October 3, 2016
junyong@uwm.edu, www.junyongkim.com

Abadie, Alberto, Alexis Diamond and Jens Hainmueller (2010), Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program, *Journal of American Statistical Association* 105 (490), 493–505

Alberto Abadie
MIT

Alexis Diamond
Central European University

Jens Hainmueller
Stanford

1 Introduction

- Two problems that limit comparative case study research
  - Ambiguity about how comparison units are chosen
  - Uncertainty about the ability of the control group to reproduce the counterfactual outcome trajectory that the affected units would have experienced in the absence of the intervention or event of interest
- Advocate the use of data-driven procedures to construct suitable comparison groups
- SYNTHETIC CONTROL METHODS (SCM) makes explicit
  - Relative contribution of each control unit to the counterfactual of interest
  - Similarities between the unit affected by the event or intervention of interest and the synthetic control

2 Synthetic Control Methods for Comparative Case Studies

- Observe J+1 regions
- Only the first region is exposed to the intervention of interest, so that we have J remaining regions as potential controls
- The first region is uninterruptedly exposed to the intervention of interest after some initial intervention period

$$
\begin{aligned}
Y_{it}^{N} &= \text{outcome without intervention} \\
Y_{it}^{I} &= \text{outcome with intervention} \\
\alpha_{it} &= Y_{it}^{I} - Y_{it}^{N} = \text{effect of intervention} \\
\Rightarrow \alpha_{1t} &= Y_{1t}^{I} - Y_{1t}^{N} = Y_{1t} - Y_{1t}^{N}, \qquad t \geq \text{intervention moment } T_0
\end{aligned}
$$

- Factor model for Y_{it}^N is

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t^\top \mathbf{z}_i + \boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i + \varepsilon_{it}, \qquad \cdots (1)$$

$\delta_t =$ unknown comman factor with constant factor loadings
$\mathbf{z}_i =$ vector of covariates
$\boldsymbol{\theta}_t =$ parameters
$\boldsymbol{\lambda}_t =$ vector of factors
$\boldsymbol{\mu}_i =$ vector of factor loadings
$\varepsilon_{it} =$ transitory shocks

- Weighting vector $\mathbf{w}=(w_2,\ldots,w_{J+1})^\top$ s.t. $w_i \geq 0$ and $\sum w_i = 1$, then the synthetic control is

$$\sum_{j=2}^{J+1} w_j Y_{jt} = \delta_t + \boldsymbol{\theta}_t^\top \sum_{j=1}^{J+2} w_j \mathbf{z}_j + \boldsymbol{\lambda}_t^\top \sum_{j=1}^{J+2} w_j \boldsymbol{\mu}_j + \sum_{j=1}^{J+2} w_j \varepsilon_{jt}$$

- Suppose that there are $\mathbf{w}^*=(w_2^*,\ldots,w_{J+1}^*)$ (i.e. the RECIPE!) such that

$$\sum_{j=1}^{J+2} w_j^* Y_{j1} = Y_{11}, \qquad \sum_{j=1}^{J+2} w_j^* Y_{j2} = Y_{12}, \qquad \cdots, \qquad \sum_{j=1}^{J+2} w_j^* Y_{jT_0} = Y_{1T_0}, \qquad \text{and} \sum_{j=1}^{J+2} w_j^* \mathbf{z}_j = \mathbf{z}_1$$

- For $t \geq T_0$, the estimator of α_{1t} is

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=1}^{J+2} w_j^* Y_{jt} = \sum_{j=1}^{J+2} w_j^* \sum_{s=1}^{T_0} \boldsymbol{\lambda}_t^\top \left(\sum_{n=1}^{T_0} \boldsymbol{\lambda}_n \boldsymbol{\lambda}_n^\top \right)^{-1} \boldsymbol{\lambda}_s (\varepsilon_{jf} - \varepsilon_{1s}) - \sum_{j=1}^{J+2} w_j^* (\varepsilon_{jt} - \varepsilon_{1t})$$

- Why SCM?
  - DID model allows for the presence of unobserved confounders but restricts the effect of those confounders to be constant in time, so they can be eliminated by taking time differences
  - In this case, however, the effects of confounding unobserved characteristics to vary with time. Under this model, taking time differences does not eliminate the confounders

$$\sum_{j=1}^{J+2} w_j^* \mathbf{z}_j = \mathbf{z}_1, \qquad \text{and} \sum_{j=1}^{J+2} w_j^* \boldsymbol{\mu}_j = \boldsymbol{\mu}_1$$

  - SCM would provide an unbiased estimator of Y_{1t}^N

- Implementation

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{z}_1^\top & \bar{y}_1^{k_1} & \cdots & \bar{y}_1^{k_m} \end{pmatrix}^\top = \text{preintervention characteristics}$$

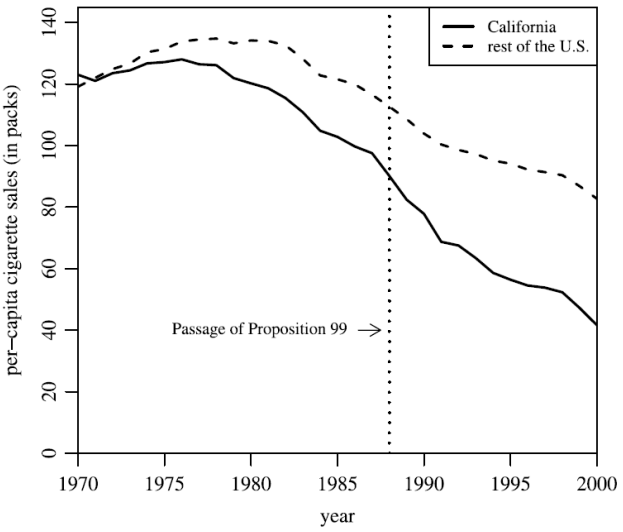
$$\mathbf{X}_0 = \begin{pmatrix} \mathbf{z}_2 & \cdots & \mathbf{z}_{J+1} \\ \bar{y}_2^{\mathbf{k}_1} & \cdots & \bar{y}_{J+1}^{\mathbf{k}_1} \\ \vdots & \vdots & \vdots \\ \bar{y}_2^{\mathbf{k}_m} & \cdots & \bar{y}_{J+1}^{\mathbf{k}_1} \end{pmatrix}$$

$$\mathbf{w}^* \equiv \underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{x}_1 - \mathbf{X}_0\mathbf{w}\|_\mathbf{V} = \sqrt{(\mathbf{x}_1 - \mathbf{X}_0\mathbf{w})^\top \mathbf{V} (\mathbf{x}_1 - \mathbf{X}_0\mathbf{w})}$$

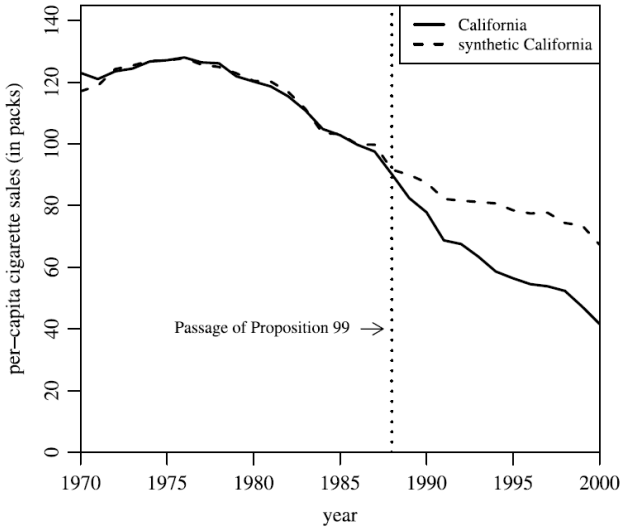
  - An optimal choice of **V** (symmetric and positive definite) assigns weights to linear combinations of the variables in $\mathbf{X}_0$ and $\mathbf{x}_1$ to minimize the mean square error of the synthetic control estimator
  - Given **V**, $\mathbf{w}^*(\mathbf{V})$ can be computed using data from the training period; then the matrix **V** can be chosen to minimize the mean squared prediction error produced by the weights $\mathbf{w}^*(\mathbf{V})$ during the validation period

3 Estimating the Effects of California's Proposition 99

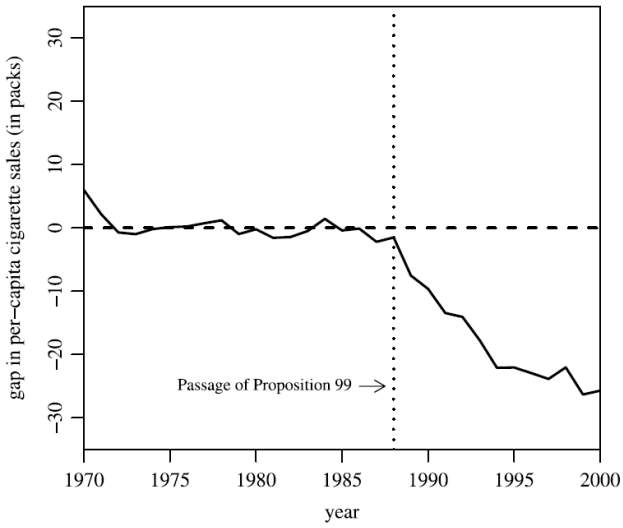
- California's Proposition 99 (1988)
  - The first modern-time large-scale tobacco control program in the US
- State-level panel data from 1970 to 2000
- Synthetic California is constructed as a weighted average of potential control states
  - With weights chosen so that the resulting synthetic CA best reproduces the values of a set of predictors of cigarette consumption in CA before the passage of Proposition 99
- Outcome of interest: Y_{it}=Annual per capita cigarette consumption at the state level (packs)
- Predictors: Average retail price of cigarettes, per capita state personal income (logged), the percentage of the population age 15–24, per capita beer consumption



CA vs. other states



CA vs. synthetic CA (TRACKS BETTER!)



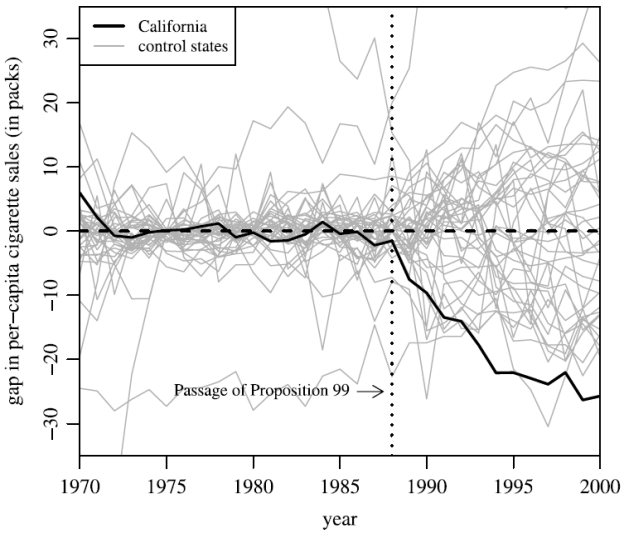
Gap between CA and synthetic CA

Table 1. Cigarette sales predictor means

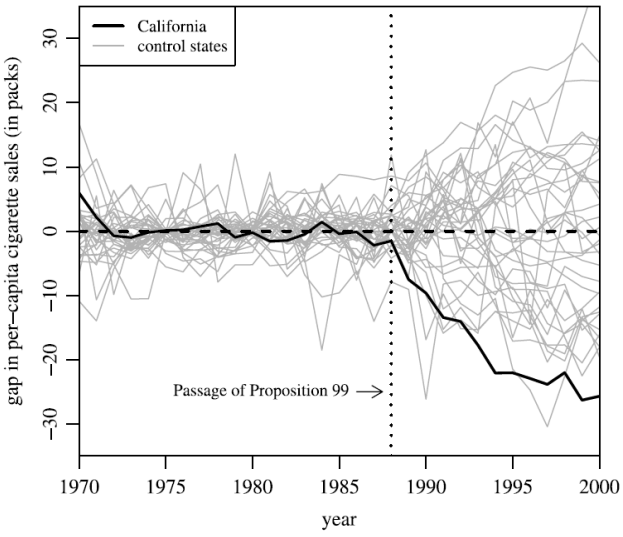
Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20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

(CA vs. other states) vs. (CA vs. synthetic CA)

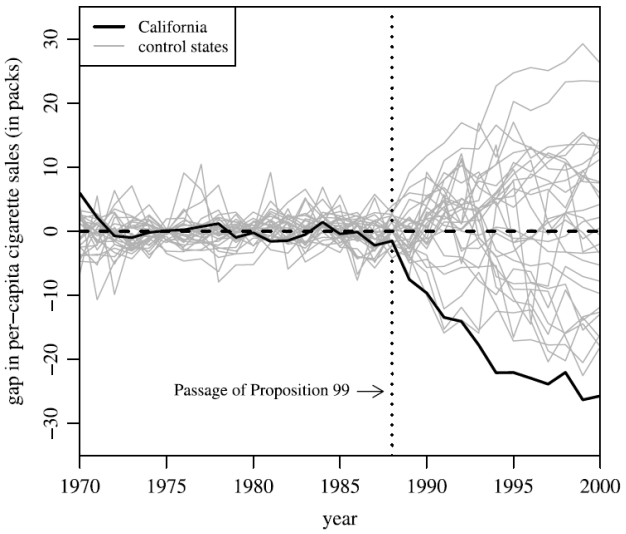
- Per capita sales in the synthetic CA very closely track the trajectory of this variable in CA for the entire pre-Proposition 99 period
- Estimates: The difference between per capita cigarette sales in California and in its synthetic version after the passage of Proposition 99
    - Results: For the entire 1989–2000 period cigarette consumption was reduced by an average of almost 20 packs per capita
    - A decline of approximately 25%
    - Robust regardless of which and how many predictor variables
- Placebo studies: Applying SCM to other states



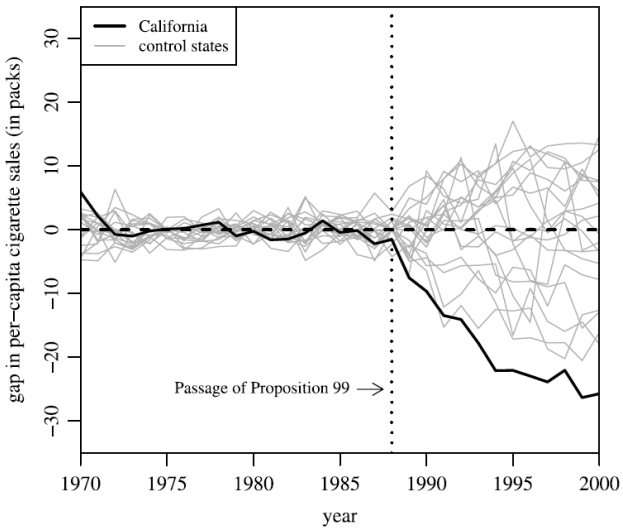
Cases of other states (38)



After removing ≥×20 MSPE (34)

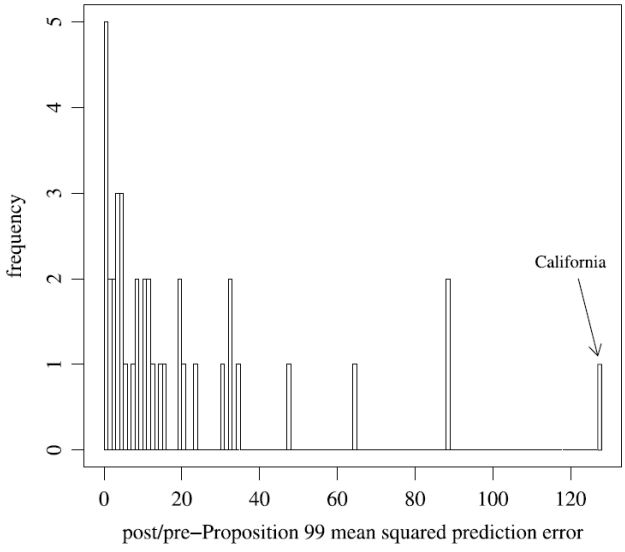


After removing ≥5× MSPE (29)



After removing ≥×2 MSPE (19)

- After removing unpredictable states (in terms of MSPE; i.e. nearly impossible to generate counterfactual data), the negative effect in CA is the lowest of all
    - The probability of estimating a gap of the magnitude of the gap for CA under a random permutation of the intervention in data is 5% (=1/20); a test level typically used in conventional tests of statistical significance
- The distribution of the ratios of post/pre-Proposition 99 MSPE
    - CA: post-Proposition 99 MSPE=130×pre-Proposition 99 MSPE



Post-Proposition effects are about 130 times bigger than pre-Proposition behaviors

4 Conclusion

- What plagues the empirical implementation of comparative case studies?
    - Inferential challenges
    - Ambiguity about the choice of valid control groups
- Advocate the use of data-driven procedures to select synthetic comparison units in comparative case studies
- Demonstrate the applicability of SCM by studying the effects of Proposition 99
    - A large-scale tobacco control program that CA passed in 1988
    - Results: Negative effects are much larger than prior estimates have reported

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

Recipes ($\mathbf{w}^*$) for synthetic CA