Analyzing Vote Choice Data

Assignment 1 - Deadline: May 3, 2023

TA: Francesco Raffaelli (francesco.raffaelli@politics.ox.ac.uk)

1 R-PACKAGES

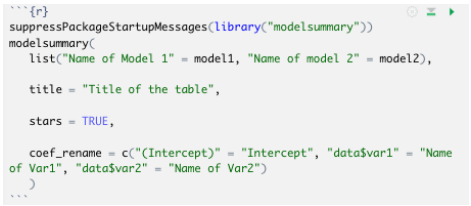
Before starting, let us review four `R-packages`, and their functions, that will be useful for this assignment - and in your research life!

- `suppressPackageStartupMessages` - When you load an `R`-package you may find yourself with quite a bit of messages in your console. If you want to avoid this, you can use this function, which uses the `library` function itself as its object.
- `haven` - This (very simple) package is used to import different statistical formats in `R`. In other words, you can use it to import `dta`, `SPSS`, or `SAS` files in `R`. For this goal, the main function you should remember is `read`

Figure 1. `haven`

- `modelsummary` - This package is used to report regression models into nice, tidy tables. It even allows you to put more regression models (or more specifications of the same models) into the same table (using `list()`, to visualize statistically significant coefficients (using `stars = TRUE`), and to modify variables names (using `coef rename()`).

Figure 2. `modelsummary`



- `ggplot2` - This package is all you need for data visualization in `R`! It has a lot of functions, but the general idea is that (a) you specify the data your are using, (b) you use `aes()` to specify your variables of interest

(i.e., your Y and your X) and other aesthetics, and (c) you finally pick the type of graph you want using `geom`. For example, to graph an histogram, you may use something like:

Figure 3. `ggplot2`

- Finally, function `glm` from package `nnet` is what you need if you want to run a logit (or probit) model

Figure 4. `nnet`

2 The Paper

This assignment focuses on dependent variable choice and on the use of logit models *versus* MNL models. In order to do so, we will use as reference Abou-Chadi and Hix's *Brahim Left versus Merchant Right? Education, class, multiparty competition, and redistribution in Western Europe* (2021), which argues that the education divide in Europe cannot be analyzed by looking at the right *versus* left blocks but that researchers must take into consideration different party families.

3 The Dataset

Dataset `ESS2018GER.dta` contains some variables for Germany from the most recent wave of the European Social Survey (2018). You need to upload it on `R` as you will use it for this assignment. In particular, the variables that we have selected for you are as follows:

- `cntry`: Country
- `ppltrst`: Trust towards others
- `prtvede1`: Party voted in last national election
- `gincdif`: Attitudes towards income redistribution
- `freehms`: Attitudes towards LBGT+ community

- `imwbcnt`: Attitudes towards immigrants
- `blgetmg`: Respondent belongs to ethnic minority
- `gndr`: Respondent's gender

Upload the assignment on Canvas by the aforementioned deadline. Rename the pdf document obtained from the R-markdown as follows:

"AVCD-Assignment1-YOURLASTNAME"

4 Exercise 1

1. First and foremost, explore and clean the data:

   (a) Get a sense of which variables and how many observations are present in the dataset, and make sure that all the variables are in the correct format: re-scale them when you find it appropriate. You may also want to use function `summary` to get a sense of how many missing values (if any) are present and if some variables are more problematic than others

   (b) Compute the mean of relevant variables and report it in a table

   (c) Make sure that all the dichotomous variables are in a "viable" format (0 / 1)

   (d) Represent the distribution of the variable measuring respondents' years of formal education. Label the axes, pick a title, and add a solid, blue line at the median value and a red, dashed line at the mean value. What do we learn about the distribution by looking at the mean and the median?

2. Following Abou-Chadi and Hix (2021), hence **A-CH**, create your outcomes variables. Use both possible operationalizations: the dichotomous left *versus* right (à-la-Piketty), and the quadripartite (à-la-ACH)

   (a) Does the distribution of years of education vary between those voting left and those who do not? For simplicity reasons, group years of education in four possible classes (Hint: use the dummy variable à-la-Piketty)

   (b) Estimate the relationship between voting left and years of education. (Hint: use the dummy variable à-la-Piketty)

   (c) Make sure to write down your estimation equation **formally**, and to comment the coefficients

   (d) Why have you chosen this model? How is it different from an OLS model?

   (e) Run the same model again, but adding the socio-economic covariates and controlling for individuals' level of social capital. Once again, write down the estimation equation **formally** and do not forget to comment your results after reporting them in a single, tidy table

3. Some theory:

   (a) Look back at the results you have gotten in Exercise 1.2, comment them, and explain how they relate to Abou-Chadi and Hix (2021)

   (b) Does education matter? Why? Take into consideration one possible channel through which education affects voting behaviour. Make sure to discuss whether it is a compositional or a contextual effect, that is, whether individuals with certain political attitudes and preferences self-select into/out of education, or whether education shapes and changes political attitudes and preferences

5 Exercise 2

1. You shall keep using the same dataset from the previous exercise. Test whether there is a positive relationship between voting the Libertarian Left and years of education

2. Is the aforementioned relationship stronger than that between education and voting for the Mainstream Left? Compare the two specifications in the same tidy table, and explain which criteria you have used in your assessment

3. Restrict the dataset to left-wing voters. Are preferences for income redistribution a good predictor of voting for both the Mainstream Left and the Libertarian Left? Compare your results (again, in the same table, if possible), and discuss them.

4. What does it happen when you add attitudes towards immigrants and sexual minorities as covariates?

   (a) Report the estimation equation **formally**

   (b) Comment the coefficients, reporting the two model specifications in the same, tidy table

   (c) What do you learn from this? Are preferences for redistribution *always* a good predictor of left voting?

5. Go back to the main dataset: you want to explore the relationship between preferences for redistribution and voting behavior *in general*:

   (a) What model would you want to run? And why?

   (b) Run the appropriate model (with and without the aforementioned attitudinal covariates) and report your baseline category of choice. What is the point of having a baseline category? Why do MNL models have it while logit models do not?