Predict the popular vote outcome of the 2020 American federal election

Sirui Xu(1004963603), Yiwen Feng(1004890932), Jiayi Bao(1004109868), Yutong Yuan(1004725103)

November,2,2020

Model

The 2020 US presidential election is the current topical issue of global concern. In this project, we will predict the popular vote outcome of the 2020 American Federal Election by fitting a logistic regression model with the sampling method of post-stratification through using the data analysis software, Rstudio. In our logistic model, we will use three predictors: age(equal and larger than 18), gender, and race respectively to predict the vote outcome of this election. In this study, we are applying post-stratification techniques to partition the voting population into several different cells based on some characters. Then, we are using the sample to calculate the estimator within each cell by applying the logistic regression model. We will do the further description of the model and the process of calculating the estimator in the following parts of model specifics and Post-Stratification.

Model Specifics

For predicting the election vote outcome, we fitted a logistic regression model which used the logistic function to estimate the binary dependent variable. In this study, we selected the variables of age, gender, and race, because those three elements are the obvious demographic attributes that have a high probability to affect their voting decision due to different life experience, gender and culture background. In our model, the reason we used age rather than the age group is that we can make a more precise estimation for every one unit change in our estimator age. The age range of our sample population is from 18 to 99. The gender is male and female and fifteen races are included. By applying the function of glm() with the family of binomial distribution, we obtain the the logistic regression model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{Male} + \beta_3 x_{race} + \epsilon$$

Where p represents the probability of observations who will vote for Donald Trump, and $(1-p)$ means the probability of the observations who will not vote for him. β_0 represents the intercept of the model. β_1 represents the slope of age in the model, which means that as the age increases one unit, we will expect the log-odds of voting for Donald Trump will increase by β_1 units if all the other elements remain unchanged. β_2 represents the slope of the dummy variable of gender, which means that if the value of β_2 is larger than 0, we can say that male are more likely to vote for Trump. β_3 represents the slope of race, which means that as Xrace increases by one unit, the log-odds would increase by β_3 units. In addition, if the value of β_3 is positive, we can say that this race is more willing to vote for Trump.

```
# Creating the Logit Model for Trump
model <- glm(vote_trump ~ age + gender+race_ethnicity,
             data=survey_data, family="binomial")

# Model Results for Trump
summary(model)
```

```
##
## Call:
## glm(formula = vote_trump ~ age + gender + race_ethnicity, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4050  -1.0921  -0.5216   1.1698   2.3194
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     -0.917029   0.240626  -3.811
## age                                              0.010473   0.001695   6.178
## genderMale                                       0.441927   0.054951   8.042
## race_ethnicityAsian (Asian Indian)              -0.659646   0.331966  -1.987
## race_ethnicityAsian (Chinese)                   -1.229315   0.376347  -3.266
## race_ethnicityAsian (Filipino)                  -0.064751   0.387698  -0.167
## race_ethnicityAsian (Japanese)                  -1.216854   0.605554  -2.009
## race_ethnicityAsian (Korean)                    -0.429913   0.659728  -0.652
## race_ethnicityAsian (Other)                     -0.959310   0.482980  -1.986
## race_ethnicityAsian (Vietnamese)                -0.603008   0.657357  -0.917
## race_ethnicityBlack, or African American        -1.901436   0.262931  -7.232
## race_ethnicityPacific Islander (Guamanian)      13.917555 535.411219   0.026
## race_ethnicityPacific Islander (Native Hawaiian) -0.485851   0.730270  -0.665
## race_ethnicityPacific Islander (Other)          -13.172019 187.870532  -0.070
## race_ethnicityPacific Islander (Samoan)         -12.874374 378.497582  -0.034
## race_ethnicitySome other race                    -0.675519   0.256212  -2.637
## race_ethnicityWhite                               0.084584   0.232584   0.364
##                                                 Pr(>|z|)
## (Intercept)                                     0.000138 ***
## age                                             6.47e-10 ***
## genderMale                                      8.82e-16 ***
## race_ethnicityAsian (Asian Indian)              0.046913 *
## race_ethnicityAsian (Chinese)                   0.001089 **
## race_ethnicityAsian (Filipino)                  0.867359
## race_ethnicityAsian (Japanese)                  0.044485 *
## race_ethnicityAsian (Korean)                    0.514626
## race_ethnicityAsian (Other)                     0.047008 *
## race_ethnicityAsian (Vietnamese)                0.358974
## race_ethnicityBlack, or African American        4.77e-13 ***
## race_ethnicityPacific Islander (Guamanian)      0.979262
## race_ethnicityPacific Islander (Native Hawaiian) 0.505856
## race_ethnicityPacific Islander (Other)          0.944104
## race_ethnicityPacific Islander (Samoan)         0.972866
## race_ethnicitySome other race                   0.008375 **
## race_ethnicityWhite                             0.716103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8243.9  on 6100  degrees of freedom
## Residual deviance: 7623.2  on 6084  degrees of freedom
## AIC: 7657.2
```

```
##
## Number of Fisher Scoring iterations: 12
```

```
# Creating the Logit model for Biden
model2 <- glm(vote_biden ~ age + gender+race_ethnicity,
            data=survey_data, family="binomial")
```

Post-Stratification

In order to obtain the estimated proportion of voters who voted for Donald Trump, we used the post-stratification techniques to partition the voting population into different cells based on age, gender, and race. Then we counted the number of observations for each cell and formed a new column named n in the dataset of census_data. Then we filtered out all the observations who are older than 18 years old (legal age for voting the election) and finally got the census_data with 1405 observations. Next, we used the function of predict() to obtain the proportion of each cell who voted for Trump, which formed another column named estimate. After that, we applied the formula that $\hat{y}$ is equal to sum(estimate*n)/sum(n) to obtain the value of the predictor. In this study, in order to do a comparison between the candidates, we calculated two predictors for Donald Trump and Joe Biden respectively, which we would compare the values of these two predictors and finally obtained the results whether who has a larger probability to win the 2020 election.

```
#Win chance for Trump
census_data<- census_data %>%
  mutate(age=age+2)%>%
  filter(age >=18)

census_data$estimate <-
  model %>%
  predict(newdata = census_data, type = "response")

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

```
## # A tibble: 1 x 1
##   alp_predict_Trump
##               <dbl>
## 1             0.467
```

```
#Win chance for Biden
census_data$estimate <-
  model2 %>%
  predict(newdata = census_data, type = "response")

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

```
## # A tibble: 1 x 1
##   alp_predict_Biden
##               <dbl>
## 1             0.321
```

Results

By observing the summary table, we obtained the Logistic Regression Model for predicting the logistic proportion of winning the 2020 election for Donald Trump.

```
summary(model)
```

```
##
## Call:
## glm(formula = vote_trump ~ age + gender + race_ethnicity, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4050  -1.0921  -0.5216   1.1698   2.3194
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     -0.917029   0.240626  -3.811
## age                                              0.010473   0.001695   6.178
## genderMale                                       0.441927   0.054951   8.042
## race_ethnicityAsian (Asian Indian)              -0.659646   0.331966  -1.987
## race_ethnicityAsian (Chinese)                   -1.229315   0.376347  -3.266
## race_ethnicityAsian (Filipino)                  -0.064751   0.387698  -0.167
## race_ethnicityAsian (Japanese)                  -1.216854   0.605554  -2.009
## race_ethnicityAsian (Korean)                    -0.429913   0.659728  -0.652
## race_ethnicityAsian (Other)                     -0.959310   0.482980  -1.986
## race_ethnicityAsian (Vietnamese)                -0.603008   0.657357  -0.917
## race_ethnicityBlack, or African American        -1.901436   0.262931  -7.232
## race_ethnicityPacific Islander (Guamanian)      13.917555 535.411219   0.026
## race_ethnicityPacific Islander (Native Hawaiian) -0.485851   0.730270  -0.665
## race_ethnicityPacific Islander (Other)          -13.172019 187.870532  -0.070
## race_ethnicityPacific Islander (Samoan)         -12.874374 378.497582  -0.034
## race_ethnicitySome other race                    -0.675519   0.256212  -2.637
## race_ethnicityWhite                               0.084584   0.232584   0.364
##                                                 Pr(>|z|)
## (Intercept)                                     0.000138 ***
## age                                             6.47e-10 ***
## genderMale                                      8.82e-16 ***
## race_ethnicityAsian (Asian Indian)              0.046913 *
## race_ethnicityAsian (Chinese)                   0.001089 **
## race_ethnicityAsian (Filipino)                  0.867359
## race_ethnicityAsian (Japanese)                  0.044485 *
## race_ethnicityAsian (Korean)                    0.514626
## race_ethnicityAsian (Other)                     0.047008 *
## race_ethnicityAsian (Vietnamese)                0.358974
## race_ethnicityBlack, or African American        4.77e-13 ***
## race_ethnicityPacific Islander (Guamanian)      0.979262
## race_ethnicityPacific Islander (Native Hawaiian) 0.505856
## race_ethnicityPacific Islander (Other)          0.944104
## race_ethnicityPacific Islander (Samoan)         0.972866
## race_ethnicitySome other race                   0.008375 **
## race_ethnicityWhite                             0.716103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8243.9  on 6100  degrees of freedom
## Residual deviance: 7623.2  on 6084  degrees of freedom
## AIC: 7657.2
##
## Number of Fisher Scoring iterations: 12
```

From the above summary table, we could see that $\hat{\beta}1$ is approximately equal to 0.0105, which means that as people increase in age, they are more likely to vote for Donald Trump. $\hat{\beta}2$ is approximately equal to 0.4419, which presents that males are more likely to vote for Donald Trump. In addition, by observing the estimated βs for the variable of race, we could see that the observations who are in the Asian area (including Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, and other Asian people),African American, and most of the areas in Pacific Islander (including Native Hawaiian, Samoan, and others in Pacific Islander) are less likely to vote for Donald Trump. While Guamanian (Pacific Islander) and white people are more likely to vote for Donald Trump, especially observations in Guam, whose estimated value for β is 13.9176, which means that they are highly supporting Donald Trump to win in this election. In addition, the p-value for the variables, age, genderMale, race_ethnicityAsian (Asian Indian), race_ethnicityAsian (Chinese), race_ethnicityAsian (Japanese), race_ethnicityAsian (Other), race_ethnicityBlack, and race_ethnicitySome other race, are less than 0.05, which means that these variables are significant to the response of logistic proportion. Through applying the method of Post-Stratification sampling, we divided the samples into several cells according to age, gender, and race to form the data set census_data. Then we calculated the estimated proportion for voting Donald Trump by substituting the data in each cell into the model we just obtained. Next, through calculating the mean of all estimated proportions we got the final estimator $\hat{y}$ for Donald Trump is 0.4670519. For comparison, we fitted a Logit Model for another popular candidate, Joe Biden. Then we repeated the above steps and obtained the estimator $\hat{y}$ for Joe Biden is 0.3208666.

Discussion

First, we cleaned the data by removing the part of people who would not vote, have no right to vote or not assigned data, and then change categorical data to binary data to build a model. We also chose three variables which were age, gender and race and wanted to investigate those factors' influences on the choices of voting. We used two logistic regression models with the cleaned census data, and survey data and our y was the probability of the proportion of voters who would vote for Donald Trump or Joe Biden divided by who would not vote for him. From the model, we could get the relationship between each factor to the votes by the p-values and estimates of the variables. From the post-stratification technique, we could get a result of Donald Trump or Joe Biden winning the election of 2020 US president. In conclusion, We think Donald Trump will win the primary vote. Based on our model, the estimated proportion of voters in favor of voting Donald Trump is 46.7%, and Joe Biden is 32.1%. Trump will have more votes than Biden so that he would probably be reappointed consecutively. In Trump's model, the estimate of age for Trump model is 0.01 and estimate of gender male is 0.44, which denotes that the elderly and males are more likely to vote for Trump. Of the 3,142 counties in the United states in 2019, 57.3% had a median age between 40.0 and 49.9 years and 6.8% had a median age 50 or older.(Bureau, 2020) Because of the aging of the U.S., Trump has plenty of older age supporters. Furthermore, the gender ratio in the United State in 2020 is 97.948 males per 100 females so that at the approximately same level of gender ratio Trump will get more support by male. (Gender ratio in the United States) Talking about race, most of them are negative estimates except Guamanian. However, 76.3% of American are white and it is the majority of the voting population. (U.S. Census Bureau QuickFacts: United States) Thus, Trump is expected to have more votes.

Weaknesses

About the limitations of the model, firstly, the variables we chose are age, gender, and race, which may be related to the privacy of some observations. Therefore, the results of the survey may have response error,

which may slightly affect the reliability of the final results. Moreover, people might have second thoughts on the candidate they would like to vote for in the end, so their choices could be different compared with the previous survey, which causes response error here as well. Thirdly, the p-values of some variables such as race_ethnicityAsian (Filipino), and race_ethnicityPacific Islander (Guamanian)) in the model are more than 0.05. It means that those variables are not significant to the response logistic proportion, so the predicted value may have errors.

Next Steps

From the above results, we know that most of the observations from Asia, Pacific Island and African American are less likely to vote for Donald Trump. While the white people and Guamanian are the group of people who are supporters for him. From these analysis, we want to further research some related information to explain the difference in the proportion of supporters between races, for example, the immigration policy in the USA, the current news about racism or the local policy for the foreign people and so on.

References

Gender ratio in the United States. (n.d.). Retrieved from http://statisticstimes.com/demographics/country/us-sex-ratio.php

Bureau, U. C. (2020, June 25). 65 and Older Population Grows Rapidly as Baby Boomers Age. Retrieved from https://www.census.gov/newsroom/press-releases/2020/65-older-population-grows.html

U.S. Census Bureau QuickFacts: United States. (n.d.). Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045219