Introduction to Synthetic Control Method

Instructor: Ma, Jun

Renmin University of China

June 9, 2022

- 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* (ADH10)
- Abadie, A., Diamond, A., Hainmueller, J., 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* (ADH15)
- Matlab code for SCM: https://web.stanford.edu/~jhain/synthpage.html

Synthetic Control Method (SCM)

- "Arguably the most important innovation in the policy evaluation literature in the last 15 years" by Athey and Imbens (2017, *Journal of Economic Literature*)
- The SCM model framework and methodology were introduced in ADH10 and ADH15. Both received more than 4000 citations.
- A vast recent literature on SCM (methodology and empirical studies).
- Two views: 1. an extension of difference-in-differences for aggregate-level data; 2. a data-driven method for quantitative comparative case studies.

Difference-in-Differences (D-i-D) Model

- To analyze the effect of a policy intervention on the outcome, the D-i-D approach uses the data of the treated and controlled groups, before and after the policy intervention.
- $Y_t^{(d)}$: potential outcomes for an arbitrary individual in the population.
- $t = 0$: pre-intervention period; $t = 1$: post-intervention period; $d = 0$: no intervention; $d = 1$: intervention.
- In the SCM literature, treatment/intervention/event/policy are often used interchangeably.

- D_t: a dummy variable, e.g., whether or not influenced by the policy intervention. In the pre-intervention period, $D_0 = 0$. In the post-intervention period, some get influenced by the intervention.
- Observe: $Y_t = D_t Y_t^{(1)} + (1 - D_t)\, Y_t^{(0)}$.
- Common trend assumption (CTA): $\mathrm{E}\left[Y_1^{(0)} - Y_0^{(0)} \mid D_1 = 1\right] = \mathrm{E}\left[Y_1^{(0)} - Y_0^{(0)} \mid D_1 = 0\right]$. In the absence of the treatment, the average outcome for the treated and the average outcome for the non-treated would have experienced the same variation over time.
- $\mathrm{E}\left[Y_1^{(0)} \mid D_1 = 1\right]$ is a counterfactual quantity.

- Under the CTA,

$$\begin{aligned}
\text{ATT} &= \mathrm{E}\left[Y_1^{(1)} - Y_1^{(0)} \mid D_1 = 1\right] \\
&= \mathrm{E}\left[Y_1^{(1)} \mid D_1 = 1\right] \\
&\quad - \left\{\mathrm{E}\left[Y_1^{(0)} - Y_0^{(0)} \mid D_1 = 0\right] + \mathrm{E}\left[Y_0^{(0)} \mid D_1 = 1\right]\right\} \\
&= \mathrm{E}\left[Y_1^{(1)} - Y_0^{(0)} \mid D_1 = 1\right] - \mathrm{E}\left[Y_1^{(0)} - Y_0^{(0)} \mid D_1 = 0\right] \\
&= \mathrm{E}\left[Y_1 - Y_0 \mid D_1 = 1\right] - \mathrm{E}\left[Y_1 - Y_0 \mid D_1 = 0\right].
\end{aligned}$$

The average treatment effect (at $t = 1$) on the treated is identified.

A Linear Structural Panel Model

- The D-i-D model is often related to the following (structural) linear panel data model. The outcome for the i-th individual is generated by

$$Y_{i,t} = \tau_{i,t} D_{i,t} + \mu_i + \delta_t + \epsilon_{i,t},$$

  where $D_{i,0} = 0$, $\forall i$, δ_t is a non-random time fixed effect that is common across individuals, μ_i is a time invariant individual fixed effect (μ_i and $D_{i,1}$ can be correlated) and $\epsilon_{i,t}$ is a random shock: $\mathrm{E}\left[\epsilon_{i,t} \mid D_{i,1}\right] = \mathrm{E}\left[\epsilon_{i,t}\right] = 0$.
- The potential outcomes in are $Y_{i,t}^{(0)} = \mu_i + \delta_t + \epsilon_{i,t}$ and $Y_{i,t}^{(1)} = \tau_{i,t} + \mu_i + \delta_t + \epsilon_{i,t}$. $\tau_{i,t} = Y_{i,t}^{(1)} - Y_{i,t}^{(0)}$ is the individual treatment effect.

- The CTA is satisfied:

$$\mathrm{E}\left[Y_{i,1}^{(0)} - Y_{i,0}^{(0)} \mid D_{i,1} = 0\right] = \mathrm{E}\left[Y_{i,1}^{(0)} - Y_{i,0}^{(0)} \mid D_{i,1} = 1\right] = (\delta_1 - \delta_0) .$$

- The ATT at $t = 1$ ($\mathrm{E}\left[\tau_{i,1} \mid D_{i,1} = 1\right]$):

$$\begin{array}{rcl}
\mathrm{E}\left[Y_{i,1} \mid D_{i,1} = 1\right] & = & \mathrm{E}\left[\tau_{i,1} \mid D_{i,1} = 1\right] + \mathrm{E}\left[\mu_i \mid D_{i,1} = 1\right] + \delta_1 \\
\mathrm{E}\left[Y_{i,0} \mid D_{i,1} = 1\right] & = & \mathrm{E}\left[\mu_i \mid D_{i,1} = 1\right] + \delta_0 \\
\mathrm{E}\left[Y_{i,1} \mid D_{i,1} = 0\right] & = & \mathrm{E}\left[\mu_i \mid D_{i,1} = 0\right] + \delta_1 \\
\mathrm{E}\left[Y_{i,0} \mid D_{i,1} = 0\right] & = & \mathrm{E}\left[\mu_i \mid D_{i,1} = 0\right] + \delta_0
\end{array}$$

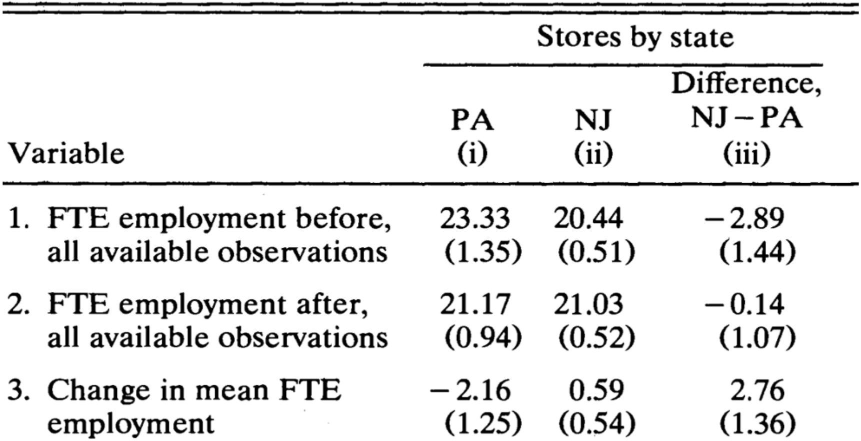
and

$$\mathrm{E}\left[\tau_{i,1} \mid D_{i,1} = 1\right] = \mathrm{E}\left[Y_{i,1} - Y_{i,0} \mid D_{i,1} = 1\right] - \mathrm{E}\left[Y_{i,1} - Y_{i,0} \mid D_{i,1} = 0\right]$$

A D-i-D Example: Card and Kruger (1994)

- Card and Kruger (1994) estimates the effect of a minimum wage increase in New Jersey on employment using a D-i-D model.
- In April 1992 NJ increased the state minimum wage from $4.25 to $5.05. The neighboring state, Pennsylvania, had minimum wage stayed at $4.25. Survey data on more than 400 fast food stores both in NJ and PA, in February and November (before and after intervention).
- The outcome: full-time-equivalent employment.

Card and Kruger (1994), Table 3.



	Stores by state		
			Difference,
	PA	NJ	NJ – PA
Variable	(i)	(ii)	(iii)
1. FTE employment before,	23.33	20.44	–2.89
all available observations	(1.35)	(0.51)	(1.44)
2. FTE employment after,	21.17	21.03	–0.14
all available observations	(0.94)	(0.52)	(1.07)
3. Change in mean FTE	–2.16	0.59	2.76
employment	(1.25)	(0.54)	(1.36)

SCM as an Extension of D-i-D

- Card and Kruger (1994) uses individual-level survey data (for small businesses). What if we have only aggregate data (i.e., average number of employees in all small businesses in February and November)?
- Policy interventions are often implemented at an aggregate level. Aggregate administrative data are often available and easily accessible than individual-level survey data.
- In the Card and Kruger (1994) example, standard errors reflect and estimate the uncertainty from the sampling error. For this example, it seems that the standard error should be taken as zero if aggregate data are available (no sampling error), which means we have zero uncertainty?
- There is still uncertainty about the treatment effect, even when we use aggregate data. We do not have perfect information about potential outcomes and are facing uncertainty about the ability of the control group to reproduce the counterfactual for the treated units.

Quantitative Comparative Case Studies

- Quantitative comparative case studies use aggregate data from one treated unit and a small set of control units (often aggregate units, i.e., city/state/country), where a policy intervention affects one unit, but not others.
- Compare the evolution of an outcome for the treated unit to the evolution of the same outcome for some control units that are deemed as being “similar” to the treated unit.
- E.g., in ADH15, the treated unit is the former West Germany. You may think of USA or Austria as being similar to West Germany. But selection of a control unit/group often incurs subjectivity.
- SCM: a data-driven approach to “synthesize” a suitable control unit.
- A “synthetic” convex combination (weighted average) of control units may do a better job of reproducing the characteristics and the evolution of the outcome of the treated unit than any one unit alone.

SCM Model Framework

- For units $i = 1, ..., J + 1$, Unit 1 is the treated unit. Units 2 to $J + 1$ are the "donor pool" (potential comparison/control units).
- Time periods: $t = 1, ..., T$; pre-treatment periods: $t = 1, ..., T_0$; post-treatment periods: $t = T_0 + 1, ..., T$.
- Unit 1 is treated or affected by the policy intervention starting in period $T_0 + 1$.
- Which units should be included in the donor pool?
    - Units whose outcome is determined in the same way as the treated unit.
    - Control units should not become treated in any of the post-treatment period.
    - No spillover effect from treatment in any of the post-treatment period.

- In the potential outcome framework, we never observe both $Y_{i,t}^{(0)}$ and $Y_{i,t}^{(1)}$.
- Let $Y_{i,t} = D_{i,t}Y_{i,t}^{(1)} + \left(1 - D_{i,t}\right) Y_{i,t}^{(0)}$ be the observed outcome.
- We observe:

$$\begin{bmatrix} Y_{1,T} & Y_{2,T} & \cdots & Y_{J+1,T} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1,T_0+1} & Y_{2,T_0+1} & \cdots & Y_{J+1,T_0+1} \\ Y_{1,T_0} & Y_{2,T_0} & \cdots & Y_{J+1,T_0} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1,1} & Y_{2,1} & \cdots & Y_{J+1,1} \end{bmatrix} = \begin{bmatrix} Y_{1,T}^{(1)} & Y_{2,T}^{(0)} & \cdots & Y_{J+1,T}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1,T_0+1}^{(1)} & Y_{2,T_0+1}^{(0)} & \cdots & Y_{J+1,T_0+1}^{(0)} \\ Y_{1,T_0}^{(0)} & Y_{2,T_0}^{(0)} & \cdots & Y_{J+1,T_0}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1,1}^{(0)} & Y_{2,1}^{(0)} & \cdots & Y_{J+1,1}^{(0)} \end{bmatrix}.$$

- The quantity of interest is the treatment effect on Unit 1 from period $T_0 + 1$ to T:

$$\tau_{1,t} = Y_{1,t}^{(1)} - Y_{1,t}^{(0)} = Y_{1,t} - Y_{1,t}^{(0)}$$

  for $t = T_0 + 1, ..., T$, i.e., the treatment effect on the treated unit in the post-treatment periods.

- The problem is that we do not observe $Y_{1,t}^{(0)}$, for $t = T_0 + 1, ..., T$, and we want to estimate it using observed outcomes in the donor pool.
- Actually, we can do a D-i-D:

$$\begin{aligned} \widehat{\tau}_{1,t} &= \left(Y_{1,t} - Y_{1,T_0}\right) - \left(\frac{1}{J}\sum_{i=2}^{J+1} Y_{i,t} - \frac{1}{J}\sum_{i=2}^{J+1} Y_{i,T_0}\right) \\ &= Y_{1,t} - \left\{Y_{1,T_0} + \frac{1}{J}\sum_{i=2}^{J+1}\left(Y_{i,t} - Y_{i,T_0}\right)\right\}. \end{aligned}$$

  Note that it gives equal weights $(1/J)$ to any unit in the donor pool.

- In contrast, SCM gives weights to a sub-group in the donor pool in a data-driven manner.
- An SC is a vector of weights $\boldsymbol{w} = (w_2, ..., w_J)$ associated with each of the available J donor units, which satisfy: $\sum_{i=2}^{J} w_i = 1$ and $w_i \geq 0 \; \forall i$.
- The goal is to select $\boldsymbol{w}$ such that the characteristics of the treated unit are best resembled by the characteristics of the SC.

- $\boldsymbol{X}_i = \left(X_{i,1}, \ldots, X_{i,k}\right)^\top$: the k-dimensional pre-intervention characteristics for the treated unit ($i = 1$) or the control unit ($i = 2, \ldots, J + 1$).
- The pre-intervention characteristics contain pre-intervention outcomes $\left(Y_{i,1}, \ldots, Y_{i,T_0}\right)^\top$ and possibly other predictors $\boldsymbol{Z}_i$ of the post-intervention outcome:

$$\boldsymbol{X}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,T_0} \\ \boldsymbol{Z}_i \end{pmatrix}.$$

- $\mathbf{X}_0 = [\boldsymbol{X}_2, \ldots, \boldsymbol{X}_{J+1}]$: the $k \times J$ matrix containing the characteristics of the control units.

- We want to choose $\boldsymbol{w}$ such that treatment/control units are similar in terms of:
  - Pre-treatment outcomes: $Y_{1,t} \approx \sum_{j=2}^{J+1} w_j Y_{j,t}$.
  - Covariates that are predictive of post-intervention outcomes: $Z_1 \approx \sum_{j=2}^{J+1} w_j Z_j$.
- For some $k \times k$ diagonal matrix $\mathbf{V} = \text{diag}\{v_1, ..., v_k\}$ ($v_j \geq 0$ $\forall j$), denote $\|\boldsymbol{x}\|_{\mathbf{V}} = \sqrt{\boldsymbol{x}^\top \mathbf{V} \boldsymbol{x}}$.
- Denote $\Delta_J = \left\{(w_2, ..., w_{J+1}) \in \mathbb{R}^J : \sum_{i=2}^{J} w_i = 1; w_i \geq 0 \, \forall i\right\}$.
- Given $v_1, ..., v_k$, ADH10 proposes to choose the SC $\boldsymbol{w}^* = \left(w_2^*, ..., w_{J+1}^*\right)^\top$ to minimize $\|\boldsymbol{X}_1 - \mathbf{X}_0 \boldsymbol{w}\|_{\mathbf{V}}^2$ subject to the constraint $\boldsymbol{w} \in \Delta_J$:

$$\begin{aligned}\boldsymbol{w}^* &= \underset{\boldsymbol{w} \in \Delta_J}{\text{argmin}} \, \|\boldsymbol{X}_1 - \mathbf{X}_0 \boldsymbol{w}\|_{\mathbf{V}}^2 \\ &= \underset{(w_2, ..., w_{J+1})^\top \in \Delta_J}{\text{argmin}} \sum_{m=1}^{k} v_m \left(X_{1,m} - w_2 X_{2,m} - \cdots - w_{J+1} X_{J+1,m}\right)^2 .\end{aligned}$$

- v_m is a weight that reflects the importance/predictive power of the m-th variable that we use to measure the distance between treated and control units.

- The SC estimator:

$$\widehat{\tau}_{1,t} = Y_{1,t} - \sum_{i=2}^{J+1} w_i^* Y_{i,t}$$

  for $t = T_0 + 1, ..., T$.

- Note that the constraints $\left(w_2^*, ..., w_{J+1}^*\right)^\top \in \Delta_J$ prevent interpolation outside of the support of the data, i.e., the counterfactual $\sum_{i=2}^{J+1} w_i^* Y_{i,t}$ cannot take a value larger than the maximum or smaller than the minimum of $\left\{Y_{2,t}, ..., Y_{J+1,t}\right\}$, observed for the control units.
- Note that the SC depends on the weights put on pre-intervention characteristics $\mathbf{V}$: $\boldsymbol{w}^* = \boldsymbol{w}^* (\mathbf{V})$.

How to Choose $\mathbf{V}$?

- ADH10 proposes to set $(v_1, ..., v_k)$ to minimize the mean square prediction error (MSPE) over the pre-treatment periods:

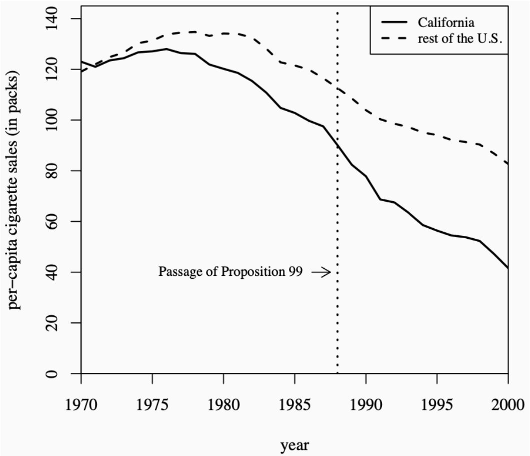
  $$\text{MSPE}(\mathbf{V}) = \sum_{t=1}^{T_0} \left\{ Y_{1,t} - \sum_{i=2}^{J+1} w_i^* (\mathbf{V}) Y_{i,t} \right\}^2 .$$

  Over the pre-treatment periods, the computed treatment effect must be zero, since the treatment has not been implemented yet.
- ADH15 proposes cross-validation.
  - Split the pre-intervention T_0 periods into initial training periods $t = 1, ..., t_0$ and subsequent validation periods $t = t_0 + 1, ..., T_0$.
  - Given any $\mathbf{V}$, for each validation periods, $t = t_0 + 1, ..., T_0$, compute $Y_{1,t} - \sum_{i=2}^{J+1} w_i^* (\mathbf{V}) Y_{i,t}$, where $\left(w_2^* (\mathbf{V}), ..., w_{J+1}^* (\mathbf{V})\right)^\top$ solves $\min_{\boldsymbol{w} \in \Delta_J} \|\boldsymbol{X}_1 - \mathbf{X}_0 \boldsymbol{w}\|_{\mathbf{V}}^2$ with $\boldsymbol{X}'s$ measured in the training periods $t = 1, ..., t_0$.
  - Choose $\mathbf{V}$ to minimize $\sum_{t=t_0+1}^{T_0} \left\{Y_{1,t} - \sum_{i=2}^{J+1} w_i^* (\mathbf{V}) Y_{i,t}\right\}^2$. Use the resulting $\mathbf{V}$ and the predictors for the last t_0 periods before the intervention to calculate $\boldsymbol{w}^* (\mathbf{V})$.

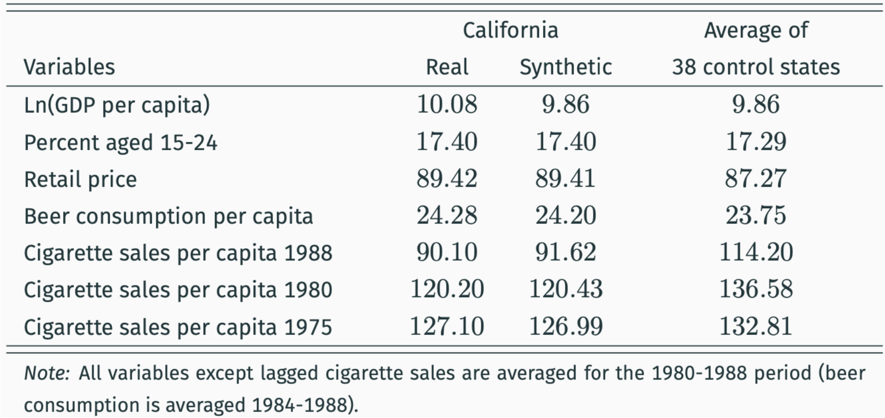
ADH10: Background and Results

- Proposition 99: the first modern-time large-scale tobacco control program in USA.
- In 1988, California passed comprehensive tobacco control legislation. This was a package of measures that included a tax increase, more spending to anti-smoking health initiatives and anti-smoking media campaigns.
- ADH10 investigates the effect of this legislation on cigarette consumption in California using SCM.
- Outcome variable: Per capita cigarette sales (packs).
- Time: 1970 to 2000, $T_0 = 1988$.
- All states which passed similar legislation in 1989-2000 are excluded from the donor pool.

- Predictors of smoking prevalence are: average retail price of cigarettes, per capita state personal income (logged), the percentage of the population age 15–24, and per capita beer consumption. These variables are averaged over the 1980–1988 period and augmented by adding three years of lagged smoking consumption (1975, 1980, and 1988).
- To evaluate the effect of Proposition 99 on cigarette smoking in California, the central question is how cigarette consumption would have evolved in California after 1988 in the absence of Proposition 99.



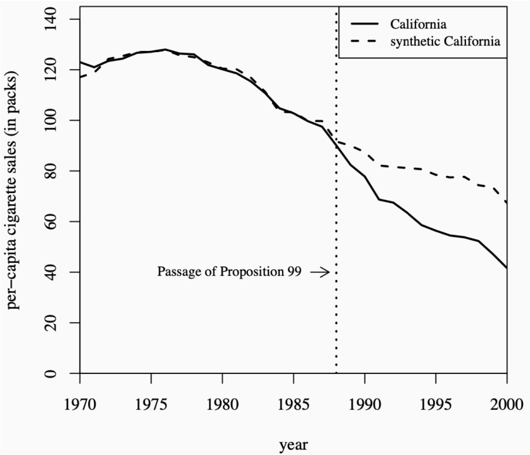
- The rest of USA may not provide a suitable comparison group for California to study the effects of Proposition 99 on per capita smoking. Even before the passage of Proposition 99 the time series of cigarette consumption in California and in the rest of USA differed notably.



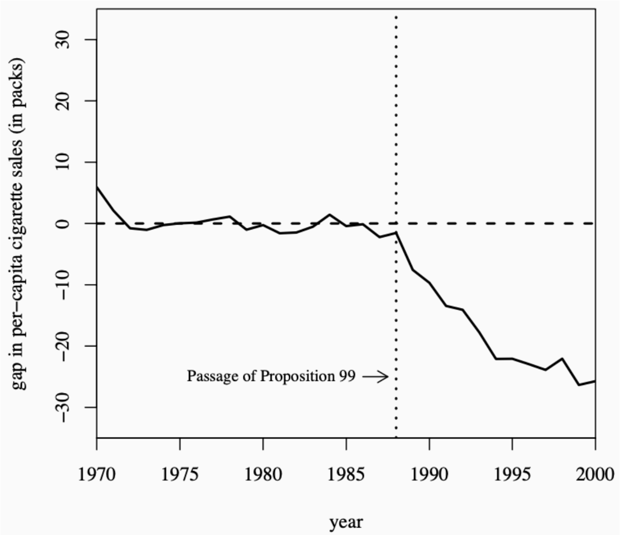
	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20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

- The average of states that did not implement a large-scale tobacco-control program in 1989–2000 does not seem to provide a suitable control group for California.



- Sparse weights: Colorado 0.164; Conneticut 0.069; Montana 0.199; Nevada 0.234; Utah 0.334; Rest of states 0.
- Per capita sales in the synthetic California very closely track the trajectory of this variable in California for the entire pre-Proposition 99 period.


gap in per−capita cigarette sales (in packs)
30
20
10
0
−10
−20
−30
Passage of Proposition 99 ⟶
1970
1975
1980
1985
1990
1995
2000
year

Bias Bound

- ADH10 gives a result on the bias of the SC estimator.
- ADH10 assumes a factor model:

$$Y_{i,t}^{(0)} = \delta_t + \boldsymbol{Z}_i^\top \boldsymbol{\theta}_t + \boldsymbol{\lambda}_t^\top \boldsymbol{\mu}_i + \epsilon_{i,t},$$

  where δ_t is a time fixed effect, $\boldsymbol{\theta}_t$ is a vector of time varying parameters, $\boldsymbol{Z}_i$ are the observed covariates, $\boldsymbol{\lambda}_t \in \mathbb{R}^F$ are unobserved common factors, $\boldsymbol{\mu}_i \in \mathbb{R}^F$ are unobserved factor loadings and $\epsilon_{i,t}$ are unobserved transitory shocks.
- $\boldsymbol{\lambda}_t$ are macroeconomic factors that affect each unit differently through $\boldsymbol{\mu}_i$.

Assumption

The transitory shocks $\left\{\epsilon_{i,t}\right\}_{i=1,\dots,J+1;t=1,\dots,T}$ are i.i.d. across both i and t, with mean zero and variance σ^2. For some even integer m, $\rho_m = \mathrm{E}\left[\left|\epsilon_{i,t}\right|^m\right] < \infty$.

Assumption

Match is perfect: $\|\boldsymbol{X}_1 - \mathbf{X}_0\boldsymbol{w}^*\|_{\mathbf{V}}^2 = 0$. *The SC perfectly reproduces the treated unit. In most applications, this condition holds approximately. I.e.,* $\|\boldsymbol{X}_1 - \mathbf{X}_0\boldsymbol{w}^*\|_{\mathbf{V}}^2$ *is small.*

Assumption

Let $\xi(M)$ *denote the smallest eigenvalue of* $M^{-1}\sum_{t=T_0-M+1}^{T_0}\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t^\top$.
$\xi(M) \geq c_\xi > 0$, *for any positive integer* M. $|\boldsymbol{\lambda}_t|_\infty \leq \overline{\lambda}$, $\forall t = 1, \ldots, T$ *(for* $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $|\boldsymbol{x}|_\infty = \max\{|x_1|, \ldots, |x_n|\}$*).*

Theorem

The bias is bounded by:

$$\mathrm{E}\left[\widehat{\tau}_{1,t} - \tau_{1,t}\right] \leq C(m)^{1/m}\left(\frac{F\overline{\lambda}^2}{c_\xi}\right)J^{1/m}\max\left\{\frac{\rho_m^{1/m}}{T_0^{1-1/m}}, \frac{\sigma}{\sqrt{T_0}}\right\}\xrightarrow[T_0\uparrow\infty]{} 0,$$

where $C(m) = \mathrm{E}\left[(R-1)^m\right]$ *with* $R \sim \mathrm{Poisson}(1)$.

- If there are many periods before the intervention, then matching on the pre-intervention outcomes allows us to control for heterogeneous response to unobservables ($\boldsymbol{\mu}_i$).
- It is easy to see that if the SC perfectly reproduces $\boldsymbol{\mu}_1$, i.e., $\boldsymbol{\mu}_1 = \sum_{i=2}^{J+1} w_i^* \boldsymbol{\mu}_i$, then the bias of the SC estimator would be zero.
- $Y_{i,t}^{(0)}$ is a function of $\boldsymbol{Z}_i$ and $\boldsymbol{\mu}_i$. So matching $Y_{i,t}^{(0)}$ is equivalent to matching $\boldsymbol{\mu}_i$. $\boldsymbol{\mu}_1 = \sum_{i=2}^{J+1} w_i^* \boldsymbol{\mu}_i$ holds approximately. Only units that are alike in both $\boldsymbol{Z}$ and unobserved $\boldsymbol{\mu}$ could produce similar patterns/trends of outcome over extended periods before the intervention.
- The bias decays to zeros as $T_0 \uparrow \infty$. So in the research design, it would be better to have data in more pre-intervention periods.

Inference/Placebo Test

- Inference/hypothesis testing in the SC context is a bit tricky. In most applications, J is relatively small. So the usual asymptotic framework is no longer appropriate.
- ADH10 uses aggregate data from 38 states. It makes little sense to think of the data as being a sample of individuals from a large population. In this case, the sample is the same as the population.
- ADH10 proposes to adopt a framework where the uncertainty does not come from sampling but comes from assignment of treatment.
- ADH10 and ADH 15 propose to use a modification of the classical inference method called “permutation test”.

Permutation Test in a Simple Framework

- n observations on the outcome: $Y_i = D_i Y_i^{(1)} + (1 - D_i) Y_i^{(0)}$, $D_i \in \{0, 1\}$.
- Assume that $\left\{\left(Y_i^{(1)}, Y_i^{(0)}\right)\right\}_{i=1}^{n}$ are all fixed but unobserved. The reason that the observed outcomes are random is that D_i is random and determines which of the two potential outcomes is observed.
- Fisher's sharp null hypothesis: $\mathrm{H}_0 : Y_i^{(1)} = Y_i^{(0)}, \forall i$.
- Denote $\overline{Y}^{(1)} = \sum_{i=1}^{n} D_i Y_i / \sum_{i=1}^{n} D_i$, $\overline{Y}^{(0)} = \sum_{i=1}^{n} (1 - D_i) Y_i / \sum_{i=1}^{n} (1 - D_i)$ and $n_1 = \sum_{i=1}^{n} D_i$. Calculate $T = \left|\overline{Y}^{(1)} - \overline{Y}^{(0)}\right|$. T is likely to be large if H_0 is false. What is the critical value?
- Suppose $n = 6$. Observed assignment vector is $(1, 1, 0, 0, 0, 0)$. If H_0 is true and the assignment vector had been $(1, 0, 0, 1, 0, 0)$, the observed outcomes would not change since $Y_i = Y_i^{(1)} = Y_i^{(0)}$.

- The collection of possible assignment vectors: $\mathcal{D} = \left\{ \boldsymbol{d} = (d_1, ..., d_n) \in \{0, 1\}^n : \sum_{i=1}^{n} d_i = n_1 \right\}$ which contains $C_n^{n_1}$ vectors. For each assignment vector $\boldsymbol{d} \in \mathcal{D}$, let $T(\boldsymbol{d})$ be the corresponding statistic (absolute difference between treatment and control group means).
- Fisher's exact p-value:

$$p - \text{value} = \frac{1}{C_n^{n_1}} \sum_{\boldsymbol{d} \in \mathcal{D}} 1 \left(T(\boldsymbol{d}) \geq T \right).$$

  Reject H_0 if p-value is less than significance level.
- This mode of inference is known as "permutation inference".
  - Calculate the "true" test-statistic under the actual treatment assignment.
  - Calculate the permutation distribution of the test-statistic under alternative treatment assignments.
  - Assess whether the "true" test-statistic is unlikely under the permutation distribution.

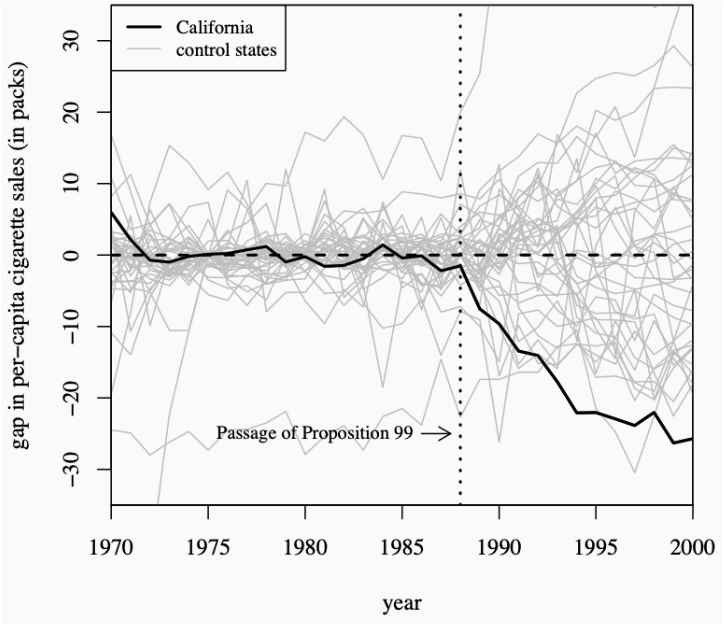
Permutation Test for SCM

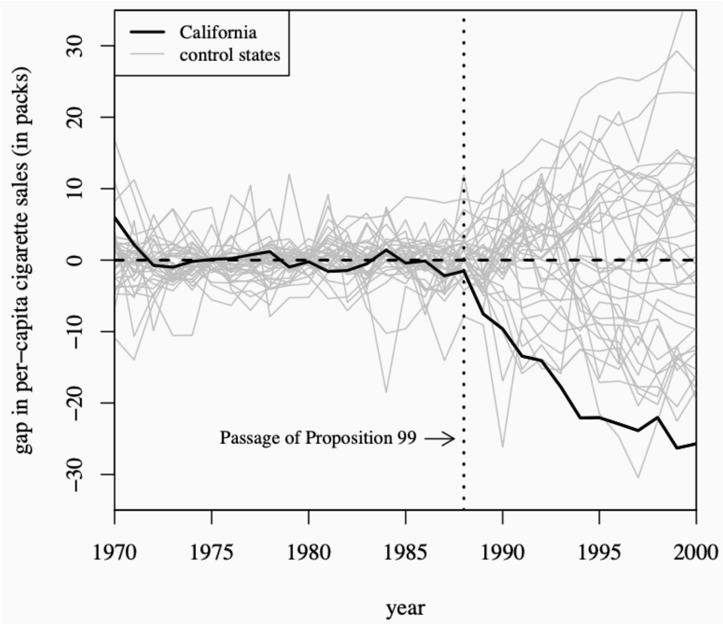
- Only one treated unit. The null hypothesis is $\mathrm{H}_0 : Y_{i,t}^{(0)} = Y_{i,t}^{(1)},\ t = T_0 + 1, ..., T, \forall i.$
- Iteratively assign treatment to each unit (treated and donor pool), $i = 1, ..., J + 1$, calculate

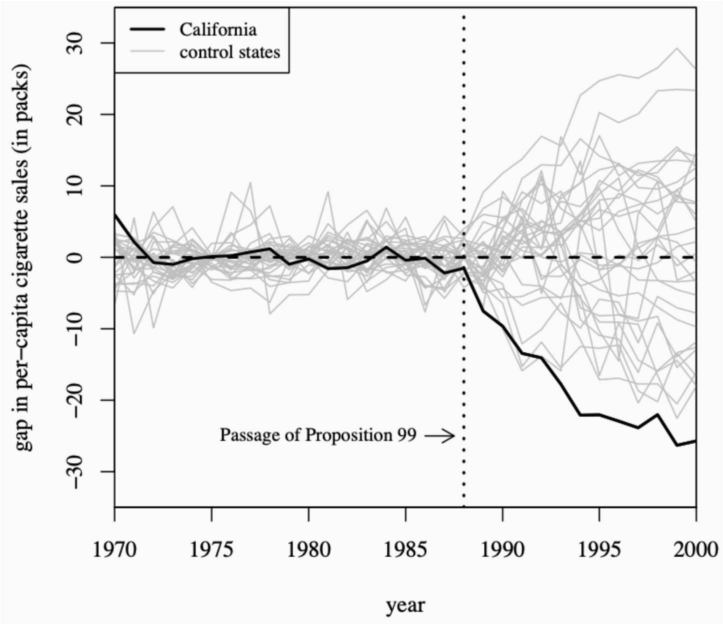
$$\begin{aligned} R_i^{\texttt{post}} &= \sqrt{\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{i,t} - \widehat{Y}_{i,t}\right)^2} \\ R_i^{\texttt{pre}} &= \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{i,t} - \widehat{Y}_{i,t}\right)^2} \end{aligned}$$

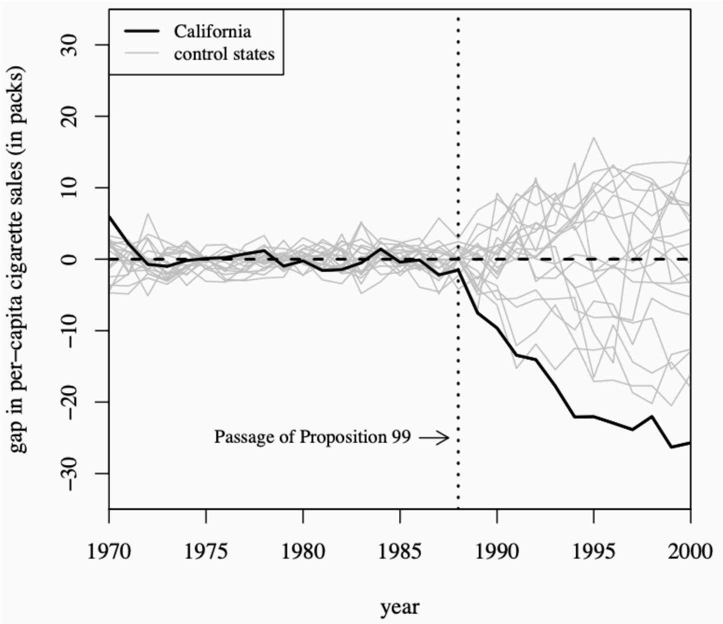
where $\widehat{Y}_{i,t}$ is the outcome on period t produced by a synthetic control when unit i is coded as treated and using all other J units to construct the donor pool.

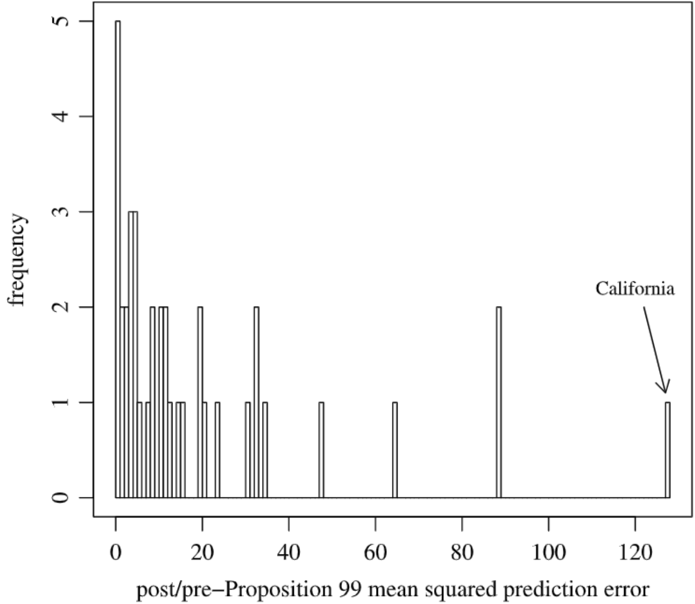
- More confident that the treatment effect is different from zero when the estimated post-intervention treatment effects ($R_1^{\texttt{post}}$) are larger.
- Less confident when the estimated pre-intervention treatment effects ($R_1^{\texttt{pre}}$) are larger (a large $R_1^{\texttt{post}}$ could be due to inability of the SC to reproduce counterfactuals).
- The treatment effect is deemed to be significantly different from zero if $R_1^{\texttt{post}}$ is extreme relative to the permutation distribution of $\left\{R_i^{\texttt{post}}\right\}_{i=1}^{J+1}$. The p-value is $(J+1)^{-1}\sum_{i=1}^{J+1} 1\left(R_1^{\texttt{post}} \geq R_i^{\texttt{post}}\right)$.
- One potential complication with this procedure is that even if an SC is able to closely fit the trajectory of the outcome before intervention, the same may not be true for all units in the donor pool.
    - Discard the i-th donor unit if $R_i^{\texttt{pre}}$ is substantially larger than $R_1^{\texttt{pre}}$.
    - Use $r_i = R_i^{\texttt{post}}/R_i^{\texttt{pre}}$. The p-value is $(J+1)^{-1}\sum_{i=1}^{J+1} 1\left(r_1 \geq r_i\right)$.


California
control states
gap in per−capita cigarette sales (in packs)
30
20
10
0
−10
−20
−30
Passage of Proposition 99 ⟶
1970
1975
1980
1985
1990
1995
2000
year


California
control states
gap in per−capita cigarette sales (in packs)
30
20
10
0
−10
−20
−30
Passage of Proposition 99 →
1970
1975
1980
1985
1990
1995
2000
year

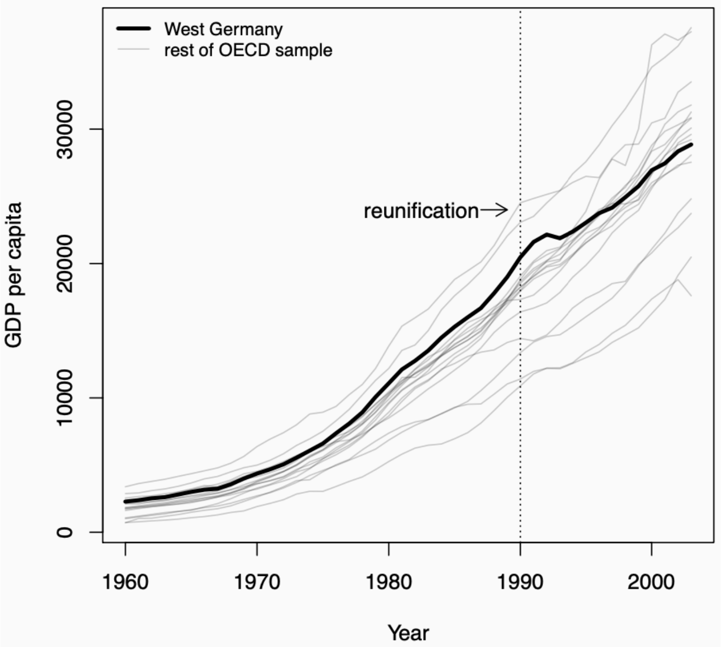

California
control states
gap in per–capita cigarette sales (in packs)
30
20
10
0
−10
−20
−30
Passage of Proposition 99
1970
1975
1980
1985
1990
1995
2000
year

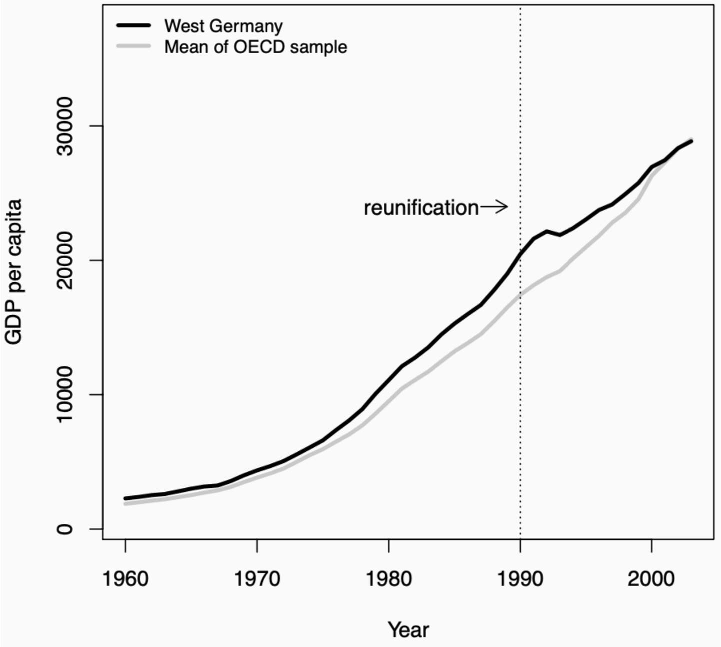

California
control states
gap in per−capita cigarette sales (in packs)
30
20
10
0
−10
−20
−30
Passage of Proposition 99 ⟶
1970
1975
1980
1985
1990
1995
2000
year


frequency
0
1
2
3
4
5
California
0
20
40
60
80
100
120
post/pre−Proposition 99 mean squared prediction error

ADH15 Background and Results

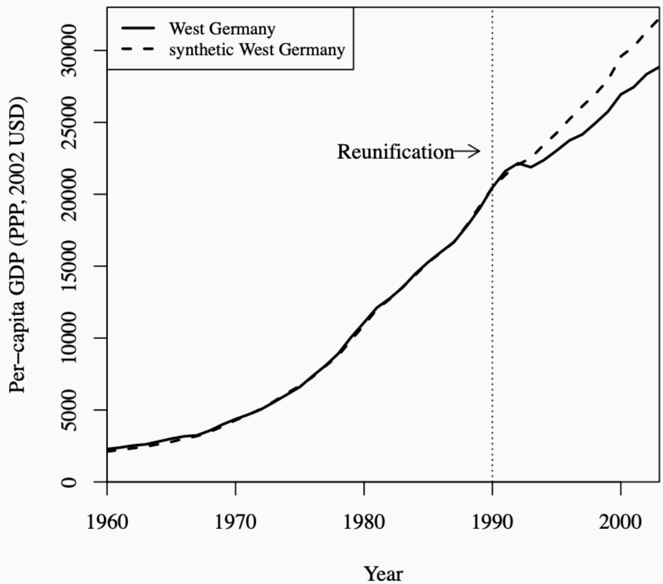
- What were the economic effects of reunification on the West German economy? Many economic historians argue that reunification had large negative economic costs, but identification is difficult because there is no obvious country with which we can compare the growth trajectory of West Germany.
- ADH15 estimate the effects of reunification by comparing the actual time series for West Germany with an SC.
- Outcome: GDP per capita (inflation adjusted).
- Time: 1960 to 2003. Reunification took place in 1990.
- Predictive variables: pre-intervention GDP per-capita, Investment rate, Trade openness, Schooling, Inflation rate, Industry share.
- Sparse SC weights: Austria 0.42; USA 0.22; Japan 0.16; Switzerland 0.11; Netherlands 0.09; Rest 0.

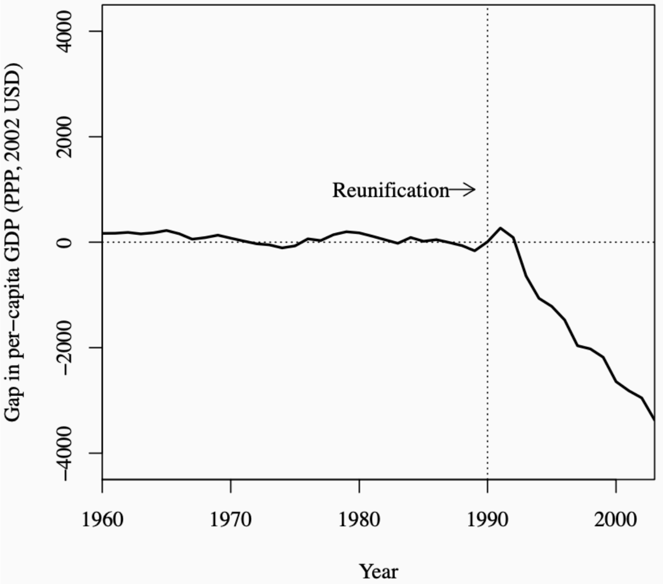

West Germany
rest of OECD sample
reunification
GDP per capita
0
10000
20000
30000
1960
1970
1980
1990
2000
Year

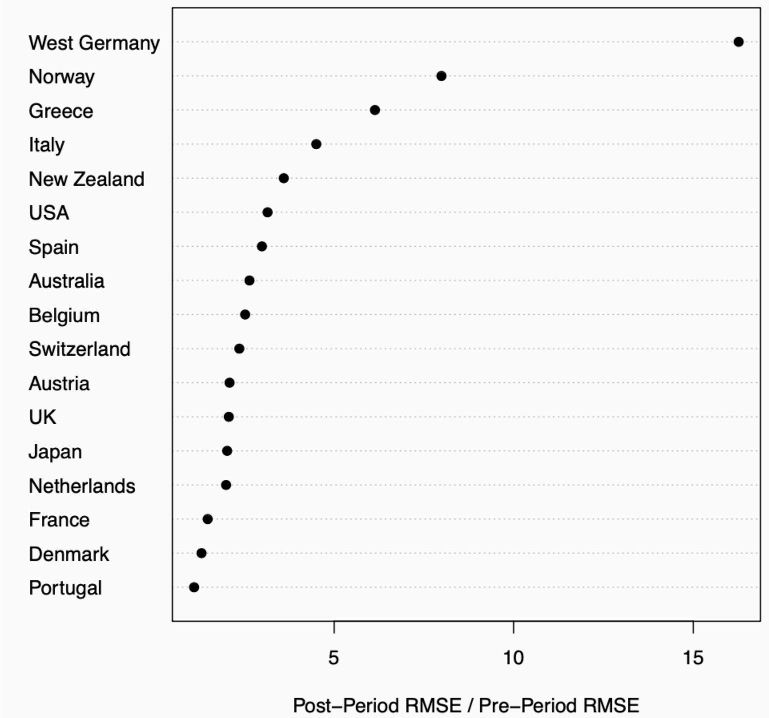

West Germany
Mean of OECD sample
reunification→
GDP per capita
0
10000
20000
30000
1960
1970
1980
1990
2000
Year

GDP predictor means:

	Treated	Synthetic	Rest of OECD Sample
GDP per-capita	15808.9	15802.2	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

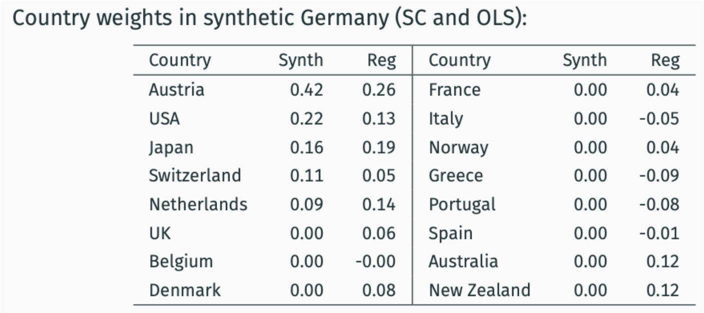

West Germany
synthetic West Germany
Reunification
Per−capita GDP (PPP, 2002 USD)
0
5000
10000
15000
20000
25000
30000
1960
1970
1980
1990
2000
Year


Gap in per−capita GDP (PPP, 2002 USD)
4000
2000
0
−2000
−4000
Reunification
1960
1970
1980
1990
2000
Year


West Germany
Norway
Greece
Italy
New Zealand
USA
Spain
Australia
Belgium
Switzerland
Austria
UK
Japan
Netherlands
France
Denmark
Portugal
5
10
15
Post−Period RMSE / Pre−Period RMSE

SC vs Regression

- $\mathbf{Y}_0$: $(T - T_0) \times J$ matrix of post-intervention outcomes of donor units; $\mathbf{X}_0$**:** $k \times J$ matrix of predictive variables of donor units; $\boldsymbol{X}_1$: $k \times 1$ vector predictive variables of the treated unit; $\overline{\mathbf{X}}_0$ and $\overline{\boldsymbol{X}}_1$: augmented $\mathbf{X}_0$ and $\boldsymbol{X}_1$ with a row of ones.
- A regression estimator of the counterfactual $\left\{Y_{1,t}^{(0)} : t = T_0 + 1, \ldots, T\right\}$ is $\widehat{\boldsymbol{B}}^\top \overline{\boldsymbol{X}}_1$, where $\widehat{\boldsymbol{B}} = \left(\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0^\top\right)^{-1} \overline{\mathbf{X}}_0 \mathbf{Y}_0^\top$.
- The regression estimator uses a linear combination: $\widehat{\boldsymbol{B}}^\top \overline{\boldsymbol{X}}_1 = \mathbf{Y}_0 \boldsymbol{W}^{\texttt{reg}}$, where $\boldsymbol{W}^{\texttt{reg}} = \overline{\mathbf{X}}_0^\top \left(\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0^\top\right)^{-1} \overline{\boldsymbol{X}}_1$.
- Regression weights $\boldsymbol{W}^{\texttt{reg}}$ sum to one but may be outside $[0, 1]$. The estimated counterfactual can be outside of the support of the data. SC weights are non-negative and sum to one.
- Regression guarantees perfect fit: $\overline{\mathbf{X}}_0 \boldsymbol{W}^{\texttt{reg}} = \overline{\boldsymbol{X}}_1$, even if the donor units are completely dissimilar in their characteristics to the treated unit.
- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics.



Country weights in synthetic Germany (SC and OLS):

Country	Synth	Reg	Country	Synth	Reg
Austria	0.42	0.26	France	0.00	0.04
USA	0.22	0.13	Italy	0.00	-0.05
Japan	0.16	0.19	Norway	0.00	0.04
Switzerland	0.11	0.05	Greece	0.00	-0.09
Netherlands	0.09	0.14	Portugal	0.00	-0.08
UK	0.00	0.06	Spain	0.00	-0.01
Belgium	0.00	-0.00	Australia	0.00	0.12
Denmark	0.00	0.08	New Zealand	0.00	0.12

- Note that SC weights are sparse and regression weights are not.
- Sparsity plays an important role for the interpretability and evaluation of reliability of the results.

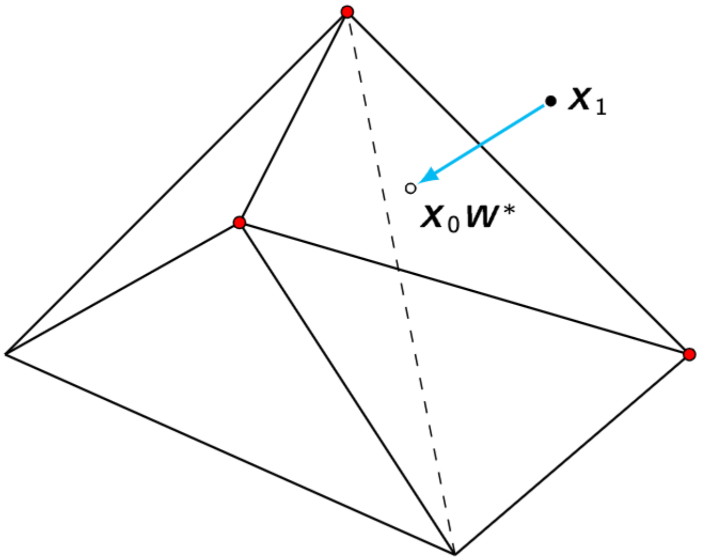


Figure 2: Projecting $\boldsymbol{X}_1$ on the convex hull of $\boldsymbol{X}_0$

- SC makes transparent the actual discrepancy between the treated unit and the convex combination of donor units that provide the counterfactual of interest. In the case when it is not possible to approximate the characteristics of the treated unit using a weighted average of the donor units, ADH advise against using SC.
- SC provides a data-driven procedure to select a comparison unit. SC makes explicit the contribution of each donor unit to the counterfactual of interest. Transparency of the counterfactual allows the use of the expert knowledge or information that is not incorporated in the research design to evaluate the validity of SC.
- For instance, Austria, Netherlands and Switzerland have large weights. If economic growth in these countries was negatively affected by the German reunification during the 1990-2003 period (perhaps because West Germany diverted demand and investment from these countries to East Germany), this would imply that SC estimates a lower bound on the magnitude (absolute value) of the negative effect of the German reunification on per capita GDP in West Germany.

Augmented SCM

- ADH's original proposal is to use SCM only when the fit on pre-intervention outcomes is good.
- Ben-Michael, Feller and Rothstein (2020)'s augmented SCM (ASCM) is an extension of SCM to setting when such pre-intervention fit is infeasible.
- The ASCM estimator of $Y_{1,t}^{(0)}$ $(t = T_0 + 1, ..., T)$ is:

$$\widehat{Y}_{1,t}^{(0)\texttt{ASC}} = \sum_{i=2}^{J+1} w_i^{\mathsf{SC}} Y_{i.t} + \left(\widehat{m}_{1,t} - \sum_{i=2}^{J+1} w_i^{\mathsf{SC}} \widehat{m}_{i,t} \right),$$

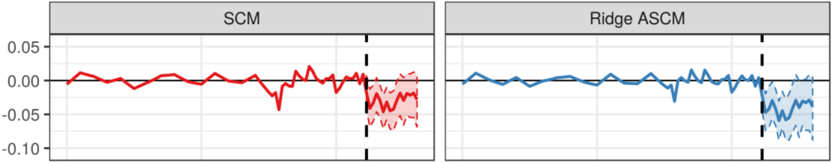
where w_i^{SC} are the SC weights, $\widehat{m}_{i,t} = \widehat{\eta}_0^{\texttt{ridge}} + \boldsymbol{X}_i^\top \widehat{\boldsymbol{\eta}}_1^{\texttt{ridge}}$,

$$\left(\widehat{\eta}_0^{\texttt{ridge}}, \widehat{\boldsymbol{\eta}}_1^{\texttt{ridge}} \right) = \underset{(\eta_0, \boldsymbol{\eta}_1)}{\operatorname{argmin}} \sum_{i=2}^{J+1} \left(Y_{i,t} - \eta_0 - \boldsymbol{X}_i^\top \boldsymbol{\eta}_1 \right)^2 + \lambda^{\texttt{ridge}} \left\| \boldsymbol{\eta}_1 \right\|^2 .$$

- The ASCM weights: $\widehat{Y}_{1,t}^{(0)\texttt{ASC}} = \sum_{i=2}^{J+1} w_i^{\texttt{ASC}} Y_{i.t}$, where

$$w_i^{\texttt{ASC}} = w_i^{\texttt{SC}} + \left(\boldsymbol{X}_1 - \mathbf{X}_0 \boldsymbol{w}^{\texttt{SC}}\right)^\top \left(\mathbf{X}_0\mathbf{X}_0^\top + \lambda^{\texttt{ridge}}\mathbf{I}_k\right)^{-1} \boldsymbol{X}_i.$$

  $w_i^{\texttt{ASC}}$ can be outside of $[0, 1]$.
- The ASCM achieves strictly better pre-treatment fit than SCM: $\left\|\boldsymbol{X}_1 - \mathbf{X}_0\boldsymbol{w}^{\texttt{ASC}}\right\| < \left\|\boldsymbol{X}_1 - \mathbf{X}_0\boldsymbol{w}^{\texttt{SC}}\right\|$.
- The ridge parameter $\lambda^{\texttt{ridge}}$ is selected by using cross validation.


SCM
Ridge ASCM