Predicting the USA 2020 Federal Election

A Generalized linear model used to determine the popular vote of the 2020 election.

Prinsa Gandhi

November 2, 2020

Abstract

This report will be analyzing the Democrazy Fund + UCLA nationscape "Full Data Set" to predict the popular vote of the 2020 American Federal Election with post-stratification perfomed using the American Community Surveys (ACS) data of 2018. This report will use model selecion techniques to determine a prediction model for the binary response variables *"vote_trump"* and *"vote_biden"*, and try to predict the outcome using the ACS data. The datasets have been cleaned and used for exploratory data analysis, to aid in creating the model. The results are very interesting, and it would be very interesting to compare these results to the federal election results.

The code and data supporting this analysis is avaliable at:

https://github.com/username1-p/Elections-Model

Introduction

I used R to perform my analysis on the survey data (@citeR). The data used to create my model is Democracy Fund + UCLA Nationscape 'Full Data Set' (Tausanovitch, Chris and Lynn Vavreck). This dataset had many interesting variables and information that could be used. In order to test my data and obtain my prediction result, I performed post-stratification with a data extract from IPUMS USA, with variables that I used in my model (IPUMS USA, University of Minnesota, www.ipums.org). I created a model to determine the votes for Donald Trump, and a model to obtain the votes for Joe Biden, and the results are very interesting.

Model

In my analysis, the goal is to predict the popular vote outcome for the 2020 American Federal Election (https://www.usa.gov/voting). I will describe the model used, and the post-stratification method used in order to obtain my estimates.

Model Specifics

The model I have chosen to use is a generalized linear model which is a binary logistic regression model. I created 2 different models, in order to predict a result for response variables *vote_trump* or *vote_biden*. I decided to choose this model because my outcome variable in the model is a binary response variable which I had given value of either 0 or 1, and thus, a logistic regression would make most sense. My binary response variable is dependent on some predictor variables which I obtained using model selection techniques. Before model selection, I determined which variables I had to use which were present in my survey data and census

data, because they had to match in order to perform post-stratification. I performed both AIC and BIC model selection, and determined the most significant variables. I compared models using the ANOVA in order to determine the model with the least residuals, and to determine significant predictors one-by-one after the AIC and BIC selection techniques. The link used in this model is the logit link. The assumptions for this model are satisfied. The cases are independent in both the survey and census data, since each person was interviewed once. The dependent variable is distributed by a binomial distribution. The model I chose is shown below:

My model is based on the predictors race, hispanic or not, state, education, foreign born or not, gender, and age group. I decided to use age group because it made sense to group related age groups together, and would simplify the model. Typically, the people in the age groups I made would have similarities between them, and this would be important for voting. The model converges and has significant pvalues. This model had the smallest residual deviance.

$$logit(\pi_i) = \beta_0 + \beta_1 x_{AgeGroup} + \beta_2 x_{Race} + \beta_3 x_{hispanic} + \beta_4 x_{state} + \beta_5 x_{education} + \beta_6 x_{ForeignBorn} + \beta_7 x_{gender}$$

which is modeling

$$logit(\pi_i)$$

the log-odds of the probability of success for the response variable Y(either vote_trump or vote_biden) dependent on the explanatory variables such as gender, and the rest of the variables shown. The β_0 is the model intercept, and the $\beta_1, .., \beta_7$ are the model coefficients for each of the predictors.

Post-Stratification

Post-stratification is a statistical technique which allows a model to be used on a new dataset, and allows estimates to be corrected according to their sampling weights. Then, an estimate is calculated for the entire population. The estimate can be calculated for the prediction model, which can be very useful in modeling behaviour or other various outcomes, and determining an estimate to make important future decisions.In my analysis, this allows me to estimate the proportion of voters who will vote for Donald Trump and in the second calculation, the proportion of voters who will vote for Joe Biden.

I will be using a post-stratification technique which allows me to use my model on the census dataset to predict estimates for certain groups that I have chosen. I will use the weight of each estimate for each group and determine the population level estimate, and this is very useful to allow a model to be used for prediction purposes and weight each estimate. This means bias will decrease because underrepresented groups will be adjusted for with the sample weight, and also leads to smaller variance. This leads to a more accurate prediction.

An important note is that before the cells were split, I removed the datapoints with ages less than 18 because they are not able to vote, and 1 datapoint of age 95+, because it was a new factor level that my original model did not have. There were some additional states that were introduced in the census data, which my original survey data did not have, thus, I could not use those new states, since my model did not have them as a factor level.

In this model, first I split the cells by age group, gender, state, race, education, whether Hispanic or not, and whether foreign born or not. I will partition the census data into demographic cells by grouping them based on my predictors which are age group, gender and state. These groups of these cells allow an estimate for each possible group (each datapoint/row) by using my model to obtain and estimate, and then I weight the estimate by its relative proportion within the whole population and aggregate all estimates to population level estimates. I weight each estimate by its population size of the bin, and find the sum of all these values for each bin. Finally, I divide this summation by the total population count.

Grouping the Cells

I decided to split my cells based on the variables that I used in my model, because they are all extremely significant variables statistically as well as based on real life outcomes. I chose the three most significant variables to group the cells. For example, state is important because it is shown that certain states are more likely to vote for certain parties. The age group is important, because different ages might have different perspectives and would vote based on what issues are important to them. The gender is important because many voters may want to vote based on issues relating to gender such as gender equality, and this would influence voter outcome as well. Thus, my cell groupings are based on age group, gender, and state because I thought these variables were the most significant to group.

Results

Here are my model results shown in a condensed format.

Table 1-Logistic Model for Vote_Trump

```
broom::tidy(logmodeltrump)
```

```
## # A tibble: 77 x 5
##    term        estimate std.error statistic p.value
##    <chr>          <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)   -0.375     0.981    -0.383  0.702
##  2 stateAL       -0.166     0.760    -0.218  0.827
##  3 stateAR        0.132     0.786     0.167  0.867
##  4 stateAZ       -0.380     0.744    -0.511  0.609
##  5 stateCA       -0.797     0.730    -1.09   0.275
##  6 stateCO       -0.474     0.756    -0.628  0.530
##  7 stateCT       -1.47      0.782    -1.89   0.0594
##  8 stateDC       -0.769     0.881    -0.873  0.382
##  9 stateDE       -0.930     0.841    -1.11   0.269
## 10 stateFL       -0.368     0.731    -0.504  0.615
## # ... with 67 more rows
```

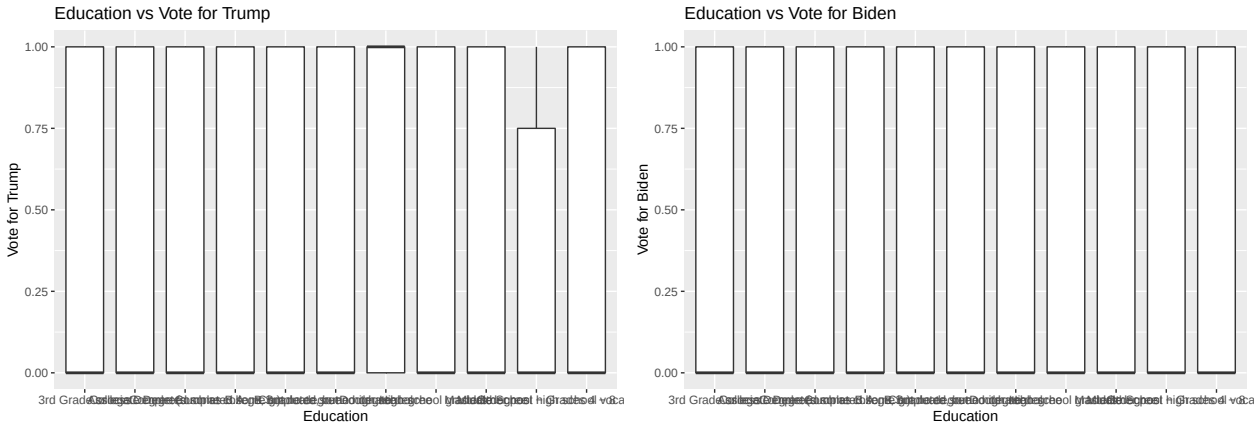
Table 2-Logistic Model for Vote_Biden

```
broom::tidy(logmodelbiden)
```

```
## # A tibble: 77 x 5
##    term        estimate std.error statistic p.value
##    <chr>          <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)   -1.20      1.05     -1.14    0.254
##  2 stateAL        0.266     0.856     0.311   0.756
##  3 stateAR       -0.342     0.894    -0.383   0.702
##  4 stateAZ        0.391     0.841     0.465   0.642
##  5 stateCA        0.734     0.828     0.886   0.376
##  6 stateCO        0.423     0.852     0.497   0.619
##  7 stateCT        1.13      0.861     1.32    0.188
##  8 stateDC        1.45      0.950     1.53    0.127
##  9 stateDE        1.09      0.913     1.19    0.233
## 10 stateFL        0.484     0.830     0.584   0.559
## # ... with 67 more rows
```

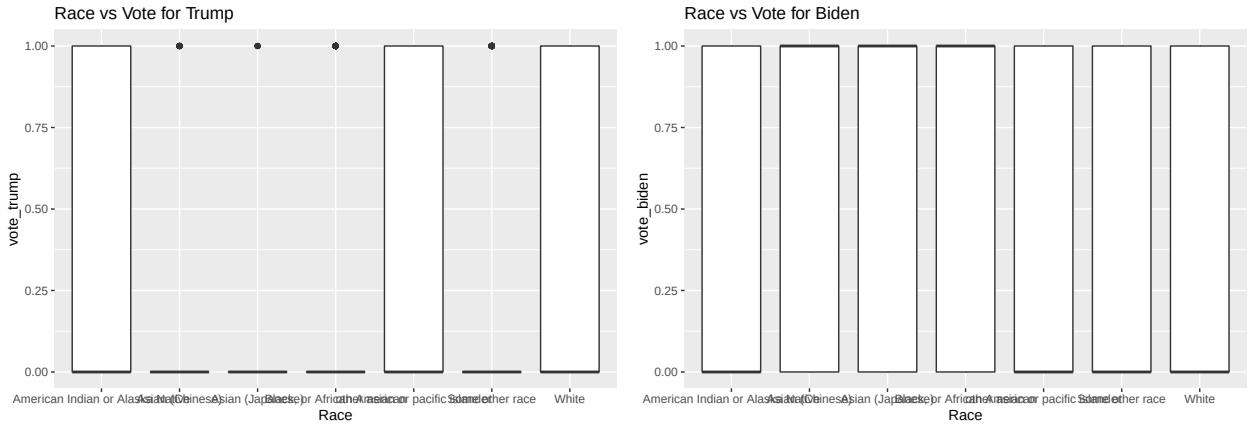
My estimate of the proportion of voters in that will vote for Donald Trump is 0.39, thus 39%. The estimate for the proportion of voters that will vote for Joe Biden is 0.388. These numbers are obtained from the post-stratification analysis of the estimate for response variable *vote_trump* and *vote_biden*. These estimates were obtained from my binary logistic regression model, which took into account the variables state, education, gender, race, age group, Hispanic or not, and foreign born or not.

Figure 1



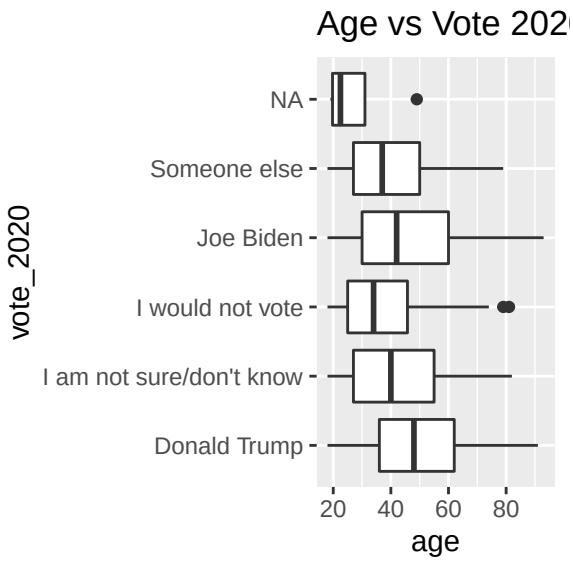
There are some interesting results obtained while performing exploratory data analysis. While reviewing Figure 1, it seems that all education groups seem to vote for Biden, but the middle school category does not prefer to vote for Trump.

Figure 2



While analyzing figure 2, it can be seen that Race is quite an important predictor in determining whom one will vote for. It seems that Race is not impacting votes for Biden, but it is very much impacting the votes for Trump.

Figure 3



While analyzing figure 3, it can be seen that age is impacting one's voting decision. It seems that the younger age groups of 20-40 are unsure, would not vote, or would vote for someone else. The age group of 40-60 are equally spread between voting for Trump or voting for Biden.

Discussion

The results of this analysis are quite interesting. In summary, the binary logistic regression models are created using the survey dataset. The model predictors are determined after numerous steps such as AIC, BIC, and Exploratory data analysis. Later, these models are used on the census dataset to create estimates for each group level, and I use the post-stratification technique to determine a population level estimate.

Finally, based on my results it is very difficult to accurately predict a winner, because my results have given an estimate of 0.39 of voters in favour of Trump, and 0.388 in favour of Biden. However, since Trump has the higher estimated proportion of voters in his favour, I will predict that Trump will win the election. This is because even small decimal amounts make huge differences in large populations, and my analysis on the census data consisted of a sample of 45,564 observations. Thus, after numerous data analysis, the GLM that I used predicts Trump to win the election.

Weaknesses

There are a few weaknesses that affect the results of my study.Firstly, I notice that about 10% of respondents had the response that they were unsure who they were going to vote for, which really affects the outcome variable in my GLM, because there is less data to use for the predictive model. Secondly, there survey dataset has numerous survey responses that would have lead to a far more accurate model. For example, the survey had the questions regarding whom one voted for in 2016, and their vote intention for 2020, as well as numerous policy questions that one may find important.However, I did not use these as predictors for my model, because I wanted to perform post-stratification with variables that were as similar as could be to the census dataset.This lead to a decrease in predictive power for my model. Furthermore, I tried to keep the most significant variables in my model. I decided not to further reduce variables because this was a large dataset, and I did not want to risk over fitting to the data set based on very few predictors, which would lead to a false analysis.

Next Steps

I believe that the next steps can be to obtain predictions using other tools such as random forest prediction, or other machine learning techniques. This would allow the survey data to be used further with additional important variables. In terms of improving this model, I would try to obtain more census data, with information such as whom one voted for in 2016, or a survey asking about opinions regarding important policies. I would try to use this data to determine if there are different results during post-stratification.

Additionally, I would compare the actual election results to the results from my prediction. I would perform an analysis using that information. For example, I would try to see which party won majority in each state and compare that to my model coefficients and determine if there were major discrepancies. I would try to do a survey right after this election asking important questions such as "Who did you vote for" and "which state are you from",etc... These results can prove extremely important, because it can be studied and compared to the pre-election survey and determine if there were major discrepancies and a major change in opinions by those who take the survey. This can prove important in trying to predict future election results, because it can help determine the changes in voter behavior. The data can be used during analysis of future elections.

References

- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=46a716b2-7321-4fcf-9ee6-8987a584a253].
- Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset][usa_00002.dta.gz]. Minneapolis, MN:IPUMS, 2020. https://doi.org/10.18128/D010.V10.0
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016
- Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown.