Causal Inference III

Mixtape Session

Prof. Scott Cunningham

Roadmap

- Synthetic Control
    - Original synthetic control method
    - Augmented Synthetic Control
    - Synthetic difference-in-differences

First outline

1. **Abadie's synth**: We will start here as the foundation, understand it inside and out, practice with it, and try to learn good practices, as well as its bias
2. **Augmented synth**: Bias correction to address imbalance bias
3. **Synthetic diff-in-diff**: Further modifications of the biased synth, combines diff-in-diff and synth

Synthetic Control

- Abadie and Gardeazabal (2003) introduced synthetic control in the AER in a study of a terrorist attack in Spain (Basque Country) 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

- Synthetic control is a constrained minimization problem where the target goal is the minimization of a vector of squared gaps in pre-treatment characteristics
- Choice vector is an endogenous weights that are constant per control group unit over time and range from [0,1).
- 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 non-negatively weighted average of donor pool units that on closely resemble 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 $1, 2, \ldots, T$

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

Potential outcomes notation

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

Group-time ATT with only one treated group

Treatment effect parameter is defined as dynamic ATT where

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

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

Optimal weights

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

Optimal weights differ by another weighting matrix

Abadie, et al. consider

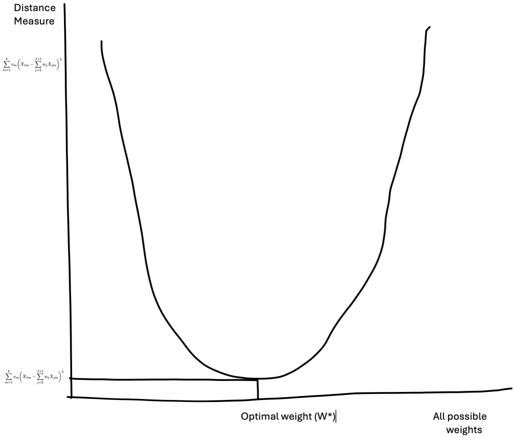
$$||X_1 - X_0W|| = \sqrt{(X_1 - X_0W)'V(X_1 - X_0W)}$$

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

Similarity to distance minimization

- Bears some resemblance to nearest neighbor matching though I don't want to oversell that
- There is a unique solution that selects weighting minimizing the distance between the covariates comparison characteristics and treatment group
- Let's look at an example from nearest neighbor matching of minimizing Euclidean distance to help firm the ideas
- https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY_LJVzOYvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing

Illustration of the ”optimal weight”



More on the V matrix

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

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

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

Choice of V is critical

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

Estimating the V matrix

Choice of $v_1, \ldots, v_k$ can be based on

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

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

Cross validation

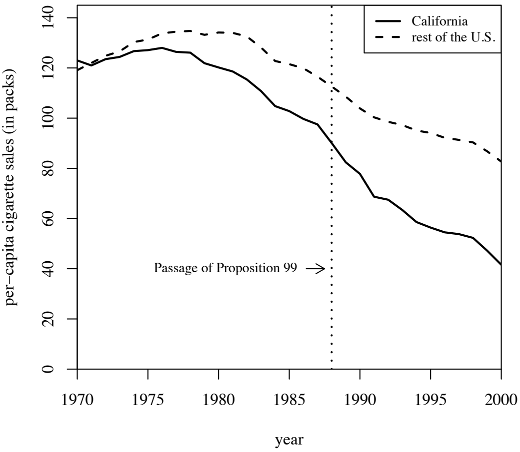
- Abadie recommends cross validation for selecting the covariates
- Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
- For any given V, calculate $W^*(V)$ in the training period.
- Minimize the MSPE of $W^*(V)$ in the validation period

Let's look at an example from the 2010 JASA paper with Hainmueller and Diamond

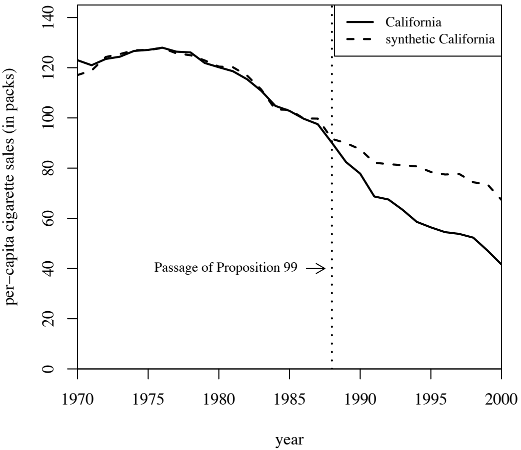
Example: California's Proposition 99

- 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



Sparsity and Synthetic Control Weights

Table 2. State weights in the synthetic California

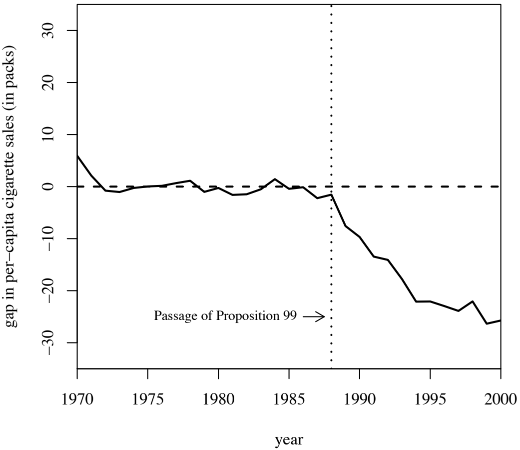
State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Predictor Means: Actual vs. Synthetic California

Variables	California 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



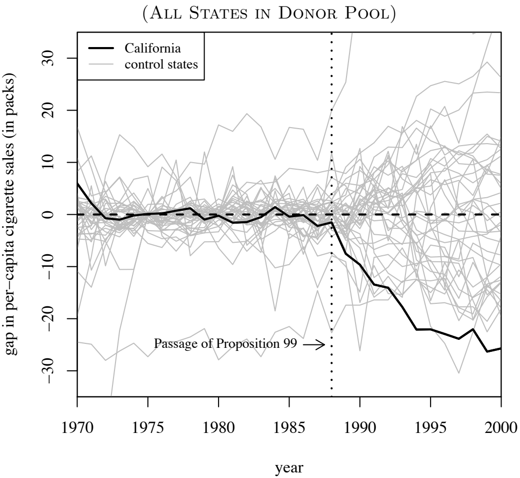
Inference

- 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:

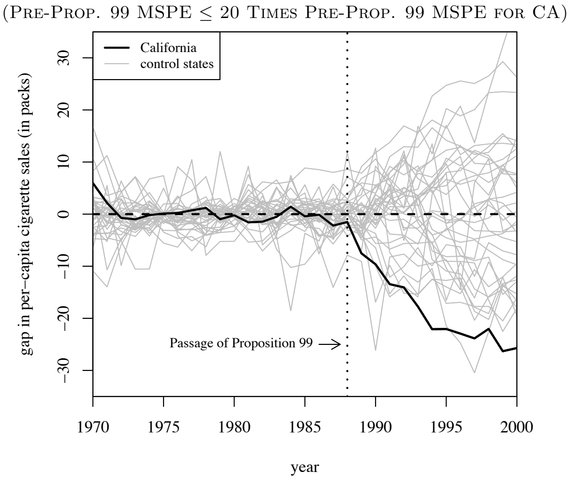
$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

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

Smoking Gap for CA and 38 control states

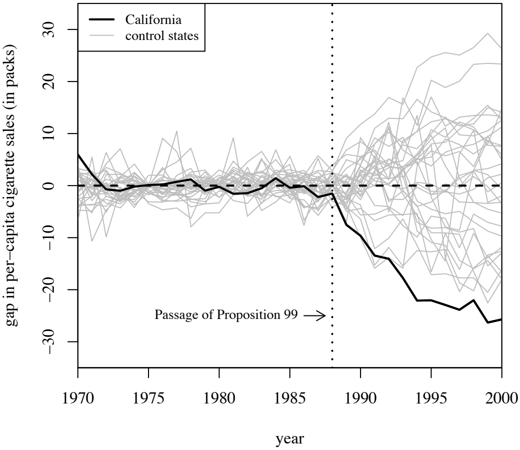


Smoking Gap for CA and 34 control states

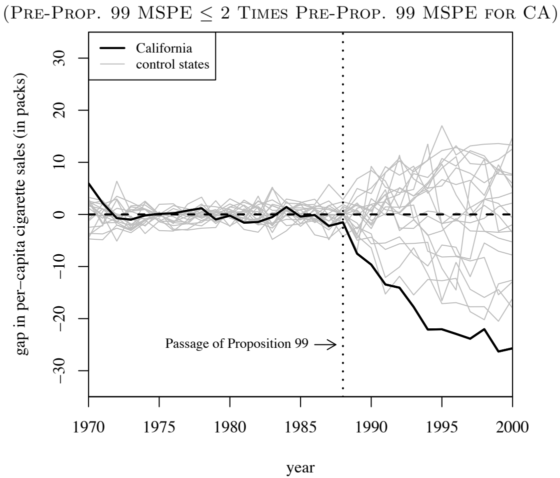


Smoking Gap for CA and 29 control states

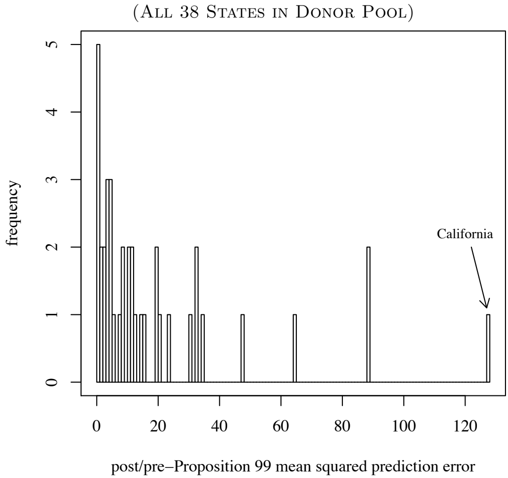
(PRE-PROP. 99 MSPE $\leq$ 5 TIMES PRE-PROP. 99 MSPE FOR CA)



Smoking Gap for CA and 19 control states



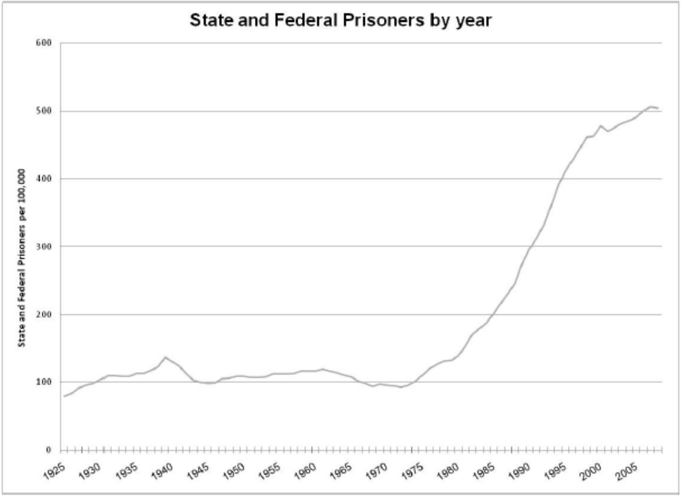
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Replication 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



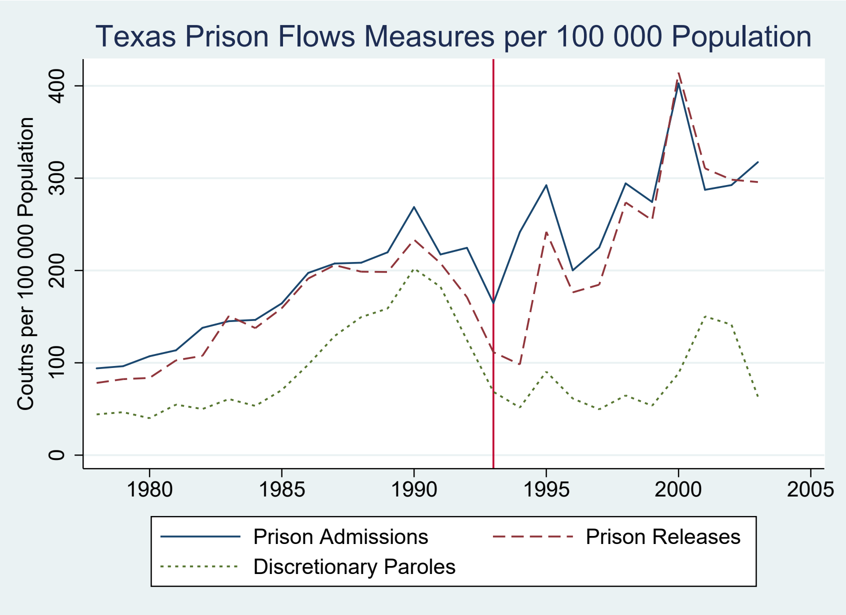
Source: www.albany.edu/sourcebook/tost_6/html

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

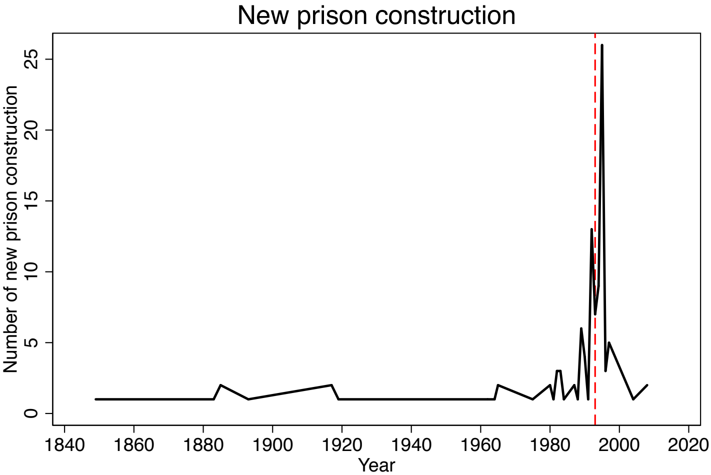


Texas Prison Flows Measures per 100 000 Population
Coutns per 100 000 Population
0
100
200
300
400
1980
1985
1990
1995
2000
2005
Prison Admissions
Prison Releases
Discretionary Paroles

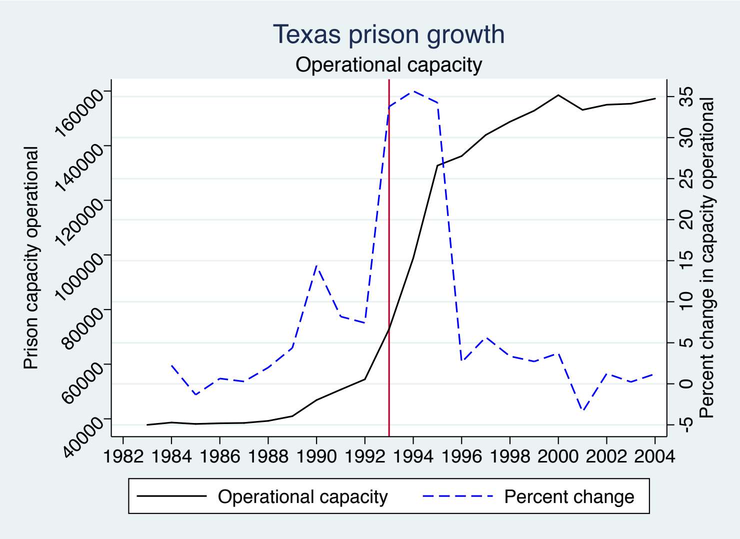
Texas prison boom

Governor Ann Richards (D) 1991-1995

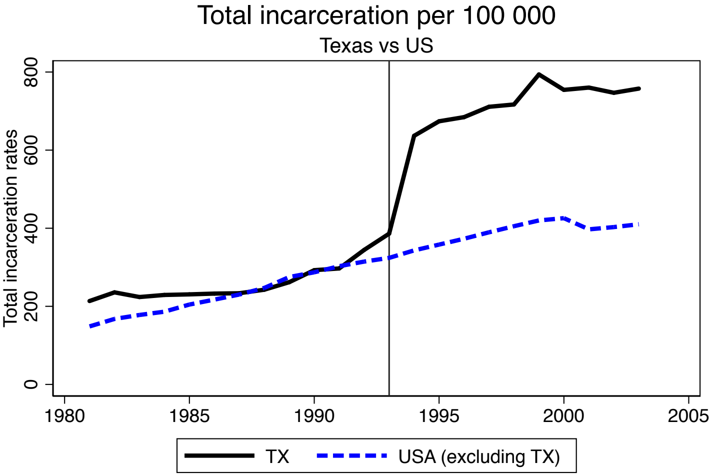
- Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
- Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
- Building of new prisons (private and public)



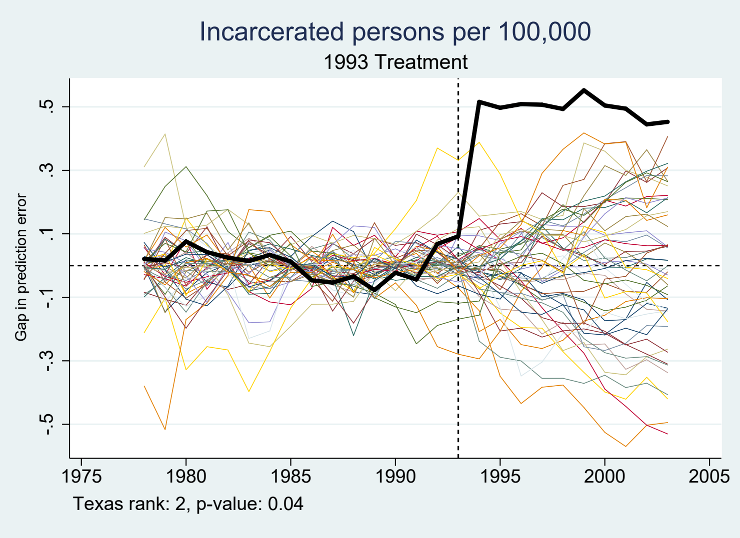
Red dashed line is 1993



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



1993 starts the prison expansion


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

Coding together

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

Summarizing

- Method is simple: take donor pool units and find a combination of units, weight them, that minimize a distance function subject to two weight constraints $W*$ and $V*$.
- But what is the bias and what can we do when we think the bias is severe?
- We now move towards that now by examining the nature of the bias in synthetic control

Imperfect Fit

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

Augmenting synthetic control

1. Original synthetic control needs perfect fit and so will be biased in practical settings as it won't be the case we get weights constrained to be on the convex hull
2. Augmentation of the synthetic control estimator uses an outcome model to estimate the bias caused by covariate imbalance
3. Outcome model is a penalized ridge regression which will provide new weights we use to reweight the original synthetic control (bias adjustment in the spirit of Abadie and Imbens (2011)

Augmenting synthetic control

4. When synth is imbalanced, augmented synth will reduce bias reweighting and bias correction, and when synth is balanced, they are the same
5. When synth is balanced, the augmented and original synth are identical (but in practice, they won't be identical)
6. 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

Optimal weights

Synth minimizes the following norm:

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

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

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

Predicting counterfactuals

Synth minimizes the following norm:

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

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

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

Penalizing the weights with ridge

Synth minimizes the following norm:

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

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

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

Convex hull

Synth minimizes the following norm:

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

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

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

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

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

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

Slight change in synth notation

- Assume that our outcome, Y_{jt}^0, follows a factor model where $m(.)$ are pre-treatment potential outcomes:

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

- Since $\widehat{m(.)}$ estimates the post-treatment outcome, it can be viewed as an estimate of matching bias
- Procedure then becomes analogous to bias correction for inexact matching (Abadie and Imbens 2011)

Bias correction

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

- Recall from earlier by Abadie, et al. (2010) and Ferman and Pinto (2021) the same point made which is that as T grows, the synthetic control achieves balance, not by fitting on the idiosyncratic noise (which is on average zero in large samples), but on the unobserved heterogeneity in the factor model
- Thus when the weights do achieve exact balance, the bias of synthetic control decreases with T

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 even with large T

Estimating the bias

- Adjust the synthetic control to adjust for poor fit pre-treatment with an estimate of the "matching bias" for both the treatment group and the weighted average donor pools (Abadie and Imbens 2011)
- We will use our outcome regression model, $\widehat{m}_{jT}$, to estimate the post-treatment potential outcome Y^0_{jT} which is recall unobserved for treatment group
- So there is in other words two steps involved: estimate the synthetic control finding optimal donor pool weights, then estimate the matching bias using ridge regression and adjust, similar to bias correction in Abadie and Imbens 2011

Setup of the estimator

$Y_1^{aug,0}$ is the augmented potential outcome based on synthetic control (first term) and adjustments for matching imbalance (second and third terms):

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

Interpreting line 1

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

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

Interpreting line 1

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

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

Interpreting line 1

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

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

Interpreting line 2

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

(2) The second line is a double robust representation of the estimator equal to a regression based outcome model plus weighted residuals

Form of outcome model

- To estimate $\widehat{m}$, recall it is an extrapolation of Y^0 based on covariates X (ignoring subscripts)
- But this can have overfitting problems, so they introduce a ridge regression
- Issue will be with regards to the hyperparameter so they'll suggest cross validation
- It's inside this second "reweighting" stage (or bias adjustment) that the negative weighting comes and it comes from the outcome model itself extrapolating

Ridge Augmented SCM

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

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

Ridge Augmented SCM

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

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

Go back to that weighting but use the ridge parameters

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

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

The ridge weights are key to the augmentation

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

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

Ridge will allow negative weights via extrapolation

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

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

Summarizing and some comments

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

Augmented synth vs original synthetic control

- 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

Replication

- We will now show code that does it in R and Stata
  - → R: ”augsynth” by the authors
  - → Stata: ”allsynth” that does several (including augsynth)
- Two examples to hopefully illustrate the bias and improvements and the lack of bias and no change in another
  - → Smoking: augmented synth improves it
  - → Prisons: augmented synth does nothing

The Basic Idea of SDID

- Combines ideas from synthetic control and difference-in-differences.
  - → Matches control units to treated units with pre-treatment weights.
  - → Addresses biases from imperfect matching.
- Introduces time weights to prioritize key pre-treatment periods.
- Balances precision from synthetic control with broader trends from DID.

How SDID Works

- Uses unit and time fixed effects for bias reduction.
  - → Accounts for unit-specific traits (e.g., state-level differences).
  - → Captures common period effects (e.g., economic booms).
- Derives time weights directly from the data.
- Flexible tool for causal inference in complex panel data.

Intuition Behind SDID

- Approximates parallel trends by reweighting data.
  - → Unit weights align control and treated units pre-treatment.
  - → Time weights balance trends across periods.
- Reduces reliance on strong assumptions about pre-existing trends.
- Enhances robustness when pre-existing trends don’t align perfectly.

Assumptions in Synthetic Difference-in-Differences (SDID)

In both synthetic control and difference-in-differences, assumptions about pre-treatment trends are central. SDID combines these methods to leverage their strengths:

- Uses pre-treatment trends to reweight treated and control groups.
- Constructs a comparison group that mimics the treated group's trends.
- Relies on four key assumptions for identification.

Assumption 1: Error Properties

- Errors are independent, identically distributed Gaussian vectors.
- Covariance matrix has bounded eigenvalues.
- Assumption is stricter than traditional DiD or synthetic control, which only assume error independence.

Assumption 2: Sample Sizes

- Large sample sizes are critical for SDID:
  - $\rightarrow$ Number of control units (N_{co}) and pre-treatment periods (T_{pre}) must be large.
  - $\rightarrow$ Product of treated units (N_{tr}) and post-treatment periods (T_{post}) should also be large.
- Balance between T_{pre} and N_{co} ensures effective reweighting.

Assumptions 3 and 4

Assumption 3: Systematic Component Properties

- The systematic component (L) has a limited number of large singular values.
- This low-rank structure captures broad patterns while reducing noise.
- Goes beyond traditional DiD and synthetic control.

Assumption 4: Weighting Properties and L

- "Oracle" weights derived from L reduce systematic bias.
- Ensures parallel trends after reweighting without requiring perfect balance.

Steps in SDID Estimation

- Identify unit weights for control units to match treated pre-treatment trends.
- Select time weights to balance pre- and post-treatment trends.
- Combine weights in regression to estimate the ATT.

SDID Estimation Process

- Calculates unit and time weights to optimize comparison groups.
- Uses regularization to prevent overfitting.
  - $\rightarrow$ Aligns with one-period outcome changes for untreated units.
- Solves constrained least-squares for unit weights ($\hat{w}$).

Estimation of SDiD

Synthetic DiD combines elements of DiD and synthetic control:

- Compute the regularization parameter to align with typical one-period outcome changes:

$$\Delta_{it} = Y_{i(t+1)} - Y_{it} \quad \text{(unexposed units).}$$

Unit and Time Weights

- **Unit Weights ($\hat{w}$):**
  - → Match treated and control units pre-treatment.
  - → Allows for a constant gap, relaxing traditional SC constraints.
- **Time Weights ($\hat{\lambda}$):**
  - → Balance temporal dynamics.
  - → Prioritize informative pre-treatment periods.

Estimating Weights

- Unit Weights:
  - → Estimated via constrained least squares on pre-treatment data.
  - → Weights are non-negative, sum to one, and allow for a level shift with regularization.
- Time Weights:
  - → Estimated via constrained least squares on control data.
  - → Prioritizes periods just before treatment without spreading weights broadly.

Estimation of SDiD: Unit Weights

- Estimate unit weights $\widehat{w}$ to define a synthetic control unit.
- Equation:

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

- Unlike traditional synthetic control, allows for an intercept term, relaxing perfect matching requirements.

Estimation of SDiD: Time Weights

- Estimate time weights $\widehat{\lambda}$ to align synthetic pre-treatment periods:

$$\widehat{\lambda}_1 + Y_{1,\text{pre}}\widehat{\lambda} \approx Y_{1,\text{post}}$$

- Focuses on identifying pre-treatment periods most informative for post-treatment behavior.

Combined Estimation

- SDID estimator combines unit and time weights.
- Weighted regression minimizes residuals across units and time.
- Ensures counterfactual outcomes align with treated outcomes pre-treatment.

Outcome Model in SDID

- Observed outcomes (Y) consist of:
  - $\rightarrow$ Systematic component (L).
  - $\rightarrow$ Treatment effect $(D \circ \delta)$.
  - $\rightarrow$ Idiosyncratic error (E).
- $D \circ \delta$ isolates ATT by selecting treated units and periods.
- L captures trends not explained by treatment or noise.

Oracle Weights

- Theoretical ”ideal” weights for unit ($\tilde{\omega}$) and time ($\tilde{\lambda}$).
- Minimize bias by balancing systematic component (L).
- Data-driven weights ($\hat{\omega}, \hat{\lambda}$) approximate oracle weights.
  - $\rightarrow$ Ensure pre-treatment balance.
  - $\rightarrow$ Converge to oracle weights as sample size increases.

Outcome Model

- Outcome is a combination of systematic trends, treatment effects, and noise:

$$Y = L + D \circ \delta + E$$

- Components:
  - → L: systematic trends across units and time ($L = \Gamma\Upsilon^\top$)
  - → $D \circ \delta$: treatment effects applied selectively
  - → E: idiosyncratic noise
- L represents shared patterns unaffected by treatment or randomness.

Role of Systematic Component L

- Captures trends unrelated to treatment or random noise.
- Potential source of confounding if not properly controlled.
- Treated and control groups must balance L pre-treatment for valid comparison.

Oracle Weights

- Theoretical weights $(\tilde{\omega}, \tilde{\lambda})$ minimize bias in estimating treatment effects.
- Balance systematic trends (L) across treated and control groups.
- Provide a benchmark for robust causal inference.

Data-Driven Weights

- Empirical weights $(\hat{\omega}, \hat{\lambda})$ approximate oracle weights.
- Constructed to balance treated and control trends pre-treatment.
- Approximation improves with larger sample sizes.

Weighted Regression for SDiD

- Final estimation uses weighted DID regression:

$$\text{argmin}_{\tau,\mu,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \, \widehat{w}_i^{\text{SDiD}} \widehat{\lambda}_t^{\text{SDiD}}$$

Regression Comparison: SC, DiD, SDiD

- Synthetic Control (SC):

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

- Difference-in-Differences (DiD):

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

- Synthetic Difference-in-Differences (SDiD):

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

Key Features of SDiD

- Combines synthetic control's matching with DID's parallel trends adjustment.
- Introduces unit and time weights for improved balance.
- Does not require perfect pre-treatment matching or strictly parallel trends.

Decomposing the Bias of SDiD

$$\widehat{\tau}^{sdid} - \tau = \varepsilon(\widetilde{w}, \widetilde{\lambda}) + B(\widetilde{w}, \widetilde{\lambda}) + \widehat{\tau}(\widehat{w}, \widehat{\lambda}) - \widehat{\tau}(\widetilde{w}, \widetilde{\lambda})$$

- **Oracle noise**: Variance introduced by weights and sample limitations.
- **Oracle confounding bias**: Systematic differences not removed by weights.
- **Deviation from oracle**: Approximation errors in empirical weights.

Oracle Noise

- Noise arises from random variation in data.
- Small when weights $(\hat{w}, \hat{\lambda})$ are small and sample sizes are sufficient.
- Ensures noise does not dominate the estimator.

Oracle Confounding Bias: Units and Time

- **Units (Rows)**:

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

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

  Ensures weights remove bias from systematic differences across units.

- **Time (Columns)**:

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

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

  Captures temporal dynamics to align treated and control groups.

Doubly Robust Property

- Sufficient for one set of weights (unit or time) to generalize well.
- Combination of oracle unit and time weights can compensate for failures in one dimension.
- Provides resilience against systematic confounding.

Deviation from Oracle

- SDID approximates oracle weights when:
  - → Oracle weights generalize well on training sets.
  - → Regularization is appropriately tuned.
- Ensures SDID estimator is close to theoretical benchmark.

Practical Implications

- Combines theoretical rigor with empirical flexibility.
- Balances systematic trends in pre-treatment data.
- Achieves robust causal inference for ATT estimation.

Practical Considerations: Pre-treatment Trends

- Visually inspect pre-treatment trends to check for alignment between treated and control groups.
- Use plots to ensure parallel trends are approximately valid before treatment.
- Address any unusual patterns or discrepancies early.

Practical Considerations: Weights and Assumptions

- Balanced weights:
  - → Ensure $\hat{\omega}$ and $\hat{\lambda}$ are not overly concentrated.
  - → Adjust regularization parameters if needed.
- Assess parallel trends:
  - → Contextual knowledge remains crucial.
  - → Account for potential confounders.

Practical Considerations: SEs and Heterogeneous Effects

- Standard errors:
  - → Select bootstrap or other approaches suited to your data structure.
  - → Be cautious with small treated samples.
- Heterogeneous effects:
  - → Consider if ATT is the correct focus.
  - → Explore alternative approaches for widely varying effects.

Key Takeaway

- Synthetic DiD offers a practical, flexible, and robust approach for causal inference in complex panel data.
- Balances strengths of synthetic control and difference-in-differences while mitigating their weaknesses.
- Poised to become a valuable tool in causal panel analysis.

Concluding Remarks: Synthesis of Methods

- Combines synthetic control (matching precision) with difference-in-differences (parallel trends).
- Addresses challenges in synthetic control's convex hull constraint.
- Regularization allows for approximate matches with intercept terms.

Concluding Remarks: Robustness and Applications

- Doubly robust:
  - → Performs well as long as unit or time weights succeed.
- Adaptable to cases where:
  - → Diff-in-diff assumptions are weak.
  - → Synthetic control fit is imperfect.
- Links to augmented synthetic control for two-way bias reduction.

Practical Problems

- **Underfitting**: Cannot achieve parallel pre-treatment trends.
  - → Solution: Explore more or better controls or alternative methods.
- **Omitted Variable Bias**: External factors coincide with treatment, leading to identification failure.
- **Overfitting**: Synthetic control perfectly matches pre-treatment but fails post-treatment.
  - → Analogous to RDD functional form issues.

How to Rule Out Overfitting: Oracle Weights

- Estimator matches an ”oracle” that avoids overfitting by design.
- Oracle weights minimize expected squared error, not just in-sample error.
- Weights are robust to noise in the data.

Properties of SDID

- Approximately unbiased and normally distributed under large samples.
- Optimal variance, estimable via clustered bootstrap.
- Robust to noise and systematic confounding.

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`

Abadie on the value and use of synthetic control

Synthetic controls provide many practical advantages for the estimation of the effects of policy interventions and other events of interest. However, like for any other statistical procedure (and especially for those aimed at estimating causal effects), the credibility of the results depends crucially on the level of diligence exerted in the application of the method and on whether contextual and data requirements are met in the empirical application at hand. In this article, I

Summarizing

- Synthetic control was developed for the comparative case study; it is a kind of matching estimator with an underlying factor model for its identification (not parallel trends)
- Advancements have been made along multiple dimensions – bias adjustments, demeaning, as well as exploring more general structures than just factor models
- It is now a more robust, general causal panel method but the assumptions needed to justify it need ”due diligence”

Closing remark

- Focus on the treatment assignment mechanisms carefully to help understand how unobserved time varying confounders may be threatening your results, pay close attention to issues around observable matching bias, remember the importance of the ”long panel”
- Extrapolation based on the negative weighting should be done with the idea of bias reduction (augmented ridge), not simply for the purpose of fitting
- Good luck!