PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Applied Econometrics
R Project - Case 1 - Part 01

Kelian IDRISSOU and Ibrahim FOUSFOS

Table of Contents :

Section	Page
1. Research question and Theoretical framework	1-2
2. Available data inspection and exploration	2-3
3. Data Manipulation	4-5
4. Annexe	6-7

1. Provide a short summary of the question at hand.

a) Formulate a clear, concise research question that relates to the case in exam

In this case, the bank wants to evaluate the efficiency of a meeting with a salesperson, and wonder if it can effectively increase clients' saving behaviour, especially with contributions to retirements. To do so, a random subset of clients is invited to a meeting on time 2 and their saving behaviour is observed over five periods (from t=0 to t=4, the meeting happens on t=2) before and after the intervention. Our economist's goal is to measure the causal effect of the meeting on savings, age, gender and prior saving habits.

We will ask then :

Does a meeting with a salesperson causally increase a client's savings and more specifically, their retirement savings?

b) Propose a hypothesis to test empirically with ideal and available data

H0: The meeting with the salesperson has no effect on retirement savings → The change in saving rate for the treatment group is equal to zero
H1: The meeting with the salesperson has a real effect on retirement saving → The change in the saving rate for the treatment group is different from zero

c) Clearly state your prior expectations, ground such expected results and effects based on economic and financial theory

Our prior expectation is that clients who meet a salesperson will, on average, increase their retirement savings compared to those who do not. This effect should become visible starting from period (t=2), when the meeting takes place assuming that the meeting effectively influences financial behaviour.

However to properly identify the causal effect of the meeting we must consider potential confounding factors such as income and gender factors. According to standard economic reasoning individuals with higher income have a greater capacity to save more even if their consumption behaviour is similar to others. The existing disparities between gender income makes men more expected to have higher saving levels than women.
To isolate the true treatment effect, we rely on the Rubin Causal Model (1974) that we saw in class. In this framework, each client ' i ' has two potential outcomes :

Y i (1) : savings level if the client attends the meeting
Y i (2) :savings level if the client does not attend the meeting

The Observed outcome can be written as: Y i = Di Yi(1) + (1-Di) Yi(0)
With Di = 1 → The client attended the meeting
Di = 0 → The client did not attend the meeting.

The casual effect of the meeting for each client will be :
□i = Yi(1) - Yi(0)

However, we can only observe one of these two potential outcomes for each client, this is the core of the theory. The average treatment effect (ATE) can be identified only if the assignment of the meeting is random and independent of potential outcomes, ensuring that treated and control groups are comparable in all relevant aspects (income, saving, gender…).

Under these conditions we expect to have a significant positive average treatment effect (ATE) of the salesperson meeting on clients'savings, once differences in income and gender are properly controlled.

2) Available data inspection and exploration

a) Import the dataset in R with a proper name

We upload the data from the EPI and we store it in a dedicated folder in our laptop. We then call it by creating a data base called : "groupe_41"
groupe_41= read.csv('/Users/ibrah/projet/group41.csv')

b) List the variables in the just imported dataset. What variables are useful to test the working hypothesis and answer the research question stated above? What are the outcome variables that you want to consider?

After importing the database, we can list all the variables using the command: names (groupe_41)

"id"	"time"	"female"	"age"	"yincome"	"savings"	"retirement"	"meeting"

The Useful Variables to test the working hypothesis and answer the research question are :

- Meeting ; Savings ; Retirement ; yincome ; female ; age

The outcome Variables that we want to focus on (Dependant Variables) are :

- savings → to study the general saving behaviour
- retirement → to study the specific effect on retirement savings

c) Provide a summary statistic for the whole dataset: variables' classes, percentiles, mean, median, size of the dataset

Statistics	id	time	age	yincome	savings	retirement

min	1	0	28.00	13297	0	0.00
1st Qu	2501	1	36.00	40734	4726	11.14
Median	5000	2	52.00	51127	8530	176.68
Mean	5000	2	52.52	51999	10100	644.47
3rd Qu	7500	3	69.00	61616	13497	757.10
Max	10000	4	87.00	130032	66966	14326.75

Statistics	Females	Meeting
Mode	logical	logical
FALSE	23810	29530
TRUE	26190	20470

d) What variables identify the unit of observation? Shortly describe the structure of the dataset, such as groupings and repeated observations

The variables that uniquely identify each observation are (id, time). We can observe the same client over five different periods (t=0,1,2,3,4), where we record data on income (yincome), savings and retirement for both treated groups (meeting = true) and non treated groups (meeting = false), according to their group characteristics (gender, age) over time.

e) To have a sense of the data, show the first 10 rows of the dataset

id	time	female	age	yincome	savings	retirement	meeting
1	0	TRUE	77	49335.82	7400.37	222.01	TRUE
1	1	TRUE	78	50698.14	7604.72	228.14	TRUE
1	2	TRUE	79	52138.99	11611.70	2195.80	TRUE
1	3	TRUE	80	53417.91	12858.05	2482.92	TRUE
1	4	TRUE	81	54897.08	13817.96	2695.92	TRUE
2	0	FALSE	35	28475.30	1537.67	4.61	FALSE
2	1	FALSE	36	29909.55	1884.30	7.54	FALSE
2	2	FALSE	37	31344.48	1567.22	4.70	FALSE
2	3	FALSE	38	32725.83	818.15	1.64	FALSE

2	4	FALSE	39	34097.32	1397.99	4.19	FALSE

3- Some Manipulation :

a) Is there any transformation on the available variables that provides useful information

Yes, we can calculate the saving rate as the ratio between total savings and the yearly income in order to know how much this rate fluctuates between two dates and how much of the client's income is saved each year.

$$\text{Saving\_rate} = \frac{\text{savings}}{\text{yincome}}$$

We can also make a retirement rate to compare retirement saving behaviour across clients with different income levels.

$$\text{retirement\_rate} = \frac{\text{retirement}}{\text{yincome}}$$

Taking the Logarithm of the income, savings and retirement variables will help us to linearize relationships and reduce the influence of extreme values.

$$\text{log\_savings} = \log(\text{savings} + 1); \text{log\_yincome} = \log(\text{yincome} + 1)$$

b) Create a separate dataset including rows from 100 to 200 of some variables you want to experiment with and carry out the transformations.
Answered in the data code

c) Now check the summary statistics only for these newly created variables: do they significantly change from the original statistics?

Statistic	log_yincom	log_savings	log_retirement	log_saving_rate	log_retirement_rate
Min	10.24	7.119	0.000	0.6896	0.0000
1st Qu	10.66	8.646	2.249	0.8053	0.2157
Median	10.82	8.920	5.234	0.8317	0.4865
Mean	10.80	8.889	4.544	0.8225	0.4190
3rd Qu	10.95	9.421	6.620	0.8574	0.6011
Max	11.23	9.910	7.982	0.8821	0.7244

Ratios saving rate and retirement rate are normalized by income, meaning their values typically range between 0 and 1. This normalization greatly reduces variance compared to absolute amounts which directly depend on income levels.

Logarithmic variable transformation mitigates the impact of extreme values by compressing large observations. The distribution becomes more symmetric making statistics such as the mean and the standard deviation more stable.

As an example, le mean of y income before any transformation is around 50,000 corresponding to a log value of 10.8
It transforms the variable from an absolute scale to a relative one, compressing large differences between high-income individuals and reducing skewness in the distribution.

d) Go back to the original dataset: is there a variable that marks the administration of a policy intervention, an exogenous decision, or a threshold? If yes find or create a dummy variable that takes 1 when that unit is affected
The dataset includes a variable that captures an exogenous policy intervention that is called "Meeting". This dummy variable equals 1 for clients who were assigned to the follow-up meeting with a salesperson and 0 otherwise.

e) Provide the share of dummy variables taking 1 and 0 for the cases just mentioned

Meeting	Share proportion
TRUE = 29530	40,94%
FALSE =20470	59,06%

ANNEXES:

A quick description of our data :

```
> describe(groupe_41)
           vars     n     mean       sd   median  trimmed      mad      min       max     range skew kurtosis    se
id            1 50000  5000.50  2886.78  5000.50  5000.50  3706.50     1.00  10000.00   9999.00 0.00    -1.20 12.91
time          2 50000     2.00     1.41     2.00     2.00     1.48     0.00      4.00      4.00 0.00    -1.30  0.01
female        3 50000      NaN       NA       NA      NaN       NA      Inf      -Inf      -Inf   NA       NA    NA
age           4 50000    52.52    19.26    52.00    52.52    25.20    18.00     87.00     69.00 0.01    -1.20  0.09
yincome       5 50000 51999.09 15609.98 51126.76 51304.65 15474.14 13296.67 130031.58 116734.91 0.48     0.36 69.81
savings       6 50000 10100.38  7426.61  8529.63  9147.58  6254.00     0.00  66966.26  66966.26 1.57     4.12 33.21
retirement    7 50000   644.47  1112.16   176.68   385.33   261.94     0.00  14326.75  14326.75 3.21    15.09  4.97
meeting       8 50000      NaN       NA       NA      NaN       NA      Inf      -Inf      -Inf   NA       NA    NA
```

The names of the initial variables:

```
> names(groupe_41)
[1] "id"         "time"       "female"     "age"        "yincome"    "savings"    "retirement" "meeting"
```

Summary of the new variables:

```
 log_yincome     log_savings    log_retirement  log_saving_rate  log_retirement_rate
Min.   :10.24   Min.   :7.119   Min.   :0.000   Min.   :0.6896   Min.   :0.0000
1st Qu.:10.66   1st Qu.:8.646   1st Qu.:2.249   1st Qu.:0.8053   1st Qu.:0.2157
Median :10.82   Median :8.920   Median :5.234   Median :0.8317   Median :0.4865
Mean   :10.80   Mean   :8.889   Mean   :4.544   Mean   :0.8225   Mean   :0.4190
3rd Qu.:10.95   3rd Qu.:9.421   3rd Qu.:6.620   3rd Qu.:0.8574   3rd Qu.:0.6011
Max.   :11.23   Max.   :9.910   Max.   :7.982   Max.   :0.8821   Max.   :0.7244
```

Summary of the all the variables:

```
> summary(subset_data)
       id          yincome         savings        retirement       meeting         saving_rate
 Min.   :20.0   Min.   :28072   Min.   : 1235   Min.   :   0.000   Mode :logical   Min.   :0.0390
 1st Qu.:25.0   1st Qu.:42493   1st Qu.: 5687   1st Qu.:   9.483   FALSE:61        1st Qu.:0.1240
 Median :30.0   Median :50068   Median : 7480   Median : 187.526   TRUE :40        Median :0.1637
 Mean   :30.4   Mean   :50060   Mean   : 8707   Mean   : 569.219                   Mean   :0.1621
 3rd Qu.:35.0   3rd Qu.:57166   3rd Qu.:12347   3rd Qu.: 750.098                   3rd Qu.:0.2080
 Max.   :40.0   Max.   :75684   Max.   :20132   Max.   :2928.752                   Max.   :0.2660
 retirement_rate     log_yincome     log_savings     log_retirement  log_saving_rate  log_retirement_rate
 Min.   :0.000000   Min.   :10.24   Min.   :7.119   Min.   :0.000   Min.   :0.6896   Min.   :0.0000
 1st Qu.:0.000280   1st Qu.:10.66   1st Qu.:8.646   1st Qu.:2.249   1st Qu.:0.8053   1st Qu.:0.2157
 Median :0.003888   Median :10.82   Median :8.920   Median :5.234   Median :0.8317   Median :0.4865
 Mean   :0.010415   Mean   :10.80   Mean   :8.889   Mean   :4.544   Mean   :0.8225   Mean   :0.4190
 3rd Qu.:0.012528   3rd Qu.:10.95   3rd Qu.:9.421   3rd Qu.:6.620   3rd Qu.:0.8574   3rd Qu.:0.6011
 Max.   :0.047960   Max.   :11.23   Max.   :9.910   Max.   :7.982   Max.   :0.8821   Max.   :0.7244
```

Table of the dummy variable:

```
> table(groupe_41$meeting)

FALSE  TRUE
29530 20470
> prop.table(table(groupe_41$meeting))

 FALSE   TRUE
0.5906 0.4094
```

Script R du Projet

```r
### we call all the library needed
library(dplyr)

### loading the data set from a specified path in groupe_41
groupe_41 <- read.csv("C:/Users/ibrah/OneDrive/Documents/ cole /S7/econometrie appliquee/projet/group41.csv")

### some simple command to inspect the data set
head(groupe_41)
dim(groupe_41)
summary(groupe_41)


####### importing the data from line 100 to 200 and only the columns specified ("id", "yincome", "savings", "retirement", "meeting") into a subset named subset_data
subset_data = groupe_41 [100:200 , c("id", "yincome", "savings", "retirement", "meeting")]
head(subset_data)

## simple rate metrics without a transformation of the data
subset_data$saving_rate = subset_data$savings / subset_data$yincome
subset_data$retirement_rate = subset_data$retirement / subset_data$yincome

## logarithmic transformation on the data
subset_data$log_yincome = ifelse(subset_data$yincome == 0,log(subset_data$yincome + 1), log(subset_data$yincome))
subset_data$log_savings = ifelse(subset_data$savings == 0,log(subset_data$savings + 1), log(subset_data$savings))
subset_data$log_retirement = ifelse(subset_data$retirement == 0,log(subset_data$retirement + 1), log(subset_data$retirement))

## logarithmic rates
subset_data$log_saving_rate = subset_data$log_savings / subset_data$log_yincome
subset_data$log_retirement_rate = subset_data$log_retirement / subset_data$log_yincome
subset_data

### summary of the subset_data only on the new created variables ("log_yincome", "log_savings", "log_retirement", "log_saving_rate", "log_retirement_rate")

summary(subset_data[, c("log_yincome", "log_savings", "log_retirement", "log_saving_rate", "log_retirement_rate")])

### dummy share of the exogenous variable
table(groupe_41$meeting)
prop.table(table(groupe_41$meeting)))
```