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

Ruotong Wang & Shengnan Mao

November 2, 2020

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

Ruotong Wang and Shengnan Mao

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

The study is to predict the overall popular vote of the 2020 American federal election, and we use a post-stratification technique to do the calculation. We will use a linear regression model to model the proportion of voters who will vote in 2020 between Donald Trump and Joe Biden. This is a basic statistical model, so this would only help us to know the predicted result but not the real result. We will only use all of the variables (age group, gender, education, employment, state, and race), which are recorded as factors, to model the probability of voting in 2020. The simple linear regression model I am using is:

$$vote_{2020} = \beta_0 + \beta_{agegroup} x_{agegroup} + \beta_{gender} x_{gender} + \beta_{education} x_{education} + \beta_{employment} x_{employment} + \beta_{state} x_{state} + \beta_{race} x_{race} + \epsilon$$

Where $vote_{2020}$ represents the proportion of voting for Donald Trump in 2020. β_0 represents the intercept of the linear regression model, the probability of voting for Donald Trump while voter's age is less than 20, the gender is female, education state is between grade1 to 3, the employment status is employed, the state is Alaska, the race is American Indian or Alaska Native. $\beta_{agegroup}$ represents that when other variables do not change, the voter increases one unit in the age group, the probability of voting for Donald Trump is expected to increase by $\beta_{agegroup}$. β_{gender} is the dummy variable of gender. We could only see β_{male} in model1 because that female is seen by a reference for baseline. While the gender changed from female to male and other variables do not change, the probability of voting for Donald Trump is expected to increase by β_{male}. $\beta_{education}$ is the dummy variable of education. It represents that while other variables do not change, every time the voter changes into another state, the probability of voting for Donald Trump is expected to increase by the corresponding $\beta_{education}$. $\beta_{employment}$ is also a dummy variable of employment. It represents that when the voter changes in the corresponding employment groups, the probability of voting for Donald Trump is expected to increase by the corresponding $\beta_{employment}$. Like the above variables, β_{race} is also a dummy variable. It represents that when the voter changes in the corresponding race groups, the probability of voting for Donald Trump is expected to increase by the corresponding β_{race}.

Post-Stratification

In order to estimate the proportion of voting for Donald Trump in 2020 we need to perform a post-stratification analysis. The post-stratification is a statistical technique to estimate for the under-representative study population in the target population, and it is the method for correcting the sampling weights. Since the sampling data we use in this study does not cover every voter within the nation, the post-stratification can help to provide a better coverage of the population. Thus, post-stratification is a useful technique for us to estimate the probability. The explanatory variables we use in this model include a voter's age, gender, employment, race ethnicity, education, and state. These variables exist in both survey data and census data, and these variables are all useful factors for us to model the probability of the election. The response variable is vote_2020 from the survey data. It is noticeable that explanatory variables all belong to demographic data. Also, we create the cell in the sample data, and we choose race and gender which has 14 cells in total. Race and gender can be seen as two important factors because we can see how these certain variables relate to the voting for Donald Trump, and we can see which type of race would be in favour of voting for Donald Trump in this case.

Results

Through all analysis and calculations, we get that Donald Trump has 317 votes, and Joe Biden has 221 votes from the total 538 electoral colleges. $\hat{vote}_{2020}$ is the predicted result with the proportion of voting for Donald Trump from the two datasets. Thus, we estimate that the proportion of voters who are willing to vote for Donald Trump to be 0.589. In the meanwhile, we can also estimate that the proportion of voters who prefer to vote for Joe Biden to be 0.411. We do the estimation based on the post-stratification technique under the linear regression model, which is looking for the proportion of voting for Donald Trump in 2020 and consists of each voter's age group, gender, education, employment, state, and race.

Discussion

To predict the proportion of voting for Donald Trump in the 2020 American federal election, we need to use the census data collected in 2018 and the data of the survey in 2020. There are many variables in those two datasets, but we choose some variables that we think are significant to forward study. The variables we choose are the voter's age, gender, race, state, education state and employment state. Based on the chosen variables in the survey 2020, we made a linear regression model. Then we know the impact of each variable to the proportion of voting for Donald Trump is different. We also made a post-stratification analysis based on the model we built before. From this analysis, we know that the electoral college in each state would vote for which candidate. Then we could predict the final result of the 2020 American federal election.

Overall, based on the sample data cleaning and post-stratification analysis of voting for Donald Trump, the proportion of voters who vote for Trump is estimated to be 0.589. Thus, we predict that Donald Trump will have a larger probability to win this year's election.

Weaknesses

In terms of the data, the survey data is not large enough, and then it can not lead to conducting an accurate estimation. Also, since the survey was done in June, it may lose some time-effectiveness. There are two special states - Maine and Nebrask, they have their own ways to calculate the election results, but we assume these two states are as same as other states in this study. In terms of our study, the cell splits in the sample data from the post-stratification part may be not good enough, then it is going to lead to an unrepresentative calculation of the total votes based on the personal weights.

Next Steps

Regarding the above weaknesses, we can collect and use large survey data because it can allow a better analytic result. Also, the data should be latest and accurate. Since Maine and Nebrask have different ways

of election decision, we can build and calculate votes based on states separately instead of using the same model, and then it will let the prediction be closer to the actual election results. When splitting cells, we can split them into thinner parts, and it can bring a more detailed model and more precise calculation results.

References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). URL: https://www.voterstudygroup.org/publication/nationscape-data-set

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