LECTURE 10. SYNTHETIC CONTROL

Dr. Tom O'Grady

In this final lecture, we will examine a very new technique known as synthetic control. The main use of synthetic control is in studying a treatment that unfolds over time in one place (often a country) and nowhere else, particularly in small samples where there may be only a small number of other places to which we can potentially compare the treated unit. The key idea behind it is that we construct a *synthetic counterfactual* for the treated unit. It consists of a weighted combination of the potential comparison units, where the weighted combination is chosen so as to best approximate what would have happened to the treated unit in the absence of the treatment. It potentially has a lot of advantages over difference-in-differences and fixed effects, but requires a lot of data. We'll look at the motivations behind it, how to calculate the synthetic counterfactual, and how to conduct inference.

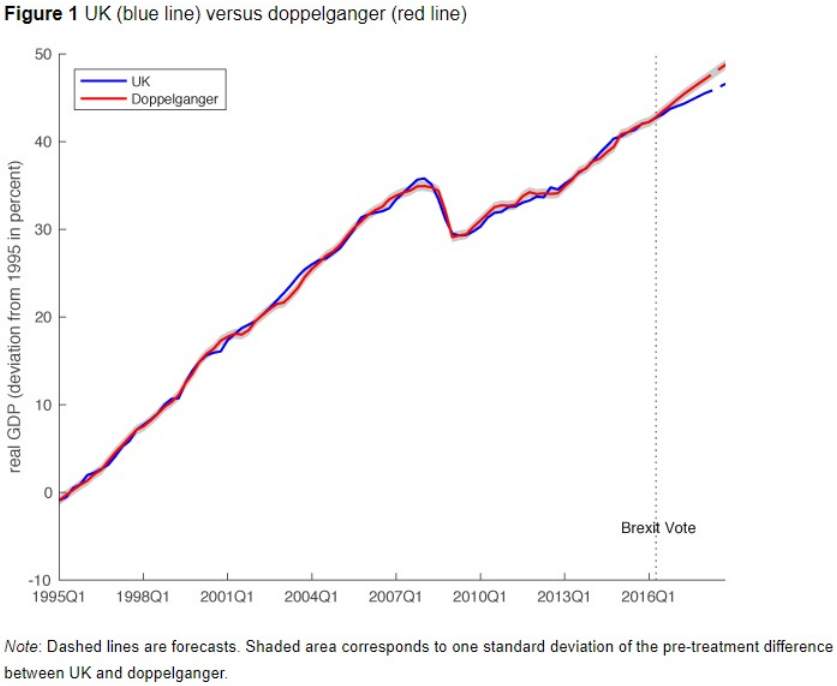
1. Motivation for Synthetic Control

Some key examples of the use of synthetic control including studying the economic effects of one-off events like regime change, German reunification, terrorist attacks and natural disasters. Figure 1 shows a very recent implementation of the technique from the Vox blog, asking what impact Brexit has had on economic growth.[1]. The blue line shows the UK's actual GDP growth since 1995, while the red line shows the GDP growth of the UK's synthetic control (which they call the 'doppelganger'), representing the authors' best guess of what would have happened to the UK in the absence of Brexit. Their synthetic control consists of a weighted combination of the economies of Canada, the USA, Japan and Hungary, selected by the synthetic control algorithm from a 'donor pool' of thirty OECD countries. Notice that prior to Brexit, the synthetic control provides an almost perfect approximation to British GDP growth. That provides some reassurance that it genuinely reflects what would have happened in the absence of Brexit, after it happened. Their central estimate is of a cumulative loss of around 1.3% of GDP, equivalent to £300m per week in output since the vote.

Why use synthetic control to answer a question like the impact of Brexit? One reason is that in a case like this, there may be limits to how useful regression analysis like difference-in-differences or fixed effects can be when we only have one treated country/state (etc.) and a few untreated countries/states. Another reason is that traditional qualitative comparison methods used for small samples also have big problems. We'll discuss each in turn.

[1]See http://voxeu.org/article/300-million-week-output-cost-brexit-vote

Figure 1: **The Economic Impact of Brexit**



1.1. Comparison to Difference-in-Differences and Fixed Effects

Difference-in-differences with two periods and two groups relies on the availability of disaggregated data within the groups to conduct testing. In the minimum wages example, our two groups were New Jersey (treatment) and Pennsylvania (control), but we had data on a large number of individual restaurants within them. Synthetic control can be used instead when we only have data at the level of the groups themselves, which might leave only a few observations over time and space (e.g. 50 states at most), potentially too few to conduct any regression analysis at all. However, synthetic control can also be useful even when disaggregated data is available, because it can help us to choose our control group. Why should Pennsylvania be the 'best' control comparison for New Jersey? Couldn't other states also serve as counterfactuals? Synthetic control offers a data-driven method for choosing the best comparison state(s).

The same question can be asked in reverse about the opposite extreme, fixed effects, where we would simply include all 50 states in the regression analysis. This implicitly assumes that *all* other states can help construct a counterfactual for the treated state - but how do we know that? As Abadie, Diamond and Hainmueller point out in the paper on German reunification on this week's reading list, regression implicitly places a weight on each country in the analysis which may bear little relation to how good each country is as a counterfactual. Why not choose the best set of comparison countries first, instead of letting the regression choose for us?

A further point of departure is that synthetic control relies on weaker assumptions, as we'll see below. Both difference-in-differences and fixed effects control for all time-invariant confounders, but not time-varying confounders. Time-varying confounders must be assumed not to exist: the parallel trends assumption. No such assumption is needed in synthetic control. If done properly, synthetic control creates a synthetic counterfactual that is 'balanced' with the

treatment case both in terms of time-constant *and* time-varying confounders.

1.2. Comparison to Qualitative Methods of Comparison

Another, more traditional approach to studying one-off events is to use the method of *small-n qualitative comparison*, which until quite recently was very widely used for causal inference in political science and sociology. This sort of approach goes back to the nineteenth century and the work of John Stuart Mill, who advocated using the **Method of Difference:**

> "If an instance in which the phenomena under investigation occurs and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

In other words, he advocated comparing a treatment case to another control case that is as similar as possible in every respect except the treatment, forming a counterfactual. Any subsequent difference between the two cases is then attributed to the treatment.

In reality, the chances that this method leads to valid causal inferences are very slim. All of the problems we've seen in this course of trying to find a counterfactual in the absence of randomisation are only magnified when we choose a single comparison case. It is very unlikely that the treatment and control cases will have the same potential outcomes. Countries, for example, are very unique. There is no single country that is like the UK in every way except for not going through Brexit. This is compounded by the fact that in many actual applications, comparison cases are often chosen in an arbitrary or haphazard way, with the choice of comparison country over other potential controls justified by researchers' subjective assessments of affinity rather than explicit balance tests or other metrics. In the worst-case scenario, it is perfectly possible to cherry-pick a comparison case with full knowledge of how that would affect the outcome of the study, as Abadie et al (2010) point out on the second page. An important part of honesty and transparency in non-experimental research is to make decisions about research *design* without knowing how that would affect the *outcome*, allowing a null result to emerge if a null result is in fact correct and true.

A related issue also pointed out by Abadie et al (2010) is that small-n qualitative comparison offers no room for uncertainty. There can be no standard error with only two observations, yet there must be a great deal of uncertainty about whether or not the comparison case is really a valid counterfactual, and hence about the size or direction of any treatment effect that emerges from the analysis.

1.3. The Advantages of Synthetic Control

In contrast to other methods, synthetic control offers:

- A transparent, data-driven, approach to choosing the optimal counterfactual that is not based on hunches or post-hoc rationalisation, and where the outcome of the study will not be known before the synthetic control is calculated
- An approach to uncertainty that reflects uncertainty about the suitability of the counterfactual, even with only a small number of observations.

- Balance is achieved on both time-constant and time-varying confounders automatically. We don't need to test for the presence of parallel trends
- The possibility of combining quantitative and qualitative analysis in a transparent way. Qualitative analysis has many advantages that quantitative analysis cannot match, including the ability to delve deeply into the history of cases to examine causal mechanisms at work. In-depth case-studies are often a good complement to quantitative analysis. Synthetic control allows us to estimate causal effects quantitatively, but also offers a data-driven way to select particular cases to delve into more deeply (e.g. the most similar or most different cases).[2]

2. Estimating the Synthetic Control

How should we go about choosing the optimal combination of comparison cases to form a synthetic control? One way to understand the method is that it's a bit like combining difference-in-differences with matching, where we match our treated unit to the potential control units using time-invariant characteristics and lagged outcomes.

2.1. Defining the Weights and the Treatment Effects

Concretely, suppose we have:

- Units (countries, states etc.) $i = 1 \dots J$
- Time periods $1 \dots T$
- A treatment that affects unit 1 from time T_0 onwards, where T_0 is between 1 and T
- Units $2 \dots J$ that could potentially be counterfactuals for unit 1
- Observed outcomes Y_{it}
- Time-invariant characteristics of the units Z_i, where Z_i is a vector
- Potential outcome Y_{1t}^N that would be observed for unit 1 at time t without the treatment
- Potential outcome Y_{1t}^I that would be observed for unit 1 at time t with the treatment

The treatment effects of interest are therefore given by:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N, \quad for\ all\ t > T_0$$

and after T_0, the actual outcome for unit 1 is:

$$Y_{1t} = Y_{1t}^N + \tau_{1t}$$

where we observe Y_{1t}^I but not Y_{1t}^N after T_0. As always in this course, the key task is to fill in these missing potential outcomes using information from other comparable units, in order to form a counterfactual. Suppose there exists some vector $W^* = (w_2^* \dots w_J^*)$ where $w_i^* \geq 0$ for all

[2]See Evan Lieberman, "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99 (3): 435-452 for a great approach to combining quantitative and qualitative analysis.

$i > 1$ such that:

$$\sum_{i=2}^{J} w_i^* Z_i = Z_1$$

and:

$$\sum_{i=2}^{J} w_i^* Y_{i1} = Y_{11}, \sum_{i=2}^{J} w_i^* Y_{i2} = Y_{12}, \cdots \sum_{i=2}^{J} w_i^* Y_{iT_0} = Y_{1T_0}$$

and:

$$\sum_{i=2}^{J} w_i^* = 1$$

In words, suppose that there exists an optimal set of weights w_i^* specifying a weight between 0 and 1 for every non-treated unit, where w_i^* must sum to 1. A weighted combination of the time-invariant characteristics of all control units should be equal to the actual time-invariant characteristics of the treated unit, and in every period up to T_0 when the treatment occurs, a weighted combination of the outcomes of all control units should be equal to the actual outcome of the treated unit. These weights allow us to recover the exact outcomes and characteristics of the treated units using only the non-treated units. It is important to choose both the variables in Z and the set of control units carefully:

- The variables in Z should be chosen so as to reflect as many important determinants of the outcome as possible (e.g. determinants of economic growth). Note that in many cases we will have time-variant information available, such as on investment in each pre-treatment year. In such a case, the synthtic control algorithm simply takes an average of the variable over the pre-treatment period, converting it into a time-constant variable.

- The control units should be chosen from amongst units whose outcome is determined in the same way as the treated unit. In practice this means picking from a similar pool. It is reasonable to assume that other rich OECD countries have their growth determined by the same set of variables as the UK, but that assumption would be more tenuous for African countries. In addition, the control countries should not experience the same treatment as the treated unit after T_0, or they cannot provide a valid counterfactual. For instance, Abadie, Diamond and Hainmueller (2010) rule out other US states that also enacted tobacco control measures. The control units must really be control units, not other treated units.

If such a W^* exists, then Abadie, Diamond and Hainmueller (2010) show that if the true model of the outcome is also comparable for all units in the sample, an approximately unbiased estimator of the treatment effects is given by:

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{i=2}^{J} w_i^* Y_{it}, \quad for\ all\ t > T_0$$

That is, the treatment effect in time t is given by the difference between the actual outcome for the treated unit and a counterfactual outcome constructed by weighting the outcomes of

all other units in the same period. The second term is our **synthetic control**, providing a counterfactual for the treated unit.

2.2. Estimating the Weights

In reality, it is unlikely that any W^* exists that perfectly fulfils the above criteria. Instead, we must choose an optimal W^* that minimises the distances between the synthetic Z_1 given by $\sum_{i=2}^{J} w_i^* Z_i$ and the actual Z_1, between the synthetic Y_{11} given by $\sum_{i=2}^{J} w_i^* Y_{i1}$ and the actual Y_{11}, and so on up to Y_{1T_0}.

Let's say that the $(M \text{ x } 1)$ vector $X_1 = (Z_1, Y_{11}, Y_{12} \dots Y_{1T_0})'$ for the treated unit contains its characteristics and pre-treatment outcomes (M variables in total) and the $(M \text{ x } J-1)$ matrix X_0 contains the same information in each column for all the untreated units. Then, we can choose W^* as the W that minimises:

$$\sum_{m=1}^{M} v_m (X_{1m} - X_{0m} W)^2$$

where v_m is a weight reflecting the importance of variable m in minimizing the overall distance between the vector of the treated unit's actual characteristics (X_1) and those of the synthetic control ($X_0 W$). In the standard `Synth` software in R, the v values are chosen optimally so as to generate the closest fit between the treated outcome and synthetic control before the treatment. Technically, the weights v are chosen by R's `Synth` package so as to minimise the pre-treatment **Root Mean Squared Prediction Error**:

$$RMSPE = \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{i=2}^{J} w_i^* Y_{it} \right)^2 \right]^{\frac{1}{2}}$$

Overall, in a spirit that is similar to matching, the algorithm tries to find the optimal weighted counterfactual (the synthetic control) that minimises the distance, in terms of time-invariant characteristics and pre-treatment outcomes, between the treated unit and the synthetic control.

2.3. Comparison to other Techniques

The algorithm produces a counterfactual (the synthetic control) that should be near-identical to the treated unit both in terms of time-constant and time-varying confounders. This contrasts with fixed effects and difference-in-differences, where balance on time-constant confounders is guaranteed but not on time-varying confounders, and we must collect circumstantial evidence in the form of parallel trends in the hope that there are no time-variable confounders. Why does synthetic control do this? The 2010 Abadie et al paper contains a formal mathematical proof which you can review if interested. Intuitively, the reason is that by finding a counterfactual whose outcomes are virtually the same before the treatment *and* whose time-constant characteristics (Z) are virtually the same, then implicitly we have also achieved balance on whatever makes up the difference between the outcome and Z in each period. This difference can only consist of time-varying confounders and idiosyncratic errors. Another way of putting this is that only units that have very similar time-constant *and* time-varying characteristics could possibly produce the same trajectory of outcomes over time.

2.4. Potential Problems or Drawbacks

Synthetic control seems to have many advantages, but problems can arise:

1. A lot of data is preferable from before the treatment. This ensures that a good 'match' is achieved on pre-intervention outcomes. But in some cases such data may not be available. That doesn't mean that we can't technically estimate a synthetic control, but it should lessen our confidence in the validity of the findings.

2. The outcomes of potential control units and the treated unit need to be generated by the same process. For example, in the Brexit case, we need to assume that the determinants of economic growth are the same in all countries. Otherwise, there is no reason to believe that a weighted combination of them provides a reasonable counterfactual. This means that really unique one-off events in unusual cases might be difficult to analyze. Otherwise, it is best to restrict the 'donor pool' of potential control units to a similar subset of units to begin with.

3. We need to assume that there are no shocks to the control units after the treatment date. Otherwise, their weighted outcomes cannot provide a counterfactual for the treated unit. In the case of Brexit, we must ensure that the control units did not receive some shock to economic output (like war, a natural disaster, etc.) that the UK would not have experienced in the absence of the treatment.

4. Spillover effects after the treatment might be a major problem. For instance, in the case of German reunification, the authors' model relies heavily on Austria in the synthetic control. But Austria is a major trading partner of Germany. If Austria's growth suffered after reunification as a result of low German growth, then a comparison to Austria is likely to overstate the treatment effect due to the counterfactual being too low.

3. Placebo Tests and Inference

It is not immediately obvious how to carry out inference in synthetic control analysis because it is not obvious why there should be uncertainty about the results, as Abadie, Diamond and Hainmueller point out in both papers on the reading list. Uncertainty in typical statistical studies relates to sampling error. We observe only some fraction of the population, and there is uncertainty over whether the results we obtained (e.g. for an estimated regression function) could have arisen due to chance alone, from a particularly unusual sample. But in synthetic control analysis, we typically do observe the whole population of interest, and we have aggregate data for each unit.[3] Note that the arguments here could also be applied to fixed effects regressions that cover the whole population of interest. In those cases, an appeal is often made to the idea of a 'super-population' of possible outcomes had history turned out differently, of which the observed data is only one actual outcome. Many statisticians find this argument unconvincing, including the authors of the papers on this week's reading list. As they point out, even when we observe the whole population of interest, uncertainty surely remains about whether or not the synthetic control provides a reasonable counterfactual.

[3]Measurement error could still introduce uncertainty, but that is never accounted for in traditional standard errors

They advocate the use of two types of placebo tests. We'll focus on one of them here, which they call "in-space" placebos in the 2015 paper. The idea behind these placebo tests is to estimate a series of other synthetic control analyses, where the treatment variable is swapped out for each of the control units in turn. Loosely speaking, under a null hypothesis of no genuine treatment effect, we would expect a large estimated actual treatment effect to have arisen by chance alone. Then, there is no reason to believe that the estimated effect should be distinguishable from the effect that is produced by artificially re-assigning the treatment to any one of the controls. If our actual estimated treatment effect falls within the range of the placebo effects, then it is likely to have arisen by chance alone. If the estimated treatment effect is large relative to almost all of the placebo effects, then we should have a lot of confidence that it didn't arise due to chance alone.

More concretely, as in randomisation inference, we can assume under the null hypothesis that we observe all the potential outcomes, since the treatment effect is assumed to be zero in all cases. The null distribution then tells us what a 'typical' estimated effect looks like in a case where in reality there is no effect. Then, we can produce a p-value by evaluating how many of the placebo treatment effects fall above or below the actual treatment effects.

4. Steps in Applying Synthetic Control in practice

To carry out synthetic control analysis in practice, the work-flow looks roughly like this:

1. **Decide on a set of control units (the donor pool)** by choosing control units whose outcomes are determined in a similar way to those of the treated unit, and which are unlikely to be subject to post-treatment shocks or spillover effects

2. **Decide on a set of pre-treatment characteristics** Z that are important in determining the outcome. Note that in practice these can also be determined iteratively, by attempting to add further variables to the analysis and determining whether or not they improve the pre-treatment fit between the treated unit and the synthetic control

3. **Decide on the start and end date of the analysis.** Ideally, a long pre-treatment period would be used.

4. **Fit the synthetic control model and assess goodness of fit** between the treated unit and the synthetic control before the treatment. Visual inspection is usually sufficient. Non-parallel trends in the period just before the treatment occurs would be particularly concerning. If the fit is poor, you could try (i) adding more variables to Z, (ii) adding more pre-treatment periods or (iii) adding more control units (but see (1)). In some cases it may not be possible to find a good counterfactual, in which case synthetic control analysis is not possible.

5. **Calculate the treatment effects** cumulatively or per-period.

6. **Conduct inference using placebo tests**