Estimation of 2020 United States presidential election

Zezhou Han, Yingchen Tan, Xueqi Wang, and Xiwen Ran

Nov 2nd, 2020

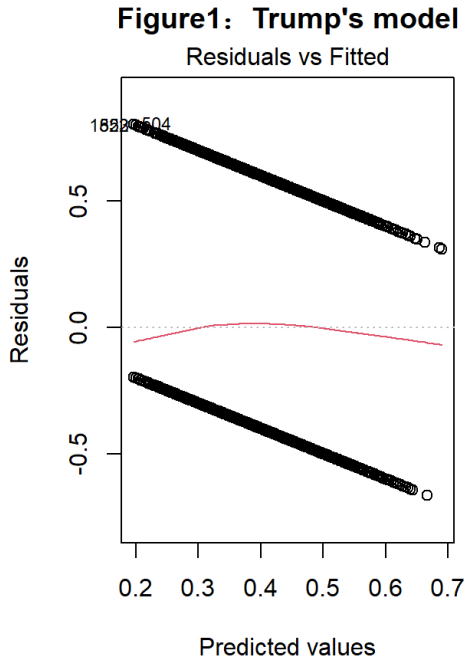
Model

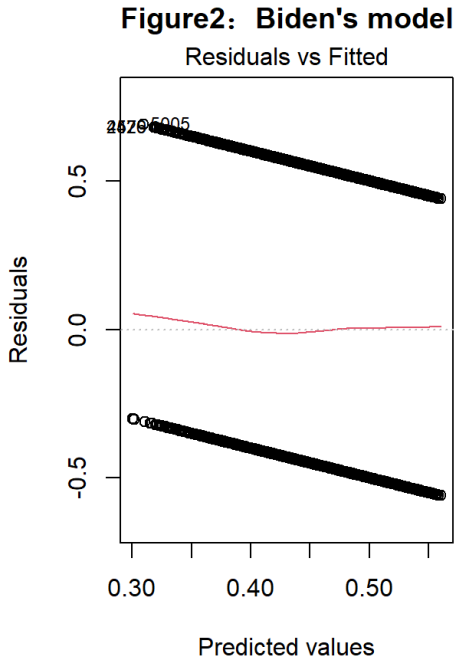
Table 1: Summary of Trump's model

Trump	Estimate	pvalue
age	0.0038623	< 2e-16
Male	0.1310671	< 2e-16
Northeast	-0.0253746	0.19261
South	0.0460387	0.00636
West	0.0460387	0.07405

Table 2: Summary of Biden's model

Biden	Estimate	pvalue
age	-0.001493	8.31e-05
Male	-0.098060	1.37e-14
Northeast	0.017776	0.373
South	-0.034527	0.046
West	0.011539	0.549





We used R to create two logistics regression models to estimate the probability of people choosing to vote for Trump or Biden. We considered people's age, gender and region are the important factors affecting people voting, so we chose the variables of age, gender, and census_region in our survey_data to be the explanatory variables. From Figure 1 and 2, we can see the red lines does not go over the data points range and they are not straight line, which means using logistics regression assumption for both probabilities are appropriate. Here are our formulas in 4 digits:

$log(p_{Trump}/(1-p_{Trump})) = 0.0039X_{age} + 0.1311X_{Male} + (-0.0254)X_{Northeast} + 0.0460X_{South} + (-0.0335)X_{West} + 0.1609 + \epsilon$
$log(p_{Biden}/(1-p_{Biden})) = (-0.0015)X_{age} + (-0.0981)X_{Male} + 0.0178X_{Northeast} + (-0.0345)X_{South} + 0.0115X_{West} + 0.5690 + \epsilon$
for all $X_{age} \geq 0$, X_{Male}=0 or 1, X_{South}=0 or 1 and X_{West} =0 or 1.

In our formula, p_{Trump} means the probability of people choosing to vote for Trump, while p_{Biden} means the probability of people voting for Biden. X_{Male} is the value of if a person is male. $X_{Northeast}$ is the value of a person's region is in the Northeast of America, X_{South} is the value of a person's region is in the South, and X_{West} is the value of a person's region is in the West. The age span of the sample is large, which is from 18

to 93. Since people who vote for Trump may be more conservative, then the older citizens may prefer to vote for Trump, and those who vote for Biden had the opposite. Also, Trump won votes in more states in 2016 (CNN, 2017), then the region should affect people's voting decisions. As most of the blue-collar workers in manufacturing are men, gender also may measure the probability of people voting Trump. From the Table 1 and 2, we can see the p-values of two models for the null hypothesis $X_{Male} = 0$, $X_{age} = 0$ and $X_{South} = 0$ are smaller than 0.05, so we can reject them. It means that people's age, gender and whether they are in the South region are correlated to the probability of people to vote for Trump or Biden. However, the p-value of $X_{Northeast}$ and X_{West} for both models is greater than 0.05, which explains whether people living in the Northeast or the West region may not affect the probability of people to vote for Trump or Biden.

Post-Stratification

Post-stratification is a statistical technique which is common for survey analysis. It is useful for us to know the weight of incorporating population distributions of variables with estimates. In this case, we first grouped the data by variable region, age and gender separately as cells. Thereinto, we clean the region in census data from 9 divisions to 4 region areas, as same as survey data (Northeast, Midwest, South and West). Then we used essentially the estimate $\hat{y}_j$ for each cell, and N_j representing the population size of the jth cell. According to the 2016 Presidential Election Results updated by the New York Times in 2017, although Trump's total number of votes was less, there were more regions that supported him that led to his victory. This was the main reason we would like to use the region data as cells, which could be effective for this year's election predictions.

Results

The results are $\hat{y}^{PS}_{Trump}$=0.6041376, and $\hat{y}^{PS}_{Biden}$=0.6079877.

We estimated that the proportion of voters in favour of voting for Trump to be 0.604. This was based on our post-stratification analysis of the proportion of voters in favour of Trump modeled by a logistic model, which accounted for age, gender and region. Also, we used the same technique as the logistic model and the same variables to estimate the proportion of voters in favour of voting for Biden to be 0.608. Basically speaking, it was obvious that the two sides were evenly matched in such a huge data set. Things get interesting, and both sides have almost the same probability of winning the election.

Discussion

Results from this study indicated that older people and males were more likely to vote for Trump. The study has shown Biden with an advantage among young voters and female voters. This study further confirmed that the population based on different regions can affect the percentage of voters who would vote for Trump. Voters in the southern region were more likely to support Trump.

Based on the results we have achieved, Trump's chance of winning was about the same as Biden. The estimated proportion of voters in favour of voting for Biden was 0.608, and the estimated proportion of voters in favour of voting for Trump was 0.604. The prediction we made was that Trump would probably win the primary vote. This is because the 2016 Presidential Election results, reporting by CNN politics (2017), showed that despite Clinton won the popular vote with 48.5 percent, Trump won the 2016 election due to the electoral votes he received. We predicted that Trump would win the election due to the electoral college outcome. However, we still believed that it was difficult to conclude who would win because of the slight difference between two estimated proportions.

Weaknesses & Next Steps

A weakness of this study is that it only examined four main regions in America. We did not have enough data for detailed regional predictions. The analysis would become more reliable if we determine all of the variables that are related to the outcome. Since there are nine divisions in America, the results can be improved by producing more precise predictions.

The results we have delivered so far suggest the need for further research of the actual election results. The American Association for Public Opinion Research (AAPOR) conducted a post-hoc analysis of the 2016 presidential election for summarizing the accuracy of 2016 pre-election polling data, as well as reviewing the variation by different prediction methods. All things considered, a post-hoc analysis can help us better understand the variables influencing the election outcome and improve estimation accuracy in future elections.

References

1. 2016 Presidential Election Results: Donald J. Trump Wins. (2017, August 9). Retrieved November 02, 2020, from https://www.nytimes.com/elections/2016/results/president (https://www.nytimes.com/elections/2016/results/president)
2. AAPOR to Examine 2016 Presidential Election Polling. (2016, November 9). Retrieved November 02, 2020, from https://www.aapor.org/Publications-Media/Press-Releases/Archived-Press-Releases/AAPOR-to-Examine-2016-Presidential-Election-Pollin.aspx (https://www.aapor.org/Publications-Media/Press-Releases/Archived-Press-Releases/AAPOR-to-Examine-2016-Presidential-Election-Pollin.aspx)
3. Presidential Results. (2017, February 16). Retrieved November 02, 2020, from https://www.cnn.com/election/2016/results/president (https://www.cnn.com/election/2016/results/president)
4. RStudio [Computer software]. Retrieved November 02, 2020, from https://rstudio.cloud/projects (https://rstudio.cloud/projects)
5. 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
6. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=9ffb9163-8cd2-4d9b-ae75-acd8845e7891 (https://www.voterstudygroup.org/downloads?key=9ffb9163-8cd2-4d9b-ae75-acd8845e7891).

Appendix

This file was made using the R markdown package. All code used in this paper can be accessed from within the code blocks of the markdown document. All the methodology we have developed are learned in STA304H1F of University of Toronto (St. George campus).