Assignment 2

Benjamin Jarvis & Richard Öhrvall

February 06, 2024

Assignment 2 (600–1,000 words)

Your assignment is to **individually** carry out the tasks described below. 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 file. The pdf file should not include your code, just the output and your written answer. Remember that you can control the output from the code chunks by setting echo, messages, etc. to FALSE.

The assignment is due on Thursday, February 15, at 09.00. You submit in Lisam, under assignments. The deadline is sharp, 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.

Please note that LiU takes plagiarism very seriously. Plagiarism can both be to hand in assignments that in parts are identical to others student's assignment or other work, to not refer to other peoples' texts and ideas properly, and more. Please consult LiU's guidelines at https://liu.se/en/article/plagiering-upphovsratt.

The data

In this assignment, you will use data from the the General Social Survey (GSS). The GSS is an opinion survey covering a wide variety of cultural, political, economic, and other social topics that has been running for decades in the US. We will use data from the GSS that is contained in the SocViz package, developed by Kieran Healy (install.packages("socviz")).

Documentation for the data can be found at https://kjhealy.github.io/socviz/reference/gss_sm.html and through the link provided there. In this task, you can ignore sampling weights. Your final answers should be provided in the form of a written report, sectioned by problem. You should use complete sentences and paragraphs in your report, accompanied by well-labeled figures and tables. Read all of the problems carefully before proceeding and consider the overall design of your report before you assemble your final .qmd file.

```
library(tidyverse)
library(socviz)

# Get the data
gss <- gss_sm
```

Voting for Obama.

In this assignment, you will explore the relationship between people's happiness and their propensity to vote for Barack Obama in the 2012 US presidential election, using the 2016 General Social Survey.

Begin by describing your study and present some hypothesis on the relationship between happiness and voting for Obama. The hypothesis does not have to be based on any research, just discuss what relationship you expect to find.

Data preparation and description

First, prepare the data for analysis. Create a new variable that takes the value 1 if the individual's mother *or* father has bachelor or a graduate degree, and 0 otherwise (if information on both parents' highest education is missing (`NA`) then this new variable should also be `NA`). Filter the data to select observations for only those who have non-missing (i.e., not `NA`) values on the education background variable you have just created, as well as for the `obama`, `age`, `sex` and `happy` variables.

Then calculate descriptive statistics for all five variables (the new education background variable, plus `age`, `sex`, `happy` and `obama`), showing the shares of observations for each categorical variable, and the mean, and standard deviations for any continuous covariates. Do this for the full sample, and separately for Obama and non-Obama voters (i.e., over different values of the `obama` variable). Put *all* of the descriptive statistics in a *single table*, making sure that the number of observations is also clearly presented.

Provide a brief discussion of data you are using and the summary statistics presented in the table.

Model estimation and odds ratios

Estimate three logistic regression models where voting for Obama (the `obama` variable) is the dependent variable and the following are explanatory variables:

a) only the happiness variable.

b) same as a), plus the new parental education variable.
c) same as b), plus sex and age variables.

Make sure that the three models are based on the same number of observations and that you use as many observations as possible given this constraint. For the happiness variable, the category “Not Too Happy” should be the reference category.

Present one table with the estimates from all three models shown as odds ratios with confidence intervals. Make sure to include key model statistics (e.g., N and the log-likelihood) at the bottom of the table.

Present a graph with the odds ratios from model c, with confidence intervals and without the intercept.

Make sure that the table and the figure have informative titles and relevant information in the comments.

Describe and interpret the results.

Predicted probabilities

Create one graph that visualizes the predicted probability of voting for Obama, including confidence intervals, across different ages (on the x-axis) and for the three levels of the happiness variable: Very happy, Pretty happy, and Not too happy. Base these predictions on Model C. For the purposes of constructing the graph, assume that you are producing predictions for a female with at least one college educated parent.

Describe and interpret the results.

Model fit

The models you have estimated are nested. Explain how the models are nested (which is nested in which?) and why. Perform likelihood ratio tests comparing model b to model a, and model c to model b. Also calculate Nagelkerke’s pseudo-R^2 and the share of observations correctly predicted for all three models.

Present the above fit statistics in a single, well-formatted and labelled table.

Describe and interpret the results of these model fit comparisons.

Conclusion

Summarise your findings and make some conclusion in relation to the hypothesis your presented.