Predicting the overall popular vote of the 2020 American federal election (Donald Trump v.s. Joe Biden) by MRP

Jiayang Wang & Zhanhe Zhang & Yanyi Wang & Tenglong Mai

November 2

Model

We are interested in predicting the overall popular vote of the 2020 American federal election using a regression model with post-stratification, which is a statistical method to predict population by sample data. Since survey data is randomly sampled (by UCLA survey data user guide), we choose to use a logistic regression model with post-stratification instead of multilevel regression model with post-stratification(MRP) because MRP is mostly used for predicting population by non-representative data. In the following subsections, we will describe the model specifics and the post-stratification.

Model Specifics

We firstly build a logistic regression model that is used to model the probability of response variable existing by explanatory variables. In our case, we want to predict the probability of a voter voting for Joe Biden by other predictors, including age group, gender, education, state, household income and race in our survey data, and we build a logistic regression model in R with the following formula:

$$\mathbf{log}(\frac{\mathbf{Prob\hat{B}iden}}{1-\mathbf{Prob\hat{B}iden}}) = -1.26066 - 0.69430\mathbf{Age21to35} + \ldots - 1.18370\mathbf{AgeAbove80} - 0.41207\mathbf{GenderMale} - 0.31210\mathbf{EduHighSchool} + \ldots - 0.29276\mathbf{EduMaster} + 1.24862\mathbf{StateAL} + \ldots + 1.94068\mathbf{StateWY} + 0.66605\mathbf{IncomeUnder14{,}999} + \ldots - 0.15864\mathbf{IncomeAbove250K} + 2.44468\mathbf{RaceBlack} + \ldots + 0.17463\mathbf{RaceWhite}$$

Where $\mathbf{Prob\hat{B}iden}$ is the expected proportion of vote for Joe Biden. - 1.26066 is the intercepts which means when all other predictors are 0, the log-odds is - 1.26066, where we define $\mathbf{log}(\frac{\mathbf{Prob\hat{B}iden}}{1-\mathbf{Prob\hat{B}iden}})$ is the log-odds, and then we can calculate the probability of voting for Joe Biden by log-odds. Besides, **GenderMale** is a dummy variable of gender, where female is the baseline for variables of sex, which means when other predictors are unchanged (i.e. agegroup, education, etc.) and the gender of the voter changed from female to male, the log-odds will decrease by 0.41207. Similarly, agegroup, education, state, income and race are all dummy variables that can be interpreted as when other predictors are unchanged, if one of the variable changes from 0 to 1, it means the voter changes from the baseline to the corresponding variable with corresponding coefficients, and we expect the log-odds changes by that coefficient. Additionally, we use education as a predictor for voting probability because we believe education has influence on voting proportion; however, when testing whether the coefficients of variables of education equals to 0, all of the p-values are greater than the significance level of 5%, so we consider coefficients of education variables are not statistically different from 0, which means education is a bad predictor for predicting vote probability. And when considering predictors of age and age group, we conduct two different logistic regression and find that the AIC of the model using age and age group with other predictors unchanged are 5454.9 and 5443.3 separately, where smaller value means better model performance.

Post-Stratification

We continue our study by conducting a post-stratification analysis to estimate the proportion of voting for Joe Biden, where post-stratification aggregates cell-level value by weighting each cell by its relative proportion in the population, the sampling weights are adjusted so that they sum to the population sizes within each post-stratum (by PennState). This calculation results in removing bias because of nonresponse and underrepresented groups in the population. In our case, the census data consists of variable person weight and states, so post-stratification helps to calculate the predicted votes for Joe Biden and Donald Trump. We firstly divide census data into 51 cells by states, and then we apply our logistic regression model on census data to predict the probability of the vote in the census data in each state by predictors of age group, gender, education, state, household income and race. Then we sum the person weight in each state as the population for each state, and then we weight each proportion estimate by the corresponding population of that state and sum them together and divide that by the total population. This process can be denoted as $\hat{Y}^{PS}$, which also can be shown in formula: $\hat{Y}^{PS} = \frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$, where N_j is the population for a state, and $\hat{y_j}$ is the estimated weight of the vote proportion for that state. Besides, we use another method to predict the vote proportion for Joe Biden, combining with the election policy of the United States. We firstly apply our logistic regression model on census data to predict the probability of the vote in the census data in each state, then we convert this probability into a prediction of a specific candidate, when the probability is greater or equal to 0.5, it is considered to vote for Joe Biden, and otherwise Donald Trump. Then we calculate the total vote for each candidate by taking person weight in census data into account. Considering the policy for American federal election, it does not count the total votes across the country. Instead of this, the candidate with more votes in the state got all votes of the electoral college in that state[1], and the candidates who gained the most electoral college votes won the election. So we group our data by state and calculate the total votes from the electoral college in each state. And finally, summarising the total electoral votes in each state of two candidates, we have the estimated election poll result.

Results

When testing whether the coefficients of variables of agegroup and gender equal to 0(Null hypothesis), we see that all of the p-values are less than significance level of 5%, so we reject our null hypothesis, say that the coefficients of agegroup and gender variables are not 0, which means agegroup and gender are good predictors for our logistic regression model. However, the p-values of education predictor and some variables of state, household income and race predictors are greater than significance level of 5%, so we consider coefficients of these variables are not statistically different from 0, which means they are bad predictors for predicting vote probability. Besides, when diagnosing our logistic regression model, we construct a confusion matrix to visualize the model performance:

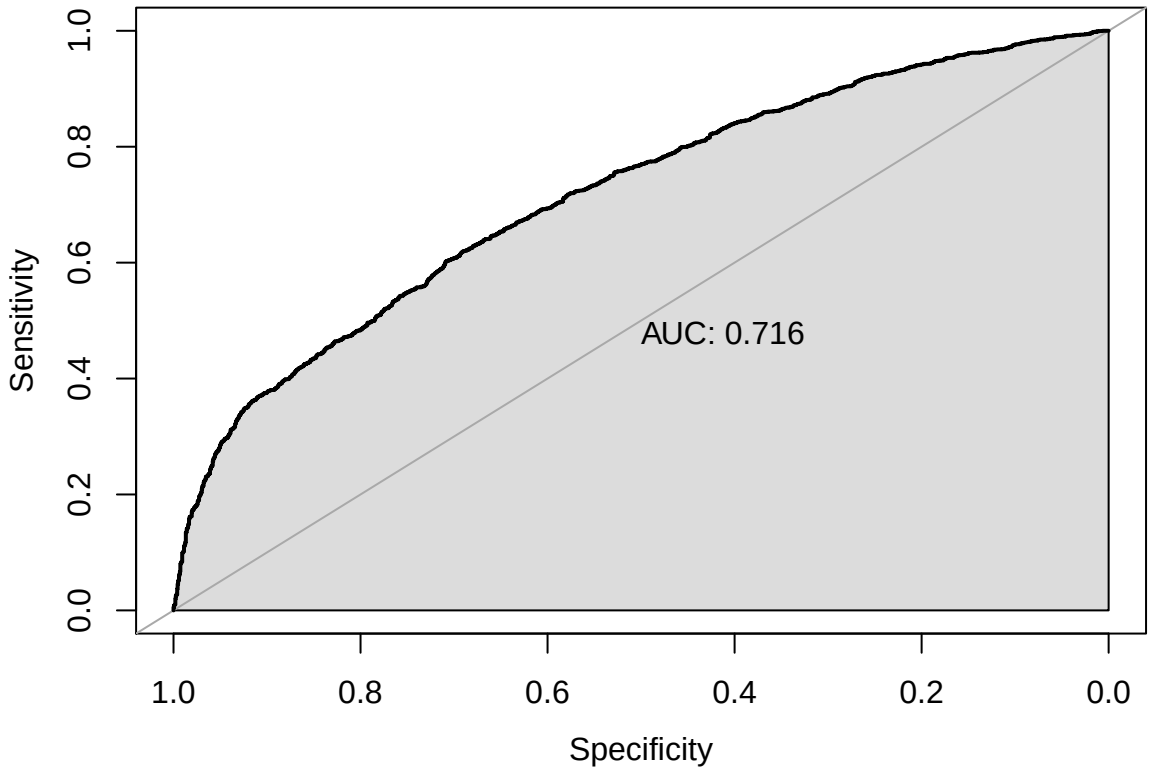
Table.1.

```
## $table
##               Reference
## Prediction     Donald Trump Joe Biden
##   Donald Trump         1398       843
##   Joe Biden             652      1403
```

From table 1 we can observe that the model captures the true number of votes for Donald Trump is 1398, whereas the ture number of the votes for Joe Biden is 1403 in total 4296 observation with a 65.2% accuracy. Besides, we plot an AUC curve(Area under the curve) also to visualize the model performance:

[1]This election policy does not include Maine and Nebraska; however, we assume these two states have the same election rule due to missing data

Fig.1.



We see that the area under the curve is 0.716, which means there is 71.6% probability that our model will predict the true result. Therefore, based our findings, we can say that our logistic regression model is good in general, but not perfect enough. In addition, when conducting the post-stratification analysis, we divide census data into 51 cells based on states, and apply our logistic regression model on census data. Then we sum the person weight in each state as the population for each state, and we weight each proportion estimate by the corresponding population of that state and sum them together and divide that by the total population, and finally we calculate our $\hat{Y}^{PS} = 0.55$ by applying formula $\hat{Y}^{PS} = \frac{\sum_j N_j \hat{y}_j}{\sum_j N_j}$, where N_j is the population for a state, and $\hat{y}_j$ is the estimated weight of the vote proportion for that state. So we can predict the voting probability for Joe Biden by $\hat{Y}^{PS}$, meaning the estimated votes proportion is 55%, where the estimation is based on our post-stratification analysis of the proportion of voters in favor of Joe Biden modeled by a logistic regression model, which accounted for age group, gender, education, state, household income, and race. Besides, when predicting for vote proportion of Joe Biden in conbining with the U.S. election policy, we have 304 votes for Joe Biden and 234 votes for Donald Trump out of total 538 votes in electoral college, representing 56.5% proportion of votes for Joe biden.

Discussion

Summary

Our study use a logistic regression model with post-stratification to predict who can win the 2020 American federal election between Joe Biden and Donald Trump. We firstly create a logistic regression model in survey data with predictors, including age group, gender, education, state, household income and race, then we conduct a post-stratification analysis on census data to estimate the proportion of voting for Joe Biden, where we apply our logistic regression model on census data for each state to predict the votes proportion; and then we calculate $\hat{Y}^{PS} = 0.55$. In addition, we use another method to predict the vote proportion for

Joe Biden, considering the U.S. election policy, where we stratify census data into states and calculate the total electoral votes to identify who wins the election, and we have result that Donald Trump wins 234 votes and Joe Biden wins 304 votes.

Conclusion

Based on table 1 in result, our logistic regression model captures the true number of votes for Donald Trump is 1398 and 1403 votes for Joe Biden in total 4296 observation in survey data. It is obvious that there is 65.2% accuracy that our logistic model can capture the true voting results, which is fine. And the ROC curve shows that the probability of our model for predicting the results correctly is 71.6% meaning that our model is reliable. As we predict in electoral college by post-stratification, Donald Trump wins 234 votes and Joe Biden wins 304 votes. We can see that the probability of Joe Biden wins the election is 56.5%, whereas our $\hat{y}^{PS}$ is 0.55, which is used to estimated the true voting proportion, meaning that the proportion of voters will vote for Joe Biden is 55%. By comparison, two results have slight difference of 1.5%, which is minor, so we can safely conclude that Joe Biden is predicted to win the election by our model. Additionally, the newest election poll from CNN shows that Biden has 52% of poll result, which is close to our prediction, so that our method is appropriate.

Weaknesses

There are also some limitations in our study. Firstly, we only consider the case that all voters vote for either Donald Trump or Joe Biden; However, they are not the only candidates from Republican Party and Democratic Party, which means the estimated results may be larger than the true proportion. Secondly, we assume that all of the people will vote if they do not specify no in survey data, which may lead to the prediction result to be idealistic. In fact, voters who give up voting might not be ignored because the performance of two candidates is not satisfactory for some people, and it may affect the accuracy of our prediction result. Similarly, when cleaning the dataset, we drop some missing data (NAs) for better modeling, but it indeed decreased the observation size, which may also lead to the inaccurate of our model and predicted results. Besides, we find that education is a bad predictor when modeling maybe because we conbine some education variables together when mapping the style of two dataset for better modeling; however, it may also affect the accuracy of our prediction results. In addition, when considering the U.S. election policy in our method, we assume Maine and Nebraska has the same election policy with other states because the information of congressional district is missing in census data; however, this is not true. In these two states, candidate who has the most votes in state has two electoral votes, then compare congressional districts within the state, and winners from each congressional district receive an electoral vote, and this will lead to a minor error in the final electoral vote results, which may be the reason of the small difference between the estimate of electoral vote and $\hat{y}^{PS}$. What's more, the census data is collected in 2018, so our prediction is lack of timeliness.

Next Steps

We are going to focus on some aspects for future improvement. Firstly, we can use predictor similar to education but is more accurate to predict voting proportion, so that our model will have better performance. Besides, for the electoral vote calculation method, we can separate Maine and Nebraska from other states to calculate the number of votes, which may reduce the error and make our prediction be more accurate. Finally, in terms of timeliness, we can add 2 years on the age in census data because our census data is in 2018 but now 2020. This can make our model be more reliable to predict the result in 2020; however, this may not be considered as a good solution, so we can make a survey and collect new data for future elections.

References

1. Post-stratification and further topics on stratification: PennState. (n.d.). Retrieved November 02, 2020, from https://online.stat.psu.edu/stat506/lesson/6/6.3
2. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. (2016). caret: Classification and Regression Training. R package version 6.0-71. https://CRAN.R-project.org/package=caret
3. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." BMC Bioinformatics, 12, 77.
4. Presidential Candidates. (2020, November 2). Retrieved November 02, 2020, from https://www.cnn.com/election/2020/candidates
5. Astrachan, L. (2020, September). Democracy Fund + UCLA Nationscape User Guide. Retrieved November 02, 2020, from https://www.voterstudygroup.org/uploads/reports/Data/Nationscape-User-Guide_2020sep10.pdf

Appendix

GitHub repo: https://github.com/DJSmallOcean/STA304-PS3