Ensimag – Grenoble INP – UGA
Modélisation statistique et analyse de données
Christophe Dutang
Olivier François

Année 2023-2024

christophe.dutang@grenoble-inp.fr
olivier.francois@univ-grenoble-alpes.fr

⚠ General guidelines for TPs

Each team shall upload its report on Teide before the deadline indicated at the course website. Please **include the name of all members of the team** on top of your report. The report should contain graphical representations. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the **questions in order and refer to the question number in your report**. Computations and graphics have to be performed with `R`.

The report should be written using the `Rmarkdown` format. **PLEASE USE THE TEMPLATE ON CHAMILO**. In Teide, you are asked to submit both the `rmd` and the `html` files. In the `html` file, you should limit the displayed `R` code to the most important instructions.

TP: analyzing math score in PISA studies

PISA is the OECD's Programme for International Student Assessment. PISA measures 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges for 85 countries.

We want to study the relation between score in mathematics, in science, in reading and in social-economic group based on 2018 data. We use the dataset `PISA2018subset.csv` available on chamilo. `PISA2018subset.csv` contains a sample student subset containing scores and other information from the triennial testing of 15 year olds around the globe. Variable to be used are described below

- `year`: Year of the PISA data. Factor.
- `country`: Country 3 character code. Note that some regions/territories are coded as country for ease of input. Factor.
- `gender`: Gender of the student. Only `"male"` and `"female"` are recorded. Factor.
- `math`: Simulated score in mathematics. Numeric.
- `read`: Simulated score in reading. Numeric.
- `science`: Simulated score in science. Numeric.
- `escs`: Index of economic, social and cultural status. Numeric.

Load dataset via

```
> PISA2018 <- read.csv("data/PISA2018subset.csv", stringsAsFactors = TRUE)
> FR2018 <- subset(PISA2018, country == "FRA")
```

We first look at direct effects for `FR2018`.

1. Visualize the joint distribution of `math`, `read`, `science`. Commands to use: `pairs()`, `plot()`.
2. Is the score in `math` explained by the score in `read`?
   Start with `cor()` without fitting a linear model. Then fit a linear model `lm()` and interpret the regression with `summary(lm())`.
3. Same question for the score in `science`.
4. Do preceding results mean the causality between `read` and other scores?

Now we look at other effects for `FR2018`.

5. We want to adjust the estimation against the socio-eonomic group `escs` of the family. Perform the regression of `math` against `escs`. Display data and the regression line and discuss the part of the explained variance.
6. Is the score in `read` linked to `escs`? Perform the regression of `read` against `escs`. Display the regression line and discuss the part of the explained variance.
7. Now we look at crossed effects. Perform the full regression of `math` against both `read` and `escs` and comment.

Finally, we investigate the gender and the country effect.

8. Compare the result in `math` between girls and boys for `FR2018`. Commands to use: `boxplot(math ~ gender, data=FR2018)`, `t.test`, `shapiro.test`.
9. Same question for the score in `read` for `FR2018`.
10. Do conclusions remain identical for other countries? Make previous analysis on `PISA2018`.