Synthetic Control Discussion

INFO/STSCI/ILRST 3900: Causal Inference

13 Nov 2024

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)

Synthetic Control: big idea

Group Activity: In groups of 2-3...

- Icebreaker: share your name and something you're looking forward to over winter break
- Compare and contrast synthetic control with matching (what is similar? what is different?)
- Compare and contrast synthetic control with difference in differences (what is similar? what is different?)
- When would you use synthetic control versus difference in differences versus 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

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
- With matching, typically looking for the single most “similar” unit or average of “k closest” units
- Synthetic control: we actually create an *artificial* (i.e. synthethic) unit to match the treated unit to

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
- With matching, typically looking for the single most "similar" unit or average of "k closest" units
- Synthetic control: we actually create an *artificial* (i.e. synthethic) unit to match the treated unit to
- This "match" is a weighted combination of control units
- Example: "Synthetic" Travis Kelce

$$Y_{t,Synthetic}^{\text{NS}} = .5 \times Y_{t,Mahomes} + .25 \times Y_{t,Bosa} + .25 \times Y_{t,Jefferson}$$

Synthetic control and Difference and Difference

- Both have observations pre and post treatment

Synthetic control and Difference and Difference

- Both have observations pre and post treatment
- Diff-in-Diff: requires parallel trends assumption
- Synthetic control: similar assumption, parallel trends holds for synthetic unit

Synthetic control and Difference and Difference

- Both have observations pre and post treatment
- Diff-in-Diff: requires parallel trends assumption
- Synthetic control: similar assumption, 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

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

- 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

- X_1: vector of pre-treatment covariates for the (eventually) treated unit (including some pre-treatment observations)
- X_0: matrix of corresponding of covariates for the donor pool
- Let V be a diagonal matrix
  - element v_{ii} is weight for covariate i representing how important that covariate is in the matching
  - we get to pick V first
- Select weights w_j to minimize

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

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

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

- Where does V show up in the equation above?

Picking weights

- Different V lead to different optimal weights $w(V)$
- There are different ways to choose V
- 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

- Different V lead to different optimal weights $w(V)$
- There are different ways to choose V
- 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

Group Activity: In the same groups as before, discuss why we want the pre-treatment outcomes of the synthetic unit and treated unit to match.

Synthetic Control - Application

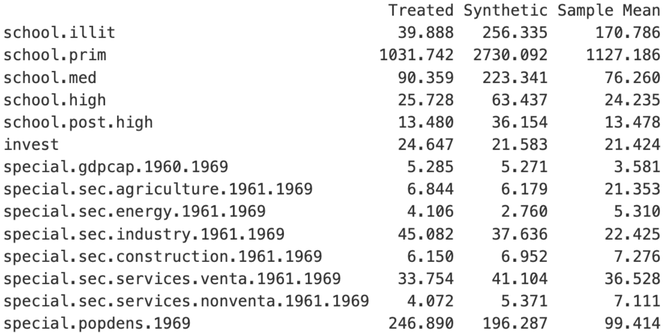
Research Question: Does violent conflict affect economic output?

- 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

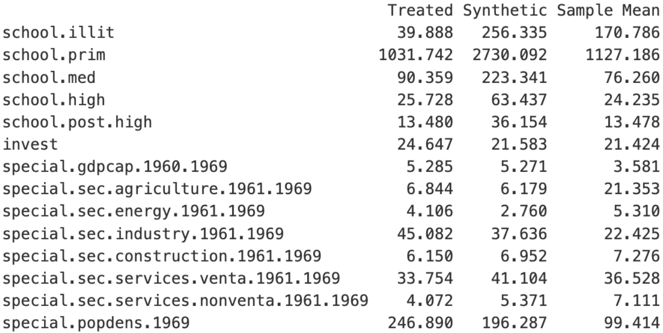


```
                                          Treated Synthetic Sample Mean
school.illit                               39.888   256.335     170.786
school.prim                              1031.742  2730.092    1127.186
school.med                                 90.359   223.341      76.260
school.high                                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

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

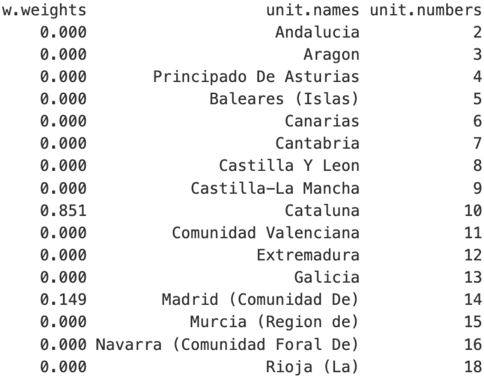


```
                                        Treated Synthetic Sample Mean
school.illit                             39.888   256.335     170.786
school.prim                            1031.742  2730.092    1127.186
school.med                               90.359   223.341      76.260
school.high                              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
```

Group Activity: Same groups as before: what would covariate balance look like here? Does the balance seem good here?

Evaluating our Synthetic Control

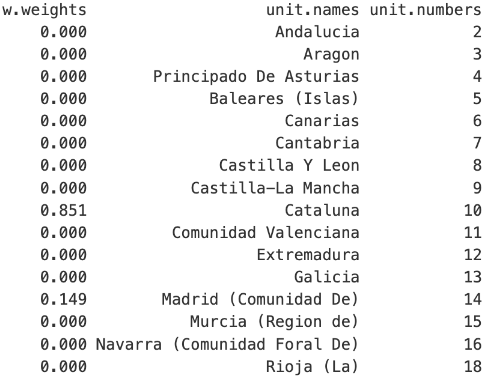
We let an optimization algorithm pick weights. Then, we can actually look at the weights!



```
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
```

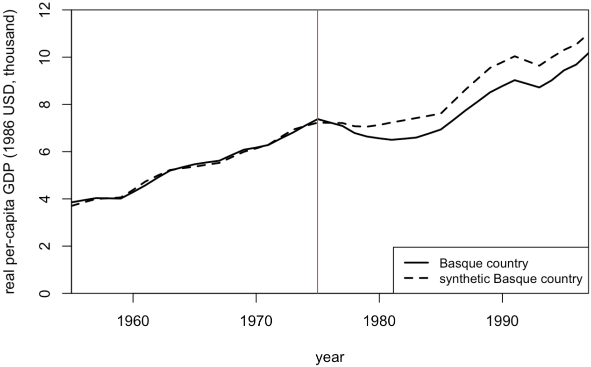
Group Activity: Same groups as before: what do you notice about the weights? how do we interpret this?

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?

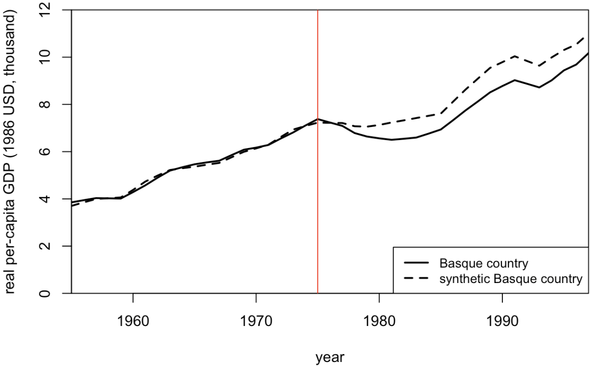
Is there a Causal Effect?

- 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?

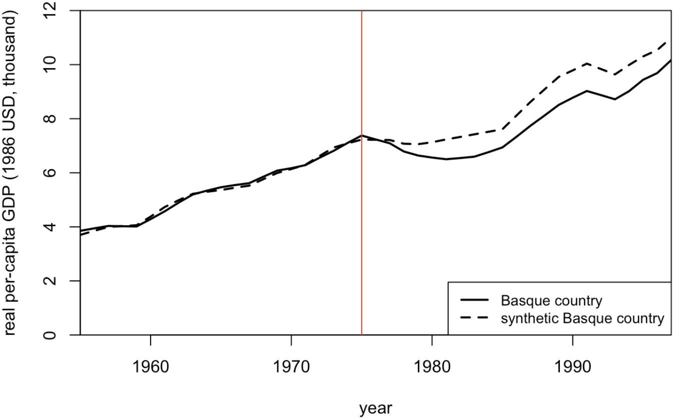
- Goal: estimate the causal effect of violent conflict on economic output
- How do we determine if there really is a causal effect?
- Compare economic output of the Basque Country region to our synthetic control unit after the terrorist attacks began



- 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?


real per-capita GDP (1986 USD, thousand)
year
Basque country
synthetic Basque country