Synthetic Control Discussion

INFO/STSCI/ILRST 3900: Causal Inference

8 Nov 2023

Reminders and Announcements

- Peer reviews for HW5 are **due tomorrow** by 5pm
- If you weren't assigned a peer review, we won't count it against you
- HW 6 up tomorrow;
  - DID and Synthetic Control
  - Due next Thursday (11/16) by 5pm
  - No peer reviews
- Project Write-up due Tuesday 11/21 by 5pm

Synthetic Control: big idea

Synthetic Control: big idea

- Many pre- and post-treatment periods in the data
- Treated unit is "unique"
- Not so many units in control group

Synthetic Control: big idea

- Many pre- and post-treatment periods in the data
- Treated unit is "unique"
- Not so many units in control group
- 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)

Synthetic Control: big idea

- Many pre- and post-treatment periods in the data
- Treated unit is "unique"
- Not so many units in control group
- 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)
- Allows for estimating time-varying trends

Discussion questions

Compare and contrast the following methods: matching, differences in differences (DID), and synthetic control.

- When might you prefer DID to synthetic control or vis-versa?
- When might you prefer matching to synthetic control or vis-versa?

Synthetic control and Matching

Synthetic control and Matching

In some ways, synthetic control can be seen as a specific form of matching

- Predict unobserved potential outcome using observed outcome of "similar" units
- Can choose "matches" (i.e., weights) to match untreated outcomes (of eventually treated unit)

Synthetic control and Matching

In some ways, synthetic control can be seen as a specific form of matching

- Predict unobserved potential outcome using observed outcome of "similar" units
- Can choose "matches" (i.e., weights) to match untreated outcomes (of eventually treated unit)
- Synthetic control differs in how weights are chosen
- Data across time (longitudinal) so we also observed untreated outcomes of (eventually) treated unit

Synthetic control and Matching

In some ways, synthetic control can be seen as a specific form of matching

- Predict unobserved potential outcome using observed outcome of "similar" units
- Can choose "matches" (i.e., weights) to match untreated outcomes (of eventually treated unit)
- Synthetic control differs in how weights are chosen
- Data across time (longitudinal) so we also observed untreated outcomes of (eventually) treated unit
- Can directly match to minimize pre-treatment fit

Synthetic control and Difference and Difference

Synthetic control and Difference and Difference

- Both have observations pre and post treatment
- Diff-in-Diff requires parallel trends assumption

Synthetic control and Difference and Difference

- Both have observations pre and post treatment
- Diff-in-Diff requires parallel trends assumption
- In synthetic control, we have a similar assumption, but parallel trends holds for synthetic unit
- Generally, Diff-in-Diff has fixed set of comparison units using prior knowledge (i.e., NJ vs PA)

Synthetic control and Difference and Difference

- Both have observations pre and post treatment
- Diff-in-Diff requires parallel trends assumption
- In synthetic control, we have a similar assumption, but parallel trends holds for synthetic unit
- Generally, Diff-in-Diff has fixed set of comparison units using prior knowledge (i.e., NJ vs PA)
- Synthetic control, we can start with a large "donor pool" and select weights using data

Picking weights

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

$$\sum_{t<T_0} \left(Y_{t,1} - \sum_j w_j Y_{t,j} \right)^2$$

- If there are many units in the donor pool, and not very many pre-treatment periods this may overfit to our data
- Why?

Picking weights

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

$$\sum_{t<T_0} \left(Y_{t,1} - \sum_j w_j Y_{t,j} \right)^2$$

- If there are many units in the donor pool, and not very many pre-treatment periods this may overfit to our data
- Why?

Potential solution: also include other covariates and don't include every pre-treatment observation

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

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

- Not including all pre-treatment observations in the original minimization may guard against overfitting
- Including all pre-treatment observations in vector of pre-treatment covariates may reduce bias but increase variance

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!?

- Similar to matching, we will use a bunch of data to construct our synthetic control. This will include regional economic activity, population levels, average education levels, etc.
- Also like matching, we want our treated unit and our synthetic control to have super similar average values for these variables.

	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 also want to see which regions in Spain contribute the most to our synthetic control unit

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 also want to see which regions in Spain contribute the most to our synthetic control unit

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

Evaluating our Synthetic Control

Similarly, we want to see which of our input variables are the most important for constructing our control unit

	v.weights
school.illit	0.039
school.prim	0.001
school.med	0
school.high	0
school.post.high	0
invest	0
special.gdpcap.1960.1969	0.041
special.sec.agriculture.1961.1969	0.24
special.sec.energy.1961.1969	0.022
special.sec.industry.1961.1969	0.248
special.sec.construction.1961.1969	0.006
special.sec.services.venta.1961.1969	0.011
special.sec.services.nonventa.1961.1969	0.049
special.popdens.1969	0.343

Evaluating our Synthetic Control

Similarly, we want to see which of our input variables are the most important for constructing our control unit

	v.weights
school.illit	0.039
school.prim	0.001
school.med	0
school.high	0
school.post.high	0
invest	0
special.gdpcap.1960.1969	0.041
special.sec.agriculture.1961.1969	0.24
special.sec.energy.1961.1969	0.022
special.sec.industry.1961.1969	0.248
special.sec.construction.1961.1969	0.006
special.sec.services.venta.1961.1969	0.011
special.sec.services.nonventa.1961.1969	0.049
special.popdens.1969	0.343

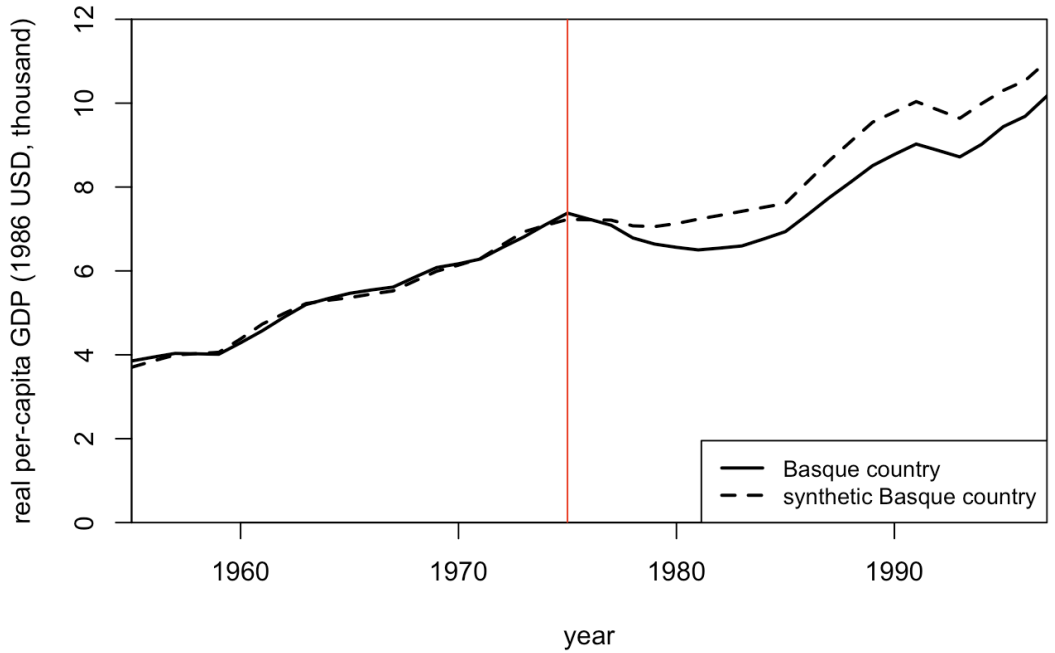
- We can see that four of our variables have a weight of 0. They don't contribute at all.
- The most important variable here is the population density of each region!

Is there a Causal Effect?

- The whole goal here is to estimate the causal effect of violent conflict on economic output.
- How do we determine if there really is a causal effect?

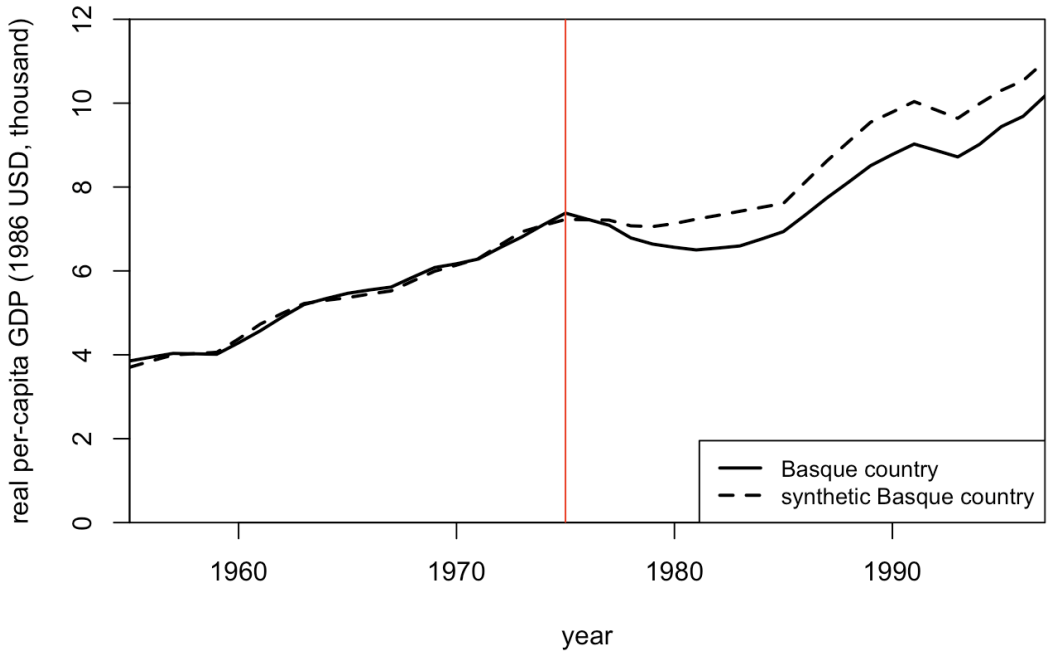
Is there a Causal Effect?

- The whole goal here is to estimate the causal effect of violent conflict on economic output.
- How do we determine if there really is a causal effect?
- We can look at the economic output of the Basque Country region compared to our synthetic control unit and see how it changes after the terrorist attacks began.



Is there a Causal Effect?

- The whole goal here is to estimate the causal effect of violent conflict on economic output.
- How do we determine if there really is a causal effect?
- We can look at the economic output of the Basque Country region compared to our synthetic control unit and see how it changes after the terrorist attacks began.



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

Code for this Example

On the website, there is a fully-worked example using this data.