

SYNTHETIC CONTROL

Juan Estrada, Ph.D

Emory University

October 27, 2023

MOTIVATION

CALIFORNIA'S PROPOSITION 99

- ▶ Tobacco control legislation in 1988
- ▶ **Intervention:** (1) increased cigarette taxes, (2) spurred clean air ordinances, (3) founded anti-smoked media campaigns, (4) tax revenues went to anti-tobacco projects

MOTIVATION

CALIFORNIA'S PROPOSITION 99

- ▶ Tobacco control legislation in 1988
- ▶ **Intervention:** (1) increased cigarette taxes, (2) spurred clean air ordinances, (3) founded anti-smoked media campaigns, (4) tax revenues went to anti-tobacco projects
- ▶ **Question:** does Proposition 99 have a causal effect on the per capita cigarette sales in California?

MOTIVATION

CALIFORNIA'S PROPOSITION 99

- ▶ Tobacco control legislation in 1988
- ▶ **Intervention:** (1) increased cigarette taxes, (2) spurred clean air ordinances, (3) founded anti-smoked media campaigns, (4) tax revenues went to anti-tobacco projects
- ▶ **Question:** does Proposition 99 have a causal effect on the per capita cigarette sales in California?
- ▶ **Data structure:** Aggregate characteristics (prices, income, demographics, etc.) and outcomes for California (**treated group**) and another 38 states (**control group**)



- ▶ How can we estimate the causal effect of the intervention (policy)?

MOTIVATION

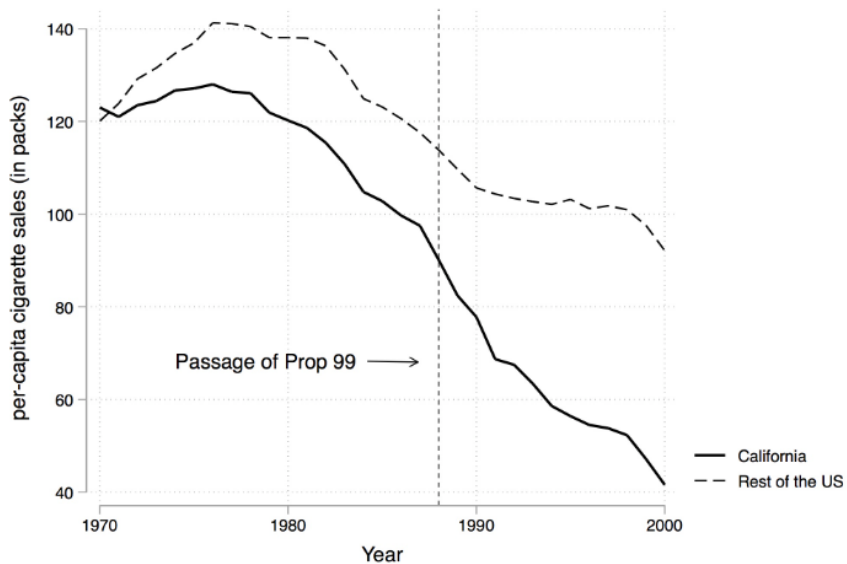
DIFFERENCE IN DIFFERENCE REQUIREMENTS

- ▶ A group of units are treated at the same time and a group of untreated units
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units.
- ▶ Sometimes it requires large sample for inference.

MOTIVATION

OUTCOME TRENDS

Figure. Evolution of Outcomes for the Treatment and Controls



MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units.

MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units. **Valid**
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units.

MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units. **Valid**
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units. **Unknown**

MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units. **Valid**
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units. **Unknown**

$$\begin{aligned}\hat{\delta}_{TC}^{2 \times 2} &= \underbrace{E[Y_T^1 | \text{Post}] - E[Y_T^0 | \text{Post}]}_{\text{ATT}} \\ &+ \underbrace{[E[Y_T^0 | \text{Post}] - E[Y_T^0 | \text{Pre}]] - [E[Y_C^0 | \text{Post}] - E[Y_C^0 | \text{Pre}]]}_{\text{Non-parallel trends bias in } 2 \times 2 \text{ case}}\end{aligned}$$

- ▶ We can argue the potential validity by looking at the pre-treatment trends.

MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units. **Valid**
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units. **Unknown**

$$\begin{aligned}\hat{\delta}_{TC}^{2 \times 2} &= \underbrace{E[Y_T^1 | \text{Post}] - E[Y_T^0 | \text{Post}]}_{\text{ATT}} \\ &+ \underbrace{[E[Y_T^0 | \text{Post}] - E[Y_T^0 | \text{Pre}]] - [E[Y_C^0 | \text{Post}] - E[Y_C^0 | \text{Pre}]}}_{\text{Non-parallel trends bias in } 2 \times 2 \text{ case}}\end{aligned}$$

- ▶ We can argue the potential validity by looking at the pre-treatment trends. (arguably) **Invalid**
- ▶ Sometimes it requires large sample for inference.

MOTIVATION

ARE THE DID ASSUMPTIONS VALID?

- ▶ **Simple framework (2×2 case):** a group of units are treated at the same time and there is a group of untreated units. **Valid**
- ▶ **Parallel trends:** the counterfactual evolution (trend) of the outcome for the *treated* units is parallel to the **observed** evolution of the *control* units. **Unknown**

$$\begin{aligned}\hat{\delta}_{TC}^{2 \times 2} &= \underbrace{E[Y_T^1 | \text{Post}] - E[Y_T^0 | \text{Post}]}_{\text{ATT}} \\ &+ \underbrace{[E[Y_T^0 | \text{Post}] - E[Y_T^0 | \text{Pre}]] - [E[Y_C^0 | \text{Post}] - E[Y_C^0 | \text{Pre}]]}_{\text{Non-parallel trends bias in } 2 \times 2 \text{ case}}\end{aligned}$$

- ▶ We can argue the potential validity by looking at the pre-treatment trends. (arguably) **Invalid**
- ▶ Sometimes it requires large sample for inference. **Invalid**

A PRIMER ON SYNTHETIC CONTROL

IMPORTANCE AND MAIN IDEA

- ▶ Originally proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the goal to estimate the effect of **aggregate interventions**.

A PRIMER ON SYNTHETIC CONTROL

IMPORTANCE AND MAIN IDEA

- ▶ Originally proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the goal to estimate the effect of **aggregate interventions**.
- ▶ Athey and Imbens (2017) said it was “arguably the most important innovation in the **policy evaluation** literature in the last 15 years.”
- ▶ Unifies **comparative case studies** focusing on the analysis of a single unit → qualitative analysis focuses on inductive reasoning.

A PRIMER ON SYNTHETIC CONTROL

IMPORTANCE AND MAIN IDEA

- ▶ Originally proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the goal to estimate the effect of **aggregate interventions**.
- ▶ Athey and Imbens (2017) said it was “arguably the most important innovation in the **policy evaluation** literature in the last 15 years.”
- ▶ Unifies **comparative case studies** focusing on the analysis of a single unit → qualitative analysis focuses on inductive reasoning.
- ▶ **Insight:** a combination of comparison units is often better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- ▶ You can think of Synthetic Control as generalizing DID (Abadie et al., 2010). [▶ Argument](#)

A PRIMER ON SYNTHETIC CONTROL

WHAT PROBLEMS DOES IT SOLVE?

Cuba, Miami, and the Mariel Boatlift (Card, 1990)

- ▶ In 1980, 125,000 Cubans emigrated to Florida over six months after Castro announced that anyone wishing to leave Cuba could do so.

A PRIMER ON SYNTHETIC CONTROL

WHAT PROBLEMS DOES IT SOLVE?

Cuba, Miami, and the Mariel Boatlift (Card, 1990)

- ▶ In 1980, 125,000 Cubans emigrated to Florida over six months after Castro announced that anyone wishing to leave Cuba could do so.
- ▶ Card, 1990 used this as a **natural experiment** to evaluate whether inflows of immigrants depress wages and the employment of natives in local labor-markets.

A PRIMER ON SYNTHETIC CONTROL

WHAT PROBLEMS DOES IT SOLVE?

Cuba, Miami, and the Mariel Boatlift (Card, 1990)

- ▶ In 1980, 125,000 Cubans emigrated to Florida over six months after Castro announced that anyone wishing to leave Cuba could do so.
- ▶ Card, 1990 used this as a **natural experiment** to evaluate whether inflows of immigrants depress wages and the employment of natives in local labor-markets.
- ▶ **Approach:** use individual-level data on wages and unemployment in Miami and chose four comparison cities (Atlanta, Los Angeles, Houston, and Tampa–St. Petersburg) as controls.
- ▶ Card, 1990 uses DID and finds no effect on wages or native unemployment.

Any potential issues with Card's (1990) approach?

A PRIMER ON SYNTHETIC CONTROL

WHAT PROBLEMS DOES IT SOLVE?

Card's (1990) study was a comparative case study that used aggregated individual data to identify the effect of an aggregate intervention.

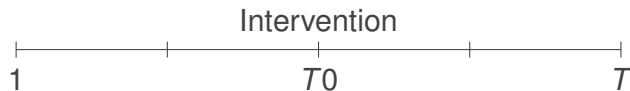
Potential Issues:

1. The selection of the control group was *ad hoc* and subjective. There was not any discussion about validity of the control group selection process.
2. Standard errors reflect sampling variance (from averaging wages and unemployment) not uncertainty about the *ability of the control group to reproduce the counterfactual of interest*.

The synthetic control estimator is a way of addressing both issues simultaneously!

FORMALIZATION

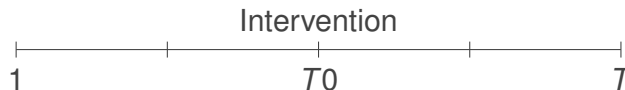
FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots T$.

FORMALIZATION

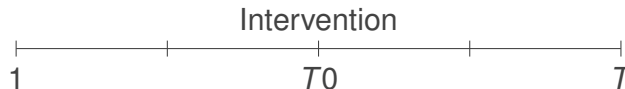
FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots, T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots, T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).

FORMALIZATION

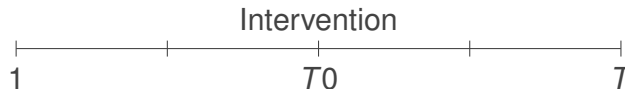
FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).
- ▶ Y_{it}^I is the outcome for unit i at time t if unit i is exposed to the intervention.

FORMALIZATION

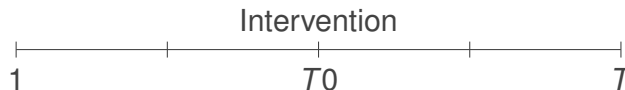
FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).
- ▶ Y_{it}^I is the outcome for unit i at time t if unit i is exposed to the intervention.
- ▶ Y_{it}^N is the outcome for unit i at time t in the absence of the intervention.

FORMALIZATION

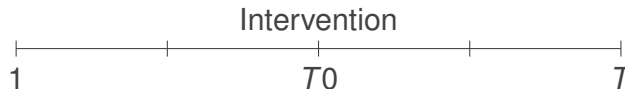
FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots, T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots, T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).
- ▶ Y_{it}^I is the outcome for unit i at time t if unit i is exposed to the intervention.
- ▶ Y_{it}^N is the outcome for unit i at time t in the absence of the intervention.
- ▶ Estimand of interest $\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$ for $t > T_0$.

FORMALIZATION

FRAMEWORK

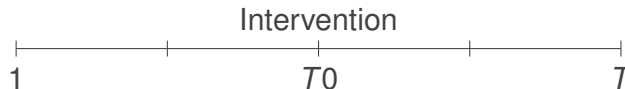


- ▶ We observe $J + 1$ units in periods $1, 2, \dots T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).
- ▶ Y_{it}^I is the outcome for unit i at time t if unit i is exposed to the intervention.
- ▶ Y_{it}^N is the outcome for unit i at time t in the absence of the intervention.
- ▶ Estimand of interest $\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$ for $t > T_0$.

What is an estimand? What does it mean for τ_{1t} to be indexed by t ? What is the difference between Y_{1t}^I and Y_{1t} ?

FORMALIZATION

FRAMEWORK



- ▶ We observe $J + 1$ units in periods $1, 2, \dots, T$.
- ▶ Unit **one** is exposed to the intervention of interest (**treated**) during periods $T_0 + 1, \dots, T$.
- ▶ The remaining J units are an untreated reservoir of potential controls (the **donor pool**).
- ▶ Y_{it}^I is the outcome for unit i at time t if unit i is exposed to the intervention.
- ▶ Y_{it}^N is the outcome for unit i at time t in the absence of the intervention.
- ▶ Estimand of interest $\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$ for $t > T_0$.

What is an estimand? What does it mean for τ_{1t} to be indexed by t ? What is the difference between Y_{1t}^I and Y_{1t} ? **Can we observe all outcomes?**

FORMALIZATION

SOLUTION FOR THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

How to estimate Y_{1t}^N ? What solves the fundamental problem of causal inference in DID?

FORMALIZATION

SOLUTION FOR THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

How to estimate Y_{1t}^N ? What solves the fundamental problem of causal inference in DID?

- ▶ Synthetic Control approximates Y_{1t}^N by a **weighted average** of units from a **donor pool**.

FORMALIZATION

SOLUTION FOR THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

How to estimate Y_{1t}^N ? What solves the fundamental problem of causal inference in DID?

- ▶ Synthetic Control approximates Y_{1t}^N by a **weighted average** of units from a **donor pool**.
- ▶ 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**.

FORMALIZATION

SOLUTION FOR THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

How to estimate Y_{1t}^N ? What solves the fundamental problem of causal inference in DID?

- ▶ Synthetic Control approximates Y_{1t}^N by a **weighted average** of units from a **donor pool**.
- ▶ 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**.
- ▶ \mathbf{X}_1 is a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit.
- ▶ \mathbf{X}_0 is a $(k \times J)$ matrix with the same variables for the unaffected units.
- ▶ \mathbf{X}_1 and \mathbf{X}_0 can include lagged outcomes!

How do we choose \mathbf{W} ?

FORMALIZATION

CRITERIA TO CHOOSE THE OPTIMAL WEIGHTS

- ▶ We choose the weights to get the best fit for the pre-treatment characteristics of the treated unit (\mathbf{X}_1) using the controls' characteristics (\mathbf{X}_0). Mathematically, we minimize:

$$\|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}\| = \sqrt{(\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W})' \mathbf{V} (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W})}$$

- ▶ The matrix \mathbf{V} is $(k \times k)$, symmetric and positive semidefinite

FORMALIZATION

CRITERIA TO CHOOSE THE OPTIMAL WEIGHTS

- ▶ We choose the weights to get the best fit for the pre-treatment characteristics of the treated unit (\mathbf{X}_1) using the controls' characteristics (\mathbf{X}_0). Mathematically, we minimize:

$$\|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}\| = \sqrt{(\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W})' \mathbf{V} (\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W})}$$

- ▶ The matrix \mathbf{V} is $(k \times k)$, symmetric and positive semidefinite (**positive semidefinite?**)
- ▶ Typically \mathbf{V} is diagonal: v_1, \dots, v_k (**interpretation?**), implying the minimization (**why?**):

$$\sum_{m=1}^k v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

- ▶ The choice of \mathbf{V} is relevant. People usually use cross-validation to minimize prediction MSE.

▶ Process

EXAMPLE

WEIGHTS FOR PROPOSITION 99

State	Weight	State	Weight
Alabama	0	Nebraska	0
Arkansas	0	Nevada	0.234
Colorado	0.164	New Hampshire	0
Connecticut	0.069	New Mexico	0
Delaware	0	North Carolina	0
Georgia	0	North Dakota	0
Idaho	0	Ohio	0
Illinois	0	Oklahoma	0
Indiana	0	Pennsylvania	0
Iowa	0	Rhode Island	0
Kansas	0	South Carolina	0
Kentucky	0	South Dakota	0
Louisiana	0	Tennessee	0
Maine	0	Texas	0
Michigan	0	Utah	0.334
Minnesota	0	Vermont	0
Mississippi	0	West Virginia	0
Missouri	0	Wisconsin	0
Montana	0.199	Wyoming	0

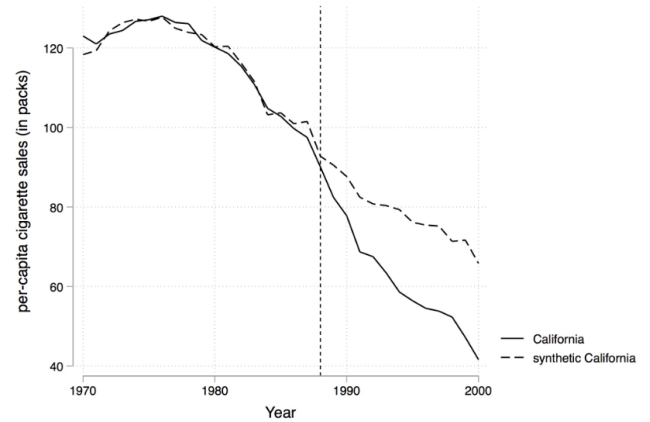
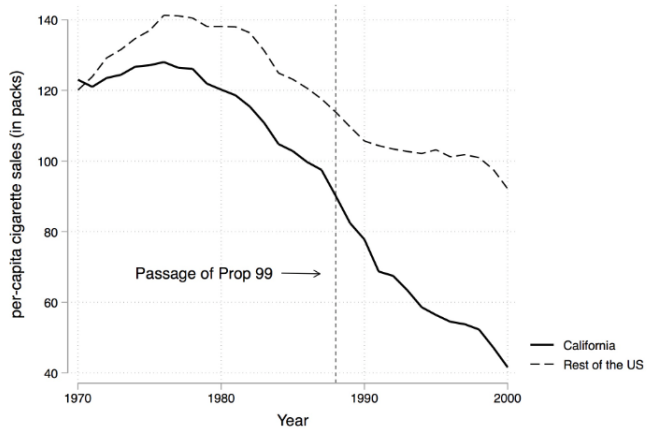
EXAMPLE

BALANCE TABLE

Variables	Real California	Synthetic	Average 38 States
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.4	17.4	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.2	23.75
Cigarette sales per capita 1988	90.1	91.62	114.2
Cigarette sales per capita 1980	120.2	120.43	136.58
Cigarette sales per capita 1975	127.1	126.99	132.81

EXAMPLE

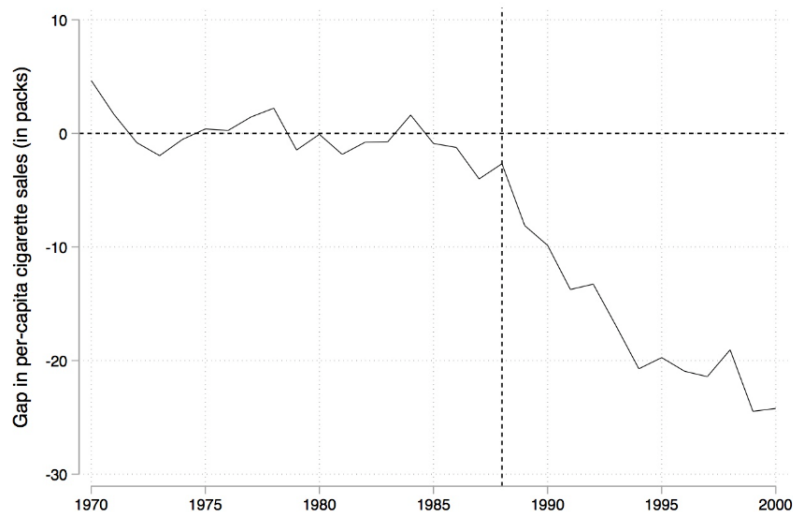
GRAPHICAL RESULTS



EXAMPLE

SYNTHETIC CONTROL ESTIMATOR

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



WHEN TO USE SYNTHETIC CONTROL

CONTEXTUAL REQUIREMENTS

1. **Volatility of the outcome:** we only have one treated unit (no average smoothing). The problematic volatility comes from the errors in the factor model (**why?**)
2. **Availability of a comparison group:** drop units that may have suffered large idiosyncratic shocks to the outcome of interest during the study period. Keep only similar units!
3. **No anticipation:** can bias the results of the estimation because units behave differently anticipating the intervention (**solution?**)
4. **No interference:** stable unit treatment value assumption (**tension with 2?**)
5. **Convex hull condition:** the differences in the characteristics of the affected unit and the synthetic control are small.

WHEN TO USE SYNTHETIC CONTROL

DATA REQUIREMENTS

1. **Aggregate data:** it is possible to use microdata to create the aggregate values (**inference?**)
2. **Pre-intervention information:** bias is bounded by a function that is inversely proportional to the number of pre-intervention periods [▶ Technically](#)
3. **Post-intervention information:** interventions may take time to emerge. The post-intervention period should be large enough to capture delays

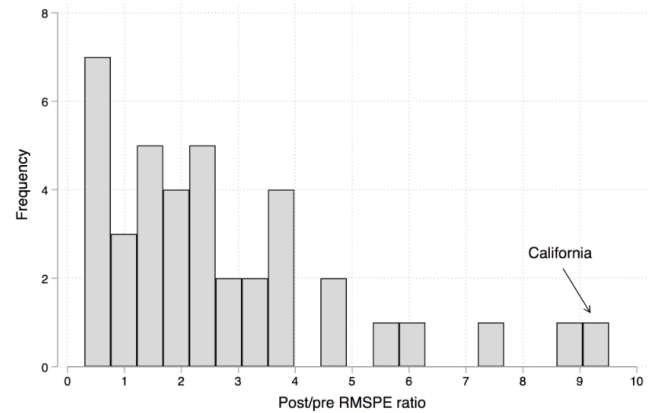
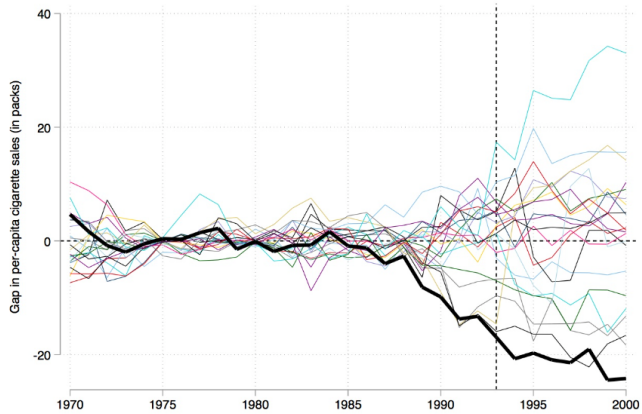
INFERENCE

PERMUTATION

- ▶ Abadie et al. (2010) propose inference 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. [▶ P-value Calculation](#)

INFERENCE

INFERENCE FOR THE PROPOSITION 99



EXTRA TOPICS

ADVANTAGES OVER LINEAR REGRESSION

Linear Regression Estimator

- ▶ Use the panel data structure
- ▶ \mathbf{Y}_0 is the $(T - T_0) \times J$ matrix of post-intervention outcomes for the donor pool
- ▶ $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_0$ are \mathbf{X}_1 and \mathbf{X}_0 plus constants
- ▶ The counterfactual estimator for Y_{1t}^N is $\hat{\mathbf{B}}^\top \bar{\mathbf{X}}_1$ for $\hat{\mathbf{B}} = \left(\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0^\top \right)^{-1} \bar{\mathbf{X}}_0 \mathbf{Y}_0^\top$
- ▶ **Linear combination of outcomes:** $\mathbf{Y}_0 \mathcal{W}^{reg}$ for $\mathbf{W}^{reg} = \bar{\mathbf{X}}_0' \left(\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0' \right)^{-1} \bar{\mathbf{X}}_1$
- ▶ **You can show:** $\iota^\top \mathcal{W}^{reg} = 1$ (regress ι on \mathbf{X}_0) and weights may be outside of $[0, 1]$

EXTRA TOPICS

ADVANTAGES OVER LINEAR REGRESSION

Linear Regression Estimator

- ▶ Use the panel data structure
- ▶ \mathbf{Y}_0 is the $(T - T_0) \times J$ matrix of post-intervention outcomes for the donor pool
- ▶ $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_0$ are \mathbf{X}_1 and \mathbf{X}_0 plus constants
- ▶ The counterfactual estimator for Y_{1t}^N is $\hat{\mathbf{B}}^\top \bar{\mathbf{X}}_1$ for $\hat{\mathbf{B}} = \left(\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0^\top \right)^{-1} \bar{\mathbf{X}}_0 \mathbf{Y}_0^\top$
- ▶ **Linear combination of outcomes:** $\mathbf{Y}_0 \mathcal{W}^{reg}$ for $\mathbf{W}^{reg} = \bar{\mathbf{X}}_0' \left(\bar{\mathbf{X}}_0 \bar{\mathbf{X}}_0' \right)^{-1} \bar{\mathbf{X}}_1$
- ▶ **You can show:** $\iota^\top \mathcal{W}^{reg} = 1$ (regress ι on \mathbf{X}_0) and weights may be outside of $[0, 1]$

Synthetic Control Advantages:

1. **No extrapolation:** weights are in the interval $[0, 1]$ (**why the name?**)
2. **Transparency of fit:** $\mathbf{X}_0 \mathbf{W}^{reg} = \mathbf{X}_1$ may hold approximately (**in regression?**)
3. **Transparency of counterfactuals:** weights are explicitly calculated

EXTRA TOPICS

EXTENTION OF THE DID METHOD

- ▶ Think we directly wanted to model the counterfactual outcome Y_{1t}^N [▶ Introduction](#)

- ▶ Choose a linear factor model given by:

$$Y_{jt}^N = \delta_t + \theta_t \mathbf{Z}_j + \lambda_t \mu_j + \varepsilon_{jt}$$

- ▶ How is this an extension of the DID model?

EXTRA TOPICS

EXTENSION OF THE DID METHOD

- ▶ Think we directly wanted to model the counterfactual outcome Y_{1t}^N [▶ Introduction](#)

- ▶ Choose a linear factor model given by:

$$Y_{jt}^N = \delta_t + \theta_t \mathbf{Z}_j + \lambda_t \mu_j + \varepsilon_{jt}$$

- ▶ How is this an extension of the DID model? The term λ_t is a set of common factors that are different across unit
- ▶ What do we need for parallel trends to hold?

EXTRA TOPICS

EXTENSION OF THE DID METHOD

- ▶ Think we directly wanted to model the counterfactual outcome Y_{1t}^N [▶ Introduction](#)

- ▶ Choose a linear factor model given by:

$$Y_{jt}^N = \delta_t + \theta_t \mathbf{Z}_j + \lambda_t \mu_j + \varepsilon_{jt}$$

- ▶ How is this an extension of the DID model? The term λ_t is a set of common factors that are different across unit
- ▶ What do we need for parallel trends to hold? conditional in $\mathbf{Z}_j \rightarrow \delta_t + \theta_t \mathbf{Z} + \lambda \mu_j$
- ▶ Synthetic Control relaxes the parallel trends assumption

EXTRA TOPICS

SYNTHETIC CONTROL BIAS

Under the factor model you can show: [Data Requirements](#)

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_{jt}^N = R_{1t} + R_{2t} + R_{3t}, \quad E(R_{1t}) = R_{2t} = 0$$

$$E |R_{1t}| \leq C(p)^{1/p} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\}$$

- ▶ $\bar{m}_p^{1/p}$ and $\bar{\sigma}$ are both moments of the errors ε_{jt} and depend on the scale of those shocks
- ▶ The bias is controlled by the ratio between the scale of the shocks and the number of pre-intervention periods!

EXTRA TOPICS

SYNTHETIC CONTROL BIAS

Under the factor model you can show: [Data Requirements](#)

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_{jt}^N = R_{1t} + R_{2t} + R_{3t}, \quad E(R_{1t}) = R_{2t} = 0$$

$$E |R_{1t}| \leq C(p)^{1/p} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\}$$

- ▶ $\bar{m}_p^{1/p}$ and $\bar{\sigma}$ are both moments of the errors ε_{jt} and depend on the scale of those shocks
- ▶ The bias is controlled by the ratio between the scale of the shocks and the number of pre-intervention periods!
- ▶ **Intuition:** matching \mathbf{Z}_1 and $\boldsymbol{\mu}_1 \rightarrow$ unbiased estimation. If $\mathbf{X}_1 = \mathbf{X}_0 \mathbf{W}^*$, SC matches \mathbf{Z}_1 . SC can't directly match $\boldsymbol{\mu}_1$ but indirectly via matching pre-intervention outcomes

EXTRA TOPICS

SYNTHETIC CONTROL BIAS

Under the factor model you can show: [▶ Data Requirements](#)

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_{jt}^N = R_{1t} + R_{2t} + R_{3t}, \quad E(R_{1t}) = R_{2t} = 0$$

$$E |R_{1t}| \leq C(p)^{1/p} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\}$$

- ▶ $\bar{m}_p^{1/p}$ and $\bar{\sigma}$ are both moments of the errors ε_{jt} and depend on the scale of those shocks
- ▶ The bias is controlled by the ratio between the scale of the shocks and the number of pre-intervention periods!
- ▶ **Intuition:** matching \mathbf{Z}_1 and $\boldsymbol{\mu}_1 \rightarrow$ unbiased estimation. If $\mathbf{X}_1 = \mathbf{X}_0 \mathbf{W}^*$, SC matches \mathbf{Z}_1 . SC can't directly match $\boldsymbol{\mu}_1$ but indirectly via matching pre-intervention outcomes
- ▶ If SC cannot match $\boldsymbol{\mu}_1$ but we observe close match in pre-intervention outcomes \rightarrow differences in errors compensate for the differences in unobserved factor loadings! (**so what?**)

EXTRA TOPICS

SELECTION OF PREDICTORS IMPORTANCE (\mathbf{V})

Choose \mathbf{V} such that it minimizes the Mean Square Prediction Error (MSPE):

$$\sum_{t \in \mathcal{T}_0} (Y_{1t} - w_2(\mathbf{V})Y_{2t} - \dots - w_{J+1}(\mathbf{V})Y_{J+1t})^2$$

For some set \mathcal{T}_0 of **pre-intervention periods**. Steps in Abadie et al., 2015:

1. Divide the pre-intervention periods: **training** and **validation** periods
2. Let $\tilde{w}_2(\mathbf{V}), \dots, \tilde{w}_{J+1}(\mathbf{V})$ be the SC in the **training** period for some \mathbf{V} . Calculate the MSPE with respect to the Y_{1t} in the **validation** period: $\sum_{t \in \mathcal{T}_0} (Y_{1t} - \tilde{w}_2(\mathbf{V})Y_{2t} - \dots - \tilde{w}_{J+1}(\mathbf{V})Y_{J+1t})^2$
3. Choose the value of \mathbf{V}^* minimizing the MSPE in 2 (not unique)
4. Use \mathbf{V}^* and the predictors data in the validation period $\mathbf{W}^* = \mathbf{W}(\mathbf{V}^*)$ [► Weights Selection](#)






EXTRA TOPICS

P-VALUE CALCULATION FOR INFERENCE

- ▶ The original proposal is to use an analogous method to **randomization inference**
 - ▶ Re-assign the treatment to every untreated unit, recalculates the coefficients, and collects them into a distribution used for inference
1. Iteratively apply the synthetic control method to each unit in the donor pool and obtain a distribution of placebo effects.
 2. Calculate the Root MSPE for each placebo for the pre-treatment period
 3. Calculate the Root MSPE for each placebo for the post-treatment period
 4. Compute the ratio of the post- to pre-treatment RMSPE
 5. Sort this ratio in descending order from greatest to highest
 6. Calculate the treatment unit's ratio in the distribution as $p = RANK / TOTAL$

California is first out of 38 states $\rightarrow p = 1/38 = 0.026$ ▶ Permutation Inference

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). **The economic costs of conflict: A case study of the basque country.** *American Economic Review*, 93(1), 113–132.
-  Athey, S., & Imbens, G. W. (2017). **The state of applied econometrics: Causality and policy evaluation.** *Journal of Economic perspectives*, 31(2), 3–32.
-  Card, D. (1990). **The impact of the mariel boatlift on the miami labor market.** *Illr Review*, 43(2), 245–257.