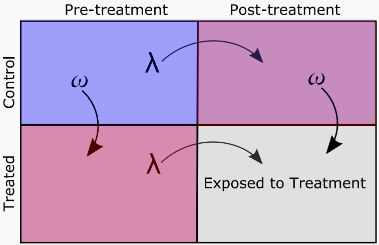
Synthetic Difference-in-Differences

David Hirshberg, Stanford University

Emory QTM. January 20, 2020.



References:

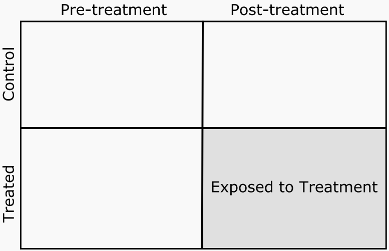
Synthetic Difference-in-Differences. arXiv 2020.
Arkhangelsky, Athey, H, Imbens, and Wager.

Least Squares with Error in Variables. Working Paper.

Panel Data is Everywhere

Onset	Treatment	Increased?
1979-81 Card	Miami’s workforce gains $45,000$ Cuban refugees	Unemployment
1989 Abadie et al.	California levies a 25 cents/pack Cigarette Tax	Health
Summer 1993	OK Soda released in select areas	Coke Sales
2015	Berkeley levies a 1 cent/ounce soda tax	Health
April 7, 2020	Wisconsin holds an election	Death
All the time, 2020	States open and close gyms, barber shops, etc.	Death

At its simplest, it looks like this



	Pre-treatment	Post-treatment
Control		
Treated		Exposed to Treatment

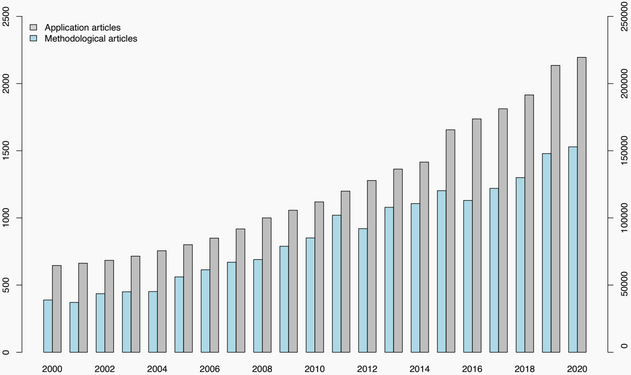
That’s what this talk is about

Where there's Panel Data, there's Diff-in-Diff

Differences-in-Differences estimation has become an increasingly popular way to estimate causal relationships.

How Much Should We Trust Difference in Differences Estimation?
Bertrand, Duflo, and Mullainathan [2004]

Published Diff-in-Diff Papers



Outline

Difference-in-Differences Concepts

Synthetic Difference-in-Differences

Theory of Identification and Inference

Panel Data Models and Reality

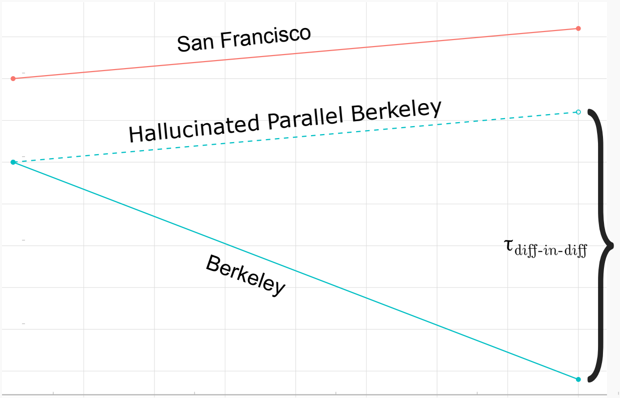
Difference-in-Differences Concepts

The Diff-in-Diff Idea

When Berkeley implemented a soda tax,
we compared to San Francisco

> *While Berkeley, the first U.S. city to pass a "soda tax," saw a substantial decline of 0.13 times/day in the consumption of soda in the months following implementation of the tax in March 2015, neighboring San Francisco, where a soda-tax measure was defeated, saw a 0.03 times/day increase*

Absent treatment, Berkeley might have increased like SF.



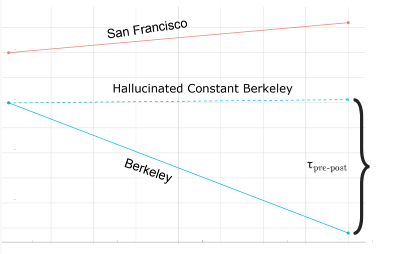
−	+
This is just speculation	No real evidence against it

This is what causal inference in observational studies is about.

- We can't know we're right.
- At best, we make claims the evidence doesn't rule out.

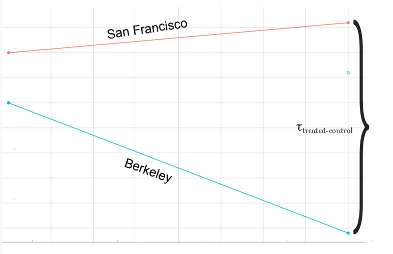
Some alternatives don’t meet this standard

Absent treatment, maybe
nothing would change in Berkeley.



Flaw: this isn’t what happened in SF.

Absent treatment, maybe
Berkeley would be just like SF.



Flaw: this wasn’t true before treatment.

Testing the Diff-in-Diff Premise

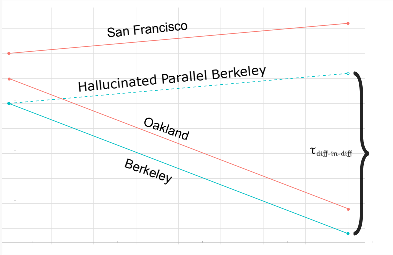
Absent treatment, Berkeley would have increased like SF.

When we have more cities, we can (sort of) test this.

- The choice to compare to SF seems arbitrary. Why not Oakland?
- If that choice makes a difference, our estimate seems arbitrary too.
- If it doesn't, we can feel a little more confident.

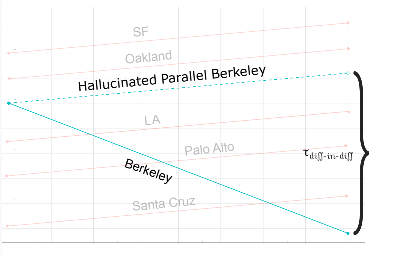
This doesn't really test our premise — we can't — but it's suggestive.

Bad: Controls follow Different Trends.



If we'd compared to Oakland,
we'd have estimated zero!

Good: Controls follow Parallel Trends.



No matter who we compare to,
we get the same estimate.

Or: Comparing controls, we estimate zero.

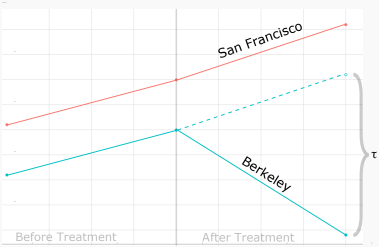
Or: Comparing controls, we estimate zero.

Testing the Diff-in-Diff Premise

When we have more time periods, we can do something similar.

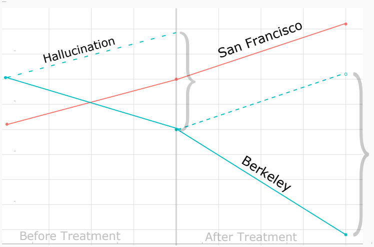
- If we’d done the same comparison last year, would diff-in-diff have worked?
- Treament *was* absent last year, so we know we should estimate zero.

Good: Pre-Treatment Parallel Trends.



When there’s no treatment, we’d have estimated zero.

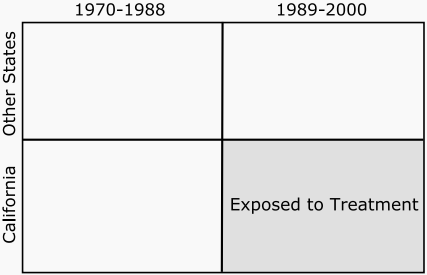
Bad: Different Pre-Treatment Trends!



When there’s no treatment, we’d have estimated a large effect.

Diff-in-Diff with Larger Panels

In 1989, California imposed a 25 cents/pack excise tax on cigarettes.
We estimate the effect on smoking in California after implementation.



	1970-1988	1989-2000
Other States		
California		Exposed to Treatment

We compare pre and post-treatment **averages** of treated units and controls to estimate the **average treatment effect under exposure** — the effect where and when treatment happened.

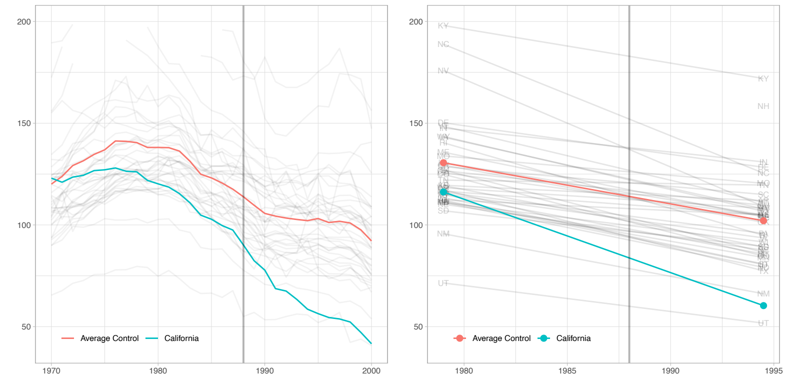
We'll doubt this estimate if it's sensitive to arbitrary-seeming choices of

1. the pre-treatment periods included.
2. the control units included.

When You Use ‘All the Data’, Checks Fail: We don’t get Parallel Trends

Abadie et al. [2010] considers data from 1970-2000 with most US states as controls.

- Choice of pre-treatment periods matters. The difference between CA and the average control is different in the 70s and the 80s.
- Choice of control units matters. The average trends for many control states are roughly parallel, but not all.



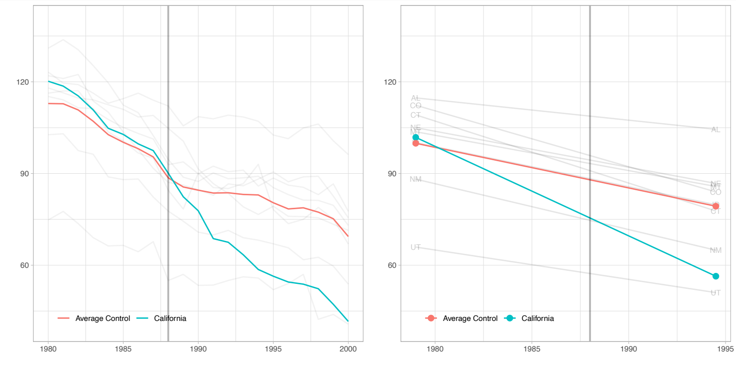
What we do when our checks fail

1. Give up. Hope they don't in your next project.
2. Try to fix it by preprocessing data, check again, iterate.
   - subset the data to including comparable control units.
   - choosing a reasonable time window.

   Some people are great at this, but it is hard to do well!

I chose a subset of controls and a time window. Things look better. How do you feel?

- remember: unless I made this up a-priori, valid statistical inference isn't simple.
- to prevent **p-hacking**, we need to account for 'multiple looks' at the data.



Automating this is a good idea

It's easy

- Tell your computer what you'd check.
- And let it 'preprocess' for you.

It has many benefits

1. It cuts down on work for everybody.
    - Doing preprocessing
    - Describing and justifying preprocessing
    - Reading descriptions of preprocessing
2. It's transparent and reproducible
    - **No p-hacking**: we can theoretically account for automated 'preprocessing'.
    - Because the method description tells the 'whole story', there's less appeal to authority.
3. It does a better job than you would. It can consider more possibilities.
    - California is like neither Nevada or Utah, but it is like $(2/3)$ Nevada + $(1/3)$ Utah.

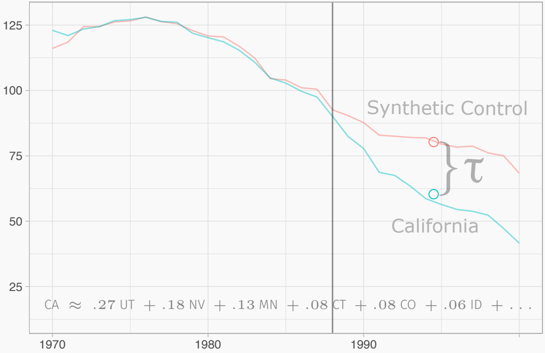
Automated preprocessing is easier to use and to trust.
It's a better community standard.

Automation Warm-Up: The Synthetic Control Estimator [Abadie et al., 2010]

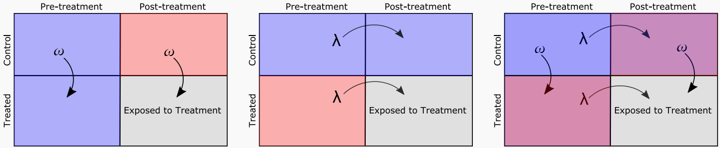
1. We'll start by automating a simple treated/control comparison.
2. We replace the kind of thing we're good at with something a computer is good at.
   - Human: choosing a subset of controls to average
   - Computer: choosing a weighted average of controls – a 'synthetic control'
3. We ask that this average tracks California pre-treatment. This is just regression.

$$\text{california}_t \approx \sum_i \omega_i \cdot \text{control}_{it}$$

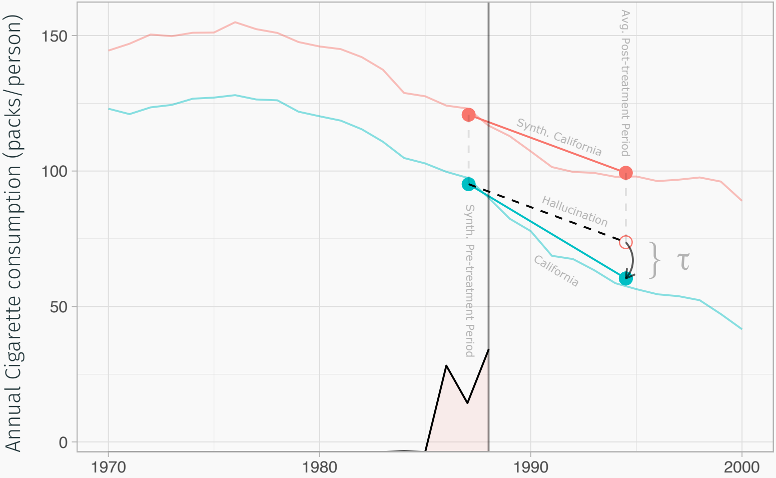
4. If this fits, we attribute post-treatment differences to treatment.



- Synthetic Control
  1. Using pre-treatment data, we learn an average of controls that's predictive of California.
  2. Assuming this relationship remain valid post-treatment, we use the same average of controls to impute treatment-free observations for California.
- Forecasting
  1. Using controls, we learn an average of periods forecasting what we see post-treatment.
  2. Assuming this relationship remain valid for the treated, we use the same average of periods to impute treatment-free observations for California.
- Synthetic Diff-in-Diff
  1. We do both synthetic control and forecasting and combine via diff-in-diff.
  2. Only one of these relationships has to remain valid.
  3. Constant offsets get *differenced out*: our synthetic control can be *parallel to* California.



The Synthetic Difference-in-Differences Diagram



Synthetic Difference-in-Differences

The Synthetic Difference-in-Differences Estimator (SDID)

Synthetic diff-in-diff is diff-in-diff with a synthetic control and pre-treatment period.

1. Estimate unit weights $\hat{\omega}$ defining a **synthetic control unit** using pre-treatment data.

$$\hat{\omega}_0 + \hat{\omega}^T Y_{co,pre} \approx Y_{\overline{tr},pre}.$$

2. Estimate time weights $\hat{\lambda}$ defining a **synthetic pre-treatment period** using control data.

$$\hat{\lambda}_0 + Y_{co,pre}\hat{\lambda} \approx Y_{co,\overline{post}}.$$

3. Apply diff-in-diff to the resulting synthetic 2×2 panel

	Synthetic Pre-Treatment	Average Post-treatment
Synthetic Control	$\hat{\omega}^T Y_{co,pre}\hat{\lambda}$	$\hat{\omega}^T Y_{co,\overline{post}}$
Average Treated	$Y_{\overline{tr},pre}\hat{\lambda}$	$Y_{\overline{tr},\overline{post}}$

Estimating the Weights

1. We estimate the weights defining the **synthetic control unit** via constrained least squares on the pre-treatment data.

$$\hat{\omega} = \underset{\omega_0, \omega}{\operatorname{argmin}} \left\| \omega_0 + \omega^T Y_{co,pre} - Y_{\overline{tr},pre} \right\|^2 + \zeta^2 T_{pre} \|\omega\|^2$$

$$\text{s.t. } \omega_i \geq 0, \quad \sum_{i=1}^{N_{co}} \omega_i = 1.$$

We require the synthetic control be a weighted average [as in Abadie et al., 2010]

- each unit's weight is nonnegative
- collectively, their weights sum to one

2. We estimate the weights defining the **synthetic pre-treatment period** via constrained least squares on the control data.

$$\hat{\lambda} = \underset{\lambda_0, \lambda}{\operatorname{argmin}} \left\| \lambda_0 + Y_{co,pre}\lambda - Y_{co,\overline{post}} \right\|^2$$

$$\text{s.t. } \lambda_t \geq 0, \quad \sum_{t=1}^{T_{pre}} \lambda_t = 1.$$

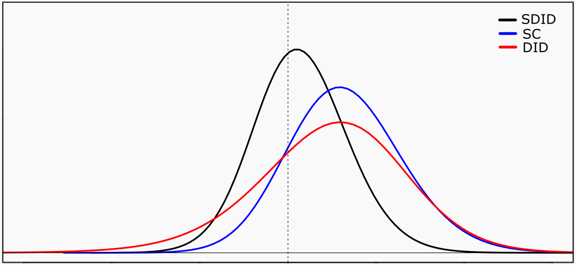
We impose analogous constraints.

Informal Theorem

In large square-ish panels with far fewer treated units than controls:

1. SDID is approximately unbiased and normal.
2. Its variance is optimal and estimable via clustered bootstrap.

Simulation Study



Distribution of errors in simulation based on Bertrand, Duflo, and Mullainathan [2004].

Outcome is Log Wage; Assignment based on Minimum Wage. See Section 1

Application: California's Cigarette Tax

- California is not an average state.
- California in the 90s is not California in the 70s.
- The more we account for that, the less impact we attribute to its 1989 cigarette tax.[1]

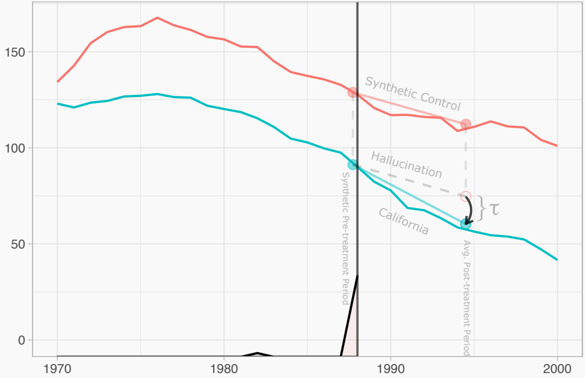
	diff-in-diff	synthetic control	synthetic diff-in-diff
Estimated Decrease annual packs per capita averaged over 1989-2000	27.4	19.8	13.4

[1] For details, see Section 2.

Theory of Identification and Inference

Theory of Identification and Inference

Potential Problems

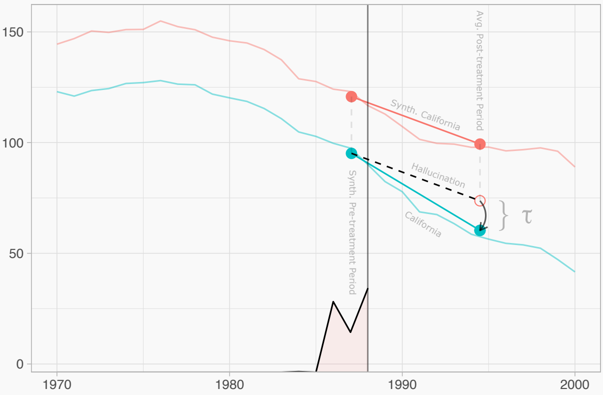


Underfitting

- We cannot get a parallel synthetic control.
- To do better, we'd need more/better controls or a fancier method.
- This is visible. When it comes up, we know to keep working.

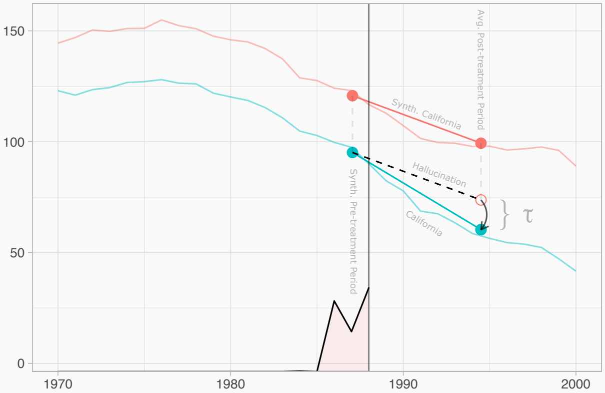
Above we underfit when using only Southeastern states as controls for California.

$$\text{CA} \not\approx 0.36\ \text{NC} + .32\ \text{LA} + .32\ \text{GA} + \omega_0.$$



Novel Confounding

- After treatment begins, something else shifts the relationship between the treated states and the states in the synthetic control.
    - e.g., if California's wildfires worsened after it passed the cigarette tax.
- To distinguish this from a treatment effect, we'd need more information.
- This is a causal problem — statistical theory can't help us.



Overfitting

- We get a parallel synthetic control, but it's an illusion.
  It just looks good because the plot shows a line fit to its training data.
- Its **comparability** to the treated unit **doesn't generalize post-treatment**.
  This is invisible: failure to generalize looks like a treatment effect.
- If we're willing to assume a model, statistical theory can rule this out.

A Simple Model to Distinguish ‘Signal’ from ‘Noise’

$$Y_{it} = L_{it} + \tau_{it} \cdot \text{treated}_{it} + \varepsilon_{it} \quad \text{where} \quad \mathrm{E}[\varepsilon \mid \text{treated}] = 0.$$

- L: deterministic ‘signal’ matrix of *noiseless* control potential outcomes.
- τ: deterministic matrix of treatment effects.
- ε: Noise matrix with iid gaussian (or similar) rows.
  - We have autocorrelation over time.
  - But no correlation between units.
- We’re estimating the average of τ on the exposed block, $\bar{\tau} = \tau_{\overline{tr},\overline{post}}$. That’s the average effect of treatment when and where it happened.

We’ll consider treatment assignment fixed.
All that will be random is the noise.

A Strategy to Rule Out Overfitting

- We'll show our estimator is equivalent to an **oracle estimator** that **can't overfit**.
- This oracle uses unit and time weights that **do not depend on the noise**.
- The weights we actually estimate minimize squared error;
  the oracle weights minimize **expected** squared error.

$$\tilde{\omega} = \operatorname*{argmin}_{\omega_0,\omega} \; \mathrm{E}_{\varepsilon} \left\| \omega_0 + \omega^T Y_{co,pre} - Y_{\overline{tr},pre} \right\|^2 + \zeta^2 \, T_{pre} \|\omega\|^2,$$

$$\tilde{\lambda} = \operatorname*{argmin}_{\lambda_0,\lambda} \; \mathrm{E}_{\varepsilon} \left\| \lambda_0 + Y_{co,pre}\lambda - Y_{co,\overline{post}} \right\|^2.$$

$$\text{s.t.} \quad \omega_i \geq 0, \quad \sum_{i=1}^{N_{co}} \omega_i = 1, \qquad \lambda_t \geq 0, \quad \sum_{t=1}^{T_{pre}} \lambda_t = 1$$

- The oracle's error is easy to characterize because these weights are non-random.
- We can't actually use it the oracle — it's not possible to compute it.
- But we prove equivalence in a sense that makes this irrelevant.
- When equivalence holds, our claims about the oracle hold for the real estimator.

Theory of Identification and Inference

Oracle Behavior

The Oracle is Nice

The oracle estimator is just a weighted average of the elements of the panel Y.

$$\tilde{\tau} = Y_{\overline{tr},\overline{post}} - \tilde{\omega}^T Y_{co,\overline{post}} - Y_{\overline{tr},pre}\tilde{\lambda} - \tilde{\omega}^T Y_{co,pre}\tilde{\lambda}.$$

This makes analysis easy. Its error separates cleanly into

- A bias component: replace Y with the signal L
- A noise component: replace Y with the noise ε.

It has everything you could want.

1. Approximate normality.
2. Low bias under plausible assumptions.
3. Optimal variance, estimable via the Bootstrap.

Normality and Inference

The oracle's noise component is a simple weighted average of mean-zero noise.

$$\tilde{\tau} - \bar{\tau} - \widetilde{bias} = \left(\varepsilon_{\overline{tr},\overline{post}} - \varepsilon_{\overline{tr},pre}\tilde{\lambda}\right) - \tilde{\omega}^T \left(\varepsilon_{co,\overline{post}} - \varepsilon_{co,pre}\tilde{\lambda}\right).$$

As noise for different units is independent:

- This average will be approximately normal by CLT.
- We can estimate variance by unit-clustered bootstrap.

Bias

The oracle estimator's bias is caused by changes in the fit of the oracle weights from *training* to *generalization*.

This change is small if:

- either set of weights fit well during training and generalize
  - ...from pre to post for the unit weights $\tilde{\omega}$
  - ...from control to treated for the time weights $\tilde{\lambda}$
- neither does, but the errors one makes are predicted by the other.

$$
\begin{aligned}
\widetilde{bias} &= \underset{\text{counterfactual post-treatment bias of } \tilde{\omega}}{\left(L_{\overline{tr},\overline{post}} - \tilde{\omega}^T L_{co,\overline{post}} - \tilde{\omega}_0\right)} \quad - \quad \underset{\text{bias of } \tilde{\omega} \text{ over the synthetic pre-treatment period}}{\left(L_{\overline{tr},pre} - \tilde{\omega}^T L_{con,pre} - \tilde{\omega}_0\right) \tilde{\lambda}} \\
&= \underset{\text{counterfactual treated-unit bias of } \tilde{\lambda}}{\left(L_{\overline{tr},\overline{post}} - L_{\overline{tr},pre}\tilde{\lambda} - \tilde{\lambda}_0\right)} \quad - \quad \underset{\text{bias of } \tilde{\lambda} \text{ on the synthetic control unit}}{\tilde{\omega}^T \left(L_{co,\overline{post}} - L_{co,pre}\tilde{\lambda} - \tilde{\lambda}_0\right)}.
\end{aligned}
$$

Optimal Variance

- The oracle time-weights predict post-treatment noise.
- That helps them minimize expected squared error.
- In particular, they converge to the post-on-pre noise autoregression vector ψ.

$$\tilde{\lambda} = \underset{\lambda_0, \lambda}{\operatorname{argmin}} \left\| \lambda_0 + L_{co,pre}\lambda - L_{co,\overline{post}} \right\|^2 + N_{co} \|\Sigma^{1/2}(\lambda - \psi)\|^2$$

$$\text{s.t. } \lambda_t \geq 0, \quad \sum_{t=1}^{T_{pre}} \lambda_t = 1 \text{ where } \psi \text{ satisfies } \operatorname{E}[\varepsilon_{i,\overline{post}} \mid \varepsilon_{i,pre}] = \varepsilon_{i,pre}\psi.$$

- This lets us do **better** than we could if we'd observed treatment effect plus noise

$$\tau_{it} + \varepsilon_{it} \text{ for exposed observations } it.$$

- That's essentially the variance of vanilla diff-in-diff.
- Our oracle time-weights get rid of the predictable part of this noise.
- It variance is that of the **least squares estimator** for $\bar{\tau}$ based on observations of

$$\tau_{it}\,\text{treated}_{it} + \varepsilon_{it} \text{ for all } it$$

Theory of Identification and Inference

Oracle Equivalence

Theorem [Arkhangelsky, Athey, H, Imbens, and Wager, 2020]

In ideal circumstances, the difference between the real and oracle SDID estimator is asymptotically negligible relative to the oracle's standard deviation in panels with

1. comparable numbers of control units and pre-treatment periods,
2. few post-treatment periods,
3. fewer treated units is than the square root of the number of controls.

This fits with the California cigarette tax example.

- 38 control states, 19 years of pre-treatment data, and 1 treated unit.

One aspect might throw you. More treated units is 'worse'.

- This is because we want the difference between the real and oracle estimators to be **smaller than** the oracle's standard deviation.
- When we add treated units, both decrease. Error does improve.
- But the oracle standard deviation can decrease faster,
  leaving too little room for other sources of error to 'disappear' in the noise.

Ideal Circumstances

Circumstances are ideal if the signal matrix L

1. admits a 'good' oracle synthetic control and synthetic pre-treatment period
2. is 'not too complex'

'Good' oracle synthetic controls/periods **fit** the signal **well** and are **diffuse**

- the oracle unit weights $\tilde{\omega}$ should distribute mass over enough control units,
- the oracle time weights $\tilde{\lambda}$ should, after fitting the noise autoregression vector ψ, distribute the rest of its mass over enough time periods.

Qualitatively, these are **overlap** assumptions: they hold if

- many control units are comparable to treated ones,
  e.g., if selection of treatment is randomized (possibly non-uniformly).
- many pre-treatment periods are comparable to post-treatment ones,
  e.g., if onset of treatment is randomized (possibly non-uniformly).

Ideal Circumstances

Circumstances are ideal if the signal matrix L

1. admits a 'good' oracle synthetic control and synthetic pre-treatment period
2. is 'not too complex'

A 'not too complex' signal is one that looks **different from** a matrix of **noise**

- formally, I mean approximable by a moderate-rank matrix with moderate error.
- moderate meaning smaller than the square root of the number of control units.

Qualitatively, this means units follow **mixtures of** relatively few **trends**.

- e.g., a state's behavior is not idiosyncratic, but characterized by its blend of industries, environments, cultures, etc.

Proof Intuition

Deviation from the oracle is essentially *bilinear* in the weight differences.

$$\begin{aligned}\hat{\tau} - \tilde{\tau} &\approx (\hat{\omega} - \tilde{\omega})^T \; L_{co,pre} \; (\hat{\lambda} - \tilde{\lambda}) \\ &\leq \|\hat{\omega} - \tilde{\omega}\| \Big\| L_{co,pre}(\hat{\lambda} - \tilde{\lambda}) \Big\|.\end{aligned}$$

Cauchy-Schwarz bounds depend on *prediction error* and *coefficient error*.
We characterize these using a version of the 'slow rate' analysis for the lasso.

Key ideas [H, 2020]

1. Including more controls won't hurt you.
   - The set of weights we optimize over — nonnegative and summing to one — is small. Error is essentially insensitive to its dimension.
2. Less than 'ideal circumstances' can be a problem. Error gets worse when:
   - the signal is too complex
   - the fit and dispersion of the oracle weights is poor

Panel Data Models and Reality

Our Options

There's a wide variety of methods for
estimating treatment effects in panel data.

1. Synthetic Difference-in-Differences and methods like it
2. Longitudinal studies methods from Biostatistics
   [van der Laan and Robins, 2003]
3. Nonseparable panel methods from Econometrics
   [Chernozhukov, Fernández-Val, Hahn, and Newey, 2013]

This is confusing. We have a bunch of methods and one task.

- It's not clear how to compare them.
  - Each assumes and attempts to exploit some structure in the data.
  - Related theory and simulations tend to assume this structure exists.
- It's reasonable to have many approaches. Panel data is many things.
- But to choose between them, we need to understand them in more general terms.

Longitudinal vs. Panel Methods

Longitudinal Approach

- We carefully compare complex treatment trajectories.
- Randomness arises from actual treatment randomization.
- We adjust for confounding assuming the patient's response depends on their medical history alone.

SDID-Type Panel Approach

- We assume simple, often additive, effect of treatment.
- Randomness arises from hard-to-interpret 'additive noise'.
- We adjust for confounding assuming rich shared structure relating units (mixtures of trends).

- In the longitudinal approach, we have a **faithful model of reality**: we estimate clearly defined treatment effects, relying on actual randomization of treament.
- With SDID, we adjust for confounding using a **rich model of shared structure**.
- We should find ways to synthesize the **best parts** of both approaches.
    - Getting identification via actual (or at least conceptual) randomization
    - Adjusting for confounding with rich shared structure

Thank you!

```
arxiv.org/abs/1812.09970
github.com/synth-inference/synthdid
```

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 2010.

Dmitry Arkhangelsky, Susan Athey, H, Guido Imbens, and Stefan Wager. Synthetic difference in differences. *arXiv*, 2020.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 2004.

David Card. The impact of the Mariel boatlift on the Miami labor market. *ILR Review*, 1990.

Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel models. *Econometrica*, 2013.

H. Least squares with error in variables. 2020.

H and Stefan Wager. Augmented minimax linear estimation. *arXiv:1712.00038*, 2017.

H and Stefan Wager. Debiased inference of average partial effects in single-index models: Comment on Wooldridge and Zhu. *Journal of Business & Economic Statistics*, 2020.

H, Arian Maleki, and Jose Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv:1901.10296*, 2019.

Vitor Hadad, H, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv:1911.02768*, 2019.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 2013.

Mark van der Laan and James Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.