Prediction of proportion of voters will vote Joe Biden in 2020 U.S.Election

Shihan Wang 1005165063 Wenyu Shu 1004951082 Kefan Cai 1004819949

Nov.1st, 2020

Model

In order to predict the outcome of the US. election 2020, which will occur on 3 November to see who will be the next President of the United States(The Visual and Data Journalism Team, 2020). A multiple logistic regression model and the post-stratification technique would be a good choice to predict the outcome. This will be elaborated in the following sub-sections.

Model Specifics

Multiple logistic Regression Model is used to model a binary response variable(y), based on predictors(xi). To be more specific, p is denoted as the case when $y = 1$, represents the probability of an event occurring, which means vote Binden in this case. $Beta0$ is considered as an intercept(constant) while $Beta1$ to $Beta4$ represents the coefficient correspond to each predictor x. The left side of the equation is called the logit function.

Mathematic notation:

$$log(p/1-p) = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{race} + \beta_3 x_{sex} + \beta_4 x_{employment}$$

Multiple logistic regression model can be considered as an appropriate model to predict 2020 US. election outcomes since y are expected as dichotomous. In order to adopt this model, four predictors including age groups, race, sex, and employment were chosen. As the notation provided, $Beta1$ meaning one unit increase in the age group will lead to $Beta1$ increases in the probability of voting Biden.(same logic for $Beta2$, $Beta3$ and $Beat4$).

##model 1

```
## # A tibble: 10 x 5
##    term                                estimate std.error statistic
 p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>
   <dbl>
##  1 (Intercept)                           -0.819     0.235     -3.48
5.03e- 4
##  2 sexmale                               -0.287     0.0521    -5.51
```

```
3.52e- 8
##  3 raceblack/african american/negro      1.60      0.246      6.50
8.14e-11
##  4 racechinese                           1.23      0.322      3.80
1.45e- 4
##  5 racejapanese                          1.73      0.542      3.19
1.45e- 3
##  6 raceother asian or pacific islander   0.847     0.268      3.16
1.59e- 3
##  7 raceother races or mixed-blood        0.763     0.252      3.03
2.45e- 3
##  8 racewhite                             0.399     0.236      1.69
9.09e- 2
##  9 employmentnot in labor force          0.159     0.0584     2.73
6.39e- 3
## 10 employmentunemployed                 -0.0674    0.0864    -0.780
4.35e- 1
```

##AIC for model 1

```
## [1] 8510.657
```

##model 2

```
## # A tibble: 12 x 5
##    term                              estimate std.error statistic
 p.value
##    <chr>                                <dbl>     <dbl>     <dbl>
   <dbl>
##  1 (Intercept)                         -0.826     0.236     -3.51
4.53e- 4
##  2 age_groupsenior                      0.142     0.0707     2.00
4.50e- 2
##  3 age_groupyouth                       0.0651    0.0842     0.772
4.40e- 1
##  4 raceblack/african american/negro     1.59      0.246      6.45
1.14e-10
##  5 racechinese                          1.21      0.322      3.76
1.69e- 4
##  6 racejapanese                         1.70      0.542      3.13
1.73e- 3
##  7 raceother asian or pacific islander  0.841     0.268      3.13
1.72e- 3
##  8 raceother races or mixed-blood       0.757     0.252      3.01
2.65e- 3
##  9 racewhite                            0.379     0.236      1.61
1.08e- 1
## 10 sexmale                             -0.285     0.0522    -5.47
4.56e- 8
## 11 employmentnot in labor force         0.0921    0.0672     1.37
1.70e- 1
```

```
## 12 employmentunemployed                  -0.0723    0.0865    -0.836
4.03e- 1
```

##AIC for model 2

```
## [1] 8510.562
```

##Anova

```
## Analysis of Deviance Table
##
## Model 1: vote_Biden ~ sex + race + employment
## Model 2: vote_Biden ~ age_group + race + sex + employment
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      6457     8490.7
## 2      6455     8486.6  2   4.0943   0.1291
```

Based on the AIC and P-value, model 2 is selected as a better model(will be elaborate in the result section).

Post-Stratification

Since the objective is to predict the ratios of the population who will vote for Biden in the U.S. 2020 Election, a post-stratification technique is a good choice. The model2 described in the above model section includes four predictors, it will be explained in detail:

Firstly, in terms of the data age, people who are under 18 were removed since they have no right to vote, then the remaining data were classified into three groups: youth(18~24), adult(25~59), and senior(older than 60). This is because it will make the data tidier. A new data named age group was created and has three different cells.

Then, data race was recategorized into seven diverse cells including: white, black/african american/negro,american indian or alaska native,other races or mixed-blood,chinese,japanese,other asian or pacific islander. Due to the large population in each group, this categorization will make the results more convincing. Choosing this data is because it is likely to influence the voter outcome.

For the remaining two data: sex and employment, they were classified into female and male, employed, unemployed, and not in labor force respectively. Therefore each of these them contains two cells and three cells.

The last step should be group by the above cells, meaning multiply cells described above($3 * 7 * 2 * 3 = 126$) to get a total of 126 cells. Then by applying the model2, the ratios of voters in each cell can be estimated. By sum these estimated value with the Corresponding population of each cell, and then divided by the whole population size. The final proportion of the voters who prefer voting Biden can be calculated.

Mathematic notation:

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

where $\sum N_j$ represents the population size of j^{th} cell, $\hat{y}_j$ means the estimate in each cell that were constructed and $\hat{y}^{PS}$ represents the estimate y based on post-stratification.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.421
```

Results

The logistic regression model estimates that the probability of voters chooses to vote for Joe Biden in the 2020 American Presidential Election,

$$\hat{y}^{PS}$$

, is 0.421. This model uses the post-stratification method to calculate the proportion of the voters who prefer voting for Joe Biden based on their sex, age groups, race, and employment status from the survey data of the Nationscape Data Set and the census data of the American Community Surveys.

The model checking of comparing AIC and using ANOVA test both shows the result that the logistic regression model (model2), with 4 variables (sex, age groups, race, employment), is better than the logistic regression model (model1), which contains 3 variables (sex, race, employment). Specifically, the AIC value of model2 (8510.6) is smaller than model1 (8510.7), and the ANOVA p-value of model1 and model2 is larger than 0.05, which also corroborates model2 is better than model1.

Discussion

In this analysis, the multiple logistic regression model and the post-stratification technique are used to predict the vote of the 2020 American federal election.

Firstly, the variables of sex, age, race, and employment status have been chosen to be the four predictor variables to build the logistic regression model. Next, to improve the model, post-stratification is used here. The first step of post-stratification is removing the data of people who don't have voting power to get new clean data. Then the new data can be selected as three groups: youth(18~24), adult(25~59), and senior(older than 60). Furthermore, race type is divided in seven types as white, black/african american/negro, american indian or alaska native, other races or mixed-blood, chinese, japanese, other asian or pacific islander.

Gender types are female and male and employment status is devided as employed, unemployed, and not in the labor force.

Therefore, 126 cells can be created by the steps above to fit the model2 and calculate the final proportion by the formula. The result is 0.421, it means that there are 42.1% of people who prefer voting for Joe Biden. So that 42.1% of people will vote for Joe Biden can be estimated from the analysis.

Weaknesses

One weakness is that reliability maybe not strong enough to predict the result. Since the data is for 2018, some changes including environmental change and personal change may happen in 2 years. For the environment change, some factors such as migration, change in age, death, and so on can cause a change in the number of citizens and their status of the voting power. For personal reasons, people's minds may change now and take a different choice from 2018.

Next Steps

Since the analysis still has some limitations, some further steps can be taken to improve the model. For example, the data can be updated by a newer one since 2018 is two years ago. Moreover, in this analysis, only four variables have been taken but actually, there are many other factors that would affect the result. So a model can include more variables that can be better in predicting the result.

References:

R & R packages:

```
##
## To cite package 'dplyr' in publications use:
##
##   Hadley Wickham, Romain François, Lionel Henry and Kirill Müller
##   (2020). dplyr: A Grammar of Data Manipulation. R package version
##   1.0.2. https://CRAN.R-project.org/package=dplyr
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {dplyr: A Grammar of Data Manipulation},
##     author = {Hadley Wickham and Romain François and Lionel {
##               Henry} and Kirill Müller},
##     year = {2020},
##     note = {R package version 1.0.2},
##     url = {https://CRAN.R-project.org/package=dplyr},
##   }
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and
 Winston Chang and Lucy D'Agostino McGowan and Romain François and Garr
ett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Ku
hn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and
 Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel
 and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wil
ke and Kara Woo and Hiroaki Yutani},
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }

##
## To cite package 'haven' in publications use:
##
##   Hadley Wickham and Evan Miller (2020). haven: Import and Export
##   'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1.
##   https://CRAN.R-project.org/package=haven
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {haven: Import and Export 'SPSS', 'Stata' and 'SAS' File
s},
##     author = {Hadley Wickham and Evan Miller},
##     year = {2020},
##     note = {R package version 2.3.1},
##     url = {https://CRAN.R-project.org/package=haven},
##   }

##
## To cite package 'broom' in publications use:
##
##   David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert
##   Statistical Objects into Tidy Tibbles. R package version 0.7.0.
##   https://CRAN.R-project.org/package=broom
##
## A BibTeX entry for LaTeX users is
##
```

```
##   @Manual{,
##     title = {broom: Convert Statistical Objects into Tidy Tibbles},
##     author = {David Robinson and Alex Hayes and Simon Couch},
##     year = {2020},
##     note = {R package version 0.7.0},
##     url = {https://CRAN.R-project.org/package=broom},
##   }

##
## To cite R in publications use:
##
##   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/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computin
g},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2019},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please cite
 it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

#survey data: Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

#census data: 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

#Data cleaning code: Phillips, N. (2018, January 22). YaRrr! The Pirate's Guide to R. Retrieved November 02, 2020, from https://bookdown.org/ndphillips/YaRrr/dataframe-column-names.html

#Model: The Visual and Data Journalism Team. (2020, November 01). US election 2020 polls: Who is ahead - Trump or Biden? Retrieved November 02, 2020, from https://www.bbc.com/news/election-us-2020-53657174