Lab Session: Difference-in-differences

Cristian Huse

Introduction

This lab session is about **Difference-in-Differences (DD)**. The structure of the session is as follows:

- Preparation;
    - Load libraries, setwd, load data (.dta format)
- Example 8. DD in a **regression** framework
- Example 9. DD in a **multivariate regression** framework
- Example 10. Household fixed-effect estimates for DD
- Bonus: Example 11. Calculating DD by taking the difference between before-after diference in the treatment and control groups.

Please note that this lab session focuses on the basic setup whereby one has two time periods (pre- and post-treatment) and two groups (treatment and control). There are more complex designs which we aim to address later on.

Preparation

```
## Initialize ####

rm(list=ls())

#Load libraries
library(clubSandwich) #for vcovCR
library(fixest) #for feols etc
library(haven) #for read_dta
library(modelsummary) #for neat tables
library(panelr) #for panel_data & widen_panel
library(plm) #for plm
library(tidyverse)

## Set working directory
#setwd("INSERT PATH OF FOLDER WHERE YOU SAVED THE DATASET WITH / or \\")
setwd("C:/Users/huse-admin/Dropbox/CRISTIAN/Teaching/Cursos_Meus/Teaching_2021/EPE/Lab6_DD")

## open data
```

```
#Open the cleaned data set
#set path for data
evaluation <- file.path(getwd(), "Data", "evaluation.dta")

#import .dta file
evaluation.df <- read_dta(evaluation)
```

Difference-in-differences

The standard DD estimator within the regression framework is the **Two-Way Fixed-Effects (TWFE)** estimator. Assuming two periods (pre- and post-treatment) and two groups (treatment and control), the estimate of interest is the one associated to the interaction between the post-treatment and treatment group dummies, thus the **two-way** reference. (See lecture slides for details.)

In our particular setting, we will compare the change in health expenditures over time between enrolled and non-enrolled households in the treatment localities (note the **panel structure**, i is followed over time and can be at either treatment or control group, ideally both). The corresponding regression model is as follows:

$$Y_{igt} = \beta_1 P_i + \beta_2 t + \delta P_i t + \beta_3 X_{igt} + \alpha_g + \theta_t + \varepsilon_{igt}$$

where $g = T, C$ denotes group, i denotes individual, Y is the outcome, $t = 0$ and $t = 1$ are baseline and follow-up periods (thus a treatment indicator), P is the treatment dummy, α_g is a group fixed-effect, θ_t is a period fixed-effect. One can also include characteristics for the treatment and control groups in both time periods (X_{igt}).

Crucially, our main coefficient of interest is δ, which provides the ATE of the programme under regularity conditions. (Please see the lecture for details.)

```
#Select the relevant data and create necessary variables
dd.df <- subset(evaluation.df, treatment_locality == 1)
dd.df$eligible_round <- dd.df$eligible*dd.df$round
```

Example 8. Difference-in-Differences in a Regression Framework

This example illustrates how to implement the DD-TWFE estimator using R and data for the HISP case. The assumption regarding the data is that *"we have data only from localities where the program has been offered. In these localities, we have data both for households that participate in the program, as well as households that do not to participate. Data are available for a baseline survey collected before the program, and a follow-up survey collected after the program."*

In what follows, we will estimate the same specification using different commands/options, e.g., creating the interaction variable between participation in the program and the time at which the data are measured (**enrolled_round**) or defining the interaction when calling the function, and using **lm()** or **feols()**. In either case, the outcome variable is regressed on this interaction term, plus indicators of whether the household participated in the programme and the time at which each data point is observed. The results document that health expenditures for households that enrolled in the program were $8.16 lower than among households that did not enroll.

```
# lm + standard
ex8_lm1 <- lm(health_expenditures ~ eligible_round + round + eligible,
              data = dd.df)
```

	ex8_lm1	ex8_lm2	ex8_feols1
(Intercept)	20.791***	20.791***	20.791***
	(0.172)	(0.172)	(0.172)
eligible_round	−8.163***		
	(0.319)		
round	1.513***	1.513***	1.513***
	(0.356)	(0.356)	(0.356)
eligible	−6.302***	−6.302***	−6.302***
	(0.193)	(0.193)	(0.193)
eligible × round		−8.163***	−8.163***
		(0.319)	(0.319)
Num.Obs.	9919	9919	9919
R2	0.344	0.344	0.344
Std.Errors	C: locality_identifier	C: locality_identifier	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

```r
# lm + interaction
ex8_lm2 <- lm(health_expenditures ~ eligible:round + round + eligible,
              data = dd.df)
# feols + interaction
ex8_feols1 <- feols(health_expenditures ~ eligible:round + round +
                      eligible,
                    cluster = ~ locality_identifier, data = dd.df)

# Note how modelsummary "fixes" the standard errors to be clustered at the locality level
# In fact, specifying the option cluster wasn't required when calling feols

models8 <- list("ex8_lm1" = ex8_lm1, "ex8_lm2" = ex8_lm2, "ex8_feols1" = ex8_feols1)
modelsummary(models8,
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

Note above the reference to the clustered standard errors (at the locality level) and how **eligible_round** and eligible:round are treated as different variables. The table can be somewhat improved as follows.

```r
# A nicer-looking table
cm <- c('eligible_round' = 'eligible x round',
        'eligible:round' = 'eligible x round',
        'eligible' = 'eligible',
        'round' = 'round'
        )

modelsummary(models8,
             coef_map = cm,
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

	ex8_lm1	ex8_lm2	ex8_feols1
eligible x round	−8.163***	−8.163***	−8.163***
	(0.319)	(0.319)	(0.319)
eligible	−6.302***	−6.302***	−6.302***
	(0.193)	(0.193)	(0.193)
round	1.513***	1.513***	1.513***
	(0.356)	(0.356)	(0.356)
Num.Obs.	9919	9919	9919
R2	0.344	0.344	0.344
Std.Errors	C: locality_identifier	C: locality_identifier	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

```
# Make sure to compare the layouts of the tables
```

Example 9. Difference-in-Differences in a Multivariate Regression Framework

So far, we haven't included any controls (X_{igt} above). However, this might be important to increase precision of the estimates and in cases when balancing across groups is not great.

```
# lm + standard
ex9_lm1 <- lm(health_expenditures ~ eligible_round + round + eligible + age_hh +
                age_sp + educ_hh + educ_sp + female_hh + indigenous + hhsize +
                dirtfloor + bathroom + land + hospital_distance,
              data = dd.df)

# feols
ex9_feols1 <- feols(health_expenditures ~ eligible_round + round + eligible + age_hh +
                      dirtfloor + bathroom + land + hospital_distance,
                  cluster = ~ locality_identifier, data = dd.df)
# Table
# Raw, full table (labels could be improved, see below)
models9 <- list("ex9_lm1" = ex9_lm1, "ex9_feols1" = ex9_feols1)
modelsummary(list(ex9_lm1,ex9_feols1),
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

In this particular case, the inclusion of controls doesn't seem to have an effect on the ATE estimates. Note, however, the increased **R-squared**.

As before, the table could also be displayed reporting only the main estimates as follows.

```
# Nicer
cm <- c('eligible_round' = 'eligible x round',
        'eligible:round' = 'eligible x round',
        'eligible' = 'eligible',
        'round' = 'round'
```

	Model 1	Model 2
(Intercept)	27.395***	27.395***
	(0.553)	(0.552)
eligible_round	−8.161***	−8.161***
	(0.320)	(0.320)
round	1.451***	1.451***
	(0.356)	(0.356)
eligible	−1.513***	−1.513***
	(0.130)	(0.130)
age_hh	0.080***	0.080***
	(0.011)	(0.011)
age_sp	−0.020	−0.020
	(0.013)	(0.013)
educ_hh	0.060**	0.060**
	(0.029)	(0.029)
educ_sp	−0.077**	−0.077**
	(0.034)	(0.034)
female_hh	1.104***	1.104***
	(0.316)	(0.315)
indigenous	−2.312***	−2.312***
	(0.236)	(0.236)
hhsize	−1.995***	−1.995***
	(0.039)	(0.039)
dirtfloor	−2.300***	−2.300***
	(0.163)	(0.163)
bathroom	0.500***	0.500***
	(0.158)	(0.158)
land	0.091***	0.091***
	(0.029)	(0.029)
hospital_distance	−0.003	−0.003
	(0.003)	(0.003)
Num.Obs.	9919	9919
R2	0.552	0.552
Std.Errors	C: locality_identifier	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

	ex9_lm1	ex9_feols1
eligible x round	−8.161***	−8.161***
	(0.320)	(0.320)
eligible	−1.513***	−1.513***
	(0.130)	(0.130)
round	1.451***	1.451***
	(0.356)	(0.356)
Num.Obs.	9919	9919
R2	0.552	0.552
Std.Errors	C: locality_identifier	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

```
        )
modelsummary(models9,
             coef_map = cm,
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

Example 10. Household Fixed Effect Estimates for Difference-in-Differences

One could actually do more than control for heterogeneity as above. If the data is a **panel** (sometimes also called **longitudinal data**), we can observe all individuals in both groups and in both time periods. This allows us to control for **individual fixed-effects**. That is, we are now able to control for time-invariant heterogeneity at the individual level instead of only demographics.

Importantly, it is not always the case that the panel is **balanced** (or rectangular), i.e., all individuals are observed every period. The typical case one is observe is the one where the panel is **unbalanced**. For instance, individuals might die, move, which means they don't "generate data" any longer. What this means in practice is that one needs to keep track of individuals and time periods to correctly calculate differences and averages in this setting. Or, in our case, to inform R who these are. (We will do this within **fixest** and **plm** = linear models for panel data.)

To understand what you are doing, please make sure you read the variable descriptions and/or open the data to understand who are the "individuals" and "time periods" in this particular case.

```
# Within fixest
#?panel
panel(data = dd.df, panel.id = c("household_identifier", "round"))
```

```
## # A tibble: 9,919 x 23
##    locality_identif~ household_ident~ treatment_local~ promotion_local~ eligible
##                <dbl>            <dbl>            <dbl>            <dbl>    <dbl>
##  1                26                5                1                1        1
##  2                26                5                1                1        1
##  3                26               11                1                1        1
##  4                26               11                1                1        1
##  5                 26                13                1                1        1
##  6                 26                13                1                1        1
##  7                 26                16                1                1        1
##  8                 26                16                1                1        1
##  9                 26                21                1                1        1
## 10                 26                21                1                1        1
## # ... with 9,909 more rows, and 18 more variables: enrolled <dbl>,
## #   enrolled_rp <dbl>, poverty_index <dbl>, round <dbl>,
## #   health_expenditures <dbl>, age_hh <dbl>, age_sp <dbl>, educ_hh <dbl>,
## #   educ_sp <dbl>, female_hh <dbl>, indigenous <dbl>, hhsize <dbl>,
## #   dirtfloor <dbl>, bathroom <dbl>, land <dbl>, hospital_distance <dbl>,
## #   hospital <dbl>, eligible_round <dbl>
```

As before, the idea to estimate a fixed-effects model is to create a variable that is equal to 1 only for enrolled households in the follow-up period. The coefficient of this variable is our impact estimate, which takes value −$8.16.

```r
# Generate the interaction dv(enrolled==1)*dv(round==1)
# Different ways are possible
# 1.library(dplyr)
#?mutate
dd.df <- mutate(dd.df, xtenrolled = ifelse(enrolled == 1 & round == 1, 1, 0))
# Base R -- old school
dd.df$xtenrolled2<-(dd.df$enrolled == 1)*(dd.df$round == 1)
tmp<-sum(abs(dd.df$xtenrolled2-dd.df$xtenrolled))
#Exactly the same!

# feols
ex10_feols1 <- feols(health_expenditures ~ xtenrolled + round +
                       eligible,
                     cluster = ~ locality_identifier,
                     data = dd.df)

# Table
cm <- c('eligible_round' = 'eligible x round',
        'eligible:round' = 'eligible x round',
        'xtenrolled' = 'eligible x round',
        'eligible' = 'eligible',
        'round' = 'round'
        )
modelsummary(list(ex10_feols1),
             coef_map = cm,
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

```r
ex10_plm <- plm(health_expenditures ~ xtenrolled + round + eligible, data = dd.df,
                model = "within",
                index = c("household_identifier", "round"))

modelsummary(ex10_plm,
             coef_map = cm,
```

	Model 1
eligible x round	−8.163***
	(0.319)
eligible	−6.302***
	(0.193)
round	1.513***
	(0.356)
Num.Obs.	9919
R2	0.344
Std.Errors	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

	Model 1
eligible x round	−8.163***
	(0.319)
Num.Obs.	9919
R2	0.244

* p < 0.1, ** p < 0.05, *** p < 0.01

```
             vcov = vcovCR(ex10_plm, cluster = dd.df$locality_identifier, type = "CR1S"),
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

BONUS – Example 11. Calculating Difference-in-Differences Estimates by Taking the Difference between Before-After Difference in the Treatment and Comparison Groups

(Data manipulation will be focused on in the future)

"The same single-difference can be estimated manually by computing the differences in the variables over time. One way to do this is to first reshape the dataset from long to wide, so that each row of data includes only one unit. In the example below, you can keep only the variables you need for the estimation. We set the program participation variable equal to 0 at baseline, and manually calculate the difference between health expenditures and program participation in the baseline and follow-up rounds. We then run a regression of the difference in the outcome variable over time on the treatment dummy. The impact estimate is exactly the same as with the fixed-effect panel estimate (-8.16)."

```
#?subset
ex11.df <- subset(dd.df, select = c("health_expenditures", "treatment_locality", "locality_identifier",
  panel_data(id = household_identifier, wave = round) %>%
  widen_panel(separator = "_", varying = c("health_expenditures", "enrolled")) %>%
  mutate(dy = health_expenditures_1 - health_expenditures_0,
         enrolled_0 = 0,
         dp = enrolled_1 - enrolled_0)

ex11_lm <- lm(dy ~ dp, data = ex11.df)
```

	Model 1
(Intercept)	1.513***
	(0.356)
dp	−8.163***
	(0.319)
Num.Obs.	4959
R2	0.159
Std.Errors	C: locality_identifier

* p < 0.1, ** p < 0.05, *** p < 0.01

```
modelsummary(ex11_lm,
             vcov = ~ locality_identifier,
             stars = c("*" = .1, "**" = .05, "***" = .01),
             fmt = 3,
             gof_omit = "AIC|BIC|Log.Lik.|R2 Adj.|R2 Within|R2 Pseudo|F")
```

References

Gertler, Paul J.; Martinez, Sebastian; Premand, Patrick; Rawlings, Laura B.; Vermeersch, Christel M. J. (2016). Impact Evaluation in Practice, Second Edition, Technical Companion (Version 1.0). Washington, DC: Inter-American Development Bank and World Bank.