amU FEG Aix Marseille Université

amse école d'économie d'aix-marseille aix-marseille school of economics

Introduction to programming for data analysis
Teacher: U. Aiounou, E. Gallic, K. Kouadio[1]

Master in Economics
1st year, Semester 1

Tutorial #3: Linear regressions and statistical tests

This tutorial focuses on regression models and statistical hypothesis testing. After completing Tutorial #2, you should be able to create basic data visualizations and compute summary statistics. Building on these skills, you are now ready to run regression analyses and apply statistical modeling techniques.
In Tutorial #1, you downloaded datasets from Eurostat, merged them into a single dataset in R, and exported the result as a CSV file.

Objectives of the tutorial

In this tutorial, you will use the dataset from exercise 2 of Tutorial #1 to:

1. create visualizations to examine the relationship between the two main variables (per capita GDP and overall life satisfaction),
2. estimate regression models, including both linear and nonlinear specifications for panel data.

Q1. Import the dataset that you saved at the end of exercise 2 of Tutorial #1.

```
library(tidyverse)
library(stargazer)
tb <- ...
```

Scatterplots

A scatterplot visually represents the relationship between two variables. In the current dataset, it is useful to examine how per capita GDP relates to overall life satisfaction.

GDP and life satisfaction

Q2. Create a scatterplot to display the relationship between per capita GDP and overall life satisfaction. Label the x-axis "Per capita GDP" and the y-axis "Overall Life Satisfaction". Use the title "Per capita GDP vs. Overall Life Satisfaction" for the graph.

Q3. Add a linear regression line to the scatterplot.

Q4. Add a quadratic curve to the scatterplot.

Changes in GDP and life satisfaction

Q5. Create two new columns, `gdp_yoy` and `life_satisf_yoy`, which represent the within-country year-on-year changes in GDP and life satisfaction, respectively.

Q6. Create a scatterplot to illustrate the relationship between `gdp_yoy` and `life_satisf_yoy`.

[1] dimitri.aiounou[at]univ-amu.fr, ewen.gallic[at]univ-amu.fr, kla.KOUADIO[at]univ-amu.fr

- Place the year-on-year variation in life satisfaction on the y-axis and label it "Yearly Change in Life Satisfaction".
- Place the year-on-year variation in per capita GDP on the x-axis and label it "Yearly Change in Per Capita GDP".
- Title the graph "Relationship Between Changes in GDP and Life Satisfaction".
- Add a linear regression line to the scatterplot.

OLS Regression

Q7. Examine and evaluate the following instructions. Then, explain what they do.

```r
model_lm <- lm(life_satisf ~ gdp, data = tb)
summary(model_lm)
```

Q8. Examine and evaluate the following instructions. Then, explain what they do.

```r
stargazer(model_lm, type = "latex", out = "tables/model_lm.tex")
```

Fixed effects regression

Q9. Extend the previous regression model by including country fixed effects to account for unobserved, time-invariant differences across countries.

```r
model_country_fe <- ...
summary(model_country_fe)
```

Q10. Extend the country fixed-effects model by adding year fixed effects to control for common shocks across all countries in a given year.

```r
model_country_time_fe <- ...
summary(model_country_time_fe)
```

Q11. Examine and evaluate the following instructions. Then, explain what they do.

```r
stargazer(
    model_lm, model_country_fe, model_country_time_fe,
    type = "latex",
    out = "tables/regression_results.tex",
    title = "Correlation between per capita GDP and life satisfaction",
    column.labels = c("Baseline OLS Reg.", "Country FE","Country and Year FE"),
    covariate.labels = c("per capita GDP"),
    add.lines = list(
        c('Country FE', '-', 'Yes', 'Yes'),
        c('Year FE', '-', '-', 'Yes')
    )
)
```

Non linear regression

We herein consider non linear regression models with a dummy as dependent variable, such as probit and logit regression models.

Q12. Compute the median of life satisfaction.

```
median_life_satisf <- ...
```

Q13. In the dataset, create a binary variable named `life_satisf_above_med`, which takes the value 1 if life satisfaction is above the median and 0 otherwise.

```
tb <- ...
```

Q14. Run a **probit** regression of `life_satisf_above_med` on `gdp`, and display the summary of the results.

```
model_probit <- ...
summary(model_probit)
```

Q15. Run a **logit** regression of `life_satisf_above_med` on `gdp`, and display the summary of the results.

```
model_logit <- ...
summary(model_logit)
```

Statistical tests

Q16. Compute the median of per capita GDP.

```
median_gdp <- ...
```

Q17. In the dataset, create a binary variable named `gdp_above_med`, which takes the value 1 if per capita GDP is above the median and 0 otherwise.

```
tb <- ...
```

Q18. Examine and evaluate the following instructions. Then, explain what they do.

```
group1 <- tb |>
    filter(gdp_above_med == 0) |>
    pull(life_satisf)

group2 <- tb |>
    filter(gdp_above_med == 1) |>
    pull(life_satisf)
```

Q19. Perform a Student's t-test to compare the mean life satisfaction between observations with GDP above the median and those with GDP at or below the median.

```
t_test_result <- ...
print(t_test_result)
```

Q20. Extract the test statistic and the p-value from the results of the t-test.

```
test_statistic <- ...
p_value <- ...
```

Q21. Examine and evaluate the following instructions. Then, explain what they do.

```
latex_table <- paste0(
    "\\begin{tabular}{l c}\n",
    "\\hline\n",
    "& Difference in life satisfaction across groups" ,
    " \\\\ \n",
    "Test Statistic & ",
    format(test_statistic, digits = 3),
    " \\\\ \n",
    "P-value & ",
    format.pval(p_value, digits = 3),
    " \\\\ \n",
    "\\hline\n",
    "\\end{tabular}"
)
file_path <- "tables/table_results_t_test_1_v1.tex"
writeLines(latex_table, file_path)
```