The next U.S. president? Donald Trump or Joe Biden?

Shlok Somani & James Bai

November 2, 2020

Github Repo : https://github.com/shloksomani/-STA304-PS3

Model

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

Model Specifics

We will be using two logistic regression models to model the proportion of voters who will vote for Donald Trump and Joe Biden respectively. Since a logistic regression is a statistical model that uses logistic function to model a binary response variable, so here it would be good to use it to predict whether the voter would vote for Trump or Biden as it there is only two outcomes. We will only be using age, gender, race ethnicity and the household income to predict the outcome of the 2020 American federal election. We have chosen these as our predictor variables because different age and gender could affect on who the voters decides. Race ethnicity would play a role because of how president Trump handled Black Lives Matter would change each race's opinion on Him. Also the household income would affect the outcome because both presidential candidates both have different plans on taxes. The logistic regression model we are using for Trump is:

$$log(\frac{\hat{p}_{Trump}}{1-\hat{p}_{Trump}}) = \beta_{T0} + \beta_{T1}x_{age} + \beta_{T2}x_{gender} + \beta_{T3}x_{race\_ethnicity} + \beta_{T4}x_{household\_income}$$

The logistic regression model we are using for Biden is:

$$log(\frac{\hat{p}_{Biden}}{1-\hat{p}_{Biden}}) = \beta_{B0} + \beta_{B1}x_{age} + \beta_{B2}x_{gender} + \beta_{B3}x_{race\_ethnicity} + \beta_{B4}x_{household\_income}$$

Where $\hat{P}$ represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, and is the probability of voting for Trump or Biden if the voter the voter didn't give any information on them. Additionally, $\beta_{1,...,4}$ represents the slope of the respective predictors. So, depending on what age group, gender, race ethnicity and household income range the voter belongs to, we expect $\beta_{1,...,4}$ to be different for each model.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump and Joe Biden we need to perform a post-stratification analysis. A post-stratification analysis is basically finding the sum of the same

models with different prediction multiplying the population related that said predictor, dividing that by the sum of the population related to each predictor. This is useful at predicting the outcome of the presidential election because it would represents the whole population. Here we create cells based of each explanatory variables in our model ages groups, income brackets, race and gender. Using the model described in the previous sub-section we will estimate the proportion of voters in each predictor bin. We will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_i}{\sum N_j}$$

$$\hat{p} = \frac{e^{\hat{y}^{PS}}}{1 + e^{\hat{y}^{PS}}}$$

Results

We estimate that the proportion of voters in favour of voting for Donald Trump to be $\hat{y}^{PS}_{Trump} = 0.3868$. The proportion of voters for Joe Biden to be $\hat{y}^{PS}_{Biden} = 0.4111$. This is based off our post-stratification analysis of the proportion of voters in favour of each presidential candidate modeled by a logistic regression model, which accounted for ages groups, income brackets, race and gender. You can notice that the estimation does not add up to 1, we think that this is the case because around 10% of the observation of the survey data didn't specify which president they were going to vote for. Overall, we think that Joe Biden would be the favours going into this presidential election.

Discussion

This research focuses on the two majority competitors in the upcoming 2020 federal elections of the United States of America. The candidates - Joe Biden and Donald Trump - are the primary subjects in our model as we use various factors to make an educated prediction on who has a higher probability of winning the elections. We used survey data provided by Democracy Fund + UCLA Nationscape and implemented post-stratification analysis on Census data provided by IPUMS in order to correctly estimate the proportion of people voting for either Trump or Biden. To further segment our data and identify voting behaviours, we used four independent/explanatory variables namely age groups, income brackets, race and gender with regards to their relevancy. Our results gave us definitive answers, in that, there was a significant impact of the four variables on the voting outcomes.

Additionally, the other axis is a binary variable called "voteTrump" or "voteBiden." With the help of multilevel regression with post-stratification, we use this binary variable to estimate the proportion of voters in favour of either candidate to most accurately predict the outcomes of the 2020 federal election.

Weaknesses

Despite our best efforts to reduce biases and eliminate factors that could potentially be skewing our model, there are several weaknesses that we have identified within our model:

- While we have included variables that we deemed were the biggest influencers in the decision making of voters, a detailed analysis of this topic would ideally include more census data surrounding the demographics and income levels of voters. Though our current results offer us a reasonably reliable trend with regards to the topic, the lack of further data gives us a limited perspective on voter behaviour overall.

- Additionally, due to discrepancies within the definitions of variables; such as the options provided to respondants under the 'race' identifier, there were approximately 6000 of 3M respondents that fell between a variety of categories. In an ideal study of this topic, consistent variables and response options would be given to survey respondents.
- Lastly, we were unable to find information regarding their choice and actions during the previous elections. Several studies have consistently proven that past behaviour is one of the best indicators and predictor of current or future decision-making. The lack of this particular variable significantly weakens the results of our model and would be an area to improve upon for further studies.

Next Steps

Based on the results of this model and analysis, we can take our research further in a few ways to enhance the utilization and accuracy of this concept.

- Examine the results from this year's election to understand the accuracy of our model and take this as an opportunity to identify weak links within the current setup.
- Include further details within variables and also include other candidates running in the election. Other factors that could potentially make a difference in a survey include the term of the candidacy, i.e. Trump is running for his re-election versus Biden who is a first-time candidate.

References

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0

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

Appendix

```
# Creating the Model
modelTrump <- glm(voteTrump ~ ageCategory + gender + race_ethnicity + household_income,
             data = survey_data, family = "binomial")

modelBiden <- glm(voteBiden ~ ageCategory + gender + race_ethnicity + household_income,
             data = survey_data, family = "binomial")
# Model Results (to Report in Results section)
modelSummaryTrump <- broom::tidy(modelTrump)
modelSummaryBiden <- broom::tidy(modelBiden)


## Post Stratification

census_data <- census_data %>%
  filter(ageCategory != "Respondent Skipped") %>%
  filter(ageCategory != "Not eligible to vote")

### Trump
census_data$logodds_estimate_trump <-
  modelTrump %>%
  predict(newdata = census_data)

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

trump <- census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict = sum(alp_predict_prop)/sum(n))



### Biden
census_data$logodds_estimate_biden <-
  modelBiden %>%
  predict(newdata = census_data)

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

biden <- census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict = sum(alp_predict_prop)/sum(n))


trumpProb <- as.numeric(trump)
bidenProb <- as.numeric(biden)

print(modelSummaryTrump)
```

```
## # A tibble: 22 x 5
```

```
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                              -0.974    0.259      -3.76 1.71e- 4
##  2 ageCategory25 to 34                       0.450    0.116       3.89 1.01e- 4
##  3 ageCategory35 to 44                       0.761    0.113       6.71 1.89e-11
##  4 ageCategory45 to 54                       0.840    0.119       7.08 1.46e-12
##  5 ageCategory55 to 64                       0.813    0.117       6.98 3.00e-12
##  6 ageCategory65 to 74                       0.844    0.121       6.96 3.37e-12
##  7 ageCategory75 or older                    1.14     0.173       6.61 3.76e-11
##  8 genderMale                                0.356    0.0553      6.44 1.16e-10
##  9 race_ethnicityBlack, or African Americ~  -1.78     0.257      -6.91 4.98e-12
## 10 race_ethnicityChinease                   -1.29     0.372      -3.46 5.45e- 4
## # ... with 12 more rows
```

```
print(modelSummaryBiden)
```

```
## # A tibble: 22 x 5
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                             -0.762      0.262    -2.90   3.71e- 3
##  2 ageCategory25 to 34                     -0.0352     0.0960   -0.366  7.14e- 1
##  3 ageCategory35 to 44                     -0.111      0.0967   -1.15   2.52e- 1
##  4 ageCategory45 to 54                     -0.263      0.104    -2.54   1.11e- 2
##  5 ageCategory55 to 64                     -0.00829    0.101    -0.0822 9.34e- 1
##  6 ageCategory65 to 74                      0.113      0.106     1.07   2.85e- 1
##  7 ageCategory75 or older                  -0.156      0.166    -0.939  3.48e- 1
##  8 genderMale                              -0.283      0.0531   -5.33   9.57e- 8
##  9 race_ethnicityBlack, or African Americ~  1.57       0.247     6.35   2.08e-10
## 10 race_ethnicityChinease                   1.23       0.324     3.81   1.41e- 4
## # ... with 12 more rows
```