Prediction of 2020 American Federal Election

Shuoyu Chen, Yiling Song

November 2, 2020

Prediction of 2020 American Federal Election

Shuoyu Chen, Yiling Song

November 2, 2020

Model

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 are using random intercept models with following properties: separately logistic regression models of whether or not the person vote for Trump/whether or not the person vote for Biden, based on age and gender, using region to model the intercepts. A random intercepts model is a model in which intercepts are allowed to vary, and therefore, the scores on the dependent variable for each individual observation are predicted by the intercept that varies across groups. The general model we use in this report is:

$$Y_{vote\_T/B} = \beta_{int\_race} + \beta_{age}x_{age} + \beta_{genderM}x_{genderM} + \epsilon$$

$$\beta_{int\_race} = \beta_{intercept} + effect_{race}$$

Where $Y_{vote\_T/B}$ represents the proportion of voters who will vote for Donald Trump/Joe Biden. Similarly, $\beta_{intercept}$ is the intercept of the model, and is the probability of voting for Donald Trump/Joe Biden at age 0 and being female. The coefficient of the age variable is β_{age}, which represents change in odds. That is, as age of every voters increases by one unit, the probability of voting for Donald Trump/Joe Biden will increase by β_{age} in odds. For the gender being male variable, the coefficient is $\beta_{genderM}$, which means the log odds of the probability of voting for Donald Trump/Joe Biden will increases by $\beta_{genderM}$ if gender of the voter is male.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump we need to perform a post-stratification analysis. The post-stratification is a common technique in survey analysis for incorporating population distributions of variables into survey estimates. We use demographics to "extrapolate" how entire population will vote. $\hat{y}^{ps} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$ where $\hat{y_j}$ is the estimate in each cell and N_j is the population size of the j^{th} cell based off demographics. Here we create cells based off different race, more precisely, "white", "black/african american/negro" and "others". The reason of choosing race as our cells is that it can obviously show which candidate is prefered by specific races of people. Using the model described in the previous sub-section we will estimate the proportion of voters in each race 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.

Results

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_trump ~ age + gender + (1 | Race)
##    Data: survey_data
##
##      AIC      BIC   logLik deviance df.resid
##   7991.8   8018.9  -3991.9   7983.8     6471
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.2985 -0.8594 -0.4699  1.0374  3.9584
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Race   (Intercept) 0.6697   0.8183
## Number of obs: 6475, groups:  Race, 3
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.929987   0.480606  -4.016 5.93e-05 ***
## age          0.012679   0.001657   7.652 1.98e-14 ***
## genderMale   0.444969   0.053809   8.269  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##            (Intr) age
## age        -0.146
## genderMale -0.053 -0.003
```

```
## # A tibble: 3 x 2
##   Race                         alp_predict
##   <chr>                              <dbl>
## 1 black/african american/negro       0.100
## 2 others                             0.262
## 3 white                              0.445
```

- This is our model summary and $\hat{y}^{ps}$ result for Trump. More clearly, $Y_{vote\_T} = -1.9300 + 0.0127x_{age} + 0.4450x_{genderM} + \epsilon$, Where Y_{vote_T} represents the proportion of voters who will vote for Donald Trump, -1.9300 is the probability of voting for Donald Trump at age 0 and being female, 0.0127 is the probability of voting for Donald Trump will increase in odds as age of every voters increases by one unit, and the log odds of the probability of voting for Donald Trump will increase by 0.4450 if gender of the voter is male. And we estimate that the proportion of voters of race white in favour of voting for Trump is 0.4448, proportion of voters of race black/african american/negro is 0.1001, and proportion of voters of others races is 0.2623.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_biden ~ age + gender + (1 | Race)
##    Data: vote_Biden_data
##
##      AIC      BIC   logLik deviance df.resid
##   8555.4   8582.5  -4273.7   8547.4     6471
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5465 -0.8211 -0.7103  1.1845  1.4419
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Race   (Intercept) 0.2472   0.4972
## Number of obs: 6475, groups:  Race, 3
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.040762   0.297349   0.137    0.891
## age          0.002277   0.001592   1.430    0.153
## genderMale  -0.289967   0.051861  -5.591 2.25e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##            (Intr) age
## age        -0.218
## genderMale -0.070 -0.021
```

```
## # A tibble: 3 x 2
##   Race                          alp_predict
##   <chr>                               <dbl>
## 1 black/african american/negro        0.658
## 2 others                              0.463
## 3 white                               0.373
```

- This is our model summary and $\hat{y}^{ps}$ result for Biden. More clearly, $Y_{vote\_B} = 0.0408 + 0.0023x_{age} - 0.2900x_{genderM} + \epsilon$, Where Y_{vote_B} represents the proportion of voters who will vote for Joe Biden, 0.0408 is the probability of voting for Joe Biden at age 0 and being female, 0.0023 is the probability of voting for Joe Biden will increase in odds as age of every voters increases by one unit, and the log odds of the probability of voting for Joe Biden will decrease by 0.2900 if gender of the voter is male. And we estimate that the proportion of voters of race white in favour of voting for Biden is 0.3731, proportion of voters of race black/african american/negro is 0.6585, and proportion of voters of others races is 0.4629.

Discussion

Summary & Conclusion

In conclusion, for Trump, our estimated proportions of voters of race white, black/african american/negro, and others in favour of voting for him are 0.4448, 0.1001, and 0.2623 separately. For Biden, our estimated proportions of voters of race white, black/african american/negro, and others in favour of voting for him are 0.3731, 0.6585, and 0.4629 separately. Based on above results, we predict that Biden will win the election since obviously, though for voters of race white, 0.3731 is smaller than 0.4448, for voters of the two other groups of races, no matter 0.6585 or 0.4629, both are larger than 0.1001 and 0.2623.

Weaknesses & Next Steps

In our logistic regression model, we did only two main races, white and black/african american/negro, since the categories of races in both survey dataset and census dataset were totally different and complex. The race of voters whose ethnicity neither white nor black/african american/negro is defined by 'others' in our model. We believe that it is also worth to discuss the situation of voting among those ethnicities.
For further research, we can compare the actual election results to those we predicted above. Then it is necessary to do another data analysis or just a survey for citizens about the actual results to help get a better estimation for the election results in the future.

References

1. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=f2b380d1-bf0f-41f4-a8f3-e4c820f70fff.

2. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [2018 5-year ACS]. Minneapolis, MN: IPUMS, 2020. Retrieved from https://doi.org/10.18128/D010.V10.0.

3. Alexander (2019, Dec. 3). Rohan Alexander: Getting started with MRP. Retrieved from https://rohanalexander.com/posts/2019-12-04-getting_started_with_mrp/