Winner Pridiction for US 2020 Presidential Election

Fei Yang(1004847696), Junwen Yuan(1004136696),Ruoqiuyan Zhang(1003926718),Yuxin Li(1002330998)

02/11/2020

The code and data supporing this analysis is available at https://github.com/Vickyli6762/STA304PS3

Model

We are interested in the 2020 American federal election, thus we want to predict the vote outcome using statistical analysis. Post-stratification technique is applied to have multi-level modelling.

- Model Specifics

The models we used are logistic regression models to determine certain factors of voters that would have an effect on whether Donald Trump or Joe Biden would win the election, respectively. Five variables were taken into consideration, which are whether voters finish college, and their age group, gender, racial group, and income. There are two models, one for the prediction of voting for Donald Trump, and the other for the prediction of voting for Joe Biden. The reason why logistic models are chosen is, they can be used to model binary response variables, which are suitable for the prediction of whether citizens vote for candidates. Two models were used, because they can reflect more comprehensive predictions of people's voting choice between the two candidates. The logistic models we used are:

$$\log\frac{p_t}{1-p_t} = \beta_0 + \beta_1 X_{fc} + \beta_2 X_{age} + \beta_3 X_{gender} + \beta_4 X_{racial} + \beta_5 X_{income}$$

$$\log\frac{p_b}{1-p_b} = \beta_0 + \beta_1 X_{fc} + \beta_2 X_{age} + \beta_3 X_{gender} + \beta_4 X_{racial} + \beta_5 X_{income}$$

Where p_t is the probability of Trump would win the election, and p_b is the probability of Biden would win the election. β_0 represents the intercept of the model. β_1, β_2, β_3, β_4, and β_5 represent the change in log odds for every one unit increase in X_{fc}, X_{age}, X_{gender}, X_{racial}, and X_{income}, respectively. X_{fc} is the binary variable represents whether voters finished college. X_{age} is the voters' age group, which has four levels: 19-24, 25-44, 45-65, and older than 65. X_{gender} represents voters' gender. X_{racial} represents voters' racial group, which has multiple levels, including white, Asian, Japanese, etc. The variable X_{income} describes voters' income level, which is grouped into less than 25,000, 25,000 to 75,000, 75,000 to 125,000, 125,000 to 250,000, and above 250,000.

- Post-Stratification

To investigate the probability of people who attempted to select Donald Trump, we would like to use the technique of Post-stratification. First, we have selected 5 variables from the 'American Community Surveys(ACS)' dataset which are cleaned during 2018 5-years. Post-stratification is used to estimate investigation on survey sampling which is always based on population, and it reduces the variance of the estimate to give out more precise results. If people apply this technique on their models correctly, it not only reduces the variance of the estimate, but it also raises the sample's representativeness. Therefore, there will be more confidence in our population's factors that are selected; for example, in this study, we have chosen education (finish_college), age (age_group), race (racial_group), income and gender as our factors. Also, the five variables are used to create cells in the process of post-stratification. To be more specific, there are 2 separate groups in education, 4 groups in age, 7 groups in race, 4 groups in income and 2 groups for gender. By multiplying the numbers of groups for each variable, the results demonstrate that there are 448 cells in this investigation. Then we will estimate the proportion of people in each education, age, race , income and gender bin. Finally, we will calculate each proportion estimate by the respective population size of that bin and sum those values and divide that by the entire population size.

Results

Based on the models and post-stratification techniques, we estimate that the overall probability of Donald Trump winning the election is 0.428, and the overall probability of Joe Biden would win the election is 0.421. Both probabilities accounted for the five variables included in the logistic regression models, which are whether voters finish college, voters' age group, gender, racial group, and income level.

Discussion

- Summary

In summary, we discuss the recently popular topic that the 2020 U.S. Federal Election between Donald Trump and Joe Biden by logistic regression models. Logistic regression models could help us forecast the choice of citizens more accurately. Since America involves all kinds of racial groups, class and age groups, here we introduce post-stratification to investigate five variables from "American Community Surveys(ACS)", which are whether voters finish college, and their age group, gender, racial group, and income.

- Conclusion

After estimating the vote probability from census data, our results show the probability of Trump is around 0.428 and Biden is around 0. 421. And Trump is slightly higher than Biden, which is around 0.007. Hence, our prediction is that Trump will win the 2020 U.S. Federal Election.

- Weaknesses

Although the data set involves five variables to predict the U.S Federal Election, there are still potential influencing factors of citizens' background or experience existing such as religion and attitude towards policy. And people who live in different states still influence their vote choice, because different states have different political and economic factors. In addition, there are different importance levels for different states. For example, California has 55 votes counting and if Donald Trump wins 50.01% during referendums in California, then he will gain a whole 55 vote in the election. Moreover, there is potential result that the Candidate wins referendums but fail to be elected president. In 2016, although Hillary Clinton won more public votes than Donald Trump, Trump won 304 votes compared to 232 votes for Hillary Clinton finally (BBC News). And the survey data only involves 6467 entries, and the data set we explored is only 448, which does not involve all citizens in America, the result might be influenced due to a small sample group.

- Next Steps

In order to improve the result more accurately, the next steps we may add each-state as one more variable in our data set. For example, Although a person in Georgia might vote for Joe Biden, Donald Trump wins Georgia, then Donald Trump might win Georgia. In addition, each day might have different results about election, but it is impractical to collect data every day and select sample data, then we could build a prediction model for accurate results by the average of vote from each week or each month.

Appendix

https://www.voterstudygroup.org/publication/nationscape-data-set https://usa.ipums.org/usa/index.shtml https://www.ushistory.org/gov/4b.asp

BBC News. “US Election 2016”. https://www.bbc.com/news/election/us2016/results

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Github, https://github.com/Vickyli6762/STA304PS3

New: Second Nationscape Data Set Release. (2020, October 30). Retrieved November 01, 2020, https://www.voterstudygroup.org/publication/nationscape-data-set.

Team, M. (2020). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 01, 2020, https://usa.ipums.org/usa/index.shtml.

What Factors Shape Political Attitudes? (n.d.). Retrieved November 29, 2020, https://www.ushistory.org/gov/4b.asp

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