10(b). Synthetic Control and Synthetic DiD

ISS5096 || ECI

Jaewon (“Jay-one”) Yoo

National Tsing Hua University

Outline

1. Synthetic Control Methods

2. Synthetic Diff-in-Differences

3. Empirical Demonstration in R

1/ Synthetic Control Methods

Synthetic Controls

- Abadie and Gardeazabal (2003) use a DID approach for "quantitative case studies."

- Application: effect of an intervention in a single country/state at one point in time.

- Basic idea: 1 treated group, many controls.
  - Compare the time-series outcomes in the treated group to the control.
  - But which control group should you use?
  - Many possible choices and they may not be comparable to the treated.

- **Synthetic control**: use a convex combination of the controls to create a synthetic control.
  - Choose the weights that minimize the pretreatment differences between treated and synthetic control.

Intervention Study

	Time period						
	1	2	$\cdots$	T_0	T_0+1	$\cdots$	T
Treated unit ($i=1$)	0	0	0	0	1	1	1
Control group ($i=2,\ldots,J+1$)	0	0	0	0	0	0	0

- Treatment:
  - All units untreated for T_0 periods.
  - Unit 1 starts treatment at T_0, continues until T.
- Potential outcomes:
  - $Y_{it}(1)$: potential outcome at time t if i had been in the treated group.
  - $Y_{it}(0)$: potential outcome at time t if i had been in the control group.
  - No pre-intervention impacts: $Y_{it}(1) = Y_{it}(0)$ for all $t \leq T_0$.
- $\mathbf{X}_i$ is an $r \times 1$ vector of (pretreatment) covariates.
- Treatment effects: $\tau_{it} = Y_{it}(1) - Y_{it}(0)$
- Goal: estimate $\left(\tau_{1,T_0+1}, \ldots, \tau_{1,T}\right)$.

Missing Counterfactuals

- By consistency, for $t > T_0$:

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

- Need to impute missing potential outcomes, $Y_{1t}(0)$.

- **Synthetic control**: Choose weights $(w_2, \ldots, w_{J+1})'$ such that:
  - $w_j \geq 0$ and $\sum_j w_j = 1$.
  - For all $t \leq T_0$ minimize

$$\left| Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right|, \qquad \left| \mathbf{Z}_1 - \sum_{j=2}^{J+1} w_j \mathbf{Z}_j \right|$$

  - Can also add a penalty for how dispersed the weights are.

- We hope this implies for $t > T_0$: $\sum_{j=2}^{J+1} w_j Y_{jt} \approx Y_{1t}(0)$

Without Synthetic Controls

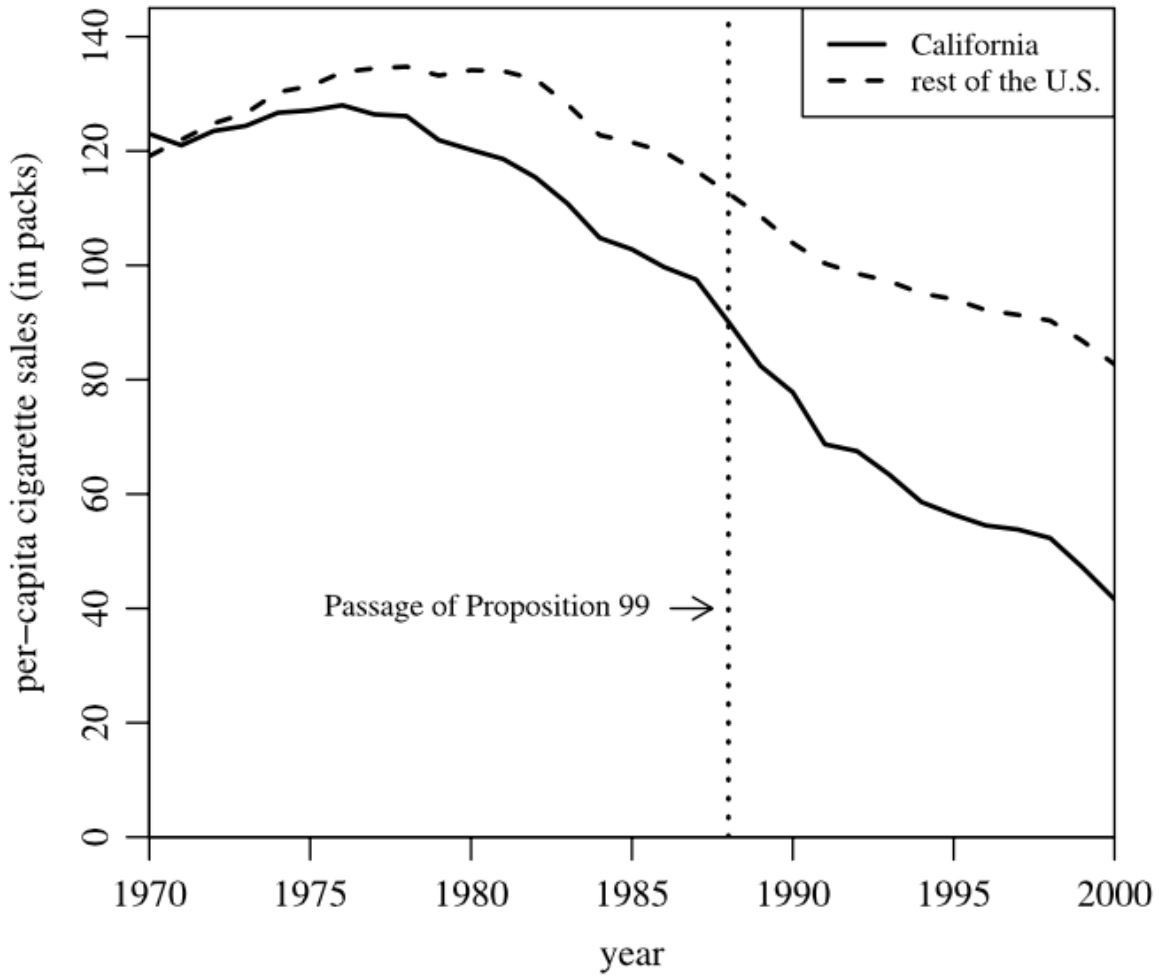


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

With Synthetic Controls

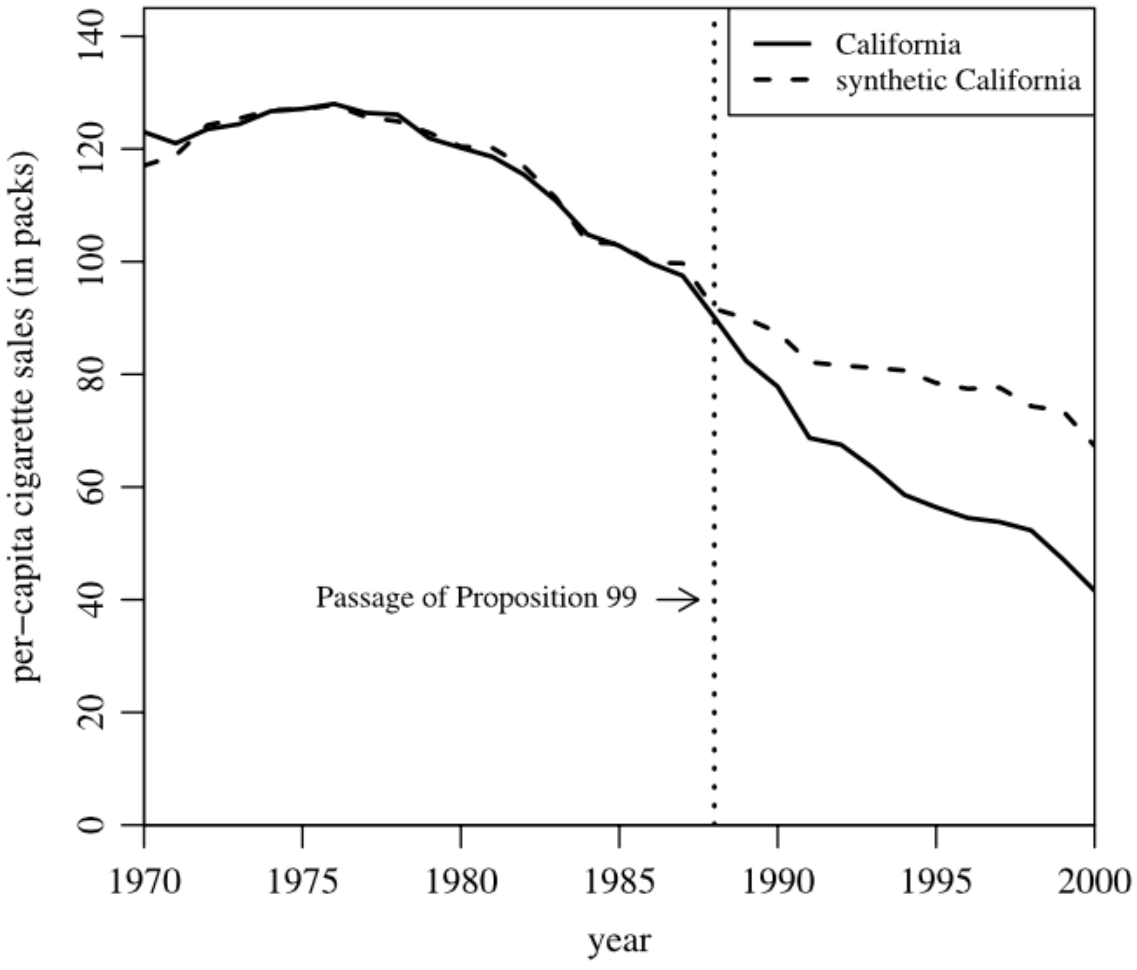


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

Weights

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

Inference

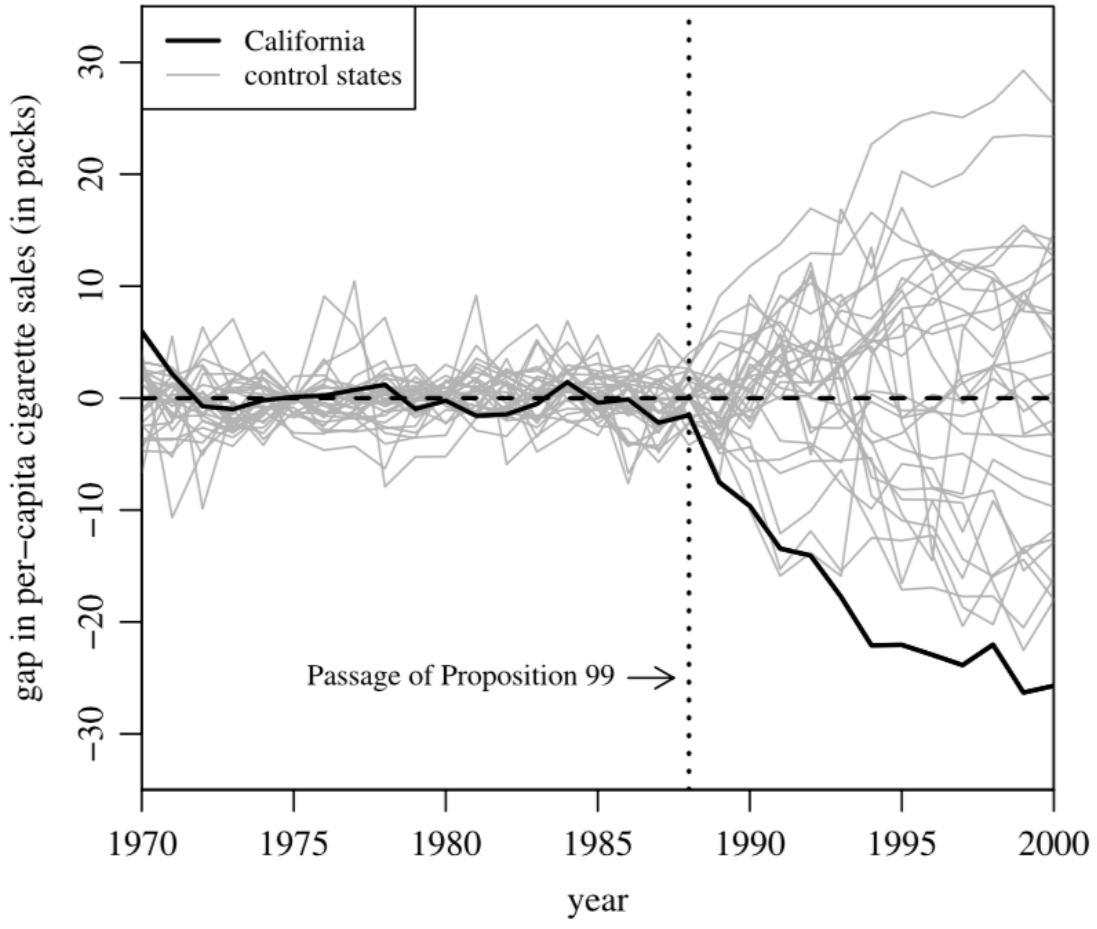


Figure 6. Per-capita cigarette sales gaps in California and placebo gaps in 29 control states (discards states with pre-Proposition 99 MSPE five times higher than California's).

Synthetic Control Justification

- ADH provide two **model-based** justifications for SC.
- **Model 1**: Interacted factor model

$$Y_{it}(0) = \mathbf{X}_i'\boldsymbol{\beta}_t + \alpha_i + \delta_t + \boldsymbol{\lambda}_t\boldsymbol{\mu}_i + \varepsilon_{it}$$

  - $\boldsymbol{\beta}_t$ are time-varying coefficients on covariates.
  - $\boldsymbol{\lambda}_t$ is a $1 \times F$ vector of common factors.
  - $\boldsymbol{\mu}_i$ is a $F \times 1$ vector of factor loadings.
  - $\boldsymbol{\lambda}_t\boldsymbol{\mu}_i$ allows time-varying confounding in a structured way.
  - Common time shocks affect each unit in a time-constant way.

- **Model 2**: autoregressive model without fixed effects

$$Y_{i,t+1}(0) = \alpha_t Y_{it}(0) + \boldsymbol{\beta}_{t+1}\mathbf{X}_{i,t+1} + u_{i,t+1}$$
$$\mathbf{X}_{i,t+1} = \gamma_t Y_{it}(0) + \boldsymbol{\Pi}_t\mathbf{X}_{it} + \mathbf{v}_{i,t+1}$$

- Either fixed effects OR lagged dependent variables, not both.

SCM Properties

- Suppose perfect balancing weights exist $(w_2^*, \ldots, w_{J+1}^*)$ such that:

$$\sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t} \qquad \sum_{j=2}^{J+1} w_j^* \mathbf{X}_j = \mathbf{X}_i$$

- Let $\widehat{Y}_{1t}(0) = \sum_{j=2}^{J+1} w_j^* Y_{jt}$ for post-intervention periods.

- Under Model 1, $\widehat{Y}_{1t}(0) \longrightarrow Y_{1t}(0)$ as $T_0 \longrightarrow \infty$.
  - As length of pre-intervention period grows, estimates get better.

- Under Model 2, $\mathbb{E}\left[\widehat{Y}_{1t}(0)\right] = \mathbb{E}[Y_{1t}(0)]$.
  - Unbiased only based on one pre-treatment periods.
  - But it assumes away unmeasured confounding!

- Outside of those models: ?????

Bias Correction

- When pre-treatment fit is imperfect $\rightsquigarrow$ significant bias in SCM.
- **Augmented SCM**: use regression models to correct for bias.
  - Let $\widehat{m}_{it} = \widehat{m}_{it}(\overline{Y}_{i,t-1})$ be predicted values for a regression of post-treatment outcomes on pre-treatment outcomes.
  - Augment estimator (Ben-Michael et al., 2021, JASA):

$$\widehat{Y}_{1t}^{\text{aug}}(0) = \sum_{j=2}^{J+1} w_j Y_{jt} + \left(\widehat{m}_{1t} - \sum_{j=2}^{J+1} w_j \widehat{m}_{jt} \right)$$

  - Can add covariates fairly easily.
- Very similar to bias correction in matching.

Generalizing to More Treated Units

- Two estimation methods to generalize to any number of treated units.
- **Interactive fixed effects**: $Y_{it}(0) = \mathbf{X}'_{it}\beta + \alpha_i + \delta_t + \boldsymbol{\lambda}_t\boldsymbol{\mu}_i$
  - Instead of weights, directly estimate IFE using iterative procedure:
    1. Treat IFE terms as fixed and fit parametric part on untreated units to get new $\widehat{\beta}$.
    2. Treat covariate coefficients as fixed and use factor analysis to estimate IFE terms.
    3. Repeat until convergence.
- **Matrix completion** methods (Athey et al., 2021)
  - Treat matrix of control POs, $\mathbf{Y}(0)$, as missing data problem.
  - Estimate lower-rank matrix $\mathbf{L}$ as best approximation to observed parts of $\mathbf{Y}(0)$ subject to regularization.

2/ Synthetic Diff-in-Differences

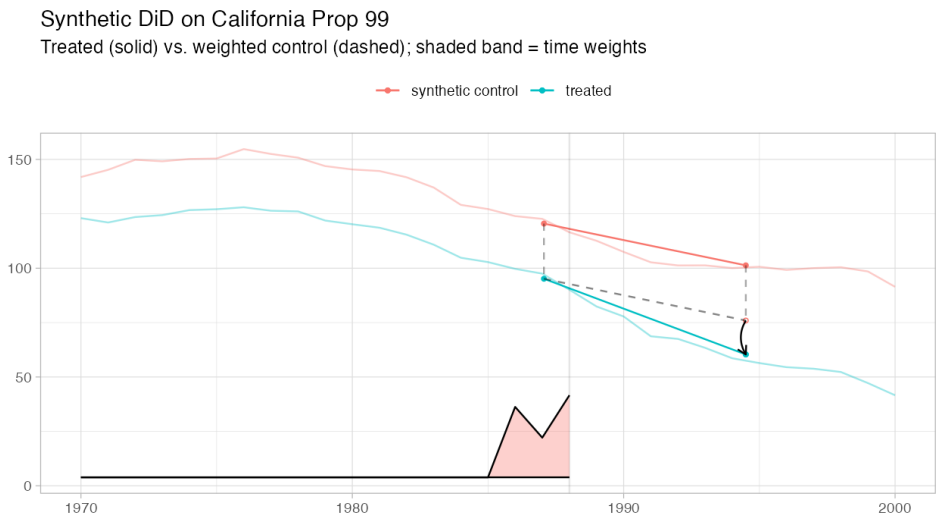
Synthetic Diff-in-Differences (Arkhangelsky et al., 2021)

- Hybrid: **TWFE regression weighted by both unit weights** (SC-style) **and time weights** (new).
- Pragmatic robustness: **attenuates sensitivity to mild parallel-trends violations**.
  - Roth (2022); Roth et al. (2023): pre-trend tests have low power; conditioning on passing distorts inference $\rightsquigarrow$ SDiD avoids strict reliance on PT in the counterfactual.
- Estimator solves:

$$\widehat{\tau}^{\text{sdid}} = \arg\min_{\tau,\mu,\alpha,\beta} \sum_{i,t} \widehat{\omega}_i^{\text{sdid}}\, \widehat{\lambda}_t^{\text{sdid}} \left(Y_{it} - \mu - \alpha_i - \beta_t - \tau D_{it}\right)^2$$

  - $\widehat{\omega}_i$ aligns pre-trends across units; $\widehat{\lambda}_t$ balances pre- vs. post-treatment periods.
- Recent applied uses: Berman & Israeli (2022); Lambrecht et al. (2023); ongoing research on platform self-regulation.

SDiD vs. DiD vs. SC on California Prop 99



Three estimators on the *same* Cal Prop 99 panel via `synthdid` R package:

- DiD (TWFE): $\widehat{\tau} = -27.3$
- Synthetic Control: $\widehat{\tau} = -19.6$
- **Synthetic DiD**: $\widehat{\tau} = -15.6$ (placebo SE $= 8.26$, 95% CI $[-31.8, +0.6]$)

With 1 treated unit, "TWFE" = California vs. *unweighted* donor average $\rightsquigarrow$ counterfactual mismatches CA (level + trend) $\rightsquigarrow$ inflated $|\widehat{\tau}|$. SC's donor weights match CA's pre-period $\rightsquigarrow$ $|\widehat{\tau}|$ shrinks. SDiD adds time weights $\rightsquigarrow$ shrinks further. Each refinement yields a more conservative ATT.

Solid turquoise: treated; solid red: synthetic control; shaded band: time weights.

3/ Empirical Demonstration in R

Empirical Demonstration: Roadmap

- One demonstration using **California Proposition 99** (Abadie, Diamond & Hainmueller, 2010) via the `tidysynth` R package.
- Steps shown with full console input + output:
  1. Build the SC pipeline (predictors + lagged outcomes + weights).
  2. Inspect donor weights and predictor balance.
  3. Visualize California vs. Synthetic California (trends + gap).
  4. Placebo inference via RMSPE ratio across all donor units.
- Companion comparison with SDiD on the same data is in Section 2 above (`synthdid` R package).

Stage 1: Synthetic California Pipeline

```
> library(tidysynth); data(smoking)  # 39 states, 1970-2000, per-capita cigarette sales
> sc <- smoking |>
+   synthetic_control(outcome=cigsale, unit=state, time=year,
+                     i_unit="California", i_time=1988, generate_placebos=TRUE) |>
+   generate_predictor(time_window=1980:1988,
+                      ln_income = mean(lnincome, na.rm=TRUE),
+                      ret_price = mean(retprice, na.rm=TRUE),
+                      youth     = mean(age15to24, na.rm=TRUE)) |>
+   generate_predictor(time_window=1984:1988, beer = mean(beer, na.rm=TRUE)) |>
+   generate_predictor(time_window=1975, cigsale_1975=cigsale) |>
+   generate_predictor(time_window=1980, cigsale_1980=cigsale) |>
+   generate_predictor(time_window=1988, cigsale_1988=cigsale) |>
+   generate_weights(optimization_window=1970:1988) |> generate_control()
```

- Predictors: pre-period averages of income, retail price, youth share, beer consumption + three lagged outcomes (1975, 1980, 1988).
- `generate_placebos=TRUE` runs the same SC on every donor unit (for inference).

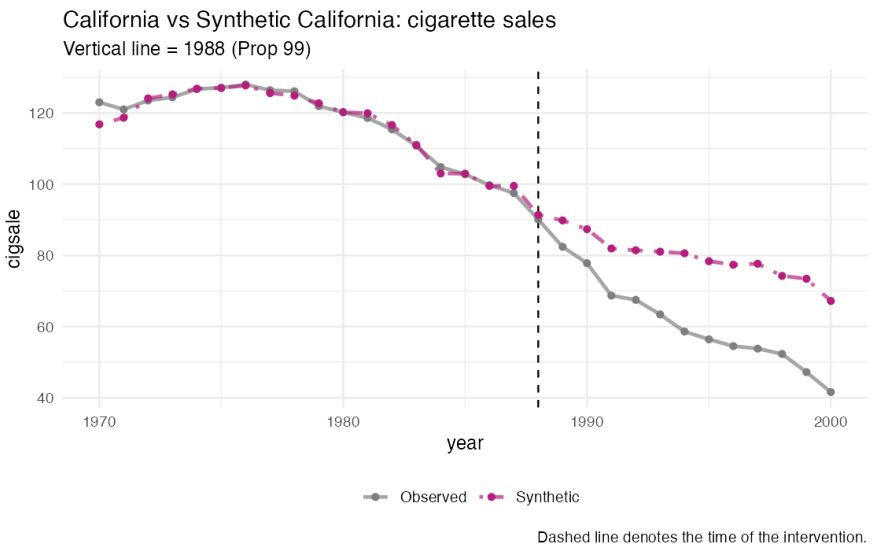
Stage 2: Donor Weights and Predictor Balance

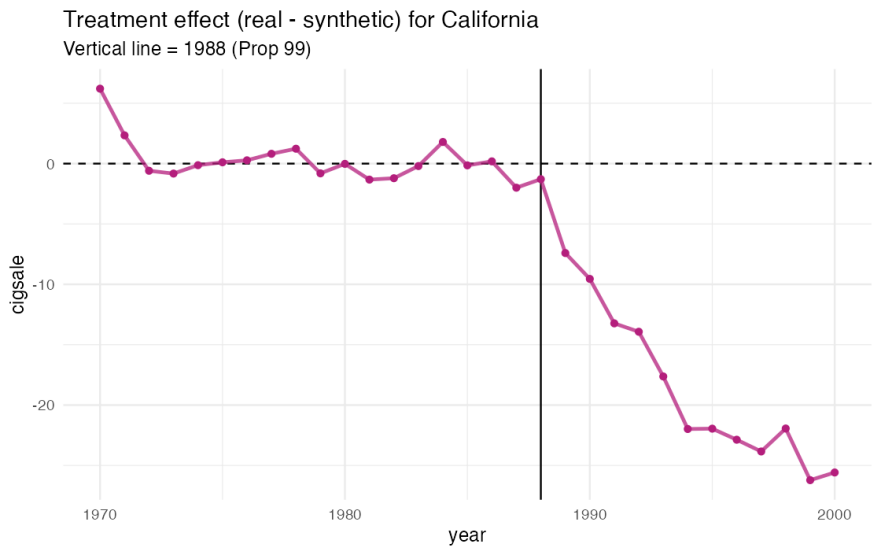
```
> sc |> grab_unit_weights() |> filter(weight > 0.01) |> arrange(desc(weight))
# A tibble: 5 x 2
  unit        weight
  <chr>        <dbl>
1 Utah        0.342
2 Nevada      0.238
3 Montana     0.209
4 Colorado    0.149
5 Connecticut 0.0617

> sc |> grab_balance_table()
# A tibble: 7 x 4
  variable     California synthetic_California donor_sample
1 ln_income        10.1                  9.85         9.83
2 ret_price        89.4                 89.4          87.3
3 youth             0.174                0.174        0.173
4 beer             24.3                 24.2          23.7
5 cigsale_1975    127.                 127.          137.
6 cigsale_1980    120.                 120.          138.
7 cigsale_1988     90.1                 91.4         114.
```

- Synthetic California is mostly Utah + Nevada + Montana + Colorado + Connecticut.
- Pre-treatment balance is near-perfect on lagged outcomes; donor sample averages are far from California (e.g. `cigsale_1988`: $90 \neq 114$).

Stage 3: California vs Synthetic California

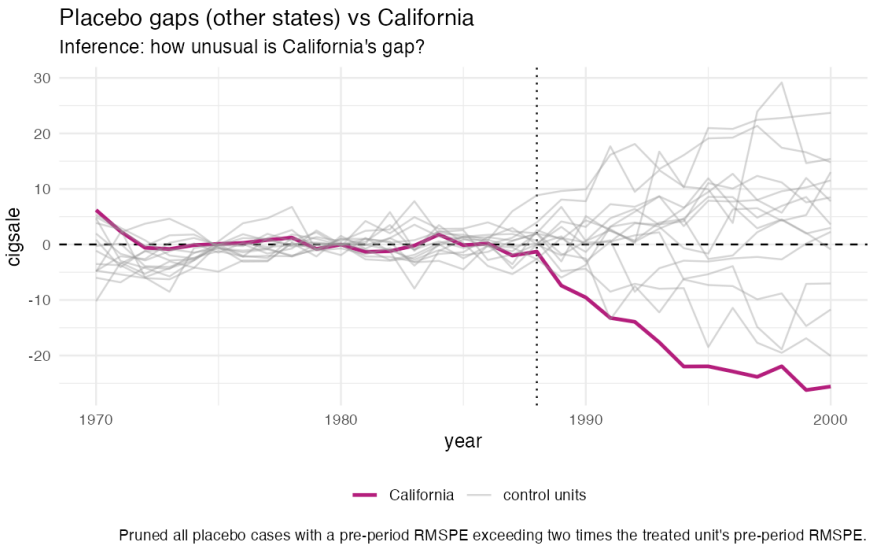


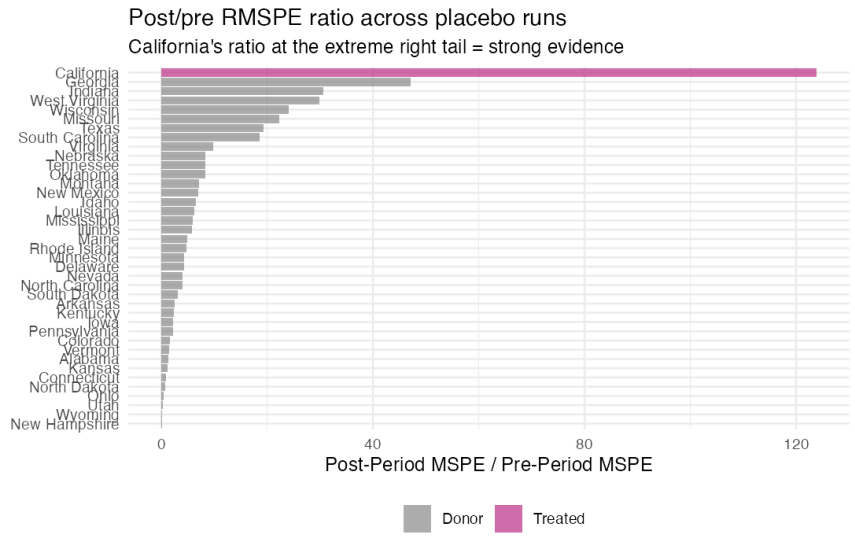


Left: California (solid) tracks synthetic California (dashed) almost perfectly pre-1988, then diverges sharply post-Prop 99. **Right**: estimated gap $\widehat{Y}_{1t} - \sum_j w_j Y_{jt}$. Average post-1988 gap $= -18.85$ packs/capita.

Stage 4: Placebo Inference (RMSPE Ratio)

```
> sc |> grab_significance() |> head(5)
# A tibble: 5 x 8
  unit_name      type    pre_mspe post_mspe mspe_ratio  rank fishers_exact_pvalue z_score
  <chr>          <chr>      <dbl>     <dbl>      <dbl> <int>                <dbl>   <dbl>
1 California     Treated     3.17      392.      124.       1               0.0256    5.97
2 Georgia        Donor       3.79      179.       47.2      2               0.0513    2.07
3 Indiana        Donor      25.2       770.       30.6      3               0.0769    1.13
4 West Virginia  Donor       9.52      284.       29.8      4               0.103     1.08
5 Wisconsin      Donor      11.1       268.       24.1      5               0.128     0.984
```





California's post/pre MSPE ratio (124.0) is the **largest of 39 placebos** ⇝ Fisher exact $p = 0.026$, $z = 5.97$. Inference here is design-based: no parametric SE, just permutation across donor units.

Back to the Main Deck!

Contact Information:
jaewon.yoo@iss.nthu.edu.tw
https://j1yoo.github.io/