# Formative Exercise: Linear Regression in Notable

THIS EXERCISE DOES NOT CONTRIBUTE TO YOUR COURSE GRADE
SUBMISSION DEADLINE: 23:59 THURSDAY 1ST FEBRUARY 2024
Word limit: 1,500 words

For this exercise you will use data from the European Social Survey[1], which can be found in the datafile `ESS.dat`. This file is a Stata 12 datafile. You can open it in notable in the same way as in lab class 1.

In this survey, questions are asked of individuals, who are nested within countries, so really multilevel models (or some other technique which accounts for this clustering) should be used to analyse the data. However, the purpose of this assignment is to revise the concepts of multiple linear regression in order to familiarise yourself with Noteable and the R programming language, as such, you will only be asked to fit single level models.

The datafile does not include all the variables from the survey, but rather a small selection. The selection of variables includes four possible explanatory variables: **gender**, the respondent's sex; **age**, their age; **eduyrs**, an indication of their level of education (years of education completed); and **income**, their household's income (in 12 categories but you can treat this as a continuous variable if you wish). There are ten possible outcome variables, each representing the relative importance that the respondent places on a value. The values (and names of the variables[2]) are **Stimulation**, **Hedonism**, **Achievement**, **Universalism**, **Conformity**, **Tradition**, **Benevolence**, **SelfDirection**, **Power**, and **Security**. Details of the items used to measure the importance placed by the respondents on each value are available in the document `ESS_Values.pdf`.

You should choose an outcome and an explanatory variable. You may choose whichever of the ten values variables you like to be your outcome, and whichever you like from among **age**, **education**, or **income** to be your explanatory variable.

Your task is to investigate the relationship between your explanatory variable and your outcome using Noteable, and submit a brief report of your results. The report should draw on results generated using Noteable (or if you prefer R) and should include:

- Which variable you have chosen as your outcome and which you have chosen as your explanatory variable (you do not have to justify your choice, although you may explain the reasons for it if you wish)

[1] ESS Round 1: European Social Survey Round 1 Data (2002). Data file edition 6.4. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data for ESS ERIC.

[2] Except for **Self-Direction**, which is **SelfDirection** without the hyphen in the dataset to conform to software naming restrictions

- Descriptive statistics relating to these chosen variables: their mean, standard deviation, minimum and maximum values, and the number of individuals for whom the variable is missing
- The number of individuals in the dataset
- A scatter plot of the outcome variable against the explanatory variable
- A brief description of what the plot shows about the relationship between them
- Details of whether you transformed (or centered) any variables and, if so, why?
- Interpretation of the results of this regression:
  - The value of the intercept
  - What this means
  - The value of the coefficient of the explanatory variable
  - What this means (e.g. "there is an increase of 0.7 in the importance placed on Hedonism for every extra year of education")
  - Whether the effect is significant
- A second regression model showing a single level regression of your outcome which includes an interaction between your chosen explanatory variable and gender (check the coding of gender and recode if needed).
- Interpretation of the results as before (all parameters should be interpreted)
- Present a third model to investigate whether the results of your first model change when you include a second explanatory variable equal to the square of your original explanatory variable.
- Whether the effect of the linear term is significant
- Whether the effect of the quadratic term is significant
- A graph of illustrating the results of your third model
- Any diagnostic graphs you feel need to be interpreted to ensure the reliability of your results.

These elements may appear in whatever order you think sensible. Each need only be covered by a sentence or two. The report should flow together as a coherent whole.

When assessing significance, use a threshold of 0.05 for the *p*-value.