Predict The Outcome of The 2020 U.S. Presidential Election

Ziqing Lyu(1005022093), Xue Shan(1004173262), He Ma(1003080541), Yaozhong Zhang(1003915409)

2020.11.02

Model

In this problem set, we are interested in predicting the popular vote outcome of the 2020 American federal election (https://hdsr.mitpress.mit.edu/). In order to predict the outcome of the popular vote, we will use a post-stratification technique on the census data. In addition, in order to employ this post-stratification technique, we need an estimate of the voter outcome for each post-stratification cell. So we will use a model to calculate this estimate based on the survey data collected from the voter study group. In the following sub-sections, we will describe the model specifics and the post-stratification calculation.

Data Cleaning

Before building the model, we need to clean up and adjust both original datasets. In the survey data, we first removed all the people who did not register and did not vote, and then removed all the missing values, remaining 4508 observations. Then in the census data, people under the age of 18 are also removed because they do not meet the voting regulations. We are interested in using age, race, and state as predictors, so we have to adjust these three variables to the same style in these two data. In order to facilitate the establishment of our model, we added a new variable to the datasets, which divides the age into different groups under 20 years old, 21-40 years old, 41-60 years old, 61-80 years old, and over 80 years old. Then in these two datasets, the observations in state and race are also replaced by the same names. Finally, select the variables that we want from the two data to create two new datasets for modeling and prediction.

Model Specifics

We plan to use a logistic regression model or a logistic multilevel regression model to model the proportion of voters who will vote for Donald Trump. After we tested these two models, we chose the model with the smaller AIC value as our final model. In statistics, AIC is used to compare different possible models and determine which model is best for the data. Therefore, we use the logistic regression model since it has a smaller AIC value. We choose the race, age group, and state as our predictor since these factors might be important factors for us to study whether voters will vote for Trump. Therefore our formula of logistic regression model will be:

$$log(p/(1-p))$$
$$= \beta_0 + (\beta_1 x_1 + \ldots + \beta_6 x_6) + (\beta_7 x_7 + \ldots + \beta_{10} x_{10}) + (\beta_{11} x_{11} + \ldots + \beta_{60} x_{60})$$

Where p represents the probability of voters who will vote for Donald Trump. In addition, $\beta_0$represents the intercept of the model, and $x_1 ... x_6$ represent the dummy variables for the race, $x_7 ... x_{10}$ represent the dummy variables for the age group, and $x_{11} ... x_{60}$represent the dummy variables for the state. For β_1, it means holding all other explanatory variables in the model fixed, if the voter's race changes from baseline variable to dummy variable x_1, log odds of the probability of voter will vote Donald Trump will increase by β_1. Also, other coefficients have similar meanings.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump, we will perform a post-stratification analysis. Post-stratification is a commonly used technique in survey analysis to incorporate the overall distribution of variables into survey estimates. Here we choose the variable "state" to create the cells. Since in the American electoral system, which candidate gets the most votes in each state, then that candidate will get the votes of all the electors in the state. Therefore, "state" is likely to influence voter outcomes. Furthermore, we will use the model described in the previous sub-section to estimate the proportion of voters in each state. Then we will weigh each proportion estimate within each cell by the respective population size of that cell. Finally, sum those values and divide that by the entire population size, we will get the proportion of voters who will vote for Donald Trump.

Results

Model

Here is our summary of our model with some none significant variables:

Table1: summary table for the logistic model with significant variables on predicting the probability of voters who will vote for Donald Trump

```
# A tibble: 12 x 5
   term                               estimate std.error statistic  p.value
   <chr>                                 <dbl>     <dbl>     <dbl>    <dbl>
 1 raceBlack, or African American       -2.45      0.363     -6.76 1.38e-11
 2 raceChinese                          -1.62      0.485     -3.35 8.12e- 4
 3 raceOther asian or pacific islander  -0.795     0.382     -2.08 3.74e- 2
 4 raceOther race                       -0.947     0.360     -2.63 8.54e- 3
 5 Age_group21 to 40                     0.649     0.182      3.56 3.76e- 4
 6 Age_group41 to 60                     0.978     0.183      5.33 9.70e- 8
 7 Age_group61 to 80                     0.804     0.186      4.33 1.50e- 5
 8 Age_groupabove 80                     1.01      0.392      2.57 1.00e- 2
 9 stateCT                              -2.34      1.17      -2.00 4.55e- 2
10 stateMA                              -2.34      1.16      -2.03 4.27e- 2
11 stateNM                              -2.47      1.25      -1.97 4.87e- 2
12 stateVT                              -4.03      1.54      -2.61 8.97e- 3
```

As we can see from Table 1, we only include the variables in which P-value is less than 0.05, so we get the intercept is not significant since the corresponding P-value is more than 0.05. There are some extremely significant variables with P-value less than 0.01, which are raceBlack, or African American, raceChinese, raceOther race, Age_group21 to 40, Age_group41 to 60, Age_group61 to 80 and stateVT. Therefore we will use these variables as predictors, our logistic regression model will be:

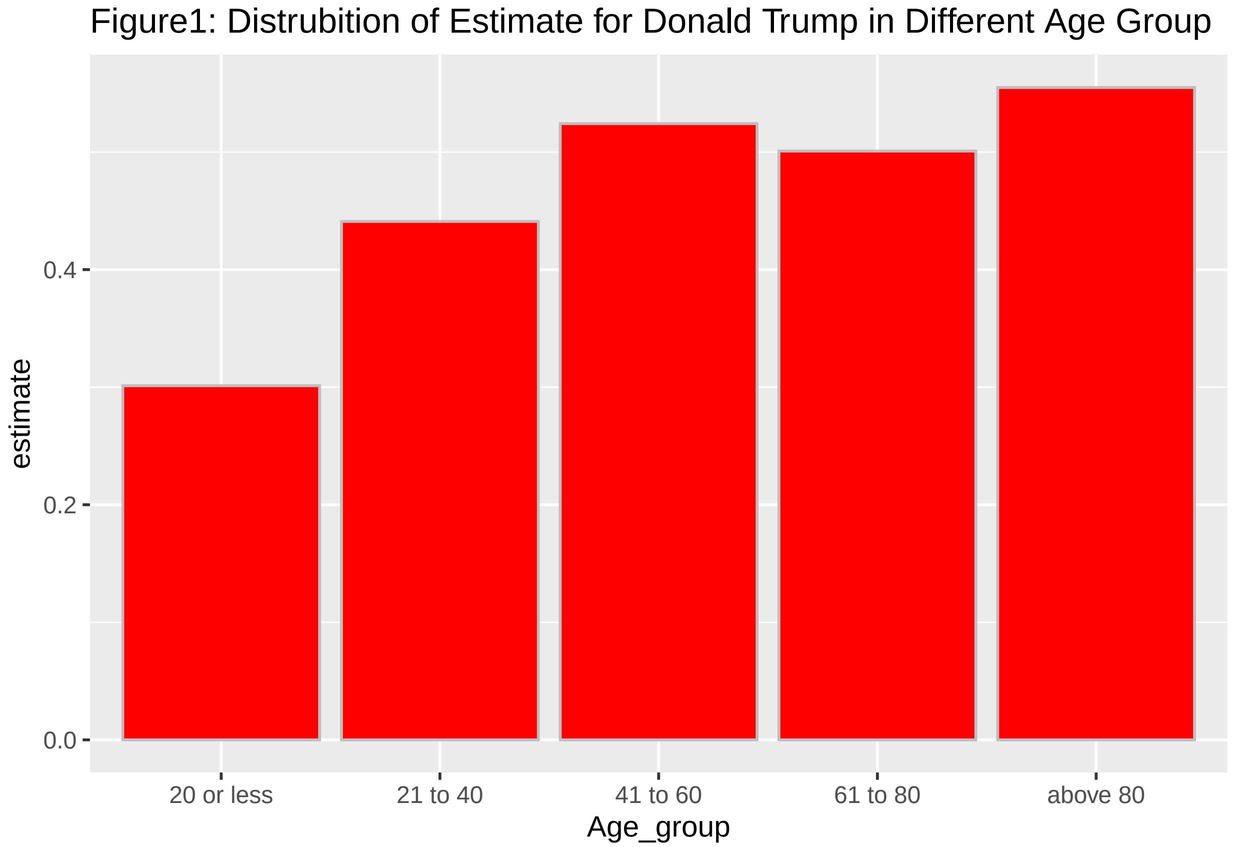
$$log(p/(1-p))$$

$$= \beta_0 + (-2.45) raceBlack, orAfricanAmerican + (-1.62) raceChinese + (-0.947) raceOtherrace + 0.649 Agegroup21to40 + 0.978 Agegroup41to60 + 0.804 Agegroup61to80 + (-4.03) stateVT$$

The following is our interpretation of the model: $\beta_1$is -2.45, which means keep everything else fixed, when the voter's race changes from raceAmerican Indian or Alaska Native to raceBlack or African American, log odds of the probability of voter will vote Donald Trump will decrease by 2.45. Also, β_2 and $\beta_3$have a similar meaning for the race. On the age group, $\beta_4$is 0.649, which means keep everything else fixed, when the voter's age group changes from Age_group20 or less to Age_group21 to 40, the log odds of the probability of voter will vote Donald Trump will increase by 0.649. Also, $\beta_5$and β_6 have similar meanings, and as well as for β_7. Furthermore, based on our post-stratification analysis of the proportion of voters in favor of Donald Trump modeled by a logistic model, which we describe above, we estimate that the proportion of voters voting for Donald Trump to be 0.483719.

Graphs

In addition, we want to analyze the difference between the distribution of the estimated proportion voting for Trump and Biden in each age group. Here are the plots:



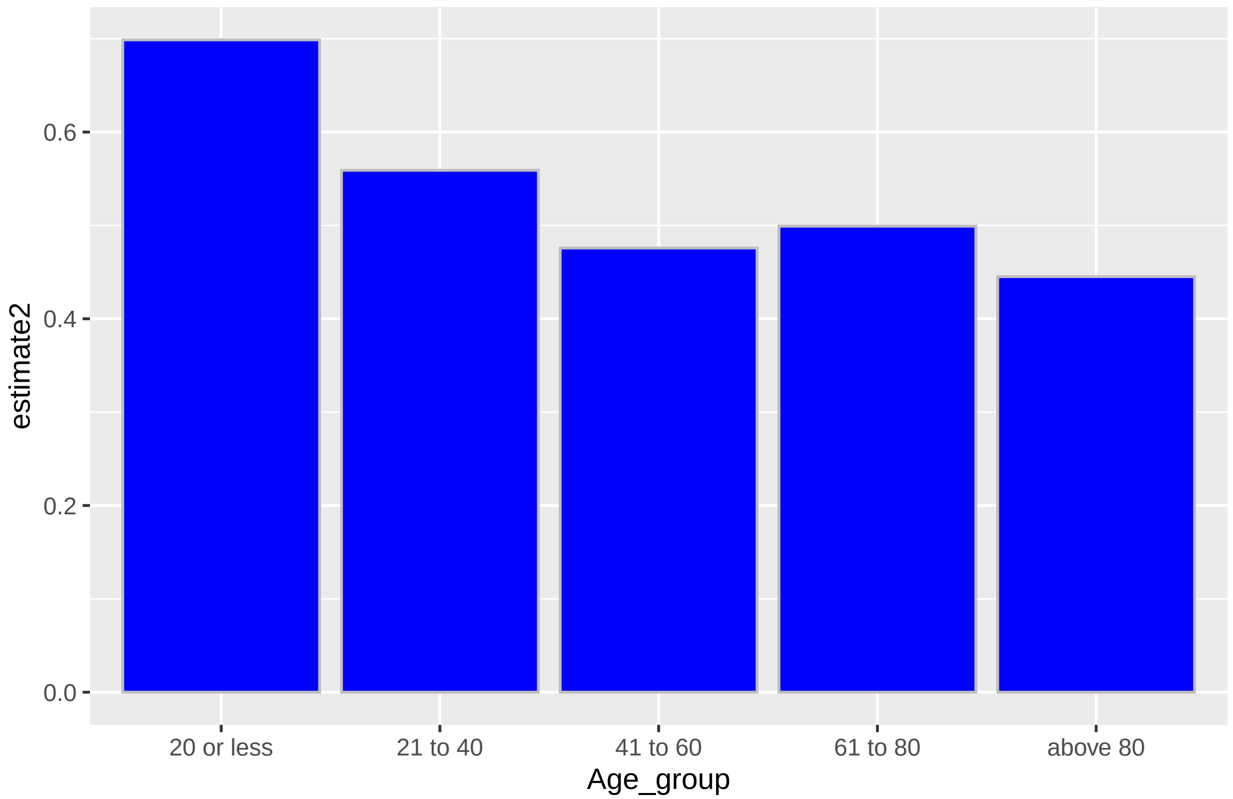


Figure2: Distrubition of Estimate for Joe Biden in Different Age Group

Figures 1 and 2 show the distribution of the estimated voting proportions of Donald Trump and Joe Biden in different age groups. It can be seen that the estimated proportion of Trump voted in Figure 1 is close to gradually rising with the increase of age group, but in Figure 2 the estimated proportion of voted Biden is gradually decreased with age group. We can clearly see that the estimated voting percentage of Joe Biden, who is 20 and under, far exceeds the estimated voting percentage of Donald Trump. In the 21-40 age group, the gap between Joe and Trump is between 0.5 and 0.55, which is the closest, but Joe Biden's votes are still higher. Among the 41 to 60 age group, Trump's estimated proportion far exceeds Joe Biden. Similarly, among the age group 61 to 80 and above, Trump has a lot higher voter turnout than Joe Biden. Except for the large gap between the age group under 20 and the age group 21-40, the gap between other age groups is relatively small. In general, the estimated proportion of Donald Trump voted and the estimated proportion of Joe Biden voted in the opposite trend as the age group increases. Also, we are interested in the distribution of estimate proportion for voting Trump on different races:

```
## Warning in estimate * n$freq: longer object length is not a multiple of shorter
## object length
```

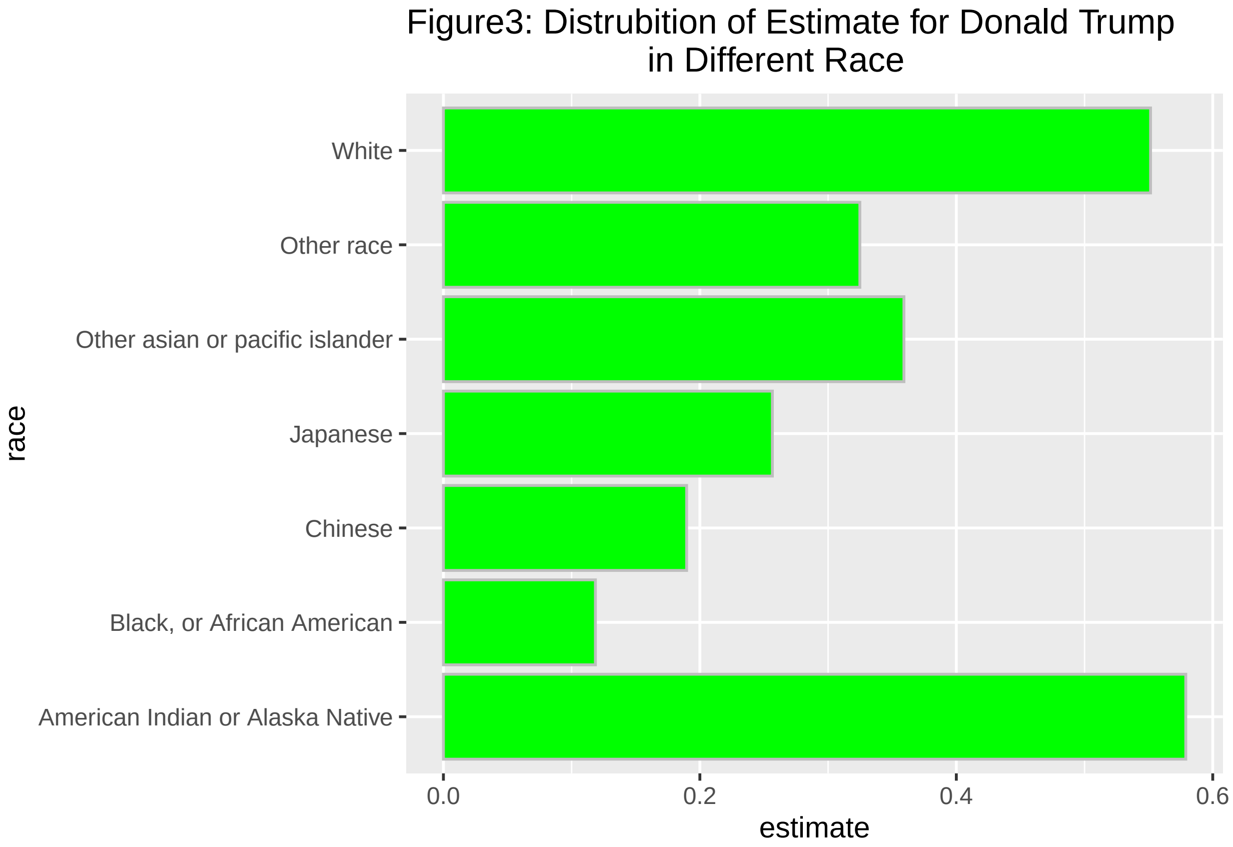


Figure 3 shows the seven ethnic groups and the distribution of Donald Trump's estimated voting proportions in each ethnic group. It can be clearly seen that the proportion of voters who are estimated to vote Trump the most are race white and race American Indians or Alaskan Native, which almost reach 60%. The most are American Indians or Alaskan Native. Conversely, the least estimated proportion voting for Trump is race black, or African American. Among other race groups , such as Japanese, Chinese, and other races, the estimated probability of voting for Trump is obviously less than that of these two groups (whites and American Indians or Alaskan Native).

Discussion

Summary & Conclusion

In this report predicting the results of the popular vote in the 2020 U.S. federal election, we first screened and processed the original data (removing irrelevant observations, selecting the required variables to create a small dataset), and then building a logistic regression model , and got the logistic regression model (Table 1). So we know that the proportion of popular vote for Donald Trump may be related to age, race, and some states. In the post stratification prediction, we selected the variable state to create a cell to predict the proportion of the popular vote for Donald Trump. Then we analyzed the distribution of the estimated proportion of voting for Trump and the estimated proportion of voting for Biden among different age groups(Figure1,2), and analyzed the distribution of the proportion of voting for Trump among different races(Figure 3).

Based on the estimated proportion of voters in favor of voting for Donald Trump being 0.483719, since the estimated proportion for voting Trump is less than 0.5, we predict that Joe Biden will win the election.

Weaknesses & Next Steps

First of all, the variables we choose are subjective. If we choose different variables, the results may be different. Then, after adjusting the data, the remaining survey data is not enough to build a model, so there will be errors in the results, and our budget results do not calculate the proportion of people. Third, we need to pay attention to the fact that the American electoral system does not calculate the total votes of the national electorate but the total votes of the electoral college. We forecast based on the national electoral votes, not the total votes of the electoral college. The result may not be accurate enough. The last weakness is that the data we selected are from June, and these reference data are not time-sensitive.

In the follow-up work, we need to further update the latest time to make the data set more timely. In order to make the results more accurate, we should calculate the total number of votes in the electoral college and calculate the proportion of each person. We can also replace some meaningful and important variables for further analysis to make the results more accurate. We can also change the modeling method. For example, using Bayesian inference may help more analysis.

References

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

(2) lme4: Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

(3) plyr: Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

(4) broom: David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1. https://CRAN.R-project.org/package=broom

(5) dplyr: Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

(6) What is AIC: https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced

(7) What is post stratification: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1993.10476368

(8) USA president election: https://en.wikipedia.org/wiki/United_States_presidential_election

(9) Mean per group in a data frame: https://stackoverflow.com/questions/21982987/mean-per-group-in-a-data-frame

(10) Count number of rows in a data frame: https://stackoverflow.com/questions/25293045/count-number-of-rows-in-a-data-frame-in-r-based-on-group

Appendix

link for Github: https://github.com/ZiqingLyu/304PS3