# Predicting the 2020 US Electrion Result using Logistic models

Xingyu Pu, Geng Li, Yuchwn Wu, Zhihuan Shao

Oct.31, 2020

# Model

In this report, we are interested in exploring a few factors that might contribute to the voting of new US president candidates and based on the Nationscape Data Set, which is collected by Democracy Fund Voter Study Group and UCLA Political Scientists Chris Tausanovitch and Lynn Vavreck through interviewing people in nearly every county, congressional district, and mid-sized U.S. city in the leadup to the 2020 election, and using those comparatively small survey data to fit a model onto census data to predict the potential outcome of 2020 US Election via post-stratification technique.

## Model Specifics

In this study, we will use logistic model with a few factors that we are interested in, including gender, race, and educational level. We are choosing those factors in several reasons:

1. **Gender** is considered since there could exist a gender preference due to the major candidates of this years election are both males (Trump and Biden), and with D.Trump's potential inappropriate behaviors to females in recent years, we suspect that gender could be one of a deciding factors.

2. **Race** is one the most important deciding factors in our opinions, based on the fact that most policies that D.Trump has implemented and suggested have a clear racial preference, especially the attitude towards immigrants and international students. Thus, we consider race to be a huge factors that could contribute to the result of election. Due to the limitation of the model, we categorize those races into 4 simple categories - White, Black or African American, Asian, and Others for analysis

3. **Educational level**, however this might not be explicitly shown as an important factor, the truth that a great number of US citizens believe the words from D.Trump is because of lacking certain higher level of knowledge or even common sense (e.g. drinking bleach to kill coronavirus), and education is one of the essential method to avoid that. Thus, we believe there should be a certain level of connections between educational levels and willingness to vote for Trump or Biden. Due to the many levels of educations, we rank the education background from the lowest (0) to the highest (10) for analysis

With the background on choosing models and components, the logistic model we are using is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{Male} + \hat{\beta}_2 X_{Black/AfricanAmerican} + \hat{\beta}_3 X_{OtherRaces} + \hat{\beta}_4 X_{White} + \hat{\beta}_5 X_{EducationLevel}$$

Where $\hat{p}$ represents the voters estimated probability of voting for the candidate. Similarly, $\hat{\beta}_0$ represents the intercept of the model, and is the logistic estimator of probability of voting for the candidate then the voter is an uneducated Asian female. Additionally, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ represent the factors that different race, gender, and educational levels can contributed to different logistic estimator, which represent the probability of voting the certain candidate respectively.

# Post-Stratification

We also need to use post-stratification technique to predict the proportion of voters that might vote for either Trump or Biden in order to complete our hypothesis. Choosing the same parameters, we divided the original census data into cells based off gender, race, and education levels. Using the model fitted above we can predict the estimated proportion of voters in each cell. Then, by weighting according to the cell size and sum those values and divided by the entire population size, we can get a glimpse of the potential proportion of voters that might vote for either Trump or Biden. The reasons that we choose those factors to split the cells have been address in Model section.

# Results

Using the survey data, we can get the predicted model for Trump and Biden individually:

Voting for Trump:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.486 + 0.440X_{Male} - 1.123X_{Black/AfricanAmerican} + 0.150X_{OtherRaces} + 0.959X_{White} + 0.020X_{EducationLevel}$$

Voting for Biden:

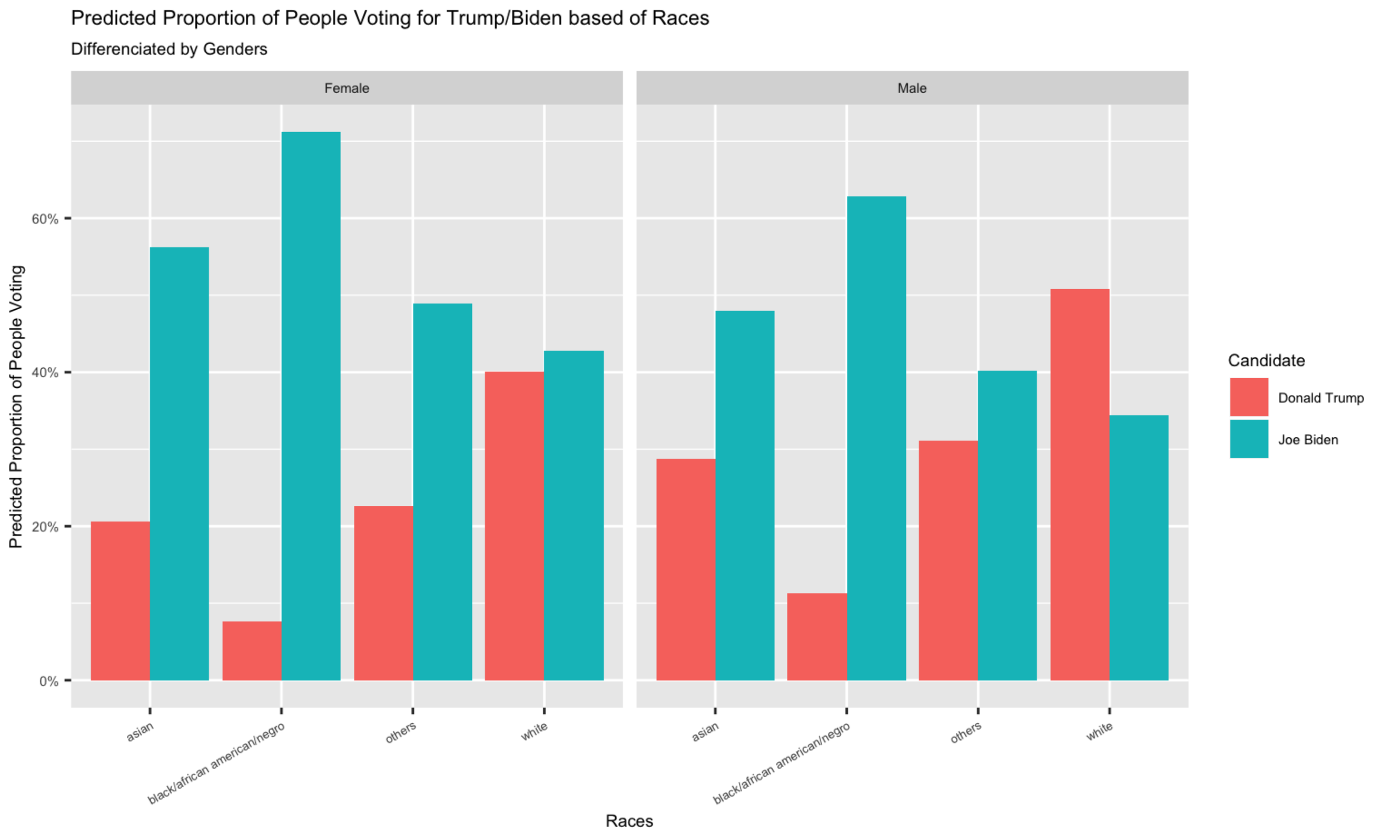
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.394 - 0.3450X_{Male} + 0.754X_{Black/AfricanAmerican} - 0.157X_{OtherRaces} - 0.500X_{White} + 0.097X_{EducationLevel}$$

And by using those models, we calculated the proportion of voters that might be willing to voter for these two candidates individually, shown as below:

| | **Republican Party** | **Democratic Party** |
|---|---|---|
| Candidate Names | Donald Trump | Joe Biden |
| Est Chance of Winning (Percent) | 39.35 | 42.37 |

We predicted that there expects to be **39.35%** of people that are willing to vote for Donald Trump, and **42.37%** of people that are willing to vote for Joe Biden. This results are based of This is based off our post-stratification analysis of the proportion of voters by a logistic model, which accounted for genders, races, and levels of education.

As for each individual parameters, a barplot is shown below:


Predicted Proportion of People Voting for Trump/Biden based of Races

Differenciated by Genders

Figure 1

As shown in Figure 1, in race wise, most races seem to be more willing to vote for Joe Biden, as well as both genders, however, the advantages seem less obvious when it comes to white voters, especially male white voters, the only exception that are more willing to vote for Donald Trump. This is as expected. Other than that, on a large scale, people are more willing to vote for Joe Biden rather than Donald Trump, however, the results from our post-stratification shows that the differences between winning rate is not as obvious as shown in the graph, most possibly due to the large population of white male voters that are willing to support Trump.

# Discussion

In this study, we performed a model fitting using survey data on a small scale of people that have taken the survey, and using post-stratification technique to predict the willingness to vote for Trump or Biden on a large scale using census data. Using logistic model allows us to predict the outcome of voting based on genders, races, and educational levels, and we discover that some of the parameters are not as important as we expected before running the analysis (e.g. educational levels).

Based off the estimated proportion of voters in favor of voting for Democratic Party being **42.37%** and for Republican Party being **39.35%**, our model successfully predicted that Joe Biden should have a slight higher chance of winning comparing to Donald Trump, however, the difference is much smaller than expected. Considering the population

# Weaknesses

However, there are still a few weakness in this analysis:

1. The categorical values we use, for example, races and educational level, are still too arbitrary. Due to the fact that the survey data and census data were using different scales, there can be some errors existing. The educational level in the survey ranges from primary school to doctoral, but in the census dataset, Grade8 to Grade12 were detailed listed as individual categories. This results in that when ranking educational levels, people in the survey that had college experience are not comparable to those in the census data (that those people had only finished high school). And this could induce some error when using the model to fit census data for prediction.
2. The models we use are still not fitted perfected, as the residuals are still quite large (detailed model summary has been shown as additional information in appendix). We haven't implemented methods to increase the accuracy of the models, thus this could also give us less accurate predictions of the results.

## Next Steps

The next steps of our analysis will include finding more important parameters that could affect the predictions of the results, and validating the model with real voting results. Some other parameters, including voting in 2016, martial status, income level, and working condition, could all contribute to impact the potential willingness to vote for Biden or Trump, as the policies of two parties are quite different. Validating our model when the real election results come out could give us a better understanding of how those parameters actually contribute to the proportion of voting, and what other significant variables should we consider if we are going to perform another similar analysis.

# References

1. 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 (https://www.voterstudygroup.org/publication/nationscape-data-set)
2. 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 (https://doi.org/10.18128/D010.V10.0)

# Appendix

# Code and data supporting this analysis is available at: https://github.com/xingyupu/PS3 (https://github.com/xingyupu/PS3)

*Logistic model for Joe Biden's voters proportion estimating:*

```
##
## Call:
## glm(formula = vote_biden ~ gender + race + edu_level, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7676  -1.0085  -0.8775   1.2712   1.7285
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.39413    0.13509  -2.918  0.00353 **
## genderMale                      -0.34482    0.05264  -6.550 5.75e-11 ***
## raceblack/african american/negro  0.75378    0.13805   5.460 4.75e-08 ***
## raceothers                      -0.15686    0.14277  -1.099  0.27190
## racewhite                       -0.50057    0.11764  -4.255 2.09e-05 ***
## edu_level                        0.09672    0.01195   8.095 5.73e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8809.5  on 6474  degrees of freedom
## Residual deviance: 8462.4  on 6469  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 8474.4
##
## Number of Fisher Scoring iterations: 4
```

*Logistic model for Donald Trump's voters proportion estimating:*

```
## 
## Call:
## glm(formula = vote_trump ~ gender + race + edu_level, family = "binomial", 
##     data = survey_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2262  -1.0184  -0.6744   1.1718   2.2991  
## 
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                      -1.48594    0.15286  -9.721  < 2e-16 ***
## genderMale                        0.43978    0.05406   8.135 4.13e-16 ***
## raceblack/african american/negro -1.12337    0.18385  -6.110 9.95e-10 ***
## raceothers                        0.15005    0.16515   0.909   0.3636    
## racewhite                         0.95890    0.13614   7.043 1.88e-12 ***
## edu_level                         0.02012    0.01218   1.652   0.0986 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 8619.4  on 6474  degrees of freedom
## Residual deviance: 8021.0  on 6469  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 8033
## 
## Number of Fisher Scoring iterations: 4
```