Assignment 4 ($<$ 2,000 words)

Benjamin Jarvis & Richard Öhrvall

March 07, 2024

In this assignment, you will examine how the age of political candidates relates to political choices in the United States, and (tenuously) the potential implications for the upcoming presidential election. To do this you will *individually* analyze data from a custom survey conducted by researchers from MIT and Georgetown University. The original citation is:

- Hainmueller J, Hopkins DJ, Yamamoto T. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*. 2014;22(1):1-30. doi:10.1093/pan/mpt024

The assignment has two main parts. The "Analysis" section describes how you should prepare and analyze your data. Writing an R script to perform these tasks is necessary to complete the assignment, but is not sufficient to receive a passing grade. The bulk of your grade will be based on your written report, as described in the "Report" section. Read all of the prompts carefully before proceeding and consider the overall design of your report as you develop your analyses and responses.

Submission

The assignment is **due on Thursday, March 14, at 10.00**. You submit in Lisam, under assignments. The deadline is strict and non-negotiable, so please do not wait until the last minute to submit. If Lisam is not working for you, e-mail your files to Ben, `benjamin.jarvis@liu.se`, before the deadline.

You should submit 2 files in Lisam: one Quarto (or R Markdown) file with your code and text **and** one file that is the rendered pdf version of the same Quarto (or R Markdown) file. The pdf file should not include code or console output, just tables, graphs and your written answers. Remember that you can control the output from the code chunks by setting echo, messages, etc. to FALSE.

Your final, compiled report (not the R script) should not exceed 2,000 words, and should ideally be less than 1,500 words. Focus on clarity and concision.

Note that this is an individual assignment and LiU takes plagiarism very seriously. Plagiarism can include handing in an assignment that is, in whole or in part, identical to another person's work, not citing other peoples' texts and ideas properly, and more. Please consult LiU's guidelines at https://liu.se/en/article/plagiering-upphovsratt.

Data

The data you are working with, contained in the file `candidate.csv` included with this assignment, are based on responses from Amazon's Mechanical Turk. In that sense, they are not representative of the US electorate and you should be extremely cautious in making inferences to the general population. But in this assignment, we will throw such caution to the wind. The MTurkers were shown a series of candidate pairings, and asked to indicate a numerical rating for each candidate (`ratingX`) and then choose the one candidate they would prefer over the other (`selectedX`). The candidates vary on a number of attributes: military service (`atmilitaryX`), religion (`atreligX`), education (`atedX`), occupation prior to politics (`atprofX`), race (`atraceX`), gender (`atmaleX`). Each respondent (given by `resID`) was shown multiple pairings, but we will ignore this for the purposes of the assignment. Instead we will simply focus on variations across choice scenarios (`choID`), regardless of respondent.

Analysis

The primary purpose of this analysis is determine whether, from an electoral/political standpoint (not a leadership or competence standpoint), concerns about Biden's age are warranted. To do this you first must analyze how age and other attributes relate to the candidate choices observed in your data. You do this using conditional logistic regression. Your main explanatory variable is age, but you should also include other variables that you think are important to candidate choice. Be sure to justify these decisions. Once you have a sufficient model, you must work out the implications of your findings for the upcoming US presidential election. You do this by producing predictions for a **data set that you make** that enumerates the differences between Donald Trump and Joe Biden in terms of the covariates in your model (and, therefore, in the choice data set). We know, the data aren't representative, but do this anyway!

1. Load the data and modify the variables as needed to treat them as continuous or factor variables, as appropriate.
2. Reshape the data for estimating a choice model.
3. Prepare a suitable table of descriptive statistics for the variables you use in your analysis. Summarize separately for the chosen and non-chosen alternative.
4. Estimate three conditional logistic regression models with candidate choice as the outcome variable. The models should, respectively, include:

- the linear effect of age as the only covariate.
- the effect of age and age-squared as your only covariates.
- the effect of age (based on whichever of the above provides a better fit) as well as the effect of any other covariates that you deem important for election outcomes.

5. Produce one table containing estimates from all three models. It should present the odds with associated confidence intervals.
6. Generate appropriate model fit statistics and put them in the regression table or another table. Include a likelihood ratio test comparing the fit of each model to the relevant, prior estimated model.
7. For all three models, produce predictions of the probability of voting for Trump or Biden. Present these predictions as a plot (not a table) as you see fit.

Report

You will be graded primarily based on your written report of your findings, **not** on the R code you wrote to obtain your findings. You must use complete sentences and paragraphs in your report. Your report should include and refer to figures and tables developed during your analysis. Figures and tables should be clearly and logically laid out, formatted, labelled, and footnoted. The quality and interpretability, not just the numerical accuracy, of your tables and figures will be a key component of your grade.

Your report should include the following sections:

- **Introduction**: briefly explain the study and your expectations regarding the main question. You don't have to do any extensive research to support your expectations, but please provide some rationale.

- **Data and Methods**: describe your data, e.g. when and how it was collected, and how you defined your variables for the purposes of modeling. This part should include your table of descriptive statistics for the variables in your models, and a brief discussion of notable patterns observed in this table. Also discuss how you modeled your outcome, and how the features of your data relate to features of the model. Here you should also discuss your model specifications (i.e., what variables are included in your analysis and how).

- **Results**: discuss your key results from your models, including appropriate interpretations of odds, p-values, model fit statistics, and predicted probabilities. Make reference to the figures/tables containing your regression results, fit statistics, and predicted probabilities.

- **Conclusions**: Summarise your main findings and compare with your expectations.