Prediction on 2020 American Election

Ziqi Gao 1003051092, Shidong Gui 1003592506, Cheng Qian 1004484569, Kaixi Zhang 1005059268

Nov.2nd, 2020

Prediction on 2020 American Election

Ziqi Gao 1003051092, Shidong Gui 1003592506, Cheng Qian 1004484569, Kaixi Zhang 1005059268

Nov.2nd, 2020

Model

Here we are interested in predicting the voting outcome of the 2020 American Federal Election. In order to do so, we decided to build a multilevel logistic model and employ the post-stratification technique with the previously mentioned models. We obtained our data from the Democracy Fund + UCLA Nationscape[1]. The reason we chose this method is that logistic regression, like all regression models, is a predictive analysis. In addition, logistic regression is used to describe data and to explain the relationship between a binary dependent variable and one or more independent variables.[2] However, in our model analysis, we need to not only consider the individual influences but also the group factor, which leads us to the multilevel logistic model method. In the meantime, the technique of post-stratification allows us to separate the data so that we can see the pattern. And because of the characteristics of the US election system, this technique helps us to predict the result more precisely. In the following subsections, we will discuss the details of our model and the application of post-stratification techniques.

Model Specifics

As mentioned above, we construct a multilevel logistic regression model using the following equations:

$$log\frac{p}{1-p} = \alpha_j + \beta_1 \hat{age} + \beta_2 \hat{gender} + \beta_3 \hat{education} + \beta_4 \hat{household income} + \epsilon$$

Where p represents the proportion of voters who will vote for Donald Trump. βs represent the slope of our dependent variables. α_j, however, represents the formula of our level 2 variable, presented by the following equation. ϵ is the error term for this estimation.

$$\alpha_j = \gamma_0 + \eta_1 Race + \mu$$

In this equation, γ is the intercept. η , which has a similar function with beta, acts as the slope for our dependent variable and μ is the error term of this estimation.

We use the vote intention variable as our response variable(represented by the left side of the first equation), and age, education, gender, and household income as our explanatory and level 1 (individual) variables (represented by the right of the first equation). Our level 2 variable is race, which is a group variable(represented

by the second equation). The ultimate goal is for our model to be able to answer the question like ' how does the probability for Trump to win the election change every additional level a person is educated' or 'does factors of age, gender, or income have influences on the probability of voting Trump'. In addition, the model can be seen as the training model for the testing model in the post-stratification which we will discuss in the next section.

Post-Stratification

Post-stratification allows us to adjust the weights so that the totals in each group are equal to the known population totals.[3] In other words, it increases the precision of our final prediction. According to the US election system, 48 out of the 50 States use a system called "Electoral College", in which each state has a set number of electoral votes, (For example, California has 55 electroal votes). The candidate who has the most individual votes in one state gets all electoral votes of that state, and the candidate who has the most electoral votes would become the winner of the election.[4] Due to each states have different numbers of electoral votes, we wish to use the post-stratification to adjust the weight between each state so that the totals in each group equal to the population. By doing so, we use individual estimations that we got from our regression models in the previous section, sorted them to the state that they belong to and computed a post-stratification weight for each state based on the American Community Survey[5].

Results

Using age, sex, education household income as independent variables, and race to be the random intercept, we can get the following equation for the logistic model for Trump:

$$log\frac{p}{1-p} = \alpha_0 + a_j + \beta_0 x_{age} + \beta_1 x_{gender} + \beta_2 x_{education_level} + \beta_3 x_{household_income} + \epsilon$$

The random intercept is given by different race, and the value of aj follows: 'american indian or alaska native' | 'black/african american/negro' |'chinese' | 'japanese' | 'other asian or pacific islander' | 'other race, nec' |'white' ——————| ————-| ————-| ————-| ——————-|————-|————- -1.255717 | -3.266396| -2.484458 | -2.105658 | -1.818034 | -1.989731| -1.183942 In this model, age and gender are significant predictors, as their p-values are extremely small, while education and most household income levels are not significant predictors in this model. The AIC and BIC are quite large for the model, with values equal to 5356.4 and 5388.1, respectively. This suggests that this model might not be very predictable. Using the same independent variables, a logistic model for Biden can also be obtained, just like the equation above. The random effect is as following 'american indian or alaska native' | 'black/african american/negro' |'chinese' | 'japanese' | 'other asian or pacific islander' | 'other race, nec' |'white' ——————| ————-| ————-| ————-| ——————-|————-|————- 1.254868 | 3.265529|2.483579 | 2.104784 | 1.817166 | 1.988873| 1.183073 Similar to the prediction model for Trump, the age and gender predictors are significant but education level and household income level are not. And again the AIC and BIC are quite large, this is not a strong model overall. According to our stratification model, we get the mean estimate of the voting probability of Trump to be 0.64 and Biden to be 0.61, while The minimum and maximum estimate for Biden are actually greater than Trump. These estimates are based on the previous logistic model and using race and gender as cells.

Discussion

The goal of our project is to predict who will win the American Federal Election in 2020 by a logistic regression model. First of all, we have built a new data set by separating the supporters between Donald Trump and Joe Biden with three variables from the origin survey data on June 25th2020, which is gender, income_level and education_level. In our logistic model, it represents that the people more likely to vote

for Donald Trump as their age increase (β_1 = 0.00878), in the male group(β_2= 0.452740), the lower level of education (β_3=0.039481). On the other hand, the people who vote for Joe Biden are more likely in a younger age group and within a female group or people have a higher level of education. Moreover, by our post-stratification model, we divided the origin census data set into 51 groups by state variable to identify the states distribution between Donald Trump and Joe Biden, and plug the new census data set into a logistic regression model to calculate the probability each candidate can win this election. According to our post-stratification model, the estimate of $\hat{p}$ in Donald Trump is higher than Joe Biden in West Coast and East Coast such as California and Florida states.

A link the Github repository is https://github.com/KaixiZhang99/STA304-PS3-Group-164

Conclusion

According to our post-stratification model, the mean estimated proportion of voters in favour of voting for Donald Trump (Conservative) is 0.6263, which is higher than the mean of estimated proportion for Joe Biden. Although the overall vote between two candidates are quite similar, the first and third quartiles of support in Donald Trump are slightly higher than Joe Biden. Based on these estimated proportions, we predict that Conservative will win the 2020 American Election.

Weaknesses

First of all, the survey data is not very large. As we build the prediction model based on the survey data, if the data set is small, there is a large chance that the prediction will be off. Also, in the prediction model, we didn't include the variable state because of the long running time in R. This should be a very strong predictor of the voting result, and missing it might cause bias. Lastly, the survey data and census data is not very up-to-date. The survey data is collected in June while the census data is collected in 2018. Old data sets might cause bias as well.

Next Steps

First thing we should do is to collect larger and up-to-date data sets. With the latest knowledge, it will give better predictions. Another survey could be collected after the election, and comparison between our model's prediction and the survey's results could be made. Also, we should include state in our model in order to be more realistic. Maybe we should code wisely and find a way to shorten the running time.

References

1. Kolenikov, Stas J. "Post-Stratification or a Non-Response Adjustment?" Survey Practice, vol. 9, no. 3, 31 July 2016, pp. 1–12., doi:10.29115/sp-2016-0014.
2. Sommet, Nicolas, and Davide Morselli. "Correction: Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS." International Review of Social Psychology, vol. 30, no. 1, 8 Sept. 2017, pp. 229–230., doi:10.5334/irsp.162.
3. Statistics Solutions. "What Is Logistic Regression?" Statistics Solutions, 9 Mar. 2020, www.statisticssolutions.com/what-is-logistic-regression/.
4. stevensteven 38911 gold badge66 silver badges1818 bronze badges, et al. "Replace Factors with a Numeric Value." Stack Overflow, Dec. 2016, stackoverflow.com/questions/34059017/replace-factors-with-a-numeric-value. UCLA. "MIXED EFFECTS LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES." IDRE Stats, stats.idre.ucla.edu/r/dae/mixed-effects-logistic-regression/.
5. US Government. "Presidential Election Process." USAGov, 13 July 2020, www.usa.gov/election.

6. Team, MPC UX/UI. "U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH." IPUMS USA, usa.ipums.org/usa/index.shtml.
7. "Data." Democracy Fund Voter Study Group, www.voterstudygroup.org/data.
8. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
9. 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
10. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
11. Alexander,Rohan, and Sam Caetano. "ProblemSet3-Templat-Logistic", 2 Nov,2020