Francisco**			
Average Effect ($\hat{\tau}$)	–107.37	65.71	–95.48
95% Confidence Interval	(–311.98, 97.24)	(–135.80, 267.22)	(–297.46, 106.50)
Panel H. **Seattle**			
Average Effect ($\hat{\tau}$)	19.85	20.72	–16.99
95% Confidence Interval	(–239.41, 279.12)	(–202.52, 243.95)	(–247.34, 213.36)
Panel I. **Washington DC**			
Average Effect ($\hat{\tau}$)	149.7	600.71	190.26
95% Confidence Interval	(–376.66, 676.05)	(86.71, 1114.72)	(–301.70, 682.22)

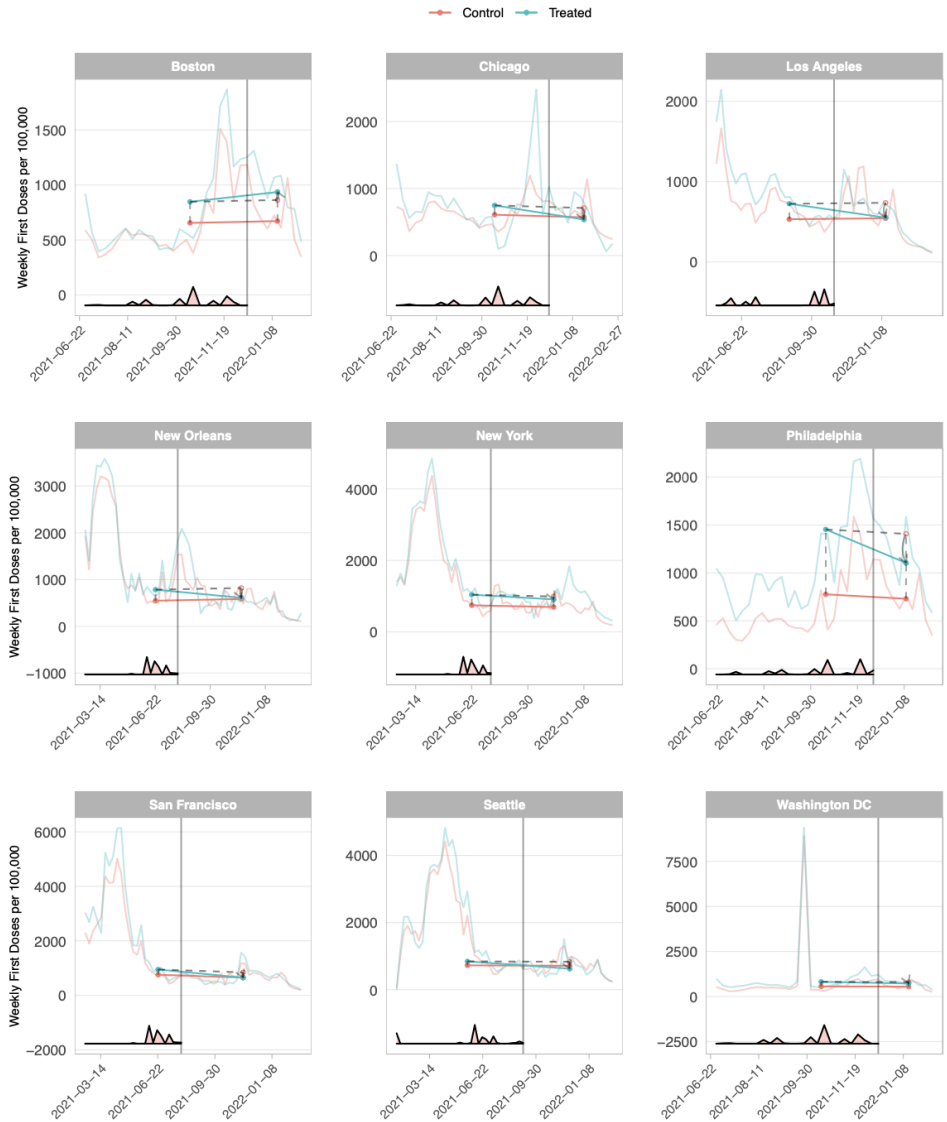
Notes: This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 cases per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ($\hat{\tau}$ from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as *** $p<0.01$, ** $p<0.05$, and * $p<0.1$.

Table 5: Announcement of Indoor COVID-19 Vaccine Mandates and COVID-19 Deaths

	Dependent Variable : *Weekly COVID*-19 Deaths per 100,000		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
Panel A. **Boston**			
Average Effect ($\hat{\tau}$)	2.32	1.65	1.38
95% Confidence Interval	(–4.75, 9.39)	(–5.97, 9.28)	(–4.76, 7.53)
Panel B. **Chicago**			
Average Effect ($\hat{\tau}$)	1.94	1.46	1.39
95% Confidence Interval	(–4.21, 8.09)	(–5.10, 8.03)	(–4.06, 6.84)
Panel C. **Los Angeles**			
Average Effect ($\hat{\tau}$)	–0.2	0.67	–0.3
95% Confidence Interval	(–5.26, 4.86)	(–3.85, 5.19)	(–4.52, 3.92)
Panel D. **New Orleans**			
Average Effect ($\hat{\tau}$)	–0.65	–2.5	–1.37
95% Confidence Interval	(–4.48, 3.18)	(–6.07, 1.07)	(–4.96, 2.22)
Panel E. **New York**			
Average Effect ($\hat{\tau}$)	–2.37	–2.66	–1.91
95% Confidence Interval	(–6.16, 1.43)	(–6.09, 0.76)	(–5.42, 1.60)
Panel F. **Philadelphia**			
Average Effect ($\hat{\tau}$)	2.76	–2.16	2.21
95% Confidence Interval	(–4.11, 9.63)	(–9.35, 5.02)	(–3.69, 8.11)
Panel G. **San Francisco**			
Average Effect ($\hat{\tau}$)	–2.19	–4.72**	–2.66
95% Confidence Interval	(–6.14, 1.76)	(–8.51, –0.93)	(–6.36, 1.04)
Panel H. **Seattle**			
Average Effect ($\hat{\tau}$)	–1.07	-1.08	–1.43
95% Confidence Interval	(–5.04, 2.91)	(–4.63, 2.47)	(–5.09, 2.22)
Panel I. **Washington DC**			
Average Effect ($\hat{\tau}$)	0.46	–0.92	0.2
95% Confidence Interval	(–6.61, 7.53)	(–8.55, 6.70)	(–5.95, 6.35)

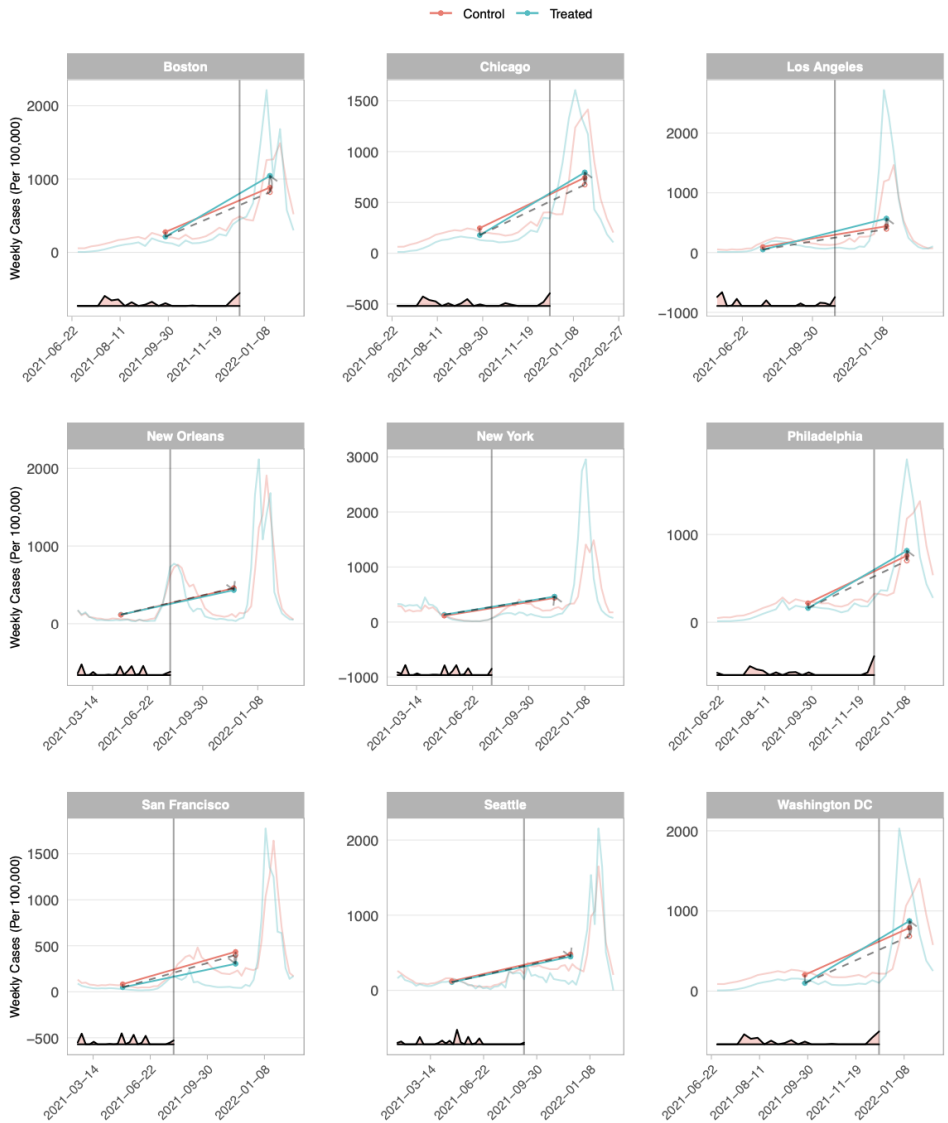
Notes: This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 deaths per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ($\hat{\tau}$ from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as *** p<0.01, ** p<0.05, and * p<0.1.

Figure 1: Trends in Weekly First Doses per 100,000 in Treated MSAs and Their Respective Synthetic Controls



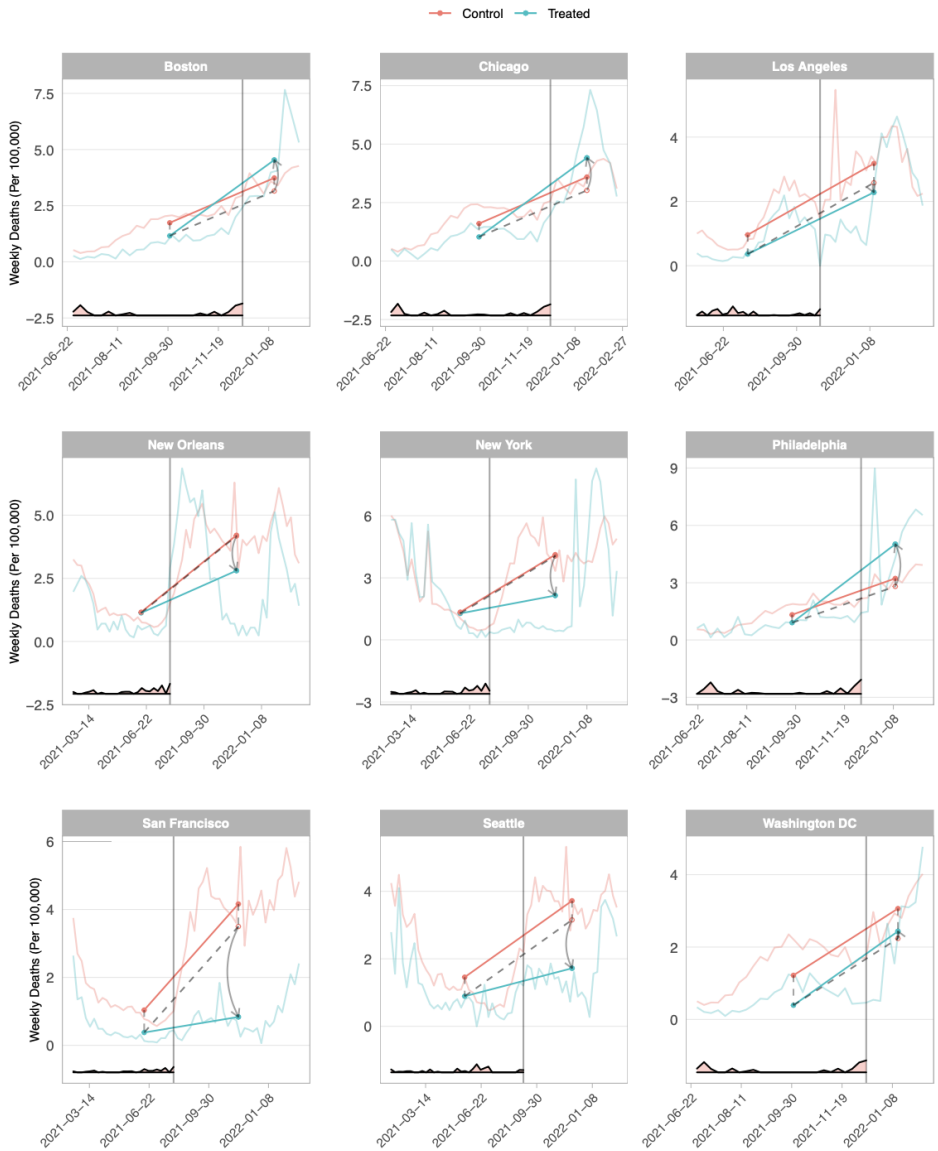
Notes: Each plot shows trends in weekly first doses of COVID-19 vaccinations per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ($\hat{\tau}$ from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

Figure 2: Trends in Weekly COVID-19 Cases per 100,000 in Treated MSAs and Their Respective Synthetic Controls



Notes: Each plot shows trends in weekly COVID-19 cases per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ($\hat{\tau}$ from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

Figure 3: Trends in Weekly COVID-19 Deaths per 100,000 in Treated MSAs and Their Respective Synthetic Controls



Notes: Each plot shows trends in weekly COVID-19 deaths per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ($\hat{\tau}$ from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

Conclusion

- They also report synth and DiD analysis as robustness – something to keep in mind is the presentation of results are subjective
- Rather than showing regression results with more controls, we tend to now see different DiD and synth estimators as the robustness
- Authors fail to find strong evidence the vaccine mandates slowed COVID-19
- What's your response?