# Prediction of US 2020 election using logistic regression with post-stratification

Erdong Zhang, Xiaoran Zhang, Weichao Wu, Haoji Ye

02/11/2020

# Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

## Model Specifics

The aim is to estimate the outcome of the election. which will be a binomial variable (e.g.1 stands for voting for Donald Trump while 0 stands for voting for Biden). In this case, applying an logistic regression model to this project will greatly fit in the reality. Compared with linear regression model，logistic regression model is capable to fit in the data more smoothly rather than creating an less-related linear model. Here is our logistic regression model:

$$y = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{sex} + \beta_3 x_{racegroup} + \beta_4 x_{educationlevel} + \beta_5 x_{householdincome} + \beta_6 x_{states} + \epsilon$$

Where $y$ represents the proportion of voters who will vote for Donald Trump. Similarly, $\beta_0$ represents the intercept of the model, and is the probability of voting for Donald Trump under these conditions: $x_{agegroup}, x_{sex}, x_{racegroup}, x_{educationlevel}, x_{householdincome}, x_{states}$ are all 0. Additionally, for $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and $\beta_6$, their respective variables are $x_{sex}, x_{racegroup}, x_{educationlevel}, x_{householdincome}, x_{states}$. Therefore, by this model, we could estimate that for everyone who are in these groups, we expect a $\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6$ change in the probability of voting for Donald Trump.

## Post-Stratification

Following our logistic regression model, a post-stratification analysis is needed to estimate the percentage of voters who will vote for Donald Trump among all voters in the U.S.. In our logistic regression model, observations are categorized based on their age, household income, sexuality, education level, race and the states where they will vote for the election. In the data, the variable called n shows the number of observations who fit in the requirements of each group in the variables mentioned above. According to the post-stratification equation, n equals to $N_j$, which means the population size of the $j^{th}$ row. Next, the $\hat{y}^{PS}$ will be calculated by the fraction, with the numerator $\sum N_j^{\hat{y}_j}$ and denominator $\sum N_j$. Calculation result shows that the proportion of voters who support Donald Trump is about 0.4175 with the result being rounded to the fourth decimal.

```
# Here I will perform the post-stratification calculation
census_data$logodds_estimate <-
  model %>%
  predict(newdata = census_data)

census_data$estimate <-
  exp(census_data$logodds_estimate)/(1+exp(census_data$logodds_estimate))

census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict=
              sum(alp_predict_prop)/sum(n))
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.417
```

# Results

After we using the post-stratification strategy, we get an estimate value of 41.74777% of voters might be in favour of voting Donald Trump in the upcoming 2020 election. This result is based on the logistic regression models by using the variables of sample voters' age group, race, family income, education level, sex and the states they lived in. By summarizing the logistic regression model results, we find that white people will prefer to vote Donald Trump than other race groups.Also, sample voters with high family income are predicted to be more likely to vote Donald Trump in the election as well.

# Discussion

# Summary:

Our main objective is to predict the result of the upcoming 2020 US election by estimating the preferences of the sample voters in different race, age group, financial status, sexuality, education level and states from the recent survey in 2020. By contracting a logistic regression model based on the recent 2020 US survey data, and a post stratification on census data of US, we estimate there might be approximately 41.7% of the voters would vote Donald Trump in election.

# Conclusion:

In conclusion, based on our estimation, we think that Biden will win the election. In our prediction, we estimate that 41.7% of the voters will vote for Trump. However, that does not mean the rest will vote for Biden. Among these 58.3% voters, there are people who did not know who to vote for. A reasonable way to think about this unstable factor is that these people will not vote, or some of them vote for Trump and some vote for Biden. Thus, the difference between the percentage of voters of Trump and Biden will remain basically the same, and Biden will win the election in our prediction.

## Weaknesses

I think the first weakness of this study will be very obviously, that is most of the people in this data set is white. Which makes it harder to predict the actual percentage of voters that vote for Donald Trump. Moreover, in the survey, there are people in the age between 80 and 90, and these people might not be able to vote in reality. Hence, this may cause bias to our prediction. Furthermore, we did not calculate the percentage of voting for Biden, and we did not figure out how many people are don't know who they will vote for. That would also make bias to our prediction.

## Next Steps

We predict Donald Trump would receive 41.7% of vote in the upcoming election, which we can compare the actual results of the election to our prediction results in order to improve our future prediction algorithm. Since our sample result is only based on certain variables of the voters, we might consider improving our estimation by adding more dependent variables like voters' martial status and their partial party. This would make our estimation on future elections more accurate.

# References

1. Rohan Alexander, Samantha-Jo Caetano. 01-data_cleaning-survey1.R
2. Rohan Alexander, Samantha-Jo Caetano. 01-data_cleaning-post-strat1.R
3. American Community Surveys (ACS). census data: US00001.dta. Retrieved from: https://usa.ipums.org/usa/.
4. Democracy Fund Voter Study Group. Survey data: ns20200625. Retrieved from: https://www.voterstudygroup.org/publication/nationscape-data-set.

# Appendix

1. data can be found in github link: https://github.com/Weichao-Wu/STA304-project-2-and-3/blob/main/01-data_cleaning-post-strat1.R https://github.com/Weichao-Wu/STA304-project-2-and-3/blob/main/01-data_cleaning-survey1.R
2. rmarkdown edition can be found in github link: https://github.com/Weichao-Wu/STA304-project-2-and-3/blob/main/Prediction%20of%20US2020%20elction%20using%20logisitic%20regression%20with%20post-stratification(1).Rmd