ESL DATA ANALYSIS BOOT CAMP 2023/2024

– Final Assignment –

Instructors: Aurélie Bertrand, Alain Guillet, Thomas Eisfeld

Deadline: December 20, 2023 at midnight. To be submitted via Moodle

Instructions and information

The goal is to create a document with **LaTeX (mandatory for ECON2MA / ETRI2MA / QEM2 students)** or **Microsoft Word (ECON2M1 / ECON2MS/G / ECON2MD students only)** containing answers to the questions. Both the data management and analysis **must be performed using R or Stata,** not Excel.

Please submit the following files to Moodle:

- the `.tex` or `.docx` file containing your answers
- the `.pdf` file created from the `.tex` or `.docx` file
- the do-file or the R script with all the code that you wrote in order to answer the different questions (please write comments and do not keep any errors in it). Do not forget to specify your working directory!

The assignment's grade is either pass or fail. All files should be submitted the latest by **December 20, 2023, at midnight via Moodle**. Every submission after the stated deadline will be considered as failed. If you have any questions concerning the assignment, please contact Thomas Eisfeld (`thomas.eisfeld@uclouvain.be`).

Data from Eurostat and the World Bank

This assignment will be based on data from the World Bank and Eurostat. In particular, please download the following four data sets:

- Unemployment, total (% of total labor force) from the World Bank: `API_SL.UEM.TOTL.ZS_DS2_en_csv_v2_5994651.csv`
- Inflation, consumer prices (annual %) from the World Bank: `API_FP.CPI.TOTL.ZG_DS2_en_csv_v2_5994714.csv`
- GDP and main components (output, expenditure and income) from Eurostat: `nama_10_gdp_page_spreadsheet.xlsx`
- List of countries in the EU (as of 2019) from Eurostat: `Country_Codes_and_Names.xlsx`

Please download the associated `.csv` and Excel files. You can access each data set by clicking on its title above. Further information about the data can be found on their associated website and in the files themselves.

1. Importing and cleaning the data

Create a data set with EU countries. The aim of this first step is to create a data set comprising a list of EU members so that you can filter other data sets later on to only contain EU membership countries. To do so, first download and import the data set "List of countries in the EU".

1. Make all variables as lowercase and keep only countries that are EU members. Make sure to also drop the United Kingdom as it not part of the EU.
2. Replace the country names of "Czech Republic" by "Czechia" and "Germany (including former GDR from 1991)" by "Germany" so that you can merge the data set with the other data later on.
3. Save the date set "EU_members.dta" in Stata or "EU_members.RData" in R.

Import and reshape panel data. Next, you have to import the data sets comprising panel data on unemployment, inflation, and GDP. To do so, download the other listed data sets.

4. After downloading them, do the following for each of the three data sets:
   (a) Import the data set.
   (b) Rename the variables so that each indicator is named in the following way: `indicator_year` (e.g., inflation in 2012 would be named `inflation_2012`).
   (c) Make sure all numeric variables are also saved as numeric variables. This might imply some cleaning (e.g., missing values are coded as ":" in the Eurostat data, so you will need to replace such values as missings before destringing them). Also make sure that "Slovak Republic" is called "Slovakia". (Hint: if you are working with R, you can use the following command for the first part of this question: `variable <- as.numeric(sub(",", ".", variable))`.)
   (d) Merge the "EU_members" data to each of the three data sets. Reduce the data to only include EU members. Make sure that all EU countries are in your data (otherwise there might be inconsistency issues across data sets).
   (e) Reshape the data set to be in the long format. Each of your final data sets will have three variables: Country, year, and the respective indicator (unemployment, inflation, or GDP). You can drop all other variables as you will not need them in your analysis later on.
   (f) Save the data set named according to the respective indicator (i.e., as "Unemployment", "Inflation", and "GDP" in `.dta` (Stata) or `RData` format (R)).

2. Analysing the data and including it in a document

Now that your data is ready, you can analyse it. Please set up a document with the sections outlined below. Paste your results in the corresponding section of your document.

Inflation. Load your inflation data set.

5. Create a bar chart showing the most recent 10 years' average inflation:

(a) Reduce the data set to comprise the most recent 10 years (i.e., 2013 – 2022).
(b) Compute the average at the country level using the `collapse` command in Stata or the `aggregate` command in R, and the median across all countries' averages. Insert the resulting table in your document.
(c) Create a bar chart including the median of the EU27 as a horizontal line. For the sake of disposition, rotate country names. Export this bar chart and include it in your document. Do not forget the caption in your document!
(d) Did Belgium's 2013 – 2022 average inflation rate fall below or exceed the EU27 median? Provide the answer in your document and list all the countries with a average inflation above the median as a table in your document.

Unemployment. Load your unemployment data set.

6. Plot in one graph the development of the unemployment rate for each EU country with an average inflation above the median for the years 2013 – 2022 in a different color. Include this line chart in your document.
7. Draw a boxplot diagram of for each country in a single plot. Make sure that each country's name is rotated for the sake of disposition. Export this graph to your document as well.

The relationship between Inflation and Unemployment. Load your inflation data set and merge it with your unemployment data set.

8. Reduce the data set to only contain 2013 – 2022 values and compute the average for that time span at the country level.
9. Generate a variable with the first three uppercase letters of each country's name using the `substr` command within `strupper` (Stata) or the `toupper` function (R).
10. Create a scatter plot showing the relationship between average unemployment (x-axis) and average inflation (y-axis). Label the countries with the first three letters of the country name. Export this plot and include it in your document. Add a brief description on the relationship you observe.

The relationship between GDP and Unemployment. Load your GDP data set and merge it with your unemployment data set.

11. Do you have GDP information on all EU countries? If not, provide a list of countries for which this information is missing in at least one of the years from 2013 to 2022.
12. Create a new variable called `log_gdp` as the natural logarithm of a country's GDP.
13. Create two tables in your document with unemployment and the GDP by year and the following percentiles and statistics: minimum, median, and the maximum.
14. Create a line plot showing Belgium's logged GDP on one y-axis over time, and Belgium's unemployment on another y-axis rate over time (see Figures below). Export this plot and include it in your document. Include a brief description on the relationship you observe. (Hint: Check out the Stata tip 93 in the Stata Journal (2010) or this note for R.)

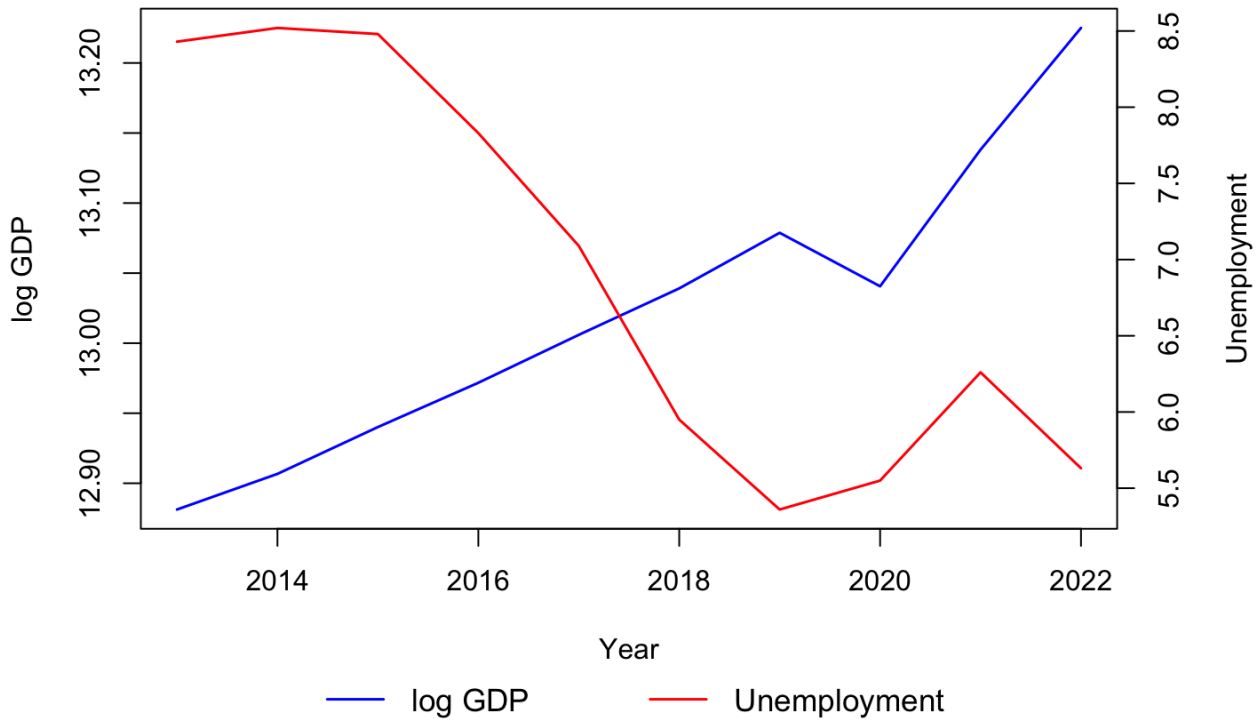

Figure 1: Log GDP and Unemployment (in R)

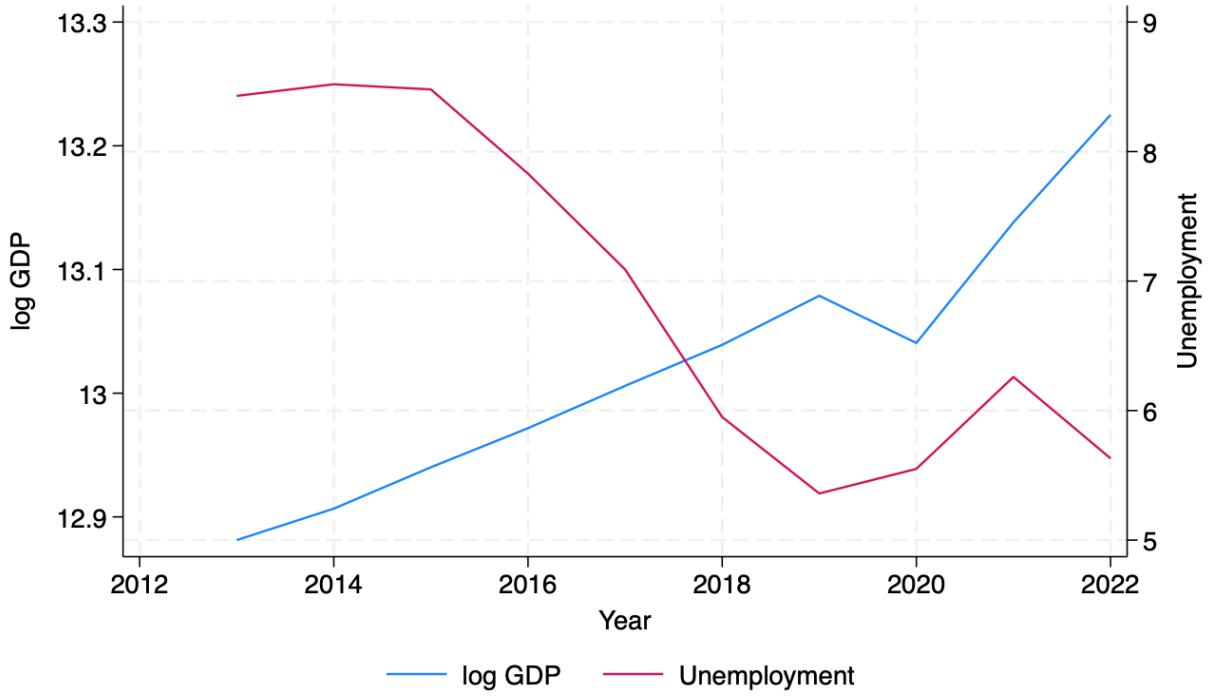

Figure 2: Log GDP and Unemployment (in Stata)