Week 5: Differences in Differences

Tanisha Mohapatra & Fernando Sanchez Monforte

HT 2025

Introduction

The intuition of the DiD strategy is to combine two simpler approaches. The first difference (before and after) *eliminates unit-specific fixed effects.* Then, the second difference *eliminates time fixed effects.* With this approach, we can, under some assumptions, obtain an unbiased estimate of a (policy) intervention.

We learned that we can break down the difference between treated and untreated units in post-treatment as the Average Treatment Effect Amongst the Treated (ATT), different time trends and selection bias. However, we can impose additional estimation assumptions to retrieve a credible estimate of this treatment effect. The key assumption in diff-in-diff studies is the so-called **Parallel Trends** or **Common Trends** assumption. This assumption states that in the absence of the treatment/policy, we should expect to see the treated and untreated units following similar trends over time. Unfortunately, we cannot comprehensively test whether this assumption holds, but we can at least conduct some analyses that provide indicative, although not sufficient, evidence for the parallel trends assumption to hold.

We also learned that we can calculate the ATT in different ways. Specifically, we learned that we can manually calculate the difference-in-difference estimator.

- Group-period interactions
- Unit and time dummies and a treatment indicator

We rely on the parallel trends assumption by assuming that time-trends are the same for both treated and untreated units.

Finally, we discussed inference and, in particular, standard errors. Specifically, we use panel data, meaning repeated observations over time, to estimate diff-in-differences models. Panel data likely exhibits serially correlated regressors and residuals. As a result, it's important to adjusted our coefficients' standard errors, e.g. by using *clustered standard errors.*

Lab Session Overview

In this seminar, we will cover the following topics:

1. “Manually” calculate the difference-in-differences estimator
2. Obtain the difference-in-differences estimator using the `lm()` function
3. Check for parallel trends before the treatment occurred
4. Use fixed effects in difference-in-differences estimations using both the `lm_robust()` and `plm()` function.
5. Conduct a placebo test

Before proeceeding further:

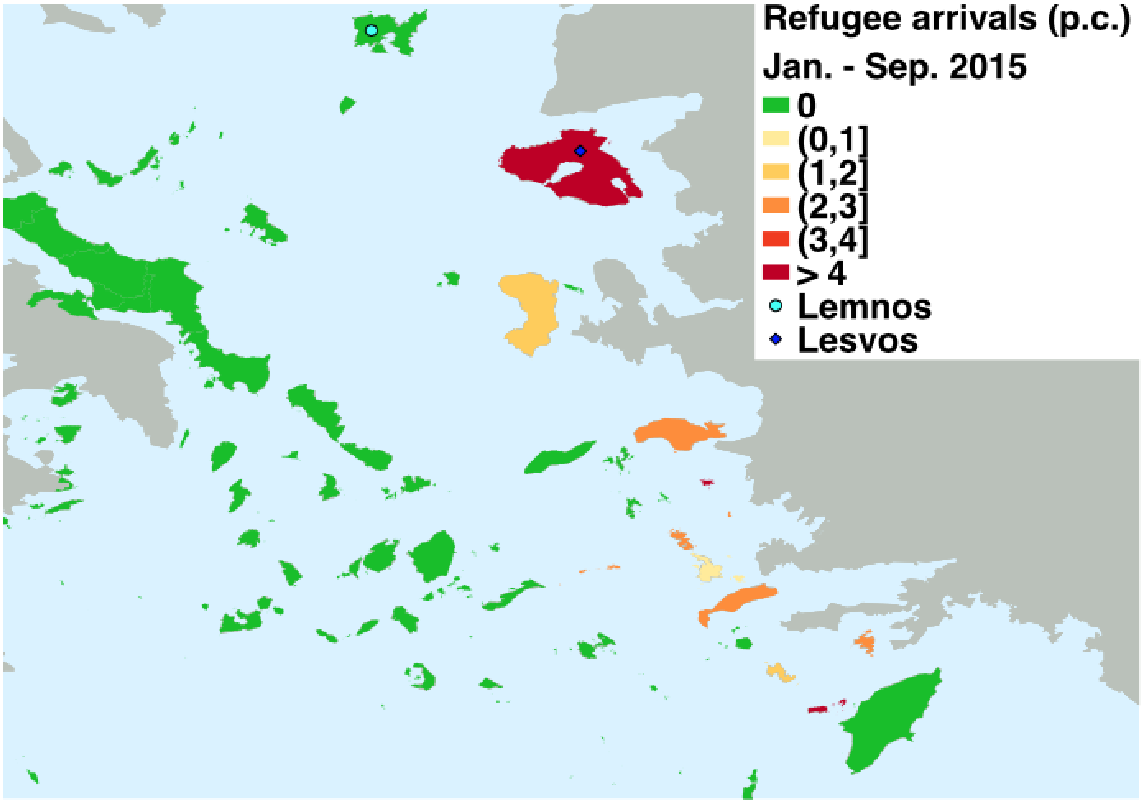
1. Create a folder called `Lab5`.
2. Download the data and the empty RMarkdown file from Canvas
3. Save both in our `Lab5` folder.
4. Open the RMarkdown file.
5. Set your working directory using the `setwd()` function or by clicking on "More".

Refugee exposure and support for the extreme right

Dinas et al. (2019)) studied the following question: *Did the influx of refugees in Greece increase support for the right-wing Golden Dawn party in 2015?*

The authors exploit that the Aegean islands close to the Turkish border experienced sudden and drastic increases in the number of Syrian refugees while other islands slightly farther away—but with otherwise similar institutional and socio-economic characteristics—did not.

The below figure shows exposure to refugees across the Aegean islands:



Estimating DiD

Group-period interactions: Here the treatment variable is equal to 1 for all the years since the unit received the treatment. Then, our coefficient of interest is captured by the interaction between the treatment and the time variable. For the example of the Golden Dawn, we can write the respective diff-in-diffs regression equation (including the interaction term's coefficient that captures the causal effect) as follows:

$$gdper_{mt} = \beta_0 + \beta_1 eventr_m + \beta_2 post_t + \beta_3 evertr_m \times post_t + u_{mt}$$

	Post = 0	**Post = 1**
Treat = 0	$\beta_0 + u_{mt}$	$\beta_0 + \beta_2 + u_{mt}$
Treat = 1	$\beta_0 + \beta_1 + u_m t$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + u_{mt}$

Specifically, we can obtain our estimate by calculating the difference of the outcome variable for both treated and untreated units and then subtract them from each other:

$$((\beta_0 + \beta_1 + \beta_2 + \beta_3 + u_{mt}) - (\beta_0 + \beta_1 + u_{mt})) - ((\beta_0 + \beta_2 + u_{mt}) - (\beta_0 + u_{mt}))$$

$$= (\beta_2 + \beta_3) - (\beta_2)$$

$$= \beta_3$$

Unit and time dummies and a treatment indicator: Alternatively, we can represent this estimation strategy for the Golden Dawn example by the following regression model:

$$gdper_{mt} = \beta_1 treatment_{mt} + \alpha_m \text{unit dummy} + \gamma_t \text{time dummy} + u_{mt}$$

For treated before treated $Treatment = 0$:

$$gdper_{mt} = \beta_1 \times 0 + \alpha_m \times 1 + \gamma_t \times +u_{mt}$$

$$gdper_{mt} = \alpha_m + \gamma_t + u_{mt}$$

For treated after treated $Treatment = 1$:

$$gdper_{mt} = \beta_1 + \alpha_m + \gamma_t + u_{mt}$$

Then, we can take the difference before and after:

$$(\beta_1 + \alpha_m + \gamma_t + u_{mt}) - (\alpha_m + \gamma_t + u_{mt})$$

$$= \beta_1$$

Dinas et al. (2019)

Context

- The refugee crisis started in the spring of 2015.
- Greece held an *election in September 2015*, right after the *first wave of refugee* arrivals. The variable `post` is coded as 1 for observations from this election and the variable `year`, somewhat unintuitively, as `2016`.
- The *previous election* had taken place only eight months prior in *January 2015*, which was before a significant number of refugees arrived. The `year` variable is coded as `2015` for observations from this election.
- Two *additional elections* were held in 2012 and 2013. The `year` variable captures observations for the election in May 2012 with the value `2012` and the election in June 2013 with `2013`.

- The units of analysis are *municipalities*.

Let's first familiarise ourselves with the data. A description of some of the variables that we will use today is below:

Variable	Description
`Year`	Election year (2012 to 2016)
`municipality`	Municipality id
`post`	A dummy variable that is 1 for the 2016 election and 0 otherwise
`treatment`	A dummy variable that is 1 if the municipality received a strong increase in refugees before the 2016 election and 0 otherwise
`evertr`	Treatment variable indicating if the municipality received refugees (= 1) or not (= 0)
`gdper`	The Golden Dawn's vote share at the municipality level

Now let's load the data from our computer by using the `read_dta()` function. We will call this data frame **greekislands**.

Let's start by checking our data as always.

Task 1 Use the `head()` function to familiarise yourself with the data set.**

Recall that the `treat` variable indicates municipalities that are treated at some point - independent of the timing -, while the `treatment` variable marks municipalities after they were treated.

As we can see, the data set covers multiple election years. Before working with the data, let's make sure we know how many and which elections are included in the data.

Task 2 How many elections, i.e. `years` are covered by the data?

Let's have a look at general voting patterns of the *Golden Dawn* over time, irrespective of treatment status of municipalities.

Task 3 Plot the vote share of the Golden Dawn Party (`gdper`) over time.

There are several ways to do this. For instance, you could plot the dispersion across municipalities by using the `boxplot` command or, alternatively, calculate average values per year. Feel free to pick the option you deem most appropriate.

Differences-in-Differences

Being aware of the general trends of the Golden Dawn's vote share is an important information about context and the party's history. However, it cannot tell us anything about the treatment effect we seek to analyse: The arrival of refugees in some Greek municipalities in the summer of 2015.

A naive observer might propose identifying this effect by looking at the differences between treated and untreated units in the post-treatment periods. Would this, however, be an appropriate representation of a possible treatment effect? It clearly would not! Comparing post-treatment differences only doesn't allows us to account for unit/municipality-specific effects and voting patterns. Treatment, after all, was not assigned randomly. We would not be able to say what the effect of the treatment is unless we can make a statement about how the treatment changed the outcome or resulted in the diversion from a previous trajectory. Using a *differences-in-differences* design allows us to do so.

Task 4 Estimate the treatment effect by calculating differences-in-differences between 2015 and 2016 using the `mean()` function.

> Hint: Calculate the differences between treated and untreated units for the years 2015 and 2016 first.

Task 5 Estimate the difference-in-difference between 2015 and 2016 using an OLS regression.

You can run a simple OLS with the interaction term of *treat* and a dummy variable for the post-treatment period as independent variables. However, you should restrict the data to the years 2015 and 2016, specifying `data = greekislands[greekislands$year>=2015,]` as argument for your OLS.

Task 6 Estimate the difference-in-differences between 2015 and 2016 using a regression with unit fixed effects.

> Hint: You can use `lm_robust()`, `feols()`, `plm()` or `lm()` with dummy variables.

Generalised Diff-in-Diff

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

Task 7 Estimate the difference-in-differences using a two-way fixed effects regression with all time periods and `treatment` as independent variable.

Task 8 Calculate robust standard errors for (i) the plm FE model, (ii) the two-way FE model and present the regression output in a single table. Also include the simple OLS model in the table.

> Note: There is no need to adjust standard errors after using `lm_robust()` as the command automatically does that.

Pararell trends

As previously mentioned, we cannot, generally, test whether the parallel trends assumption holds because we cannot, in principle, observe the vote share of the Golden Dawn among treated and untreated units for the **hypothetical scenario** of the 2015 refugee crisis **not occurring**. However, we can examine, either visually or through statistical analyses, whether the Golden Dawn's vote share trends more or less uniformly for both treated and untreated units **before the treatment**, namely the 2015 refugee crisis, occurred. Indeed, if it was really the treatment (in our case, refugees arriving on the coasts of some Greek islands in 2015) that causally affected the Golden Dawn's vote share and not anything else, then we should observe the Golden Dawn's vote share moving **in parallel** for units (in our case, Greek municipalities) that were either treated or not with refugees in 2015. However, note that detecting such parallel trends **before the treatment occurred** is only a **necessary** but not a **sufficient condition** for testing the parallel trends assumption which (hopefully) illustrates why this assumption is, principally, not testable. Nevertheless, the necessary condition is better than nothing. Indeed, at the very least, by examining parallel trends before the treatment occurred, we allow for the parallel trends assumption to be potentially **falsified**.

One way to examine these parallel trends before the treatment occurred is to plot the outcome variable for both treated and untreated units overtime, especially before the treatment/intervention occurred. To do so, we, first, need to calculate the Golden Dawn's mean vote share for both treated and untreated units. Let's do this with tidyverse code this time, specifically by using the `group_by()` and `summarise()` functions.

Task 9 Calculate the mean vote share for the Golden Dawn for the treated and the untreated units for all the elections. Store this into a new data frame called `plot.parallel`. Use tidy to create this new data frame.

The syntax and the description of the functions is provided below:

Function	Description
`new.data <- dataframe`	Assignment operator where the new data frame will be stored
`group_by()`	Group observations by a variable or set of variables

Function	Description
`summarise()`	Creates a new data frame with one ore more rows of each combination of grouping variables
`mutate()`	Allows to create new variable or modify existing ones

Let's now plot these results.

Task 10 Plot the parallel trends. Set vote share for the Golden Dawn in the y axis, year in the x-axis. Connect the data points of the two groups using a line. Place the legend of your plot at the bottom. Change the default colour. Below you will find all the functions necessary to generate this plot. Remember to use the plus sign between functions.

Function	Description
`ggplot(x = , y =, colour = )`	Map data components into the graph
`geom_point()`	To generate a scatterplot
`scale_color_manual(values=c("colorname1", "colorIwant2"))`	Replace the default palette
`theme(legend.position = "bottom")`	To place the legend at the bottom
`geom_line(aes(x=year,y=vote, color = condition, group = condition))`	To connect the dots with lines by group

Optional: Now let's look at the leads to identify any anticipatory effects. Let's imagine that the Golden Dawn back in 2012 believed that there was going to be a major humanitarian in the future. Then, they thought that they could exploit this situation to increase their electoral gains. In that case, we wouldn't be able to disentangle whether changes in vote share are due to the previous campaigning efforts on the part of the Golden Dawn or due to the influx of alyssum seekers to Greece. We can use leads to identify if there are any anticipatory effects. If we find systematic differences between treated and untreated units, this would suggest that units in one or both groups are responding to the treatment before receiving it.

EXTRA Exercise 1: Create dummy year variables equal to 1 for every year and only for the treated municipality. Call this variable leads, plus the year. For example, the lead2012, will take value 1 only for treated municipalities and only for observations of these municipalities in the year 2012. You can see an example below. use the `mutate()` function to create these new variables. You can also use the `ifelse()` function to create these dummy variables. The syntax of the `ifelse()` function is the following: `new variable = ifelse(condition, "value if condition is met", "value if the condition is not met")`. Create these dummy variables for the elections in 2012, 2013, and 2015.

EXTRA Exercise 2: Conduct the same two-way fixed-effect model that we used before, but rather than using the `treatment` variable, replace this variable with the new leads variables that you created. Ran separate estimations for each lead. Store the outputs of these regressions into different objects. Does the evidence suggest that there are any anticipatory effects? Are the results of these three models statistically significant? You can use the summary() or screenreg() functions to take a look at your results.

Now, let's plot all the two-way fixed effects models where we used the `plm()` function into a single figure.

EXTRA Exercise 3: Plot the coefficients from the leads models, plus the two-way fixed model for the 2016 election that you used in Question 7 ("twoway_FE"). Use the `plot_coef()` function to generate this plot. Add the argument `scale` and set it equal to `TRUE`. Also, include the argument `robust` and set it equal to `TRUE`. In addition to the `plot_coefs()` include the `coord_flip()` function. This function will flip the Cartesian coordinates of the plot, so we have the models (years) in the x-axis and the coefficients in the y axis. Remember to add plus sign operator between two functions. You can also add the `xlab("Year")` function to a label in the x-axis.

Placebo test

We can conduct a placebo test to evaluate whether the parallel trend holds. We are trying to show that there will be no statistically significant difference in the trend between treated and untreated municipalities.

The steps to conduct are the following:

1. Use data for periods that came before the treatment was implemented/happened.
2. Create a dummy variable for each before treatment that is equal to 1 only for that specific year and only for treated units (as we did before).
3. Estimate the difference-in-difference using `plm()` or `lm_robust()` function.
4. If you find statistically significant results, this may suggest a violation of parallel trends.

Task 14 Drop all the observations of the year of the intervention (2016). Do this using the `filter()` function. Then, create a fake treatment variable and call it `post2` and set it equal to 1 for all observations in year 2015. This variable would indicate as the hypothetical case that the municipality would received refugees in 2015.

Task 15 Conduct the same two-way fixed effect model using the `lm()` and use the `post2` variable. Store all the models in a list and plug this list inside of the `modelsummary(list)` function to report your results. Did you find statistically significant differences between treated and untreated units pre-treatment?. You can see an example of how to store multiple models in a list `list()`. Also, subset the data so one model will only conduct the placebo test using the 2012 and 2015 elections, and another model only using the observations from the 2013 and 2015 elections.