The Prediction of 2020 U.S. Presidential Election

Kaiyue Wu, Mengxin Zhao, Zikun Lei

Due Date: 11/03/2020

Code and data supporting this analysis is available at:
https://github.com/leizikun/The-Prediction-of-2020-U.S.-Presidential-Election.git

We are interested in predicting the popular vote outcome of the 2020 U.S. presidential election (Dassonneville & Tien, 2020). In this report, we used post-stratification analysis based on a logistic regression model to predict the result of the U.S. Presidential Election in 2020. The result suggests that 36.1% of voters are in favor of voting for Donald Trump.

Model

Model Specifics

We will use survey samples obtained from Democracy Fund and UCLA Nationscape and based on the sample description, survey_data was released in September 2020 and census_data was released in 2018. A logistic regression model is used in our study. We choose this model since the variable we predict, whether or not one would vote for Trump, is a boolean variable. The model was run on R.studio. We chose citizenship as one factor from the census_data while cleaning the data and decided to remove non-citizens in the U.S. since they can not vote for the presidential election. From the survey dataset, four voters' features, age, census_region, employment and race_ethnicity are selected as predictor variables. Here the four features stand for the voter's age, physical region where the voter locates, the employment status and the voter's race, respectively. The response variable is vote_trump, and it is 1 if the voter would vote for Trump else it is 0. Subscripts of X are names of predictor variables or the category if one predictor is categorical. And every β is the slope of the corresponding feature. β_0 is the intercept, which represents the general chance of people voting Trump without looking at their features. β_2 to β_4 are the slopes of the voters' region. β_5 to β_6 are the slopes of their employment status. β_7 to β_10 are the slopes of their race. The specific value for each beta is displayed in figure 1 in the result part. Age was selected since we believe that the proportion of rightists among people in the same age group varies as age increases. We also selected race since we believe that ethnic minorities prefer the Democratic party. Their living region and employment status might also affect their opinions towards the parties. And gender was excluded since we believe that gender was not likely to influence the voter outcome that much.

And here is our model:

$$
\begin{aligned}
log(\frac{(p)}{1-(p)}) =& \beta_0 + \beta_1 X_{age} + \beta_2 X_{census\_region:Northeast} + \\
& \beta_3 X_{census\_region:South} + \beta_4 X_{census\_region:West} + \\
& \beta_5 X_{employment:not\ in\ labor\ force} + \beta_6 X_{employment:unemployed} + \\
& \beta_7 X_{race\_ethnicity:Asian\ or\ pacific\ islander} + \beta_8 X_{race\_ethnicity:Black,\ or\ African\ American} + \\
& \beta_9 X_{race\_ethnicity:Someother\ race\ or\ multiraces} + \beta_{10} X_{race\_ethnicity:White}
\end{aligned}
$$

Post-Stratification

Post-Stratification is a technique that computes the weighted mean of estimates of cells that the population is split into. The weight of each cell is the ratio of the population of the cell versus the net population. This technique is applied since it is a correction of non-probability based sampling. census_data is split into cells by states. We had thought about splitting census_data into smaller cells based on more variables but we finally did not because of the winner-take-all mechanism in the presidential election. Specifically, if Trump wins the popular vote of one state, everyone in the elector college of that state will vote for Trump.

Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1330340	0.2424034	-4.6741667	0.0000030
age	0.0176194	0.0018738	9.4030075	0.0000000
as.factor(census_region)Northeast	-0.0829826	0.0852043	-0.9739254	0.3300936
as.factor(census_region)South	0.2776649	0.0740640	3.7489862	0.0001776
as.factor(census_region)West	-0.0770367	0.0832530	-0.9253324	0.3547930
as.factor(employment)not in labor force	-0.4673363	0.0752552	-6.2100214	0.0000000
as.factor(employment)unemployed	-0.2873375	0.0691304	-4.1564561	0.0000323
as.factor(race_ethnicity)Asian or pacific islander	-0.6005372	0.2588452	-2.3200633	0.0203375
as.factor(race_ethnicity)Black, or African American	-1.8701248	0.2571925	-7.2713034	0.0000000
as.factor(race_ethnicity)Some other race or multiraces	-0.5981676	0.2500901	-2.3918087	0.0167656
as.factor(race_ethnicity)White	0.2120754	0.2258450	0.9390307	0.3477150

Figure 1: Key statistical summaries of the logistic regression model.

Figure 1 displays the outcomes of the model. Some variables have large p-values(comparing to 0.05). Those variables include northeast and west in the census region and white people. These variables have little impact on the results. The other variables are significant as they have small p-value. Age, and people living in the South region, are the factors that have a positive influence on the results. People who live in the South area are most likely to vote for Trump compared to people in other regions. Some factors have a negative influence on the model, which are the categories of people not in the labor force, people who are unemployed, Asian or Pacific islander, African American, and people with some other race or multiraces. It indicates that when the voters own these features, the chance of them voting Donald Trump is relatively lower. For each categorical variable, one category is omitted. The slope of the omitted categories is 0. And the slope of other categories of the same variable indicate how they affect the response compared with the missing category. Based on the post-stratification analysis modelled by the logistic model, the proportion of voters who prefer voting him is estimated to be 36.1%.

Discussion

Summary

By using the 2018 data sample from IPUMS and 2020 Nationscape Data Set, a post-stratification analysis modelled by a logistic model was performed to predict the chance of the voters with certain features voting Donald Trump. The final result shows that 36.1% of the voters are predicted to be in favor of voting for Donald Trump. One bias we have is that there are two kinds of citizens: naturalized citizens and immigrants. But there is no category under citizen for immigrants on the website. So this model does not reflect the sounds of immigrants. Another bias is that the proportion of cell size of cells is not strictly the same as

the proportion of elector college size of states in America. In particular, here the proportion of cell size of cell in population size is approximately the proportion of population of the corresponding state in the total population of the U.S.. But the electoral college size of a state is not proportional to the state's population. So the model cannot reflect the real probability if the final_result is close to 50%.

Conclusion

As 36.1% of the voters are predicted to be in favor of voting Donald Trump based on our analysis, we think that the chance for Donald Trump winning the election is relatively small. As results show that there might be no relationship between the voters' intention and their regions and part of other variables, the effects due to these variables might not be as huge as we thought. Age, and the category of people living in the South region, are the factors that have a positive influence on the results. It implied that people who are older usually approve more of Donald Trump's way of doing things. In contrast, the categories of people not in the labor force, people who are unemployed, people who are Asian or Pacific islander, people who are black or African American, and people with some other race or multiraces have a negative influence on the model. This shows that people who own these features lean towards leftists.

Weaknesses

Since the census data we choose are from 2018, it may be outdated and results may vary. Due to the large outbreak of COVID-19 in 2020, people's voting intentions might change according to Donald Trump's government's response to COVID-19 pandemic and policy changes. Since the census_data and survey_data have differences in their outputs and we need to recategorize some variables. For example, in both census_data and survey_data, we recategorized Asian people and Pacific islander to one group since census_data only contains "Asian or pacific islander" in the race_ethnicity column. This kind of change may affect the proportion outcome because not all parts of Asian and pacific islander can be considered as one race.

Next Steps

There's the possibility of lurking variables that we have yet to observe. Therefore a similar analysis will need to be conducted on various other factors in the American Community Survey 2018 Sample. Since the election is scheduled on Nov. 3rd, 2020, we will search on the internet afterwards to look up the support rate of each candidate and compare it with other analysis. An online survey will be created related to people's voting intentions in 2018 and who they vote for in 2020. We can also do a similar analysis for Biden and compute his probability of winning the election and then compare it with Trump's.

References

1. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. (2020). IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS. Retrieved from https://doi.org/10.18128/D010.V10.0

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

3. Tidyr. (n.d.). Retrieved from https://www.rdocumentation.org/packages/tidyr/versions/0.8.3/topics/drop_na

4. kableExtra. (n.d.) Retrieved from https://www.rdocumentation.org/packages/kableExtra/versions/1.2.1

5. Grouped data. (n.d.). Retrieved from https://dplyr.tidyverse.org/articles/grouping.html

6. Smoothed conditional means - geom_smooth. (n.d.). Retrieved from https://ggplot2.tidyverse.org/reference/geom_smooth.html

7. Dassonneville, R., & Tien, C. (2020, October 15). Introduction to Forecasting the 2020 US Elections. Retrieved from https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/introduction-to-forecasting-the-2020-us-elections/78235400F6BB7E2E370214D1A2307028/core-reader