Synthetic Control Discussion

INFO/STSCI/ILRST 3900: Causal Inference

12 Nov 2025

Reminders and Announcements

- Peer reviews for Pset 5 are due **Nov 18**
- Quiz 5 is on **Nov 18**
- Project groups are posted on Canvas (under Modules)

Synthetic Control: big idea

- Many pre- and post-treatment periods in the data
- Treated unit is "unique", there is no single control unit that is a direct match
- Construct synthetic unit to approximate untreated version of treated unit using weighted average of untreated units
- Pick weights to match pre-treatment characteristics (either covariates or observations)

Picking weights

- In class, we mentioned selecting weights to directly minimize pre-treatment fit

$$\underbrace{\sum_{t<T_0}}_{pre-treatment} \left(\underbrace{Y_{t,1}}_{\text{outcome of treated unit}} - \underbrace{\sum_{j} w_j Y_{t,j}}_{\text{weighted avg of control units}} \right)^2$$

- Intuition:
  - synthetic unit represents the treated unit under no treatment
  - in pre-treatment period, treated unit has not yet received treatment
  - outcomes of the synthetic unit pre-treatment should be very close to the outcomes of the treated unit pre-treatment

Picking weights

- Let X_1 denote a vector of pre-treatment covariates for the (eventually) treated unit (including some pre-treatment observations)
- Let X_0 denote the matrix of corresponding of covariates (including some pre-treatment observations) for the donor pool
- Let V be a diagonal matrix which weights how important matching each covariate is
- Select weights to minimize

$$(X_1 - X_0 W)^T V (X_1 - X_0 W) = \sum_h v_h (X_{1,h} - \sum_j w_j X_{j,h})^2$$

so that for each covariate $X_{1,h}$

$$X_{1,h} \approx \sum_j w_j X_{j,h}$$

Picking weights

- Different V lead to different optimal weights $w(V)$
- Can specify V directly (remember Mahalonobis distance?)
- Most commonly select V to minimize pre-treatment mean squared error

$$\sum_{t<T_0} \left(Y_{t,0} - \sum_j w_j(V) Y_{t,j} \right)^2$$

Picking weights

- Different V lead to different optimal weights $w(V)$
- Can specify V directly (remember Mahalonobis distance?)
- Most commonly select V to minimize pre-treatment mean squared error

$$\sum_{t<T_0} \left(Y_{t,0} - \sum_j w_j(V) Y_{t,j} \right)^2$$

- Why? because we want our synthetic version of the treated unit to actually match the treated unit's outcomes in the pre-treatment period

Picking weights

- Overfitting can also be assessed using backdating
- Pick another time period in pre-treatment period as a "fake treatment time"
- Re-run synthetic control with "fake treatment time"
- Assess how well synthetic unit predicts after "fake treatment time"

Picking weights

- Overfitting can also be assessed using backdating
- Pick another time period in pre-treatment period as a "fake treatment time"
- Re-run synthetic control with "fake treatment time"
- Assess how well synthetic unit predicts after "fake treatment time"
- Or "placebo/permutation tests" (tomorrow in Lecture)
- Run synthetic control with a control unit as the treated unit
- Compare the "effect of treatment" for units who never actually received treatment to the effect of treatment of the unit that actually did recieve it

Synthetic Control - Application

Research Question: Does violent conflict affect economic output?

- In the mid 1970's the Basque Country region of Spain was afflicted by a series of violent terrorist attacks.
- This was specific to the Basque Country region and did not affect the other regions of Spain.
- We can use Synthetic Control here! The pre-treatment period is before the terrorist attacks, and all the other regions in Spain will form our synthetic control donor pool!
- We will construct a control unit from all other regions and then compare the economic output of the Basque Country region after the terrorist attacks to our control unit.

Evaluating our Synthetic Control

How do we check if our Synthetic Control is any good!?

- Like matching, construct synthetic control using covariates, including regional economic activity, population levels, etc.
- Like matching, we want our treated unit and our synthetic control to be balanced on covariates

	Treated	Synthetic	Sample Mean
school.illit	39.888	256.335	170.786
school.prim	1031.742	2730.092	1127.186
school.med	90.359	223.341	76.260
school.high	25.728	63.437	24.235
school.post.high	13.480	36.154	13.478
invest	24.647	21.583	21.424
special.gdpcap.1960.1969	5.285	5.271	3.581
special.sec.agriculture.1961.1969	6.844	6.179	21.353
special.sec.energy.1961.1969	4.106	2.760	5.310
special.sec.industry.1961.1969	45.082	37.636	22.425
special.sec.construction.1961.1969	6.150	6.952	7.276
special.sec.services.venta.1961.1969	33.754	41.104	36.528
special.sec.services.nonventa.1961.1969	4.072	5.371	7.111
special.popdens.1969	246.890	196.287	99.414

Evaluating our Synthetic Control

We let an optimization algorithm pick weights. Then, we can actually look at the weights!

w.weights	unit.names	unit.numbers
0.000	Andalucia	2
0.000	Aragon	3
0.000	Principado De Asturias	4
0.000	Baleares (Islas)	5
0.000	Canarias	6
0.000	Cantabria	7
0.000	Castilla Y Leon	8
0.000	Castilla-La Mancha	9
0.851	Cataluna	10
0.000	Comunidad Valenciana	11
0.000	Extremadura	12
0.000	Galicia	13
0.149	Madrid (Comunidad De)	14
0.000	Murcia (Region de)	15
0.000	Navarra (Comunidad Foral De)	16
0.000	Rioja (La)	18

Evaluating our Synthetic Control

We let an optimization algorithm pick weights. Then, we can actually look at the weights!

w.weights	unit.names	unit.numbers
0.000	Andalucia	2
0.000	Aragon	3
0.000	Principado De Asturias	4
0.000	Baleares (Islas)	5
0.000	Canarias	6
0.000	Cantabria	7
0.000	Castilla Y Leon	8
0.000	Castilla-La Mancha	9
0.851	Cataluna	10
0.000	Comunidad Valenciana	11
0.000	Extremadura	12
0.000	Galicia	13
0.149	Madrid (Comunidad De)	14
0.000	Murcia (Region de)	15
0.000	Navarra (Comunidad Foral De)	16
0.000	Rioja (La)	18

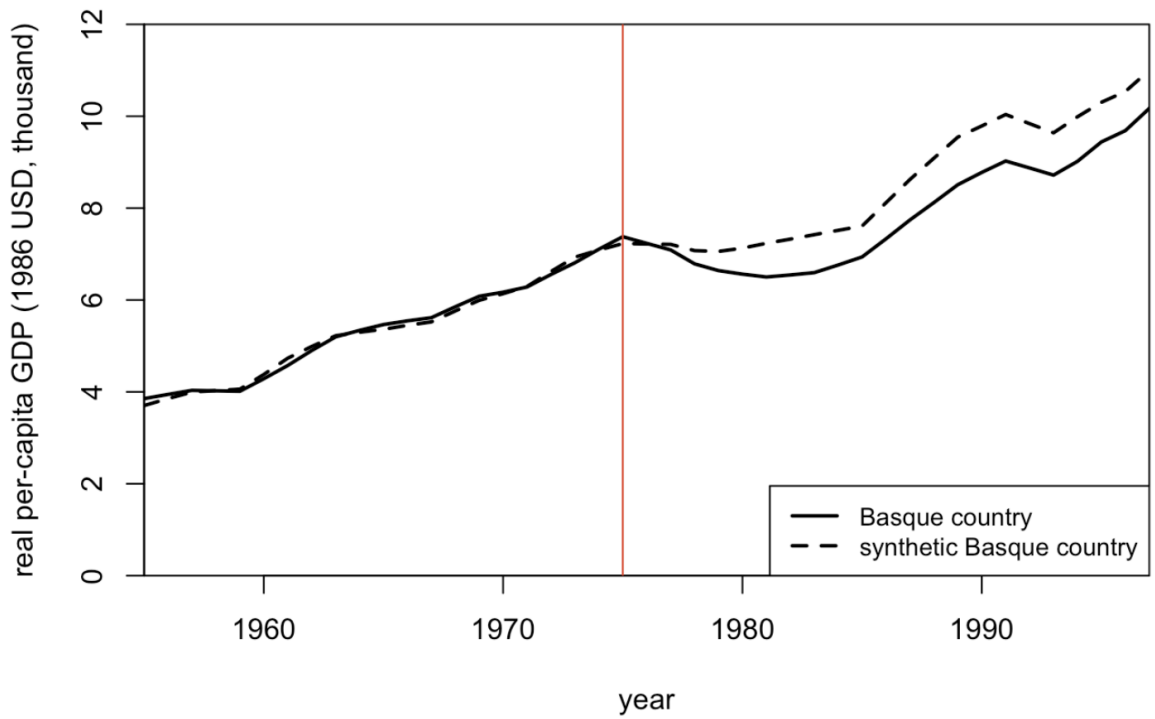
- We can see that only *two* regions contribute at all to our synthetic control

Synthetic Control versus Regression- Interpretability

- By restricting weights in synthetic control to be non-negative and sum to one, we introduce sparsity
- By sparsity, we mean many weights equal 0
- Also, with this restriction, makes the synthetic control easy-to-interpret
- Example: Basque Country in Spain is about 85% Cataluna and about 15% Madrid
- Could use regression instead without restricting the weights, but then you don't get sparsity and you may get negative weights... what does it mean for a region to be negative percent of another?

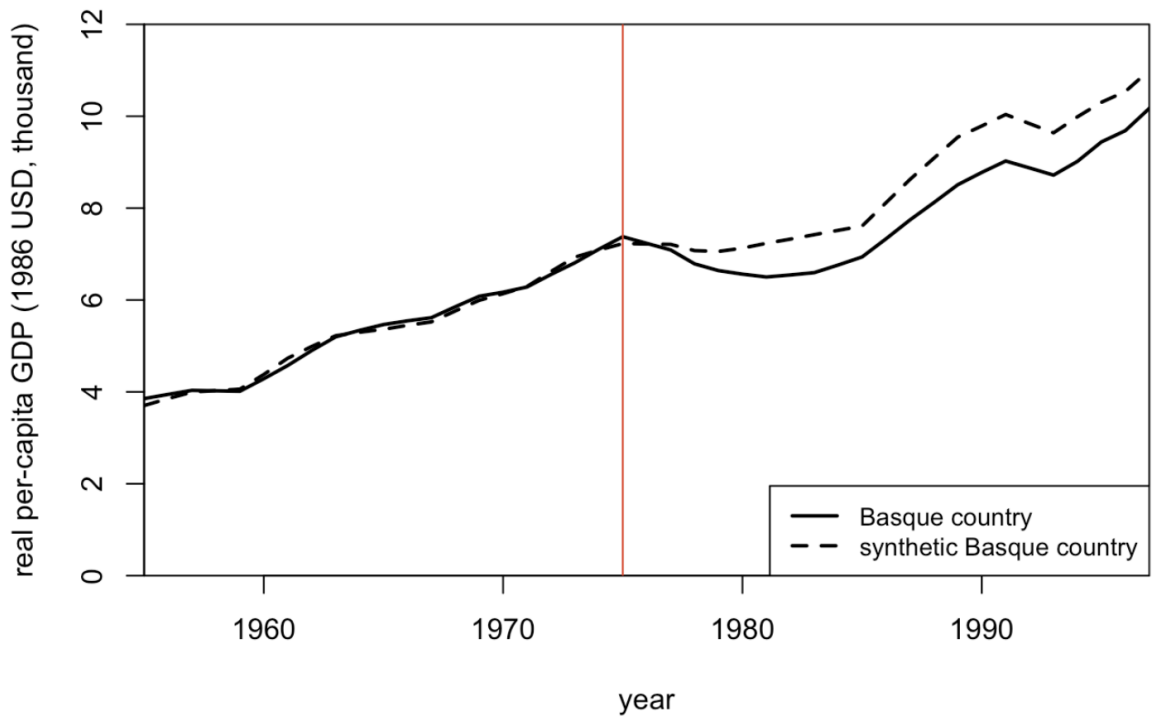
Is there a Causal Effect?

- Goal: estimate the causal effect of violent conflict on economic output
- How do we determine if there really is a causal effect?
- Compare economic output of the Basque Country region to our synthetic control unit after the terrorist attacks began



Is there a Causal Effect?

- Goal: estimate the causal effect of violent conflict on economic output
- How do we determine if there really is a causal effect?
- Compare economic output of the Basque Country region to our synthetic control unit after the terrorist attacks began



- This trend indicates that economic output dropped by quite a bit as a result of the violent conflict!

Check Your Understanding

- What do you notice about the outcomes of Basque country and synthetic Basque country in the pre-treatment period?
- Based on the post-treatment period, why might we think there is a causal effect?