Estimation of America's Election Using Logistic Model

Xinran Gu

2020/11/02

Model

This report is focusing on forecasting the 2020 US election and a logistic regression model will be used for analysis. Also, in the logistic model, age and gender are the two independent variables.After we established the model, we will employ a post-stratification technique with the same variables. In the following subsections I will describe the model specifics and the post-stratification calculation.

Model Specifics

I will be using a logistic regression model to estimate the proportion of votes Donald Trump will get. This is a naive model. I will only be using age and gender , to model the probability of voting for Donald Trump. Hence the model consists a numeric variable and a categorical variable. The reason why I choose these variables is because I believe people with different age and gender will encounter different situations in society which will affect their political party preference. Moreover, age and gender are the two variables that will not be affected by the coronavirus pandemic. I think this could ensure the accuracy of the estimation. Then,I choose to use logistic model because we are trying to predict the proportion of votes which will go to Donald Trump. If we use a linear model, we will not be able to obtain a probability as output. Moreover, the result of US election is binary, it is about whether Donald Trump will be elected again or not. Hence, logistic model is suitable to use in this report.The logistic regression model I am using is:

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

The $log(\frac{p}{1-p})$ on the right-hand side represents the proportion of voters who will vote for Donald Trump and it could also be called as log odds.The numerator p is the probability Donald Trump gets the vote and the denominator $1-p)$ is the probability Donald Trump does not get the vote. On the left hand side, β_0 represents the intercept of the model, and is the probability of a female voting for Donald Trump at age 0. We need to notice that, as we will use glm function in R to run the model, R will automatically read "Male" as 1 and "Female" as 2 for gender variable.Additionally, β_1 and β_2 represent the change in log odds for every one unit increase in x_{age} and x_{gender}.

Post-Stratification

A post-stratification analysis is performed in this report to estimated the proportion of people voting for Trump.Post-stratification is a technique that we divided sample data into strata after we randomly selected data from the population. We often use this technique when it is hard place the data into their correct strata until we finish sampling. Also, this technique allows us to minimize bias cause by underrepresented group in the population as it adjusting sampling weights. In this report, I create cells based off different ages and genders. For example, there will be 2 cells for age 18, one is for female who is 18 and the other one is for male who is 18. There are total 162 cells in this analysis according to this method.Then, I will use the logistic model described in the previous Model section to estimate the proportion of voters in every cells. After that, we will use the formula $\hat{y}^{ps} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ to estimate the proportion of voters in favor of Trump. The numerator $\sum N_j \hat{y}_j$ is the sum of estimates we got from the logistic model for each cell times the population of each cell. The denominator is the population of all cells.

Data cleaning

Since the census data we use to create strata containing data from 2014 to 2018, I have removed people who are younger than 16. This is because people who are under 18 are not allowed to vote in the election, and the remaining people in the data should all are over 18 years old in 2020.

Results

Table1	Estimate	p-value
$\hat{\beta}_0$	-1.1113	p<0.0001
$\hat{\beta}_1$	0.0145	p<0.0001
$\hat{\beta}_2$	0.5107	p<0.0001

The above table1 is a summary of logistic model for estimating proportion of people voting for Trump. From this table, we could obtain the model representation: $\hat{y} = -1.1113 + 0.0145x_{age} + 0.5107x_{gender}$.Based on this logistic model which accounted for age and gender, I have performed a post stratification analysis. The result of the analysis showed that the proportion of votes which will go to Donald Trump will be 0.4632.

Discussion

Summary and Conclusion

We have first built a logistic model to predict the proportion of people voting for Trump using two variables age and gender. As the results section said, the model representation is $\hat{y} = -1.1113 + 0.0145x_{age} + 0.5107x_{gender}$. This means the proportion of people voting for Trump will increase by 0.0145 when people's age increased by 1 year with other predictors hold constant. Additionally, males are more willing to vote for Trump as the proportion will increase by 0.5107.Moreover, as the p-vale for both our predictors are smaller than 0.0001,it suggests that age and gender do have significant influence the the proportion of votes that Donald Trump will get.

Then we have used this model to do a post stratification analysis by diving sample data into 162 cells based on age and gender. We have adjusted the weights of each cell by the cell's population and then get the final prediction $\hat{y}^{ps}$.Since the estimated proportion of voters in favor of voting for Trump is 0.4632, I predict that he will lose the election.

Weaknesses

The survey data used in this report is from June 25th 2020. Although this is the newest data set we have, the half year gap till the election does decrease the accuracy of the analysis. This weakness is even more sever this year, especially with the acceleration of the impact of the coronavirus pandemic in America. Voters might change their decision due to the actions of Donald Trump have done in recent months. Moreover, the model that used in this report is a naive model. The factors that will affect the proportion of people's willingness to vote for Trump are complicated. Especially, the coronavirus pandemic could cause many accidents.If we could add some variables that related to coronavirus like the scale of anxiety towards coronavirus pandemic will definitely increase the power of the analysis.

Next Steps

We will need to compare our results to the actual election results in order to check the accuracy of our analysis. A follow-up survey that ask who and why voters are voting to the candidates after the election should be conducted. This will help us to identify other variables that affects voter's decision, we could improve our model for future estimations of elections.However, we need to know that some variables might be only significant in this specific period due to coronavirus pandemic.

References

1.Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17,2019(version 20200814).Retrieved from [https://www.voterstudygroup.org/publication/2019-voter-survey-full-data-set].

2.Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [2014-2018, ACS 5-year]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

3.Dassonneville, R., & Tien, C. (2020). Introduction to Forecasting the 2020 US Elections. PS: Political Science & Politics, 1-5. doi:10.1017/S104909652000147X

4.Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

5.The Pennsylvania State University. (2020). 6.3 - Poststratification and further topics on stratification: STAT 506. Retrieved November 01, 2020, from https://online.stat.psu.edu/stat506/lesson/6/6.3

6.Abramowitz, A. (2020). It's the Pandemic, Stupid! A Simplified Model for Forecasting the 2020 Presidential Election. PS: Political Science & Politics, 1-3. doi:10.1017/S1049096520001389

7.Alexander, R., & Caetano, S. (2020, October 22). 01-data_cleaning-post-strat1 [R].

8.Alexander, R., & Caetano, S. (2020, October 22). 01-data_cleaning-survey1 [R].

9.Alexander, R. (2020, May 17). Telling Stories With Data. Retrieved November 01, 2020, from https://www.tellingstorieswithdata.com/

Appendix

1.Code and data supporting this analysis is available at:"https://github.com/LiviaSta/Prediction-of-US-Election".