Trump VS Biden

Ju Yoon Lee

Due Date: November 2, 2020

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

I will be using a logistic regression model with categorical variables to model the proportion of voters who will vote for Donald Trump. I will be using age, gender, race, education, and income as my explanatory variables. These variables, with the exception of age, are categorical with 2 categories, 0 and 1.

- Gender is classified as 1 if observation is male, 0 otherwise;

- race is classified as 1 if observation is white, 0 otherwise;

- education is classified as 1 if the participant has completed at least a bachelors degree, 0 otherwise;

- income is classified as 1 if the participant has an average income or higher, 0 otherwise.

The general logistic regression model I am using is

$$y = \beta_0 + \beta_1 x_{Age} + \beta_2 x_{Gender} + \beta_3 x_{Race} + \beta_4 x_{Education} + \beta_5 x_{Income} + \epsilon$$

Where y represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, the proportion of 0 year old female, non-white voters with less than average income and has not completed at least a bachelors degree who will vote for Donald Trump (This is actually not possible since age will always be greater than or equal to 18). Additionally, β_i represents the slopes of the model. So, for each categorical explanatory variable, β_i is the slope for the variable and notice that if $x_{Gender}, x_{Race}, x_{Education}, x_{Income}$ are all equal to zero, then their slopes are irrelevant and the model depends only on age.

The survey data and census data originally had multiple categories for each variable but has been reduced to 2 for the sake of simplifying the model based on the proportions of the two categories. All observations where the age of the participant is younger than the legal age to vote has been removed from the data and age is the only explanatory variable which is numeric and continuous. This reduces the data to only observations that are relevant and separates the data into very broad categories.

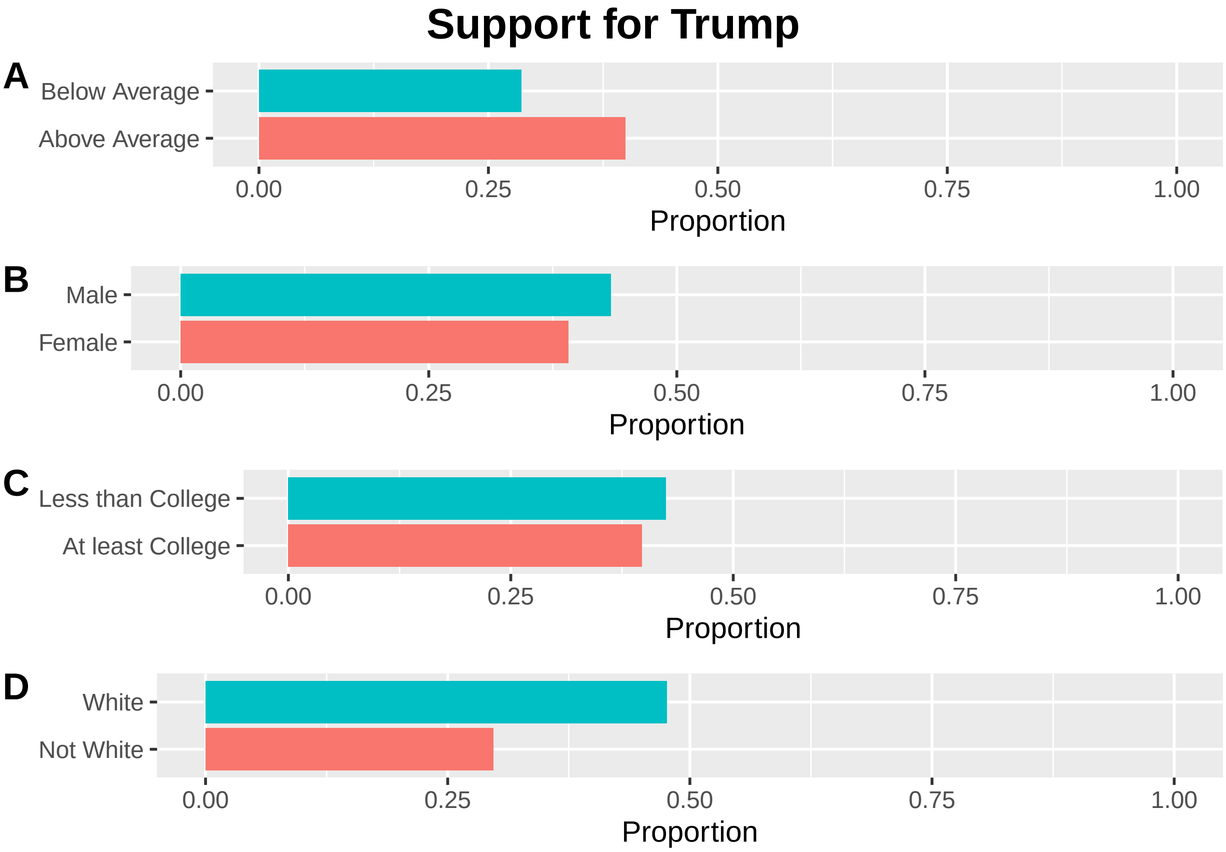
Based on the summary of the model, the slope estimate for age is 0.011720 and we can predict that age will not have any big impacts on the outcome.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Here I create cells based off different ages, race, income, education, and gender. I decided to split the cells based on age, race, income, education, and gender because differences between these factors seem to have a greater impact on the outcome of proportion of voters who will vote for Trump compared to, for example, marital status. Using the model described in the previous sub-section I will estimate the proportion of voters in each bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size. This process predicts the outcome of the popularity vote between Trump and Biden.

Results

First, we will look at the proportion of voters who will vote for Trump for each category. Below is the corresponding bar graphs where the different colours represent the different categories for each explanatory variable.



Based on the graphs, it seems that the race observes the biggest difference in proportion of voters who will vote for Trump, with the proportion of white voters supporting Trump is significantly greater than the proportion of non-white voters supporting Trump. This is something we can expect, considering that Trump has been accused of being a racist and a white supremacist. Notice that none of these values for proportion exceed 0.50 which means that a greater proportion of voters will vote for Biden. For example, the proportion of white voters who will vote for Trump is roughly 0.48 which means that roughly 0.52 of white voters will vote for Biden. Similarly, for each category, it looks as if a higher proportion of all voters will vote for Biden. Thus, we have no objections so far that Biden will win the popularity vote with a higher proportion of votes.

Our prediction of the proportion of voters who will vote for Trump based on model created using the survey and census data has an output of 0.4182. In other words, we predict that 41.82% of all voters will vote for Trump and as a result, Biden will win the popularity vote. This is consistent with our expectation of the outcome, which was that Biden will win the popularity vote.

Discussion

To summarize what we have done, we first transformed all of our multilevel explanatory variables into binary variables which greatly simplifies our data. We then made a logistic regression model based on these variables and then used the census data to predict the proportion of voters who will vote for Trump. Essentially, we categorized each observation into broad groups and counted all of the observations that are in the same groups.

Based off the estimated proportion of voters in favour of voting for Donald Trump being 0.4182, we predict that Joe Biden will win the popularity vote. However, because our data has been overly simplified by making all categorical variables binary, we lose a lot of accuracy when using this transformed data. Although this method of using simplified data provides potentially inaccurate results, it serves as a good starting point for a more in-depth analysis of the model and gives a general idea of what we can expect the results to be like. Although 0.4182 is not the exact proportion of voters who will vote for Trump and as we start adding in the multilevel categories, transforming from binary back to original, we can expect that the result will be more accurate and will be somewhat close to 0.4182. This is one of the weakness of this model. Oversimplification. This model is separated into groups that are too general and people with different properties may be categorized into the same group, thus reducing accuracy. One way to work around this issue is to make new models where each categorical variable is not binary anymore, but includes more than 2 categories if not all.

Additionally, our model has a relatively small sample size. With a total of 6,116 observations, this value is far less than the total number of voters in the US. This further reduces the accuracy of the model. Unfortunately, a larger sample size was unavailable for use as it slowed down my computer way too much. However, a simple solution for this issue is to find more participants and get a larger sample size.

Next Steps

As mentioned before, the next steps towards a more complete and accurate model is to first increase the sample size and much as possible. This will ensure the values calculated from the model will be more accurate. Additionally, we can use all of the categories for each explanatory variable instead of the simplified binary variables.

References

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

- Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

- Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

- Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr

- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1. https://CRAN.R-project.org/package=broom

- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

- Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814).