Statistical Modeling and Causal Inference with R

Week 8: DiD and Synthetic Controls

Manuel Bosancianu

Max Schaub

November 2, 2020

Hertie School of Governance

Today’s focus

- ✓ A brief look at panel data
- ✓ Differences-in-difference / DiD
- ✓ Synthetic control method

A brief look at panel data

A brief look at panel data

So far, we have not looked at observations over time. This will change in this and the next session, where we look at panel data, i.e. repeated observations from the same unit over time.

Notation: Y_{it}, X_{it} etc. – additional subscript $_t$ for time period, e.g. years 2014 and 2020

Representation 1:

Wide format table

Unit c	Y_{c2014}	Y_{c2020}	D_c
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...

One row for each unit *i*, different columns for different time periods *t*.

A brief look at panel data

Representation 2:

Long format table			
Unit c	Year t	Y_c	D_c
County A	2014	42.1	0
County A	2020	38.5	0
County B	2014	41.2	1
County B	2020	40.2	1
...	...	...	...

One row for each time period t, $t \times i$, one column for each outcome of covariate.

A brief look at panel data

Panel data interesting for causal inference

- ✓ Allows for comparisons over time
- ✓ By looking at *changes* of the same unit over time, we can
  - ✓ Compare the same unit to itself, and by so doing...
  - ✓ control for all time-invariable characteristics of this unit.
  - ✓ Compare *changes* in outcomes between *different* units
- ✓ More technical treatment of panel data during the next class

Difference-in-differences / DiD

DiD: General setting where DiD applies

A major use of panel data are difference-in-differences DiD Analyses. In the DiD approach, causal identification is established by comparing changes between treatment and control units over time. DiD can potentially be used when:

1. There is some event or measure that affects some units of a population but leaves others untouched.

✓ A violent conflict starts in some regions but not others

✓ A schooling reform is introduced in some states but not others

✓ A tax reform affects certain projects but not other

✓ A diseases starts spreading in some areas but not others

and many other settings...

2. We have panel data from one or more points in time before the event happened/the measure was implemented, and after.

A motivating example: Effect of COVID-19 on electoral outcomes

Example: Effect of COVID-19 on municipal election in Bavaria on 15 March 2020 (Leininger & Schaub, 2020)

Outcome: Vote share for the largest party (CSU) at the county level

'Treatment': COVID-19 cases (69/27 counties had cases)

Data: Electoral outcomes for 2014 and 2020, COVID-19 prevalence

	CSU vote shares		
Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...

Example: Possible counterfactuals

Given these data, what is the most convincing comparison?

A first approach would be to compare County A, which did not have a COVID-19 case with County B, which was affected by COVID-19, i.e. to use $Y_{c=A,t=2020}$ as counterfactual for $Y_{c=B,t=2020}$.

In our data: $41.2 - 42.2 = -0.9$.

	CSU vote shares		
Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...

Example: possible counterfactuals

Such a comparison is obviously affected by a myriad of confounders, e.g. rural/urban, local economic development, education levels etc.

Note: In practice, we would do this for all 96 counties, and would use regression and matching methods to try to make the treatment assignment as independent as possible from such confounding factors.

With two-period data, we can use the electoral outcome during the previous election as a powerful control variable that captures many of the idiosyncracies that make counties different in potential outcomes.

Example: Possible counterfactuals

Another simple comparison would be to look a the affected County B before and after it was affected by COVID-19, i.e. to use $Y_{c=B,t=2014}$ as counterfactual for $Y_{c=B,t=2020}$

In our data: $40.2 - 41.2 = -1$.

	CSU vote shares		
Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
...	...	...	...

Example: Possible counterfactuals

This addresses the concern with time-fixed confounders such as whether the county is more urban or rural, and also largely addresses slow-moving factors such as age structure, average level of education etc.

However, again there are many other factors that may have influenced electoral outcomes over time, e.g. the appearance of a new party (the AfD), a change in environmental consciousness, a change from a less popular to a more popular leader, etc.

Any of these could be at the root of the effect we are observing.

Example: Possible counterfactuals, DiD

The idea of the differences-in-differences approach is to address the problems of between-unit and over-time comparison by comparing over-time differences between units, i.e. we use $Y_{c=A,t=2020} - Y_{c=A,t=2014}$ as a counterfactual for $Y_{c=B,t=2020} - Y_{c=B,t=2014}$.

	CSU vote shares			
Unit	Y_{2014}	Y_{2020}	D	$\Delta Y_{2020-2014}$
County A	42.1	38.5	0	-3.6
County B	41.2	40.2	1	-1
Δ	-0.9	1.7		2.6

In our data: $(40.2 - 41.2) - (38.5 - 42.1) = 2.6$.

DiD: Basic idea

The DiD estimator captures confounders that do not vary over time because it compares units to itself.

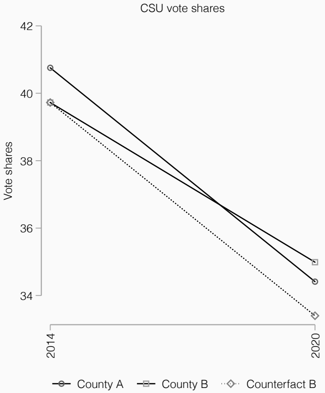
It also has the potential to capture confounders time-varying confounders that are the same for all units whether treatment or control.

The specific identifying assumption is that in the absence of the treatment (COVID-19), the outcome would have followed a time-invariant unit (here: county) effect and a time effect that is common across all counties, i.e.

$E[Y_{0ct}|D, t] = \gamma_c + \lambda_t$

DiD: Counterfactual

Graphical representation of treatment effect given the counterfactual:



DiD: Assumption

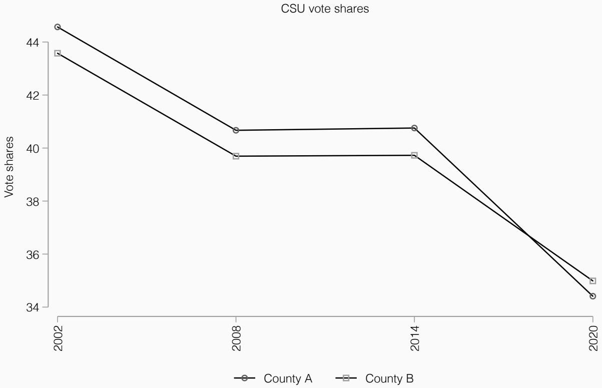
For the identifying assumption to hold, the parallel trends assumption has to be met: the claim that in the absence of treatment, treatment and control group outcomes would have moved in parallel.

This assumption in itself is not testable since we cannot observe the potential outcome $Y_{0ct+1}|D_c = 1$.

However, the parallel trends assumption implies that prior to the treatment, the units in the treatment condition should have followed the same trend – they should have moved in a similar way in response to external influences.

DiD: Parallel trends

Graphical representation of common trends prior to treatment:



DiD: Calculation

Usually, we consider more than two units. For example, in the election dataset there are 96 counties.

	CSU vote shares		
Unit	Y_{2014}	Y_{2020}	D
County A	42.1	38.5	0
County B	41.2	40.2	1
County C	47.4	36.4	0
County D	49.0	40.2	0
County E	38.2	34.3	1
County F	43.9	39.1	1
...	...	...	...

DiD: Calculation

In such cases, we use group averages to calculate the DiD, i.e.

$$DiD = \{E[Y_{1c}|D=1, t=2020] - E[Y_{0c}|D=0, t=2020]\} - \\ \{E[Y_{1c}|D=1, t=2014] - E[Y_{0c}|D=0, t=2014]\}$$

There are several ways to calculate the DiD:

1. Manually, by calculating averages for subgroups defined by *D* and *t*
2. Regressing first differences on the treatment indicator (in ‘wide’ data format)
3. Using the regression formulation of the DiD model (in ‘long’ data format)

DiD: Calculation

1. ‘Manually

$$
\begin{aligned}
DiD = {} & \{E[Y_{1c}|D=1, t=2020] - E[Y_{0c}|D=0, t=2020]\} - \\
& \{E[Y_{1c}|D=1, t=2014] - E[Y_{0c}|D=0, t=2014]\} \\
& = (34.99 - 34.41) - \\
& (39.73 - 40.76) \\
& = 1.61
\end{aligned}
$$

DiD: Calculation

2. Regression of first differences on treatment indicator (in ‘wide’ data format) using $\Delta Y_{c_{2014-2020}} = \alpha + \delta D_c + \Delta u_c$

	$CSU_{2014-2020}$
Treat	1.61^{**}
	(0.79)
Intercept	-6.34^{***}
	(0.71)
N	96
R^2	0.05

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

DiD: Calculation

3. Using the regression formulation of the DiD model, by far the most common way to calculate the DiD for the two-period, two-conditions case, i.e. : $DiD = Y_{ct} = \alpha + \beta D_c + \gamma Post_t + \delta(D_c \times Post_t) + u_{ct}$ with $Post_t$ standing for the post-treatment period, i.e. the year 2020

The coefficients estimated by the regression model translate into the observed outcomes as follows:

	$t = 2014$ (pre)	$t = 2020$ (post)
$D_c = 0$	$E[Y_{0c2014} \| D_c = 0]$	$E[Y_{0c2020} \| D_c = 0]$
$D_c = 1$	$E[Y_{1c2014} \| D_c = 1]$	$E[Y_{1c2020} \| D_c = 1]$
$D_c = 0$	α	$\alpha + \gamma$
$D_c = 1$	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta$

DiD: Regression estimators

Regression results from our example data:

	Share CSU
Treat	-1.03
	(1.56)
Post	-6.34^{***}
	(0.72)
Treat × Post	1.61^{**}
	(0.79)
Intercept	40.76^{***}
	(1.39)
N	192
R^2	0.16

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

DiD: Regression estimators

We can use these results and the insights into the interpretation of the coefficients to recover the 'manual' DiD estimator:

	$t = 2014$ (pre)	$t = 2020$ (post)
$D_c = 0$	40.76	$40.76 + (-6.34) = 34.42$
$D_c = 1$	$40.76 + (-1.03) = 39.73$	$40.76 + (-1.03) + (-6.34) + 1.61 = 35.00$

which gives: $DiD = (35.00 - 39.73) - (34.42 - 40.76) = 1.61$

Easier to just look at δ, of course...

DiD: Generalized DiD regression estimator

The DiD model can also be expressed as a two-way fixed effects model in the so-called generalized DiD:

$DiD = Y_{ct} = \lambda_c + \gamma_t + \delta D_{ct} + u_{ct}$

where λ_c are dummies for the units, γ_t for the included time periods, and D_{ct} indicates whether a unit c is in its treated or untreated state at a given point in time t.

This model also compares within-unit deviations in outcomes between units at the same time, but accommodates situations where some units have their treatment ‘switched on’ at different points in time.

DiD: Regression estimators

The regression formulation of the DiD has the advantage that we can add additional control variables, notably factors varying at the level of the unit over time. E.g.

$$DiD = Y_{ct} = \lambda_c + \gamma_t + \delta D_{ct} + \boldsymbol{\rho X_{ct}} + u_{ct}$$

Such factors ρX_{ct} could be the economic development at the county level between 2014 and 2020, population dynamics, refugee settlement, etc.

Including such predictors should increase the precision of our estimates.

The framework also allows for non-dichotomous treatments/a treatment of varying intensity e.g. the number of COVID-19 cases.

DiD: Regression estimators

DiD including controls: outcome is % CSU vote share.

Treat	-2.17
	(1.56)
Post	-7.07^{***}
	(1.05)
Treat × *Post*	1.62^{**}
	(0.80)
Diff popdens	-0.03^{*}
	(0.02)
Diff foreign	-0.03
	(0.20)
Diff unemployrate	-0.12
	(0.08)
Diff age 60 and over	0.40
	(0.31)
Diff employees	0.06
	(0.06)
Intercept	36.29^{***}
	(7.60)
N	192
R^2	0.34

Standard errors in parentheses

$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

DiD: Leads in regression estimators

The generalized formulation of the DiD framework also allows us to test the common trends assumption by including 'leads' – 'treatment effects' for the periods leading up to the actual treatment.

Practically this is done by including the interactions between the treatment indicator and year dummies, but leaving out the pre-treatment period (the year 2014 in our example) as reference period:

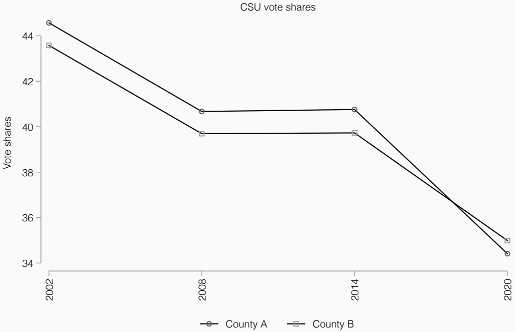
$$Y_{ct} = \lambda_c + \gamma_t + \delta_{t-3}D_{c_{2002}} + \delta_{t-2}D_{c_{2008}} + \delta_t D_{c_{2020}} + \rho X_{ct} + u_{ct}$$

If the common trends assumption holds, the coefficients for δ_{t-3} and δ_{t-2} should be small and statistically insignificant.

Another way of thinking about 'leads' is in terms of placebo treatments: you are 'assuming' that the treatment took place in 2002 or 2008 – and show that this imagined treatment has no effect!

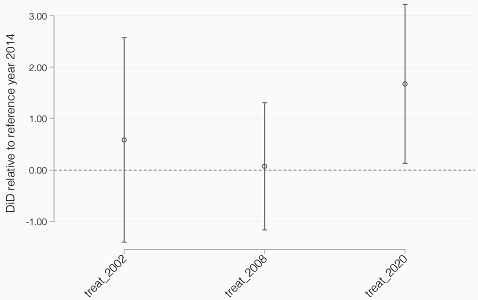
DiD: Leads in regression estimators

Recall the graphical representation of the common trends prior to treatment:



Two-period diff-in-diff

Plot of ‘leads’ in DiD framework:



Plotted ‘leads’ are the DiD between 2002 and 2014, 2008 and 2014, and the third coefficient is the difference-in-differences between 2020 and 2014, respectively.

DiD: Lags in regression estimators

In the same way, we can include ‘lag’ – ‘treatment effects’ for the periods after the actual treatment occurred (cp. Autor (2003) for the classic reference using both ‘leads’ and ‘lags’).

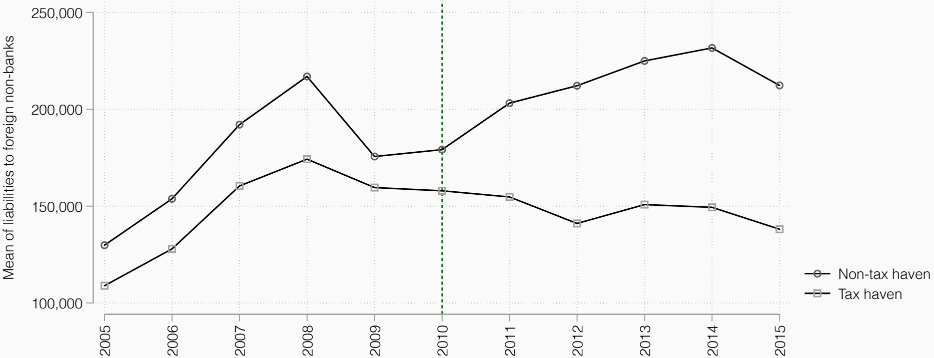
These can help to see how long-lasting the effect is. Again, this is done by adding interactions between the treatment indicator and year dummies.

The full model may then look like this (here based on an example from Hakelberg and Schaub (2018)):

$$Y_{ct} = \lambda_c + \gamma_t + \sum_{t=2005}^{2009} \delta_t D_{ct} + \sum_{t=2011}^{2015} \delta_t D_{ct} + \rho X_{ct} + u_{ct}$$

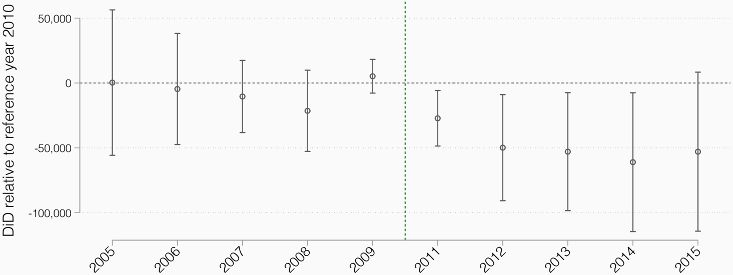
DiD: Lags in regression estimators

Example studying the effect of the FATCA (Foreign Account Tax Compliance Act) reform in the U.S. in 2011 on tax havens (Hakelberg & Schaub, 2018).



DiD: Lags in regression estimators

Graphical representation of inclusion of ‘lags’ in DiD framework from Hakelberg and Schaub (2018).



The plot shows that the ‘placebo treatments’ between 2005 and 2009 – imagined reforms prior to the actual reform in 2011 – had no effect on liabilities, whereas the real reform in 2011 had a clear, additive effect every single year 2011-2014.

DiD: Further issues

When using DiD, standard errors need to be adjusted. This is because observations from the same unit for two or more time periods are likelily to be serially correlated, i.e. not independent from each other.

For example, the vote share 2014 is correlated with the vote share 2020. The adjustment can be done by using *clustered standard errors* or *bootstrapping* (estimating the standard errors by sampling sub-samples from the data that give an estimate of the variability of the estimated DiD effect).

The DiD framework is useful for the analysis of almost any panel data set where the treatment and control group follow a common trend pre-treatment.

However, even though the common trends assumption can be relaxed by allowing for units to follow their own time trends (cp. Angrist and Pischke (2009)), most scholars would agree that DiD conceptually makes most sense when there are parallel trends.

So what if there are no parallel trends?

Synthetic control method

Synthetic control method: General setting

The synthetic control method serves to estimate the causal effect of an intervention or treatment by comparing a treated unit over time to a synthetic control unit that is constructed as a weighted average of potential control or ‘donor’ units.

Synthetic control method: General setting

The synthetic control method might be used in cases:

1. Where there is no optimal control/a lack of a common trend
2. Where the treatment unit is $n = 1$ – inspired by qualitative comparative case studies

✓ Violent conflict: terrorism in the Basque country (Abadie & Gardeazabal, 2003)

✓ Legal changes in a single state: tobacco law in California (Abadie, Diamond, & Hainmueller, 2010)

✓ Rare historical events: German reunification (Abadie, Diamond, & Hainmueller, 2015)

3. We have panel data from several points in time before the event happened/the measure was implemented, and after

Synthetic control method: Comparison with DiD

Similar idea to DiD:

- ✓ Estimate treatment effect by comparing the trends in outcomes in a treatment unit to that of control unit

Differences:

- ✓ Explicitly restricted to a single treatment unit, and a single, composite control unit
- ✓ Rather than relying on actually observed cases, constructs optimal control unit from pool of control units
- ✓ The synthetic control is contructed as a weighted average of the units in the donor pool
- ✓ Different (less formal) statistical tests to gage plausibility of estimated effect

Synthetic control method: Example

What is the effect of an anti-smoking measure in California (Proposition 99, enacted in 1989) on cigarette sales (Abadie et al., 2010)?

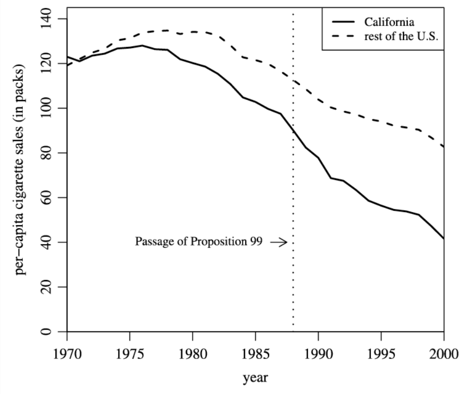
Idea: compare development in cigarette sales in California to similar state.

Problems:

- ✓ No other unit(s) with similar pre-Proposition trajectory of cigarette sales
- ✓ Single treated case (California) – how to do statistical inference/test uniqueness of effect?

Example: Lack of control group

Figure demonstrating that California cigarette sales followed different trajectory from rest of the U.S.



Synthetic control group: Distance measure

Abadie, Diamond, and Hainmueller (2010) (following Abadie and Gardeazabal 2003) address this problem by constructing a 'synthetic' California out of a weighted sample of 38 other U.S. states.

Technically, the idea is as follows:

Assume that X_1 is a vector holding covariate values from the treatment unit (California), and X_0 is a matrix holding values for the same covariates for all the control units (the 38 other states), we can find a vector W^* with weights for each control unit that minimize the distance between the treatment case and the control units.

Synthetic control group: Distance measure

Specifically, Abadie, Diamond, and Hainmueller (2010) look for a vector W^* that minimizes

$$\sum_{m=1}^{k} v_m (X_{1m} - X_{0m} W)^2$$

i.e. the squared distance between all the values m of the vector X_1 minus the value for the same covariate m in the vector of the control units X_0 times a weight v_m that captures the importance of this covariate for predicting the outcome of interest.

The weight v_m is found with methods from machine learning.

Example: Selected weights

Table showing the weights assigned to the units of the door pool to create the ‘synthetic California’

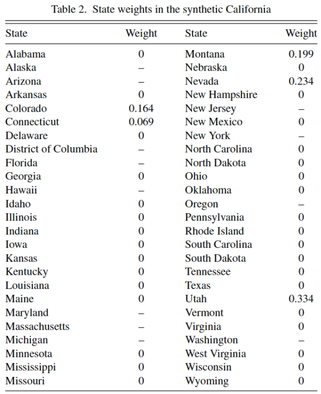
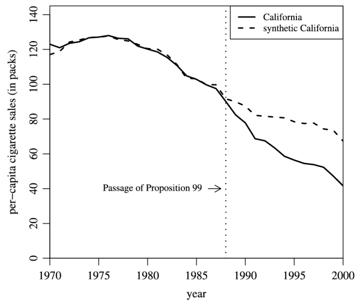


Table 2. State weights in the synthetic California

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

Example: Synthetic California

The ‘synthetic California’ closely tracks cigarette sales in the real California until the Proposition 99 is introduced, from where on the treated case and the synthetic control diverge.



From this they conclude that the treatment *caused* cigarette sales to decline.

Synthetic control group: Estimator

The treatment effect is given by the comparison of postintervention outcomes (cigarette sales post Proposition 99) between the treated unit and the synthetic control, i.e. $Y_1 — Y_0 W^*$, which is calculated for each post-treatment period t as:

$$Y_{1t} - \sum_{j=2}^{J+1} \omega_j^* Y_{jt}$$

i.e. the difference between the outcome for the treated unit ($j = 1$) minus the sum of the weighted outcomes of the control unit (indexed $j = 2$, $j = 3$, etc.).

Synthetic control method: Inference

The problem is that with such an approach is that we are left with a 1:1 comparison, meaning that usual statistical tests (which are based on comparing the *distributions* betweent the treatment and the control group) do not apply.

Instead of the usual ‘frequentist’ tests, the authors conduct a series of other tests, including “in-time placebos” (pretending the treatment took place in another year, just as is done with the inclusion of ‘leads’ in DiD), and “in-space placebos.”

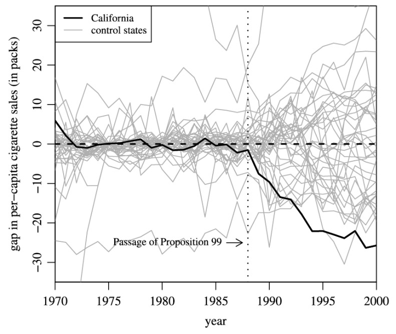
Synthetic control method: Inference

The idea of the “in-space placebo” tests is to

1. estimate hypothetical treatment effects for every potential control unit in the donor pool
2. rank treatment effects and determine which share of hypothetical treatments produce a stronger treatment effect than the one observed in the studied case
3. which then this gives us a p-value; for example, if 92% of observations produce a smaller treatment effect, the associated p-value of the treatment effect in focus (that for California) is $p = 0.08$

Example: Inference

Graphical representation of ”in-space placebo:”



The different paths show differences in cigarette sales between each control unit and it’s best fitting synthetic control over time. The treatment effect for California clearly stands out as one of the strongest effects, lending credibility to the claim that the reform here had a causal impact.

Thank you for watching, and see you next Monday!

References i

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510.

Abadie, A., & Gardeazabal, J. (2003, March). The Economic Costs of Conflict: A Case Study of the Basque Country. *The American Economic Review*, *93*(1), 113–132.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist’s companion*. Princeton: Princeton University Press.

Autor, D. H. (2003, January). Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, *21*(1), 1–42.

References ii

Hakelberg, L., & Schaub, M. (2018, September). The redistributive impact of hypocrisy in international taxation. *Regulation & Governance*, *12*(3), 353–370.

Leininger, A., & Schaub, M. (2020, April). *Strategic Alignment in Times of Crisis: Voter choice at the Dawn of the COVID-19 Pandemic* (Preprint). SocArXiv.