Predicting Election Turnout with Multilevel Regression and Poststratification

Lisa Oh, Saksham Ahluwalia, Labib Chowdhury, Eric Yuan

November 2, 2020

Model

We are interested predicting the proportion of voters who vote for Donald Trump in the 2020 US Presidential 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

A logistic regression model was chosen to model the proportion of voters who will vote for Donald Trump versus Joe Biden in the 2020 United States Presidential Election. It is represented by the following equation (Eq. 1):

$$\log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i$$

The response variable p indicates whether someone votes for Donald Trump. β_0 represents the intercept term, which represents the case where every factor level in the chosen predictors is at the initial level. Similarly, $\sum_{i=1}^{K} \beta_i x_i$ indicates the slope β_i associated with each factor level of the predictors x_i. This conveys that the slope associated with each factor level represents the change in log odds for every unit increased in the associated predictor. Also, K represents the total number of factor levels in each predictor. The predictors for our model are age, race, household income, and state.

The logistic model was chosen for this problem space due to the binary nature of the response variable `vote_trump`, where `1` represents a vote for Donald Trump and `0` represents a vote for Joe Biden (all other responses, such as 'Will not vote', were removed). The predictors were selected as we believe these are key factors when one decides which candidate to vote for, and we wanted to explore whether or not these factors are as significant as we believe them to be.The model from Eq. 1 was fit using survey data from the voter study group data collected by UCLA and Democracy Fund (Tausanovitch and Vavreck, 2020).

Post-Stratification

Afterwards, we employed the post-stratification technique. In this technique, assuming demographics information is collected, each observation in a dataset is categorized into specific bins/cells by select demographic variables. Thus, by making estimates for the variable of interest within each cell, a population-level estimate can be extrapolated by weighting each cell by its relative proportion in the dataset. Post-stratification is useful when the data is not necessarily representative of the population. This is especially common if the individuals chosen for the dataset were not selected at random (otherwise known as non-probability sampling). The formula to employ the post-stratification technique is the following (Eq 2.):

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{K} N_j \hat{y}_j}{\sum_{j=1}^{K} N_j}$$

where K is the number of cells, N_j is the size of each cell, and $\hat{y}_j$ is the cell-level estimate.

For our investigation, we partitioned our census data into cells by age, race, household income, and state. We chose household income because it is likely to influence voter outcome because of the different proposed policies by the two political parties. Variables related to dwelling were not included because they are likely dependent on household income. In total, we created 42,217 cells. Cell-level estimates of voter outcome were made by the logistic model previously described.

Results

Table 1 below shows the summary of the first 10 weights.

```
## # A tibble: 87 x 5
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                                0.380     0.402     0.946 3.44e- 1
##  2 as.factor(age_cat)26-35                    0.407     0.122     3.33  8.83e- 4
##  3 as.factor(age_cat)36-45                    0.589     0.121     4.88  1.06e- 6
##  4 as.factor(age_cat)46-55                    0.825     0.129     6.40  1.59e-10
##  5 as.factor(age_cat)56-65                    0.622     0.124     5.02  5.28e- 7
##  6 as.factor(age_cat)66-75                    0.543     0.132     4.11  3.94e- 5
##  7 as.factor(age_cat)76-85                    0.871     0.214     4.08  4.57e- 5
##  8 as.factor(age_cat)85 and over              1.07      0.722     1.49  1.37e- 1
##  9 as.factor(race)black/african american/~   -2.28      0.307    -7.41  1.26e-13
## 10 as.factor(race)chinese                    -1.43      0.436    -3.28  1.02e- 3
## # ... with 77 more rows
```

(Table 1: Summary of Model)

From Table 1, we observe the slope for the "black/african american/negro" race level has a value of -2.28, which has the greatest magnitude in our model. This shows that the log odds of a voter voting for Trump, if they identify as a member of this race, decreases by a factor of 2.28.

The result of the post-stratification technique is displayed below. This value tells us that the estimated proportion of voters in favor of Trump over Biden modeled by a Logistical Regression model is about 0.507.

```
##   alp_predict
## 1    0.506598
```

Furthermore, upon studying the census data below (Table 2) we found a sharp division in the 'age_cat' variable which represents age group. Only 36% of individuals aged 18-25 were in favour of voting for Trump compared to 61% of individuals aged 85 and over. We also observed a linearly increasing trend in 'age_cat' with respect to voting in favour of Trump. As a result, if an individual is aged 85 or over then the log odds that they vote for Trump increases by a factor of 1.07 compared to a factor of 0.41 if the individual is aged 18-25 (Table 1).

```
## # A tibble: 8 x 2
##   age_cat     alp_predict
##   <chr>             <dbl>
## 1 18-25             0.368
## 2 26-35             0.458
## 3 36-45             0.510
## 4 46-55             0.569
## 5 56-65             0.525
## 6 66-75             0.505
## 7 76-85             0.578
## 8 85 and over       0.617
```

(Table 2: Prediction of Proportion of Trump Votes by Age Category)

We also found individuals who identified as 'White' and 'American Indian or Alaska native' were more likely to vote for Trump compared to individuals from the 'African American' community (Table 3).

```
## # A tibble: 7 x 2
##   race                              alp_predict
##   <chr>                                   <dbl>
## 1 american indian or alaska native        0.559
## 2 black/african american/negro            0.137
## 3 chinese                                 0.239
## 4 japanese                                0.256
## 5 other asian or pacific islander         0.417
## 6 other race, nec                         0.352
## 7 white                                   0.569
```

(Table 3: Prediction of Proportion of Trump Votes by Race)

Discussion

Summary

As we are trying to model the proportion of voters who will vote for Donald Trump, we proceeded with a logistic regression model as the response variable we are looking for is of binary nature. This model was trained on four predictor variables: age, household income, residing state, and race. We then employed the post-stratification technique by partitioning the data we retrieved from the Integrated Public Use Microdata Series (Ruggles et al., 2020) into their own prospective household income group, and then applying the logistic model to the partitioned dataset.

Conclusion

From our procedure, we estimate that the proportion of voters that are in favour of voting for the Republican party is 0.507; we predict that the Republican party will win the 2020 election and Donald Trump will be the next president.

Weaknesses

One of the main weaknesses of our analysis is that we disregard electoral colleges, and solely assume that having the popular vote will determine the winner of the presidential election. There has been five times in US history where the president did not have the popular vote, but still won the election: John Quincy Adams in the 1824 election, Rutherford B. Hayes in the 1876 election, Benjamin Harrison in the 1888 election, George W. Bush in the 2000 election, and Donald Trump in the 2016 election (Law, 2019). Another weakness is that our model only looks at four predictors.

Next Steps

To improve upon the weaknesses of our analysis we can factor in the electoral colleges. That is, we determine what the potential swing states are based on historical voting data, and assign a weighting to these states as they will have a higher impact on the election when compared to non-swing states. Furthermore, we can also look into adding more covariates to improve on our existing model or fitting a different type of model to the survey data. It would also be interesting to compare our results with the true election results and carry out a post-hoc analysis on how to better improve our estimation in future elections.

References

Law, T. (2019, May 15). These Presidents Won Electoral College But Not Popular Vote. Retrieved November 01, 2020, from https://time.com/5579161/presidents-elected-electoral-college/

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, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=dd05a560-c382-42f1-a3e4-7e5318f9781a

Wickham, H. (n.d.). Plyr. Retrieved October 30, 2020, from https://www.rdocumentation.org/packages/plyr/versions/1.8.6/topics/revalue