Predicting the Proportion of Voters in Favour of Donald Trump in the 2020 American Federal Election

Celine Kim (Student Number: 1001251147)

2 November, 2020

Model

In order to predict the overall popular vote of the 2020 American federal election, I will use a logistic regression model with post-stratification. I am interested in predicting the proportion of voters who will vote for Donald Trump so I will be modelling a response variable on whether an individual will vote for Trump or not. The model specifics and calculations used in the post-stratification technique will be explained in the following subsections of this report.

Model Specifics

Since whether one will vote for Donald Trump or not in the 2020 American federal election is a binary outcome, a logistic regression model will be used to model this binary response variable. The aspects focused in this model are the employment status, gender, census region, racial ethnicity, age and birthplace of the voter. Firstly, according to the report “An examination of the 2016 electorate, based on validated voters” conducted by the Pew Research Center of U.S. Politics and Policy, “88% of Trump’s supporters were Whites, and the age group of Trump’s voters were higher than Clinton’s” which shows that age and racial ethnicity or birthplace influence voting decisions. In addition, a more recent 2020 report by the same source states that “attitudes of voters on race and gender in 2020 are more divided than in 2016”. Also, from a 2016 article by the Washington Post, there is a proposition that “Trump supporters are from communities with high unemployment rates” which shows that employment status and region of voters are another crucial factors to consider when predicting the result of the presidential election.

Thus, in this model, the following variables are used.

Let x_{emp} represent the employment status where 1 represents that the voter is employed and 0 otherwise.

Let x_{male} be the gender of the voter. This is a binary categorical variable where 1 represents that the voter is male and 0 if female.

Let x_{his} be a binary categorical variable where 1 represents that the voter is of Hispanic, Latino or Spanish origin and 0 otherwise.

Let x_{white} be a binary categorical variable where 1 represents that the voter’s racial ethnicity is White. Let x_{black} be a binary categorical variable where 1 represents that the voter’s racial ethnicity is Black or African American.

Let x_{30to44}, x_{45to59}, x_{60plus} be binary categorical variables where 1 represents that the voter belongs to the age group between 30 to 44, 45 to 59, 60 plus, respectively.

Let x_{bpl} be a binary variable that represents the birthplace of the voter where 1 represents that the voter was born in the United States of America.

The event of interest in this model is whether the voter will vote for for Donald Trump in the 2020 American federal election. Then let p be the probability that one will vote for Donald Trump in this election. In this model, $\log \frac{p}{1-p}$ is the log odds of one voting for Trump, the response variable.

Then this model in mathematical notation can be written as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_{emp} + \beta_2 x_{male} + \beta_3 x_{his} + \beta_4 x_{white} + \beta_5 x_{black} + \beta_6 x_{30to44} + \beta_7 x_{45to59} + \beta_8 x_{60plus} + \beta_9 x_{bpl}$$

β_0 is the intercept which represents when x_{emp}, x_{male}, x_{his}, x_{white}, x_{black}, x_{30to44}, x_{45to59}, x_{60plus} and x_{bpl} are all 0. In this specific model, this situation corresponds to the log odds of one voting for Trump when the voter is a foreign-born, unemployed female who is not of Hispanic, White or African American origin and in the age group between 18 to 29 (since voters of age of 17 and under has been excluded).

β_1 to β_9 are coefficients that represent the change in log odds for every one unit increase in the respective explanatory variables. Since all explanatory variables are binary variables, one unit increase in this case represents presence of the corresponding condition.

In the initial statistical model that can be found in the Appendix, variables such as the census region of the voter and whether the voter's racial ethnicity is American Indian, Asian or Pacific Islander were also included. However, using the p-value base of 0.05, I decided to remove such variables from the model. This is a sensible decision because people of Hispanic or African American origin in the U.S. are those in conflict with Trump while little or no conflict exists between Trump and Asians, Native Americans or Pacific Islanders.

Post-Stratification

Post-stratification is a statistical analysis technique that involves dividing data into small demographic cells, adding those cell-level estimates and extrapolating to create a population-level estimate in order to predict the response variable. This technique is useful because it enables correction for non-probability sampling errors and accurate estimate for population level predictions. Therefore, I performed a post-stratification analysis to estimate the "nationwide" proportion of voters who will vote for Donald Trump.

I created bins based on state, is_working, gender, hispanic, is_white, is_black, age_group and us_born because these were the parameters (that were also available in the census data) used to develop the logistic regression model. The variable "state" is included to perform further analysis where we make predictions on the proportion of voters per state on who will vote for Trump (please check the Additional Analysis Per State Basis). Initially, I wanted to include the voting data from the 2016 presidential election, preference of news source such as CNN or Fox, primary party that one follows as these data were available in the Democracy Fund and UCLA Nationscape 'Full Data Set', but the census data from the American Community Surveys (ACS) did not include such information so they were removed. These factors were originally considered because according to an article called "Fox News was No. 1 news source for Trump voters", "40% of those who voted for Trump relies on Fox News as the main source of news in the 2016 election" which shows that news source is another aspect to consider.

Additional Analysis Per State Basis

The United States presidential election is an indirect voting system where the citizens vote for the members of the Electoral College instead of directly choosing between the candidates. After the general election where citizens cast a vote, the electors then vote based on the state results. Finally, the presidential candidate is decided based on the results of votes of the electoral college. Thus, a more accurate prediction would require calculations per state basis. Therefore, I grouped by state and included it as a bin to conduct post-stratification analysis on vote per state.

The result of this additional analysis is in Table 3 of the Appendix.

Results

Table 1: Summary of Logistic Regression Model

Below is the summary of the logistic regression model used in this report.

```
## # A tibble: 10 x 5
##    term                           estimate std.error statistic  p.value
##    <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                      -1.91     0.146     -13.1  5.89e-39
##  2 as.factor(is_working)1            0.209    0.0596      3.50 4.59e- 4
##  3 as.factor(gender)2                0.374    0.0557      6.73 1.73e-11
##  4 as.factor(hispanic)1             -0.321    0.0866     -3.70 2.13e- 4
##  5 as.factor(is_white)1              0.519    0.0926      5.60 2.14e- 8
##  6 as.factor(is_black)1             -1.47     0.156      -9.44 3.90e-21
##  7 as.factor(age_group)ages30to44    0.536    0.0828      6.48 9.35e-11
##  8 as.factor(age_group)ages45to59    0.681    0.0874      7.80 6.36e-15
##  9 as.factor(age_group)ages60plus    0.710    0.0896      7.92 2.38e-15
## 10 as.factor(us_born)1               0.408    0.120       3.41 6.53e- 4
```

Since the p-value of variables in the logistic regression model are all below 0.05, this provides grounds to reject the null hypotheses that β_0 to β_9 = 0.

This model can be mathematically expressed as:

$$\log \frac{p}{1-p} = -1.9063 + 0.2089x_{emp} + 0.3745x_{male} - 0.3208x_{his} + 0.5188x_{white} - 1.4687x_{black} + 0.5364x_{30to44} + 0.6814x_{45to59}$$

$$+0.7100x_{60plus} + 0.4077x_{bpl}$$

Table 2: Results from Post-Stratification Analysis

```
## [1] -0.4973671
```

```
## [1] 0.3781596
```

The results from the post-stratification analysis are that $\log \frac{p}{1-p}$ = -0.4973671 and $\hat{y}^{PS}$ = 0.3781596.

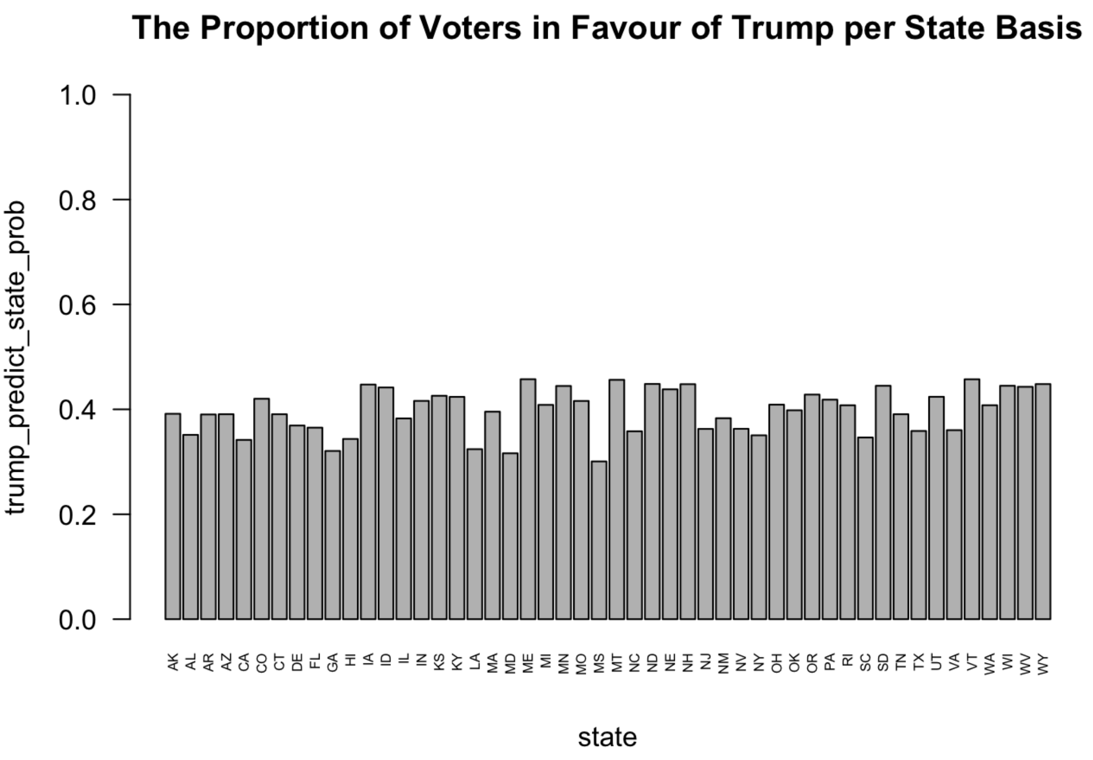
Using the post-stratification analysis of the proportion of voters in favour of Trump modelled by a logistic regression model, I estimate that the proportion of voters in favour of voting for Trump to be 0.3782. The logistic regression model used in this analysis accounts for the employment status, gender, racial ethnicity, and age group of the voter and whether the voter was born in the U.S. or not.

The result of the prediction was trump_predict_overall which is the log odds of voting for Donald Trump. Then, I converted the log odds to the probability by calculating exp(trump_predict_overall) /(1+exp(trump_predict_overall)).

Table 3 of the Appendix is the result from the additional analysis performed to estimate the proportion of voters in favour of Trump per state basis. As shown in Table 3 of the Appendix, the log odds of voting for Donald Trump for all states is less than 0, meaning that the proportion will be less than 0.5 in all states.

Figure 1: Barplot of The Proportion of Voters in Favour of Trump per State Basis

Below is the barplot of the proportion of voters in favour of Trump per state basis using the result of the additional analysis.



Discussion

From the survey data obtained from the Democracy Fund and UCLA Nationscape ‘Full Data Set’, I selected the employment status, gender, census region, racial ethnicity, age and birthplace as factors since they were shown to have influence on voting decisions. Then, I built a logistic regression model that predicts voting in favour of Donald Trump using these data at an individual level. Then, we conducted post-stratification analysis using census data from 2018, to predict proportion of voters nationwide and per state basis (used in the Additional Analysis Per State Basis) in favour of Donald Trump.

In conclusion, Donald Trump will lose not only the overall nationwide popular vote of the 2020 American federal election but also in every state. As shown in the nationwide election results of Table 2, only 0.379 will vote for Trump, and in calculations based on each state in Table 3 of the Appendix, the proportion in each state is less than 0.5. The maximum proportion is 0.4574 in Maine. The proportion was between 0.4 and 0.5 in about 50% of the states, and the other 50% of the states had proportions below 0.4. Thus, it is not likely that Trump will win. Therefore, regardless of which voting system the U.S. is using, it is likely that Donald Trump will lose the 2020 presidential election.

Weaknesses

There are other important parameters that have great influence on predicting the result of the 2020 presidential election including the voting data from the 2016 election, the main preference for news sources as mentioned before, opinions on how the current government is coping with the coronavirus outbreak and personal thoughts on policies such as the Green New Deal and deportation of unregistered immigrants. The citizens' opinions on such regulations have great influence on the voting decisions. For example, while the Green New Deal is supported by the Democratic Party, Trump opposes the regulation. According to an article by the Washington Post (cited in #7 of the Reference Section), the opinion is divided on whether the country needs to cut down on the use of fossil fuels between Democrats and Republicans and "most Americans oppose paying trillions of dollars by tax to reach the goals of the regulation". However, these variables were not included in the analysis because such data were not available in census data provided by ACL. Even though these variables might have greater impact than the variables we studied, since the census data includes more objective information that does not inquire specifically on voting or political issues, I was not able to utilize these parameters. Especially, the vote rate per state in the previous election might be a very important factor when calculating the proportion of voters per state. In addition, since this model uses data based on June 25th, 2020, the model does not take into account recent changes that happened after June 25th including Trump getting COVID-19 and election debates between September and October which might have influenced or changed people's voting decisions.

Next Steps

If the data on the voting rates per age group in each state is available, this information can be used to re-adjust the proportion per bin because the population proportion does not represent the proportion of those who actually vote. This additional information can yield a more accurate prediction. In addition, since the actual data will be available after the election occurs, post hoc analysis can be performed to examine if each variable used in this analysis are critical variables that influence the voting trend and analyze to improve the model for better prediction in the future elections.

References

1. Tausanovitch, C., & Vavreck, L. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200131). Retrieved from https://www.voterstudygroup.org/downloads?key=b5385d6e-a729-4a43-8ab8-4c6d6fca3100 (https://www.voterstudygroup.org/downloads?key=b5385d6e-a729-4a43-8ab8-4c6d6fca3100).

2. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. (2020) 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)

3. Pew Research Center. (2018, August 09). An examination of the 2016 electorate, based on validated voters. Retrieved October 29, 2020, from https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/ (https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/)

4. Pew Research Center. (2020, September 10). Voters' Attitudes About Race and Gender Are Even More Divided Than in 2016. Retrieved October 30, 2020, from https://www.pewresearch.org/politics/2020/09/10/voters-attitudes-about-race-and-gender-are-even-more-divided-than-in-2016/ (https://www.pewresearch.org/politics/2020/09/10/voters-attitudes-about-race-and-gender-are-even-more-divided-than-in-2016/)

5. Aytaç, S., & Rau, E. (2019, April 18). Trump supporters vastly overestimate unemployment - and they blame politicians for it. Retrieved October 29, 2020, from https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/02/trump-supporters-vastly-overestimate-unemployment-and-they-blame-politicians-for-it/ (https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/02/trump-supporters-vastly-overestimate-unemployment-and-they-blame-politicians-for-it/)

6. Sutton, K. (2017, January 18). Fox News was No. 1 news source - for Trump voters. Retrieved November 01, 2020, from https://www.politico.com/blogs/on-media/2017/01/study-fox-news-is-no-1-news-source-for-trump-voters-233773 (https://www.politico.com/blogs/on-media/2017/01/study-fox-news-is-no-1-news-source-for-trump-voters-233773)

7. Clement, S., & Grandoni, D. (2019, December 04). Americans like Green New Deal's goals, but they reject paying trillions to reach them. Retrieved November 01, 2020, from https://www.washingtonpost.com/climate-environment/2019/11/27/americans-like-green-new-deals-goals-they-reject-paying-trillions-reach-them/ (https://www.washingtonpost.com/climate-environment/2019/11/27/americans-like-green-new-deals-goals-they-reject-paying-trillions-reach-them/)

8. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

9. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom (https://CRAN.R-project.org/package=broom)

Appendix

Below are the libraries and data used for this analysis.

```
library(tidyverse)
library(broom)

# Loading in the cleaned survey Data
survey_data <- read_csv("/Users/celinekim/Desktop/ps3/survey_data.csv")

# Loading in the cleaned census Data
census_data <- read_csv("/Users/celinekim/Desktop/ps3/census_data.csv")
```

Below is the initial model created for the analysis. The explanatory variables used in this initial model are is_working, gender, census_region, hispanic, is_white, is_black, is_american_indian, is_asian_or_pacific_islander, age_group and us_born. The variables that have been excluded in the final report are census_region, is_american_indian and is_asian_or_pacific_islander.

```
# The initial model
initial_model<-glm(as.factor(vote_trump_2020) ~  as.factor(is_working) +
                     as.factor(gender) + as.factor(census_region) +
                     as.factor(hispanic) + as.factor(is_white) +
                     as.factor(is_black) + as.factor(is_american_indian) +
               as.factor(is_asian_or_pacific_islander) + as.factor(age_group) +
             as.factor(us_born), data=survey_data, family="binomial")
```

Below is the summary data of the initial model that was not included in the final report.

```
# Model Results
broom::tidy(initial_model)
```

```
## # A tibble: 15 x 5
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                               -2.04      0.198    -10.3   6.56e-25
##  2 as.factor(is_working)1                     0.224     0.0600     3.74  1.87e- 4
##  3 as.factor(gender)2                         0.398     0.0561     7.10  1.26e-12
##  4 as.factor(census_region)2                  0.0770    0.0858     0.897 3.70e- 1
##  5 as.factor(census_region)3                  0.412     0.0772     5.34  9.48e- 8
##  6 as.factor(census_region)4                  0.0520    0.0857     0.607 5.44e- 1
##  7 as.factor(hispanic)1                      -0.352     0.0926    -3.80  1.43e- 4
##  8 as.factor(is_white)1                       0.473     0.130      3.65  2.65e- 4
##  9 as.factor(is_black)1                      -1.59      0.181     -8.74  2.38e-18
## 10 as.factor(is_american_indian)1             0.262     0.259      1.01  3.11e- 1
## 11 as.factor(is_asian_or_pacific_islander…  -0.154     0.189     -0.813 4.16e- 1
## 12 as.factor(age_group)ages30to44             0.543     0.0832     6.53  6.70e-11
## 13 as.factor(age_group)ages45to59             0.691     0.0879     7.86  3.73e-15
## 14 as.factor(age_group)ages60plus             0.726     0.0902     8.05  8.40e-16
## 15 as.factor(us_born)1                        0.381     0.124      3.07  2.18e- 3
```

Below is the logistic regression model used for the analysis in the final report.

```r
# Creating the Model
model <- glm(as.factor(vote_trump_2020) ~  as.factor(is_working) +
             as.factor(gender) + as.factor(hispanic) + as.factor(is_white) +
             as.factor(is_black) + as.factor(age_group) + as.factor(us_born),
           data=survey_data, family="binomial")
```

Below is the post-stratification analysis.

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

census_data <- census_data %>%
  mutate(trump_predict_prop = estimate*cell_prop_of_division_total)

logit2prob <- function(logit){
  odds <- exp(logit)
  prob <- odds / (1 + odds)
  return(prob)
}

trump_predict_overall <- sum(census_data[, 'trump_predict_prop'], na.rm=TRUE)
trump_predict_overall_prob <- logit2prob(trump_predict_overall)
```

Table 3: Results of Additional Analysis Performed Per State Basis

```
## # A tibble: 50 x 3
##    state trump_predict_state trump_predict_state_prob
##    <chr>               <dbl>                    <dbl>
##  1 AK                 -0.441                    0.392
##  2 AL                 -0.613                    0.351
##  3 AR                 -0.446                    0.390
##  4 AZ                 -0.444                    0.391
##  5 CA                 -0.655                    0.342
##  6 CO                 -0.322                    0.420
##  7 CT                 -0.444                    0.391
##  8 DE                 -0.536                    0.369
##  9 FL                 -0.553                    0.365
## 10 GA                 -0.751                    0.321
## 11 HI                 -0.647                    0.344
## 12 IA                 -0.212                    0.447
## 13 ID                 -0.234                    0.442
## 14 IL                 -0.478                    0.383
## 15 IN                 -0.339                    0.416
## 16 KS                 -0.299                    0.426
## 17 KY                 -0.307                    0.424
## 18 LA                 -0.735                    0.324
## 19 MA                 -0.424                    0.396
## 20 MD                 -0.771                    0.316
## 21 ME                 -0.171                    0.457
## 22 MI                 -0.370                    0.409
## 23 MN                 -0.223                    0.445
## 24 MO                 -0.339                    0.416
## 25 MS                 -0.844                    0.301
## 26 MT                 -0.175                    0.456
## 27 NC                 -0.583                    0.358
## 28 ND                 -0.207                    0.448
## 29 NE                 -0.248                    0.438
## 30 NH                 -0.209                    0.448
## 31 NJ                 -0.564                    0.363
## 32 NM                 -0.476                    0.383
## 33 NV                 -0.563                    0.363
## 34 NY                 -0.617                    0.350
## 35 OH                 -0.368                    0.409
## 36 OK                 -0.413                    0.398
## 37 OR                 -0.289                    0.428
## 38 PA                 -0.329                    0.418
## 39 RI                 -0.373                    0.408
## 40 SC                 -0.635                    0.346
## 41 SD                 -0.221                    0.445
## 42 TN                 -0.444                    0.391
## 43 TX                 -0.581                    0.359
## 44 UT                 -0.307                    0.424
## 45 VA                 -0.574                    0.360
## 46 VT                 -0.171                    0.457
## 47 WA                 -0.373                    0.408
## 48 WI                 -0.221                    0.445
## 49 WV                 -0.229                    0.443
## 50 WY                 -0.208                    0.448
```