Predict the Overall Popular Vote of the 2020 American Federal Election

Yingren Luo(1004194873), Li Si Xuan (1004941325),Youwen Xu (1006675493), Alvin Gong(1005121556)

2020/11/02

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election based on individual-level survey data (5) and post-stratification data (7). To do this we are employing a post-stratification technique. In the following sub-sections, we will describe the model specifics and the post-stratification calculation.

Model Specifics

We will be using a logistics regression model to model the probability of voters who will vote for Donald Trump. "Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary response variable" (4). The reason we choose this model is that the vote intention variable is binary, and "logistic regression is suitable when the outcome of interest is binary" (4).

We will be using age (numeric variable), gender (categorical variable), Asian Pacific race (dummy variable), white race (dummy variable), black race (dummy variable) and other race (dummy variable) to model the probability of voting for Donald Trump. The logistics regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 x_{age}+\beta_2 x_{genderMale}+\beta_3 x_{race\_asian\_pacific}+\beta_4 x_{race\_white}+\beta_5 x_{race\_black}+\beta_6 x_{race\_other}+\epsilon$$

Where p represents the probability of voting for Donald Trump. β_0 represents the intercept of the model, in this case, there is no practical interpretation for the intercept since only people over 18 have the right to vote. β_1 represents the slope of the model. So, for every additional unit increase in age, we expect the log odds of the probability of voting for Donald Trump to increase by a β_1, given other predictors hold constant. β_2 represents the log of odds ratio between the female group and male group. β_3 represents the log of odds ratio between the people who are Asian Pacific race and the people who aren't Asian Pacific race. β_4 represents the log of odds ratio between the people who are white race and the people who aren't white race. β_5 represents the log of odds ratio between the people who are black race and the people who aren't black race. β_6 represents the log of odds ratio between the people who are other race and the people who aren't other race.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump, we need to perform a post-stratification analysis. "Post-stratification is a method for adjusting the sampling weights, usually to account for underrepresented groups in the population." It is useful because it can decrease bias and tend to result in smaller variance estimates. (6)

Here we create cells based off different ages, gender and race. We choose "age" and "gender" because they are likely to influence voter outcome because of people with different ages and gender have different ideas and political attitudes. Additionally, we choose "race" because it is also likely to influence voter outcome because of people with different race have different ethnic culture, which might make them have different political attitudes. For instance, especially for people who are black, their votes must depend on presidential candidates' attitudes on racism. Also, our variable is not including *vote_2016* because it is not available in the census data.

Using the model described in the previous sub-section we will estimate the proportion of voters in each age, gender and race bin. We will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

Results

Table 1 - Summary table of logistics regression model

Coefficients	Estimate	Std. error	z value	Pr(>
(intercept)	-0.237873	0.275066	-0.865	0.387157
age	0.007514	0.001803	4.168	3.07e-05
genderMale	0.435979	0.058730	7.423	1.14e-13
race_asian_pacific	-1.059570	0.297177	-3.565	0.000363
race_white	-0.156013	0.266489	-0.585	0.558253
race_black	-2.259823	0.293754	-7.693	1.44e-14
race_other	-0.894669	0.289589	-3.089	0.002005

From the *Table 1*, we can see that the estimated logistics regression model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0+\hat{\beta}_1 x_{age}+\hat{\beta}_2 x_{genderMale}+\hat{\beta}_3 x_{race\_asian\_pacific}+\hat{\beta}_4 x_{race\_white}+\hat{\beta}_5 x_{race\_black}+\hat{\beta}_6 x_{race\_other}$$

$$= -0.237873 + 0.007514 x_{age} + 0.435979 x_{genderMale} - 1.059570 x_{race\_asian\_pacific} - 0.156013 x_{race\_white}$$

$$-2.259823 x_{race\_black} - 0.894669 x_{race\_other}$$

The $\hat{\beta}_0$ is the estimated intercept of the estimated logistic regression model, which is -0.237873. The standard error of $\hat{\beta}_0$ is 0.275066. The null hypothesis, H_0 is $\beta_0 = 0$, while the alternative hypothesis, H_a is $\beta_0 \neq 0$. Since the p-value is 0.387157, which is larger than 0.05, there is no evidence against the null hypothesis, H_0.

The $\hat{\beta}_1$ is the estimated coefficient of age of the estimated logistic regression model, which is 0.007514. The standard error of $\hat{\beta}_1$ is 0.001803. The null hypothesis, H_0 is $\beta_1 = 0$, while the alternative hypothesis, H_a is $\beta_1 \neq 0$. Since the p-value is 3.07e-05, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_2$ is the estimated coefficient of gender, whether is male or not, of the estimated logistic regression model, which is 0.435979. The standard error of $\hat{\beta}_2$ is 0.058730. The null hypothesis, H_0 is $\beta_2 = 0$, while the alternative hypothesis, H_a is $\beta_2 \neq 0$. Since the p-value is 1.14e-13, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_3$ is the estimated coefficient of race, whether is Asian Pacific or not, of the estimated logistic regression model, which is -1.059570. The standard error of $\hat{\beta}_3$ is 0.297177. The null hypothesis, H_0 is $\beta_3 = 0$, while the alternative hypothesis, H_a is $\beta_3 \neq 0$. Since the p-value is 0.000363, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_4$ is the estimated coefficient of race, whether is white or not, of the estimated logistic regression model, which is -0.156013. The standard error of $\hat{\beta}_4$ is 0.266489. The null hypothesis, H_0 is $\beta_4 = 0$, while the alternative hypothesis, H_a is $\beta_4 \neq 0$. Since the p-value is 0.558253, which is larger than 0.05, there is no evidence against the null hypothesis, H_0.

The $\hat{\beta}_5$ is the estimated coefficient of race, whether is black or not, of the estimated logistic regression model, which is -2.259823. The standard error of $\hat{\beta}_5$ is 0.293754. The null hypothesis, H_0 is $\beta_5 = 0$, while the alternative hypothesis, H_a is $\beta_5 \neq 0$. Since the p-value is 1.44e-14, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_6$ is the estimated coefficient of race, whether is other or not, of the estimated logistic regression model, which is -0.894669. The standard error of $\hat{\beta}_6$ is 0.289589. The null hypothesis, H_0 is $\beta_6 = 0$, while the alternative hypothesis, H_a is $\beta_6 \neq 0$. Since the p-value is 0.002005, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

Based on the estimated logistics regression model considering the variable, *age*, *gender*, and *race*, we use the post-stratification analysis to calculate the estimated proportion of voters who will vote for Donald Trump, $\hat{p}$, is 0.487377.

Discussion

According to individual-level survey data (5) and post-stratification data (7), we used post-stratification technique to predict the result of the 2020 American federal election. Our model based on *age*, *gender*, and *race*, and we got that the estimated proportion of voters who voted for Donald Trump will be 0.487377. This may mean that Donald Trump is more likely to lose the election though our analysis.

Overall, we predict that Joe Biden may have a higher chance of winning the election, since from the data we have obtained, he has a certain advantage, which is 0.512623. In addition, racism plays a central role in American life. So whether the relationship between races can be handled well is an aspect that people will pay attention to. In a Pew Research poll in June, 35 percent of people have confidence in Trump, and about 48 percent of Biden (1).

Weaknesses

Our data may be biased due to insufficient background information of this organization and then perhaps this organization has a political bias. And we had to remove some variables when we made the model since some of them are not available in the census data, there may not be enough sufficient variables to determine a good model. Therefore, it may affect the performance of the model. If we can add some other variables such as *education*, *foreign-born* or *state*, our data and model should be improved, so we can make better predictions.

Next Steps

We will compare our prediction with the actual election results. Then we will do a post-hoc analysis to check what are the shortcomings of our data or what factors we have overlooked, so that we can better improve our prediction in future elections. For example, we can do a survey on the reasons why people choose Trump or Biden and classify the reasons, such as international relations or policies towards the people. Then we can analyze which ones have the main influence on the election results, thereby improving our prediction in the future.

References

1. Beason T, (2020, August 18). Trump and Biden couldn't be more different on the complicated issue of race. Retrieved October 31, 2020, from https://www.latimes.com/politics/story/2020-08-06/trump-biden-race-policy

2. Caetano S, Alexander R. (2020). 01-data_cleaning-survey1.R

3. Caetano S, Alexander R. (2020). 01-data_cleaning-post-strat1.R

4. Caetano, S. (2020). Introduction to Logistic Regression. Lecture.

5. Finance, Yahoo, et al. New: Second Nationscape Data Set Release. 28 Oct. 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set.

6. Poststratification — Poststratification for survey data. (n.d.). Retrieved October 30, 2020, from https://www.stata.com/manuals13/svypoststratification.pdf

7. Team, MPC UX/UI. "U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH." IPUMS USA, usa.ipums.org/usa/.

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

Appendix

GitHub link: https://github.com/Yingren-Luo/STA304-PS3