UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Applied Econometrics
R Project – Case 1 – Part 02

Ibrahim FOUSFOS

Contents

Overview 1

1) Data Import And Summary by Treatment Group 1

2) Creation Of New Variables And Data Transformations 1

3) Distribution Of Outcome Variables with Histograms and Interpretations 2

4) Scatter Plot Analysis Between Income And Savings 3

5) Creation Of A Dummy Variable And Probit Regression Analysis 4

6) OLS Estimation And Correlation Analysis 5

7) Extended OLS Regression With Covariates 7

8) Economic Interpretation And Discussion Of Results 7

Sources **8**

Overview

1) Data Import And Summary by Treatment Group

We choose Savings and retirement as our two dependent variables (Y) because they directly capture the client's saving behaviour which is the core focus of our research question.

The purpose of the study is to determine whether the meeting with a salesperson influences how much clients save in general and more specially how much they contribute to their retirement savings.

- **Savings** → Reflects the overall saving behaviour of clients. It shows whether the meeting encourages people to save more money in general.
- **retirement** → Focuses on long term financial planning and measures whether the meeting motivates clients to invest more in their retirement plans.

Together, these two variables allow us to evaluate both short term and long term saving decisions giving a more complete view of the meeting impact.

Table 1: Summary statistics by treatment group

		Retirement			Savings		
meeting	n	Mean	SD	Median	Mean	SD	Median
Control	29530	214.89	388.77	75.38	9261.31	7024.26	7780.96
Treated	20470	1264.18	1467.31	807.12	11310.81	7814.36	9641.03

2) Creation Of New Variables And Data Transformations

The variables `log_income`, `log_savings` and `log_retirement` help linearize relationships and reduce the effect of extreme values, allowing percentage-based interpretations.

We created:

- **`log_saving_rate`** → measures the proportion of income saved on a logarithmic scale. It helps understand saving intensity relative to income and compare clients with different income levels.
- **`log_retirement_rate`** → measures how much of a client's income goes to retirement savings on a logarithmic scale. It captures long-term saving effort and is less sensitive to income differences.

3) Distribution Of Outcome Variables with Histograms and Interpretations

We focus on the distribution of post-treatment savings (periods $t \geq 2$) at the individual level. The first histogram shows that savings are positively skewed, with most clients saving relatively small amounts and a few saving much more.

When we split the data by the treatment group (meeting), both distributions have a similar shape, but the clients who attended the meeting (in blue) tend to have slightly higher savings on average.

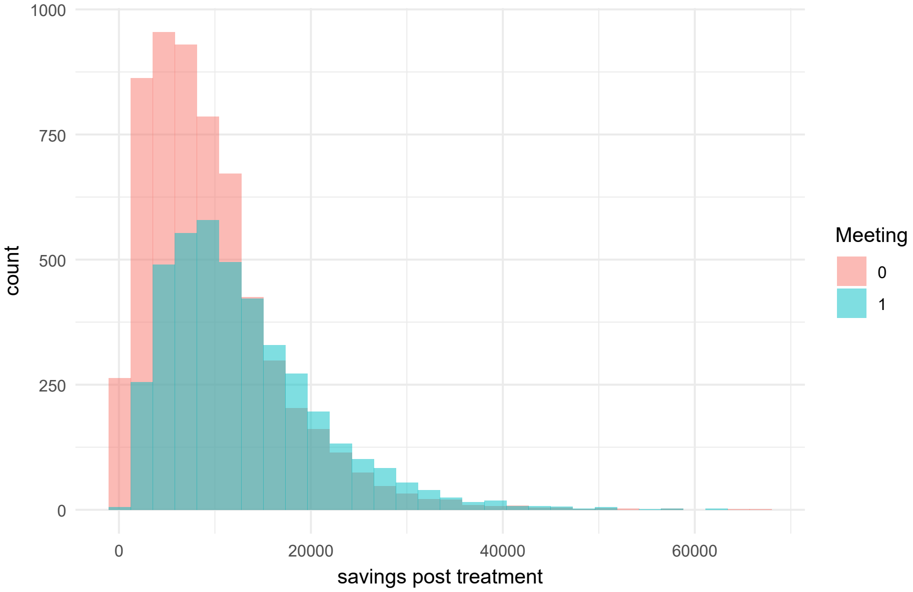

Figure 1: Distributions by treatment

When weighting the data to give each group the same total mass, the shape of the distribution becomes clearer and comparable across groups. The weighted histogram shows that although the two groups have similar savings patterns, the treated group maintains a small upward shift in saving levels.

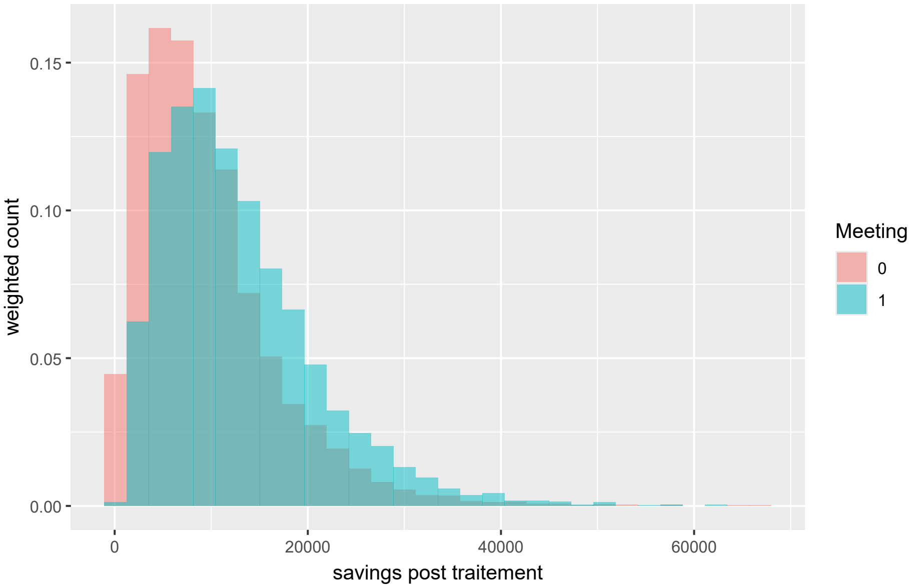

Figure 2: Distributions by treatment with weights

4) Scatter Plot Analysis Between Income And Savings

The scatter plot of *log savings* against *log income* shows a clear positive relationship, meaning that clients with higher income tend to save more. The relationship is economically consistent with the life-cycle theory (Modigliani and Brumberg, 1957) and the permanent income theory (Friedman, 1957). According to these theories, individuals aim to smooth consumption over time and therefore save part of their income during high-earning periods to prepare for future needs or retirement.

In the plot, the fitted trend shows that as income increases, savings also rise proportionally. The log–log specification highlights a nearly linear relationship, suggesting that the elasticity of savings with respect to income is relatively stable.

When comparing groups, clients who attended the meeting (in blue) generally lie slightly above the untreated group, showing higher savings for a given income level. This suggests that the meeting may have reinforced financial awareness or encouraged better saving habits. The difference between groups is more visible among lower-income clients, who appear to benefit more from the intervention.

Overall, the data support the theoretical prediction that income positively influences savings, while also indicating a small behavioral effect of the meeting.

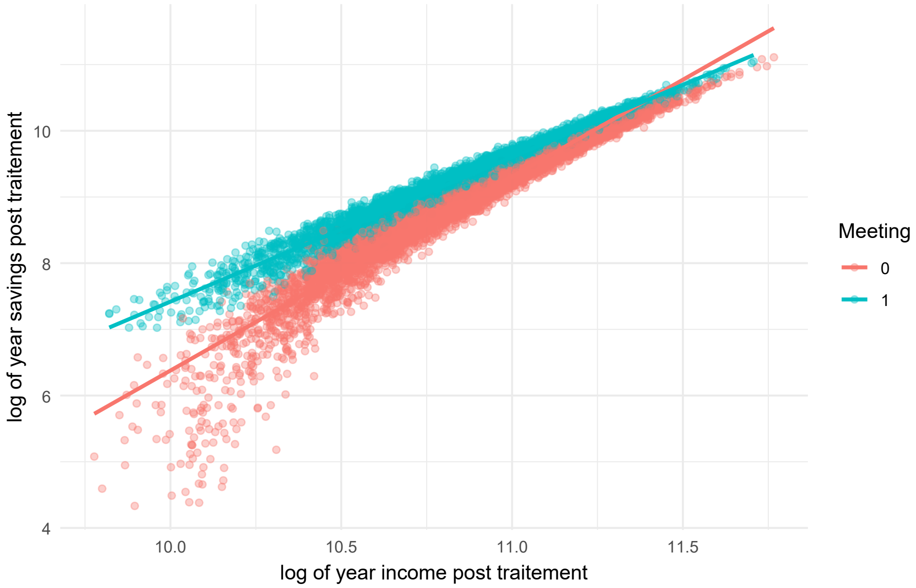

Figure 3: Income–Savings Relationship (log–log), post-treatment period

5) Creation Of A Dummy Variable And Probit Regression Analysis

To test whether the treatment (`meeting` = 1) was randomly assigned, we compared the pre-treatment characteristics of treated and control clients. We estimate a probit model to assess whether treatment assignment was random with respect to pre-treatment characteristics such as income, savings, retirement savings, and gender. From the probit estimation, we compute the *average marginal effects* (AME), which measure how a one–unit change in each explanatory variable affects the probability of being treated on average. The AME are all close to zero and statistically insignificant (all $p > 0.10$), indicating that the pre-treatment variables have no meaningful effect on the probability of being treated.

Variable	Estimate	Std. Error	z value	$\Pr(>\lvert z\rvert)$
(Intercept)	-0.22	0.10	-2.06	0.04
inc__pre	-0.00	0.00	-0.19	0.85
sav__pre	0.00	0.00	0.53	0.59
ret__pre	-0.00	0.00	-1.06	0.29
femaleTRUE	-0.02	0.03	-0.78	0.43

Table 2: Probit coefficients

The average income before the meeting is 49,864 for the control group and 50,053 for the treated group, with a difference of only 189 €; the associated p-value equals 0.55, which indicates that this difference is not statistically significant, meaning that the two groups had similar income levels before the intervention.

Table 3: Balance table: pre-treatment characteristics by treatment status

Variable	Control mean	Treated mean	Diff (1-0)
Pre-income (€)	49,863.97	50,052.83	188.86
Pre-savings (€)	8,942.85	9,025.15	82.29
Pre-retirement (€)	207.83	205.23	-2.61
Female (=1)	52.7%	51.9%	-0.8%

Mean savings are 8,943 € and 9,025 €, with a difference of 82 € ($p = 0.55$), and mean retirement savings are 208 € and 205 € with a difference of -2.6 € ($p = 0.73$). Gender composition is also balanced with 52.7% women in the control group compared to 51.9% in the treated group (χ^2 test, $p = 0.44$).

Variable	Mean..0.	Mean..1.	Diff..1.0.	t	p.value
Pre-income (€)	49863.97	50052.83	188.86	-0.60	0.55
Pre-savings (€)	8942.85	9025.15	82.29	-0.59	0.55
Pre-retirement (€)	207.83	205.23	-2.61	0.35	0.73

Table 4: Equality-of-means tests (pre-treatment variables by treatment status)

Variable	Prop (0)	Prop (1)	Chi-square	df	p-value
Female (=1)	52.7%	51.9%	0.59	1	0.44

Table 5: Chi-square test for equality of proportions (Female)

None of these differences is statistically significant (all $p > 0.10$), meaning that before the policy intervention both groups had comparable income, saving behaviour, and gender composition. These results are consistent with random assignment of the meeting, so any differences observed after the meeting can plausibly be attributed to the effect of the meeting itself.

6) OLS Estimation And Correlation Analysis

The graph of the mean savings (post) by `meeting` displays the mean *post-treatment* savings for clients who had a meeting (1) and those who did not (0), with 95% confidence intervals. It shows a clear upward shift for the treated group: on average, clients who attended the meeting saved substantially more after the intervention.

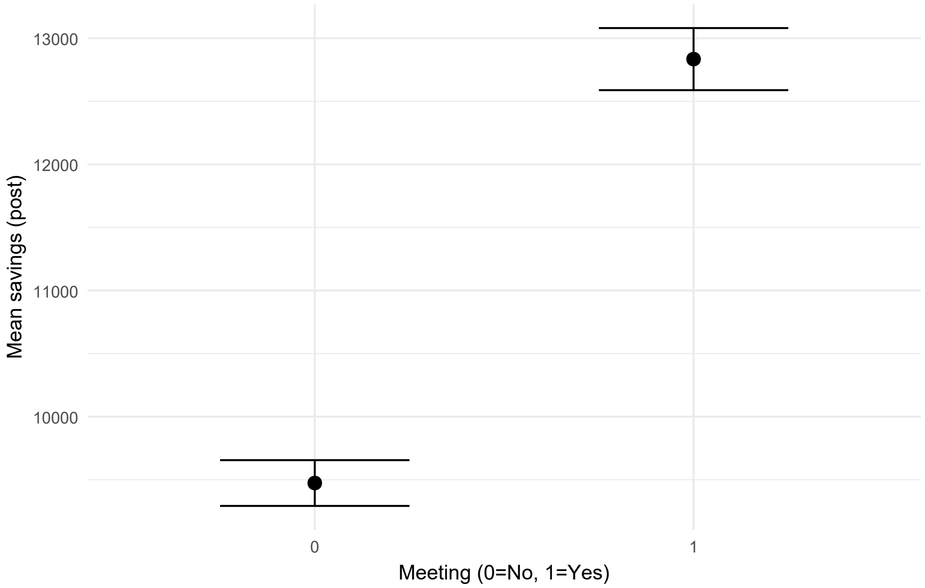

Figure 4: Mean post-treatment savings by meeting participation with 95% confidence interval

To assess the correlation between the policy intervention and saving behaviour, we estimate a simple OLS regression of post-treatment savings (`Y_post`) on the dummy variable (`meeting`).

Table 6: OLS: $\mathrm{Y}_{p}ost\ meeting$

	Dependent variable:
	$\mathrm{Y}_{p}ost$
Meeting	3,360.963***
	(152.629)
Constant	9,473.620***
	(97.659)
Observations	10,000
R^2	0.046
Adjusted R^2	0.046
F Statistic	484.903*** (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression results show a positive and highly significant coefficient for `meeting` ($\hat{\beta} = 3{,}361$, $p < 0.001$). This means that, on average, clients who attended the meeting saved about 3,361 € more after the intervention than those who did not. The coefficient is statistically significant at the 1% level, indicating strong evidence of a positive association. The intercept (9,474) represents the average post-treatment savings for the control group (clients without a meeting).

The R^2 of 0.046 suggests that the meeting explains about 4.6% of the variation in savings, meaning that while the meeting has a significant effect, most of the variation is still due to other factors.

Overall, the results indicate that the meeting has a positive and statistically significant correlation with post-treatment savings, suggesting that the intervention likely encouraged higher saving behaviour.

7) Extended OLS Regression With Covariates

To understand the relationship between the meeting intervention and post-treatment savings, we estimate an extended OLS regression including several relevant covariates that were selected as:

- **Pre-income** controls for clients' financial capacity, since higher-income individuals typically save more.
- **Pre-savings** capture baseline saving behaviour, ensuring we compare clients with similar prior habits.

In our R code, we first defined the outcome (*savings*) and its log version to allow both level and percentage interpretations. The post-treatment (`Y_post`) was computed for each client ($t \geq 2$), while pre-treatment averages (`inc_pre`, `Y_pre`) served as covariates to control for income capacity and prior saving habits.

The results shown in the table indicate a positive and highly significant coefficient for `meeting` (treated) across both specifications. In the level model, treated clients save 3,275 € more on average than those without a meeting. In the log specification, the coefficient of 0.419 suggests an increase of approximately 42% in post-savings for treated clients, controlling for income, past savings and gender.

Both pre-income and pre-savings are also strong predictors of post-treatment savings, as expected. The models show excellent fit with $R^2 = 0.985$ for levels and $R^2 = 0.944$ for logs, suggesting that most variation in post-savings is explained by these covariates.

8) Economic Interpretation And Discussion Of Results

From an economic perspective, our results are consistent with our initial expectations : clients who attended a meeting with a salesperson have significantly higher saving levels, both in total and in retirement savings. The positive and robust coefficient of the meeting variable across all models supports the hypothesis that follow-up meetings encourage stronger behaviour.

This finding aligns with the life cycle theory of Modigliani and Brumberg (1954) which suggests that individuals seek to smooth consumption and increase savings when their income or financial awareness rises. The meeting likely acted as a behavioral nudge improving financial planning and promoting long-term saving decisions.

Our extended OLS regressions confirm that this effect remains significant even after controlling for pro-income and pre-savings, meaning that the relationship is not driven by initial income differences or prior saving habits. The earlier randomness checks (T-tests and probit model) also show that the treatment was assigned independently of observable characteristics, reducing selection bias and reinforcing the internal validity of the analysis.

However, while the estimated relationship is statistically strong and economically meaningful, it is incorrect to say that it is considered fully causal without accounting for potential unobserved heterogeneity, for instance differences in motivation or financial literacy that might also influence saving behaviour.

Overall, the results are in line with theoretical expectations and indicate that the meeting intervention likely had a positive and substantial impact on the client's saving behaviour. Yet, the relationship should be interpreted as a robust correlation with plausible causal direction, rather than definitive proof of causality.

Sources

lien régression scatter

- **Applied Econometrics lecture slides :**
  - https://cours.univ-paris1.fr/course/view.php?id=44026
- **Life-Cycle Hypothesis developed by Modigliani in 1954 :**
  - https://www.lafinancepourtous.com/decryptages/finance-perso/epargne-et-placement/epargne/la-theorie-du-cycle-de-vie/
- **The Table function for making a table with R studio :**
  - https://bookdown.org/yihui/rmarkdown-cookbook/kable.html
- **To know how to use LaTeX :**
  - https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes
  - https://latex-tutorial.com/tutorials/pgfplotstable/
- **To know how to make a Histogram with ggplot :**
  - https://www.datacamp.com/tutorial/make-histogram-basic-r
  - https://r-graph-gallery.com/histogram_several_group.html
- **To know how to scatter :**
  - https://www.datacamp.com/doc/r/scatterplot-in-r

  - https://www.r-bloggers.com/2020/07/create-a-scatter-plot-with-ggplot/

- **To know how to make a regression :**
  - https://www.datacamp.com/tutorial/linear-regression-R

- **To know how to make an OLS regression :**
  - https://www.r-bloggers.com/2017/07/ordinary-least-squares-ols-linear-regression-in-r/
  - https://www.geeksforgeeks.org/r-machine-learning/ordinary-least-squares-ols-regressio
  - https://cran.r-project.org/web/packages/olsrr/vignettes/intro.html