Final Data Analysis

# Trust in the European Parliament

Statistics I, Fall 2023

**Due by December 18th, 2023 at 5pm CET. You need to submit two copies of your report: one to the dropbox at the top of the course Moodle page, and one to your TA via email. Your reports should be in PDF format. File names should start with your matriculation number, e.g. "101010_FDA.pdf". Please omit your name. Your report should not exceed 8 pages in length, including tables and figures. Please attach your R-script as an appendix. The R-script does not count towards your page limit. The assignment should be treated as a take-home exam, which means that all work must be done individually. Notes, class materials, books and other written resources may be consulted. Other students and people other than your TAs and professor may not be consulted.**

This assignment asks you to explore factors associated with **trust in the European Parliament**. The data comes from the latest round of the European Social Survey (round 10, 2020). To access the data, follow this link and download the *ESS10 – integrated file, edition 3.2* in a **csv** format. Please note that in order to download the file, you need to register with your name, email etc. The website automatically asks for this, when you click 'Download' in the top right corner (if asked to choose between 'individual' and 'institutional' registration, choose 'individual'). Once the download is complete, run the code found on page 3 of this document – this will make the data set ready for you to work with. We recommend downloading the data early to prevent last minute complications. On page 4 of this document, you will find a description of each variable in the final data set (i.e. your codebook). The codebook also specifies which variables you should treat as interval-level variables, and which variables you should treat as categorical variables.

## Task:

Choose an independent variable of interest and formulate a research question, a null hypothesis, and an alternative hypothesis for how the variable relates to trust in the European Parliament. Include relevant descriptive statistics.

Design a statistical model to test your hypothesis and interpret any effects that you observe. As part of your model, you should identify 3-5 control variables and justify why each of them need to be included. Remember that theory should guide your model specification. If theoretically necessary control variables are not available, discuss the likely direction of the bias. Display your results in a regression table (e.g. using *stargazer*). Do not paste in rough R-output.

Be as convincing as you can in your argumentation. Test several models to explore under- and over-specification and discuss why your chosen model is the most convincing. Check for outliers and if present, discuss how they could affect your results. Run diagnostic tests to check for other pitfalls and show your most important diagnostic plots and/or output. Justify your model specification (is it a linear relationship? Is it an interaction?)

Grading will be reflective of the clarity and persuasiveness of your statistical analysis, your substantive argumentation, and the visualisation you employ in support. Note that no literature review or bibliography is necessary.

## R code: Run this code in R to get the data set ready!

```
## 1. Load data:

ESS <- read.csv("ESS10.csv")

## 2. Trim data set to only include the set of variables you can choose from:

ESS <- ESS[ , c("cntry", "agea", "gndr", "eisced", "brncntr", "ppltrst", "nwspol",
                "trstprl", "trstep", "stflife", "imbgeco", "hhmmb", "respc19")]

## 3. Rename the variables to make them easier to work with:

library(dplyr)
ESS <- rename(ESS, country=cntry, age=agea, gender=gndr, education=eisced,
              bornincountry=brncntr, soctrust=ppltrst, news=nwspol,
              parltrust=trstprl, eptrust=trstep, lifesat=stflife,
              immig=imbgeco, household=hhmmb, covid=respc19)

## 4. Recode NA values.
# All NA values are by default coded as values in ESS. With the following lines of
# code, you tell R which values reflect NAs:

ESS[, c(2,4,6:12)][ESS[, c(2,4,6:12)] > 50] <- NA
ESS[, c(3,5,13)][ESS[, c(3,5,13)] > 5] <- NA

## 5. Recode dummy variables and define them as factors.
# The dummy variables are by default coded as 1,2 instead of 0,1 in ESS. Further,
# they are not automatically read as factors (i.e. as categorical variables).
# The following lines of code will fix this:

# Gender:
ESS$gender <- ifelse(ESS$gender==2, 1, 0)
ESS$gender <- as.factor(ESS$gender)
# "Female" is now 1, "Male" is 0.

# Born in country:
ESS$bornincountry <- ifelse(ESS$bornincountry==2, 0, 1)
ESS$bornincountry <- as.factor(ESS$bornincountry)
# "Yes" is now 1, "No" is 0.

# Covid:
ESS$covid <- ifelse(ESS$covid==1, 1, 0)
ESS$covid <- as.factor(ESS$covid)
# "Yes" is now 1, "No" is 0.
```

## Codebook:

| *Name of variable* | *Description* | *Additional information* |
|---|---|---|
| **eptrust** (dependent variable) | 'Please tell me on a score of 0-10 how much you personally trust the European Parliament' | Ranges from 0 (no trust at all) to 10 (complete trust). Treat as interval-level variable. |
| **country** | Country where survey was conducted. | 22 European countries were included in the sample. Categorical variable. |
| **age** | Age of respondent. | Interval-level variable. |
| **gender** | Gender of respondent. | Recoded such that 0 = Male, 1 = Female. Dummy variable. |
| **education** | **'**What is the highest level of education you have successfully completed?' | Ranges from 1 (less than lower secondary) to 7 (higher tertiary). Harmonised across countries. Treat as interval-level variable. |
| **bornincountry** | **'**Were you born in [country]?' | Recoded such that 0 = No, 1 = Yes. Dummy variable. |
| **soctrust** | 'Generally speaking, would you say that most people can be trusted, or you can't be too careful in dealing with people?' | Ranges from 0 (you can't be too careful) to 10 (most people can be trusted). Treat as interval-level variable. |
| **news** | 'On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs?' | Answer given in hours and minutes. Interval-level variable. |
| **parltrust** | 'Please tell me on a score of 0-10 how much you personally trust [country]'s parliament?' | Ranges from 0 (no trust at all) to 10 (complete trust). Treat as interval-level variable. |
| **lifesat** | 'All things considered, how satisfied are you with your life as a whole nowadays?' | Ranges from 0 (extremely dissatisfied) to 10 (extremely satisfied). Treat as interval-level variable. |
| **immig** | 'Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?' | Ranges from 0 (bad for the economy) to 10 (good for the economy). Treat as interval-level variable. |
| **household** | 'Including yourself, how many people – including children – live here regularly as members of this household?' | Interval-level variable. |
| **covid** | 'Have you had coronavirus?' | Recoded such that 1 = Yes, 0 = No. Dummy variable. |