Prediction of the Overall Popular Vote of the 2020 American Federal Election Using Logistic Regression Model With Post-stratification

Peiyu Li

November 2, 2020

Code and data supporting this analysis is available at:

https://github.com/PeiyuBellaLi/STA304-PS3

Model

I am interested in predicting the overall popular vote of the 2020 American federal election ("2020 United States presidential election", 2020) based on a survey result. To make the prediction more accurate, I am using a post-stratification technique. I will first build a model based on the survey data, then used this model to predict the result for the census data which is divided into different cells based on my predictor variables. The calculation of the final result and the specific steps will be discussed in the following sub-section.

Data Cleaning Process

To do the post-stratification for the survey and census data, I need to match the variables levels and names in both data set. Initially, the variable race_ethnicity contains the following categories in the survey data: <White, Black, or African American,Asian (Asian Indian),Asian (Chinese), etc.>, and in the census data(race):<white, black/african american/negro, american indian or alaska native, chinese, etc.>. Then I combine some of the categories for both variables and create a new variable names race_new with levels <white, black, american indian or alaska native, other asian or pacific islander, chinese, etc.>. Similarly, I create the other two variables 'sex' and 'edu' with categories <male, female> and <8th grade or less, high school, some colleges, 5+ years of college> to map these variables in the survey and census data.

In the survey data, I keep the observations who were registered for this election and clearly stated they were going to vote, and their intention to vote are only between Trump and Biden. In this way, the model I will build will be more accurate and reasonable. For the census data, I remove people who are not eligible to vote (i.e. less than 18 years old or not a citizen) so that the prediction using the model will also be reasonable.

I divide the observations in the census data into cells based on the combinations of sex, education and race. A new variable names 'n' represents the number of people that fall into each cell. This will be helpful when doing the post-stratification step and calculating the final result.

Model Specifics

I will use R to run a logistic model to predict the proportion of people who will vote for Donald Trump or Joe Biden. The observations I kept were people who are registered to vote and clearly stated they would

vote between Trump and Biden. The predictor variables I will use are sex, race, and education level. The reasons why I choose these variables are I think people in different races are cared about which presidential candidate pursues better policies and with better attitudes towards their own races. Sex and education levels can also affect the ways how people consider this serious problem.

All three predictors are categorical variables. The logistic regression model I am using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{edu_{8thgradeorless}} + \beta_2 x_{edu_{highschool}} + \beta_3 x_{edu_{somecollege}} + \beta_4 x_{race_{black}} + \beta_5 x_{race_{chinese}}$$
$$+\beta_6 x_{race_{japanese}} + \beta_7 x_{race_{otherasianorpacificislander}} + \beta_8 x_{race_{otherrace}} + \beta_9 x_{race_{white}} + \beta_{10} x_{sex_{male}} + \epsilon$$

Where P represents the proportion of voters who will vote for Donald Trump and $log(\frac{p}{1-p})$ is the log odds. Similarly, β_0 represents the intercept of the model and is the probability of voting for Donald Trump from a female American Indian or Alaska native with of education level of 5+ years of college. Additionally, β_1 to β_10 represents the change in the probability of voting for Trump when a person's sex, education level, and race change from female, 5+ years of college and American Indian or Alaska native to other categories.

An alternative model I will build is the same logistic model but only contains the predictors race and education level, because in this case the sex differences may not have much impact on the voting as the other two variables. To choose a better one, I need to compare their AIC value. The AIC value represents the amount of information lost by a model, and we always want a model with a lower AIC value (Rawat, 2017).

To evaluate the performance of the model, the confusion matrix and ROC curve are also helpful (Rawat, 2017). The confusion matrix shows the difference between the observed value and the predicted response, which will also allow calculating of the accuracy of the model. The area under the ROC curve is also an index of accuracy, so the higher these two values are, the better the model is.

Post-Stratification

I will use the post-stratification technique to calculate the final popular vote from the census data using the model. Here I create cells based on different combinations of sex, education, and race. Since the model only contains three predictors, it is not enough to divide the cell just base on one or two predictors. The categories for these variables are not too much, which can make sure the number of observations within each cell is sufficient to calculate the final result.

Basically, post-stratification requires estimating the response within each cell, and then weight it by its proportion to the population size. So I am going to estimate the proportion of voters in each cell using the model I built, and weigh each proportion estimate in each cell by the corresponding population size. I will get the final result by adding those values and divide it by the total number of the population size. This process can be explained by this formula:

$$\hat{y}^{PS} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j}$$

where $\hat{y}$ is the estimate in each cell, N_j is the population size of the j^{th} cell.

In this case, I am looking at the proportion of people voting for Donald Trump, so the $\hat{y}^{PS}$ represents the proportion fo the popular vote for Trump.

Results

Comparing the AIC values for my originally proposed model and the alternative model with only two predictors, the original model has a lower AIC value (5610.969 compares to 5654.628). As a result, the model stated in the above sub-section is the model I will use, and the following Figure 1 shows the summary of this model. All three predictors are significant according to the p-value.

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	0.06042574	0.33848244	0.1785196	8.583150E-01
(edu)8th grade or less	-0.57714537	0.57583268	-1.0022796	3.162086E-01
(edu)high school	0.20415267	0.09849286	2.0727662	3.819405E-02
(edu)some college	-0.13807295	0.08344848	-1.6545891	9.800788E-02
(race_new)black	-2.33350551	0.36147427	-6.4555231	1.078457E-10
(race_new)chinese	-1.61638930	0.48300793	-3.3465067	8.183669E-04
(race_new)japanese	-1.27636837	0.66801174	-1.9106975	5.604346E-02
(race_new)other asian or pacific islander	-0.89011365	0.37950743	-2.3454446	1.900440E-02
(race_new)other race	-1.00922029	0.35821202	-2.8173826	4.841682E-03
(race_new)white	-0.07963269	0.33180165	-0.2400009	8.103296E-01
(sex)male	0.43635257	0.06469129	6.7451519	1.528670E-11

Figure 1: Summary of the model

The equation of this logistic model becomes:

$$log(\frac{p}{1-p}) = 0.06 - 0.58x_{edu_{8thgradeorless}} + 0.20x_{edu_{highschool}} - 0.14x_{edu_{somecollege}} - 2.33x_{race_{black}} - 1.62x_{race_{chinese}}$$
$$-1.28x_{race_{japanese}} - 0.89x_{race_{otherasianorpacificislander}} - 1.01x_{race_{otherrace}} - 0.08x_{race_{white}} + 0.44x_{sex_{male}}$$

The baseline here is a female with an education level of 5+ years of college and is an American Indian or Alaska native. If a voter is a white male who graduated from a high school, then the log odds of the probability of this person will vote for Trump will be increased 0.20 by the variable 'edu', decrease 0.08 by 'race_new' and increase 0.44 by 'sex'.

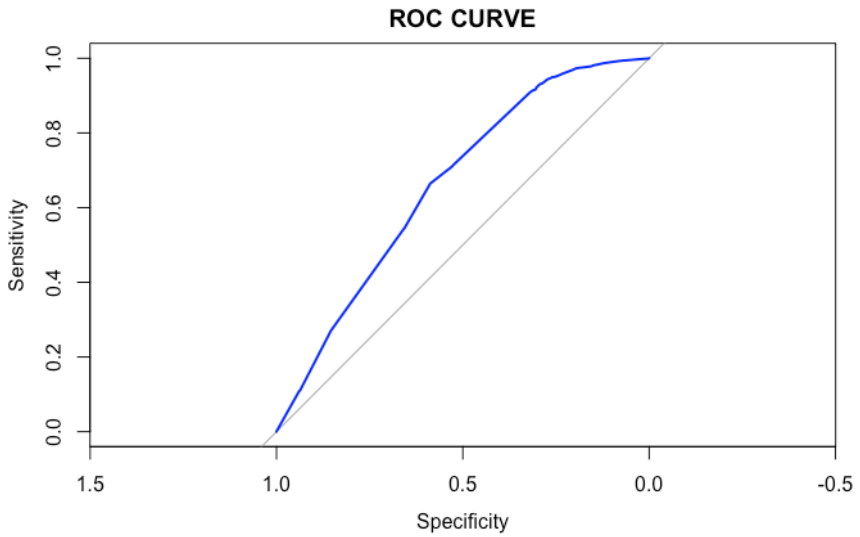

Figure 2: ROC CURVE

The accuracy of this model calculated from the confusion matrix is 0.624 which indicates this model is relatively accurate in prediction since it is higher than 0.5. The ROC curve is shown above (Figure 2)

evaluates the performance of the model, and the area under the curve is 0.6621 which also indicates the accuracy of the model. As a result, I can conclude that this model is significant and can be used to do predictions.

The $\hat{y}^{PS}$ I calculated after doing post-stratification is 0.3999, which means the prediction of the popular vote for Donald Trump is 40%. In other words, most people who are eligible to vote will vote for Joe Biden based on this logistic regression model.

Discussion

The goal for this analysis is to use the survey data obtained from Democracy Fund and UCLA Nationscape to build a regression model to predict the popular vote for the 2020 US presidential election. The census data that is used to do the post-stratification was obtained from IPUMS USA.

In summary, I built a logistic model using variables sex, education and race from the survey data to predict the proportion of vote for Donald Trump in the entire population, and the result shows only 40% of voters will vote for him. Compare to the alternative model with only two predictors (race & education), the original one is better due to a lower AIC value. Also, the accuracy of this model can be seen from the confusion matrix and the area under the ROC curve, which indicates this model is relatively accurate.

Weaknesses

When I cleaned the data, I combined some categories for the predictors in order to match the variables' levels in both survey and census data, so the total number of cells decreased. Thus the prediction based on the model would be less precise. Also, the number of observations in some of the cells may be not enough because there were only 4392 observations in the survey data after doing the cleaning steps. This can weaken the performance of the model and the post-stratification.

In addition, the predictors in this model are the ones that I am interested in, but there are definitely many more factors that could affect who people will vote for. Since the electoral vote result in the US are based on the result in each state, and the number of members of the electoral college from each state is also different, the result I got is not sufficient to predict the actual election result.

Next Steps

When the next election comes, we should collect a larger sample in order to build a better model. It is a good idea to include the variable state in the model, which I did not include in my model. Then we can do a multiple regression model with group-level 'state', and calculate the proportion estimate in each state and weight it by the proportion of members in the electoral college for each state. Thus we can get a prediction of the electoral vote which should be more reliable and useful than the popular vote.

References

2020 United States presidential election. (2020, November 1). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/2020_United_States_presidential_election

Rawat, A. (2017, October 31). Binary Logistic Regression. An overview and implementation in R. Retrieve from https://towardsdatascience.com/implementing-binary-logistic-regression-in-r-7d802a9d98fe

Ruggles,S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. IPUMS USA: Version 10.0 [2018 ACS]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, C.& Vavreck,L. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=deef61f6-ba76-4084-8df4-cef164481d1a