Investigation on Predicting the Results of the 2020 American Federal Election

Ying Tian (1005174240), Zhaowei Yao (1005333355)
Yuqing Wu (1004725737), Baoying Xuan (1004808149)

Nov 2nd, 2020

Model

In this report, we are interested in predicting the 2020 American federal election's popular vote outcome. 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

The model chosen here is the logistic regression model, where the response variable is "vote_2020_biden", and the predictor variables selected for analysis are "age", "sex"(a binary variable - equals 1 for a male respondent, and 0 for a female), and "hispan"(a categorical variable indicates the Hispanic origin of the respondent). The dependent variable "vote_2020_biden", is a binary variable, which equals 1 if the respondent wants to vote for Joe Biden in the 2020 election, and equals 0 for Donald Trump. It would be appropriate to use the logistic regression model, since it is a predictive analysis used to describe data and explain the relationship between one dependent binary variable and nominal, ordinal, interval or ratio-level independent variables *(What is Logistic Regression? 2020)*.

The notation of the logistic regression model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{hispan} + \epsilon$$

Where $log(\frac{p}{1-p})$ represents the proportion of voters who will vote for Joe Biden. β_0 represents the intercept of the model, and is the probability of voting for Joe Biden at age 0. Additionally, β_1, β_2, and β_3 represent the parameters related to the predictors - "age", "sex" and "hispan". So, for every one unit increase in age(holding other variables unchanged), the expected log odds in favor of voting for Joe Biden increase by β_1. And β_2 means that the expected log odds in favor of voting for Joe Biden in males are higher than in the female group by β_2 on average.

Post-Stratification

Post-stratification is useful because while selecting similar units into one group can be generally viewed as not only a tool to reduce the variance of the survey estimates obtained, but also a method to decrease the bias due to non-response and underrepresented groups in the population. We chose "age" as a variable because different respondents' ages will likely to influence their voting decisions. We chose the second indicator "sex" because gender may also affect the voting outcome. Lastly, we included "hispan" because the Hispanic voters have made up increasingly larger shares of the electorate in every state.

To filter voters based on the election policy in America, we used "citizen" and "age", since voters should be a U.S. citizen and 18 years old on or before election day. Then we conducted a post-stratification analysis to predict the probability of Biden winning the 2020 election. Here we created cells based on different ages/genders/Hispanic groups. The proportions of voters in each bin were estimated using the model described in the Model Specifics section. The weights in each bin were then estimated based on the population size. We divided the summed up values of the weights timing the relevant bin size, the final post-stratification prediction was then calculated by dividing the sum by the entire census population size.

Additional Information

P-values and standard error in this model would work well together, which provides us with information - whether the relationships are statistically significant. Assuming the confidence level is 95%, if the p-value for the dependent value is less than its significance level, then there is enough evidence to reject the null hypothesis for the entire population *(Frost et al., 2020)*.

Standard errors refer to beta values, which can determine if the value is significantly different from zero by evaluating the t – statistic value. The smaller the standard errors are, the more precise the estimations are conducted.

Results

Table 1 - Summary of the Logistic Model

```
##
## Call:
## glm(formula = vote_2020_biden ~ age + sex + hispan, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5689  -1.1291  -0.8936   1.1509   1.6695
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.130253   0.404139  -0.322  0.74723
## age                -0.008431   0.001697  -4.968 6.75e-07 ***
## sexmale            -0.489219   0.055266  -8.852  < 2e-16 ***
## hispanmexican       1.167401   0.408383   2.859  0.00426 **
## hispannot hispanic  0.715555   0.398178   1.797  0.07232 .
## hispanother         0.920546   0.417628   2.204  0.02751 *
## hispanpuerto rican  1.561072   1.293371   1.207  0.22744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7555.3  on 5449  degrees of freedom
## Residual deviance: 7400.0  on 5443  degrees of freedom
## AIC: 7414
##
## Number of Fisher Scoring iterations: 4
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1   0.6199227
```

$$\hat{y}^{PS}Biden = 0.6199227$$

According to the p-value calculated by the fitted logistic regression model for each parameter, it can be seen that the response variable has a significant statistical association with the age, sex, and Hispanic origin of the respondent. Among the categorical variables of "hispan", only *hispanmexican* and *hispanother* has a significant relationship with the response variable. Derived from our post-stratification analysis based on the logistic regression model, we estimate that voters' proportion in favor of voting for Joe Biden to be 0.6199(i.e. 61.99%).

Discussion

Summary

This report analyzes and estimates the popular vote outcome of the 2020 American federal election based on the 2018 5-year American Community Survey data and 2020 Census data. Methods of analysis include the logistic regression model and post-stratification. We selected three variables, including age, sex, and Hispan origin, and determined whether the association between the response (whether to vote for Joe Biden) and the term is statistically significant. Following that, we predicted the proportion of voters who chose Joe Biden based on this model.

Conclusion

Based on the estimated proportion of voters in favor of voting for Joe Biden being 61.99%, we predict that Joe Biden will win the election.

Weaknesses

In general, there are a variety of weaknesses and limitations existing in our investigation. Firstly, in terms of methodological weaknesses, our model uses data collected in 2018 to predict the voting results for the 2020 election, which is based on past results. However, situations are changing every year, even every moment. Also, as people get elder, they are likely to have a different understandings of whom they want to be their president. Therefore, the time difference between the two sets of data collection will produce a particular bias. Moreover, the survey data could not cover all the census data in our model. More specifically, the two sets of data variables cannot be matched entirely, indicating that the survey data is not enough to represent and precisely predict the voting results fully.

Next Steps

Political problems like the federal election are inextricably related to many other external factors, such as the social rules that the candidates promote, economic conditions, etc. Thus, there is much to desire in its completion method. First of all, the model that we use for prediction is the logistic model, one of the simplest ways to analyze when the data sets are linearly separable. However, the election problem is too complicated so that the simple logistic model could not accurately and precisely represents. Thus, it is necessary to use more advanced models to do further analysis. Also, as mentioned in the limitation session, various variables from the census data could not be found in the survey data. Thus, it is important to collect more data to fulfill different variables to make a more precise prediction. Lastly, it would be better to compare with the actual election results and do a post-hoc analysis of how to better improve the future estimation.

References

Frost, J., Di, Zubayda, Aleeha, Hashmi, M., Rasentsoere, K., . . . Toby. (2020, July 16). How to Interpret P-values and Coefficients in Regression Analysis. Retrieved November 02, 2020, from https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved November 02, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 02, 2020, from https://usa.ipums.org/usa/index.shtml

The Changing Racial and Ethnic Composition of the U.S. Electorate. (2020, October 22). Retrieved from https://www.pewresearch.org/2020/09/23/the-changing-racial-and-ethnic-composition-of-the-u-s-electorate/

Wickham, H. (n.d.). Welcome to the {tidyverse}. Journal of Open Source Software, 4, 1686.

What is Logistic Regression? (2020, March 09). Retrieved November 02, 2020, from https://www.statisticssolutions.com/what-is-logistic-regression/

Who Can and Can't Vote in U.S. Elections. (n.d.). Retrieved from https://www.usa.gov/who-can-vote

6.3 - Poststratification and further topics on stratification: STAT 506. (n.d.). Retrieved from https://online.stat.psu.edu/stat506/lesson/6/6.3

Appendix

Github Repo URL: https://github.com/yaozhaow/STA304-ProblemSet3-Group20