Prediction of 2020 American Federal Election

Jiarong Ye

2 November, 2020

Model

Model Specifics

I will be using a logistic regression model to model the proportion of voters who will vote for Donald Trump. The reason why the logistic regression model is used because my Y variable is binary meaning either vote for Trump or not, value 1 represents vote for Trump and value 0 represents not vote for Trump. I will be using age, gender and region to model the probability of voting for Donald Trump. The logistic regression model I am using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Age} + \beta_2 x_{Male} + \beta_3 x_{Northeast} + \beta_4 x_{South} + \beta_5 x_{West} + \epsilon$$

Where p represents the probability of the people voting for Trump. Similarly, β_0 represents the intercept of the model, and is the probability of voting for Donald Trump at age 0 from region Midwest. Additionally, for everyone one unit increase in age, we expect a β_1 increase in the probability of voting for Donald Trump. For everyone one unit who is male, we expect a β_2 increase in the probability of voting for Donald Trump as well. However, β_3, β_4 and β_5 we do not know their relationship with region Midwest, further information will be explored later.

Model 1:

```
##
## Call:
## glm(formula = vote_trump ~ age + gender + census_region, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5417  -1.0914  -0.8485   1.2027   1.6336
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.162021   0.104264 -11.145   <2e-16 ***
## age                     0.014739   0.001676   8.793   <2e-16 ***
## genderMale              0.534703   0.055710   9.598   <2e-16 ***
## census_regionNortheast -0.101839   0.087428  -1.165   0.2441
## census_regionSouth      0.199605   0.075383   2.648   0.0081 **
## census_regionWest      -0.131960   0.084776  -1.557   0.1196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7511.5  on 5449  degrees of freedom
## Residual deviance: 7314.9  on 5444  degrees of freedom
## AIC: 7326.9
##
## Number of Fisher Scoring iterations: 4
```

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Here I create cells based off different ages, genders and regions. Using the model described in the previous sub-section I will estimate the proportion of voters in each age bin. I 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. The reason why I think ages and genders are variables that influence the possibility of Trump being elected is because people in different ages and genders has different ways of thinking things, so their opinion toward Trump might be different. Regions could also be a influential variable since the country is big and development in different regions of U.S. is different, so it causes people have different living styles in different regions, so their thoughts to politics would also be different.

Table 1:

```
## Rows: 640
## Columns: 5
## $ age           <dbl> 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 1...
## $ gender        <chr> "Female", "Female", "Female", "Female", "Male", "Male...
## $ census_region <chr> "Northeast", "Midwest", "South", "West", "Northeast",...
## $ n             <dbl> 137, 103, 209, 182, 127, 82, 212, 152, 135, 89, 240, ...
## $ estimate      <dbl> 0.2692244, 0.2897246, 0.3324517, 0.2633397, 0.3860709...
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.474
```

Results

Model 1 represents the logistic regression, below is how logistic regression being expressed.

$$log(\frac{p}{1-p}) = -1.1620 + 0.0147x_{Age} + 0.5347x_{Male} - 0.1018x_{Northeast} + 0.1996x_{South} - 0.1320x_{West}$$

p represents the probability of the event of interest occurring, here it refers to the probability that Donald Trump can be re-elected in 2020. The intercept, $\hat{\beta}0$ here is -1.1620. It is also a log of odds when $x_{\hat{age}}$, $x_{\hat{Male}}$, $x_{\hat{Northeast}}$, $x_{\hat{South}}$ and $x_{\hat{West}}$ are zero. Only $x_{\hat{age}}$ here is a numerical predictor variable, others are categorical predictor variables.The positive coefficient, $\hat{\beta}1$ suggests that for every additional unit increase in age we expect the log odds of Trump being re-elected increase by 0.0147 for both female and male groups from all regions across the U.S. $\hat{\beta}2$ here indicates the difference in average value between two groups, male and female is 0.5347. It implies that men are more willing to vote for Trump than women. Let's now look at coefficients for regions. $\hat{\beta}3$, $\hat{\beta}4$ and $\hat{\beta}5$ represent the difference in average value between their corresponding region and region Midwest, the differences are -0.1018, 0.1996 and -0.1320. The two negative coefficients are $\hat{\beta}3$ and $\hat{\beta}5$. It delivers a message that less people for all genders and ages would vote for Trump from region Northeast and West comparing to people from region Midwest. However, $\hat{\beta}4$ is a positive coefficient meaning Trump has more supporters in region South than in region Midwest. Now, we have to use the logistic model to do Post-Stratification by taking advantage of demographics to extrapolate how entire population will vote. The census data I used is in 2018, but now we are in 2020 so in early data cleaning process, I changed everyone's age two years older. Only citizens of U.S who are 18 or over 18 years old are able to vote, thus I deleted all people that do not own a American citizenship and people who are not reach 18 years old are also deleted from the census data. Since variables for both survey and census have to be same so they can be used for post-stratification. However, the categorical variable, region from my census data, it is more specific than my survey data. Therefore, I assigned them into only four different regions in order to keep them same as my survey data. For example, according to the geographic map of U.S. I combined east north central div and west north central div into Midwest. After the census data is rearranged, the table 1 below Post-Stratification is made. The table shows that it partition the census data into 640 demographic cells. Each cell is a demographic to a unique combination of gender, region and age. Each cell also includes the estimate response variable, the probability of Donald Trump being re-elected that is calculated by the logistic model introduced previously. How could we get the final prediction by Post-Stratification? We first use estimate of each cell times its corresponding population size and then take the sum of the value from every single one of the cell, lastly divide it by the entire population size. Here $\hat{y}^{ps} = 0.4740$. The prediction of probability of Donald Trump being re-elected in entire population size is 47.40 %.

Discussion

Summary

I did data cleaning process for both survey and census data. I then created logistic regression model by using the survey data. The model is used to see the how could variables affect the probility of Trump being re-elected. Post-Stratification is started, I divided the census data into many demographic cell and each cell contains estimate response variable calculated based on the logistic regression model. Lastly, after some calculation the probability of Trump being re-elected in the entire population size is predicted.

Conclusion

From the results, we have found older people are more willing to vote for Trump than younger people. This is because four years ago Trump claimed he will make America great again, but today there are still many old people would believe him because of old people's nostalgia. Some young people especially college students, they do not pay attention to politics, some of them do not even know the voting process and they just abstain from voting. Young people cares about future, however just 30 percent of millennial voters (ages 24 to 39) say they feel confident that their children generation will be better than theirs has been. Young people also really cares about economy, they think that the economy has been hit hard due to Trump's trade war, they have already lost trust on Trump. Trump is tend to be more appeal among men. According to a interview, most of the Mexican-American men think that the macho allure of Mr. Trump is undeniable. He is forceful, wealthy and, most important, unapologetic. In a world where at any moment someone might be attacked for saying the wrong thing, he says the wrong thing all the time and does not bother with self-flagellation. However, most of the women think Trump is too aggressive, women dislike his behaviors and manners. Safety is something women really care. However, Trump said he is going to cut $25 billion from Social Security. Trump has many supporters from the South region since in 2016 Trump addresses issues in that have long resonated in the South: criticism of the federal government, attacks on illegal immigration, protests of foreign trade deals and pledges to bolster the military. People from West region do not really satisfied with Trump, since the 2020 wildfire in west coast has destroyed a total of 4,800 structures, including 1,145 homes and also cause serious air pollution. However, Trump barely paid attention and blamed forest management, it makes Western people disappointed.

Weaknesses & Next Steps

The census data is not up to date and it is from 2018. Many things could be happened between year 2018 and 2020. For example, some people might past away so the population size cannot represent for today. During the Post-Stratification process, we assume everyone in census data is going to vote for the election. However, in fact many people would choose to abstain from voting so lead to inaccuracy of population size. The categorical variable, regions in census data and survey data have different values. I have to make them same in order to do Post-Stratification. However, the revised values could not 100% express the original values. The next step is to wait for the real election outcome to happen, and compare the real probability to my predicted probability. If my prediction is close to the actual outcome then it means I am using the correct method and knowledge from this course. If the outcome is quite different from my prediction, then I have to read some relative expert analysis and reports and reflect on my work to find what were the problems affecting my prediction and how should I avoid them next time.

References

Miao, Hannah. Young Voters Are Poised to Be a Decisive Factor Even as Coronavirus Creates Obstacles. 27 Oct. 2020, www.cnbc.com/2020/10/26/2020-election-young-voters-could-be-a-decisive-force-despite-coronavirus.html.

Ball, Molly. "Trump's Graying Army." The Atlantic, Atlantic Media Company, 25 Oct. 2016, www.theatlantic.com/politics/archive/2016/10/trumps-graying-army/505274/.

Shah, D. (2020, September 17). US election: Are older Trump voters sticking with him? Retrieved November 02, 2020, from https://www.bbc.com/news/election-us-2020-54099242