Joe Biden With a Huge Advantage Against Donald Trump in 2020 United States Presidential Election

Jixuan Huang & Pinyang Zhou & Maoyuan Gao & Zishu Zhu

November 2nd, 2020

Model

Model Specifics

In order to predict the overall popular vote outcome of the 2020 American federal election, we build a multiple logistic regression model and make use of the post-stratification technique. This model was selected since the outcome is dichotomous with more than one independent variable. It measures the probability of voting for Donald Trump by using age group, gender, employment state, state, household income and race as our predictor variables. Here, we use age group rather than age since we are interested in the relation associated with specific groups rather than individuals. RStudio is the software that we used to run the model.

Our model is given by the formula:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 AG.1 + \beta_2 AG.2 + \beta_3 AG.3 + \beta_4 AG.4 + \beta_5 AG.5$$
$$+ \beta_6 gender.M + \beta_7 ES.N + \beta_8 ES.U + \beta_9 state.1 + ... + \beta_{58} state.50$$
$$+ \beta_{59} hhicome.1 + ... + \beta_{66} hhicome.8 + \beta_{67} race.B + ... + \beta_{72} race.W$$

where $\beta_0 = 1.75308$ is the intercept, which represents log odds of voting for Donald Trump is equals to 1.75308 when age group is 18 and under, gender is female, employment state is employed, state is AK, household income is $250,000 and above, race is American Indian or Alaska Native. $\beta_1 = -0.24737$ represents when age group change from "18 and under" to "18 to 33", log odds of voting for Donald Trump will reduce by 0.24737. The terms β_1, β_2,..., β_{72} are the change in log odds for dummy variables, in which represents the factor of odds that Y=1 within that category of X, comparing with the odds that Y=1 within the reference category.

Post-Stratification

We build a model with the post-stratification to estimate the proportion of voters in favor of voting for Donald Trump. Post-stratification is a way to 're-weight' so that the weighted totals within different cells equal the population totals. Our sample data was post-stratified on age, gender, employment status, state, household income, and race. The weight of the known population in the census data was also post-stratified by these variables. In addition, our sample size is small and there exist problems associated with low statistical power and unbalanced by the representation. Therefore, the post-stratification technique is appropriate to apply in our case and to help to obtain more accurate estimates of the population. We build a model with the post-stratification on sample data and making predictions using this model with data from the census. The variables we include are available in both sample and census data. The reasons why we include these explanatory variables in our cell split are as follows. Voters of different ages and different genders will have distinct opinions on deciding which candidate they are going to support. Thus, age and gender are likely to influence the outcome. We choose the variable "state" because states with few immigrants or

Table 1: Vote Result for Each State

State	Estimate	Result	E.C
CA	0.3908228	Joe Biden	55
FL	0.4776040	Joe Biden	29
GA	0.4938314	Joe Biden	16
IL	0.4353037	Joe Biden	20
MI	0.4550112	Joe Biden	16
NC	0.4661654	Joe Biden	15
NJ	0.4718032	Joe Biden	14
NY	0.3669131	Joe Biden	29
PA	0.4714890	Joe Biden	20
TX	0.5743236	Donald Trump	38

relied on traditional industries tend to be more conservative, as policies such as loosening the immigration restriction or pursuing new energy would potentially harm their interests. Thus, "state" is a variable that might affect each voter's decision. Moreover, Voters of different races could have different opinions on their preferred candidate since two parties take different approaches to deal with racial issues. Income level also would affect voters' choices because candidates aim to exercise their distinct taxation regulations and health insurance plans on households with different income levels.

Results

We estimate that the proportion of voters in favor of voting for Donald Trump is 0.253 and this is the post-stratification estimate for electoral votes. The group of presidential electors is formed for the purpose of electing the president and the electors come from the 50 states and the federal district. Winning the popular vote does not ensure presidential election victory, this is achieved by the electoral college system. Each state gets some electors and if a candidate wins more than 50% of the vote, the state will grant all their electoral votes to this candidate. In "Table 1" above, we list the estimates, vote result, and the numbers of electoral votes for the states that have more than 10 electoral votes. In the total 538 electoral college votes, there are 136 electors vote in favor of Donald Trump.

Discussion

Summmary

The aim of this report is to fit a logistic regression model to estimate the vote rate for the 2020 US election. The full data was collected from Democracy Fund + UCLA Nation cape and the post stratification data was collected from American Community Survey. The data was cleaned by selecting several rows that are of our interest. The rows we selected are age, gender, state, employment, household income, and race. Also, person weight is included since it plays an essential role in the election. A logistic regression model was fitted on the survey data and then was applied to the census for prediction.

Conclusion

In conclusion, we predict that Joe Biden will receive 74.72% and Donald Trump will receive 25.28% of the electoral college votes. Therefore, we predict that Joe Biden will win the election.

Weakness & Next Steps

We measured the result by comparing the probability that a state would vote for Donald trump, which has drawbacks. Notice that the average probability sometimes does not reflect the result precisely. For example, if there are three voters in a state, the probability of them to vote for Donald Trump is 40%, 40% and 73%, then the average of them voting for Donald Trump is 51%. Therefore, we conclude that all the members in electoral college of this state will vote for Donald Trump. But in fact, Donald Trump receives only 1/3 votes, which means this state actually supports Joe Biden more. Also, we did not apply any methods to evaluate the model. This might result in a biased or inaccurate model.

Therefore, in next step we should improve our model by applying multi level logistic regression to get a more confident prediction. For model evaluation and diagnostic, we can apply likelihood ratio test, pseudo R^2 or Hosmer-Lemeshow test to evaluate the goodness of fit. Since these two candidates may have very similar number of votes, we have to make our prediction more confident in order to make a more confident prediction. In addition, we will pay attention to the result of the election and how each electoral college vote and correct our misjudgements accordingly.

Reference

[1] Press, C., Finance, Y., & Newsweek. (2020, October 30). New: Second Nationscape Data Set Release. Retrieved November 03, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

[2] Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 03, 2020, from https://usa.ipums.org/usa/index.shtml

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

[4] Kun Ren and Kenton Russell (2016). formattable: Create 'Formattable' Data Structures. R package version 0.2.0.1. https://CRAN.R-project.org/package=formattable

[5] Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version2.3.1. https://CRAN.R-project.org/package=haven

[6] Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1.https://CRAN.R-project.org/package=kableExtra

[7] Grace-Martin, K., João, Anita.a, & Dina. (2018, December 14). How to Interpret Odd Ratios when a Categorical Predictor Variable has More than Two Levels. Retrieved November 03, 2020, from https://www.theanalysisfactor.com/odds-ratio-categorical-predictor/

[8] US election 2020: What is the electoral college? (2020, October 27). Retrieved November 03, 2020, from https://www.bbc.com/news/world-us-canada-53558176