2020 American federal election

Shuk Yin Chung (1003384964)

November 2, 2020

2020 American federal election

Shuk Yin Chung

November 2, 2020

Model

The aim of the model is to predict the vote outcome of the 2020 American federal election. By performing the post-stratification technique, I will describe and analyze the model specifics and the post-stratification calculation in the following.

Model Specifics

I decide to use a logistic regression model to model the voting result of 2020 American federal election. It aims to predicts the overall votes for potential candidates, who are Donald Trump and Joe Biden. I choose this model because the response variable is set to binary. It would be helpful to determine voters' intention for who they would vote for. Also, it would demonstrate the supporting rate of each candidate. The logistic regression model has only predictor variables, which can be either categorical or numerical.

Then, I apply this model to my dataset. I choose sex and region as predictor variables to model the probability of voting for Donald Trump and Joe Biden. It is because I want to investigate the vote intention of male and female. Also, I choose region because some provinces have more votes than the others. So, it might affect the result of election. Here is my model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{region}$$

However, sex and region are categorical variables. I will use dummy variable coding to change the responses to binary. Meanwhile, I add new variables to represent the existing variables. I use a new variable x_{male} to replace x_{sex}, and variables of $x_{region_{NE}}$,x_{region_S}, and x_{region_W} to represent Northeastern, Southern, Western regions. This is the updated version of my model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{male} + \beta_2 x_{region_{NE}} + \beta_3 x_{region_S} + \beta_4 x_{region_W}$$

Where p refers to the probability of voting for either Donald Trump or Joe Biden. β_0 is the intercept of the model, which is the mean for all response variables when predictor variables are on zeros. Besides, β_1,β_2,..., β_4 are the coefficients of x_1,x_2,...,x_4 respectively. They are the change in log odds. In General, the model means when the values of every predictors increase, their coefficients also increase. Therefore, the probability of voting for either Donald Trump or Joe Biden also increases.

Post-Stratification

I will perform a post-stratification calculation and analysis. It would be helpful to estimate the overall population of voting for either Donald Trump or Joe Biden during the 2020 American federal election. Then, I am able to estimate the result of election. I make cells based off sex and region as I have also chosen these two predictors from the survey data. I select sex because male and female might have different political viewpoints since their concerns could be different. I think region is an important factor that determines the probability. It is because the sizes of voters are different in each region and some provinces have larger and more influential votes. I will estimate the voting in each of Northeastern, Southern, and Western region, and male. I will also weight each estimate with regard to each bin. It is calculated by the sum of each bin's population and divided by the entire population.

Results

Firstly, I will analyze trump's model and his probability of getting voted.

Here is the summary of trump's logistic regression model:

```
##
## Call:
## glm(formula = vote_trump ~ sex + region, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1562  -1.0267  -0.8373   1.3062   1.5916
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.75316    0.06270 -12.013   <2e-16 ***
## sexMale          0.57017    0.05195  10.976   <2e-16 ***
## regionNortheast -0.11470    0.08187  -1.401   0.1612
## regionSouth      0.13291    0.07020   1.893   0.0583 .
## regionWest      -0.18238    0.07940  -2.297   0.0216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8619.4  on 6474  degrees of freedom
## Residual deviance: 8480.1  on 6470  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 8490.1
##
## Number of Fisher Scoring iterations: 4
```

Plugging in the estimated coefficients to trump's model:

$$log(\frac{p}{1-p}) = -0.753 + 0.570x_{male} - 0.115x_{region_{NE}} + 0.133x_{region_S} - 0.182x_{region_W}$$

Then, I will calculate the post-stratification by using this formula:

$$\hat{y}^{ps} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j}$$

The value is the estimate of each x variable for respective cells: $\hat{y}^{male}$ = 0.7816403 $\hat{y}^{region_{NE}}$ = 2.175856 $\hat{y}^{region_S}$ = 1.026215 $\hat{y}^{region_W}$ = 1.621484

Then, substitute the estimates into trump's model, we get: $log(\frac{p}{1-p})$ = -0.753 + 0.570*(0.7816403)* - *0.115*(2.175856) + 0.133*(1.026215)* - *0.182*(1.621484)

Calculate p: $log(\frac{p}{1-p})$ = -0.716312 $\frac{p}{1-p}$ = 10^(-0.716312) p = 0.161

Therefore, we estimate that the proportion of voters in favour of voting for Donald Trump to be 0.161. This is based off our post-stratification analysis of the proportion of voters in favor of Donald Trump modeled by a logistic regression model, which accounted for sex and region.

Next, I will analyze Biden's model and his probability of getting voted.

Here is the summary of biden's logistic regression model:

```
##
## Call:
## glm(formula = vote_biden ~ sex + region, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1540  -1.0721  -0.9258   1.2865   1.4519
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.11946    0.06041  -1.978   0.0480 *
## sexMale         -0.37256    0.05084  -7.328 2.34e-13 ***
## regionNortheast  0.05863    0.07967   0.736   0.4618
## regionSouth     -0.13336    0.06925  -1.926   0.0541 .
## regionWest       0.06417    0.07679   0.836   0.4034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8809.5  on 6474  degrees of freedom
## Residual deviance: 8746.9  on 6470  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 8756.9
##
## Number of Fisher Scoring iterations: 4
```

Plugging in the estimated coefficients to Biden's model:

$$log(\frac{p}{1-p}) = -0.119 - 0.373x_{male} + 0.0586x_{region_{NE}} - 0.133x_{region_S} + 0.0642x_{region_W}$$

Then, I will calculate the post-stratification.

Here is the estimate of each x variable for respective cells: $\hat{y}^{male}$ = 0.8565902 $\hat{y}^{region_{NE}}$ = 2.384495 $\hat{y}^{region_S}$ = 1.124616 $\hat{y}^{region_W}$ = 1.776965

Then, substitute the estimates into biden's model, we get: $log(\frac{p}{1-p})$ = -0.119 - 0.373*(0.8565902)* + *0.0586*(2.384495) - 0.133*(1.124616)* + *0.0642*(1.776965)

Calculate p: $log(\frac{p}{1-p})$ = -0.3342695 $\frac{p}{1-p}$ = 10^(-0.3342695) p = 0.317

Therefore, we estimate that the proportion of voters in favour of voting for Joe Biden to be 0.317. This is based off our post-stratification analysis of the proportion of voters in favour of Joe Biden modeled by a logistic regression model, which accounted for sex and region.

Discussion

Summary: In this report, I have used a logistic regression model to predict the probability of voters to vote for Donald Trump or Joe Biden. I have also performed a post-stratification calculation and analysis to estimate the entire population of voters and the result of 2020 American federal election.

Conclusion: Based off the estimated proportion of voters in favor of voting for Donald Trump being 0.161 and Joe Biden being 0.317, we predict that Joe Biden will win the election.

Weaknesses

In this report, I attempted to create graphs but it seems irrational and make the report exceed the page limit. It is irrational because sex and region are not numeric variables. Another weakness is that I did not split data into groups, so it is quite messy when processing a dataset with a large population size.

Next Steps

The subsequent work could be comparing this report to the actual result of 2020 American federal election, by using the Bayesian logistic regression model. A follow-up survey could ask respondents whether they are satisfied with the result of election. Then, we can do a subsequent study on voters' intention vs voters' satisfaction of the election.

References

1. Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-15.
2. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=1623c29b-7c65-4d73-b402-5c2f34ab9497 (https://www.voterstudygroup.org/downloads?key=1623c29b-7c65-4d73-b402-5c2f34ab9497).
3. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [usa_00002.dat]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0 (https://doi.org/10.18128/D010.V10.0)
4. "Codebook for an IPUMS-USA Data Extract: DDI 2.5 metadata describing the extract file 'usa_00002.dat' (ddi2-184825_usa_00002.dat-usa.ipums.org)". Minnesota Population Center. University of Minnesota, October 31, 2020.
5. "User Extract usa_00002.dat". Minnesota Population Center. University of Minnesota, October 31, 2020.