Difference-in-difference

```
library(tidyverse)
library(magrittr)
library(tidymodels)
library(haven)
library(estimatr)
library(haven)
greekislands <- read_dta("greekislands.dta")

greekislands <- greekislands %>%
  rename(evertr = treat)
```

Data

```
#install.packages("downloadthis")
library(downloadthis)

download_link(
  link = "https://github.com/bayreuth-politics-ci/CI22/raw/gh-pages/docs/data/greekislands
  output_name = "greekislands",
  output_extension = ".dta",
  button_label = "Lab 6 Data",
  button_type = "success",
  has_icon = TRUE,
  self_contained = TRUE
)

#download_link(
#  link = "https://github.com/dpir-ci/CI22/raw/gh-pages/docs/lectures/lecture4.pdf",
#  output_name = "week5",
#  output_extension = ".pdf",
#  button_label = "Lecture Slides",
#  button_type = "default",
#  has_icon = FALSE,
#  self_contained = FALSE
#)
```

Recap

The intuition of the DD 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 get an unbiased estimate of a (policy) intervention.

We learned that we can break down the difference between treated units 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 the effect of treatment. 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 that treated units would follow similar trends to the untreated ones. Unfortunately, we cannot test whether this assumption holds, but we can at least conduct some tests that would be indicated that the **Parallel Trends** holds.

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

- **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.

If you go back to the Golden Dawn, we can see how we retrieve the beta coefficient from the interaction.

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

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

Then, we can get our estimate by the calculating the difference of the outcome variable for both treated and untreated units and, and then subtract these differences:

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

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

$$\beta_3$$

- **Unit and time dummies and a treatment indicator**: Again, in the case of the Golden Dawn example, we can represent this estimation strategy using 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:

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

$$\beta_1$$

Again we are levering 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. Given that we have repeated observations, this type of data exhibits serially correlated regressors and residuals, we need to make the appropriate adjustments to calculate standard errors. One way is to address this is to use clustered standard errors.

Before starting this seminar

1. Create a folder called “lab6”
2. Download the data (you can use this button or the one at the top, or read csv files directly from github):

3. Open an R script (or Markdown file) and save it in our “lab6” folder.

4. Set your working directory using the setwd() function or by clicking on “More“. For example *setwd(“~/Desktop/Causal Inference/2022/Lab6”)*

5. Let’s install an load the packages that we will be using in this lab:

```
library(jtools) # generate plots with regression coefficients
library(stargazer) # generate formated regression tables
library(texreg) # generate formatted regression tables
library(tidyverse) # to conduct some tidy operations
library(ggplot2) # to generate plots
library(plm) # conduct one-way and two-way fixed effects
library(estimatr) #  to conduct ols and provides robust standard errors
library(lmtest) # calculates the variance-covariance matrix to be clustered by group.
library(sandwich) # to calculate heteroscedasticity-Consistent Covariance Matrix Estimatio
library(haven) # upload dta files
#install.packages("wesanderson")
library(wesanderson) # Wes Anderson palette (Let's make our plots as Indie as possible)
#install.packages("modelsummary")
library(modelsummary)
```

Seminar Overview

In this **seminar**, we will cover the following topics:
1. “Manually” calculate the difference-in-difference estimator”
2. Obtaining the difference in difference estimator using the `lm()` function
3. Check for parallel trends
4. Conduct a fixed effect in difference in difference estimation using both `lm_robust()` and `plm()` function.
4. Conduct a placebo test

Waking Up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties? (Dinas et al, 2018)

The main question of this paper is the following: **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.

We can see here on the Figure in the left that level of exposure to refugees across the Aegean islands. The figure on the right shows how sudden was the influx of refugees.

Timing

- The refugee crisis started in the spring of 2015
- Greece held an election September, 2015, right after the first wave of refugee arrival, where in the data is coded as 1 for the `post` variable and coded as year 2016
- The previous election had taken place only eight months prior in January 2015, before significant number of refugees arrived, which is coded in the data as year 2015.
- Two more elections were held in 2012. The election in May 2012 is coded as 2012 and the election in June 2013 is coded as 2013.

- The unit of analysis are municipalities

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

Variable	Description
`Year`	Election Year from 2012 to 2016
`municipality`	Municipality id
`post`	A dummy variable that switches on year 2016
`treatment`	A dummy variable indicating if the municipality received refugees on the 2016 election
`evertr`	Treatment variable indicating if the municipality received refugees
`gdper`	Vote share for the Golden Dawn at the municipality level as percentage of total votes

Now let's load the data. There are two ways to do this:

You can load the dataset from your laptop using the `read_dta()` function. We will call this data frame **greekislands**.

```
# Set your working directory
#setwd("~/Desktop/Causal Inference/2022/Lab6")
#
#setwd("/cloud/project")
```

Or you can download the data from the course website from following url: https://github.com/bayreuth-politics/CI22/raw/gh-pages/docs/data/greekislands.dta.dta.

Let's start by checking our data as always.

Exercise 1: Use the `head()` function to familiarise yourself with the data set.

Reveal Answer

```
greekislands <- greekislands %>%
   group_by(muni) %>%
   mutate(treat = max(treatment))
```

```
head(greekislands)
```

This is what we would expect. 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.

Exercise 2: How many elections, i.e. years are covered by the data?

Reveal Answer

```
unique(greekislands$year)
```

Using the `unique` command, we find that four elections are covered by the data: The elections that took place in 2012, 2013, 2015, and 2016.

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

Exercise 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.

Reveal Answer

```
# First option: Boxplot
boxplot(gdper ~ year,
        data = greekislands,
        xlab = "Election Year",
        ylab = "Golden Dawn Vote Share",
        las = 2) # This argument rotates the axis labels

# Second Option: Plot averages
# Calculating and storing means
average_data <- greekislands %>%
            group_by(year) %>%
            summarize(gd_averages = mean(gdper))

# Plotting means using ggplot

ggplot(average_data, aes(x = year, y = gd_averages)) +
  geom_point() + geom_line() + xlab("Election Year") + ylab("Average GD Vote Share") + yli
```

We can se that the vote share for the Golden Dawn party has remained somewhat stable between the 2012 and 2013 elections, and dropped in the 2015 election to an average value (per municipality) of 4.46%. In the 2016 election, the party's vote share rose substantively - to a new high of 6%.

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.

Exercise 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.

Reveal Answer

```
# Difference in means between treated and untreated in 2016 (pot-treatment).
post_difference <- mean(greekislands$gdper[greekislands$treat == 1 & greekislands$year ==
post_difference

# Difference in means between treated and untreated in 2015 (pre-treatment).
pre_difference <- mean(greekislands$gdper[greekislands$treat == 1 & greekislands$year == 2

pre_difference

# Now calculate the difference between the two differences above
diff_in_diff <- post_difference - pre_difference
diff_in_diff
```

The difference in the Golden Dawn's vote share between treated and untreated municipalities has increased in 2016. The differences-in-differences amount to `2.12`. This suggests that the treatment (i.e. the arrival of refugees) increased the vote share of the golden dawn by roughly 2%points in the affected municipalities.

While it is important to understand what exactly the difference-in-difference means, we usually do not have to calculate it manually. In fact, we can simply use an OLS to estimate differences-in-differences.

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

Hint

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.

Reveal Answer

```
# Creating a post-treatment period (i.e. 2016) dummy
greekislands$post_treatment <- greekislands$year == 2016


ols_did <- lm_robust(gdper ~ evertr * post_treatment, data = greekislands[greekislands$yea
summary(ols_did)

library(broom)

print(summary(ols_did),digits=max(3, getOption("digits") - 3))
```

The estimate for the interaction term, i.e. treated units after they were treated, corresponds to the difference-in-differences. As you can see, the OLS provides exactly the same point estimate as the manual calculation of the difference-in-differences, `2.12` - however, the OLS also provides measures of uncertainty, showing us that the estimate is significant to conventional levels. The OLS provides quite intuitive estimates: The intercept corresponds to the Golden Dawn's vote share in in untreated municipalities in 2015. Treated municipalities had, on average, a higher vote share for the party by roughly `0.62`%points. The parties vote share increased by about 1.27%points in untreated municipalities in 2016.

Note that there are multiple ways to calculate the difference-in-differences. Fixed effects are usually the preferred approach as they provide the most flexible (e.g. in terms of multiple time periods) and most efficient estimator. Let's therefore calculate a fixed effects model for the same difference-in-differences as in Exercise 5.

Exercise 6: Estimate the difference-in-differences between 2015 and 2016 using a Fixed Effects regression.

_Hint: You can use either `lm_robust()`, `plm()` or `lm()` with dummy variables.

Reveal Answer

```
# lm_robust
fe_robust <- lm_robust(gdper ~ evertr*post_treatment, fixed_effects= factor(muni), data=gr
summary(fe_robust)

# plm
fe_plm <- plm(gdper ~ evertr*post_treatment, model = "within", index = c("muni"), data=gre
summary(fe_plm)

fe_dummy <- lm(gdper ~ evertr * post_treatment + factor(muni), data = greekislands[greekis
summary(fe_dummy)
```

All three fixed effects estimators provide the same estimate - which corresponds to the one we manually estimated at the beginning of this lab. Note that we could also use the first differences estimator - in the case of two time periods it is equivalent to the fixed effects estimator. However, the fixed effects estimator provides much more flexibility when more than two time periods are being analysed.

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.

Exercise 7: Estimate the difference-in-differences using a two-way Fixed Effects regression with all time periods and `treatment` as independent variable.

Reveal Answer

```
twoway_FE <- plm(gdper ~ treatment, effect="twoways",model="within",index=c("muni","year")
summary(twoway_FE)
```

As you can see, the estimate is now slightly smaller than the one we got by comaparing 2015 and 2016 only. This is not surprising as the regression considers all pre-treatment periods now and, therefore, attributes some of the variance to other factors such as unit-specific trends. However, it is noteworthy that the change is negligible. We still estimate that the treatment effect, or ATT, is roughly above 2%points.

Exercise 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. 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.

Reveal Answer

```
### ADJUSTING STANDARD ERRORS
## plm FE model
fe_plm_se <- coeftest(fe_plm, vcov = vcovHC(fe_plm, type = "HC1",
                                         cluster = "group"))
fe_plm_se
```

```
## two-way FE model

twoway_FE_se <- coeftest(twoway_FE, vcov = vcovHC(twoway_FE, type = "HC1",
                                        cluster = c("group", "time")))
twoway_FE_se


## CREATING A REGRESSION TABLE

screenreg(l=list(ols_did, fe_plm_se, twoway_FE_se), custom.header = list("OLS (2-period)"
```

Pararell trends

As we know, we cannot test whether the parallel trends, but we can conduct some visual inspects and statistical analyses. One way to do this is to plot the outcome variable for both treated and untreated units overtime before the intervention.

Let’s first get the mean vote share for the Golden Dawn for the treated and untreated units. Let’s do using Tidyverse using the `group_by()` and `summarise()` functions.

Exercise 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 and called `plot.parallel`. Use tidy to create this new data frame.The syntax and the description of the functions below:

```
new.data <- data %>%
  group_by(variable_1, variable_2) %>%
  summarise(new_variable = mean(outcome, na,rm = TRUE)) %>%
  mutate(condition = ifelse(evertr == 1, "Treatment", "Control") )
```

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
`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

Reveal Answer

```
plot.parallel <- greekislands %>%
  group_by(evertr, year) %>%
  summarise(vote = mean(gdper)) %>%
  mutate(condition = ifelse(evertr == 1, "Treatment", "Control"))

head(plot.parallel)
```

We can see now that we have the average vote share for treated and untreated for every election year. For example, we can see that the average vote share for the Golden Dawn in untreated islands was `5.33` for the 2013 election.

Right, so we see that we have the average vote share for treated and untreated units for each election from 2012 to 2016. Now, let's plot the trends before and after the intervention.

Exercise 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 and use the Wes Anderson Palette. 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=wes_palette("Royal1"))`	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

Reveal Answer

```
ggplot(aes(x = year, y = vote, colour = condition), data = plot.parallel) +
  geom_point() + scale_color_manual(values=wes_palette("Royal1")) + theme(legend.position
```

The plot suggests that parallel trends assumption would hold in this case. The vote share for the Golden Dawn follows a similar path for treated and untreated units for at least three elections before the intervention.

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.

Exercise 11: 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.

```
greekislands <- greekislands %>%
  mutate(lead2012 = ifelse(evertr == 1 & year == 2012, 1, 0))
```

Reveal Answer

```
table(greekislands$year)

greekislands <- greekislands %>%
  mutate(lead2012 = ifelse(evertr == 1 & year == 2012, 1, 0),
         lead2013 = ifelse(evertr == 1 & year == 2013, 1, 0),
         lead2015 = ifelse(evertr == 1 & year == 2015, 1, 0))
```

Now that we have created these dummy variables, we can run a two-way fixed effect model and see if they are anticipatory effects.

Exercise 12: 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.

Reveal Answer

```
lead2012 <- plm(gdper ~ lead2012, model = "within", effect = "twoways",
            index = c("muni", "year"), data = greekislands)

lead2013 <- plm(gdper ~ lead2013, model = "within", effect = "twoways",
            index = c("muni", "year"), data = greekislands)

lead2015 <- plm(gdper ~ lead2015, model = "within", effect = "twoways",
            index = c("muni", "year"), data = greekislands)

summary(lead2015) # as an example.
```

The evidence gathered from the three models suggests there are no anticipatory effects. We don't find any statistically significant dynamics between treated and untreated municipalities on vote share for the Golden Dawn before the 2016 election, which is good!

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

Exercise 13: 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.

Reveal Answer

```
plot_coefs(lead2012, lead2013, lead2015, twoway_FE, scale = TRUE, robust = TRUE) +
coord_flip() + ylab("Year")
```

Again, we can observe that pre-treatment coefficients are nearly zero and they are not statistically significant, but then for the 2015 election, we find that the vote share for Golden Dawn has risen substantially. Thus, we found strong evidence that the increase in vote share is caused by the influx of refugees into the Greek islands.

Placebo test

We can conduct a placebo test to evaluate whether the parallel trend holds. We are trying to prove that there is no clear difference in trending tendencies 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)
2. Estimate the difference-in-difference using `plm()` or `lm_robust()` function.
3. If you find statistically significant results, this may suggest a violation of parallel trends.

Exercise 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. You can see an example below.

```
greekislands <- greekislands %>%
  filter(variable != 2016) %>%
  mutate(new variable = year variable == "year")

## or alternatively

greekislands <- greekislands %>%
  filter(variable != 2016) %>%
  mutate(new.variable = ifelse(year variable == "year", 1, 0))
```

Reveal Answer

```
table(greekislands$year)

greekislands <- greekislands %>%
  filter(year != 2016) %>%
  mutate(post2 = ifelse(year == "2015", 1, 0))
```

Great! Now let's use this dummy variable to perform the placebo test. We are essentially going to test is that there are no differential trends between treated and untreated, as we did before when we looked at anticipatory trends.

Exercise 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.

```
models <- list(
    "plm "= lm(outcome ~ treatment, model = "within", effect = "twoways", index = c("muni"
    "lm" = lm(outcome ~ outcome, data = data, subset=(year!=year)), # set year equal to ye
    "lm_robust" = lm_robust(outcome ~ treatment, data = data[data$year != year,], cluster
```

Reveal Answer

```
models <- list(
     "1215"=lm_robust(gdper ~ evertr + post2+evertr*post2, data = greekislands, subset=(ye
    "1314" = lm_robust(gdper ~ evertr + post2+evertr*post2, data = greekislands, subset=(y


modelsummary(models)
```

We find that there are no systematic differences between treated and untreated municipalities. We find near to zero coefficients.

	Pre	**Post**	Difference
Far			
Treated	4.18	5.01	0.83
Control	5.12	5.68	0.56
Difference	-0.94	-0.67	
Diff-in-Diff	**0.27**		
			0.95
Close			
Treated	5.87	8.71	2.84
Control	3.72	5.34	1.62
Difference	2.15	3.37	
Diff-in-Diff	**1.22**		