

Causal Inference II

MIXTAPE SESSION



Roadmap

TWFE Pathologies

- Brief history

- Bacon decomposition

- Simulation

Two solutions and a new decomposition

- CS

- SA

- Some opinions and an application

Synthetic control

- Abadie's non-negative weighting

- Ben-Michael, et al's Least Negative Weighting Method

- Athey, et al's Matrix Completion with Nuclear Norm Method

- Synthetic difference-in-differences

Material we hope to cover

- Differential timing and diff-in-diff
 - 1. Bacon's decomposition
 - 2. Callaway and Sant'Anna
 - 3. Sun and Abraham
- Non-negative weighting in the synthetic control
- Least negative weighting in the synthetic control
- Matrix completion and synthetic diff-in-diff

Why a day of differential timing?

- Diff-in-diff, as we said yesterday, is the most common method
- And two-way fixed effects (TWFE) panel estimators are the most common estimators
- Recent work has shown that this estimator's most common specification required constant treatment effects otherwise it was biased
- New methods have been developed in its place

Differential timing outline

We will cover some of the properties of two way fixed effects (TWFE), some solutions and my personal opinions

1. Brief review of potential outcomes and the ATT
2. Difference-in-differences equation ("four averages and three differences") and the parallel trends assumption
3. TWFE Pathologies in static specification
 - Goodman-Bacon decomposition as diagnosis of the problem
 - Callaway and Sant'Anna estimator as a cure
4. TWFE Pathologies in event study specification
 - Sun and Abraham as both a diagnosis and a cure
5. Application, practical advice and code

Beaver dam and diff-in-diff credibility crisis

- Differential timing literature is like a stick that struck a beaver's dam
- Stick made a hole causing a leak
- Gradually that hole got larger and the leak got bigger
- Eventually the dam collapsed
- That's now



Difference-in-differences credibility crisis

- I'll start with circa 2016 onward – several grad students and assistant professors found critical pathologies with TWFE and developed solutions
- Many simultaneous discoveries, some redundancies, and **sudden** awareness of the issues started happening around 2017, eventually became a massive thing
- Extreme meteoric rise, unusual for econometrics

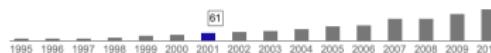
Compare with LATE paper

- Compare with Imbens and Angrist 1995 LATE in *Econometrica*
- 61 annual cites the year Imbens is denied tenure at Harvard for what would later win him a Nobel Prize

Identification and estimation of local average treatment effects

Authors Guido W Imbens, Joshua D Angrist
Publication date 1994/3/1
Journal *Econometrica: journal of the Econometric Society*
Pages 467-475
Publisher Econometric Society
Description RANDOM ASSIGNMENT OF TREATMENT and concurrent data collection on treatment and control groups is the norm in medical evaluation research. In contrast, the use of random assignment to evaluate social programs remains controversial. Following criticism of parametric evaluation models (eg, Lalonde (1986)), econometric research has been geared towards establishing conditions that guarantee nonparametric identification of treatment effects in observational studies, ie identification without relying on functional form restrictions or distributional assumptions. The focus has been on identification of average treatment effects in a population of interest, or on the average effect for the subpopulation that is treated. The conditions required to nonparametrically identify these parameters can be restrictive, however, and the derived identification results fragile. In particular, results in Chamberlain (1986), Manski (1990 ...)

Total citations [Cited by 5586](#)



Compare with synth paper

- Athey and Imbens called synth the most important innovation in causal inference of the last two decades
- Most econometrics papers, even influential ones, show slow growth
- Something was different about diff-in-diff even before the econometricians recently shifted their attention to it

[The economic costs of conflict: A case study of the Basque Country](#)

Authors Alberto Abadie, Javier Gardeazabal

Publication date 2003/3/1

Journal American economic review

Volume 93

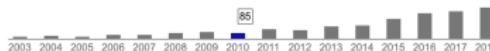
Issue 1

Pages 113-132

Publisher American Economic Association

Description This article investigates the economic effects of conflict, using the terrorist conflict in the Basque Country as a case study. We find that, after the outbreak of terrorism in the late 1960's, per capita GDP in the Basque Country declined about 10 percentage points relative to a synthetic control region without terrorism. In addition, we use the 1998-1999 truce as a natural experiment. We find that stocks of firms with a significant part of their business in the Basque Country showed a positive relative performance when truce became credible, and a negative relative performance at the end of the cease-fire.

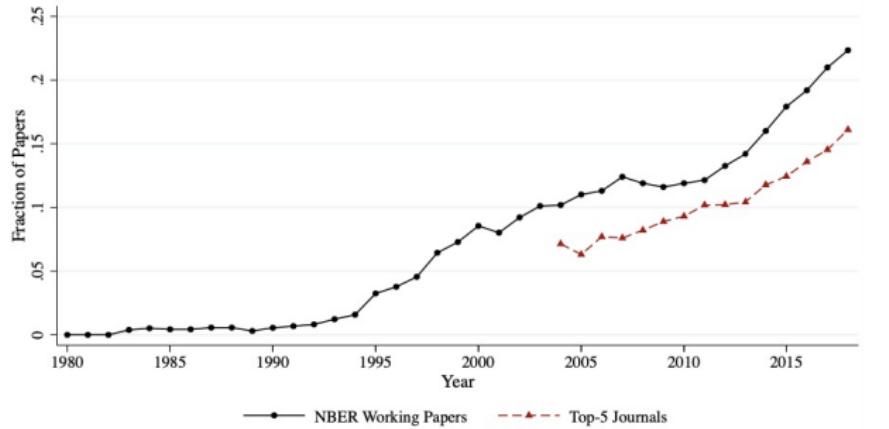
Total citations [Cited by 5368](#)



Diff-in-diff had belonged to the empiricists

Figure: Currie, et al. (2020)

A: Difference-in-Differences



With some exception (e.g., Heckman, Ichimura and Todd 1997; Abadie 2005; Bertrand, Duflo and Mullainathan 2004), econometricians had not given it much notice

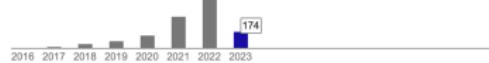
Borusyak et al

- Starts it all; written as grad students at Harvard
- Goes through many revisions, posted as working paper
- Returned to a few years ago with a third coauthor, Jahn Spiess, now R&R at Restud

Revisiting event study designs: Robust and efficient estimation

Authors Kirill Borusyak, Xavier Jaravel, Jann Spiess
Publication date 2021/8/27
Journal arXiv preprint arXiv:2108.12419
Description We develop a framework for difference-in-differences designs with staggered treatment adoption and heterogeneous causal effects. We show that conventional regression-based estimators fail to provide unbiased estimates of relevant estimands absent strong restrictions on treatment-effect homogeneity. We then derive the efficient estimator addressing this challenge, which takes an intuitive "imputation" form when treatment-effect heterogeneity is unrestricted. We characterize the asymptotic behavior of the estimator, propose tools for inference, and develop tests for identifying assumptions. Extensions include time-varying controls, triple-differences, and certain non-binary treatments. We show the practical relevance of these insights in a simulation study and an application. Studying the consumption response to tax rebates in the United States, we find that the notional marginal propensity to consume is between 8 and 11 percent in the first quarter—about half as large as benchmark estimates used to calibrate macroeconomic models—and predominantly occurs in the first month after the rebate.

Total citations Cited by 1399



"dCdH"

- First major hit (in AER), may have been in working paper in 2017 (at least 2018)
- Very thorough decomposition of the TWFE pathology, very general solution, included Stata code
- Very active and talented young team (assistant profs when this was done)

Two-way fixed effects estimators with heterogeneous treatment effects

Authors Clément De Chaisemartin, Xavier d'Haultfoeuille

Publication date 2020/9/1

Journal American Economic Review

Volume 110

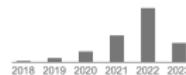
Issue 9

Pages 2964-2996

Publisher American Economic Association

Description Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they estimate weighted sums of the average treatment effects (ATE) in each group and period, with weights that may be negative. Due to the negative weights, the linear regression coefficient may for instance be negative while all the ATEs are positive. We propose another estimator that solves this issue. In the two applications we revisit, it is significantly different from the linear regression estimator. (JEL C21, C23, D72, J31, J51, L82)

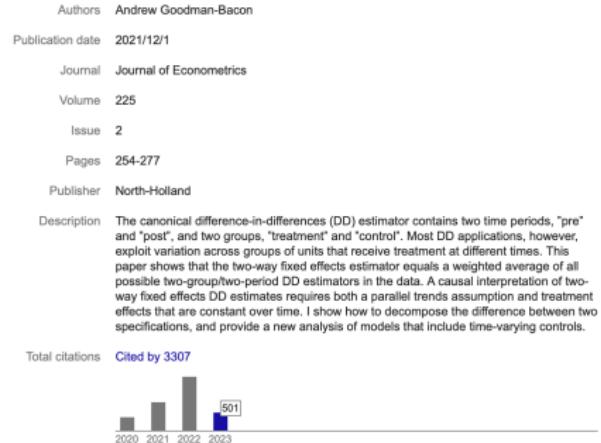
Total citations Cited by 2019



Goodman-Bacon

- Arguably the most influential in terms of bringing attention to the problem (but no solution)
- Begun while grad student at Michigan, published last of the crop
- Probably Twitter network had a role as he was very active, also not an econometrician

Difference-in-differences with variation in treatment timing



"CS"

- Second published solution to the problem, written while assistant professors at Vanderbilt and Ole Miss,
- Pedro is a UC3M alum (2015 grad) and Brantly is a Vanderbilt grad
- Both are now coauthors with Andrew Goodman-Bacon
- Introduced new terms like group-time ATT, released very tight R code ("did")

Difference-in-differences with multiple time periods

Authors Brantly Callaway, Pedro HC Sant'Anna

Publication date 2021/12/1

Journal Journal of Econometrics

Volume 225

Issue 2

Pages 200-230

Publisher North-Holland

Description In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the ...

Total citations Cited by 2378



“SA”

- Third published solution to the problem, very similar to CS
- Focus was on decomposing the event study
- Written while grad students at MIT but Sophie Sun is now an assistant professor at CEMFI!

Estimating dynamic treatment effects in event studies with heterogeneous treatment effects

Authors Liyang Sun, Sarah Abraham

Publication date 2021/12/1

Journal Journal of Econometrics

Volume 225

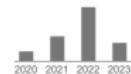
Issue 2

Pages 175-199

Publisher North-Holland

Description To estimate the dynamic effects of an absorbing treatment, researchers often use two-way fixed effects regressions that include leads and lags of the treatment. We show that in settings with variation in treatment timing across units, the coefficient on a given lead or lag can be contaminated by effects from other periods, and apparent pretrends can arise solely from treatment effects heterogeneity. We propose an alternative estimator that is free of contamination, and illustrate the relative shortcomings of two-way fixed effects regressions with leads and lags through an empirical application.

Total citations Cited by 1828



There's even more and more coming

- Gardner, Wooldridge, John Roth, and on and on
- Too many people to name at this point
- Given the large cites, we are likely to keep seeing more on this
- Probably shifting applied practice for the better but there are some growing pains

Two-way fixed effects

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

Two OLS Models

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist} \quad (1)$$

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist} \quad (2)$$

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

First equation works; second one only sometimes works

Discussion of estimate

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

- So that's the simple case; what about the differential timing case?
- If you estimate with OLS with differential timing, what does $\hat{\delta}$ correspond to?
- It also corresponds to the previous “four averages and three subtractions” – but it’s numerous of them, not just one

Decomposition Preview

- Andrew Goodman-Bacon decomposed $\hat{\delta}$ and showed it is numerically identical to a weighted average of all “four averages and three subtractions”
- But, even before we get to causality there are unusual features
- TWFE model assigns its own weights which are a function of the size of a “group” and the variance of group treatment dummies

K^2 distinct DDs

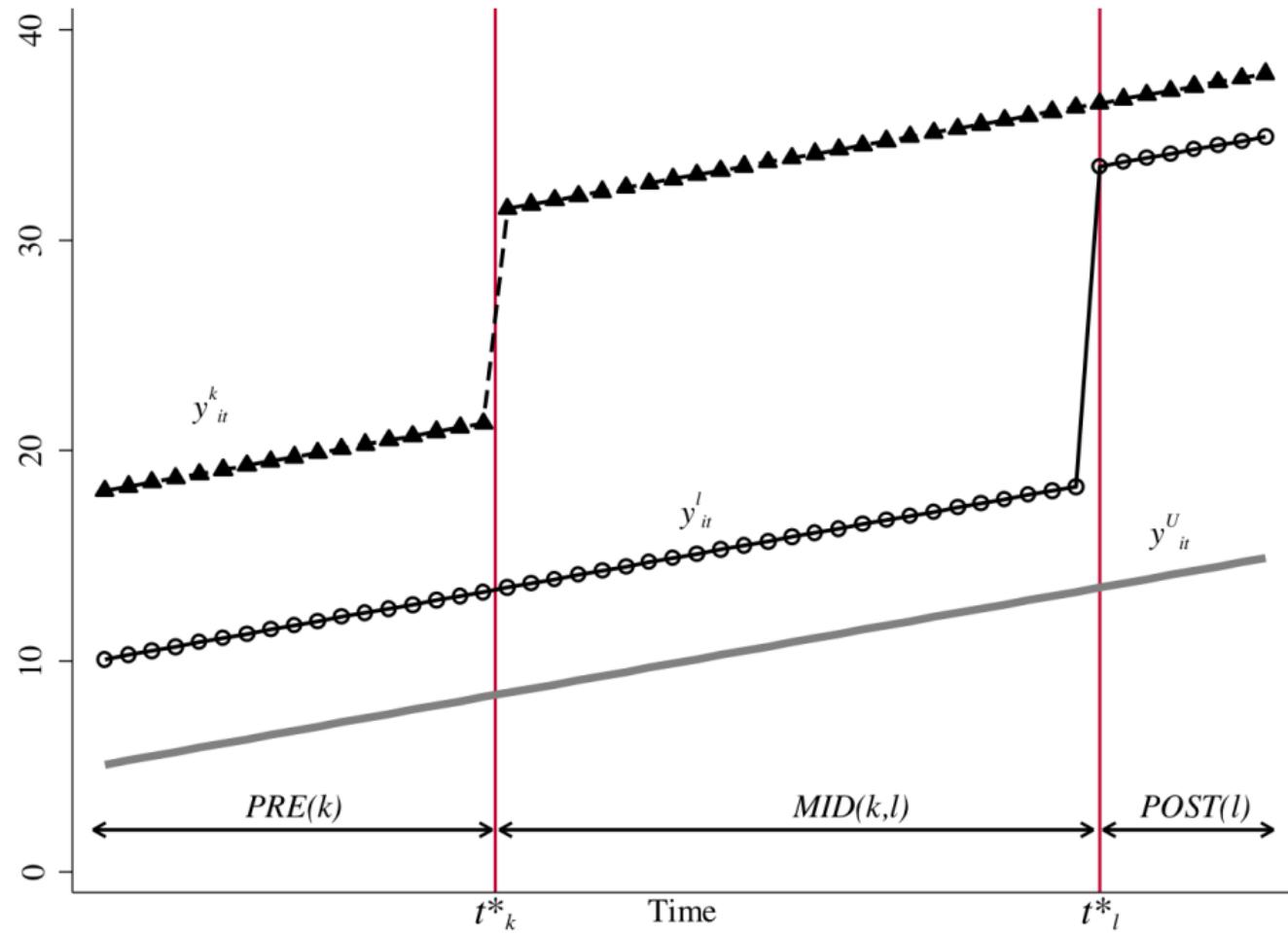
Let's look at 3 timing groups (a, b and c) and one untreated group (U).
With 3 timing groups, there are 9 2x2 DDs. Here they are:

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

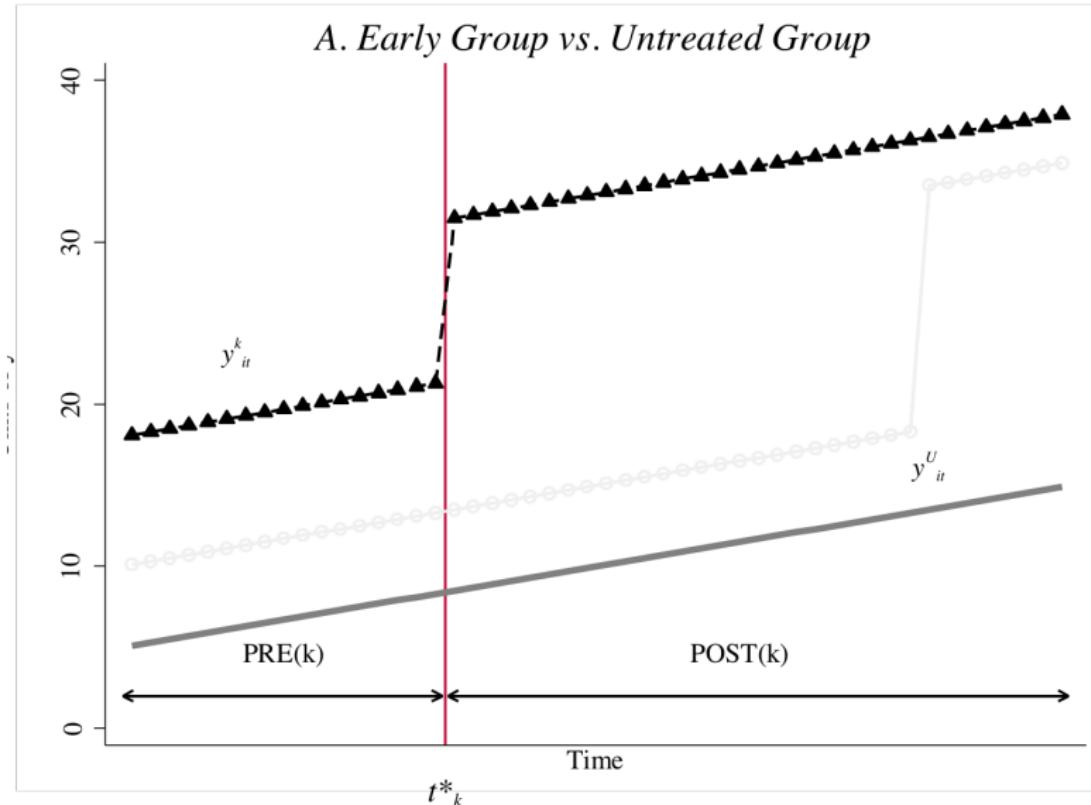
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

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

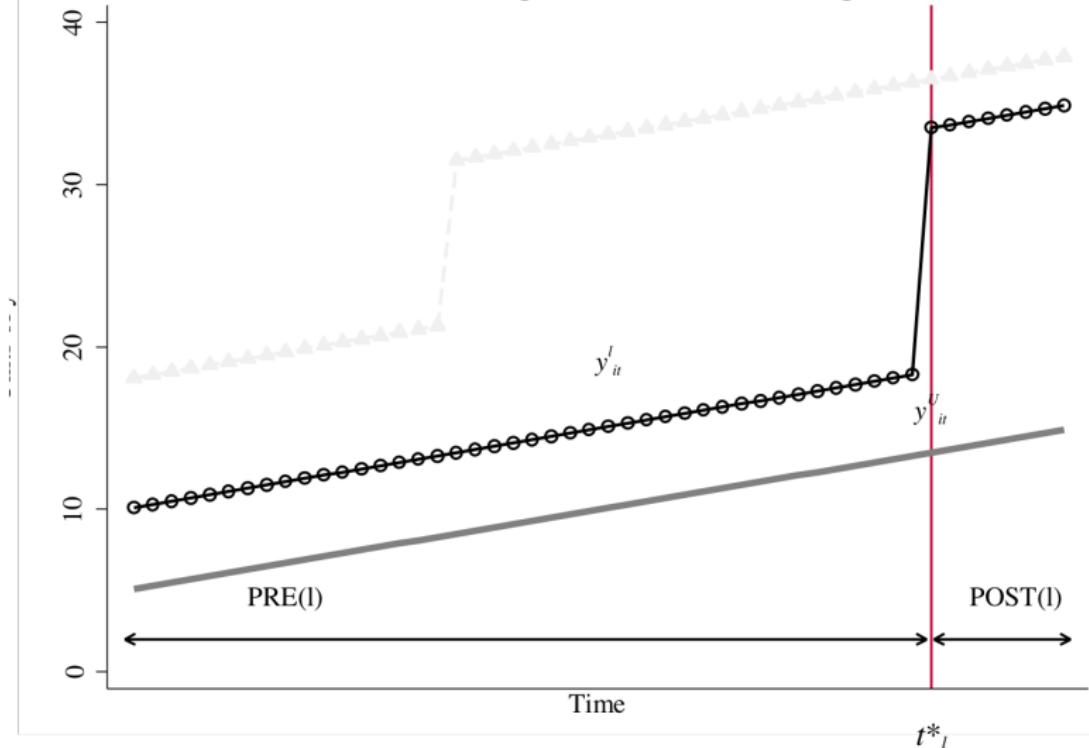


$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$

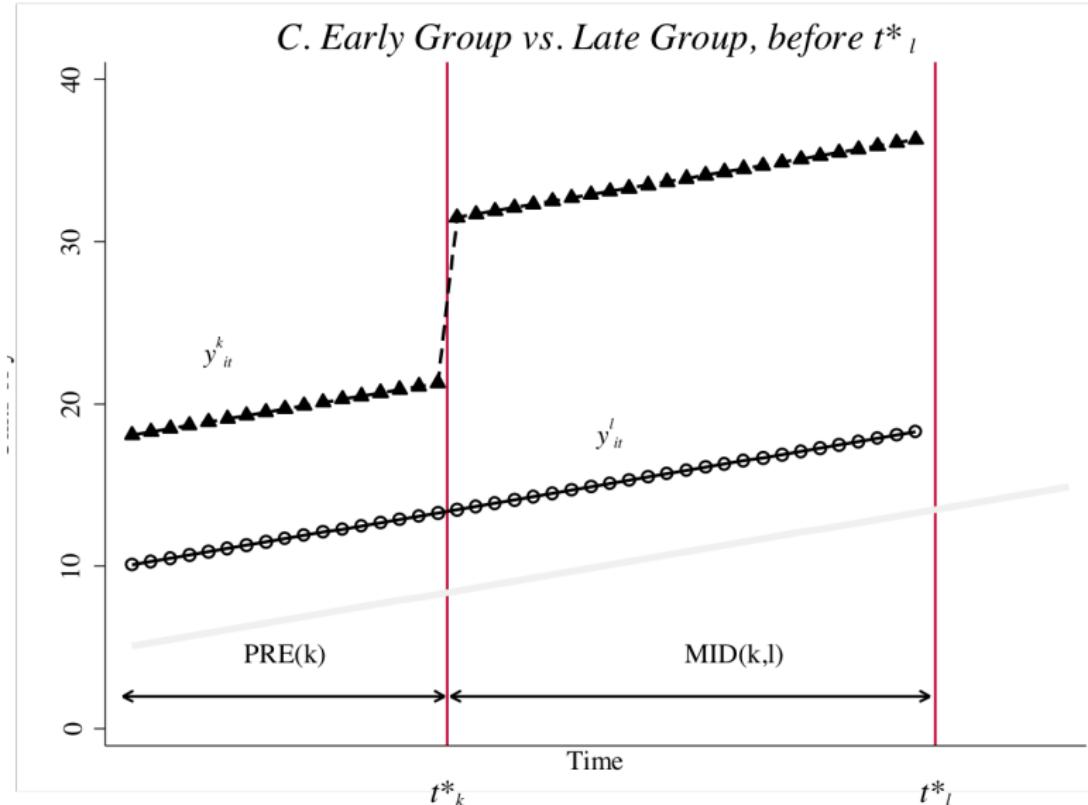


$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_l^{post(l)} - \overline{y}_l^{pre(l)} \right) - \left(\overline{y}_U^{post(l)} - \overline{y}_U^{pre(l)} \right)$$

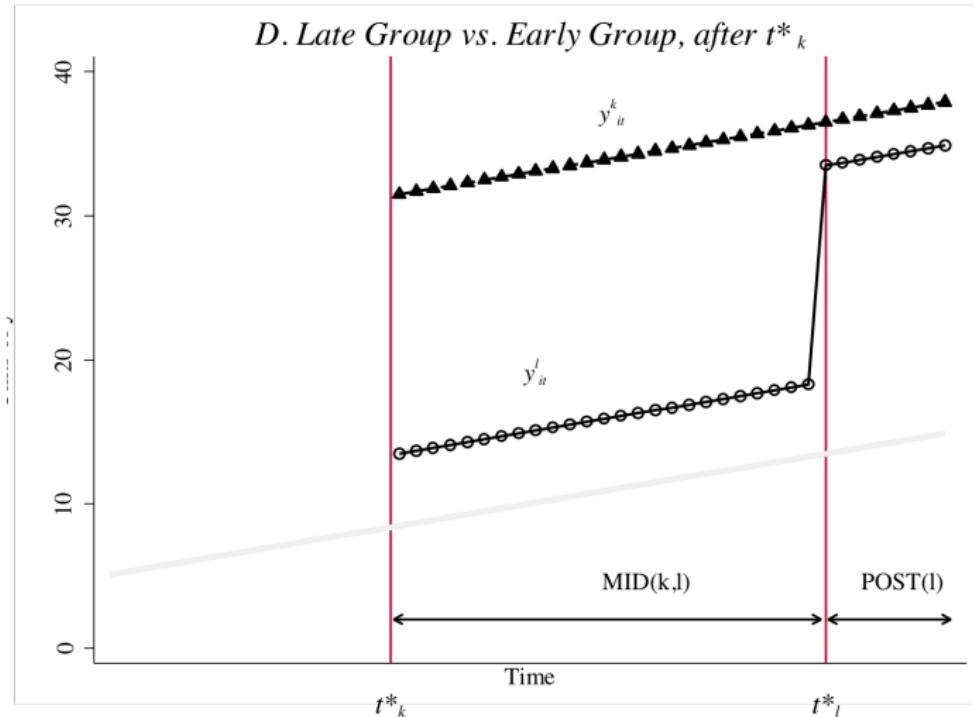
B. Late Group vs. Untreated Group



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



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



Bacon decomposition

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

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

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

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

Third, the Weights

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

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

Weights discussion

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

More weights discussion

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

Summarizing TWFE centralities

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

Back to TWFE

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

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

Moving from 2x2s to causal effects and bias terms

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

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

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

The dangerous 2x2

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

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

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

Notice all those potential sources of biases!

Potential Outcome Notation

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

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

Simulated data

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

Constant vs Dynamic Treatment Effects

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

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

Group-time ATT

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

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

Estimation

Estimate the following equation using OLS:

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

Table: Estimating ATT with different models

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

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

Bacon decomposition

Table: Bacon Decomposition (TWFE = -6.69)

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

T = Treatment; C= Comparison

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

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

Roadmap

TWFE Pathologies

- Brief history

- Bacon decomposition

- Simulation

Two solutions and a new decomposition

- CS

- SA

- Some opinions and an application

Synthetic control

- Abadie's non-negative weighting

- Ben-Michael, et al's Least Negative Weighting Method

- Athey, et al's Matrix Completion with Nuclear Norm Method

- Synthetic difference-in-differences

Callaway and Sant'Anna 2020

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

Difference-in-differences with multiple time periods

Authors	Brantly Callaway, Pedro HC Sant'Anna
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	200-230
Publisher	North-Holland
Description	In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the ...



When is CS used

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

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

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

Group-time ATT

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

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

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

CS identifies all feasible ATT(g,t)

Group-time ATT

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

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

Notation

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

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

Assumptions

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

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

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

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

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

CS Estimator (the IPW version)

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

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

Aggregated vs single year/group ATT

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

Group-time ATT

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

Question: Why didn't CS estimate all $\text{ATT}(g,t)$? What is "feasible ATT"?

Reporting results

Table: Estimating ATT using only pre-2004 data

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
<i>Feasible ATT</i>	68.33	26.81 ***	68.34***		

TWFE is no longer negative, interestingly, once we eliminate the last group (giving us a never-treated group), but is still suffering from attenuation bias.

Event study and differential timing

- Sometimes we care about a simple summary, and sometimes we care about separating it out in time and sometimes in even more interesting ways
- Event studies with one treatment group and one untreated group were relatively straightforward
- Interact treatment group with calendar date to get a series of leads and lags
- But when there are more than one treatment group, specification challenges emerge

Differential timing complicates plotting sample averages

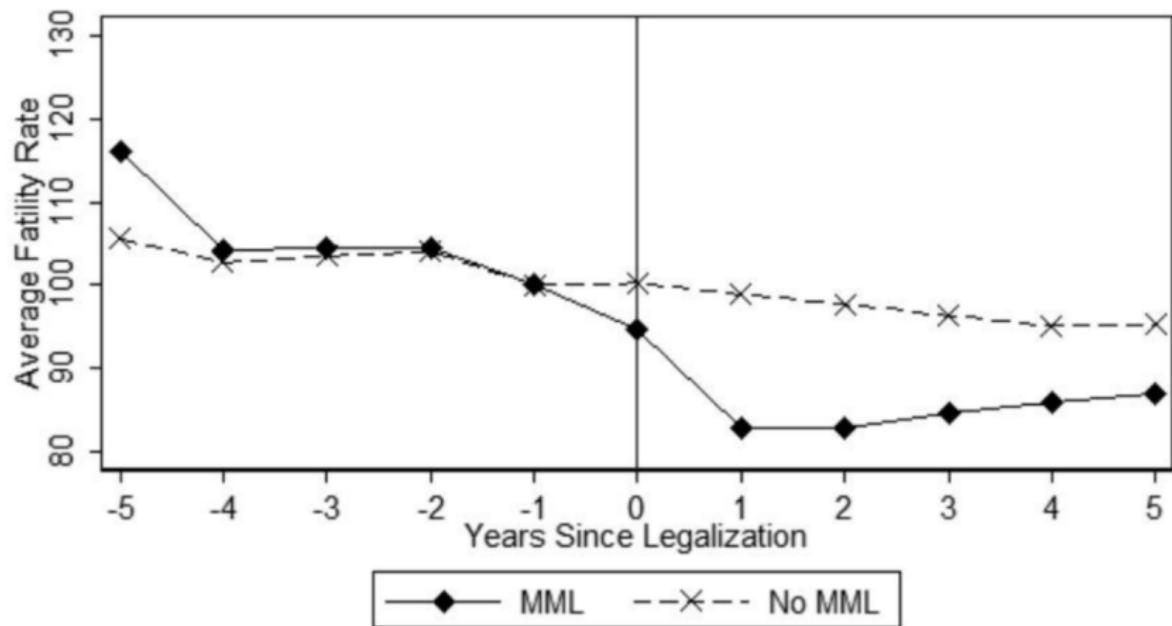


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Replicated from a project of mine

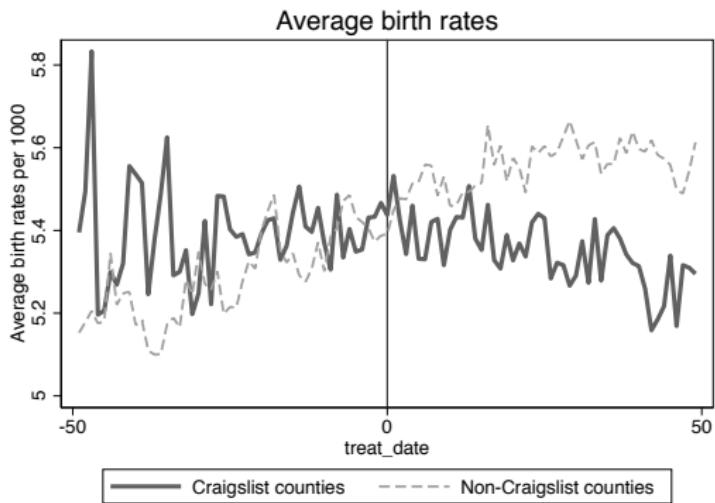
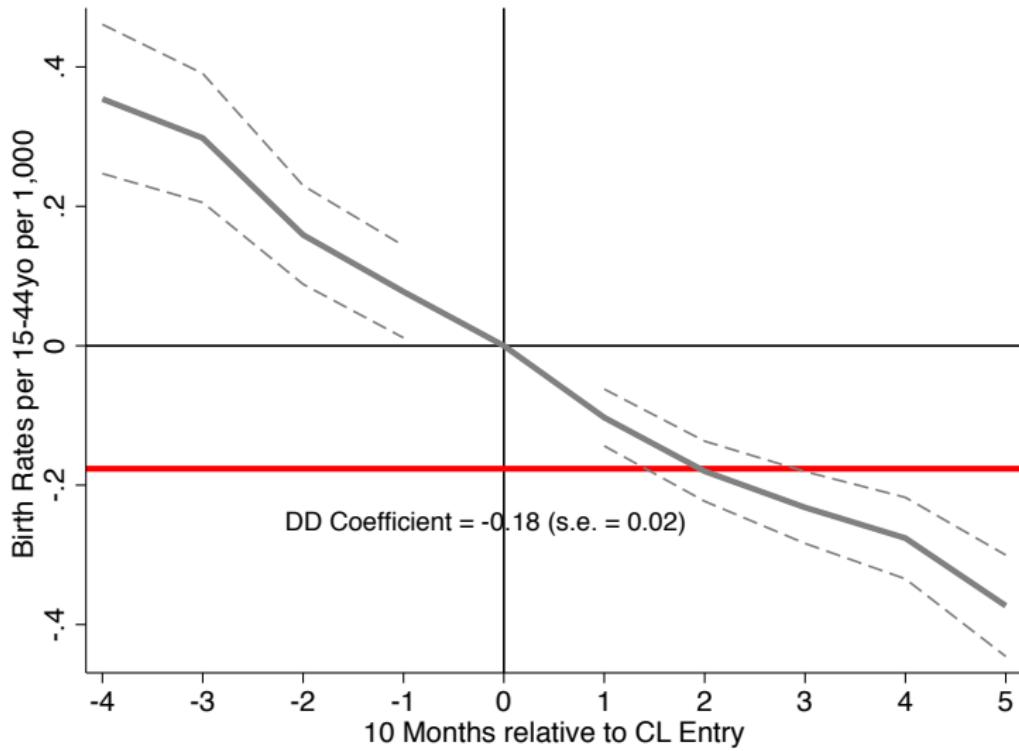


Figure: Roll out of Craigslist “personal ads” for casual intimate encounters and birth rates using the “randomized treatment assignment” approach for visualization

Event study specification with TWFE

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g \mathbf{1}\{t - E_i \in g\} + \varepsilon_{i,t}$$

Coefficient μ_g on a dummy measuring the number of years prior to or after that unit was treated.



Same data as a couple slides ago, leads don't look good, so I abandoned the project.

Bias of TWFE Event Study Specification

- Bacon only focused on the static specification, and that's where the biases due to dynamics revealed itself
- He was unable to get into the leads and lags using the FWL method he was using ("it's hard!" - Bacon)
- Sophie Sun and Sarah Abraham did though – prompted by a stray comment by their professor
- But they also unlike Bacon present a solution (which is like CS, but discovered independently)

1. SA shows a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
2. They show that the population regression coefficient is “contaminated” by information from other leads and lags (which is then later generalized by Goldsmith-Pinkham, Hull and Kolsar 2022)
3. SA presents an alternative estimator that is a version of CS only using the “last cohort” as the treatment group (not the not-yet-treated)
4. Derives the variance of the estimator instead of bootstrapping, handles covariates differently than CS, but otherwise identical

Summarizing (cont.)

- Under homogenous treatment profiles, weights sum to zero and “cancel out” the treatment effects from other periods
- Under treatment effect heterogeneity, they do not cancel out and leads and lags are biased
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Some notation and terms

- As people often **bin** the data, we allow a lead or lag l to appear in bin g so sometimes they use g instead of l or $l \in g$
- Building block is the “cohort-specific ATT” or $CATT_{e,l}$ – same as $ATT(g,t)$
- Our goal is to estimate $CATT_{e,l}$ with population regression coefficient μ_l
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- The ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome: $Y_{i,t} - Y_{i,t}^{\infty}$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as $l = \text{year} - \text{treatment date}$

Relative vs calendar event time

```
. list state-treat time_til in 1/10
```

	state	firms	year	n	id	group	treat_~e	treat	time_til
1.	1	.3257218	1980	1	1	1	1986	0	-6
2.	1	.3257218	1981	2	1	1	1986	0	-5
3.	1	.3257218	1982	3	1	1	1986	0	-4
4.	1	.3257218	1983	4	1	1	1986	0	-3
5.	1	.3257218	1984	5	1	1	1986	0	-2
6.	1	.3257218	1985	6	1	1	1986	0	-1
7.	1	.3257218	1986	7	1	1	1986	1	0
8.	1	.3257218	1987	8	1	1	1986	1	1
9.	1	.3257218	1988	9	1	1	1986	1	2
10.	1	.3257218	1989	10	1	1	1986	1	3

Definition 1

Definition 1: The cohort-specific ATT l periods from initial treatment date e is:

$$CATT_{e,l} = E[Y_{i,e+l} - Y_{i,e+l}^{\infty} | E_i = e]$$

Fill out the second part of the Group-time ATT exercise together.

TWFE assumptions

- For consistent estimates of the coefficient leads and lags using TWFE model, we need three assumptions
- For SA and CS, we only need two
- Let's look then at the three

Assumption 1: Parallel trends

Assumption 1: Parallel trends in baseline outcomes:

$E[Y_{i,t}^\infty - Y_{i,s}^\infty | E_i = e]$ is the same for all $e \in supp(E_i)$ and for all s, t and is equal to $E[Y_{i,t}^\infty - Y_{i,s}^\infty]$

Lead and lag coefficients are DiD equations but once we invoke parallel trends they can become causal parameters. This reminds us again how crucial it is to have appropriate controls

Assumption 2: No anticipation

Assumption 2: No anticipator behavior in pre-treatment periods:

There is a set of pre-treatment periods such that

$$E[Y_{i,e+l}^e - Y_{i,e+l}^\infty | E_i = e] = 0 \text{ for all possible leads.}$$

Essentially means that pre-treatment, the causal effect is zero. Most plausible if no one sees the treatment coming, but even if they see it coming, they may not be able to make adjustments that affect outcomes

Assumption 3: Homogeneity

Assumption 3: Treatment effect profile homogeneity: For each relative time period l , the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_l$.

Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn't rule out parallel trends

Event study model

Dynamic TWFE model

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g \mathbf{1}\{t - E_i \in g\} + \varepsilon_{i,t}$$

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

Interpreting $\widehat{\mu}_g$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$\begin{aligned} \mu_g &= \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Targets}} \\ &+ \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Contamination from other leads and lags}} \\ &+ \underbrace{\sum_{l \in g^{excl}} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Contamination from dropped periods}} \end{aligned}$$

Weight ($w_{e,l}^g$) summation cheat sheet

1. For relative periods of μ_g own $l \in g$, $\sum_{l \in g} \sum_e w_{e,l}^g = 1$
2. For relative periods belonging to some other bin $l \in g'$ and $g' \neq g$,
 $\sum_{l \in g'} \sum_e w_{e,l}^g = 0$
3. For relative periods not included in G , $\sum_{l \in g^{excl}} \sum_e w_{e,l}^g = -1$

Estimating the weights

Regress $D_{i,t}^l \times 1\{E_i = e\}$ on:

1. all bin indicators included in the main TWFE regression,
2. $\{1\{t - E_i \in g\}\}_{g \in G}$ (i.e., leads and lags) and
3. the unit and time fixed effects

Still biased under parallel trends

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bin g is a linear combination of $CATT_{e,l \in g}$ as well as $CATT_{d,l'}$ from other relative periods $l' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g = & \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g CATT_{e,l}}_{\text{Desirable}} \\ & + \underbrace{\sum_{g' \neq g, g' \in G} \sum_{l' \in g'} \sum_e w_{e,l'}^g CATT_{e,l'}}_{\text{Bias from other specified bins}} \\ & + \underbrace{\sum_{l' \in g^{excl}} \sum_e w_{e,l'}^g CATT_{e,l'}}_{\text{Bias from dropped relative time indicators}}\end{aligned}$$

Still biased under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$\begin{aligned}\mu_g = & \sum_{l' \in g, l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{l' \in g^{excl}, l' \geq 0} \sum_e w_{w,l'}^g CATT_{e,l'}\end{aligned}$$

Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods even *though parallel trends hold in the pre period.*

Proposition 4

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,l} = ATT_l$ is constant across e for a given l , and the population regression coefficient μ_g is equal to a linear combination of $ATT_{l \in g}$, as well as $ATT_{l' \notin g}$ from other relative periods

$$\begin{aligned}\mu_g &= \sum_{l \in g} w_l^g ATT_l \\ &+ \sum_{g' \neq g} \sum_{l' \in g'} w_{l'}^g ATT_{l'} \\ &+ \sum_{l' \in g^{excl}} w_{l'}^g ATT_{l'}\end{aligned}$$

Simple example

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. For illustrative purposes, we will include bins $\{-2, 0\}$ in our calculations but drop $\{-1, 1\}$.

Simple example

$$\begin{aligned}\mu_{-2} = & \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}} \\ & + \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}\end{aligned}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

Robust event study estimation

- All the robust estimators under differential timing have solutions and they all skip over forbidden contrasts.
- Sun and Abraham (2020) propose a 3-step interacted weighted estimator (IW) using last treated group as control group
- Callaway and Sant'anna (2020) estimate group-time ATT which can be a weighted average over relative time periods too but uses "not-yet-treated" as control

Interaction-weighted estimator

- **Step one:** Do this DD regression and hold on to $\widehat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated. The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

Interaction-weighted estimator

- **Step two:** Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

Interaction-weighted estimator

- **Step three:** Take a weighted average of estimates for $CATT_{e,l}$ from Step 1 with weight estimates from step 2

$$\hat{v}_g = \frac{1}{|g|} \sum_{l \in g} \sum_e \hat{\delta}_{e,l} \widehat{Pr}\{E_i = e | E_i \in [-l, T - l]\}$$

Consistency and Inference

- Under parallel trends and no anticipation, $\hat{\delta}_{e,l}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IW estimator is consistent for a weighted average of $CATT_{e,l}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

DD Estimator of CATT

Definition 2: DD estimator with pre-period s and control cohorts C estimates $CATT_{e,l}$ as:

$$\widehat{\delta}_{e,l} = \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} \times 1\{E_i \in C\})]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,l}$.

Software

- **Stata:** eventstudyinteract (can be installed from ssc)
- **R:** fixest with subab() option (see
<https://lrberge.github.io/fixest/reference/sunab.html/>)

Reporting results

Table: Estimating ATT

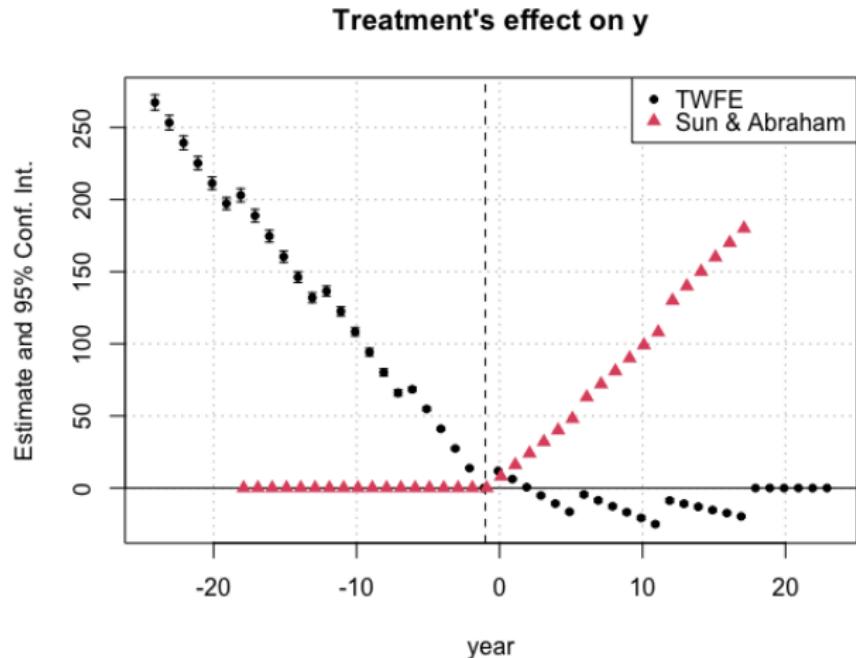
	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
<i>Feasible</i> \widehat{ATT}	68.33	26.81***	68.34***	68.33***	

Computing relative event time leads and lags

Year	Truth				Relative time coefficients		
	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)	Leads	Truth	SA
1980	0	0	0	0			
1986	10	0	0	0	t-2	0	0.02
1987	20	0	0	0	t	8	8.01
1988	30	0	0	0	t+1	16	16.00
1989	40	0	0	0	t+2	24	24.00
1990	50	0	0	0	t+3	32	31.99
1991	60	0	0	0	t+4	40	40.00
1992	70	8	0	0	t+5	48	48.01
1993	80	16	0	0	t+6	63	62.99
1994	90	24	0	0	t+7	72	72.00
1995	100	32	0	0	t+8	81	80.99
1996	110	40	0	0	t+9	90	89.98
1997	120	48	0	0	t+10	99	99.06
1998	130	56	6	0	t+11	108	108.01
1999	140	64	12	0	t+12	130	129.92
2000	150	72	18	0	t+13	140	139.92
2001	160	80	24	0	t+14	150	150.01
2002	170	88	30	0	t+15	160	159.97
2003	180	96	36	0	t+16	170	169.99
2004	190	104	42	4	t+17	180	179.98
2005	200	112	48	8			
2006	210	120	54	12			
2007	220	128	60	16			
2008	230	136	66	20			
2009	240	144	72	24			

Two things to notice: (1) there only 17 lags with robust models but will be 24 with TWFE; (2) changing colors mean what?

Comparing TWFE and SA



Question: why is TWFE *falling* pre-treatment? Why is SA rising, but jagged, post-treatment?

Advice

- DiD will remain popular for a while, and if anything all this new DiD has brought even more attention to it
- But now things are changing – how do we write the papers? Not just how do we estimate parameters
- Papers are a combination of science and rhetoric – let's look at a new one
- Braghieri, Levy and Makarin (2022), "Social Media and Mental Health", *American Economic Review*, 112(11): 3660-3693

Big picture

- Widely cited that social media causes mental health problems in youth
- Anecdotal, documentaries, but no causal evidence ("slim to none")
- Study will use staggered rollout of Facebook platform to college campuses from 2004 to 2006 to estimate the effect on aggregate mental health scores from a survey
- You be the judge, but they present what in most cases would be strong evidence that Facebook harmed college students mental health

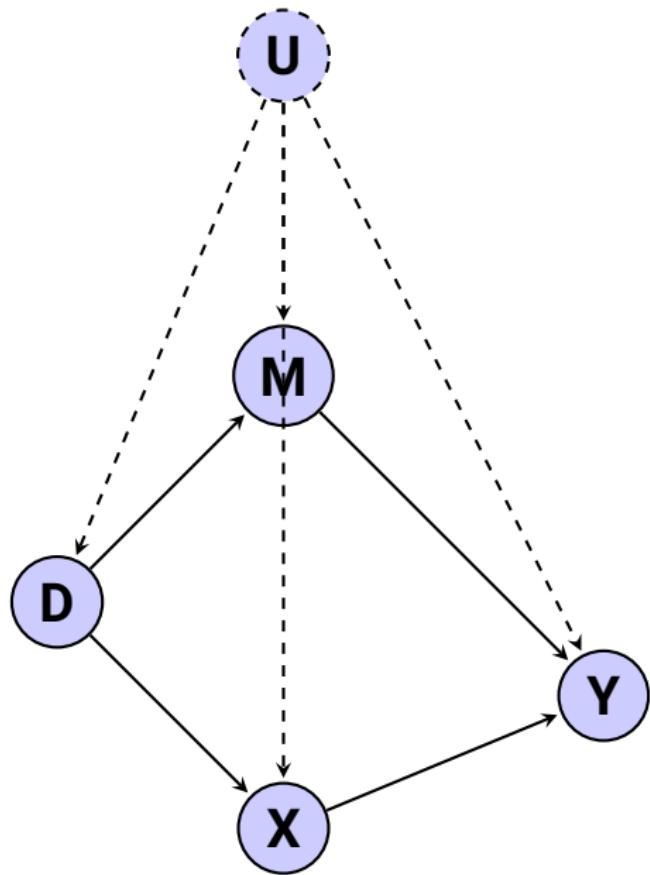
Many things to like

- Important question: mental health, suicide, review descriptive stats together
- Strong design: staggered rollout
- Event study is eye popping
- Mechanism and main results
- Very interesting dataset

Fives parts of a strong DiD

1. **Bite:** They cannot really show much here. No data on Facebook usage. More an ITT
2. **Main Results:** Estimated effect on mental health measures
3. **Mechanism:** Speculative
4. **Falsifications:** I can't really see very strong falsifications either.
5. **Event studies:** POW. Just wait

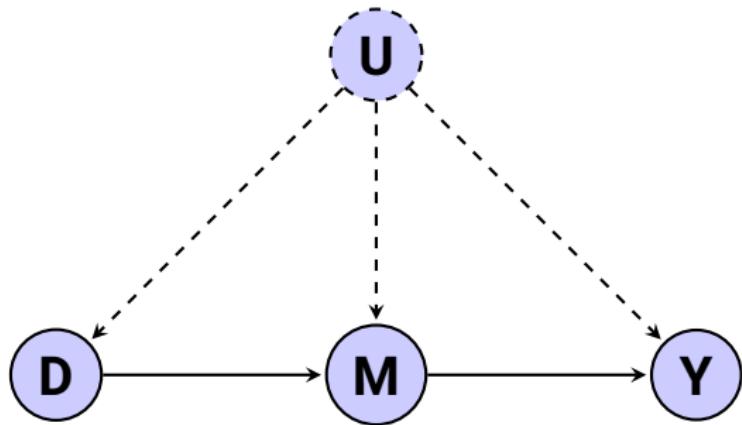
Mechanism



Mechanism

- D is the treatment variable, and the ATT is over all possible channels, but what if you want to think M is the mechanism
- When you can't rule out competing theories with falsifications, you have to try and build the case that the effect is coming through a channel
- Rule out X and provide evidence for M
- Goal here is to try and present evidence (not proof) that it's probably the story you're saying

Ruling out alternative mechanism



Mechanism

- Story is interpersonal comparisons which they try to show
- We can discuss how plausible we found it, but ask yourself at the end – did the event study help you believe it? Why/why not?

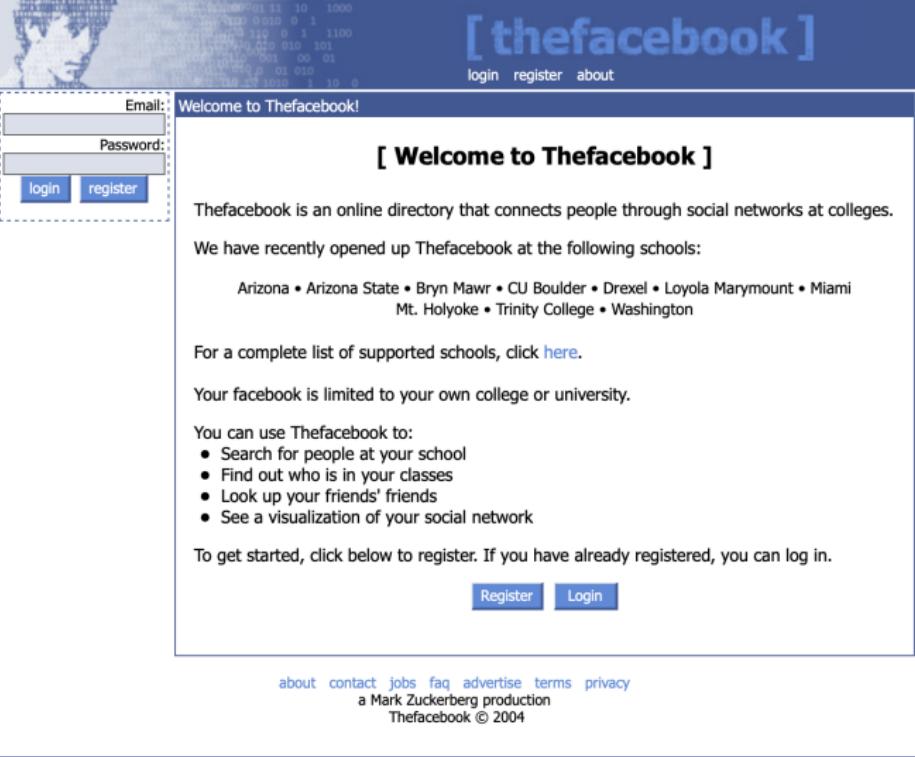
Data on Facebook

- Ingenious use of the Wayback Time Machine
- Looked at over 700 schools using Facebook screen shots
- When Facebook first mentions a school on its front page, that school is marked as having gotten Facebook

New schools being adopted

The screenshot shows the homepage of Thefacebook (now Facebook) from 2004. At the top, there's a blue header bar with the text '[thefacebook]' in white. Below it are links for 'login', 'register', and 'about'. On the left, there's a sidebar with a user profile picture and fields for 'Email:' and 'Password:', each with a corresponding input box. Below these are two blue buttons: 'login' and 'register'. The main content area has a white background. It features a large bold heading '[Welcome to Thefacebook]'. Below this, a paragraph reads: 'Thefacebook is an online directory that connects people through social networks at colleges.' Another paragraph states: 'We have opened up Thefacebook for popular consumption at:'. A list of college names follows: BC • Berkeley • Brown • BU • Chicago • Columbia • Cornell • Dartmouth • Duke • Emory • Florida • Georgetown • Harvard • Illinois • Michigan • Michigan State • MIT • Northeastern • Northwestern • NYU • Penn • Princeton • Rice • Stanford • Tulane • Tufts • UC Davis • UCLA • UC San Diego • UNC • UVA • WashU • Wellesley • Yale'. A section titled 'Your facebook is limited to your own college or university.' is present. Below it, a list of things you can do on Thefacebook includes: 'Search for people at your school', 'Find out who is in your classes', 'Look up your friends' friends', and 'See a visualization of your social network'. A note says: 'To get started, click below to register. If you have already registered, you can log in.' At the bottom of the page, there are two blue buttons: 'Register' and 'Login'. At the very bottom, there's a footer with links for 'about', 'contact', 'faq', 'advertise', 'terms', and 'privacy', followed by the text 'a Mark Zuckerberg production' and 'Thefacebook © 2004'.

New schools being adopted



Thefacebook

login register about

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have recently opened up Thefacebook at the following schools:

Arizona • Arizona State • Bryn Mawr • CU Boulder • Drexel • Loyola Marymount • Miami
Mt. Holyoke • Trinity College • Washington

For a complete list of supported schools, click [here](#).

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#) [Login](#)

about contact jobs faq advertise terms privacy
a Mark Zuckerberg production
Thefacebook © 2004

Data on college students

- NCHA Data is survey administered to college students on a semi-annual basis by American College Health Assoc
- Inquires about demographics, physical health, mental health, alcohol and drug use, sexual behaviors, and perception of these behaviors by peers
- ACHA merged a treatment indicator to each respondent based on Facebook dataset provided to them so that privacy could be maintained

Mental health

- Self-reported symptoms are standard medical practice in mental health – DSM-5 relies on self-reports such as difficulty sleeping, fatigue, feelings of guilt, suicidal ideation
- No data on Facebook or social media usage so this is ITT version of the ATT
- Respondent answers to the questions are aggregated into indices such as *poor mental health* where larger numbers are worse

Main TWFE Model

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times Facebook_{gt} + X_i \times \gamma + X_c \times \psi + \varepsilon_{icgt} \quad (3)$$

Y_{icgt} is an outcome for person i in wave t attending college c in expansion group g ; α_g is expansion-group or college fixed effects; δ_t are survey-wave fixed effects; $Facebook_{gt}$ indicates the respondents' campus has Facebook by time t at expansion group g ; X_i and X_c are individual and college-level controls; and standard errors are clustered at college level.

$\hat{\beta}$ identifies the ATT under parallel trends in the robust models

Robustness

- Main static results will all be in TWFE, but appendix shows other methods like CS and SA
- Event studies will show all models including some we haven't reviewed
- Growing popularity to show "all the robust DiD" models so that readers can see you aren't cherry picking

TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137 (0.040)	0.124 (0.022)	0.085 (0.033)	0.077 (0.032)
Observations	374,805	359,827	359,827	359,827
Survey-wave fixed effects	✓	✓	✓	✓
Facebook-expansion-group fixed effects	✓	✓		
Controls		✓	✓	✓
College fixed effects			✓	✓
FB-expansion-group linear time trends				✓

Notes: This table explores the effect of the introduction of Facebook at a college on student mental health. Specifically, it presents estimates of coefficient β from equation (1) with our index of poor mental health as the outcome variable. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation (1) without including controls; column 2 estimates equation (1) including controls; column 3, our preferred specification, replaces Facebook-expansion-group fixed effects with college fixed effects; column 4 includes linear time trends estimated at the Facebook-expansion-group level. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Column 2 also includes indicators for geographic region of college (Northeast, Midwest, West, South); such indicators are omitted in columns 3 and 4 because they are collinear with the college fixed effects. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. Standard errors in parentheses are clustered at the college level.

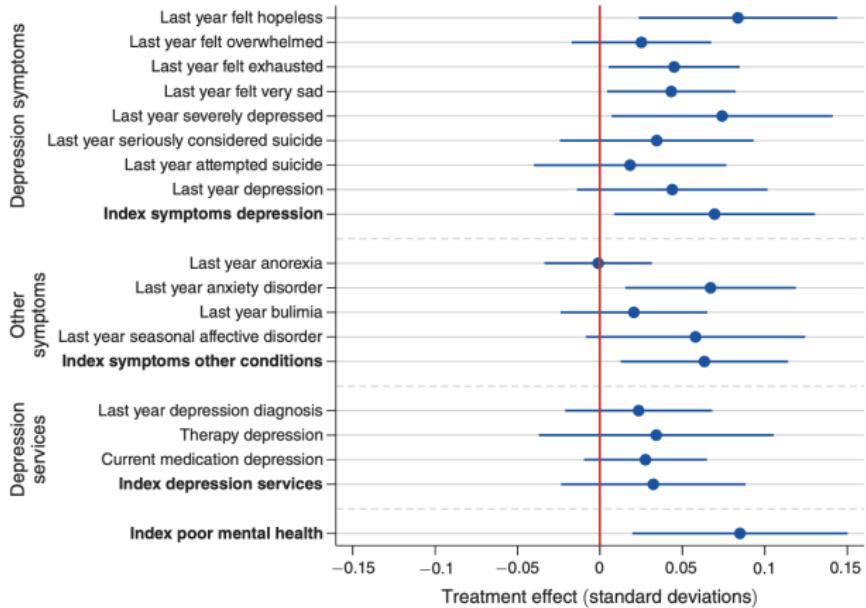


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

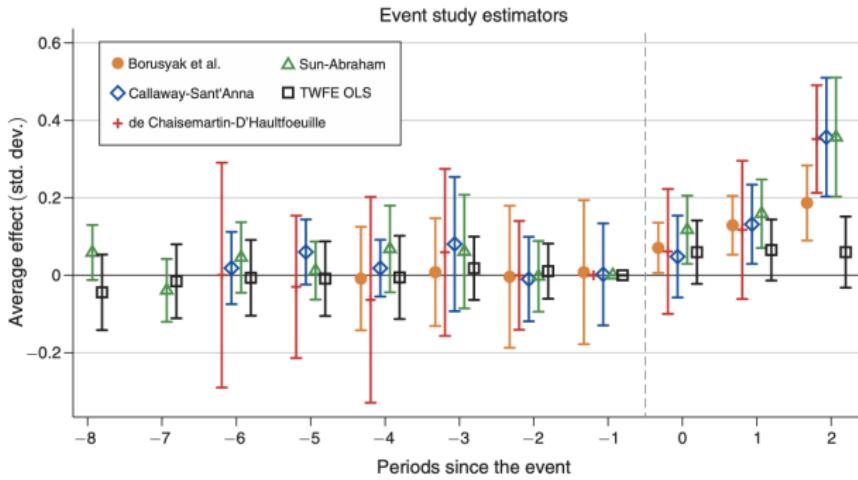


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

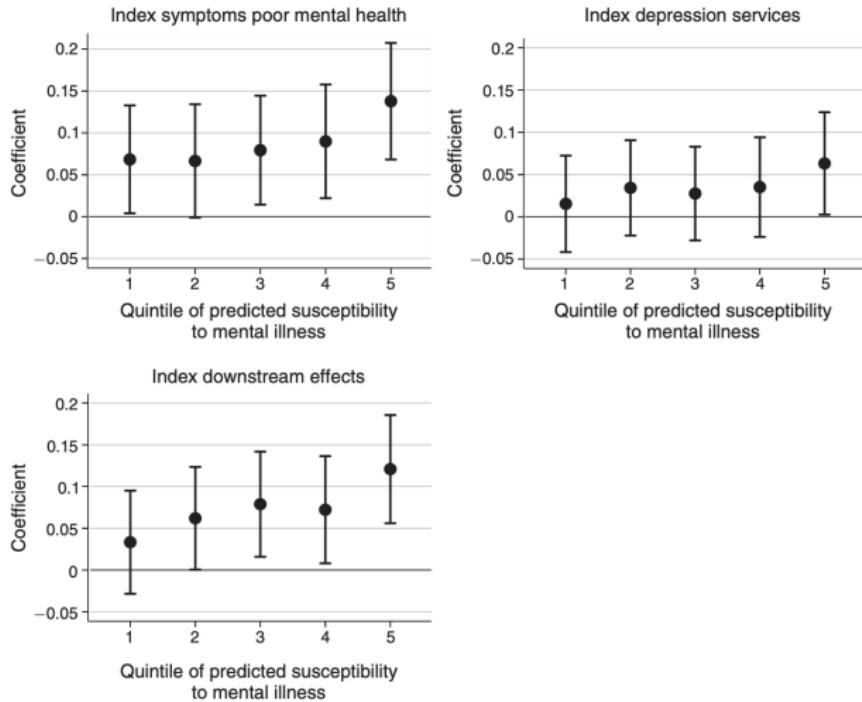


FIGURE 3. HETEROGENEOUS EFFECTS BY PREDICTED SUSCEPTIBILITY TO MENTAL ILLNESS

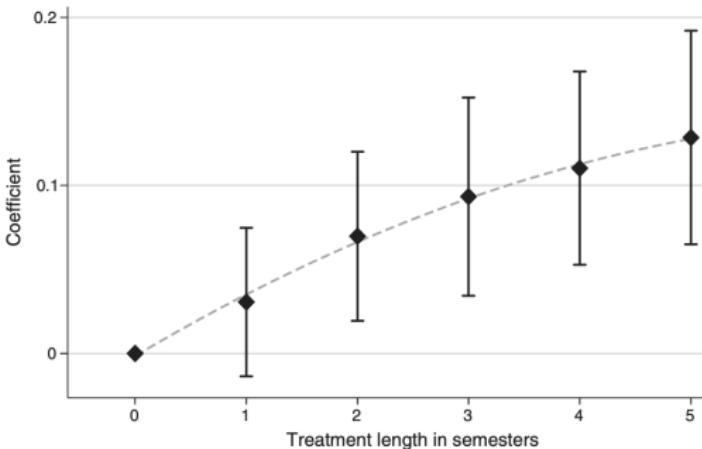


FIGURE 4. EFFECT ON POOR MENTAL HEALTH BY LENGTH OF EXPOSURE TO FACEBOOK

Notes: This figure explores the effects of length of exposure to Facebook on our index of poor mental health by presenting estimates of equation (4). The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. The dashed curve is the quadratic curve of best fit. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Students who entered college in 2006 might have been exposed to Facebook already in high school, because, starting in September 2005, college students with Facebook access could invite high school students to join the platform. Such students are excluded from the regression. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

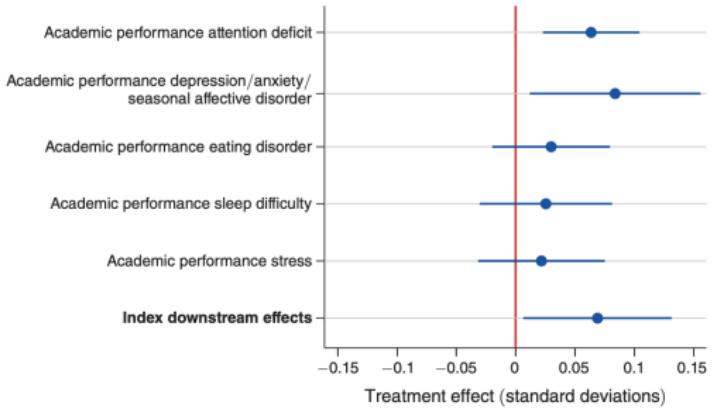


FIGURE 5. DOWNSTREAM EFFECTS ON ACADEMIC PERFORMANCE

Notes: This figure explores downstream effects of the introduction of Facebook on the students' academic performance. It presents estimates of coefficient β from equation (1) using our preferred specification, including survey-wave fixed effects, college fixed effects, and controls. The outcome variables are answers to questions inquiring as to whether various mental health conditions affected the students' academic performance and our index of downstream effects. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

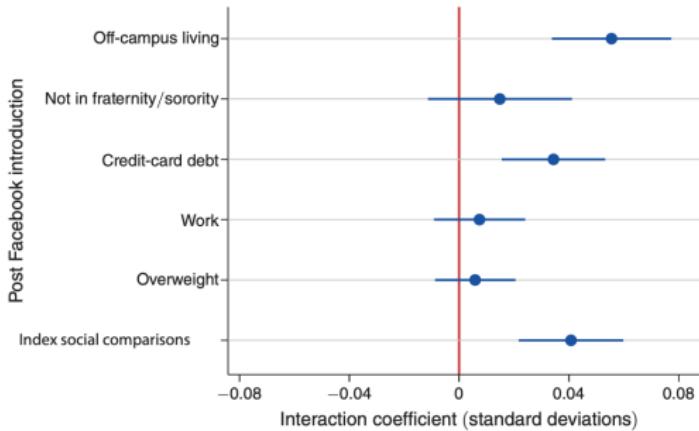


FIGURE 6. HETEROGENEOUS EFFECTS AS EVIDENCE OF UNFAVORABLE SOCIAL COMPARISONS

Notes: This figure explores the mechanisms behind the effects of Facebook on mental health. It presents estimates from a version of equation (1) in which our treatment indicator is interacted with a set of indicators for belonging to a certain subpopulation of students. The outcome variable is our overall index of poor mental health. The estimates are obtained using our preferred specification, namely the one including survey-wave fixed effects, college fixed effects, and controls. For a detailed description of the outcome, treatment, interaction, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

Comments

- Can't lose sight of the big picture – you still have to write a great paper, not just pass your exams, and going from estimation to publishing is a different but related skill
- Some of the old exhibits may not carry forward (TWFE with many columns)
- Increasingly, people are presenting a single event study graph with “all the DiD” against TWFE so as to avoid cherry picking estimators
- You should use the tool for the job, but these differences are subtle (“which parallel trends?”, “which comparison group?”)

Conclusion

- Good question, good data, and you can publish well with DiD
- Hardly definitive, but the staggered design is a solution to our inability to run the RCT
- Remember – many questions can be randomized in theory but not practice (e.g., smoking)
- Learn as much as you can and don't stop learning

Roadmap

TWFE Pathologies

Brief history

Bacon decomposition

Simulation

Two solutions and a new decomposition

CS

SA

Some opinions and an application

Synthetic control

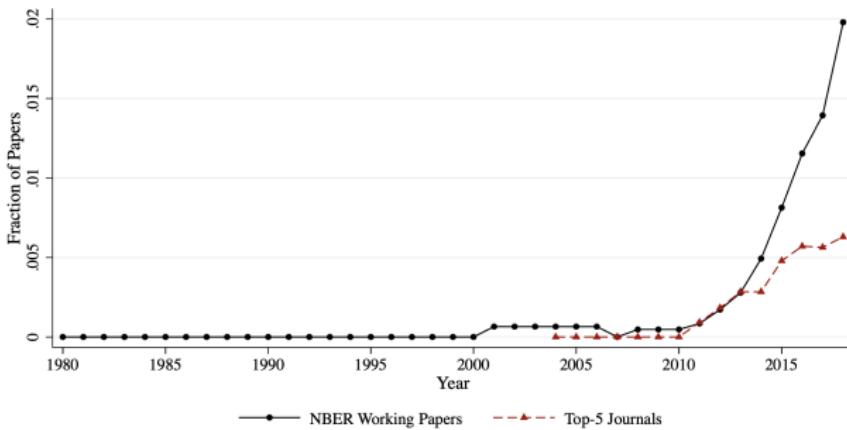
Abadie's non-negative weighting

Ben-Michael, et al's Least Negative Weighting Method

Athey, et al's Matrix Completion with Nuclear Norm Method

Synthetic difference-in-differences

D: Synthetic Control



What is synthetic control

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

What is a comparative case study

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

Qualitative comparative case studies

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

Traditional quantitative comparative case studies

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

Pros and cons

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

Description of the Mariel Boatlift

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







Selecting control groups

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

Why these four?

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

Diff-in-diff

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

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

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

Parallel trends

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

Synthetic Control

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

Researcher's objectives

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

Synthetic control method: advantages

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

Notation and setup

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

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

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

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, \dots, v_k . Then, the synthetic control weights w_2^*, \dots, w_{J+1}^* minimize:

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

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

How this works

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

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

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

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, \dots, v_k can be based on

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

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

Cross validation

Abadie suggests model selection based on training and validation:

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

Avoiding Cherry Picking Synth

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

Avoiding Cherry Picking Synth

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

With and without covariates, report p-values and event study plot

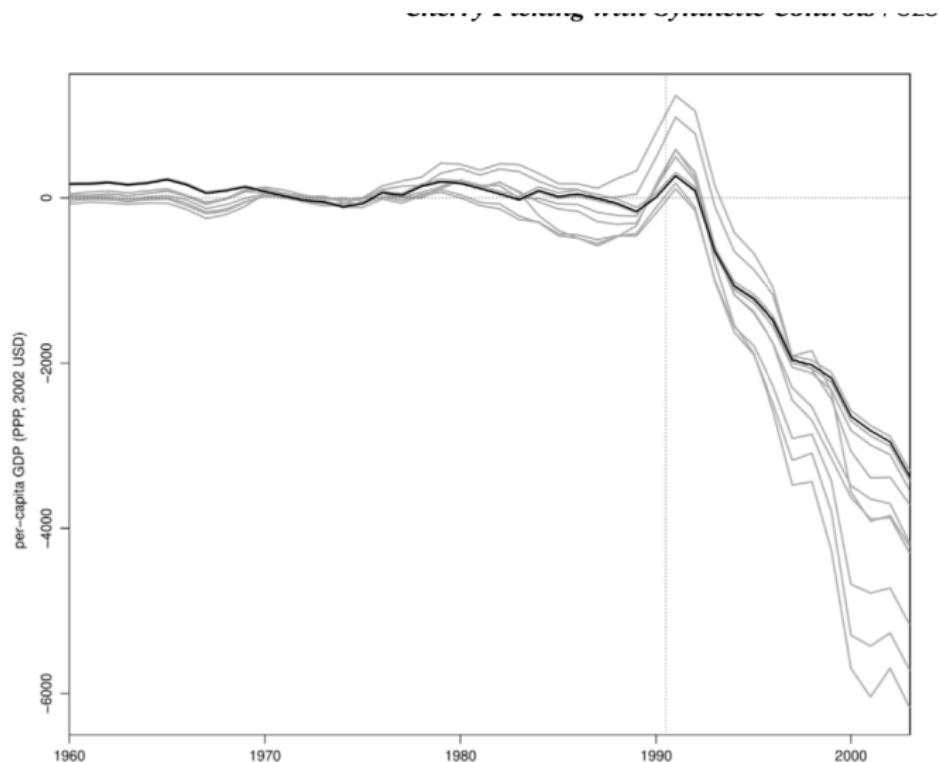
Avoiding Cherry Picking Synth

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 Synth



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.

Assumption: Y^0 is determined by factor model

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

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

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

With some manipulation

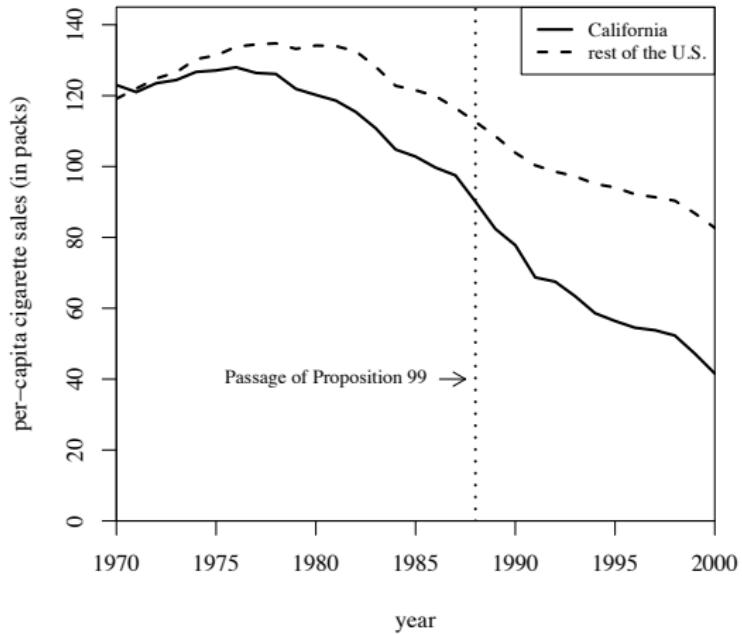
$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

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

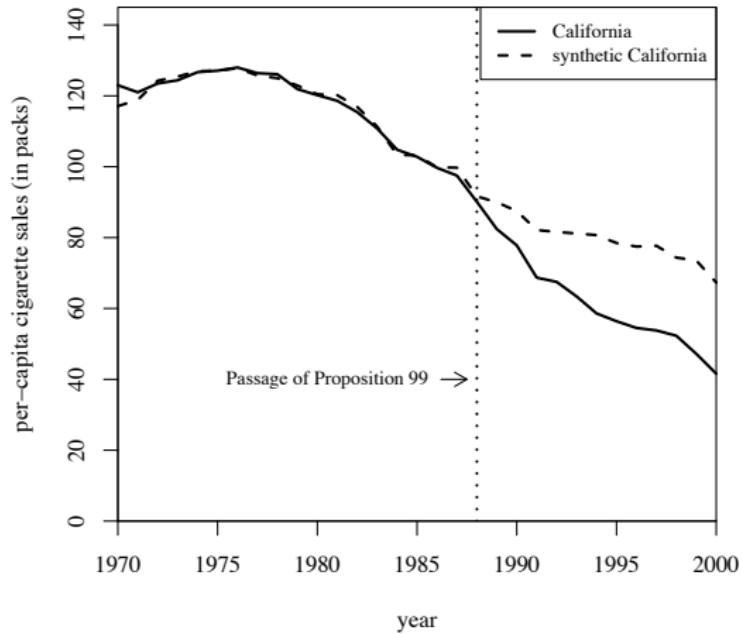
Example: California's Proposition 99

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

Cigarette Consumption: CA and the Rest of the US



Cigarette Consumption: CA and synthetic CA

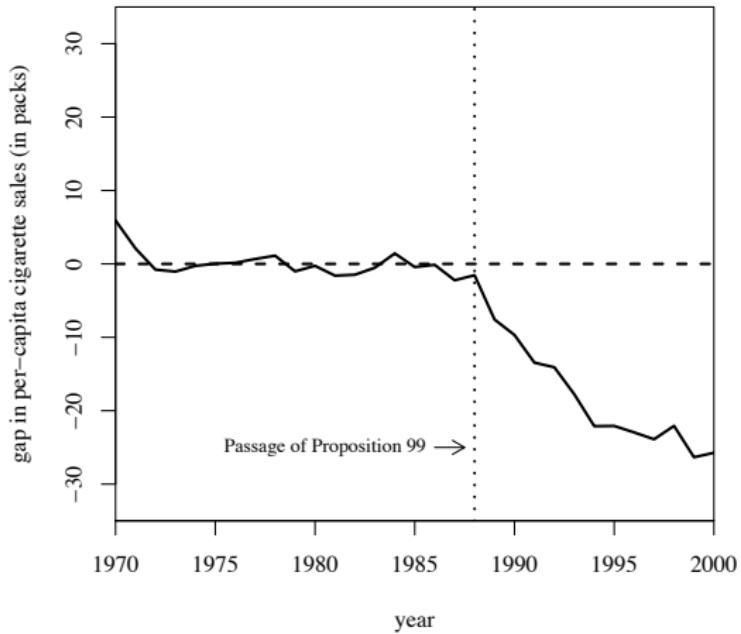


Predictor Means: Actual vs. Synthetic California

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

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

Smoking Gap between CA and synthetic CA



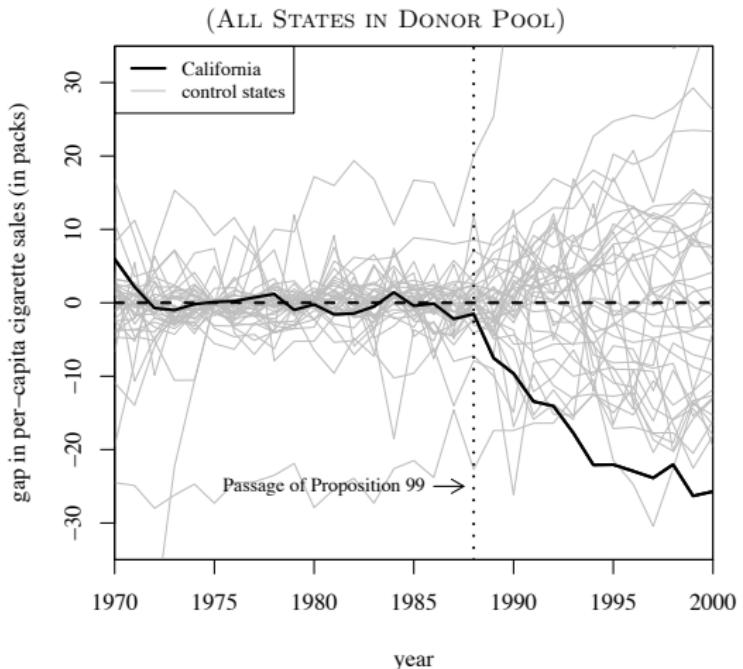
Inference

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

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

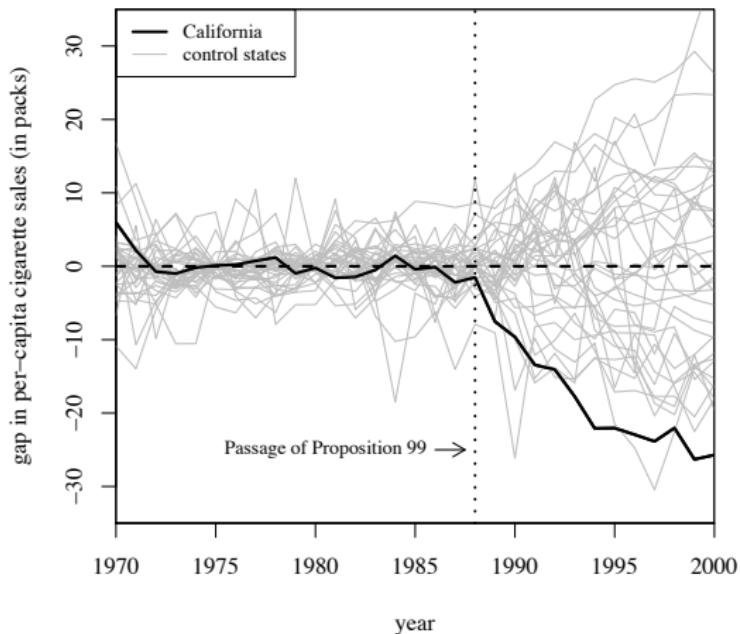
and the exact p-value is the treatment unit rank divided by J

Smoking Gap for CA and 38 control states



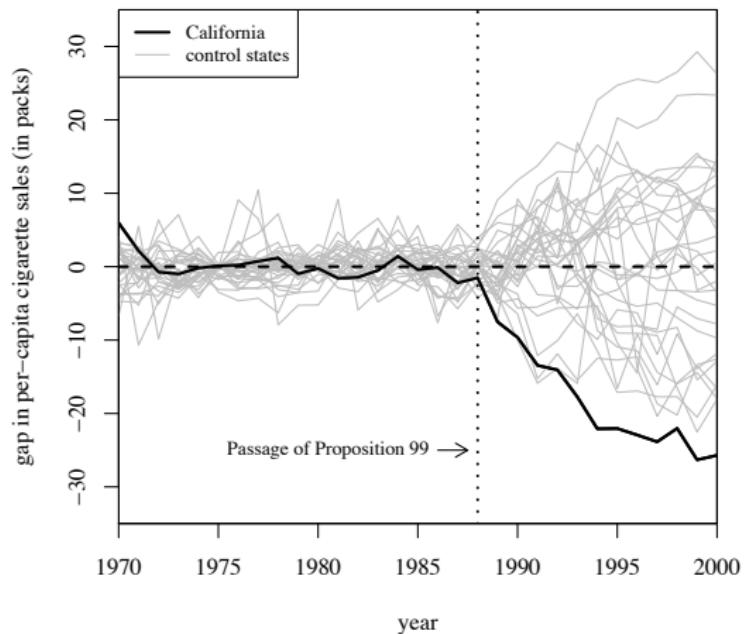
Smoking Gap for CA and 34 control states

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



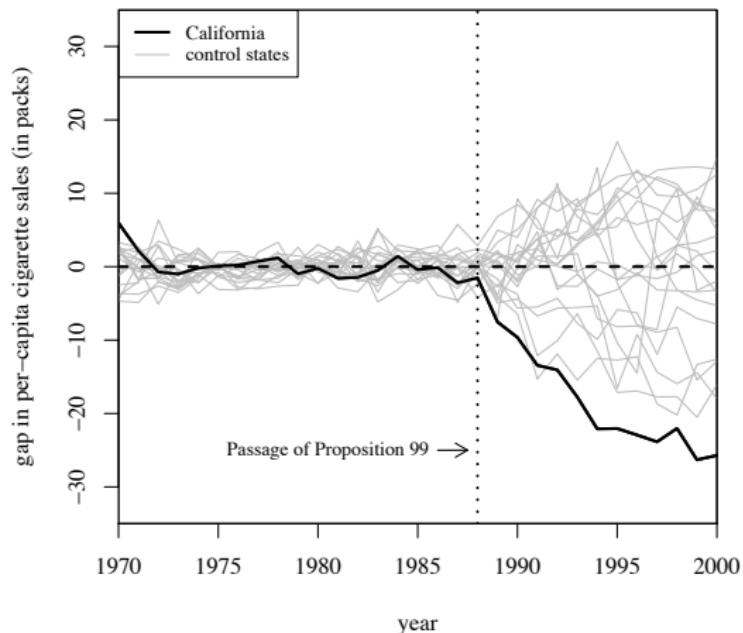
Smoking Gap for CA and 29 control states

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

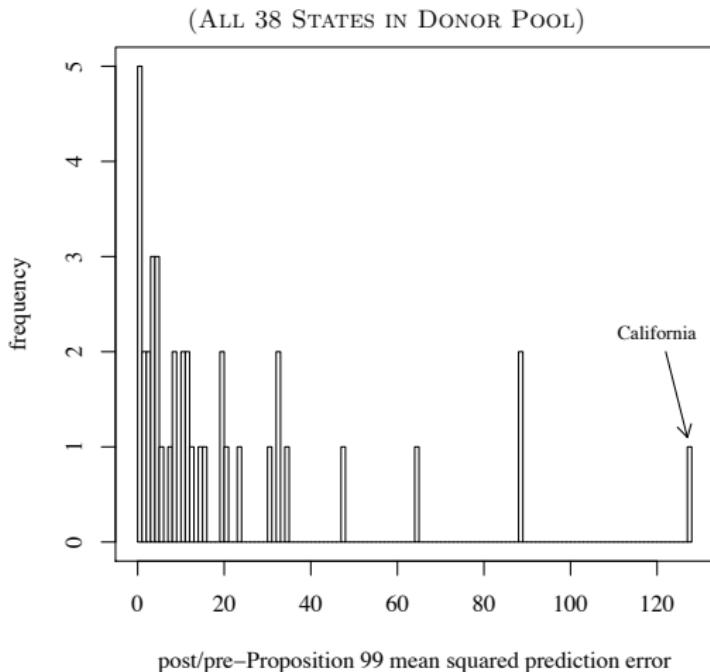


Smoking Gap for CA and 19 control states

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



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



Example

- Let's look at an example now from my own work
- Rhode Island legalized sex work in 1980 (indoor anyway) but did so accidentally
- Judge enforced it in 2003
- Use DiD and synth to estimate the effect on violence and STIs
- Also bite and mechanism

Institutional details

- Google alert in 2009 took me to an article that told this whole story
- There is no evidence that anyone knew about the 1980 legalization
- Appears to be unintentional legalization due to poorly worded amendments to §11-34
- July/August 2003 “Operation Rubdown” and judicial decriminalization
 - Wave of arrests of massage parlor employees.
 - Judge dismisses charges against massage parlor employees and police stop arresting indoor sex workers

"Accidental" and unknown legalization

- Senator John Revens Jr. (2009) on the 1980 General Assembly
 - "they would never sponsor a bill decriminalizing prostitution if they knew what it was. No way. Not in a million years" (Arditi 2009)
- Senator John F. McBurney III (from the 1980 GA)
 - "We probably vote on 500 bills a year. ... Legislators didn't know what they were voting for."
- Unusual lack of awareness of the 1980 law from 1980–2005
 - Not a single mention of the law change in any newspaper until Breton (2005)
 - No legal scholars mention it (e.g., Posner and Silberbaugh 1996)

Discovery of Loophole

- 1997 *State vs. DeMagistris*: §11-34-8 and §11-34-8.1 apply only to street prostitution
- July/August 2003 “Operation Rubdown” and judicial decriminalization
 - Wave of arrests of massage parlor employees.
 - Judge dismisses charges against massage parlor employees and police stop arresting indoor sex workers

Learning

- Police stop arresting indoor sex workers

"Chief of Police Tom Verdi says that until 2 years ago, Providence police were arresting alleged prostitutes inside massage parlors. But they stopped doing so after [Kiselica] persuaded District Court judges to dismiss prostitution cases based on the wording of the current law." (2005 newspaper article)

- Re-criminalization in 2009

Our study

- Tons of evidence that sex workers are raped at a higher rate than the general population
- Tons of evidence that sex workers have been the drivers of STI epidemics in areas
- But is this causal, or is it correlational?
- And is this caused by sex work or is it caused by sex work policy?
- We didn't know – humility! – until this study

Proposed Mechanism Framework: Rape

1. Decriminalization ↑ rape offenses:
 - Increase number transactions in indoor sex market by reducing costs of indoor sex work ⇒ increase prostitution related violence
2. Decriminalization ↓ rape offenses:
 - Freeing up of police personnel and resources
 - Firms invest in locks, security cameras, security personnel to reduce opportunity of premeditated client violence (Brents and Hausbeck 2005)
 - Sex workers more willing to cooperate with police/less police abuse (Church et al. 2001, Levitt and Venkatesh, 2007)
 - Could reduce male violence if prostitution is substitute for violence against women (Posner 1994)

Proposed Mechanism Framework: STIs

1. Decriminalization ↑ population STIs:
 - Increase number transactions in indoor sex market ⇒ increase the scale and growth rate of a gonorrhea epidemic
2. Decriminalization ↓ population STIs:
 - If decriminalization shifts transactions indoors to lower STI risk sex workers and/or draws in lower risk sex workers ⇒ decrease an epidemic (Kremer 1996)
 - Empirical evidence of safer indoor sex work:
 - More indoor transactions (relative to street) lower STI rates (Gertler and Shah 2011)
 - Massage parlor SWs in UK use condoms more, receive STI screens more, report fewer weekly transaction than street-based SWs (Jeal and Salisbury 2007)
 - Higher gonorrhea incidence and more requests for non-condom sex among illegal street SWs than licensed indoor SWs in Australia (Seib et al. 2009)

We collected a lot of different data

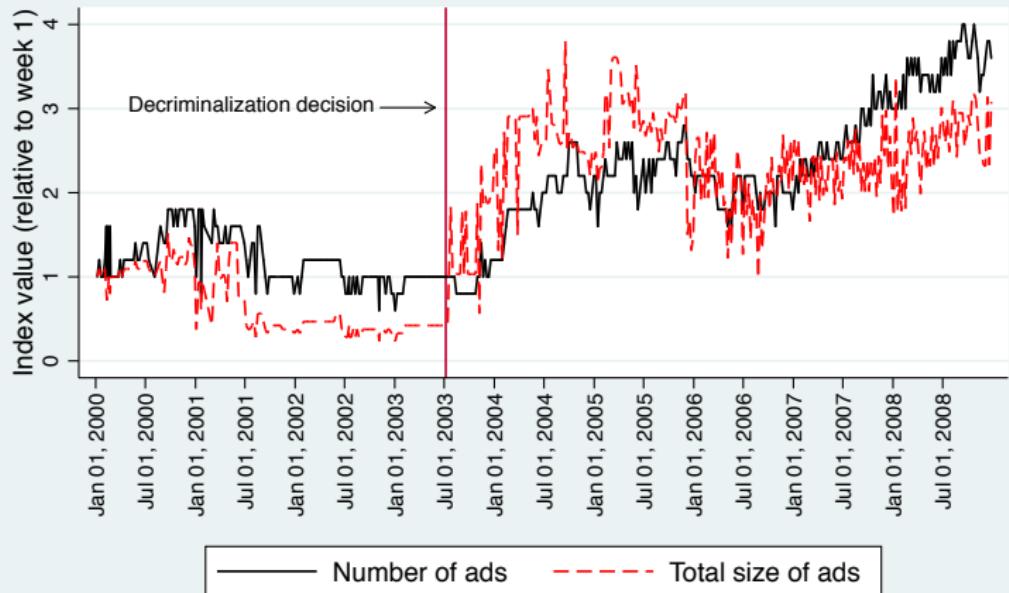
- Lots of crime data (from the FBI from police)
- Lots of data on gonorrhea (from the CDC)
- Lots of data on classified ads for massage parlors (from the Providence Phoenix)
- Lots of data on prostitution arrests (from the FBI from police)
- Lots of data on indoor sex workers reviewed online by clients (from The Erotic Review)

Evidence

1. **Bite**: Advertising and sex worker reviews
2. **Falsifications**: Other crimes
3. **Mechanism**: Entry and sex acts
4. **Main results**: DiD and synth estimates of ATT for rape and gonorrhea
5. **Visuals**: Lots of event study and synth plots

Providence Phoenix Advertising

Providence Phoenix Weekly Adult Services Ads
Number of advertisers and total size advertisements



Five unique advertisers purchased a combined 0.67 of full page at start of series.

Providence Phoenix Advertising (Before)

Marilyn & co
Adult Work Out Club, Day Spa, Fitness, Fun
(201) 291-0380
Body Building, Aerobics, Cardio

Quality Escorts
Handpicked Escorts for Your Pleasure & Relaxation
Call 401-291-0380
24 hr. Emergency Escort Service
We Offer Escort & Escort Services

**Endless Fun in
Bachelor Parties!**
Call 242-2829
401-954-3719
Lacee Modeling

**Feeling Lucky
Boy or? Picture is Real!**
401-958-8806

**10 places where
ESCORTS ad here, call**
401-273-6281 x200

Maxim Lingerie
Lacewear & Modeling Studio
Sensual Private Lingerie Modeling
Shows & Varieties & Performances
Role Playing / Cosmetology & Clothing Alterations
Assistance with Exercises & State of Ultimate Support
227 Main St., Providence, RI 02803
Phone: 401-273-6281
Fax: 401-273-6282
Employment Opportunities Available
Local Business Opportunities

Real Seduction
Satisfy your Senses!
(Call 24 hrs)
1-888-844-5988
(Leave a message)
RealProduction, Inc.
Always Hiring

WOMAN
100% Natural, Handmade
in a Small Family Operated
Business Since 1982
Specialty: Handmade
Cosmetics, Soaps, Candles
and Bath Products

-SPAS-

Spa Midori
Body wrap
Showers
Dry Sauna
Walk-In Service
7 days a week 10am-12pm
112 Union St - 1st Floor
(Providence Aerobic) Providence, RI
(401) 274-3334 • (401) 274-6661

TOKYO SPA
Hot桑拿, Exercise
Tables Showers
RI 95 B-5 to Exit 10
27 West Street • Pawtucket, RI
(401) 722-0111

1410 Rear Mineral Spring Ave, North Providence, RI
OPEN 7 DAYS A WEEK • WALK IN'S WELCOME
From 95N, Take exit 23 to Rte. 146 to the Mineral Spring Ave exit, take Left onto Mineral Spring. Building is on corner of Mineral Spring and Woodward Rd.
From 95N, Take exit 24 (Branch Ave.), take Right onto Branch Ave., go 7 lights to Woodward Rd. and take a right. Go to end of Woodward to Mineral Spring Ave - building is on the left.

ORIENTAL GARDEN SPA
770 N. Main Street • Providence
Open 7 Days a week 10am to Midnight

• Dry Sauna • Steam Sauna
• Massage • Walk-in Service

Directions: From 95 N
Exit 24 left onto Branch Avenue From
95 S Exit 24 right onto Branch Avenue
Both take right turn onto North Main
Street then 1st Right on Livingston Street

Branch Ave.
Exit 24
Main St.
Livingston St.
North Main St.
Industrial Drs.

**To place your
ad for a **SPA**
here, call
401-273-6397**

www.thephoenix.com

401.621.8609

Providence Phoenix Advertising (After)

JULY 23

SPAS
MORE SPAS ON THE NEXT PAGE

Lily Spa
Private Parking
New Hours
• Body Rub
• Body Shampoo
• Steam & Dry Sauna

Pleasant Massage Therapy
401-718-1700
STEAM & DRY SAUNA
BODY SCRUB
LICENSED MASSAGE THERAPIST

ORIENTAL GARDEN SPA
401-473-6800

DOWNTOWN SPA
Under New Management
401-553-5800

WANICKICK WELLNESS CENTER
Bodywork • Reflexology
Accupuncture • Dry Sauna
Steam Room • Grotto & Hammam

Pinetree Spa
401-718-1700
Body Scrub • Body Treatment
Lavender Massage • Therapeutic

Midori Spa
Walk-In Service
7 days a week 10am-12pm
112 Union St., 1st Floor
(Telephone Building)
Providence, RI

2000
2000
2000

Market Prices and Composition

Table 3 Effect of Decriminalization on Transaction Characteristics

Dependent variable:	Ln Price	Massage	Oral condom	Oral bare	Vaginal Sex	Anal Sex	White	Asian	Hispanic	Black
Panel A: Clustered Standard Errors										
RI decriminalization	-0.414*** (0.041)	0.231*** (0.032)	0.080*** (0.012)	-0.212*** (0.022)	-0.135*** (0.024)	-0.177*** (0.009)	0.023 (0.015)	0.178*** (0.020)	-0.012 (0.012)	0.021** (0.007)
Panel B: Permutation Tests										
RI decriminalization	-0.414	0.231	0.080	-0.212	-0.135	-0.177	0.023	0.178	-0.012	0.021
5th percentile	-0.258	-0.157	-0.186	-0.306	-0.127	-0.061	-0.141	-0.163	-0.055	-0.048
95th percentile	0.239	0.139	0.342	0.109	0.134	0.074	0.123	0.080	0.078	0.069
Two-tailed test p-value	0.09	0.05	0.60	0.28	0.09	0.05	0.60	0.05	0.47	0.47
Observations	82944	83135	83135	83135	83135	83135	83135	83135	83135	83135
Baseline mean	5.39	0.11	0.56	0.39	0.94	0.22	0.44	0.22	0.06	0
State and year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

These are DD regressions using 1999-2009 years. Individual controls include a dummy whether the provider was independent. Panel A presents clustered standard errors and Panel B presents 5th and 95th percentile confidence intervals from placebo-based inferential calculations, and p-values from a two-tailed test. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Arrests, rapes and gonorrhea

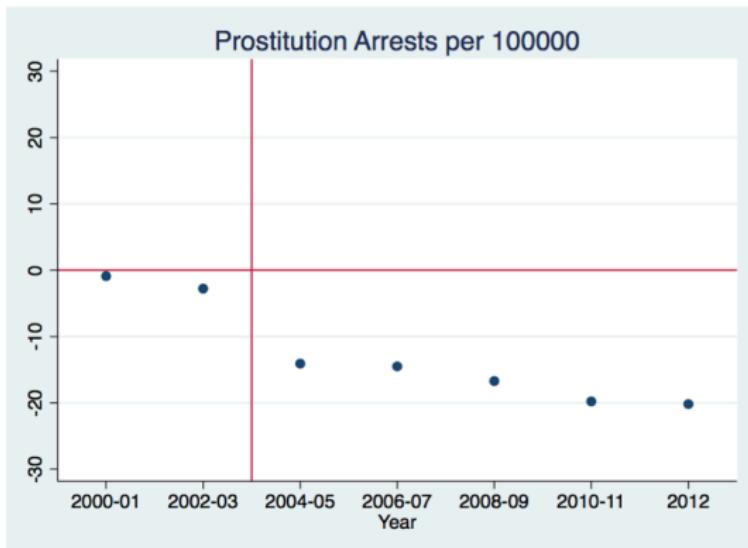
So it looks like the policy caused sex work to increase – itself new, as many don't think it can change

But Gary Becker (Rest in Peace) said our goal needs to be to minimize the social costs of crime – so did this policy increase or decrease those social costs?

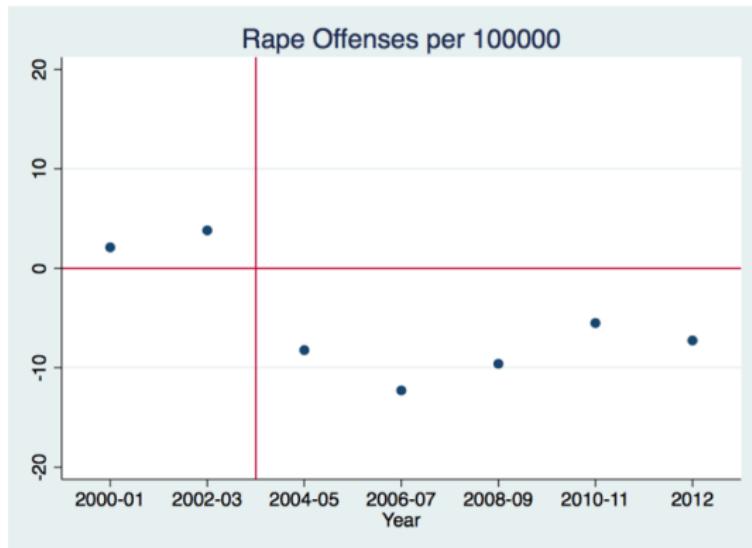
Let's look at sex worker arrests, rapes, and gonorrhea

Causal assumption: if what happens in other states is what *would have happened* ("parallel trends") in Rhode Island, then we can use the other states as substitutes for Rhode Island counterfactuals

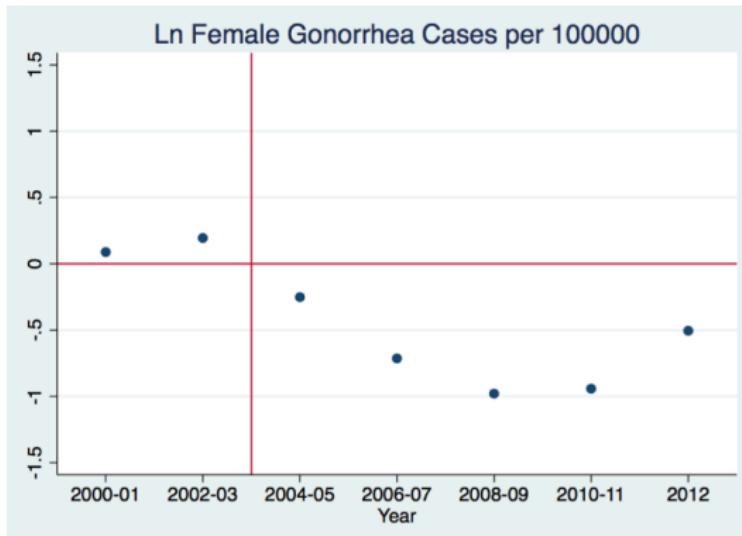
Parallel trends: prostitution arrests



Parallel trends: rape offenses



Parallel trends: gonorrhea offenses



Arrests, Rapes and Gonorrhea

Table 2 Effect of Decriminalization on Arrests, Rape and Gonorrhea

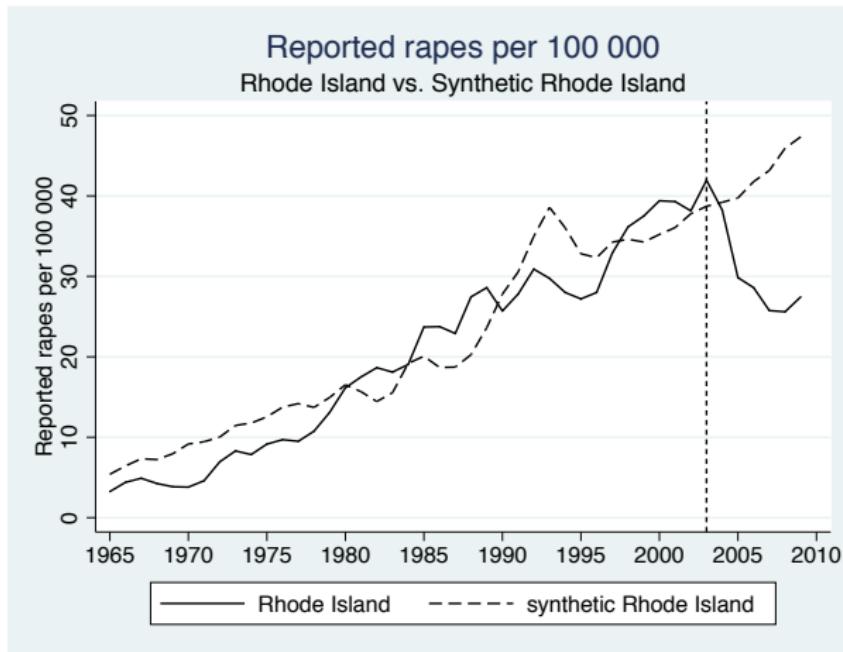
Dependent variable:	Prostitution Arrests		Rape Offenses		Ln Female Gonorrhea	
Panel A: Clustered Standard Errors						
RI decriminalization	-13.650*** (1.187)	-8.806* (3.341)	-12.607*** (0.798)	-13.712*** (1.334)	-0.762*** (0.034)	-0.633*** (0.069)
Panel B: Permutation Tests						
RI effect post-decriminalization	-13.650	-8.806	-12.607	-13.712	-0.762	-0.633
5th percentile	-12.365	-14.832	-7.548	-7.027	-0.292	-0.276
95th percentile	12.052	12.255	11.584	10.595	0.482	0.335
Two-tailed test p-value	0.08	0.35	0.04	0.04	0.04	0.04
Observations	545	545	561	561	561	561
Baseline mean	34.05	34.05	40.4	40.4	4.39	4.39
State and year FE	Yes	Yes	Yes	Yes	Yes	Yes
Time variant controls	No	Yes	No	Yes	No	Yes

These are DD regressions using 1999-2009 Uniform Crime Reports (Arrests and Rape Offenses) and CDC (Female Gonorrhea) data. Time-variant controls include female population, male population, unemployment rate, share of population below poverty line, share of population in military, share of white population, share of black population, share of population that is male and single, share of population that is female and single, share of population that is male and married, and share of population that is female and married. Panel A presents clustered standard errors and Panel B presents 5th and 95th percentile confidence intervals from permutations tests and p-values from a two-tailed test. * p<0.10, ** p<0.05, *** p<0.01

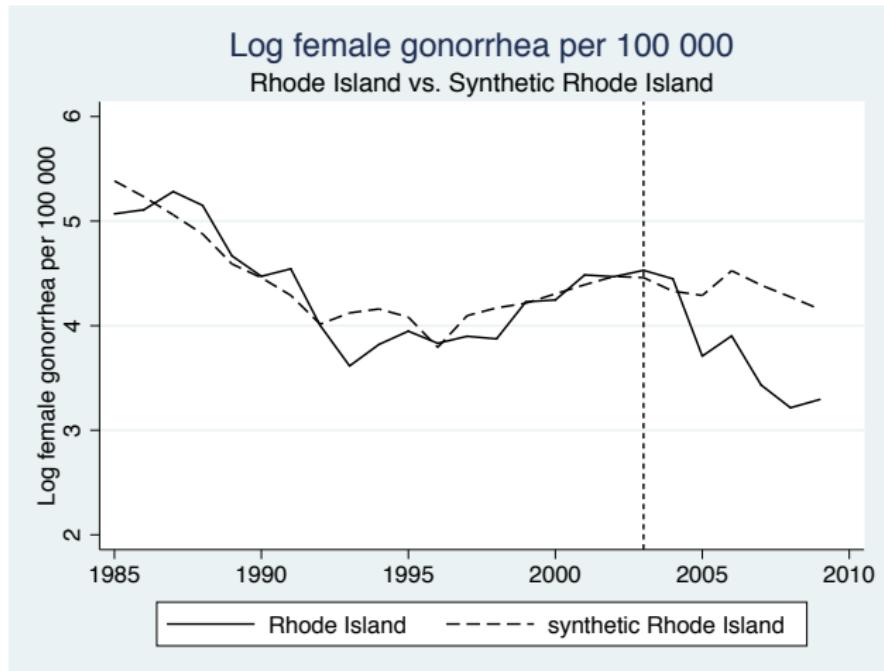
Synthetic Control Analysis

- Think about minimum wage increases – they happen in different places at different points in time
- Synthetic control estimator is useful when only one group is treated at one point in time
- Similar to what we did only 1) we don't need parallel trends anymore and 2) only uses the best states as controls
- What's the best states? The ones that had the same time path in rapes, gonorrhea and arrests before the legalization event
- "If they looked the same before, why wouldn't they look the same after?"

Synthetic Control Model: Rape



Synthetic Control Female Gonorrhea



Why did reported female rape offenses fall?

Some possible explanations

1. Did police reallocate resources away from arrests to investigating rapist? Defense attorney told us **no**
2. Did definitions of rape change? **Not at this time**
3. Did fewer sex workers get raped after decriminalization? Probably but most *reported* rapes **are not** sex workers in the first place
4. Did violent males think of rape and sex with prostitutes as substitutes? **Maybe?**

Why did female gonorrhea rates fall?

Some possible explanations

1. Safer sex workers entered the market thus diluting the propagation of gonorrhea
2. Safer sex in relevant part of sexual network thus diluting the propagation of gonorrhea
3. Spillovers to males causing feedbacks throughout the network

Compositional effects

Table 3 Effect of Decriminalization on Transaction Characteristics

Dependent variable:	Ln Price	Massage	Oral condom	Oral bare	Vaginal Sex	Anal Sex	White	Asian	Hispanic	Black
Panel A: Clustered Standard Errors										
RI decriminalization	-0.414*** (0.041)	0.231*** (0.032)	0.080*** (0.012)	-0.212*** (0.022)	-0.135*** (0.024)	-0.177*** (0.009)	0.023 (0.015)	0.178*** (0.020)	-0.012 (0.012)	0.021** (0.007)
Panel B: Permutation Tests										
RI decriminalization	-0.414	0.231	0.080	-0.212	-0.135	-0.177	0.023	0.178	-0.012	0.021
5th percentile	-0.258	-0.157	-0.186	-0.306	-0.127	-0.061	-0.141	-0.163	-0.055	-0.048
95th percentile	0.239	0.139	0.342	0.109	0.134	0.074	0.123	0.080	0.078	0.069
Two-tailed test p-value	0.09	0.05	0.60	0.28	0.09	0.05	0.60	0.05	0.47	0.47
Observations	82944	83135	83135	83135	83135	83135	83135	83135	83135	83135
Baseline mean	5.39	0.11	0.56	0.39	0.94	0.22	0.44	0.22	0.06	0
State and year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

These are DD regressions using 1999-2009 years. Individual controls include a dummy whether the provider was independent. Panel A presents clustered standard errors and Panel B presents 5th and 95th percentile confidence intervals from placebo-based inferential calculations, and p-values from a two-tailed test. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Spillovers

Dependent variable:	Prostitution Arrests		Rape Offenses		Ln Female Gonorrhea		Ln Male Gonorrhea	
Panel A: Clustered Standard Errors								
RI decriminalization	-13.633*** (1.176)	-8.706* (3.451)	-12.607*** (0.790)	-14.178*** (1.387)	-0.762*** (0.034)	-0.698*** (0.076)	-0.364*** (0.036)	-0.351*** (0.070)
Panel B: Permutation Tests								
RI decriminalization	-13.633	-8.706	-12.607	-14.178	-0.762	-0.698	-0.364	-0.351
5th percentile	-12.348	-15.330	-7.548	-7.677	-0.292	-0.289	-0.331	-0.301
95th percentile	12.072	12.352	11.584	10.655	0.482	0.371	0.482	0.397
Two-tailed test p-value	0.08	0.35	0.04	0.04	0.04	0.04	0.08	0.04

Recriminalization

- November 2009: sweeping sex trafficking legislation
- Prostitution in all its forms (including indoor) recriminalized

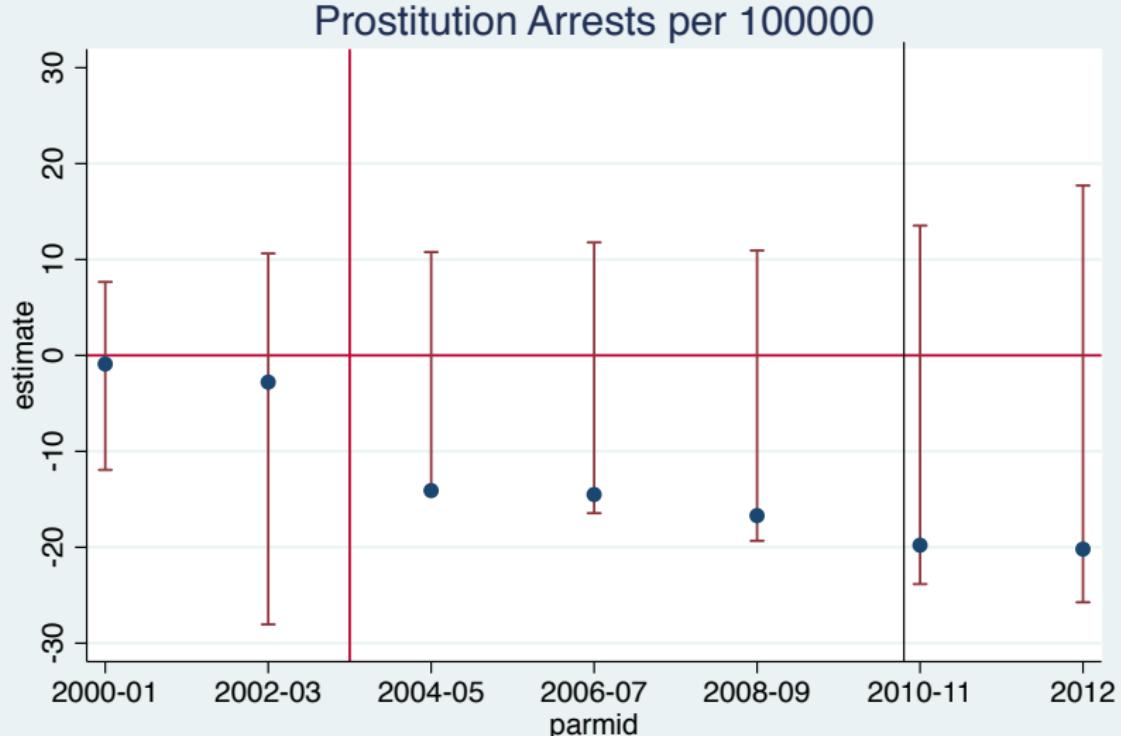
Market Prices and Composition

Table 7 Effect of Decriminalization and Recriminalization on Transaction Characteristics

Dependent variable:	Ln Price	Massage	Oral condom	Oral bare	Vaginal sex	Anal sex	White	Asian	Hispanic	Black
Panel A: Clustered Standard Errors										
RI decriminalization	-0.333*** (0.051)	0.162*** (0.034)	0.075*** (0.013)	-0.187*** (0.030)	-0.158*** (0.027)	-0.180*** (0.009)	0.062*** (0.013)	0.059*** (0.016)	0.039** (0.012)	0.050*** (0.007)
Panel B: Permutation Tests										
RI decriminalization	-0.333	0.162	0.075	-0.187	-0.158	-0.180	0.062	0.059	0.039	0.050
5th percentile	-0.283	-0.165	-0.190	-0.383	-0.164	-0.075	-0.153	-0.147	-0.061	-0.040
95th percentile	0.245	0.158	0.403	0.124	0.128	0.089	0.125	0.061	0.087	0.077
Two-tailed test p-value	0.14	0.14	0.60	0.28	0.19	0.05	0.42	0.23	0.74	0.33
Panel C: Clustered Standard Errors										
RI recriminalization	0.226*** (0.018)	-0.109*** (0.008)	0.020* (0.008)	0.003 (0.011)	0.058*** (0.008)	0.055*** (0.003)	0.060*** (0.008)	-0.237*** (0.009)	0.131*** (0.005)	0.026*** (0.005)
Panel D: Permutation Tests										
RI recriminalization	0.226	-0.109	0.020	0.003	0.058	0.055	0.060	-0.237	0.131	0.026
5th percentile	-0.255	-0.077	-0.148	-0.140	-0.088	-0.045	-0.135	-0.067	-0.041	-0.046
95th percentile	0.201	0.098	0.124	0.168	0.066	0.046	0.111	0.117	0.079	0.069
Two-tailed test p-value	0.09	0.09	0.74	1.02	0.23	0.14	0.37	0.05	0.05	0.60
Observations	159467	159805	159805	159805	159805	159805	159805	159805	159805	159805
State and year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

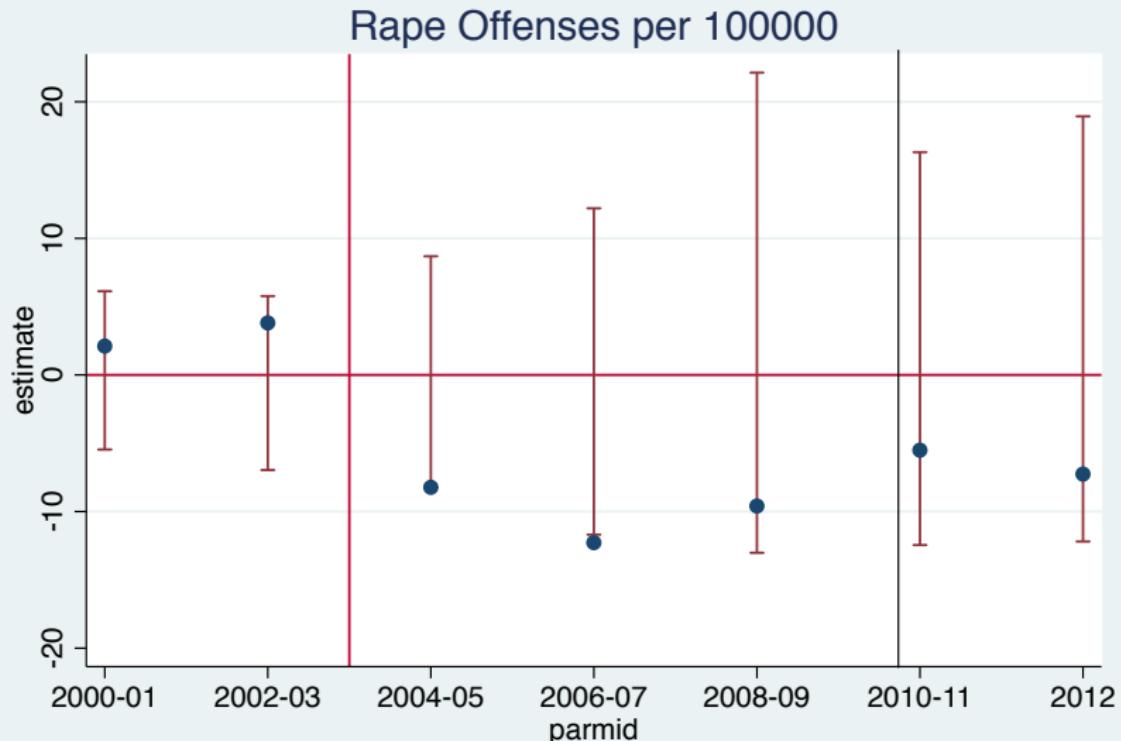
These are DD regressions using 1999-2012 The Erotic Review data. Individual controls include a dummy whether the provider was independent. Panels A and C present clustered standard errors and Panels B and D present 5th and 95th percentile confidence intervals from placebo-based inferential calculations, and p-values from a two-tailed test. * p<0.10, ** p<0.05, *** p<0.01

Recriminalization: Arrests

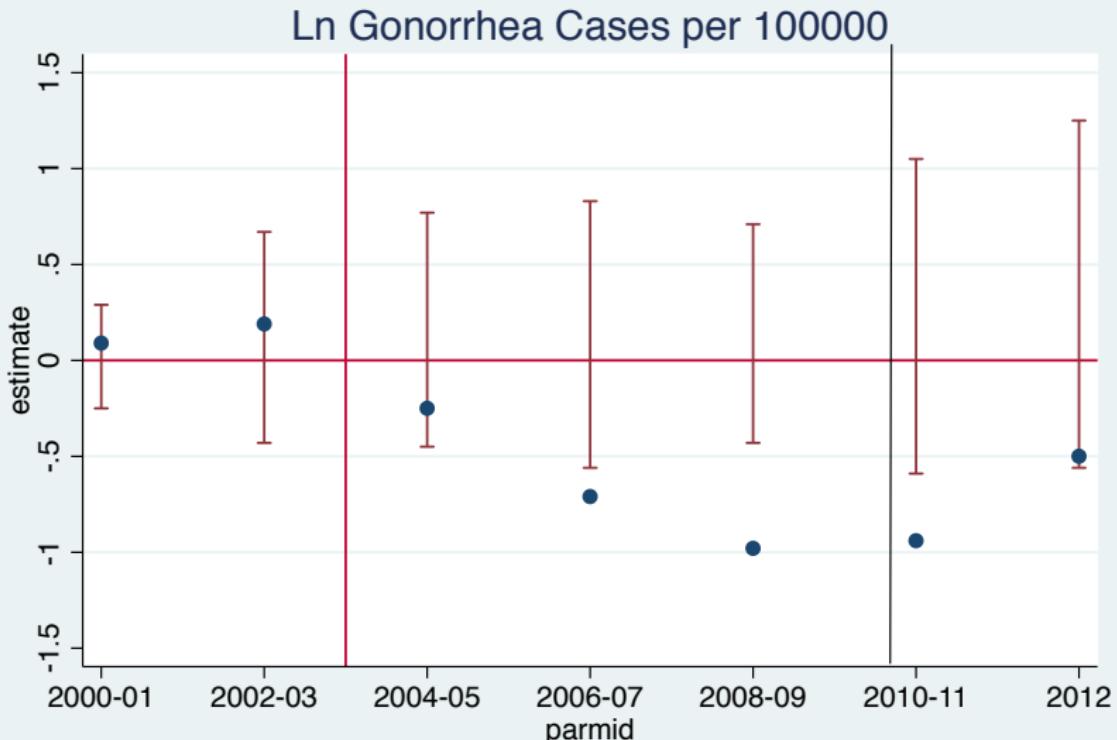


Dashed vertical bars are the sampling distribution of placebo estimates from 5th-95th percentile

Recriminalization: Rape Offenses



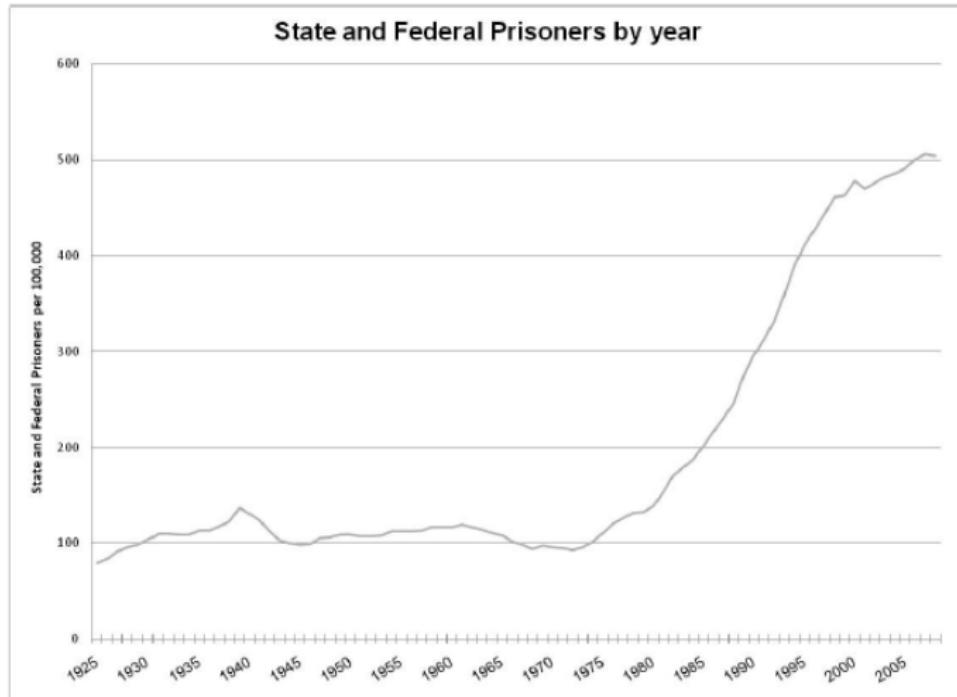
Recriminalization: Gonorrhea



Coding exercise

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

Figure 1
History of the imprisonment rate, 1925 - 2008



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

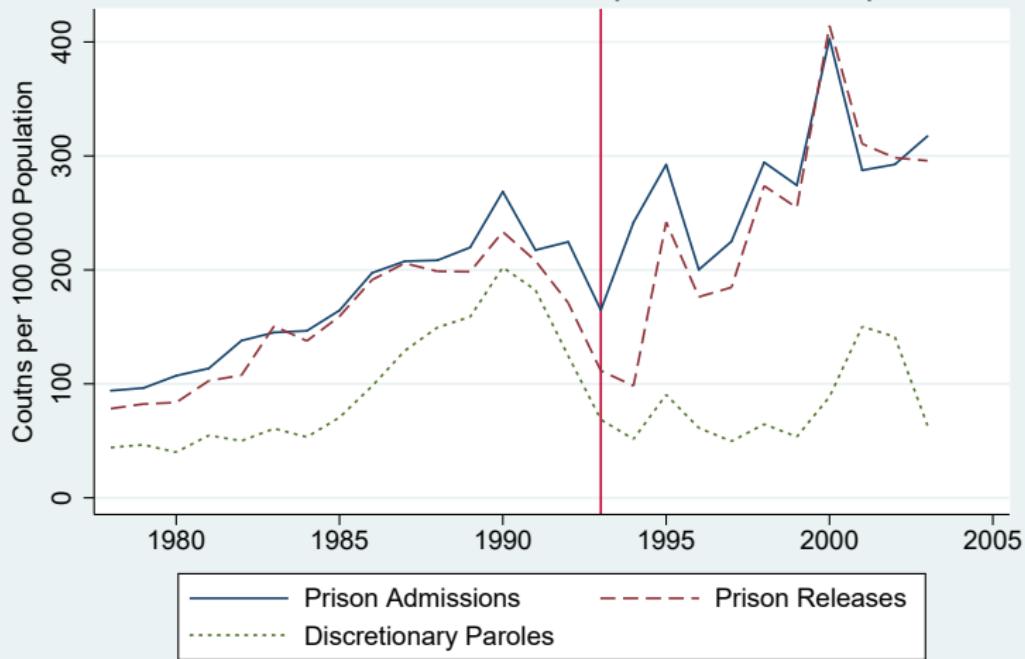
Prison constraints

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

Ruiz v. Estelle 1980

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

Texas Prison Flows Measures per 100 000 Population

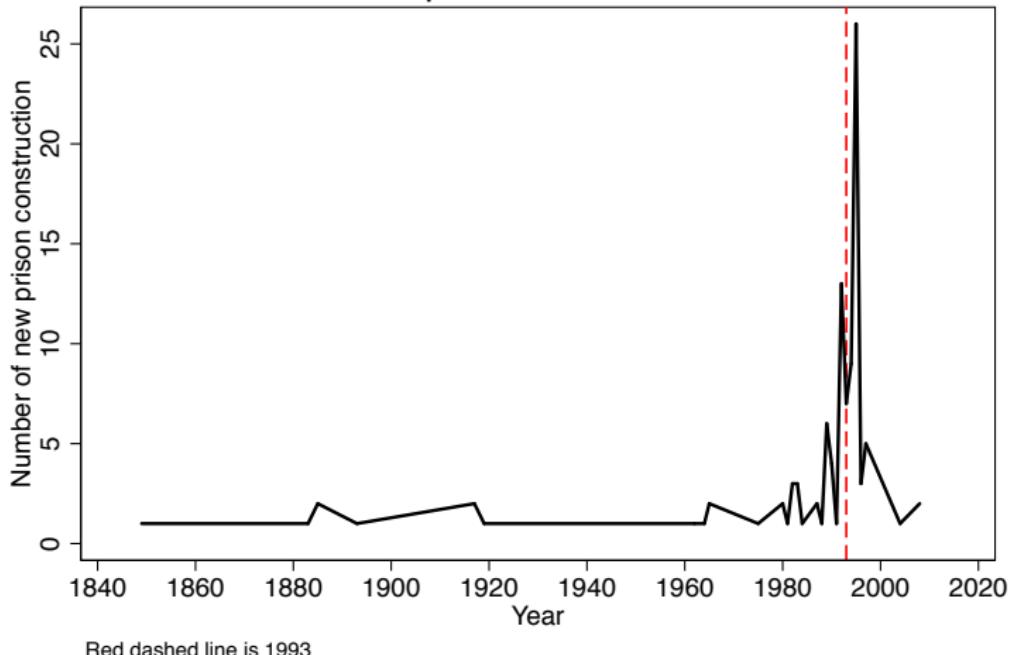


Texas prison boom

Governor Ann Richards (D) 1991-1995

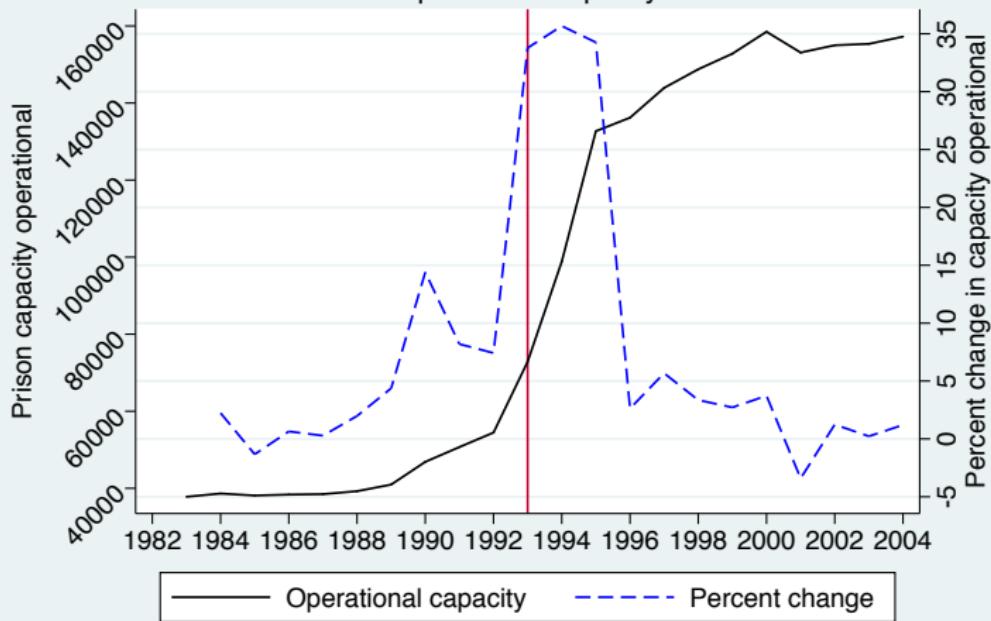
- Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
- Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
- Building of new prisons (private and public)

New prison construction

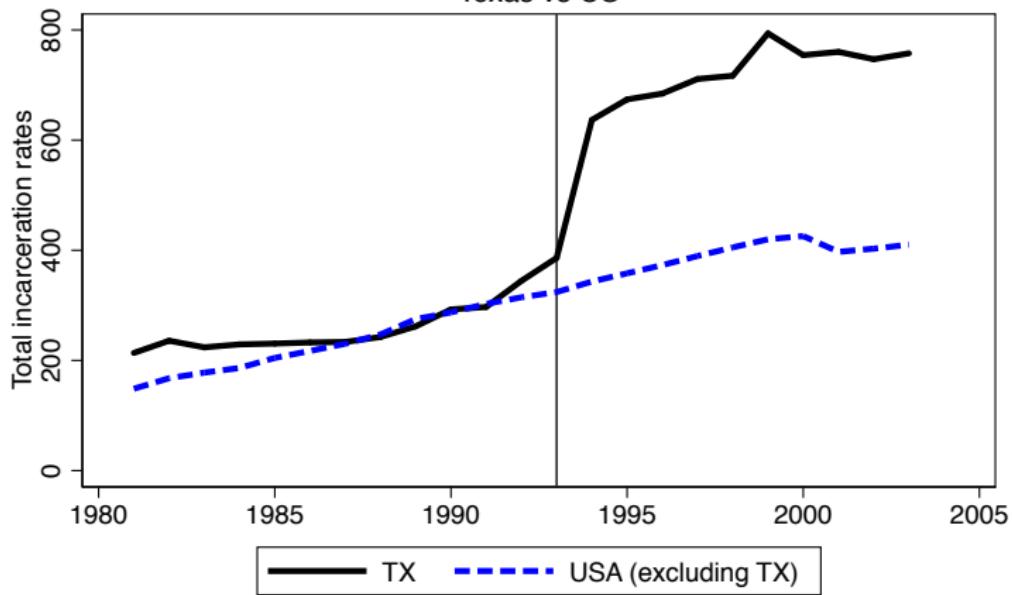


Texas prison growth

Operational capacity



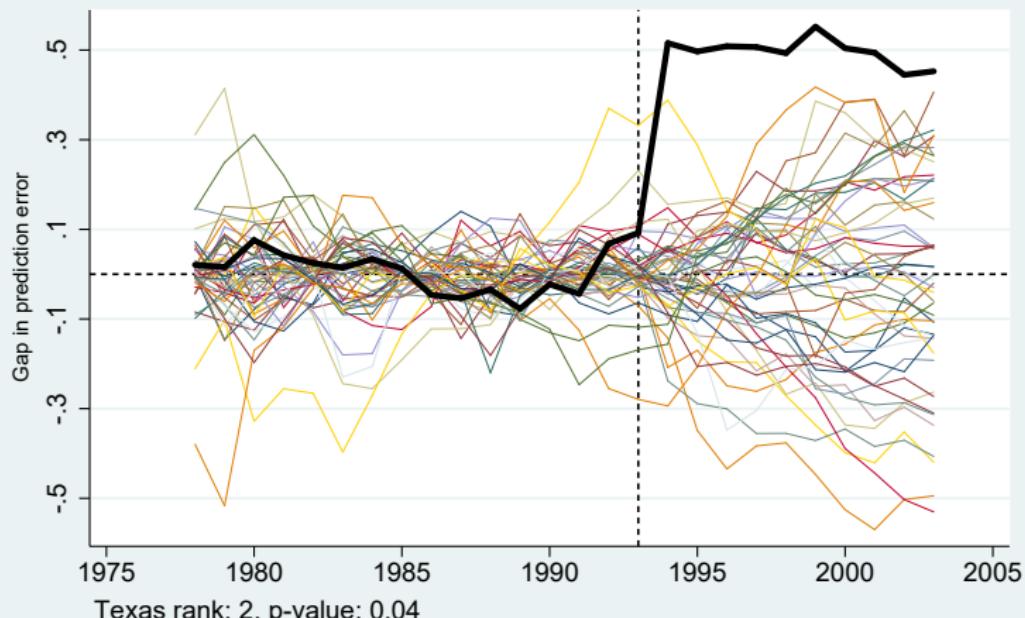
Total incarceration per 100 000 Texas vs US



1993 starts the prison expansion

Incarcerated persons per 100,000

1993 Treatment



Coding together

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

Last Comments

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

Regression vs synth weights

TABLE 1 Synthetic and Regression Weights for West Germany

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

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

Where to now?

- So synth is rising as we saw in the figure
- And developments in diff in diff maybe have lowered the *relative price* of using synth by making them not terribly different
- But the positive weights has been an issue for some who want extrapolation
- Let's look at that now

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 reduces 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

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

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:

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

$Y'_0 w^*$ (i.e., optimally weighted donor pool) is the unit 1 "synthetic control"

Predicting counterfactuals

Synth minimizes the following norm:

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

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:

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

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:

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

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 constraint.”

Convex hull

Synth minimizes the following norm:

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

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

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

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 $m(\cdot)$ are pre-treatment outcomes:

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

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

Bias correction

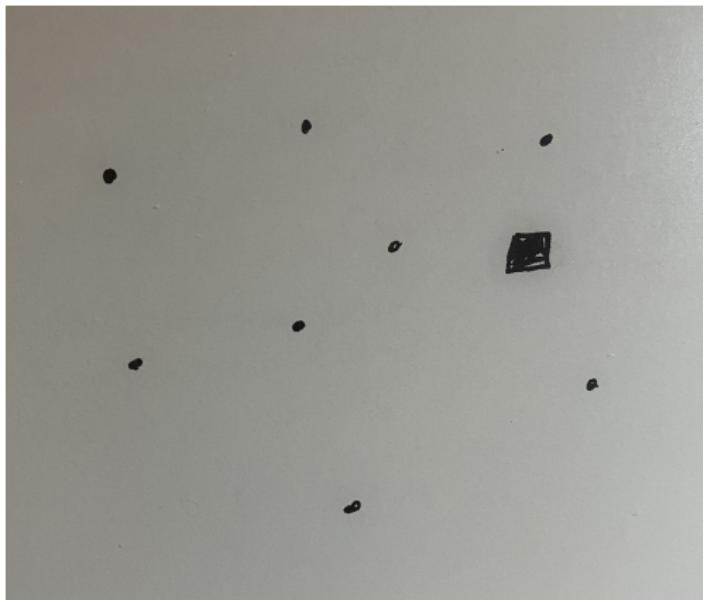
$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

- 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

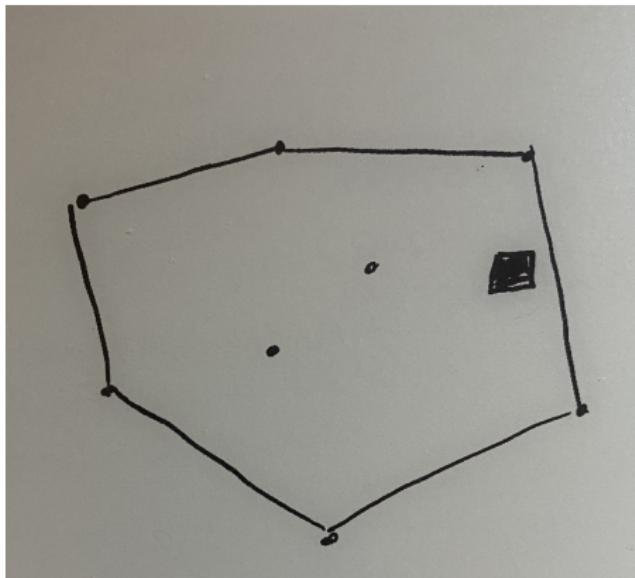
Common practice

- 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

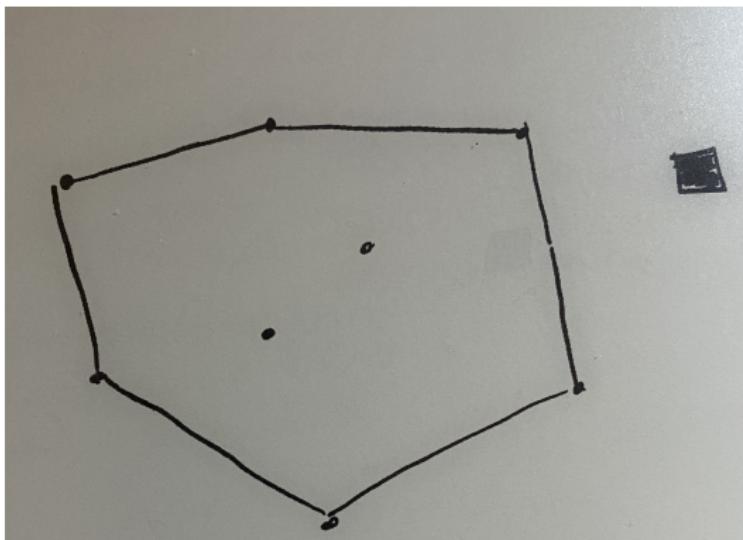
Treatment and control units



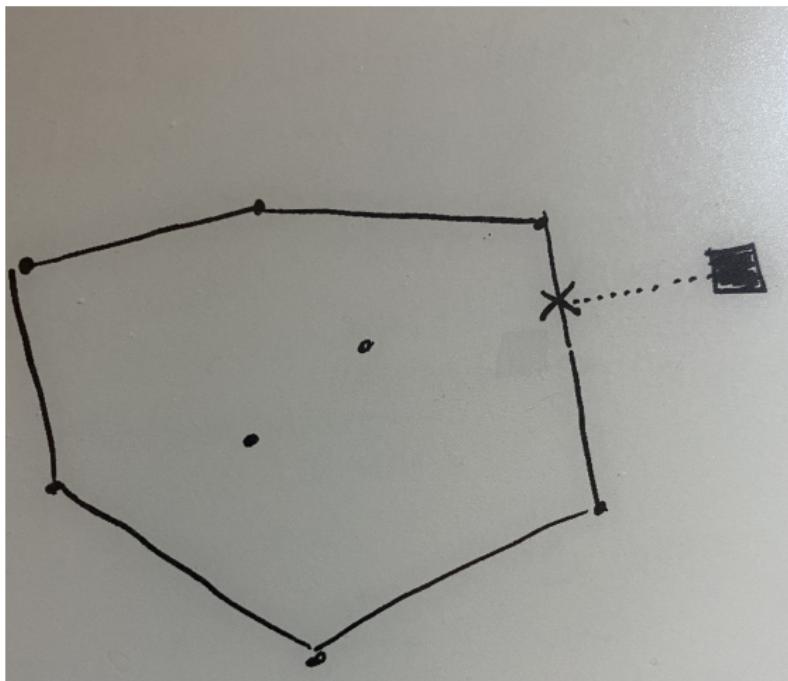
Convex hull – ideal for synth



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 \hat{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

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

Interpreting line 1

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

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

Interpreting line 1

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

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

Interpreting line 1

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

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

Interpreting line 1

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

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

Interpreting line 2

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

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

Interpreting line 2

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

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

Ridge Augmented SCM

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

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

Go back to that weighting but use the ridge parameters

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

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

The ridge weights are key to the augmentation

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

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

Ridge will allow negative weights via extrapolation

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

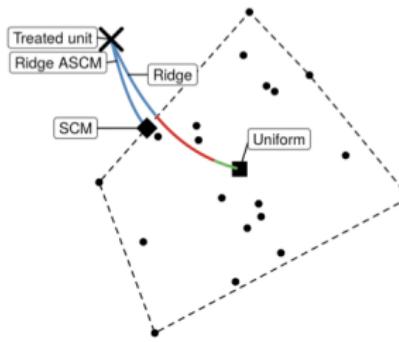
Relaxing the constraint from synth that weights be non-negative, as non-negative weights prohibit extrapolation. But we don't have synthetic control on the simplex, so we *must* extrapolate, otherwise synth will be biased.

Summarizing and some comments

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

Intuition

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



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

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

Conformal Inference

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)
 - 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

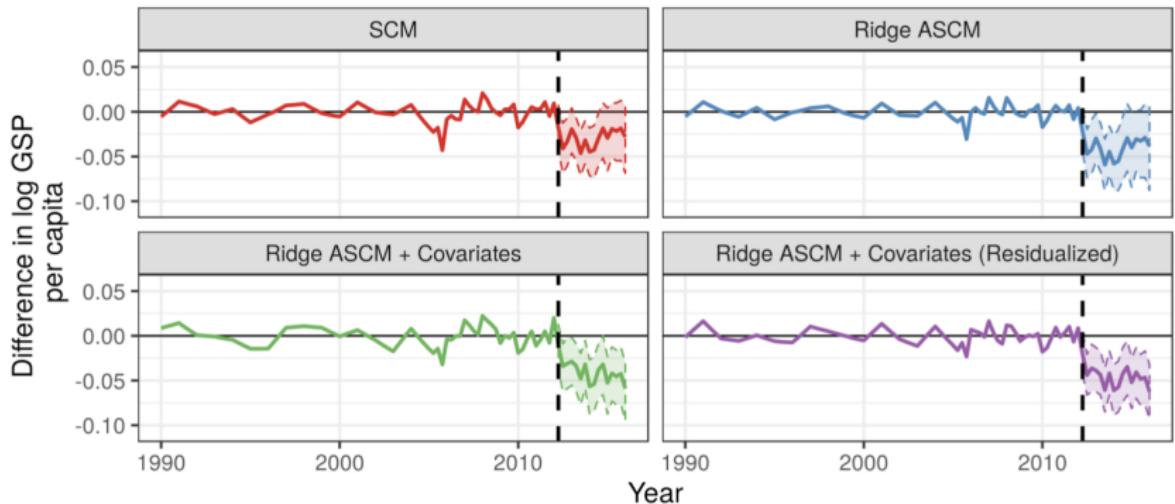
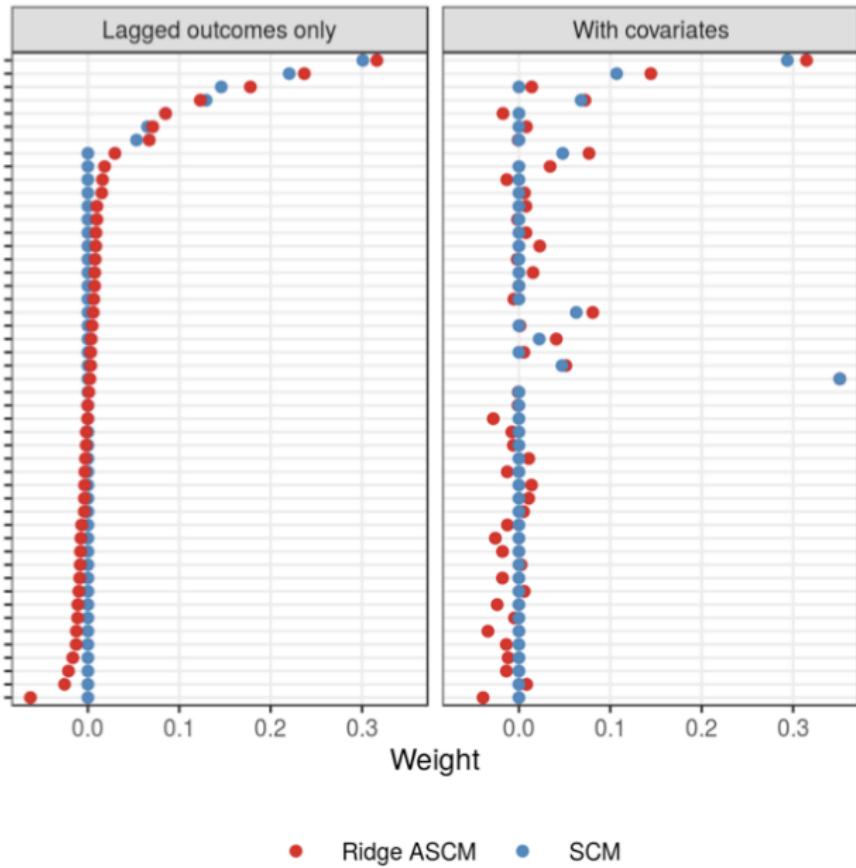


Figure 6: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.



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

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

Notice, this is the synthetic control design because a single unit (unit i) is missing Y^0 for the 3rd and t th 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 $N \gg T$, but fat matrices are ones where $T \gg N$
- 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

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

where the objective function is:

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

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

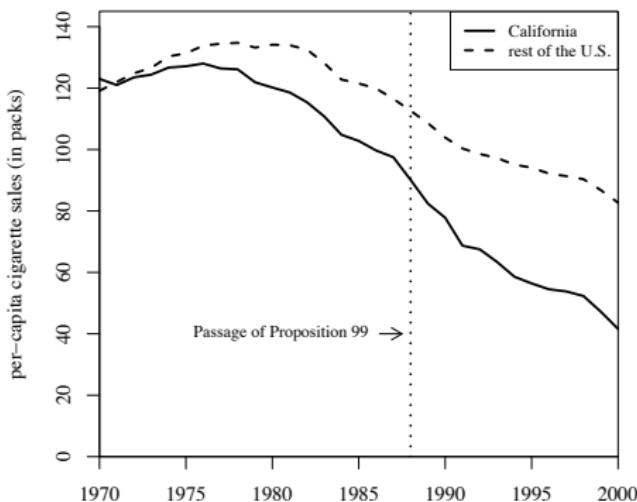
- Remember what Athey and Imbens said – “most important innovation in causal inference of the last 15 years”.
- The synthetic DiD bears some similarities to their MCNN model, but focuses on estimating weights, not the L^* matrix
- It will dominate the Abadie, Diamond and Hainmueller (2010) as they will 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, \hat{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, $\hat{\lambda}$, 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 \hat{w} defining a synthetic control unit (just like Abadie, Diamond and Hainmueller 2010) using the pre-treatment data

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

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 $\hat{\lambda}$ defining a synthetic pre-treatment period using control data

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

Estimation

4. Compute the SDID estimator via the weighted DID regression

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

Estimating the weights

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

5. Estimate the control weights, \hat{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. $\hat{\lambda}$, 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

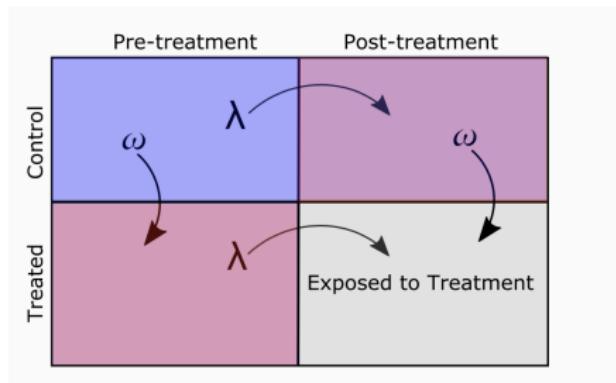
$$Y = L + \tau D + E$$

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:

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

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

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

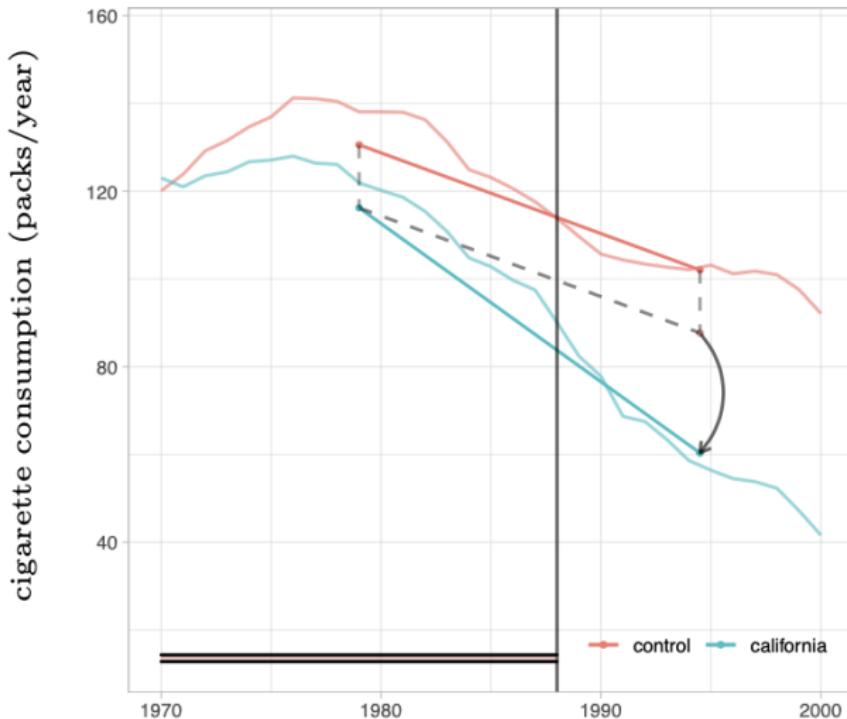
- SDiD is weighted regression with unit FEs and time FEs:

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

Formal results overview

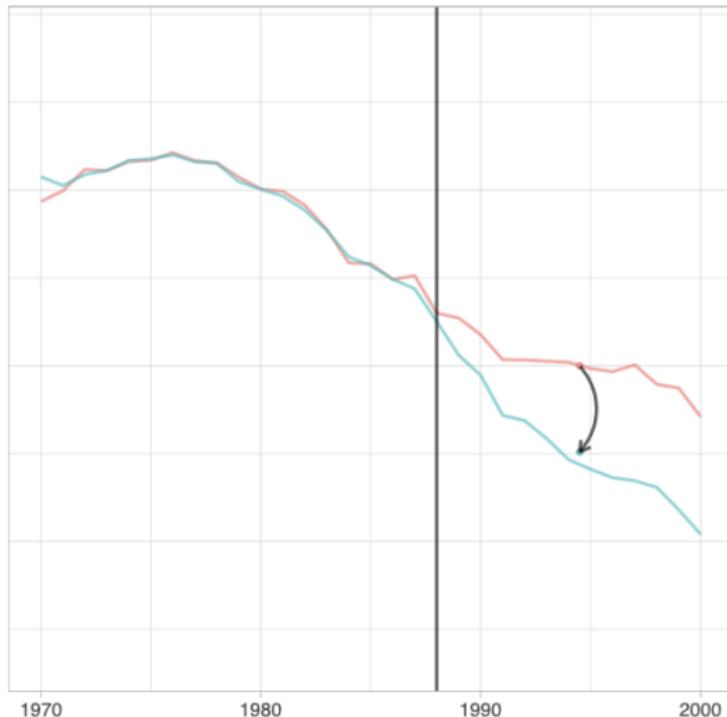
- Formal results will show SDiD is “doubly robust” (recall Sant’Anna and Zhao 2020)
- Factor model on the outcome can be a latent factor model but true model is that signal model and it’ll still be consistent
- Asymptotic normality of $\hat{\tau}^{SDiD}$
- With oracle weights, SDiD will have “good weights”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

Difference in Differences



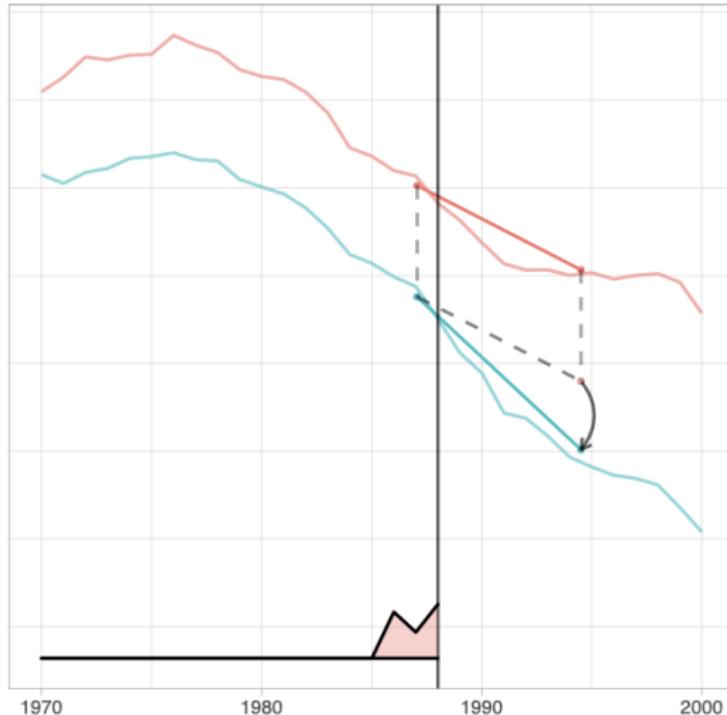
Estimated decrease: -27.3 (17.7)

Synthetic Control



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

Synthetic Diff. in Differences



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

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

So they characterize these terms

Oracle noise

First term: the oracle noise

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

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)

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

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)

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

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 SDID:

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

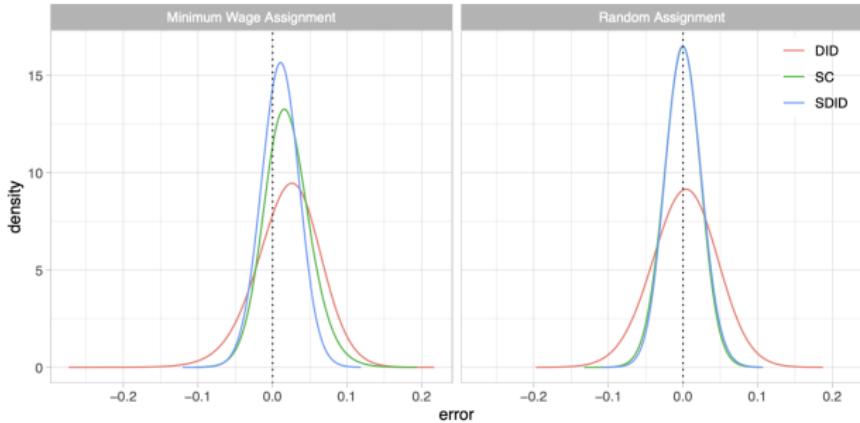


Figure 2: Distribution of the errors of SDID, SC and DID in the setting of the “baseline” (i.e., with minimum wage) and random assignment rows of Table 2.

	RMSE					Bias				
	SDID	SC	DID	MC	DIFFP	SDID	SC	DID	MC	DIFFP
Baseline	0.28	0.37	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
<i>Outcome Model</i>										
No Corr	0.28	0.38	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
No \mathbf{M}	0.16	0.18	0.14	0.14	0.16	0.01	0.04	0.01	0.01	0.01
No \mathbf{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
<i>Assignment Process</i>										
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
<i>Outcome Variable</i>										
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
<i>Assignment Block Size</i>										
$T_{\text{post}} = 1$	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 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

Variable	All MSAs			Treated MSAs			Untreated MSAs		
	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

	<i>Dependent Variable : Weekly First Doses per 100,000</i>		
	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

	<i>Dependent Variable : Weekly COVID-19 Cases per 100,000</i>		
	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)

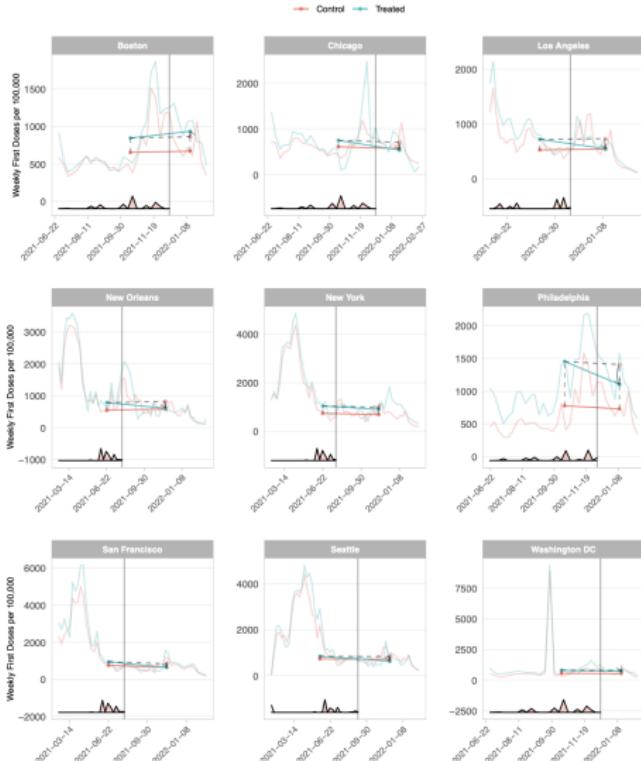
Notes: This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 cases per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ($\hat{\tau}$ from equations (2), (3), and (1)). Also 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

	<i>Dependent Variable : Weekly COVID-19 Deaths per 100,000</i>		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
<i>Panel A. Boston</i>			
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)
<i>Panel B. Chicago</i>			
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)
<i>Panel C. Los Angeles</i>			
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)
<i>Panel D. New Orleans</i>			
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)
<i>Panel E. New York</i>			
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)
<i>Panel F. Philadelphia</i>			
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)
<i>Panel G. San Francisco</i>			
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)
<i>Panel H. Seattle</i>			
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)
<i>Panel I. Washington DC</i>			
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)

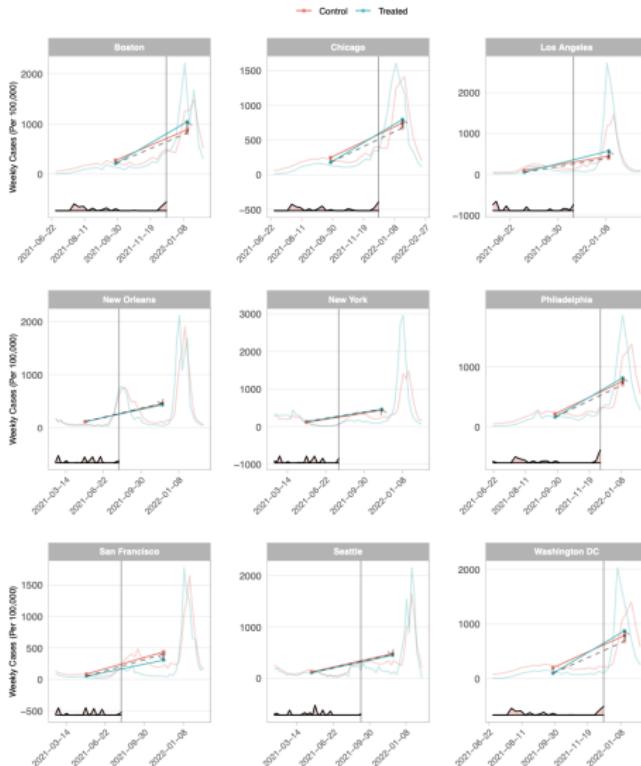
Notes: This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 deaths per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ($\hat{\tau}$ from equations (2), (3), and (1)). Also 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



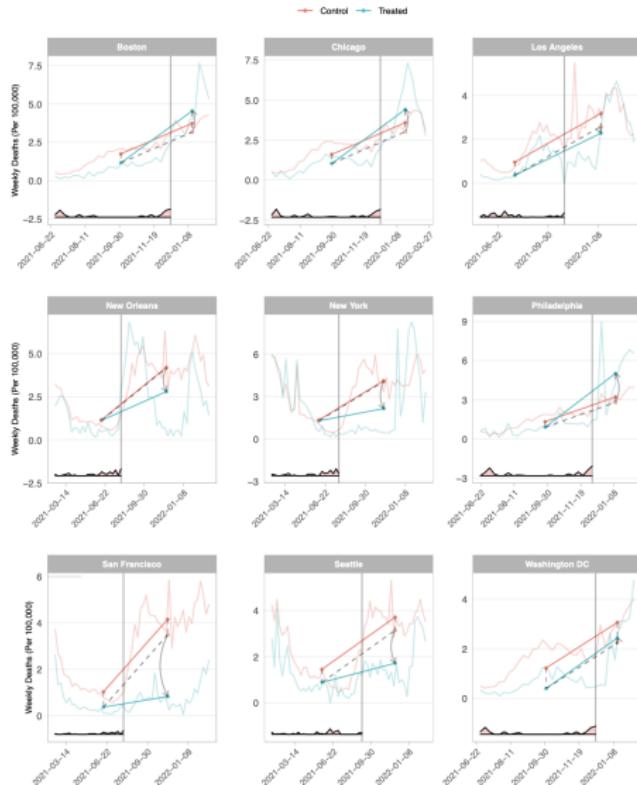
Notes: Each plot shows trends in weekly first doses of COVID-19 vaccinations per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved 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{f} 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?