Popular Vote Prediction for the 2020 American Federal Election

Kexin Qin and Diana Azriel

November 2nd, 2020

Popular Vote Prediction for the 2020 American Federal Election

Kexin Qin and Diana Azriel

November 2nd, 2020

Model

We are interested in predicting the popular vote outcome of the 2020 American federal election. We will be using two seperate models to predict the popular vote for both president Donald Trump and former vice president Joe Biden. To do this we wil be employing a multi-variable logsitic regression model and a post-stratification technique. In the following sub-sections we will describe the model specifics and the post-stratification calculation.

Model Specifics

For this report we are interested in the proportion of voters who will vote for Trump, and the proportion of voters who will vote for Biden. We chose a logistic regression for our model as the response variable that we are interested in is binary: one either votes, or does not vote. We will be using the voter's age and their race for our model. Age is recorded as a numeric variable, and race has been divided in to 6 categories: American Indian/Alaska Native, Asian (Chinese), Asian(Japanese), Black/African American, Other Asian/Pacific Islander, some other race.

$$\log(\frac{p_{trump}}{1-p_{trump}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{white} + \beta_3 x_{chinese} + \beta_4 x_{japanese} + \beta_5 x_{black} + \beta_6 x_{islander} + \beta_7 x_{other}$$

p_{trump} represents the probability that a voter will vote for Donald Trump. β_0 represents the intercept of the model, which is the probability that someone who is American Indian/Alaskan Native with age 0 would vote for Donald Trump. The intercept has no practical interpretation since people younger than 18 years old cannot vote in the US elections. β_1 represents the change in log odds of voting for Trump for one year increase in age. $\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ represents the change in log odds of voting for Trump given the voter's ethnicity.

Similarly, we built another model with the same predictor variables as the model above to model p_{biden} , the probability that a voter will vote for Joe Biden. The prime symbol is used to denote that the values for slopes and intercept for the Biden model will be different from the slopes and intercepts for the Trump model.

$$\log(\frac{p_{biden}}{1-p_{biden}}) = \beta'_0 + \beta'_1 x_{age} + \beta'_2 x_{white} + \beta'_3 x_{chinese} + \beta'_4 x_{japanese} + \beta'_5 x_{black} + \beta'_6 x_{islander} + \beta'_7 x_{other}$$

Below is a quick summary of Trump's model:

Table 1: Fitting generalized (binomial/logit) linear model: vote_trump ~ race_new + age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.045	0.23	-4.543	5.556e-06
race_newAsian (Chinese)	-1.156	0.3673	-3.147	0.001648
race_newAsian (Japanese)	-1.053	0.5997	-1.756	0.07915
race_newBlack, or African American	-1.8	0.254	-7.084	1.398e-12
race_newOther Asian or Pacific Islander	-0.4557	0.2663	-1.711	0.08708
race_newSome other race	-0.6109	0.2465	-2.478	0.0132
race_newWhite	0.2405	0.2227	1.08	0.2801
age	0.01304	0.001651	7.897	2.848e-15

Below is a quick summary of Biden's model:

Table 2: Fitting generalized (binomial/logit) linear model: vote_biden ~ race_new + age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9611	0.2415	-3.981	6.876e-05
race_newAsian (Chinese)	1.212	0.3215	3.769	0.0001637
race_newAsian (Japanese)	1.783	0.5366	3.323	0.0008909
race_newBlack, or African American	1.576	0.2453	6.424	1.328e-10
race_newOther Asian or Pacific Islander	0.8258	0.2674	3.088	0.002015
race_newSome other race	0.7527	0.2512	2.996	0.002734
race_newWhite	0.3421	0.2354	1.453	0.1462
age	0.001936	0.001597	1.212	0.2256

Data cleaning

The census data and the survey data used different categories for race, and we tried to match the two as much as possible in our cleaning process. For example, the survey data has Pacific Islanders divided into 4 categories (Guamania, Native Hawaiian, Samoan, and others) while the census data only had a category called "other Asian or Pacific Islander"; we matched these categories together. The census data had categories like "two major races" and "three or more major races" while the survey data only had a category called "some other race", so these categories were matched together as well.

Post-Stratification

In order to obtain an estimate of the proportion of voters who will vote for Donald Trump and an estimate of the the proportion of voters who will vote for Joe Biden, we need to perform post-stratification analysis. We used post-stratisfication because it allows us to use our survey data to predict how the entire US population will vote. First, we created bins based on different ages and races. We chose age because studies have shown that age has a very strong influence on voting behavior and political prefrences. Race was chosen because it plays a role in political attitudes as well. Therefore, we have good reason to believe that these two variables will help us with our predictions. After creating bins based on differnt ages and ethnicities, we used the model described previously to estimate the proportion of voters in each bin. Lastly, each of the proportion estimates were weighted by the respective population size of that bin, summed, and divided by the entire population size.

Results

With the data from the table and the post-stratisfication technique described above, we estimate that the proportion of voters in favour of voting for Donald Trump to be 0.38. This is based off our post-stratification analysis of the proportion of voters in favour of The Republican Party modelled by a logistic general linear model, which accounted for age and race.

We also estimate that the proportion of voters in favour of voting for Joe Biden to be 0.41. This is based off our post-stratification analysis of the proportion of voters in favour of The Democratic Party modelled by a logistic general linear model, which accounted for age and race.

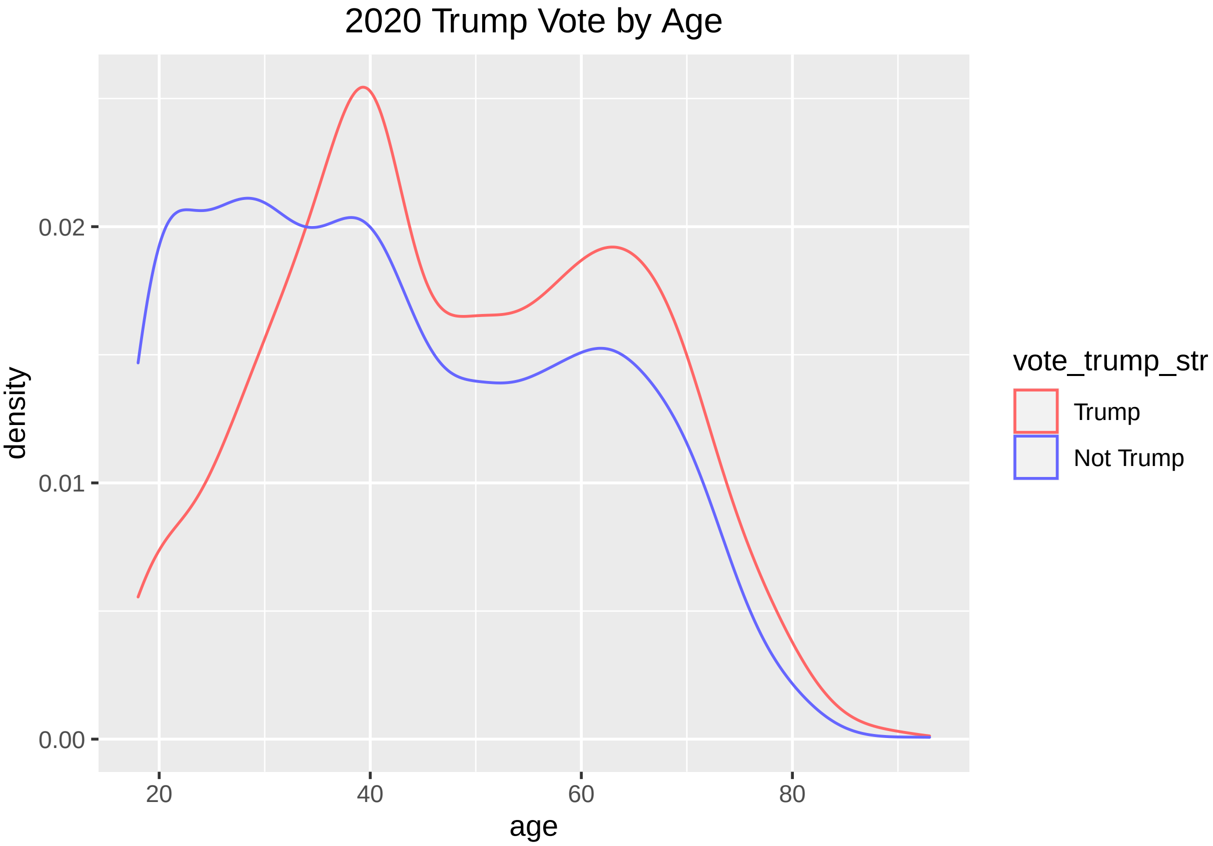

Figure 1: Trump Vote Distribution by Age

As we saw in the model results, age is a significant indictor of one's likelihood of voting for Trump. It had a positive value, which means that with increase in age, the log likelihood of one voting for Trump increases, thus the likelihood increases as well. Figure 1 above depicts the density of one voting for trump as well as the density of all other responses broken down by age. We can see that in ages 18-35, approximately, more respondents did not indicate they would vote for Trump. However, from age 35 onwards, the density curve for Trump votes lies above the other density curve. This is consistent with our model results, indicating that at larger ages, an individual is more likely to vote for Trump.

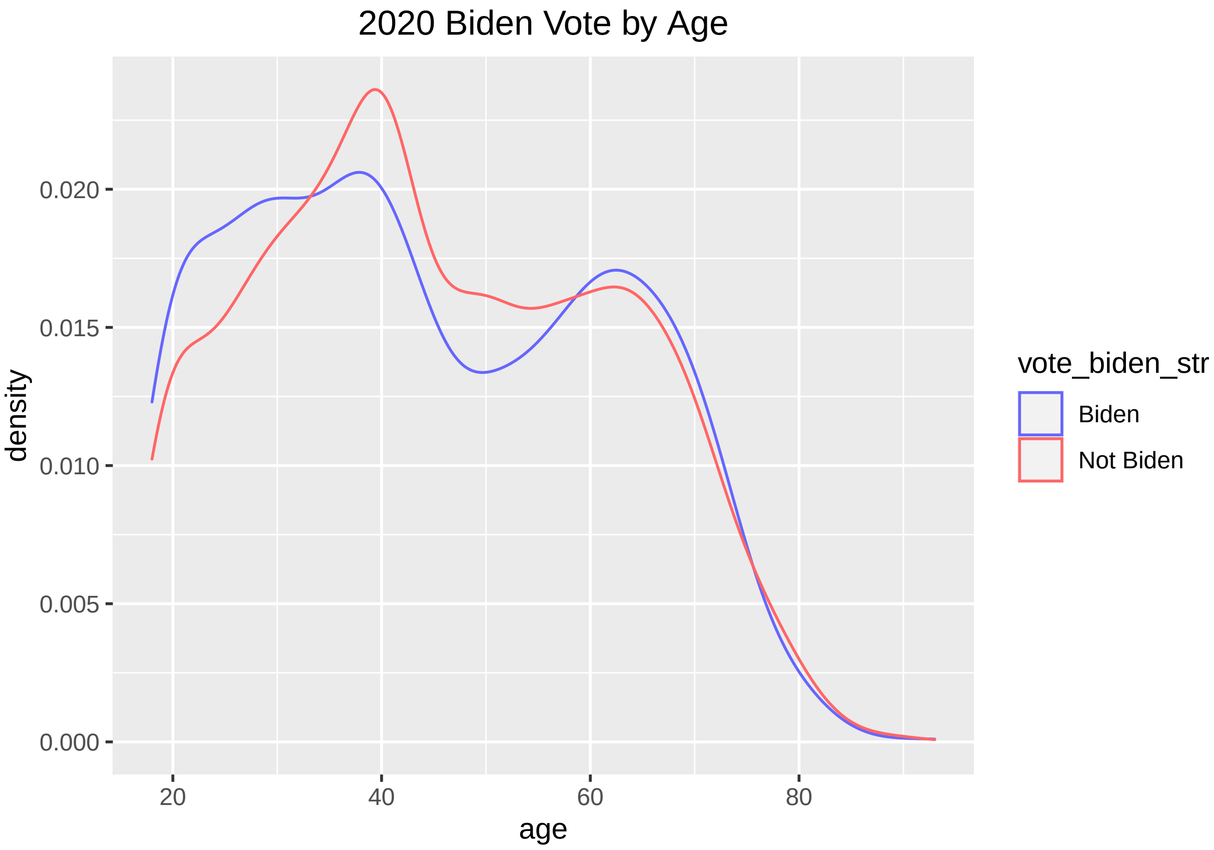

Figure 2: Biden Vote Distribution by Age

Similarly to figure 1, figure 2 shows the density of the 2020 Biden Vote by age. In Biden's model, age had a p-value much larger than 0.05, so it is not surprising that the density curve for Biden votes and all other votes are much closer together, and seem to cross each other at ages 34, 58, and 76 approximately. Younger voters are more likely to vote for Biden, while voters between ages 34 and 58, approximately, are more likely not to do so. In later ages, the curves follow a similar trend and are almost identical.

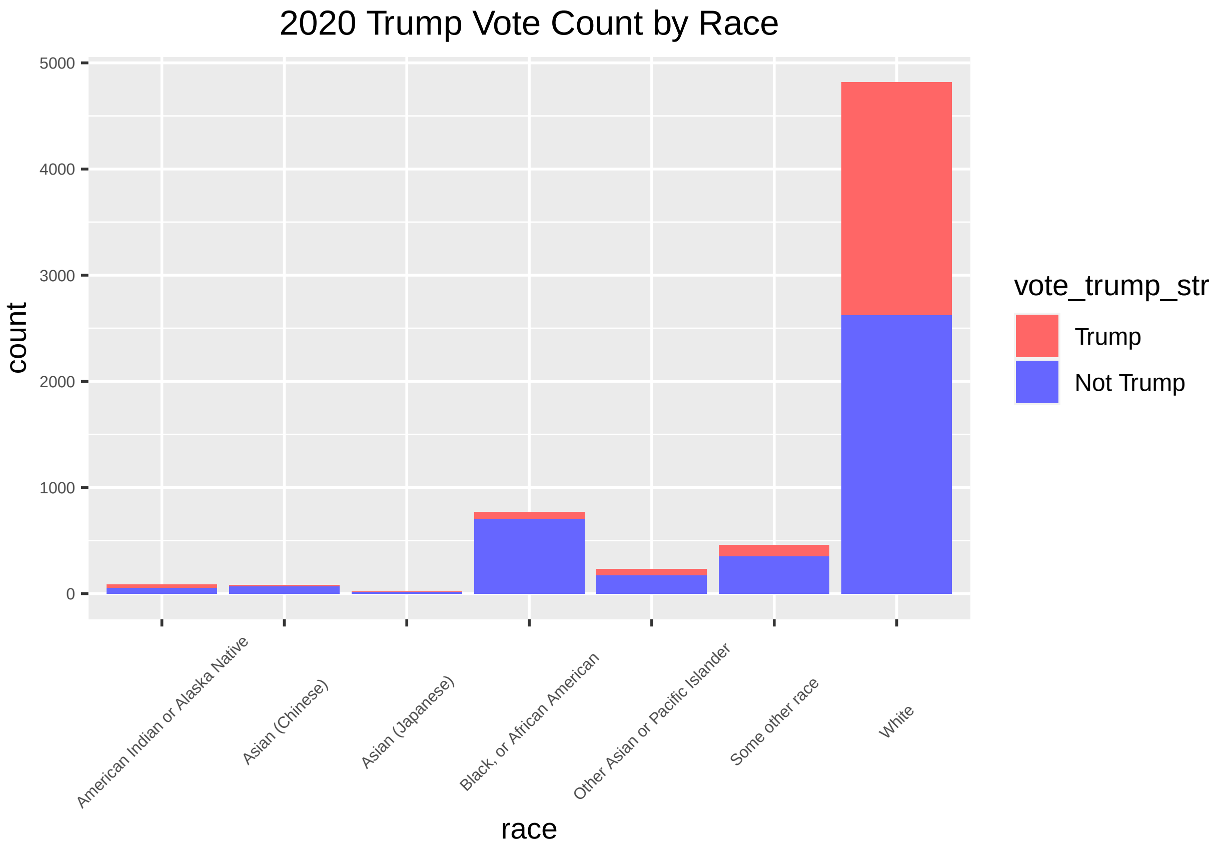

Figure 3: 2020 Trump Vote Count by Race

In the Trump model, all Betas corresponding to race are negative, except that of Whites. We can observe from Figure 3 that individuals in the "Black, or African American" category have the samllest propotion of Trump votes. In our model, their beta value is -1.8, the most negative of all betas. This indicates that if an individual belongs to that category, they are least likely to vote for Trump. Other than the voting breakdown, this graph shows us that that most individuals in the survey are white, with the second largest group being Black or African American.

Discussion

In the interest of predicting the popular vote outcome of the 2020 American federal election, we ran a logistic regression model with the response variables age and race to determine the likelihood of an individual voting for Trump or otherwise, and similarly for Biden or otherwise. Then, through post-stratification using census data, we determined the proportion of voters who will vote for each candidate.

Based off the estimated proportion of voters in favour of voting for Donalad Trump being 0.38, while the estimated proportion of voters in favour of voting for Joe Biden being 0.41, we predict that Joe Biden will win the American federal election in 2020. Through further analysis, we discovered that age was an influential indicator of whether a person will vote for Trump, but this was not the case for Biden voters. Of all race categories, Whites are most likely to vote for Trump, while Japenese Asians are most likely to vote for Biden.

Weaknesses

Some weakeness of the data include inconsistency between race categorization in the census data versus the survey data. Namely, the survey data had 15 categories for race, while the census data only had 7. In the interest of merging the two, some racial cateogries had to be combined into others, therefore losing some individual indicators in order to perform post-stratification.

According to the US Census Bureau, Hispanic/Latino is not included as a category for race because people of Hispanic origin may be of many different races. Therefore, our model was not able to take Hispanics/Latinos into account. This can also lead to us losing some individual indicators for our post-stratisfication.

Additionally, some of the p-values in our models were smaller than 0.05 and differed in the two models. For example, age was a significant indicator in the Trump model, but not in the Biden model. Ideally, a fair comparison of the final voting predictions would have come from models with similar characteristics.

Next Steps

Following the election results, this analysis should be compared to others of its kind in order to determine what factors are most influential in voting decisions because it is possible that age and race are not comprehensive in this determination. Finally, the model should be rewritten with the variables deemed most influential in mind, and rerun leading up to the next election in 2024.

References

Alexander, R. (2020). "01-data_cleaning-survey1.R". Retrieved from: https://q.utoronto.ca/courses/184060

Daróczi, G. (2014). pander: An R Pandoc Writer. R package version 0.5.1. Retrieved from: http://cran.r-project.org/package=pander

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

Hao, Z. (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. Retrieved from: https://cran.r-project.org/web/packages/kableExtra/index.html

Holland, J. (2013). Age Gap? The Influence of Age on Voting Behavior and Political Preferences in the American Electorate. Retrieved from: https://research.libraries.wsu.edu/xmlui/handle/2376/4982 (the study that showed that age influences voting behavior)

Race and Ethnicity Still Play a Role In Political Attitudes. (n.d). Retrieved from: https://iop.harvard.edu/race-and-ethnicity-still-play-role-political-attitudes

Ruggles S. et al. IPUMS USA: Version 10.0 [ACS2018]. 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=86e190d2-3cd7-4d1c-9ddd-e1931f15a2c1.