A prediction on the overvall

Yuehao Huang, Lingyi Li, Suran Wu, Siyu Chen

Nov.2. 2020

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (For citation please see reference). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

We will be using a logistic regression model to model the proportion of voters who will vote for Donald Trump in 2020. The reason we use this model is because we are interested in whether Donald Trump will be elected as president in 2020, by cleaning the data and adding a variable with response only containing 1 and 0, indicating whether people will vote for him or not, it is perfect to use logistic regression with binomial family for our prediction. Our model has two predictors. They are: gender and the employment status(note we combined all employed and unemployed categories and dropped the "N/A" and "other" response) which is recorded as a categorical variable, to model the probability of people voting for Donald Trump or not. The logistic regression model we are using is:

$$y = \beta_0 + \beta_1 x_{male} + \beta_2 x_{unemployed} + \epsilon$$

Where $\beta_0 = -0.58924$, $\beta_1 = 0.5382$, $\beta_2 = -0.38877$.

y represents the log proportion value of voters who will vote for Donald Trump vs. against him. β_0 represents the intercept of the model, meaning the probability of a employed female would have a log value of -0.58924 for in the log proportion possibility value of voting for Trump against not voting for him. Additionally, β_1,represents the slope of the model for male. For example, for any male, we would expect a 0.5382 increase in the log proportional probability value. Finally, β_2 is the slope for unemployed people. For instance, one who is unemployed would have a -0.38877 change in numbers in the log proportional probability (Please see additional information section below).

Post-Stratification

Post-stratification is useful for correction of non-probability based sampling. After getting lots of demographic data, we may categorize them by the desired variable and split them into different factor cells. Such categorizing can be performed many times to create many cells. Then, we may calculate response estimates for each cell, weight them by its relative proportion to whole data and sum them up to get a response estimate for the entire data. This way we could get a more accurate prediction based on the data we have.

In our sample data we choose gender because a study(See reference) suggests men are more libreal and this likelihood may impact the election. Employment factor is chosen for an obvious reason, the working class and the unemployed would clearly have different interests and points of view of the government. For example, the unemployed may vote for candidates who have good plans for unemployment benefits while the working class will support the candidates who implement income tax cuts. Hence we have 4 different cells - Male employed, Female employed, Female unemployed, and Male unemployed. after we have done the cleaning process on post stratification data.

Additional Information

We are using the logistic regression model here, meaning the model will give us a log ratio of $log(p/1-p)$ on how people would vote for and against Trump in this 2020 election instead of the percentage of people voting for Trump. However, we can do some calculation to get this value, simply do exponential on both sides and solve for the unknown probability p, which will give us the percentage of people voting for Trump(by prediction.)

We initially set out 4 factors that could potentially impact voters' choices as our explanatory variables. They are race, age,gender and employment status. After simulating the multiple logistic regression models, we drop factors of age and race due to high p-value as it suggests our sample does not have strong evidence against null hypothesis which is such factors may have little impact on our response outcome. Also, when we add 'not in labour force' factor in our model, it gives us a high p-value in the summary table, and the AIC score reduced 2000 when we removed 'not in labour force' which means the model without factor 'not in labour force' is better. Therefore we abandoned ages , factor 'not in the labour force' and kept gender and other employment status.

Results

Table.1

term	estimate	std.error	statistic	p.value
(Intercept)	-0.5892359	0.0507128	-11.619069	0.0000000
genderMale	0.5382029	0.0668213	8.054362	0.0000000
employmentunemployed	-0.3887700	0.1068636	-3.638001	0.0002748

We estimate that the log proportion of voters in favor of voting for Donald Trump and against him to be -0.329 after applying post stratification. Then we take exponential on both sides since it is a log proportion, what we have is 0.719 for the proportion of vote for vs. against Trump. For example, Say there are 172 people, about 72 people would vote for Trump and 100 would not. Hence we can interpret the percentage of people who would vote Trump based on the post stratification data will be 72/172 = 0.42, meaning About 42% of people would vote for Trump, which accounted for employment status and genders, modeled by our Logistic regression model.

Discussion

Summary

In order to predict the outcome of the 2020 American federal election, the multiple logistic regression model is used to predict whether Trump will be able to retain his presidential position. So our response variable will be whether people will vote for Trump, using dummy variable 1 as vote for Trump and 0 otherwise. For explanatory variables, only gender and employment status are used and as shown in Table.1, every variable in the model are significant. After the regression model is built, we use post-stratification to partition the data into different cells based on employment status and genders factors. We may obtain our expected outcome for each cell. Weight each cell by its relative proportion to population and sum them up to get the estimate for the entire data.

Conclusion

By our regression model and test by post-stratification technique, we expect 42% of voters are in favor of Trump. And by current stage, only the remaining 2 candidates(Trump and Biden) are running for the position. So we would expect Trump to lose the election.

Weaknesses

The weakness of analysis should be recognized. First of all, we only include two explanatory variables which are gender and employment status. While they both are good indicators of whether people will vote for Trump this year, there is still room to enhance the accuracy of our model by including more relevant factors. Furthermore, we could divide our data into more cells by choosing more vote-related factors. This could increase accuracy of our prediction after we have a multiple logistic model.

Next Steps

Future steps would include examining accuracy by comparing the actual outcome and our prediction. A wrong prediction suggests room for building a better model and post-stratification, and a right prediction does not necessarily guarantee a perfect model and variables chosen for splitting cells. A post-hoc analysis could be performed by simulating the model by different exploratory variables and different variables chosen for partition the data. If we get a closer value, that is, close to 1 if Trump actually wins or close to 0 otherwise, we may conclude such a variable is useful to predict election outcome and may be used for future estimation.

References

Goodyear-Grant, E., Associate Professor, & Bittner, A., Associate Professor. (2020, July 13). How sex and gender influence how we vote. Retrieved October 31, 2020, Retrieved from https://theconversation.com/how-sex-and-gender-influence-how-we-vote-106676

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [American Community Survey 2014-2018 5-year sample]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, June 25 - July 01, 2020 (version ns20200625). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra