2020 American federal election slightly in favor of Joe Biden predicted by logistic regression model with post-stratification technique

Michael Huang, Jiawei Wang, Huaqing Zhang, Qiushu Zhou

November 2, 2020

Code and data supporting this analysis is available: https://github.com/Sta304-group145/304PS3.git

Model

In this report, we are interested in predicting the popular vote outcome of the 2020 American federal election (Tausanovitch & Lynn, 2020). To do this we are employing a logistic regression model. Historically, the voting outcome differs significantly by region and race. To take into account this difference in sample and target population, we are employing a post-stratification technique to decrease the variance and bias of the predictors selected in the model.

In this analysis, we partition the population into cells based on multiple demographic and geographic characteristics and then we use the sample (survey data) to estimate the response variable (probability of voting for Donald Trump) within each cell. Finally we use census data to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population. In the following subsections we will describe the model specifics and the post-stratification calculation.

Model Specifics

We will be using a logistic regression model to model the proportion of voters who will vote for Donald Trump with the software R studio. In order to model the probability of voting for Donald Trump and obtain relatively precise results, we made some re-arrangement on these data. We are interested in whether the respondent voted Donald Trump or Joe Biden (which is a binary response) so other options like "I am not sure/don't know" , "I would not vote" and "Someone else" have been filtered out. We make use of the glm() function in the R package glm2 (Marschner,2011) to build the logistic regression model which is:

$$Pr(y_i) = log(\frac{y_i}{1-y_i}) = \beta_o + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{race} + \beta_4 x_{region} + \epsilon$$

Here y_i represents the proportion of voters who will vote for Donald Trump. And we used four predictors: age, gender, race of voter and their region. β_0 represents the intercept of the model, and is the probability of voting for Donald Trump at age 0. Additionally, β_1 represents one slope of the model. So, for everyone one unit increase in age, we expect an increase in the probability of voting for Donald Trump. Similarly, β_2 represents the slope of sex variable. In this model we expect male to have a higher probability of voting for Donald Trump. Since β_3 refers to the slope of race, and it has four categories (Asian, White, Black and other races), Asian is a base line, if existing races like Black, White and other races then it changes in probability. β_4 represents the slope of regions, and there are four categories (Midwest, Northeast, South and West), Midwest is a base line. While other regions exist, the probability of voting for Donald Trump would change.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump, we make use of post-stratification technique. Historically, the voting outcome differs significantly by region and race which shows that survey data is not representative of the whole population. Thus, we would apply a post-stratification technique to consider the difference in sample and target population. Specifically, we generate the cells by considering all combinations of age (93 categories), sex (2 categories), race (4 categories) and region (4 categories), thus partitioning the data into 2976 cells. Using the logistic regression model described in the previous subsection we will estimate the proportion of voters in each cell. We will then weigh each proportion estimate (within each cell) by the respective population size of that cell and sum those values and divide that by the entire population size. The post-stratification estimate can be noted as:

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

where $\hat{y_j}$ represents the estimate in each cell j, and N_j represents the size of the j^{th} cell in the population.

We are including all the variables to create cells because they are all likely to influence voter outcome. For instance, candidates might claim that they will roll out specific regulations or policies which are beneficial for some regions. Thus, people from different states will make different decisions on whether they should vote for this candidate or not. Race is another element which can cause large divergence. The voters with the same race tend to support one same candidate. Likewise, people from similar age range might have similar preferences on candidate selection. Also, the gender group of one candidates' voters can be influenced by which gender group the candidate is in. People are likely to support the candidate who has the same gender as them.

Analysis of Variance (ANOVA) for the logistic regression model

In order to determine the significance of four independent variable (age, sex, race and region), we apply an ANOVA table for this logistic regression model.

Results

First we built a logistic regression model, we then used the post-stratification technique to estimate the proportion of voting for Donald Trump. Consequently, we found the result $\hat{y}^{ps}$ is about 0.4718. From Table 1, the predictors: age, sex, races of black and white, and Southern regions all have extremely small p-values based on this model, such as the p-value of the variable age is $3.54 * 10^{-14}$. However, races for groups of people other than Asian and two main races; meanwhile, Northeast and Western regions have bigger p-values, which are smaller than 0.3 and bigger than 0.2.

Table 1, Model Output from R Studio

```
## # A tibble: 9 x 5
##   term                                  estimate std.error statistic  p.value
##   <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                           -1.32      0.174       -7.58 3.54e-14
## 2 age                                    0.00799   0.00181      4.41 1.02e- 5
## 3 sexmale                                0.468     0.0593       7.90 2.71e-15
## 4 race_voterblack/african american/negro -1.36      0.193       -7.07 1.56e-12
## 5 race_voterother                        0.207     0.176        1.18 2.40e- 1
## 6 race_voterwhite                        0.812     0.146        5.55 2.83e- 8
## 7 regionsNortheast                      -0.101     0.0919      -1.10 2.70e- 1
```

```
## 8 regionsSouth                          0.317     0.0802       3.96 7.62e- 5
## 9 regionsWest                          -0.0987    0.0897      -1.10 2.71e- 1
```

In order to determine the significance of four independent variable (age, sex, race and region), we apply an ANOVA table for this logistic regression model. From Table 2, it illustrates that the residual deviance of the null model with just an intercept is 7198, every addition of predictors decreases the residual deviance. For instance, when age is a predictor, the residual deviance decreases from 7198 to 7131. Similarly, as sex, race and region of voter are predictors, the residual deviance decreases into 7040, 6645, and 6605.

Table 2, ANOVA table

```
## # A tibble: 5 x 5
##   term          df Deviance Resid..Df Resid..Dev
##   <chr>      <int>    <dbl>     <int>      <dbl>
## 1 NULL          NA     NA        5199      7198.
## 2 age            1     67.3      5198      7131.
## 3 sex            1     90.4      5197      7040.
## 4 race_voter     3    395.       5194      6645.
## 5 regions        3     40.6      5191      6605.
```

Discussion

In this report, we are predicting the result of the 2020 American federal election. We are interested in whether Americans vote for Donald Trump or Joe Biden. Hence we filtered out all other responses in this question. Four predictors were selected age, sex, race and region. Estimator is the proportion of voters who will vote for Donald Trump. Since the estimator is binary, a logistic regression model was used. By viewing the p-value of each predictor of the model, it is observed that four predictors all have a significant effect on the proportion of voting for Donald Trump. Nevertheless, some categories of the predictors contain bigger p-values. In particular, we apply an ANOVA table. It shows that compared to the null model, the residual deviance of the logistic regression model decreases by the addition of these four predictors. Hence, four predictors can be used to predict the result of the 2020 American federal election. After applying the logistic model, we make use of the post-stratification technique, finally we obtain the probability of voting for Donald Trump is 0.4718.

In conclusion, since the probability of voting for Donald Trump is around 47.18%, we think the Democratic party will win the primary vote and the estimated proportion of voters in favor of voting for Joe Biden is 0.5282 which is bigger than the proportion of the republican party.

Weaknesses

As we used a logistic regression model to analyze the data, we only used 4 variables from the data, which might not be the most significant variables. This would cause some error on predicting the result of the 2020 American federal election slightly. If there exist dependent variables, it can seriously influence our prediction on the election results. Although we cleaned the raw data in the very first step, some responses from survey data are 'do not know' and other unclear responses, and we remove these responses, which might also affect the estimation slightly.

Next Steps

It is sufficient for us to select more predictors in this study, which can help reduce the errors for prediction. Also, before building the regression model, the correlation between predictors should be checked as it can

impact the prediction. Since we choose age, sex, race and region as our predictors, more predictors could be selected to build the model. In addition, AIC and BIC model selection techniques applied to select the best model for our study. Meanwhile, it is noted that the main goal of this study is to predict the US election results, then we can do some follow-up surveys on people's opinions on the election result when the election ends.

References

Barboza, I & Williams R. (2005). Post-stratification and response bias in survey data with applications in political science. Michigan State University. https://msu.edu/~barbozag/Web/poststrat.pdf

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.

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version2.3.1. https://CRAN.R-project.org/package=haven Marschner, I. C. (2011). glm2: Fitting Generalized Linear Models with Convergence Problems. The R Journal 3(2): 12-15

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2020. Broom: Convert Statistical Objects into Tidy Tibbles. https://CRAN.R-project.org/package=broom.

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

Tausanovitch, C & Lynn V.(2020). Nationscape Data Set. Democracy Fund & UCLA Nationscape. https://www.voterstudygroup.org/downloads?key=06949080-017c-4010-aae4-1259d59b5465

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting,31(3), 980-991.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

Zahorski, Alex.(2020). Multilevel Regression with Post Stratification for the National level Viber/Street poll on the 2020 Presidential Election in Belarus. Uladzimir Karatkevich National university of Belarus Miensk. https://arxiv.org/abs/2009.06615v1