DiD II & Synthetic Control Method

Causal Inference, DPIR, University of Oxford

Tanisha Mohapatra & Fernando Sanchez Monforte

HT 2025

Introduction

Before starting this seminar

1. Create a folder called `lab6`.
2. Download the data and the empty RMarkdown file (available on Canvas).
3. Save both in our `lab6` folder.
4. Open the RMarkdown file.
5. Set your working directory using the `setwd()` function or by clicking on "More". For example, this may look like this: `setwd(\"\~/Desktop/Causal Inference/2024/lab6\")`.

Seminar Overview

In this seminar, we will cover the following topics:

1. **Generalized Difference-in-Differences (Recap)**:
   - Introduction to the logic and application of Generalised DiD designs.
   - Discussion on the role of unit fixed effects in addressing unobserved heterogeneity.
2. **Synthetic Control Method**:
   - Overview of the synthetic control method for causal analysis.
   - Case study on its application to evaluate policy impacts, focusing on the Economic and Monetary Union's effect on Spain's Current Account Balance.

Generalised Diff-in-Diff (Recap)

Basic Logic of Difference-in-Differences (Diff-in-Diff) Designs

Recall that the difference-in-differences (Diff-in-Diff) approach is a quasi-experimental design used in observational studies to estimate the causal effect of a treatment or intervention. It compares the change in outcomes over time between a group that is exposed to the treatment (treatment group) and a group that is not (control group). By analysing the differences in changes between these groups, we can infer the treatment effect, assuming that the trends in outcomes for both groups would have been parallel in the absence of the treatment (parallel trends assumption). This method helps to control for confounding variables that are constant over time and affect both groups similarly, thereby isolating the impact of the treatment.

Inclusion of Unit Fixed Effects

Incorporating unit fixed effects into Diff-in-Diff designs is *essential* when there are unobserved variables that may affect the outcome and vary across units (e.g., individuals, firms, countries) but not over time. Unit fixed effects control for these time-invariant characteristics, ensuring that the comparison between treatment and control groups is not biased by such unobserved heterogeneity. By including unit fixed effects, researchers can more reliably attribute differences in outcomes to the treatment itself rather than to underlying differences between the units.

Understanding Fixed Effects

Fixed effects models are used to control for unobserved heterogeneity when this heterogeneity is constant over time but varies across units being studied. Essentially, fixed effects models remove the effect of time-invariant characteristics so that the net effect of the variables of interest can be assessed more accurately. This is done by allowing each unit (e.g., individual, company, country) to have its own intercept in the regression model, effectively controlling for all unobservable characteristics that do not change over time for each unit. The primary benefit of fixed effects models is their ability to produce unbiased estimates of causal relationships by focusing solely on the variation within units over time. This method is particularly useful in longitudinal (panel) data analysis, where the goal is to examine how changes in predictors lead to changes in outcomes within the same units over multiple time periods. By controlling for all time-invariant differences between units, fixed effects models enhance the credibility of causal inferences in observational studies.

Clustered standard errors

One important assumption when modelling panel data concerns the distribution of the coefficients' *standard errors*. Recall that standard errors in linear regression quantify the *average variability of an estimated coefficient* when drawing (hypothetical) repeated samples from the population. Or, in other words, standard errors capture how much "off", on average, an estimator (in our case, a regression coefficient) is from its "true" population parameter value (meaning, the coefficient's "true" population value that we cannot directly observe) over a repeated data-generating process. More colloquially put, standard errors provide us with a measure of *uncertainty* for our predictions. As social scientists, we are usually really interested in knowing whether an effect is likely to exist *in the population* based on our sample or whether the effect that we observe in our sample is so small (depending on sample size) that we could have also observed it purely due to random sampling error.

Specifically, when modelling panel data, we assume that observations *within* each group (meaning within units or time-points) are *independently and identically distributed* (for short, *i.i.d.*). Although this is, in the end, an empirical question, it is nevertheless easy to come up with scenarios when this assumption is violated in panel data. For example, imagine that you want to study the causal effect of economic wealth (measured as GDP per capita) on democracy (e.g. measured by the Freedom House index) on the country-level based on multiple observations per country over time. Assuming that observations *within* each group are *independently and identically distributed* means that the effect of economic wealth on democracy for each country and at each time point should not vary drastically. As you may or may not see, this assumption seems quite far-fetched. For example, one country, let's say South Korea, may have a much better statistics agency than other one, let's say Greece, resulting in South Korea's GDP per capita measure being considerably more accurate and stable than Greece's. As a result, South Korea's measure of economic wealth contains much less measurement error and bias than Greece's and is therefore considerably more predictive of its democracy score than Greece's.

We can *cluster our standard errors* to account for the problem of standard errors potentially being correlated within the same units or for the same time-points. Specifically, by using *clustered standard errors*, we deal with *heteroskedasticity* and *autocorrelation* within units.

Note that fixing the problem of standard errors failing to be independently and identically distributed within groups does not automatically fix the problem of unit- or time-specific confounders potentially biasing our regression coefficients or vice versa! Both problems may occur at the same time, which means that we would have to take care of both of them separately. The fact that Greece's statistics agency produces a more variable estimate of the country's GDP per capita measure than South Korea's statistics agency does not imply that there are no other characteristics that distinguish Greece from South Korea which may affect the relationship of economic wealth on democracy. Conversely, the fact that South Korea and Greece have different cultures which may bias the relationship between economic wealth on democracy does not automatically imply that South Korea's statistics agency necessarily produces more stable GDP measures than Greece's statistics agency.

Let's now extend our analysis from last week by including all pre-treatment periods in our analysis. The easiest way to do so is running a two-way fixed effects regression.

Exercise 1: Estimate the difference-in-differences using a two-way fixed effects regression with all time periods and `treatment` as independent variable.

> Hint: Use the `feols` function from the fixest package, like so: feols(dv~treat|unit+time, data = data, cluster = "cluster_var")

```
#  Your amazing code starts here!
```

Synthetic control

The Effect of Economic and Monetary Union on Spain's Current Account Balance – Hope (2016)

In early 2008, about a decade after the Euro was first introduced, the European Commission published a document looking back at the currency's short history and concluded that the European Economic and Monetary Union was a "resounding success". By the end of 2009 Europe was at the beginning of a multiyear sovereign debt crisis, in which several countries – including a number of Eurozone members – were unable to repay or refinance their government debt or to bail out over-indebted banks. Although the causes of the Eurocrisis were many and varied, one aspect of the pre-crisis era that became particularly damaging after 2008 were the large and persistent current account deficits of many member states. Current account imbalances – which capture the inflows and outflows of both goods and services and investment income – were a marked feature of the post-EMU, pre-crisis era, with many countries in the Eurozone running persistent current account deficits (indicating that they were net borrowers from the rest of the world). Large current account deficits make economies more vulnerable to external economic shocks because of the risk of a sudden stop in capital used to finance government deficits.

David Hope investigates the extent to which the introduction of the Economic and Monetary Union in 1999 was responsible for the current account imbalances that emerged in the 2000s. Using the sythetic control method, Hope evaluates the causal effect of EMU on current account balances in 11 countries between 1980 and 2010. In this exercise, we will focus on just one country – Spain – and evaluate the causal effect of joining EMU on the Spanish current account balance. Of the J countries in the sample, therefore, $j = 1$ is Spain, and $j = 2, ..., 16$ will represent the "donor" pool of countries. In this case, the donor pool consists of 15 OECD countries that did not join the EMU: Australia, Canada, Chile, Denmark, Hungary, Israel, Japan, Korea, Mexico, New Zealand, Poland, Sweden, Turkey, the UK and the US.

The `hope_emu.csv` file contains data on these 16 countries across the years 1980 to 2010. The data includes the following variables:

Variable Name	Description
`period`	The year of observation
`country_ID`	The country of observation
`country_no`	A numeric country identifier
`CAB`	Current account balance
`GDPPC_PPP`	GDP per capita, purchasing power adjusted
`invest`	Total investment as a % of GDP
`gov_debt`	Government debt as a % of GDP
`openness`	Trade openness
`demand`	Domestic demand growth
`x_price`	Price level of exports
`gov_deficit`	Government primary balance as a % of GDP
`credit`	Domestic credit to the private sector as a % of GDP
`GDP_gr`	GDP growth %

Use the `read.csv` function to load the downloaded data into R. For this kab, we will need the qualitative variables to be stored as character variables, rather than the factor encoding that R uses by default. For this reason, we will set the `stringsAsFactors` argument in the `read.csv` function to be false.

```
emu <- read.csv("hope_emu.csv", stringsAsFactors = FALSE)
```

Raw trends

```
# Your amazing code starts here!
```

Exercise 1: Plot the trajectory of the Spanish current account balance (`CAB`) over time (in red), compared to three other countries of your choice. Plot an additional dashed vertical line in 1999 to mark the introduction of the EMU. Would you be happy using any of them on their own as the control group?

Preparing the synthetic control

Exercise 2: The `Synth` package takes data in a somewhat unusual format. The main function we will use to get our data.frame into the correct shape is the `dataprep()` function. Use this to prepare the `emu` data. Try on your own first, and then look at the solution below.

> *Hint:* Look at the help file for this function using `?dataprep`. You will see that this function requires us to correctly specify a number of different arguments. The main arguments you will need to use are summarised in the table below:

Argument	Description
`foo`	This is where we put the data.frame that we want to use for the analysis
`predictors`	This argument expects a vector of names for the covariates we would like to use to estimate the model. You will need to use the `c()` function, and enter in all the variable names that you will be using.
`dependent`	The name of the dependent variable in the analysis (here, `"CAB"`)

Argument	Description
`unit.variable`	The name of the variable that identifies each unit (must be numeric)
`unit.names.variable`	The name of the variable that contains the name for each unit (here, `"country_ID"`)
`time.variable`	The name of the variable that identifies each time period (must be numeric)
`treatment.identifier`	The identifying number of the treatment unit (must correspond to the value for the treated unit in `unit.variable`)
`controls.identifier`	The identifying numbers of the control units (must correspond to the values for the control units in `unit.variable`)
`time.predictors.prior`	A vector indicating the time periods before the treatment
`time.optimize.ssr`	Another vector indicating the time periods before the treatment
`time.plot`	A vector indicating the time periods before and after the treatment

dataprep_out <- dataprep(foo = , predictors = , dependent = , unit.variable = , time.variable = , treatment.identifier = , # controls.identifier = , time.predictors.prior = , time.optimize.ssr = , unit.names.variable = , time.plot =)

Estimating the synthetic control

Exercise 3: Fortunately, though getting the data in the prep function correctly can be a pain, estimating the synthetic control is very straightforward. Use the `synth()` function on the `dataprep_out` object that you just created, remembering to assign the output to a new object.

> *Note:* It can take a few minutes for this function to run, so be patient!

```
# Your amazing code starts here!
```

`R` prints some details when it finishes the estimation of the synthetic control, but these are a little difficult to interpret directly. Instead, we will move on to interpreting the types of plots that we saw in the lecture.

Plotting the results

Exercise 4: Use synth's `path.plot()` and `gaps.plot()` functions to produce plots which compare Spain's actual current account balance trend to that of the synthetic Spain you have just created. Interpret these plots. What do they suggest about the effect of the introduction of EMU on the Spanish current account balance?

> *Hint:* These function takes two main arguments, and then some additional styling arguments to make the plot look nice. Look at the help file for more details.

Argument	Description
`synth.res`	This is where we put the saved output of the `synth()` function (i.e. the estimated synthetic control object)
`dataprep.res`	This is where we put the saved output of the `dataprep()` function (i.e. the data we used to estimate the synthetic control).
`tr.intake`	A number to indicate the time of the treatment intake (here, 1999)
`Xlab`	The name of the variable on the x-axis (here, `"Time"`)
`Ylab`	The name of the variable on the x-axis (here, `"Current account balance"`)

Argument	Description
`Legend`	Optional text for the legend of the plot.
`Ylim`	The range of the y-axis (here, `c(-10,5)`)

```
# Your amazing code starts here!
```

Interpreting the synthetic control unit

Exercise 5: A crucial strength of the synthetic control approach is that it allows us to be very transparent about the comparisons we are making when making causal inferences. In particular, we know that the synthetic Spain that we created in exercise 2 is a weighted average of the 15 OECD non-EMU countries in our data.

a. What are the top five countries contributing to synthetic Spain?

> *Hint:* Look in the help file for `?synth`, and read the "Value" section of that page. The value section will tell you all of the things that are returned by a function. You can access them by using the dollar sign operator that we have used in the past to extract variables from a data.frame. *Another Hint:* Fortunately, there is an easier way to extract the country weights as well as a) information on the weights assigned to each of the predictor variables in the model, and b) the balance on each predictor variable across the treated country and the synthetic country. Look at the help file for the `synth.tab()` function and apply that function to the output of the `synth()` and `dataprep()` functions from the questions above.

```
# Your amazing code starts here!
```

```
#  Your amazing code starts here!
```

b. Which variables contribute the most to the synthetic control? Is the synthetic control unit closer to the treated unit in terms of the covariates than the sample mean?

Estimating a placebo synthetic control treatment effect

Exercise 6: One way to check the validity of the synthetic control is to estimate "placebo" effects – i.e. effects for units that were not exposed to the treatment. In this question we will replicate the analysis above for Australia, which did not join EMU in 1999.

a. In constructing synthetic Australia, we must exclude Spain – the *actual* treatment unit – from the analysis. Before you repeat the steps above for Australia, create a new data.frame that doesn't include the Spanish observations.

> *Hint:* Here you will want to select all rows of the `emu` data for which the `country_ID` variable is *not* equal to `"ESP"`.

```
# Your amazing code starts here!
```

```
# Prepare the data for Australia


# Estimate the new synthetic control

# Plot the results
```

b. Now repeat the steps above to estimate the synthetic control for Australia.

c. What does the estimated treatment effect for Australia tell you about the validity of the design for estimating the treatment effect of the EMU on the Spanish current account balance?

```
# Define function for calculating the RMSE
rmse <- function(x,y){
  sqrt(mean((x - y)^2))
}

# Define vector for pre/post-intervention subsetting

## Spain
# Extract the weights for synthetic spain

# Calculate the outcome for synthetic spain using matrix multiplication

# Extract the true outcome for spain

# Calculate the RMSE for the pre-intervention period for spain

# Calculate the RMSE for the post-intervention period for spain



## Australia
# Extract the weights for synthetic Australia

# Calculate the outcome for synthetic Australia using matrix multiplication

# Extract the true outcome for Australia

# Calculate the RMSE for the pre-intervention period for Australia

# Calculate the RMSE for the post-intervention period for Australia
```

d. Compare the treatment effects from the Australian synthetic control analysis and the Spanish synthetic control analysis in terms of the pre- and post-treatment root mean square error values.