Prediction of the Overall Popular Vote of the 2020 American Federal Election

Huiyan Li; Wenyu Qu; Bingzhen Wan; Tongxin Zeng

November 2, 2020

Model

Model Specifics

In the effort of discovering the potential winner of the 2020 US Presidential Election, we will be using three logistic regression models to model the proportion of voters who will vote for Donald Trump, Joe Biden, and people uncertain of who to vote. The variables we used for all three models include age, gender, race, and employment status of voters. The logistic regression model expression for all three models is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{race} + \beta_3 x_{employment}$$

Where $log(\frac{p}{1-p})$ represents the proportion of voters who will vote for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3. The β_0 represents the intercept of the model, and is the probability of voting for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3 when age equals to 0, gender is female, the race is American Indian or Alaska Native race, and employment status is employed. Additionally, β_1 represents for every one unit increase/decrease in age, we expect a β_1 increase/decrease in the probability of voting for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3. Similarly, β_2 represents for every one unit increase/decrease in the gender of male, we expect β_2 increase/decrease in the probability of voting for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3. Then, β_3 represents for every one unit increase/decrease in race, we expect β_3 increase/decrease in the probability of voting for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3. Lastly, β_4 represents for every one unit increase/decrease in employment, we expect β_4 increase/decrease in the probability of voting for Donald Trump in model 1/Joe Biden in model 2/Uncertain Vote in model 3.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump/Joe Biden/Uncertain Vote, we need to perform a post-stratification analysis. Post-stratification is a data analysis technique for correcting known differences between sample and target populations that partitions the data into thousands of demographic cells and then estimates the response variable (y) for each cell. Then "it aggregates the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the pollution" (Wang et al., 2014). We can denote post-stratification estimate as

$$\hat{y}^{PS}$$

, where y is the outcome of interest. Post-stratification is useful because it allows us to make corrections for poorly representative samples or non-probability sampling. Here we create cells based on different ages, gender, races, and employment status. The reason we choose race is, according to President Obama's Presidential Election, "race is the biggest factor," which illustrates that racial diversification can play an integral part in the presidential election. And the reason we choose employment status is that it may significantly impact their voting behaviors. For example, an unemployed individual may have a lower optimistic level of voting for those who support increasing taxes because their living expenses will increase.

Now, we will calculate

$$\hat{y}^{PS}$$

of Donald Trump/Joe Biden/Uncertain Vote to estimate the proportion of voters in favour of voting for Donald Trump/Joe Biden/Uncertain Vote. Here we create cells based on different ages, gender, race, and employment status. Using the logistic regression model described in the previous sub-section, we will estimate the proportion of voters in each combination of age, gender, race, and employment status bins. We will then calculate the proportion estimate within each bin and 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.

Results

Result Table

–	Donald Trump	Joe Biden	Uncertain Vote
$\hat{y}^{PS}$	0.393	0.528	0.077

As you can see from the result table above,

$$\hat{y}^{PS}_{trump} = 0.393$$

,

$$\hat{y}^{PS}_{biden} = 0.528$$

, and

$$\hat{y}^{PS}_{uncertainvote} = 0.077$$

. The interpretaion of these results are as follows:

Donald Trump

The estimated result of the proportion of voters who are more likely to vote for Donald Trump is 39.3%. This is based on our post-stratification analysis of the proportion of voters who is more likely to vote for Donald Trump modelled by the logistic regression model 1 above, which accounted for age, gender, race, and employment status.

Joe Biden

The estimated result of the proportion of voters who are more likely to vote for Joe Biden is 52.8%. This is based on our post-stratification analysis of the proportion of voters who is more likely to vote for Joe Biden modelled by the logistic regression model 2 above, which accounted for age, gender, race, and employment status.

Uncertain

The estimated result of the proportion of voters who are uncertain of who to vote for is 7.7%. This is based on our post-stratification analysis of the proportion of voters who are unsure of who to vote by the logistic regression model 3 above, which accounted for age, gender, race, and employment status.

By comparing the proportion of voters who will vote for Donald Trump and Joe Biden. We can see that the proportion of voters who will vote for Joe Biden is 13.5% higher than Donald Trump, which indicates Joe Biden is more likely to win the 2020 American federal election without considering uncertain votes. Although there are still 7.7% of undecided votes, even if everyone unsure about voting eventually decided to vote for trump, trump still needs a proportion of 5.8% more voters to win Biden.

Discussion

Summary

In this study, we are interested in predicting the 2020 American federal election's popular vote outcome. We selected age, gender, employment status, and race as our base variables after careful consideration and reading the material regarding previous US Presidential Election. We employed data cleaning, modelling, and post-stratification technique. We filtered off voters who are not eligible to vote (i.e. less than 18 years old or not American citizens), voters who are not registered, and voters who do not intend to vote. Then we categorized age into five different groups. Each age group's gap is 20. and mapped the category specification of gender, employment status and race between survey and census data during the data cleaning process. Later, we created three logistic regression models to model the proportion of voters who will vote for Donald Trump, Joe Biden, and people uncertain of who to vote for. We then applied these models to calculate the post-stratification of Donald Trump, Joe Biden, and people uncertain of who to vote for. From the post-stratification result, we would predict the potential winner for the 2020 US Presidential Election; Donald Trump or Joe Bidden.

Conclusions

From the logistic regression model's summary statistics, we observed that the estimated value/slope of age for Trump is 0.0114, the estimated value/slope of age for Biden is -0.006854, which shows that older people are more likely to vote for Trump instead of Biden. We also observed that the estimated value/slop of males for Trump is 0.4043; the estimated value/slope of males for Biden is -0.299756, which shows that males are more likely to vote for Trump over Biden, which supports our hypothesis for gender behaviour in voting. Moreover, we observed that all the estimated values of each racial group for Trump are negative, suggesting that compared with the base level race "the American Indian or Alaska Native race", the proportion of other races voting for Trump decreases as the number of votes increase. Contradictorily, for Biden, each racial group's estimated value/slope is positive, suggesting that compared with the base level race "the American Indian or Alaska Native race", the proportion of other races votes Biden increases as the number of vote increase. Additionally, we observed that all the estimated values/slope of each employment group for Trump are negative, suggesting that compared with the base level employment status - "employed", the proportion of people with other employment status(i.e. unemployed or not in the labour force) vote for Trump decreases as the number of vote increase. Some potential reasons for this situation may include but are not limited to Trump's policy of building walls along the USA-Mexico borderline and increase tariff hurts many small and medium businesses, resulting in many unemployment. Contradictorily, for Biden, each employment status group's estimated values/slope is positive, suggesting that compared with the base level employment status - "employed", the proportion of people with other employment status votes Biden increases as the number of vote increase.

According to the result of post-stratification, we conclude that most people will vote for Joe Biden, which Bidden has 13.5% more voter proportion than Trump. Therefore, the potential winner for the 2020 US Presidential Election is Joe Biden.

Weaknesses

One of the main weaknesses of our model appears during the process of data selection. First of all, there are "time difference" between our survey data and census data, which people who are under 18 when completing the survey but later aged older than 18 when conducting the census are not considered in our model. Therefore, one weakness is the data inaccuracy, which some data become unavailable to gather. Then, we only select people with "naturalized citizen" citizenship. However, there are some people whose parents are citizens while they are born outside of the US. Thus, those people might have the right to vote but are not considered as "naturalized citizen." Hence, we might miss those people in our data.

Another weakness exists in our modeling process. In post-stratification, we apply the models we get from survey data to census data to estimate the proportion of voters in each combination of age, gender, race, and employment status cell. Then, we calculate the proportion estimate within each bin then calculate their average based on their weights. However, this method is flawed. For example, there might exist a situation like there are three voters whose probability to vote Trump are 0.45, 0.45, 0.80 respectively, and if we try weighted average, the result yields 56.7%. If we choose 50% as the threshold, the actual probability of voting for Trump should be 33.3%. Hence, we might get an incorrect result.

Next Steps

With regard to the weakness related to probability, we can approach to discover the probability from another perspective. For example, we might estimate the probability of voting for Trump based on the different states in the US to estimate and predict the total probability of voting for Trump. According to our estimation, we only use age, race, gender, employment as independent variables. In order to obtain a more accurate and more precise result, we can try adding more variables such as education level and personal income in the logistic regression model.

Furthermore, those values predicted in our model are only estimated results. Therefore, after we retrieve the actual election results, we can compare these two results and conduct further analysis. For example, we can analyze whether some variables have correlations with each other and whether some specific factors have special results and how that could happen. With a better variable selection, a better regression model, and other improvements, we could apply them in future selection so that we would find a more precise estimation result in the future.

References

- Albert-Deitch, Cameron, “How Much the Trade War Is Affecting Small Businesses, According to a New Survey”, Inc. Accessed on NOV. 1, 2020
- Alexander, Rohan; Caetano, Sam, “Data Cleaning Survey”, Created 22 October 2020. Lisence MIT
- Alexander, Rohan; Caetano, Sam, “Data Cleaning Post-stratification”, Created 22 October 2020. Lisence MIT
- Aphalo, Pedro “na.omit”, from photobiology v0.10.4 RDocumentation
- Astro Statistic Department, “Logarithms and Exponentials”, Pennsylvania State University. astrostatistics.psu.edu/su07/R/html/base/html/Log.html
- Caetano, S. (2020). Week 6: Multilevel Regression & Poststratification. STA304 Survey, Sampling and Observational Data. University of Toronto.
- CAWP, “THE GENDER GAP Voting Choices In Presidential Elections”, Center for American Women and Politics (CAWP), Eagleton Institute of Politics, Rutgers University. Cawp.rutgers.edu/sites/default/files/resources/pr
16-12_wvwatch.pdf
- ENDMEMO, “R sub Function”, endmemo.com © 2020, Accessed on Nov. 2, 2020
- Grafstein, Robert, “The Impact of Employment Status on Voting Behavior” The Journal of Politics Vol. 67, No. 3 (Aug., 2005), pp. 804-824
- Harvard University, “Race and Ethnicity Still Play a Role in Political Attitudes”, Harvard University Kennedy School of Political Science iop.harvard.edu/race-and-ethnicity-still-play-role-political-attitudes
- Fox, John, “predict”, from car v3.0-9, RDocumentation
- Hadley Wickham [aut, cre], Evan Miller [aut, cph] (Author of included ReadStat code), RStudio [cph, fnd], “Package ‘haven’ ”, Cran-r-project.org
- Pew Research Center, “AN EXAMINATION OF THE 2016 ELECTORATE, BASED ON VALIDATED VOTERS”, PRC US Politics and Policy, Created AUGUST 10, 2018
- R-core R-core@R-project.org, “remove”, From base v3.6.2, by RDocumentation

- R-core R-core@R-project.org, "numeric", From base v3.6.2, by RDocumentation
- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=6fe6300d-ebf0-4c1f-8f9c-8be80050fd0f
- Tidyverse, "Tidyverse packages", Powered by R. tidyverse.org/packages/Accessed on Nov. 2, 2020
- Wickham, Hadley, "rename", From base plyr v1.8.6 RDocumentation
- Wikipedia contributors, "Political career of Donald Trump," Wikipedia, The Free Encyclopedia en.wikipedia.org/wiki/Political_career_of_Donald_Trump
- 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
- Wang, W., et al., Forecasting elections with non-representative polls. International Journal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.06.001

Appendix

Github Repo link: https://github.com/liviasidealab/US-Election-2020