General Prediction of Donald Trump will lose in 2020 American Federal Election

Hanrui Dou, Hanjing Huang, Hairuo Wang, Xuan Zhong - Group 161

2 November 2020

Code and data supporting this analysis is available at: https://github.com/HairuoWang/STA304-PS3-GP161

Model

We are interested in predicting the popular vote outcome of the 2020 American federal election by the survey data from Democracy Fund + UCLA Nationscape. Therefore, in the below sub-section, we are going to build a model with four variables from survey data, and check the effect of the variables on the probability of voting for Donald Trump or not.

As well, we are going to make the post-stratification analysis base on census data from American Community Surveys 2018 5-year. We will perform the proportion of voting Donald Trump of each cell and also calculate the prediction probability of voting Donald Trump.

Additionally, we use r script to clean the survey data from Democracy Fund + UCLA Nationscape and census data from the American Community Survey. We are going to use the cleaned data to predict the vote outcome of the 2020 American federal election further.

Model Specifics

We have two options to build model, which are the linear regression model and logistic regression model. As we fitted two models with r markdown and compare the standard error, p-value and estimates in both models, we know that most values are similar in the two models, and the p-value of variables in the logistic model is less than the p-value in the linear model. Moreover, since the dependent variable 'Vote trump' is a binary variable, according to slides "Logistic Regression Intro" from Week 5, "logistic regression is suitable when the outcome of interest is binary." Therefore, we think that the logistic model will have more accurate outcomes than the linear model. Furthermore, logistic regression, also known as logarithmic model, it is used to model dichotomous outcome variables. In the logarithmic model, the log probability of the outcome is modeled as a linear combination of predictors.(https://stats.idre.ucla.edu/r/dae/logit-regression/) Additionally, most of the selected variables are categorical variable and the predictor variables of logistic regression model could be numerical and categorical. Thus, we think that the logistic regression model could fit the data and perform the model result better.

We will use a logistic regression model to model the proportion of voters who will vote for Donald Trump. As well, we are going to use 4 variables, which are sex, age, regions and employment status of the voters. The formula of the general model then will be:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age} + \beta_3 x_{northeast} + \beta_4 x_{south} + \beta_5 x_{west} + \beta_6 x_{unemployed} + \beta_7 x_{notinlaborforce}$$

Base on the model, we could see the relationship between prediction variables and probability of voting Trump directly. The model result is more straightforward than the plots of prediction variables and dependent variable, and the model result has specific numbers to perform the relationships, where $log(\frac{p}{1-p})$ represents

change of proportion of voters who will vote for Donald Trump in log odds. For β_0, it represents the intercept of the logistic model, which is the proportion of voting Trump in log odds when all of the variables are 0. (However, that is impossible since the minimum value of prediction variable of age is 17. x_{age} could not be zero.) As well, β_1 represents the slope of the model. So, β_1 is the average difference in log odds of voting Trump between the category for which voter is male or female. β_2 represents the slope of the model. So, for every additional unit increase in age we expect the log odds of proportion of voting Trump to increase by β_2. β_3 represents the slope of the model. So, β_3 is the average difference in log odds of voting Trump between votes from voters in the northeast region and those in the non-northeast regions. β_4 represents the slope of the model. So, β_4 is the average difference in log odds of voting Trump between votes from voters in the south region and those in the non-south regions. β_5 represents the slope of the model. So, β_5 is the average difference in log odds of voting Trump between votes from voters in the west region and those in the non-west regions. β_6 represents the slope of the model. So, β_6 is the average difference in log odds of voting Trump between unemployed voters and not unemployed voters. Finally, β_7 represents the slope of the model. So, β_7 is the average difference in log odds of voting Trump between votes from voters in the labor force and voter that is not in the labor force.

It is important to select useful variables that affect the probability of voting for Donald Trump. For this reason, we build the full models and reduced model to check the AIC for checking whether the variables are necessary to fit the model and not overfitting, which will affect the outcome of the model. AIC of the model with sex, age, region, and employment status is the least, as a result, we selected to build the model with these four variables. For the variable "age", we use ages rather than age groups since we want to check that in pace with the increase of age, whether the probability of voting for Donald Trump will increase or decrease. As well, if we use age groups then there will be too many variables of the model and that is hard for us to perform the result of each age group specifically in a short time. Moreover, we create a new column to arrange a detailed region of the voters into four general regions. Since we found that the correlation between the detailed region and the probability of voting for Donald Trump is weaker than the general region. Therefore, we think that we should focus on data that have a stronger correlation with outcome. As well, if we use the detailed region, there will be too many parameters in the logistic model, and that will take too long time for performing the model. Additionally, we created a new column and arrange employment status into three classifications. The original version of the survey data includes too many classifications and some options are confusing for us to identify whether the voters are employed or not in the labor force. Base on the lecture note from http://www.fbe.hku.hk/~wsuen/teaching/labor/intro.html, we rearrange the classification of employment status which is more convenient to build a model and analysis for next step.

According to the residual plot of the logistic model, we know that the model is a good model with constant variance since the plot does not show any obvious curvature on a trend of residuals. As well, the QQ-plot of the model follows the logistic function well. To check if the model is multicollinearity, we will check the vif for each variable. VIF of each variable is not greater than 4, which represents all of the prediction variables are independent.

Post-Stratification

In order to estimate the probability of voter who votes for Trump, we need to perform an post-stratification analysis. Since we could check the relationship between the prediction variable and the probability of voting of individual voter by the model above, and we want to use the relationship to estimate the probability of voting for Trump in entire population. Thus, we are going to create cells based on different ages, sex, region, and employment status. According to the logistic model that we built, we will estimate the proportion of voters in each age bin with different sex bins, different region bins, and different employment status bins. For example, we are going to show the estimated proportion of voters votes for Trump who are male and 17 years old, living in the midwest region, and are employed. As well, we will show the estimated proportion of voters votes for Trump who are male, 17 years old, living in the midwest region, and are not employed.

Additionally, we will weight each proportion estimate by the respective population size and sum those values up, then divided that by the population size, and the formula of post-stratification will be:

$$\hat{Y}^{ps} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$$

Where $\hat{y_j}$ on the numerator is the estimate of each bin, and Nj on the numerator is the number of voter in each bin. N_j on the denominator represents the entire population size.

We selected all of the variables in the model to estimate the proportion of voting Trump of the entire population. First of all, we choose the variables which are contained in both survey data and census data when we selected data at the beginning. Base on the model result, that shows age and gender of the voters have strong relationship with the probability of voting Trump. Although the p-value of 'region' and 'employment status' shows that variables do not have extremly strong relationship with the outcome compare with age and gender. However, as we mentioned in above section, AIC of model that have four variable could estimate the outcome more accurate. In additional, according to the historical dataset of 2016 election result on CNN and Statista website, there are obviosuly difference of voting result between different region and if the voter have income respectively. Therefore, we use the variables in model to perform post-straitification calculation.

The additional Information that help us to research

When we cleaned the data, we remove the age from less than 1 year old and 90+ because the number of voters of those age groups is too small to influence the outcome of the voting. Therefore, it is not necessary to analyze those age groups and we will ignore them. As well, we remove all of the 'NA' options since it is useless for checking the relationship between the prediction variable and the outcomes of voting. 'NA' options may also lead to biased outcomes, so it is necessary to be removed from the data. In addition, we also rearrange the detailed version of "region" and "employment status" from general data and drop the option without any voters. We also want to make the variables from survey data similar to those from census data which could make it more accurate when we estimate the proportion of voting for Trump by post-stratification analysis. On the other hand, if we used detailed data to build the model, the relationship between some parameters and the outcome is not apparent and the result of the model will be hard to understand for the readers.

In order to conclude the prediction of election, we will fit a similar model for probability of voting biden with all of the same variable. As well, we will perform the post-stratification analysis to check the estimated proportion of voting Biden. However, we will only mainly focus on the model result of probability of voting Trump.

Results

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.406
```

Base on the result of the post-stratification analysis of the proportion of voting for Donald Trump in the model section, we estimate that the proportion of voters in favor of voting for Donald Trump to be 0.4056. Moreover, the model is modeled by the logistic regression model, which is predicted by age, sex, region, and employment status of the voters.

Table 1: Coefficients of Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5894154	0.1016552	-15.635350	0.0000000
as.factor(sex)male	0.5070483	0.0536267	9.455139	0.0000000
age	0.0214433	0.0017343	12.363959	0.0000000
as.factor(region_new)northeast	-0.1380027	0.0836627	-1.649513	0.0990426

	Estimate	Std. Error	z value	Pr(>|z|)
as.factor(region_new)south	0.1701992	0.0720267	2.363001	0.0181276
as.factor(region_new)west	-0.1434901	0.0812611	-1.765790	0.0774311
as.factor(empstat)not in labor force	-0.2987315	0.0628499	-4.753094	0.0000020
as.factor(empstat)unemployed	-0.2961766	0.0908727	-3.259249	0.0011171

We can see clearly from Table 1 that there are 3 of the estimated parameters being positive and 4 being negative. According to the p-value of the variables, we could see that the most of parameters have a small p-value except 'region'. The result shows that the variable 'region of the voters' will not influence the outcome of voting strongly, compared with the other variables. For every unit increase in the variable, positive estimate parameters will lead to an increase in the probability of voting for Donald Trump and negative estimate parameters will lead to a decrease in the probability of voting for Donald Trump. Among the 3 positive estimates, the estimates for sex is the highest, which is about 0.507, and the p-value of 'sex' is extremely small. Thus that means if the voter is a male and then the probability of voting for Trump will increase. The result also shows that the gender of the voter will have a strong influence on their voting results; Among four negative estimates, the estimates for "whether or not a voter in the labor force" is the lowest, which is about -0.299 and the p-value is very small as well, which means that the voters are not in the labor force will not vote Trump in the election and that has a strong negative relationship with the outcome of voting Trump. Moreover, the value of the estimated parameter of β_0 doesn't mean anything since some of the variables are impossible to be all zero, such as "age".

Now since we are very curious about the relationship between "age" and the votes, we are going to make a boxplot of the variable "age" and the Result of Voting for Trump or not. Based on what we talked about above, from the table (model result), we can only see that increases in age will lead to increases in the probability of voting for Trump, and "age" has a strong negative relationship with the outcome of voting for Trump, but not the age distribution of voters who vote for Trump or other details, that is why we need to do make boxplot for further analysis.

Figure 1 − Boxplot of Ages and the Result of Voting Trump or not

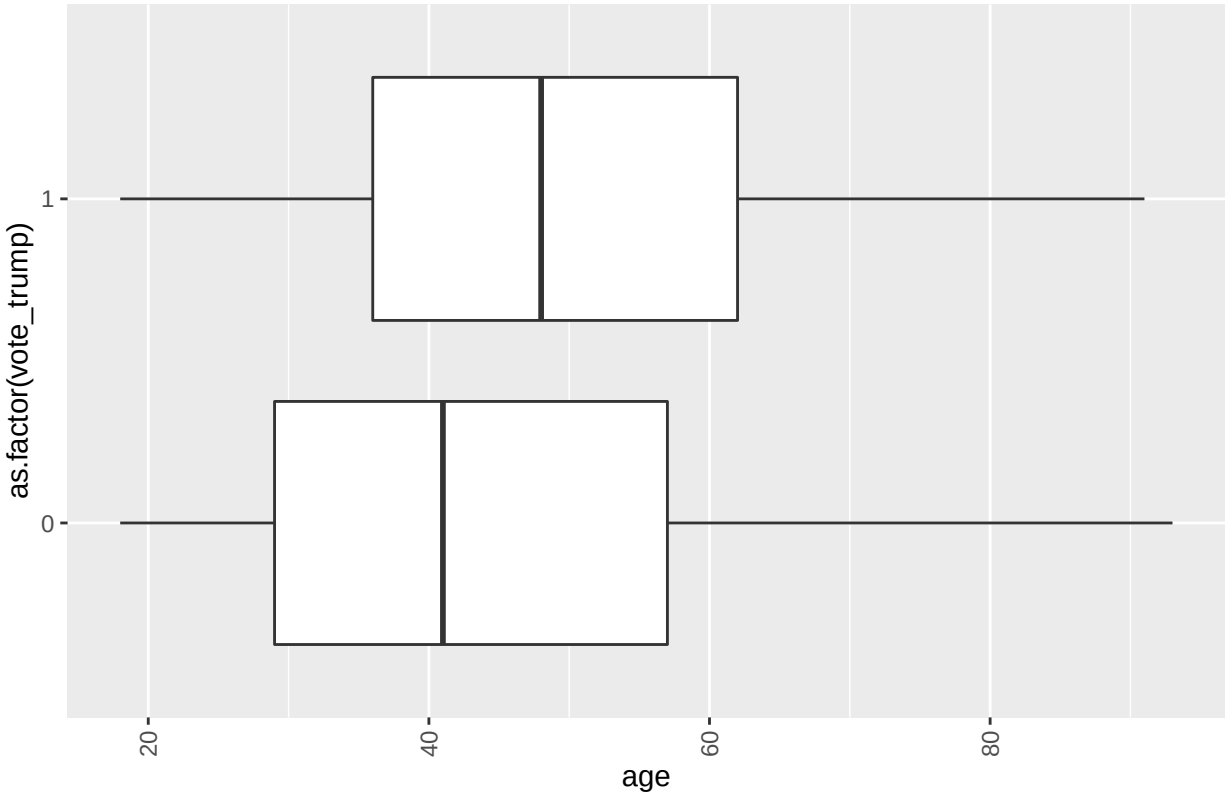

Figure 1 shows the differences of voting Trump between ages.

Based on the boxplot, voters who vote for Donald Trump are older than those who do not vote for Donald Trump on average, since we can see the lower bound, upper bound, and mean value of age are all older for people who vote for Donald Trump.

Discussion

We use the cleaned dataset from Decocracy Fund + UCLA NationsScape to build logistic model for estimating the probability of voting Trump with the different characterizations of individual voters. As well, we use the cleaned and selected dataset from IPUMS to perform the post-stratification analysis. Base on the model result of individual voters, and estimate the proportion of probability of voting Trump of the entire population. When we cleaned the data, we found that the most variables are reasonable and there is no variable with repeated meaning. As well, the number of 'NA' is acceptable and that means it will not lead to serious non-response biases to the dataset. In additional, the options for each variables are detailed and not awkward. Therefore, the estimate results based on the model and post-stratification analysis will be more accurate. When we build the logistic model without multi-level, we selected four variables which are age, sex, region and employment status of the voters.

According to a new Wall Street Journal/NBC News poll of registered voters, it shows that women voters prefer Democratic nominee Joe Biden more than President Trump by a large margin. Trump won a majority of white women in 2016, while that support has eroded over the past four years. He used the term "suburban women" on one of his canvassing,which insulted women when asking for support. His actions may cause women to resent him and cause his election to be unsatisfactory.

Overall, based on the result of estimated proportion of voters who would like to voting to Trump is 0.4056. As well, the result represents there are 59.44% voters who would vote to other party or not vote. Recall the result of post-stratification analysis of other model for Biden, the estimated proportion of voters who would

like to voting Biden is 0.4192. Therefore, we predict that Joe Biden will win the election.

Weaknesses

We have figured out the relationship between characterization of individual voters and the outcome of voting, as well we get the estimate proportion of voting and make the prediction base on the results. Nevertheless, there are some limit and drawback of the analysis. We only fit four variables of the model because of the limit of time and technique. Thus we might miss some other important variables, and that could impact the outcome of voting strongly. Missed variables could lead to the estimate proportion of the outcome of election is not premise. On the other hand, we drop the detailed factors of some variables and replace it with general factors since they are more convenient to build model. The detailed factors may result smaller p-values which represents stronger connection with the dependent variable, and it can represent the thoughts and information of the voters more premises. Besides, since we have not studied the model check of logistic model systematically yet, the analysis for model check and diagnostic are not sufficient.

Next Steps

There are some limits and drawbacks about the analysis and estimation so far, we will spend more time on cleaning data to get the variables which are in both survey data and census data. For example, we could use income groups of voters and education of the voters. Such that we could add more independent variable for fitting model and get more premise. Meanwhile, we will perform model check while we fit more variables and avoid the overfitting and multicolinearity issues of the model. In addition, we will try to fit multi-level model to stratified by different states level. Since we could see from the presidential result on CNN, the different state would have the totally different result of voting.

Additionally, we could analysis the relationship between prediction variables and the actual election results. We could find the key variables which could influence the election results strongly, and use the variables to improve the estimation of the next election better.

References

- 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].
- Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Lecture Note 1. (n.d.). Retrieved October 29, 2020, Retrieved from http://www.fbe.hku.hk/~wsuen/teaching/labor/intro.html
- John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- 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.
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

- N.A. (2004). BIOST515: Lecture 14 Diagnostics and model checking for logistic model [PowerPoint Slides]. Retrieved from https://courses.washington.edu/b515/l14.pdf
- R Pruim (2016). Mathematics in R Markdown. Retrieved from https://rpruim.github.io/s341/S19/from-class/MathinRmd.html
- CNN: Presidential Results. (2016, Feburary 16). Retrieved from https://edition.cnn.com/election/2016/results/president
- Statista: Election 2016 exit polls: percentage of votes by income Published by Statista Research Department, Nov 9, 2016 percentage of votes by income. (2016, November 9). Retrieved from https://www.statista.com/statistics/631244/voter-turnout-of-the-exit-polls-of-the-2016-elections-by-income/
- Collins, E. (2020, October 15). Biden Has 11-Point Lead Over Trump Less Than Three Weeks to Election Day. Retrieved from https://www.wsj.com/articles/biden-has-11-point-lead-over-trump-less-than-three-weeks-to-election-day-11602734461
- Weaver, C. (2020, October 16). 'Please like me': Donald Trump loses ground with suburban women. Retrieved from https://www.ft.com/content/ec950743-c277-4b49-b39c-cc1978208da7
- Andrew. (2020, January 10). Linear or logistic regression with binary outcomes. Retireved from https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/.