Lecture 12: Matching and Sythetic Control Method

Introduction to Econometrics,Spring 2023

Zhaopeng Qu

Nanjing University Business School

June 07 2023

Matching

Introduction

- Recall: The main identification strategy of OLS regression is **Control**, ie. puting **covariates** into the regression as control variables.
- The main identifying assumption of an OLS regression is
  - **Conditional Independence Assumption(CIA)**: which means that if we can "balance" covariates X then we can take the treatment D as randomized, thus

$$(Y_1, Y_0) \perp\!\!\!\perp D|X$$

  - Then ATE or ATT can be obtained to estimate the CEF

$$\delta = E[Y_{1i} - Y_{01} \mid X_i]$$

- Essentially the strategy compares treatment and control subjects who have **the same observable characteristics**, which is often called **Selection on observables**.
- In addition to OLS regression, **Matching** is another method based on **Selection on observables**.

Matching Estimator

- Suppose we have treated and untreated groups but the here assignment is not random. Then we can't obtain the causal effect because we don't know the counterfactural of an outcome Y_i in the treated group.
- The idea of matching method is quite simple. **What if we can construct a "reasonable" control group by selecting some(or all) samples in untreated group?** i.e Y_i^c
- then we can estimate the treatment effect

$$\hat{\delta} = \frac{1}{N_T} \Sigma_{i=1}^{N_T} (Y_i - Y_i^c)$$

  - N_T is the sample size in treatment group
  - Y_i^c is the corresponding counterfactual outcomes by matching(selecting) the sample from untreated group.

Matching Estimator: A Trainning Case

- the only covariates is X, which is used to select the "proper" counterfactuals

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
Average:		

Matching Estimator: an example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:		

Matching Estimator: A Trainning Case

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:	28.5	13982

Matching Estimator: A Training Case

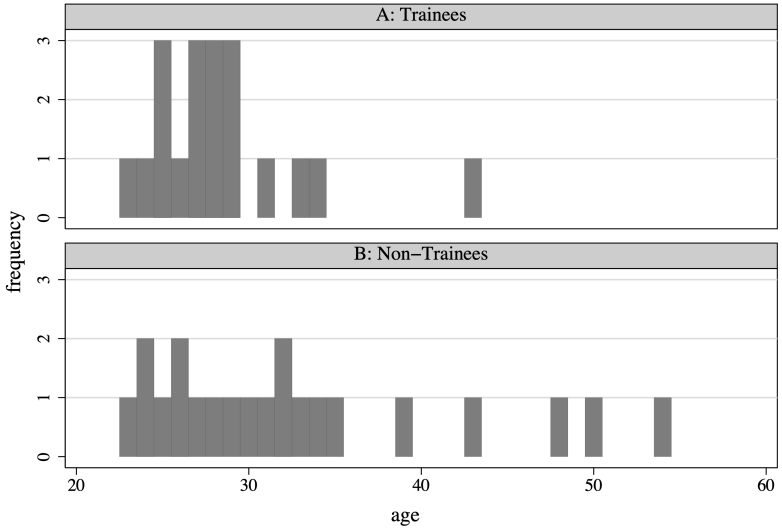
- Difference in average earnings between trainees and non-trainees
  - Before matching:

$$16426 - 20724 = -4298$$

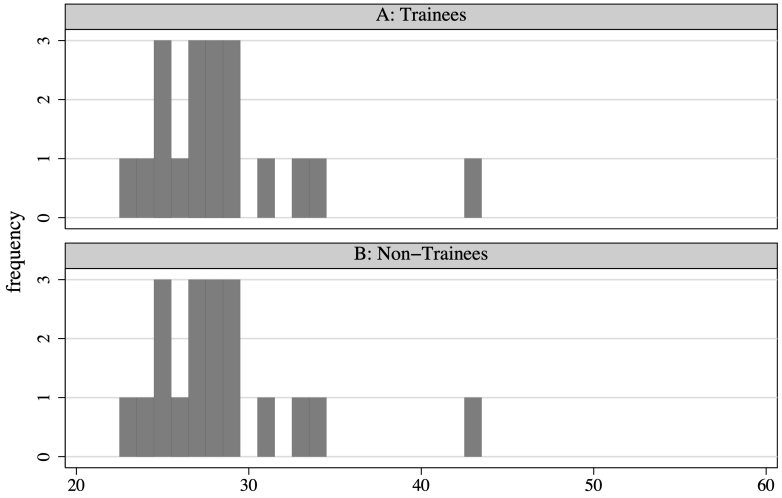
  - After mathcing:

$$16426 - 13982 = 2444$$

Age Distribution: Before Matching



Age Distribution: After Matching



Matching Estimators: Assumptions

Conditional Independence Assumption

$$(Y_1, Y_0) \perp\!\!\!\perp D|X$$

Common Support Assumption

$$0 < \Pr\left(D_i = 1 \mid X_i\right) < 1$$

Matching Estimators: Exact matching is hard

- The training case is an example of **Exact** matching which means that only units with identical covariate values are used to contruct the control group.
- But what if we have multiple covariates using to match, thus $X = (X_1, X_2, ...X_k)'$.
  - In this case, it is **impossible** to find proper units with identical values in all covariates $X_1, X_2, ...X_k$.
- Two complementary solutions running in parallel
  1. lower the accuracy of the comparison.
     - From "find a unit in the untreated group with the **same** covariate values" to "find a unit in the untreated group with **similar** covariate values."
  2. Directly reduce dimensionality by converting multiple variables into a single numerical value.

Matching Estimators: similarity between vectors

- If $X = (x_1, x_2, ...x_k)$ is a k-class vector, then the **distance** to measure "closeness" or "similarity" between two vectors such as X_i and X_j is the **Euclidean distance**

$$\| (X_i - X_j) \| = \sqrt{(X_i - X_j)'(X_i - X_j)} = \sqrt{\Sigma_{n=1}^{k}(X_{ni} - X_{nj})^2}$$

- The Euclidean distance is not invariant to changes in the scale of the X's. A more commonly used distance is the **normalized Euclidean distance**:

$$\| (X_i - X_j) \| = \sqrt{(X_i - X_j)'V_X^{-1}(X_i - X_j)}$$

where V is the symmetric and positive semidefinite variance matrix of X.

- Alternatively, a more general measure is **Mahalanobis distance**, which takes into account the correlation between variables.

$$\| (X_i - X_j) \| = \sqrt{(X_i - X_j)'\Sigma_X^{-1}(X_i - X_j)}$$

where Σ is the variance-covariance matrix of X.

Propensity Scores Matching(PSM)

Introduction

- Even we use the “distance” between vectors, the *Curse of dimensionality* makes matching on K covariates challenging.
- Rubin (1977) and Rosenbaum and Rubin (1983) develop a method that can contain those K covariates used for adjusting.
    - Propensity scores method
- Propensity scores summarize covariate information about treatment selection into a probability, thus the propensity scores.
- Then comparing units with similar estimated probabilities of treatment instead of Xs.

Defination

- A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables:

$$p(X) = Pr(D = 1 \mid X)$$

- Remind: D is a dummy variable which denotes the treatment status.

Propensity-score theorem

Propensity score theorem

If $(Y_{0i}, Y_{1i}) \perp D_i \mid X_i$, then $(Y_{0i}, Y_{1i}) \perp D_i \mid p(x_i)$ where $p(x_i) = \Pr(\mathrm{D}_i = 1 \mid X_i)$

- This theorem extends our CIA to a one-dimensional score, avoiding the curse of dimensionality.
- Conditioning on the propensity score is enough to have independence between the treatment indicator and the potential outcomes.
- To prove this theorem, we will show $E[D \mid Y_1, Y_0, p(X)] = p(X) = E(D \mid X)$ i.e. D is independent of (Y_1, Y_0) after conditioning on $p(X)$.

Propensity-score theorem

Proof

$$E\left[D \mid Y_1, Y_0, p(X)\right]$$

Propensity-score theorem

Proof

$$E\left[D \mid Y_1, Y_0, p(X)\right] = \underbrace{E\left[E\left[D \mid Y_1, Y_0, p(X), X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{by LIE}}$$

Propensity-score theorem

Proof

$$
\begin{aligned}
E\left[D \mid Y_1, Y_0, p(X)\right] &= \underbrace{E\left[E\left[D \mid Y_1, Y_0, p(X), X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{by LIE}} \\
&= \underbrace{E\left[E\left[D \mid Y_1, Y_0, X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{Given } X\text{, we know } p(X)}
\end{aligned}
$$

Propensity-score theorem

Proof

$$
\begin{aligned}
E\left[D \mid Y_1, Y_0, p(X)\right] &= \underbrace{E\left[E\left[D \mid Y_1, Y_0, p(X), X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{by LIE}} \\
&= \underbrace{E\left[E\left[D \mid Y_1, Y_0, X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{Given } X\text{, we know } p(X)} \\
&= \underbrace{E\left[E[D \mid X] \mid Y_1, Y_0, p(X)\right]}_{\text{by CIA}}
\end{aligned}
$$

Propensity-score theorem

Proof

$$
\begin{aligned}
E\left[D \mid Y_1, Y_0, p(X)\right] &= \underbrace{E\left[E\left[D \mid Y_1, Y_0, p(X), X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{by LIE}} \\
&= \underbrace{E\left[E\left[D \mid Y_1, Y_0, X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{Given } X\text{, we know } p(X)} \\
&= \underbrace{E\left[E[D \mid X] \mid Y_1, Y_0, p(X)\right]}_{\text{by CIA}} \\
&= \underbrace{E\left[p(X) \mid Y_1, Y_0, p(X)\right]}_{\text{propensity score definition}}
\end{aligned}
$$

Propensity-score theorem

Proof

$$
\begin{aligned}
E\left[D \mid Y_1, Y_0, p(X)\right] &= \underbrace{E\left[E\left[D \mid Y_1, Y_0, p(X), X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{by LIE}} \\
&= \underbrace{E\left[E\left[D \mid Y_1, Y_0, X\right] \mid Y_1, Y_0, p(X)\right]}_{\text{Given } X\text{, we know } p(X)} \\
&= \underbrace{E\left[E[D \mid X] \mid Y_1, Y_0, p(X)\right]}_{\text{by CIA}} \\
&= \underbrace{E\left[p(X) \mid Y_1, Y_0, p(X)\right]}_{\text{propensity score definition}} \\
&= p(X)
\end{aligned}
$$

Propensity-score theorem

Proof

Using a similar argument, we obtain

$$E\left[D \mid p(X)\right] = E\left[E[D|X] \mid p(X)\right] = E\left[p(X) \mid p(X)\right] = p(X)$$

Propensity-score theorem

Proof

Using a similar argument, we obtain

$$E\left[D \mid p(X)\right] = E\left[E[D|X] \mid p(X)\right] = E\left[p(X) \mid p(X)\right] = p(X)$$

Then

$$E\left[D \mid p(X)\right] = E\left[D \mid Y_1, Y_0, p(X)\right]$$

Propensity-score theorem

Proof

Using a similar argument, we obtain

$$E\left[D \mid p(X)\right] = E\left[E[D|X] \mid p(X)\right] = E\left[p(X) \mid p(X)\right] = p(X)$$

Then

$$E\left[D \mid p(X)\right] = E\left[D \mid Y_1, Y_0, p(X)\right]$$

Thus

$$(Y_{0i}, Y_{1i}) \perp D_i \mid p(x_i)$$

Propensity score matching

- Based on CIA, we need only control for covariates that affect the probability of treatment to obtain the causal effect.
- Base on propensity score theorem: Actually we don't need to control all covariates, but the only one is the probability of treatment, thus the propensity scores

$$p(X) = Pr(D = 1 \mid X)$$

Propensity score matching

- How to obtain the propensity scores?
  - estimate it, thus $\hat{p}(X)$
- Many ways to do
  - Logit Model with flexible specification(with interactions)
  - Kernel regression
  - Machine learning

Matching and PSM in practice

- Matching in practice
  - both directions matching
  - 1:1 matching v.s m:1
  - With or without replacement
  - greedy or optimal technique
  - with or without a caliper width
- Choosing the “best” matching method for one’s data depends on the unique characteristics of the dataset as well as the goals of the analysis.

Matching v.s Regression

- Both matching and regression rely on CIA (selection on observables). Most biases we could suffer in regression, such as OVB, measurement error, and simultaneous causality, will not be avoided even if we use matching.
- Why we still need matching?
  - Due to its non-parametric characteristics, matching does not impose any restrictions on empirical specification or estimate specific parameters of the CEF function.
  - Regression does not account for the common support issue.
- Using matching alone is less common in economics, but it can be combined with other methods like DID and SCM.

Extensions of DID(II): Synthetic Control Method(SCM)

Basic Idea

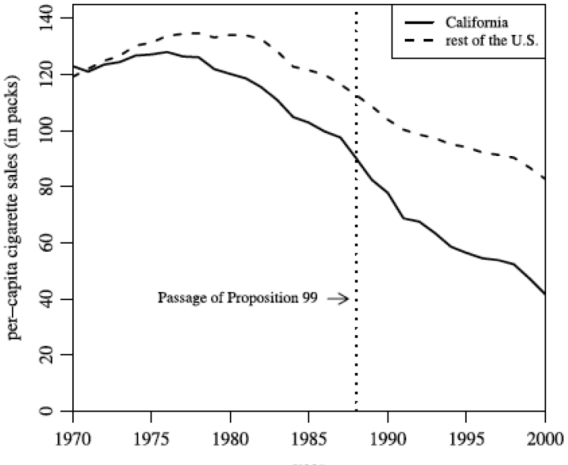
- The **synthetic control method**(SCM) were originally proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the aim to estimate the effects of aggregate interventions,
- Interventions that are implemented at an aggregate level affecting a small number of large units (such as a cities, regions, or countries), on some aggregate outcome of interest.
- The basic idea behind synthetic controls is that a combination of units often provides a better comparison for the unit exposed to the intervention than any single unit alone.
  - a data-driven procedure to use a small number of non-treated units to build the suitable counterfactuals.
- It is useful for case studies, which is nice because that is often all we have.
- Continues to also be methodologically a frontier for applied econometrics and is widely used in many field, even outside academia.

Extensions of DID: Synthetic Controls Method

- The basic idea is use (long) longitudinal data to build the **weighted average of non-treated** units that best reproduces characteristics of the treated unit over time in pre-treatment period.
- The weighted average of non-treated units is the **synthetic cohort**.
- Causal effect of treatment can be quantified by a simple difference after treatment:
  - **treated vs synthetic cohort**.

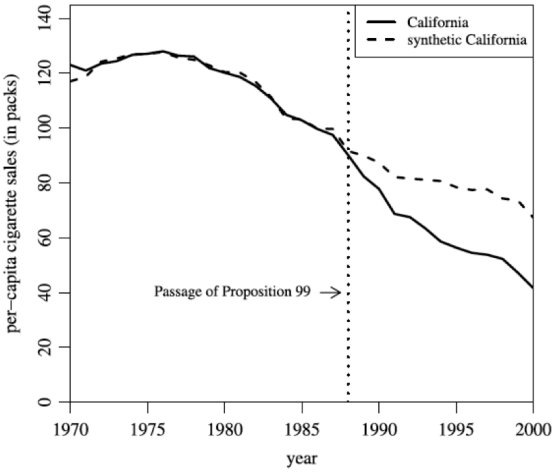
Abadie et.al(2010): Tax on Cig-Consumption

- In 1988, California passed comprehensive tobacco control legislation: Increased cigarette taxes by $0.25 per pack.



Abadie et.al(2010): Tax on Cig-Consumption

- Using 38 states that had never passed such programs as controls: **Synthetic CA**



Predictor Means: Actual vs Synthetic California

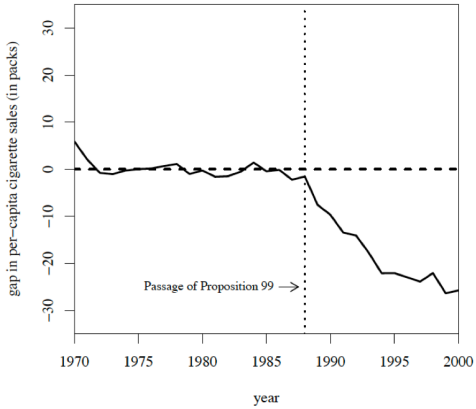
- Most observables are similar between Actual and Synthetic

	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 Application: Actual vs Synthetic California

- The treatment effect is measured by the gap in ciga-sales between Actual and Synthetic



Formalization

Formalization: The Setting

- Suppose that we obtain data for $J+1$ units: $j = 1, 2, ..., J+1$
  - Assume that the first unit ($j = 1$) is the **treated unit**, that is, the unit affected by the policy intervention of interest.
  - Then the set of potential comparisons,$j = 2, ..., J+1$ is a collection of **untreated units**, not affected by the intervention.
- Assume also that our data span T periods and that the first T_0 periods are **before the intervention**.
- Let Y_{jt} and Y_{jt}^C be the real and counterfactual outcomes of interest for unit j of $J+1$ aggregate units at time t with and without intervention.
- Then the effect of the intervention of interest for the affected unit in period $t(t > T_0)$(ATT)

$$\tau_{1t} = Y_{1t} - Y_{1t}^C$$

Formalization: The Setting

- How to reproduce Y_{1t}^{C} which is totally unobservable?
  - Use unaffected units in control groups to predict it like matching in cross-sectional data.
- More specifically, a weighted average of the units in the comparison group use to construct the potential outcome of treated units, which define as **synthetic control**.Thus,

$$\hat{Y}_{1t}^{C} = \Sigma_{j=2}^{J+1} w_j Y_{jt}$$

- Then the question is how to determine these values of the weights, w_j

Formalization: Weights

- Let more specifically, $W = (w_2, ..., w_{J+1})'$ have to satisfy two restriction conditions
  - $w_j \geq 0$ for $j = 2, ..., J+1$
  - $\Sigma_{j=2}^{J+1} w_j = 1$
- **Key Question**: how to determine these values of the weights, w_j or how to construct a proper control group?
  - eg. assigning equal weights, thus

$$w_j = \frac{1}{J}$$

  - or a fraction of the total population in the comparison group(at the time of the intervention),thus

$$w_j = \frac{N_j}{\Sigma_{j=2}^{J+1} N_j}$$

Formalization: Weights of Xs

- For each unit, j, we also observe a set of characteristics or covariates which can be use to predict the outcome Y_{jt}, denoted as $X_{1j}, ...X_{kj}$
- Let X_1 is a $k \times 1$ vector of these characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the untreated units.
- Let X_1 is a $k \times 1$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- Recall: how to measure the closeness or similarity between two vectors?

Formalization: Weight by Matching

- The rule to choose the optimal weight vector $W^* = (w_2, ..., w_{J+1})'$ will be

$$argmin_W \parallel (X_1 - X_0 W) \parallel$$

  - Thus,the optimal vector of weight W should **minimize the "distance"** between treated unit and untreated group,subject to two weight constraints.

- More specifically, *Abadie, et al(2010)* consider

$$\parallel (X_1 - X_0 W) \parallel_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where V can be some $(k \times k)$ symmetric and positive semidefinite matrix.

Formalization: More on the V matrix

- Typically, V is diagonal with main diagonal $v_1, \ldots, v_k$. Then the synthetic control weights minimize

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

  - Where v_m is a **weight** that reflects the *relative importance* that we assign to the m^{th} variable when we measure the discrepancy between the treated unit and the synthetic controls.
  - And v_m is critical because it weights directly shape w_j,which help reproducing the counterfactual outcome for the treated unit in the absence of the treatment.

Formalization: Estimating the V matrix

- Various ways to choose V
  - In practice,most people choose V that minimizes *the mean squared prediction error(MSPE)*.Thus,

$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V) Y_{jt} \right)^2$$

- If the number of pre-intervention periods in the data is “large”, then matching on pre-intervention outcomes can allow us to control for the heterogeneous responses to multiple unobserved factors.
- The intuition here is that only units that are alike on unobservables and unobservables would follow a similar trajectory pre-treatment.

A Machine learning procedure

1. Divide the pre-intervention periods(T_0) into a initial **training** period($t = 1, ...t_0$) and a subsequent **validation** period($t = t_0 + 1, ...T_0$).
2. Select a value V^* make the MSPE is small

$$\sum_{t=t_0+1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j(V) Y_{jt}\right)^2$$

3. Use the resulting V^* and data on the predictors for the last t_0 before in the intervention,$t = t_0 + 1, t_0 + 2, ..., T_0$ to calculate $w^* = w(V^*)$

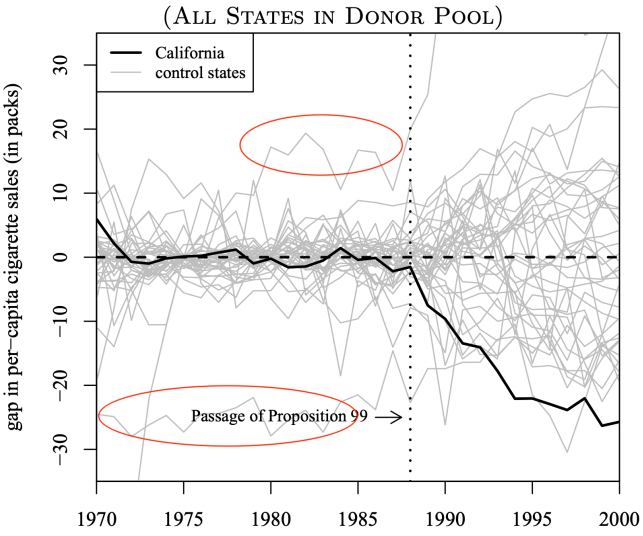
Inference

- Permutation Strategy: whether the effect estimated by the synthetic control for the unit affected by the intervention is **large** relative to the effect estimated for a unit chosen at random.
- Implementation: “randomization” of the treatment to each unit, re-estimating the model, and calculating a set of root mean squared prediction error (RMSPE) values for the pre- and post-treatment period.
- For $0 \leq t_1 \leq t_2 \leq T$ and $j = 1, 2, ..., J + 1$,let

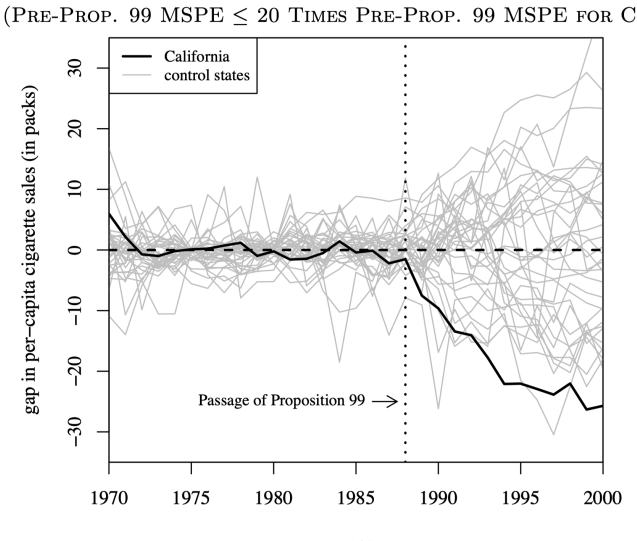
$$R_j(t_1, t_2) = \Big(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (Y_{jt} - \hat{Y}_{jt}^N)^2\Big)^{\frac{1}{2}}$$

- Some states whose pre-treatment RMSPE is considerably different than California’s can be dropped.

Inference: Dropping Sample

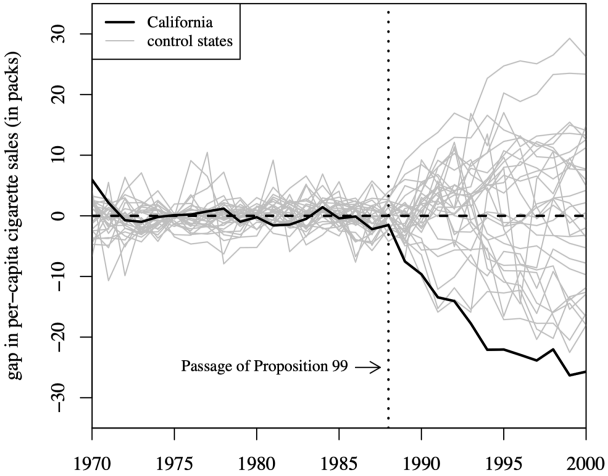


Inference: Dropping Sample



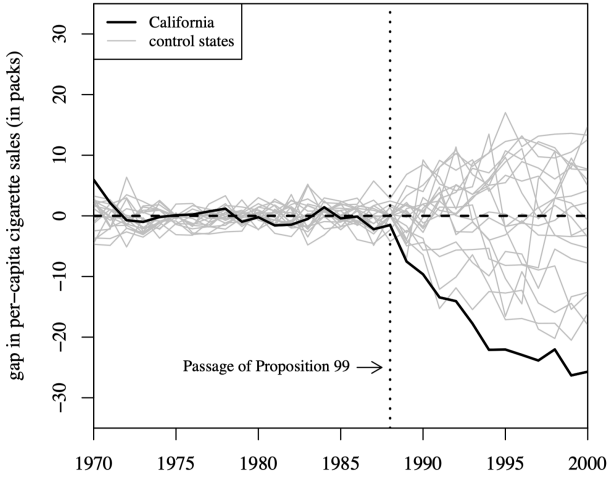
Inference: Dropping Sample

(Pre-Prop. 99 MSPE $\leq$ 5 Times Pre-Prop. 99 MSPE for CA)



Inference: Dropping Sample

(Pre-Prop. 99 MSPE $\leq$ 2 Times Pre-Prop. 99 MSPE for CA)



Inference: Procedure

1. Iteratively apply the synthetic method to each state in the unaffected group and obtain a distribution of placebo effects.
2. Calculate the RMSPE(root mean squared prediction error) for *each placebo* for the pre-treatment and post-treatment.

- Post-treatment $R_{j,post} = RMSPE_j(T_0 + 1, T)$
- Pre-treatment $R_{j,pre} = RMSPE_j(1, T_0)$

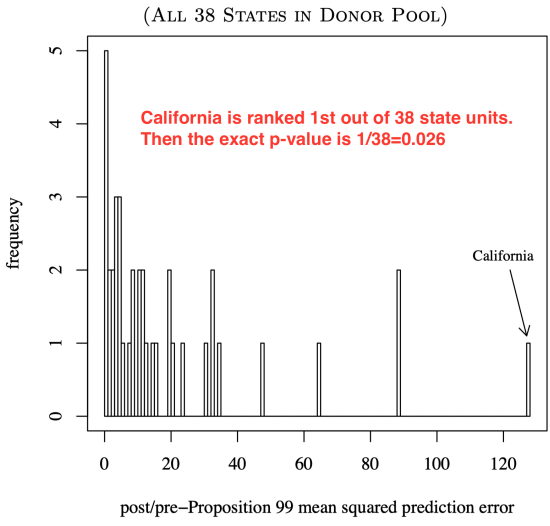
3. Compute the ratio of the post-to-pre-treatment and sort it in descending order from greatest to highest. Thus

$$r_j = \frac{R_{j,post}}{R_{j,pre}}$$

4. The exact p-value is defined as

$$p - value = \frac{rank_{th}}{J + 1}$$

Inference: P-Value

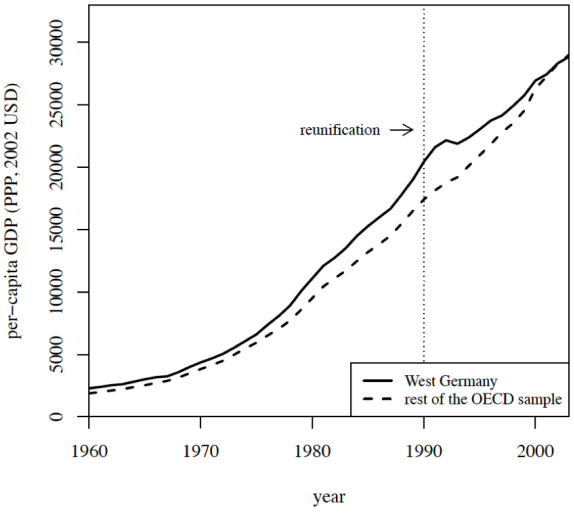


An Application: The 1990 German Reunification

The Economic Effect of the German Reunification on West Germany

- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics.
  - “What do Thailand, the Dominican Republic, Zimbabwe,Greece and Bolivia have in common that merits their being put in the same regression analysis? Answer: For most purposes, nothing at all.” (Harberger 1987)
- Application: The economic effect of “Berlin Wall” Falling,thus the 1990 German reunification,on West Germany.
- Control group is compositional restricted to 16 OECD countries

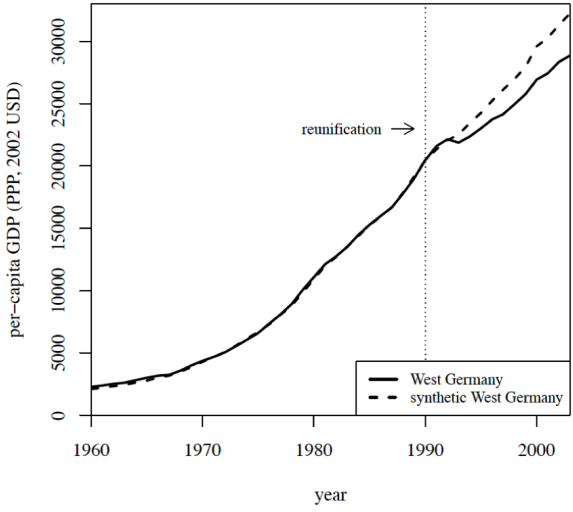
West Germany v.s. OECD countries



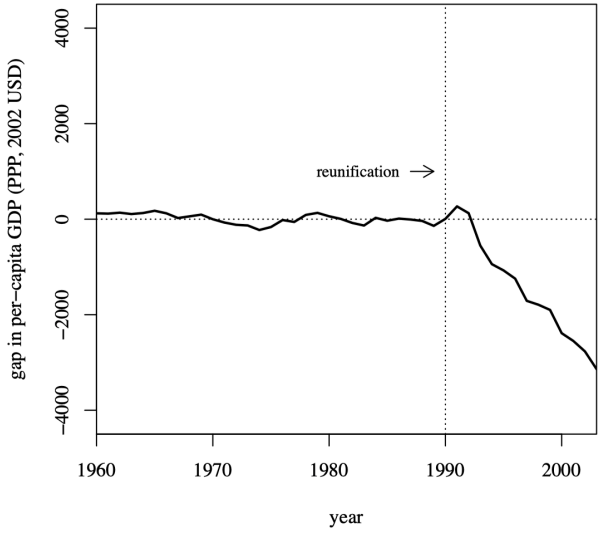
Economic Growth Predictors Means across groups

	West Germany	Synthetic West Germany	OECD Sample
GDP per-capita	15808.9	15800.9	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

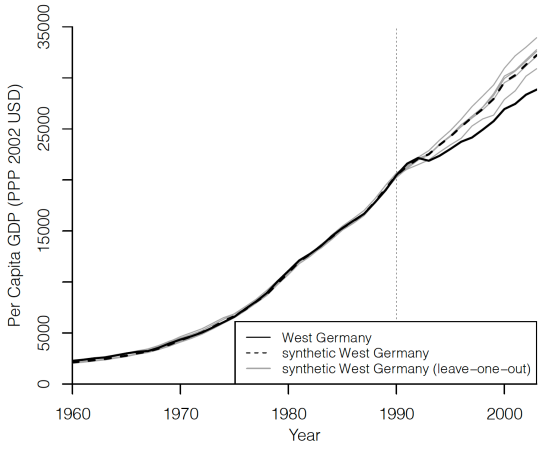
West Germany v.s Sythetic West Germany



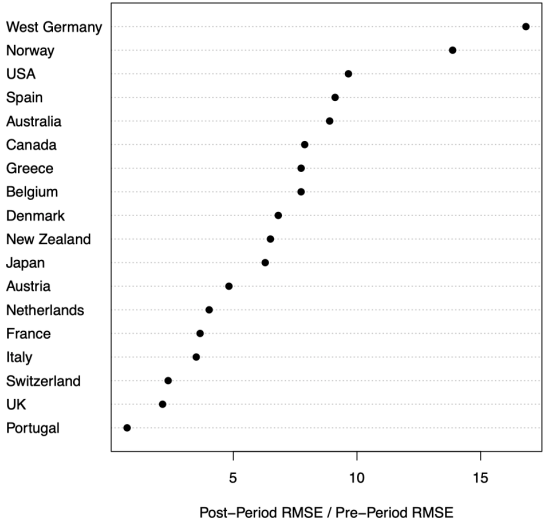
GDP Gap: West Germany and synthetic West Germany



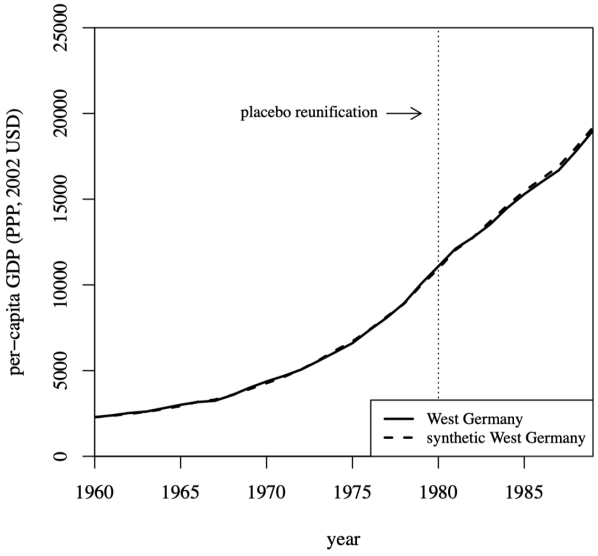
The 1990 German Reunification: Leave-one-out estimates



RMSE Test


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

Placebo Test: What if '1980' German Reunification



Wrap Up

- **Synthetic control method** provide many practical advantages for causal inference.
- The credibility of the results depends on
  - the level of diligence exerted in the application
  - whether contextual and data requirements are met

A Summary of Causal Inference Method

The goal of causal inference

- Build a reasonable counterfactual world by naturally occurring data to find or construct a proper control group is the core of econometrical methods.
- Common Idea: match similar units, and produce a proper comparison
  - OLS: gives conditional mean comparison
  - Matching: a weighted conditional mean comparison
  - IV: compares difference between instrumented and non-instrumented groups.
  - RD: compares means around the cutoff.
  - DID: compares the changes of the difference across locations.
  - SCM: compares the gaps between treated and sythetic control groups.
- All are about a a **believable** and **reliable** comparison.

Final Thoughts(Angrist and Pischeke,2008)

- A good research design is one you are excited to tell people about
  - that's basically what characterizes all research designs, whether **instrumental variable**,**regression discontinuity designs** or **difference-in-differences**,**synthetic control method** among others(**Seven Magic Weapons**).
- Causality is *easy and hard*. Don't get confused which is the hard part and which is the easy part.
- Always understand *what assumptions you must make*, be clear which parameters you are and are not identifying.
- Last but not least, Remember: **Good question is always the first priority**. Along with good research design is in the second place.
- What is a good research question?
  - interesting(people cares) and/or relevent(does matter something)
  - should not simply duplicate existing research, but instead should aim to be innovative and unique.

Though still a long way to go but now we could take a break and enjoy the landscape.