Multivariate regressions in R Markdown

Samy Zitouni

December 10, 2024

Instructions

The goal of this exercise if to perform some econometrics analysis on some data. Using what we have seen throughout this tutorial, you will present your results in an R Markdown document. Fell free to ask questions if you are lost.

We are interested in the data from an old inquiry about jobs in France, conducted by INSEE. The sample may not be representative of France.

1. Go to *https://samyztn.github.io/* and download the data named *data_tuto5.csv*. The file contains data about some french indivuals' salaries. We are particularly interested in the effect of age and the level of diplomas on salaries. Below is a table that explains the code for the variable **DDIPL**, representing the level of diploma, and for SEXE.

Code	Diploma
1	Bac +5 and above
3	Bac +3 (Licence)
4	Bac +2 (BTS, equivalent, L2, ...)
5	Bac
7	Under the Bac level

Table 1: Level of diploma for the code of DDIPL

Code	Diploma
1	Male
2	Woman

Table 2: Code for variable SEXE

2. Open a new R Markdown file (*File + Open + R Markdown*). Fill the information in the opening pop-up. Once the Markdown is created, remove any text except the header (between two sets of $---$). and the first code chunk, that define that by default, the code displays, unless you set *echo = False* in you code chunk. Change this option for echo to be always FALSE unless you specify echo = TRUE.

3. For this work, we will use different libraries. Please load them in a code chunk that does not show, neither show results. We don't want the report being polluted. You can quickly open a code chunk with *Ctrl + Alt + I*. Load the following libraries:

   - readxl (you may install it before, in the command panel with *install.packages('readxl')*)
   - ggplot2
   - dplyr
   - knitr (you may need to install)
   - psych (you may need to install)

- tidyr (you may need to install)

4. In a new code chunk, load the data and call *kable(head(data), caption = "First rows of the data").* You may add text above defining a section with *### Data*, and some text to introduce.

5. Plot a simple scatter plot of earnings as a function of the age (SALRED, AGE). What do you notice ?

6. Just for the plot, remove the first and last centiles of data. Then, plot a grid with the same graph for every diploma. How do you feel the relation between SALRED and AGE will be ?

7. Now, for the whole dataset, plot $\log SALRED$ as a function of AGE (you may need to create the variable logSAL). Then, plot a facet for every diploma, adding the line *geom_smooth(method = lm, formula = y x, color = 'darkred')* to fit a linear regression line. How well does it perform ?

8. Perform the regression of logSAL over AGE. Print the summary, and interpret the coefficients. Interpret the value of R^2.

9. Plot the residuals' density.

10. Plot a QQ plot with you residuals. What are the conclusion of the two last questions ?

11. Now, we are interested in adding some controls to the regression. Print summaries for the three following regressions:

$$logSAL = \beta_0 + \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \epsilon \quad (1)$$

$$logSAL = \beta_0 + \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \beta_4 AGE^2 + \epsilon \quad (2)$$

What do you remark about the coefficients before AGE. Could you anticipate the sign before the coefficient AGE^2

12. Now, we want to look at different interactions. The first is whether an additional year is more beneficial to men than to women.

$$logSAL = \beta_0 + \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \beta_3 (Male \times AGE) + \epsilon \quad (3)$$

Intepret the results !

13. The second one is the difference for diplomas:

$$logSAL = \beta_0 + \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \beta_3 (Male \times DDIPL) + \epsilon \quad (4)$$

Interpret the results. Compare with the results for a regression on every different level of diploma.

14. Now, let's have a look at co-linearity. In the dataframe, create a variable Female that is equal to 1 when the individual is a female. Run :

$$logSAL = \beta_0 + \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \beta_4 Female + \epsilon$$

, then run

$$logSAL = \beta_1 Male + \beta_2 AGE + \beta_3 DDIPL + \beta_4 Female + \epsilon$$

See what happens. What is the interest of doing one on top of another ? What problems could you meet ?

15. Run a regression with as much controls you can include ! See what happens.