CODERS' CORNER

CENTRE FOR THE STUDY OF AFRICAN ECONOMIES

SYNTHETIC CONTROL METHOD IN STATA

Determining the causal effect of a policy intervention is challenging: while you may be able to observe the outcome of interest with the new policy in place, you can't ever observe what would have happened had the policy not been implemented. In other words, you can't observe the counterfactual. The synthetic control method is one of many inferential methods that have been developed to deal with this challenge.

Conceptually, the synthetic control method is a combination of the difference-in-differences approach and matching methods. From these two techniques, the synthetic control method borrows the idea (1) to compare two units (one treated and one untreated) that show parallel trends in the outcome variable of interest pre-intervention and (2) to find an appropriate comparison unit by matching on observable characteristics that potentially have an effect the outcome variable.

Finding a single control unit that fulfills these criteria can be difficult in practice. Instead, a weighted average of several different control units might make for a more appropriate comparison. A key advantage of the synthetic control method is its use of data-driven procedures to construct such a comparison unit from a donor pool of untreated units which the researcher selects.

For a detailed discussion of the use of the synthetic control method for comparative case studies, take a look at the paper by Abadie, Diamond and Hainmueller (2010) on the effect of California's tobacco control program. The authors developed a package called *synth*, which lets you implement the synthetic control method fairly easily in Stata (or R, if you prefer).

To use the method, you'll need a balanced panel data set containing your treated unit and a number of potential control units that make up the donor pool. The data set should include pre-intervention and post-intervention observations on your outcome of interest and a set of explanatory variables, all of which have to be numerical variables.

First, install the package and declare your data as time series data:

ssc install synth
tsset panelvar timevar

The syntax for the synthetic control method is then simply:

synth depvar predictorvars , trunit(#) trperiod(#)

where *depvar* is your outcome variable of interest, *predictorvars* are the explanatory variables, *trunit* specifies the name of the treated unit (e.g. the country that implemented the new policy) and *trperiod* specifies the time of the treatment (e.g. the year in which the new policy was implemented). You may want to include lagged values of your dependent variable as a predictor variable to improve the pre-intervention fit.

Adding the option *figure* produces a line graph that allows you to visually inspect the differences between the treated unit and the synthetic control unit in terms of the dependent variable pre- and post-intervention.

After running *synth*, the weights assigned to individual units in the donor pool can be retrieved using the command *e(W_weights)*.

To further enhance the credibility of your analysis, you may want to do the following:

(i) Use a validation set approach: to avoid over-fitting the synthetic control unit, you may want to split your pre-intervention data into a training set and a validation set. Fit the synthetic control using just the training set and use the resulting weights to test the fit on the validation set. For details on cross-validation see Abadie, Diamond and Hainmueller (2015).

(ii) Run placebo tests: reassign the treatment to a number of different time periods and/or different units in the sample (simply by changing the *trperiod(#)* and *trunit(#)* inputs) and compare the resulting estimated effect sizes. If the effect size you observe in your original analysis is significantly larger than any of the placebo effects, there is a good chance you're onto something causal.

Lisa Martin, PRS in Economics, Keble College, Oxford
04 February 2020