Predicting the 2020 America Federal Election Result from Democracy Fund + UCLA Nationscape and IPUMS USA Datasets

Rutvik Gupta (1004939837), Elyssa Plaza (1004356760)
Yubing Xia (1005063244), Hongbo Zhou(1004832862)

November 2, 2020

Model

Our team is interested in predicting the popular vote outcome of the 2020 American federal election obtained from *Democracy Fund + UCLA Nationscape* dataset (Tausanovitch, Chris and Vavreck, Lynn. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=88d78bb4-8775-430d-bda2-27b88f955618) and the *IPUMS USA* dataset (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. Retrieved from https://doi.org/10.18128/D010.V10.0). To do this, we used a logistic regression model and employed post-stratification technique. In the following sub-sections, we will describe the model specifics and the post-stratification calculations.

Code and data supporting this analysis is available at: https://github.com/RutvikGupta/STA304-PS3

Model Specifics

Firstly, we will build a regression model using the RStudio software based on the *Democracy Fund + UCLA Nationscape* survey data.

Since the response variable, the proportion of voters who will vote for Donald Trump, is binary, we will be using a logistic regression model for modeling. After examining all the survey data variables, 7 predictor variables are selected based on our beliefs of influential factors for further examination. However, we discover that only 3 of them seem to exhibit statistical significance, thus we will only be using age, gender, race_ethnicity to model the probability of voting for Donald Trump. We sort out the race_ethnicity variable into 3 main categories – White, Black or African American and Other Race for model simplicity, and use age as a numerical variable since it is difficult to divide by age groups that are homogeneous within each group.

Table 1:

```
## # A tibble: 6 x 4
##   vote_trump gender race_ethnicity                 age
##        <dbl> <chr>  <chr>                        <dbl>
## 1          1 Female White                           49
## 2          1 Female White                           46
## 3          1 Female White                           75
## 4          1 Female White                           52
## 5          0 Female White                           21
## 6          0 Female Black, or African American    38
```

Table 1 gives an overview of our survey data for modelling purposes.

The logistic regression model we are using is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{genderMale} + \beta_3 x_{raceOther} + \beta_4 x_{raceWhite}$$

$\hat{p}$ represents the estimated percentage of voters who would vote for Donald Trump. β_0 is the measure of the model intercept and describes the probability of Donald Trump votes when the voter is a female black or African American of age at 0. β_i, where $1 \leq i \leq 4$, stands for the slope of the model. When age increases by one unit, voting for Donald Trump measured in log odds is expected to increase by β_1. The log odds of being a male Trump supporter is β_2 times that of female voters. Clear difference is also shown among different races. Other ethnicities have a β_3 time of log odds of voting for Donald Trump than of black or African Americans, compared to a β_4 multiple of Caucasians (raceWhite).

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump in the population, we need to perform a post-stratification analysis using the *IPUMS USA* dataset (Ruggles, Flood, Goeken, Grover, Meyer, Pacas and Sobek). Post-stratification partitions data into demographic cells and estimate response variables for each of the cells. Then these cell-level estimates are aggregated up to a population level,leading to a final result. This is done by weighting each cell by its relative proportion to the population (Lecture 6), as described by $\hat{y}^{PS} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$. Here we create cells by splitting based on **age, race, gender, and the state where the census was taken**. Using the model described in the previous sub-section we will estimate the proportion of voters in each bin. We will then weigh 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.

We choose to include age because it would likely influence one's vote. The time a person was born can result in different ways of thinking and beliefs. Variables gender,race and state are also involved because a vote can depend on how one likes the way a candidate interacts with their gender, race or state policy. On the other hand, we decide not to include income because it was difficult to find the total income of the household on the census data.

Table 2:

```
## # A tibble: 6 x 5
##   sex     age race                             stateicp        n
##   <chr> <dbl> <chr>                            <chr>       <dbl>
## 1 male      2 white                            connecticut   118
## 2 male      2 black/african american/negro     connecticut    13
## 3 male      2 american indian or alaska native connecticut     1
## 4 male      2 other asian or pacific islander  connecticut    10
## 5 male      2 other race, nec                  connecticut     7
## 6 male      2 two major races                  connecticut    18
```

Table 2 gives an overview of our census data for post-stratification.

Results

Table 3:

```
## # A tibble: 5 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                   -2.80      0.178       -15.7  1.19e-55
## 2 age                            0.00794   0.00195      4.07 4.67e- 5
## 3 genderMale                     0.400     0.0635       6.30 3.05e-10
## 4 race_ethnicitySome other race  1.59      0.181        8.77 1.86e-18
## 5 race_ethnicityWhite            2.35      0.159        14.7  3.71e-49
```

From the Table 3 summary of the logistic regression model, we find that the individual p-values of age, gender (Male), race ethnicity (Other) and race ethnicity (White) estimates are all lower than the significance level at 0.05. Therefore, all the predictor variables - age, gender and race ethnicity - have a statistically significant impact on voting for Donald Trump. More specifically, for one unit increase in age, we expect an increase of 0.0079 in the log odds of voting for Donald Trump. In terms of gender, the log odds of voting for Donald Trump in males is 0.3997 times that of the female population. As for race ethnicity, the log odds for supporting Donald Trump in presidential elections in other races is 1.5904 times that of amongst black or African Americans. This value is slightly different, at 2.3468, when comparing the log odds in white people than that of black or African American ethnicities.

We estimate that 45.51% of voters were in favor of voting for Donald Trump, who represents the Republican Party. With cells split by age, race, sex and the state where the census was taken, we performed post-stratification analysis on the voter proportions who are in support of the Republican Party modelled by the logistic model discussed above.

Discussion

Summary

This study aims to predict the overall popular vote of the 2020 American Federal election. A logistic regression model was built based on the *Democracy Fund + UCLA Nationscape* dataset (Democracy Fund VSG, Tausanovitch and Vavreck). Then the post-stratification method was applied on the *IPUMS USA* dataset (Ruggles, Flood, Goeken, Grover, Meyer, Pacas and Sobek). We estimate that Donald Trump, representative of the Republican Party, would collect 45.51% of the votes of the general voters population.

Possible biases are included due to the formatting of the response variable, vote_trump. Amongst the sample population, we dichotomize voters in support of Donald Trump as 1's whilst those who are in favor of the opposing candidate, Joe Biden of the Democratic Party, as 0's. Answers including "I am not sure/don't know" and "I would not vote" are excluded from our data analysis, which could potentially lead to data misrepresentations. However, these biases are minor and are beyond our reach to correct for, so for this study's purpose we will continue using the binary response variable.

Conclusions

Based on the post-stratification output, we estimate that 45.51% of the voting population will support Donald Trump in the Republican Party, and 54.49% of the public votes will go to Joe Biden, concluding the 2020 American Federal Election winner to the Democratic Party which Joe Biden represents.

This study output serves as a reflection on the general voters' opinions and therefore brings broad impacts to multiple parties. For Americans who haven't practiced their voting privileges for the 2020 Election, this analysis might assist them in their final decision. Moreover, in several states such as Michigan, Wisconsin and Connecticut, citizens are allowed to change their votes (Phelps and Karson). This analysis might give some of them a different perspective and change their votes accordingly. Moreover, the two competing parties could analyze and respectively modify their publicity strategies, aiming to win more votes before the final closing date of the citizen poll in each state.

Weaknesses

For model simplicity, during the initial stages of choosing variables we selected 1 response and 7 predictor variables from the 17 variables in the provided reduced survey dataset. This decision of excluding these variables pose weaknesses as to the overall model expressivity.

For geographical data of the respondent, we included the census region but not state or congress district information for subsequent modelling. This was due to concerns including information overlap and the level of difficulty in handling state and congress distinct inputs, as these variables have too many categories, which would make the model difficult to interpret thus decreases its usability.

We also made a judgement call to only include only the foreign_born variable and exclude the hispanic variable for collection of the political opinions from the minority groups in America. The foreign_born variable allows us to analyze the political perspectives of those who are not born in and have immigrated to America. Although addition of this variable collects standpoints from the immigrants on the 2020 Federal Election, we excluded the group of Americans of hispanic, latino or spanish origin. This population might have a different opinion on the two candidates from the immigrants, and could therefore affect the results if included.

The socioeconomic background of individual voters was represented by household income for modelling purposes. An individual's household income could represent their purchasing power, but cannot indicate their education and employment state which could also affect their stands in social class. Moreover, education level and employment status of individual voters could greatly impact their support of the competing candidates, since the two have different stands as to job market opening in general and for citizens of specific education levels. Therefore, the exclusion of education and employment variables in the model pose information loss to the accuracy of the model output.

Aside from variable selections, our model failed to filter by the registration variable, which possibly brings biases to the output. The registration variable collects whether the participants are registered to vote or not. Without filtering those who are registered to vote, we possibly included opinions to the 2020 Federal Election from people who are not eligible to vote (until they are registered). If they intend not to vote but still express their preference to a Party, this could lead to biased results.

Next Steps

The 2020 American Federal Election result will be released by January, 2021 (Wikipedia contributors). By then the actual election output could be compared with the estimated result from this study. The difference in general popularity vote of the two candidates could then be further studied. For instance, the different variables discussed in the weakness session could be included to generate another model to compare with the initial one for analyzing which variables are more statistically significant. A follow-up survey could be sent out to the identical participant tool, asking whether the respondents did follow their expressed preference in the actual election as in the initial survey or not. Additionally, another survey could be conducted for collecting voters' opinions at the 2020 Election result, which would allow us to perform a post-hoc analysis. With these modifications, we are able to improve the estimation ability of the current model and provide a more accurate output in future election studies.

References

1. 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. Retreived from https://doi.org/10.18128/D010.V10.0

2. Tausanovitch, Chris and Vavreck, Lynn. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=88d78bb4-8775-430d-bda2-27b88f955618

3. Phelps, Jordyn, and Karson, Kendall. "Can you change your vote? Trump thinks people should." *ABC News*, 27 Oct. 2020, https://abcnews.go.com/Politics/change-vote-trump-thinks-people/story?id=73854468.

4. "Timeline of the 2020 United States presidential election." *Wikipedia*, https://en.wikipedia.org/wiki/Timeline_of_the_2020_United_States_presidential_election. Accessed 31st October 2020.