

# Synthetic Controls: Methods and Practice

Alberto Abadie  
MIT

Synthetic Control Conference  
Center for Statistics and Machine Learning  
Princeton University

June 2, 2022

# Introduction

- ▶ Synthetic control methods were originally proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the aim to estimate the effects of **aggregate interventions**.
- ▶ Many events or interventions of interest naturally happen at an aggregate level affecting a small number of large units (such as cities, regions, or countries).
- ▶ Even in experimental settings micro-interventions may not be feasible (e.g., fairness) or effective (e.g., interference).

---

In this talk, I will use the terms “event”, “intervention”, and “treatment” interchangeably.

## Applications

- ▶ Synthetic controls have been applied to study the effects of right-to-carry laws (Donohue et al., 2017), legalized prostitution (Cunningham and Shah, 2018), immigration policy (Bohn et al., 2014), corporate political connections (Acemoglu et al., 2016) and many other policy issues.
- ▶ They have also been adopted as the main tool for data analysis across different sides of the issues in recent prominent debates on the effects of immigration (Borjas, 2017; Peri and Yasenov, 2017) and minimum wages (Allegretto et al., 2017; Jardim et al., 2017; Neumark and Wascher, 2017; Reich et al., 2017).
- ▶ Synthetic controls are also applied outside economics in the social sciences, biomedical disciplines, engineering, etc. (see, e.g., Heersink et al., 2017; Pieters et al., 2017).

## Applications

- ▶ Outside academia, synthetic controls have been widely adopted by multilateral organizations, think tanks, business analytics units, governmental agencies, and consulting firms.
- ▶ For example, the synthetic control method plays a prominent role in the official evaluation of the effects of the massive Bill & Melinda Gates Foundation's *Intensive Partnerships for Effective Teaching* program (Gutierrez et al., 2016).
- ▶ Widely applied at tech companies (Uber, Google, Amazon, eBay, ...)

# The Washington Post

Wonkblog

**Seriously, here's one amazing math trick to learn what can't be known**

# THE WALL STREET JOURNAL

REAL TIME ECONOMICS | ECONOMICS

**How an Analysis of Basque Terrorism Helps Economists Understand Brexit**

A method pioneered by an MIT professor has also been used to estimate the economic effect of a tobacco ban, German reunification, legalization of prostitution and gun rights

## Literature is large, and there is much I will not cover ...

- ▶ **Matrix/tensor completion:** Amjad, Shah, and Shen, (2018), Agarwal, Shah and Shen (2020), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2018), Bai and Ng (2020)
- ▶ **Bias correction:** Abadie and L'Hour (2020), Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, (2019), Ben-Michael, Feller, and Rothstein (2020)
- ▶ **Inference:** Cattaneo, Feng, Titiunik (2020), Chernozhukov, Wüthrich, and Zhu (2019a, 2019b), Firpo and Possebom (2018)
- ▶ **Functional and distributional outcomes:** Chernozhukov, Wüthrich, and Zhu (2019c), Gunsilius (2020)
- ▶ **Large- $T$ :** Botosaru and Ferman (2019), Ferman (2019), Li (2020)
- ▶ **Other related methods:** Brodersen, Gallusser, Koehler, Remy, and Scott (2015)

... and many more (and many, many, empirical applications).

## A primer on synthetic control estimation

- ▶ When the units of analysis are a few aggregate entities, a combination of comparison units (a “synthetic control”) often does a better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- ▶ A synthetic control is selected as the **weighted average** of all potential comparison units that best resembles the characteristics of the treated unit(s).

## A primer on synthetic control estimation

- ▶ 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  $J$  units are an untreated reservoir of potential controls (a donor pool).
- ▶ Let  $Y_{it}^I$  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  $T$ .
- ▶ Let  $Y_{it}^N$  be the outcome that would be observed for unit  $i$  at time  $t$  in the absence of the intervention.
- ▶ We aim to estimate the effect of the intervention on the treated unit,

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

for  $t > T_0$ , and  $Y_{1t}$  is the outcome for unit one at time  $t$ .



## A primer on synthetic control estimation

- ▶ Let  $\mathbf{W} = (w_2, \dots, w_{J+1})'$  with  $w_j \geq 0$  for  $j = 2, \dots, J+1$  and  $w_2 + \dots + w_{J+1} = 1$ . Each value of  $\mathbf{W}$  represents a potential synthetic control.
- ▶ Let  $\mathbf{X}_1$  be a  $(k \times 1)$  vector of pre-intervention characteristics for the treated unit. Similarly, let  $\mathbf{X}_0$  be a  $(k \times J)$  matrix which contains the same variables for the unaffected units.
- ▶ The vector  $\mathbf{W}^* = (w_2^*, \dots, w_{J+1}^*)'$  is chosen to minimize  $\|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}\|$ , subject to our weight constraints.
- ▶ Let  $Y_{jt}$  be the value of the outcome for unit  $j$  at time  $t$ . For a post-intervention period  $t$  (with  $t \geq T_0$ ) the synthetic control estimator is:

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

## A primer on synthetic control estimation

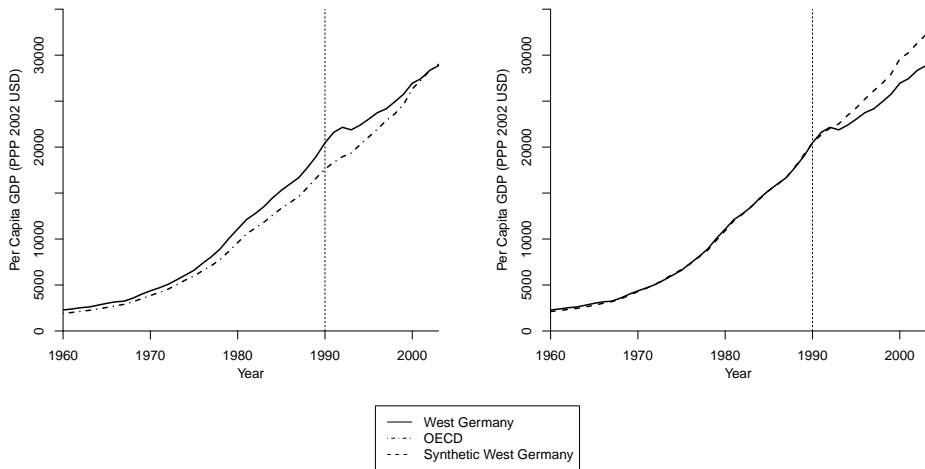
- ▶ Typically,

$$\|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}\| = \left( \sum_{h=1}^k v_h (X_{h1} - w_2 X_{h2} - \dots - w_{J+1} X_{hJ+1})^2 \right)^{1/2}$$

- ▶ The positive constants  $v_1, \dots, v_k$  reflect the predictive power of each of the  $k$  predictors on  $Y_{1t}^N$ .
- ▶  $v_1, \dots, v_k$  can be chosen by the analyst or by data-driven methods.

# A primer on synthetic control estimation

## Application: German reunification



# A primer on synthetic control estimation

Application: German reunification

	West Germany (1)	Synthetic West Germany (2)	OECD Sample (3)
GDP per-capita	15808.9	15802.24	13669.4
Trade openness	56.8	56.9	59.8
Inflation rate	2.6	3.5	7.6
Industry share	34.5	34.5	34.0
Schooling	55.5	55.2	38.7
Investment rate	27.0	27.0	25.9

*Note:* First column reports  $\mathbf{X}_1$ , second column reports  $\mathbf{X}_0 \mathbf{W}^*$ , and last column reports a simple average for the 16 OECD countries in the donor pool. GDP per capita, inflation rate, and trade openness are averages for 1981–1990. Industry share (of value added) is the average for 1981–1989. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–1984.

# A primer on synthetic control estimation

Application: German reunification

country $j$	$W_j^*$	country $j$	$W_j^*$
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

## A primer on synthetic control estimation

- ▶ Abadie et al. (2010) establish a **bias bound** under the **factor model**

$$Y_{it}^N = \theta_t \mathbf{Z}_i + \lambda_t \boldsymbol{\mu}_i + \varepsilon_{it},$$

where  $\mathbf{Z}_i$  are observed features,  $\boldsymbol{\mu}_i$  are unobserved features, and  $\varepsilon_{it}$  is a unit-level transitory shock, modeled as random noise.

- ▶ Suppose that we can choose  $\mathbf{W}^*$  such that:

$$\sum_{j=2}^{J+1} w_j^* \mathbf{Z}_j = \mathbf{Z}_1, \quad \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \quad \dots, \quad \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0},$$

with probability one. In practice, these may hold only approximately.

## A primer on synthetic control estimation

Suppose that  $E|\varepsilon_{jt}|^p < \infty$  for some  $p > 2$ . Then,

$$|E[\widehat{\tau}_{1t} - \tau_{1t}]| < C(p)^{1/p} \left( \frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\}$$

where  $F$  is the number of unobserved factors,

$$\sigma_{jt}^2 = E|\varepsilon_{jt}|^2, \quad \sigma_j^2 = \frac{1}{T_0} \sum_{t=1}^{T_0} \sigma_{jt}^2, \quad \bar{\sigma}^2 = \max_{j=2, \dots, J+1} \sigma_j^2,$$
$$m_{pjt} = E|\varepsilon_{jt}|^p, \quad m_{pj} = \frac{1}{T_0} \sum_{t=1}^{T_0} m_{pjt}, \quad \bar{m}_p = \max_{j=2, \dots, J+1} m_{pj},$$

for  $p$  even,  $|\lambda_{tf}| \leq \bar{\lambda}$  for all  $t = 1, \dots, T$  and  $f = 1, \dots, F$ , and

$$\underline{\xi} \leq \xi(M) = \text{smallest eigenvalue of } \frac{1}{M} \sum_{t=T_0-M+1}^{T_0} \lambda'_t \lambda_t.$$

## A primer on synthetic control estimation

- ▶ The bias bound is predicated on close fit, and controlled by the ratio between the scale of  $\varepsilon_{it}$  and  $T_0$ .
- ▶ In particular, the credibility of a synthetic control depends on the extent to which it is able to fit the trajectory of  $Y_{1t}$  for an extended pre-intervention period.



## A primer on synthetic control estimation

- ▶ There are no ex-ante guarantees on the fit. If the fit is poor, Abadie et al. (2010) recommend against the use of unadjusted synthetic controls. Bias correction methods have been proposed for this case.
- ▶ Settings with small  $T_0$ , large  $J$ , and large noise create substantial risk of overfitting.
- ▶ To reduce interpolation biases and risk of overfitting, restrict the donor pool to units that are similar to the treated unit.

---

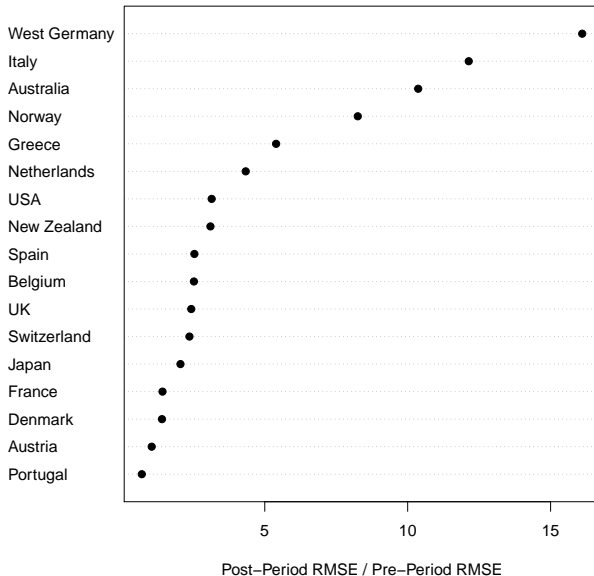
For bias-correction techniques, see Abadie and L'Hour (2020), Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, (2019), Ben-Michael, Feller, and Rothstein (2020).

## A primer on synthetic control estimation

- ▶ Abadie et al. (2010) propose a mode of inference for the synthetic control framework that is based on permutation methods.
- ▶ A permutation distribution can be obtained by iteratively reassigning the treatment to the units in the donor pool and estimating “placebo effects” in each iteration.
- ▶ The effect of the treatment on the unit affected by the intervention is deemed to be significant when its magnitude is extreme relative to the permutation distribution.

# A primer on synthetic control estimation

## Application: German reunification

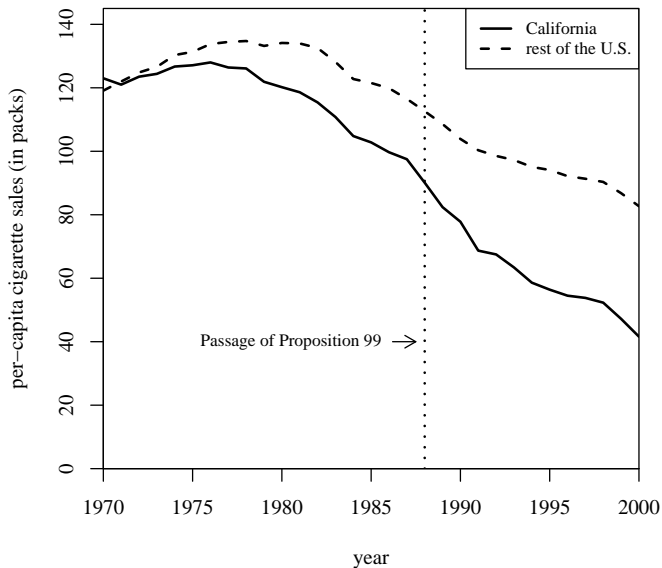


## A primer on synthetic control estimation

- ▶ The permutation distribution is more informative than mechanically looking at  $p$ -values alone.
- ▶ Depending on the number of units in the donor pool, conventional significance levels may be unrealistic or impossible.
- ▶ Often, one sided inference is most relevant.

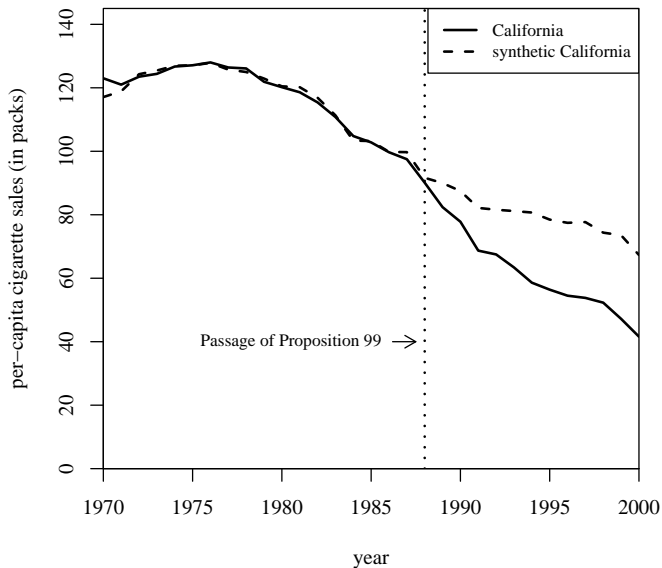
# A primer on synthetic control estimation

Application: California tobacco control program



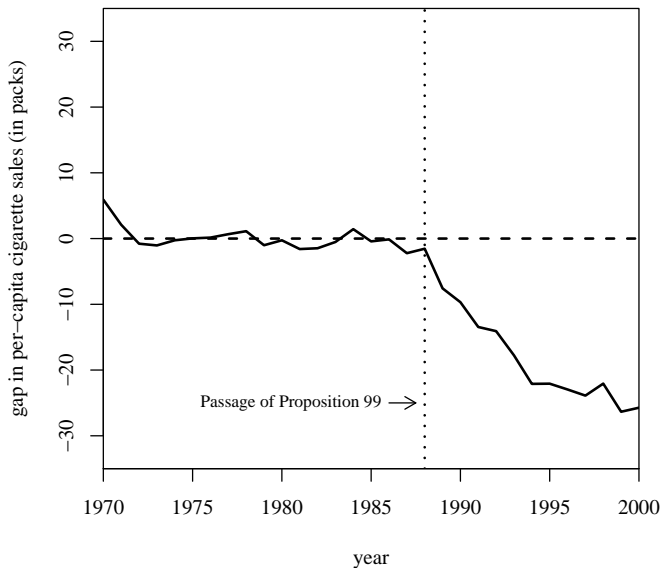
# A primer on synthetic control estimation

Application: California tobacco control program



# A primer on synthetic control estimation

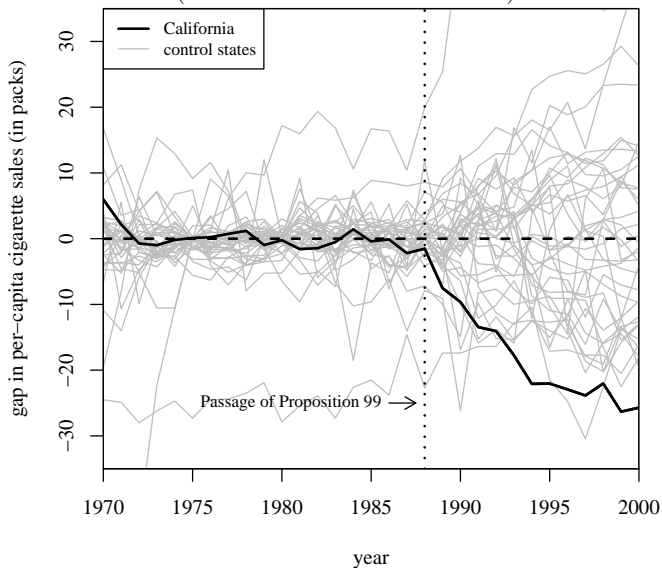
Application: California tobacco control program



# A primer on synthetic control estimation

Application: California tobacco control program

(ALL STATES IN DONOR POOL)

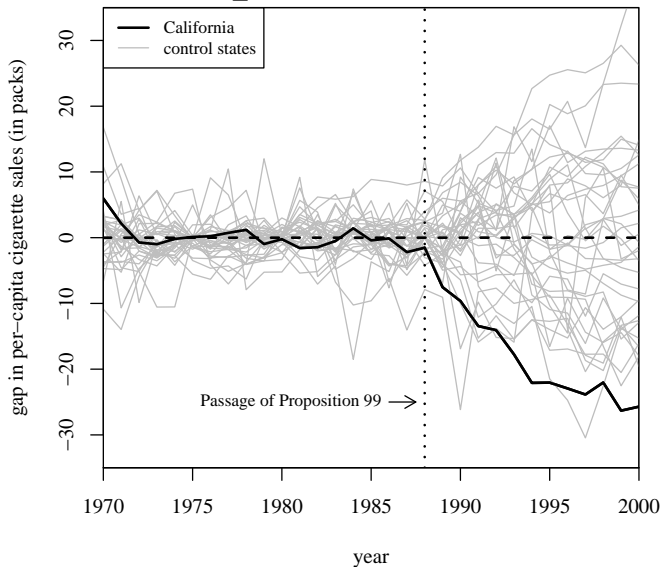




# A primer on synthetic control estimation

## Application: California tobacco control program

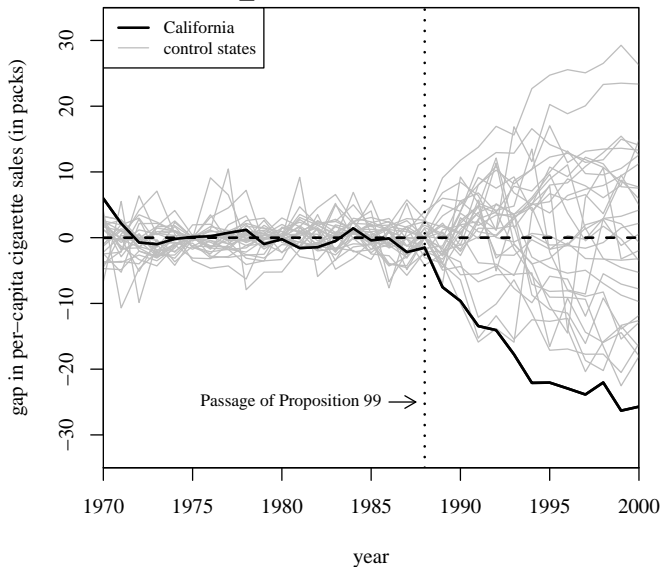
(PRE-PROP. 99 MSPE  $\leq$  20 TIMES PRE-PROP. 99 MSPE FOR CA)



# A primer on synthetic control estimation

## Application: California tobacco control program

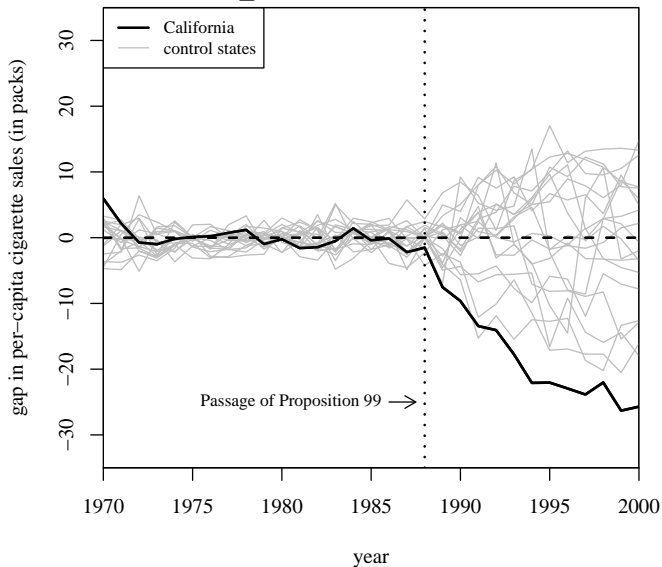
(PRE-PROP. 99 MSPE  $\leq$  5 TIMES PRE-PROP. 99 MSPE FOR CA)



# A primer on synthetic control estimation

## Application: California tobacco control program

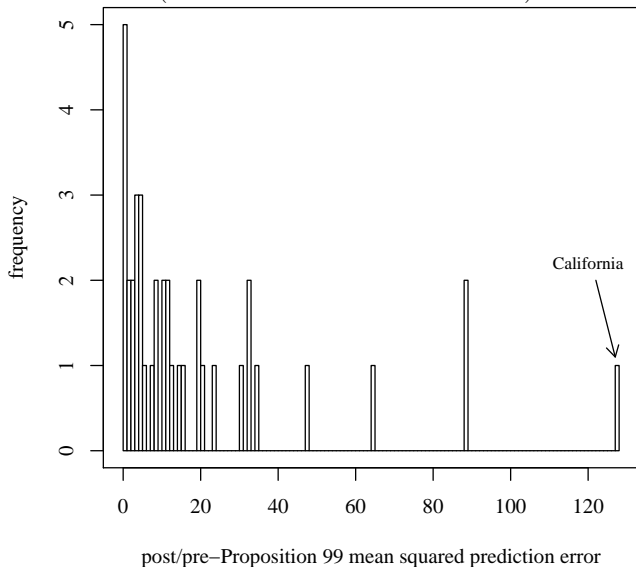
(PRE-PROP. 99 MSPE  $\leq$  2 TIMES PRE-PROP. 99 MSPE FOR CA)



# A primer on synthetic control estimation

Application: California tobacco control program

(ALL 38 STATES IN DONOR POOL)



## A primer on synthetic control estimation

- ▶ The availability of a well-defined procedure to select the comparison unit makes the estimation of the effects of placebo interventions feasible.
- ▶ The permutation method we just described does not attempt to approximate the sampling distributions of test statistics.
- ▶ Sampling-based inference is often complicated in a synthetic control setting, sometimes because of the absence of a well-defined sampling mechanism and sometimes because the sample is the same as the population.

## A primer on synthetic control estimation

- ▶ This mode of inference reduces to classical **randomization inference** (Fisher, 1935) when the intervention is randomly assigned, a rather improbable setting.
- ▶ More generally, this mode of inference evaluates significance relative to a **benchmark distribution** for the assignment process, one that is implemented directly in the data.

---

The uniform benchmark is often employed in practice, but departures from uniformity are possible (see, Firpo and Possebom, 2018).

## Why use synthetic controls?

- ▶ Compare to linear regression. Let:
  - ▶  $\mathbf{Y}_0$  be the  $(T - T_0) \times J$  matrix of post-intervention outcomes for the units in the donor pool.
  - ▶  $\bar{\mathbf{X}}_1$  and  $\bar{\mathbf{X}}_0$  be the result of augmenting  $\mathbf{X}_1$  and  $\mathbf{X}_0$  with a row of ones.
  - ▶  $\hat{\mathbf{B}} = (\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0')^{-1} \bar{\mathbf{X}}_0 \mathbf{Y}_0'$  collects the coefficients of the regression of  $\mathbf{Y}_0$  on  $\bar{\mathbf{X}}_0$ .
- ▶  $\hat{\mathbf{B}}' \bar{\mathbf{X}}_1$  is a regression-based estimator of the counterfactual outcome for the treated unit without the treatment.
- ▶ Notice that  $\hat{\mathbf{B}}' \bar{\mathbf{X}}_1 = \mathbf{Y}_0 \mathbf{W}^{reg}$ , with

$$\mathbf{W}^{reg} = \bar{\mathbf{X}}_0' (\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0')^{-1} \bar{\mathbf{X}}_1.$$

- ▶ The components of  $\mathbf{W}^{reg}$  sum to one, but may be outside  $[0, 1]$ , allowing extrapolation.

## Why use synthetic controls?

Application: German reunification

country $j$	$W_j^{reg}$	country $j$	$W_j^{reg}$
Australia	0.12	Netherlands	0.14
Austria	0.26	New Zealand	0.12
Belgium	0.00	Norway	0.04
Denmark	0.08	Portugal	-0.08
France	0.04	Spain	-0.01
Greece	-0.09	Switzerland	0.05
Italy	-0.05	United Kingdom	0.06
Japan	0.19	United States	0.13



## Why use synthetic controls?

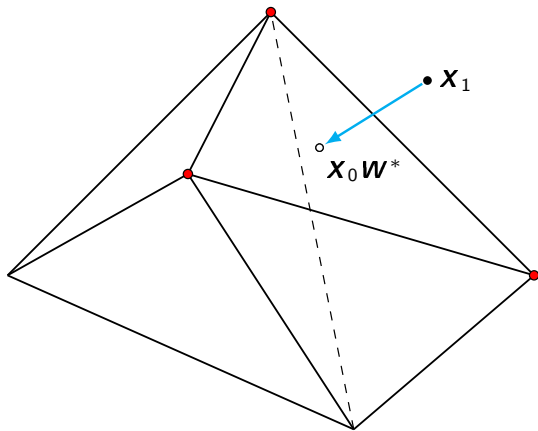
- ▶ **No extrapolation.** Synthetic control estimators preclude extrapolation outside the support of the data.
- ▶ **Transparency of the fit.** Linear regression uses extrapolation to guarantee a perfect fit of the characteristics of the treated unit,  $\mathbf{X}_0 \mathbf{W}^{reg} = \mathbf{X}_1$ , even when the untreated units are completely dissimilar in their characteristics to the treated unit. In contrast, synthetic controls make transparent the actual discrepancy between the treated unit and the convex hull of the units in the donor pool,  $\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}^*$ .
- ▶ **Safeguard against specification searches.** Synthetic controls do not require access to post-treatment outcomes in the design phase of the study, when synthetic control weights are calculated. Therefore, all design decisions can be made without knowing how they affect the conclusions of the study.

## Why use synthetic controls?

- ▶ **Safeguard against specification searches (cont.)**  
Synthetic control weights can be calculated and pre-registered before the post-treatment outcomes are realized, or before the actual intervention takes place, providing a safeguard against specification searches and  $p$ -hacking.
- ▶ **Transparency of the counterfactual.** Synthetic controls make explicit the contribution of each comparison unit to the counterfactual of interest.
- ▶ **Sparsity.** Because the synthetic control coefficients are proper weights and are sparse, they allow a precise interpretation of the nature of the estimate of the counterfactual of interest (and of potential biases).

# Why use synthetic controls?

Sparsity: Geometric interpretation



## Why use synthetic controls?

- ▶ In some cases, especially in applications with many treated units, the values of the predictors for some of the treated units may fall in the convex hull of the columns of  $\mathbf{X}_0$ .
- ▶ Then, synthetic controls are not unique or necessarily sparse.
- ▶ A modification of the synthetic control estimator that is always unique and sparse is developed in Abadie and L'Hour (2019).

## Synthetic controls for experimental design

- ▶ Suppose a ridesharing company wants to assess the impact of a new incentive pay program for drivers.
- ▶ To do so, the new incentive pay treatment will be applied as a pilot program in one market/city or in a few markets/cities.
- ▶ Which market or markets should they treat?
- ▶ Which market or markets should they use as a comparison/control?
- ▶ This is a setting where randomization of treatment may create defective designs where:
  - ▶ The treated market/markets are non-representative of the entire set of markets of interest.
  - ▶ Treated and control markets are very different in their characteristics.

## Synthetic controls for experimental design

- ▶ In these contexts, Abadie and Zhao (2021) propose:
  - ▶ Find a first set of weights that make a synthetic treated unit reproduce the features of the population of units of interest.
  - ▶ Create a synthetic control for the synthetic treated unit.
- ▶ In contrast to the observational case, in the experimental settings we have two synthetic control units: one treated and one untreated.
- ▶ Related ongoing work by Doudchenko and co-authors. Synthetic controls are widely used as experimental designs by business analytics units (e.g., Uber).

## Closing remarks

- ▶ Synthetic controls provide many practical advantages for the estimation of the effects of policy interventions and other events of interest.
- ▶ Like for any other statistical procedure (and especially for those aiming to estimate 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 (see Abadie, *JEL* 2021).

## Closing remarks

- ▶ Some open areas of research: sampling-based inference, external validity, sensitivity to model restrictions, estimation with multiple interventions, data driven selectors of  $v_h$ , mediation analysis ...
- ▶ An area of recent heightened interest regarding the use of synthetic controls is the design of experimental interventions.
- ▶ Results on robust and efficient computation of synthetic controls are scarce, and more research is needed on the computational aspects of this methodology.
- ▶ On the empirical side, many of the events and the policy interventions economists care about take place at an aggregate level, affecting entire aggregate units.