Forecasting the 2020 American Presidential Election

Ran Li, Andrei Velasevic

Monday November 2, 2020

Model

In this analysis, our goal is to predict the popular vote outcome of the 2020 American federal election. To do this we are employing a post-stratification technique with a multilevel logistic regression model. The specifics of the model and post-stratification calculation will be discussed in the following subsection:

Model Specifics

We decided to use multilevel logistic regression to model the proportion of voters who will vote for Donald Trump. In our model, we are using a variety of factors to model the probability of Donald Trump being voted for. The data for the model was taken from the Democracy Fund + UCLA Nationscape 'Full Data Set', which is an individual-level survey of *6479* observations of *18* categories. To create a meaningful model for our prediction, we took the explanatory variables to be:
1. Gender
2. Age
3. State
4. Education
5. Race
6. Employment

and build a two_level logistic regression model given by:

$$P(Y_{ij} \in \{Donald\ Trump, Joe\ Biden\}) = logit^{-1}(\beta_{0j} + \beta_1(gender) + \beta_2(age\ group) + \beta_3(employment) + \epsilon_i)$$

where Y represents the proportion of voters who will vote for Donald Trump. β_0 represents the base intercept of the model and β_1 correspond to the slope of the model in relation to the gender category, treated as a binary outcome (Male or Female).So, in example for every possible outcome of gender, we expect a β_1 change in the probability of voting for Donald Trump.However, since we believe that people in similar age, state, education, race, employment group behave similar in voting, the intercept term is dependent on these varaibles and randomness of the intercept term is modeled by the second level regression model as followed:

$$\beta_{0j} = r_{00} + r_{01}b_j^{state} + r_{02}b_j^{race} + r_{03}b_j^{eduction} + \epsilon$$

Where in the above model, β_{0j}, corresponds to the random intercept, r_{00} is an overall intercept for second level regression and r_{0i}s are the slopes in relation to the other dependent observations b_j. The reason how we set categories in these varaibles will be explained in more detail in the poststratification section.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump our group needed to perform a post-stratification analysis. The basic idea is to split data into cells and take weighted average of estimates within each cell This technique is used usually to account for underrepresented groups in the population. (https://www.stata.com/manuals13/svypoststratification.pdf) and this is very useful because it decreases non response bias and bias from underrepresented groups. It also decreases variance estimates of models.

Here cells are created based off of different ages greater than 18 (4 categories), gender(2 categories), states (51 categories, including Disctrict of Columbia), education (10 cataegories), race(7 categories), and employment (4 categories) with all possible combinations which generate a total of 114,240 cells.

The ages are chosen due to certain age groups being succeptable to certain political views. Young adults tend to go to universities where they can express themselves and possibly amplify the views of the institution. Gender can also be a factor in seeing how President Trump's views on certain gender-related issues differ from Biden's. States are an obvious choice since many states have certain affiliations with political parties. Education is important because individuals may have better understandings of more complicated political topics if they have a higher education. As we have seen so far, there have been many big events that have caught the media's attention in months leading up to the election that relate to racial differences. Due to this we decided to also include census data of racial groups. Employment is a large topic in politics as unemployment rates are often brought up.

After splitting cells, by using the model described in the "Methods" subsection, we are estimating the proportion of voters in each cell and weight each proportion estimate (within each bin) by the respective population size of that bin, formula for calculation the overall estimate of proportion voting Donald Trump is presented here:

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

Further more, after estimating for the overall result of voting, we are also interested in predicting voting result in different states, races, age, gender, etc. The formula of doing estimation is the same as above, only with the population size changed into the size of each group (For example, if we want to predict proportion of voting for Trump in New York, then we are going to divide by the population size of New York.)

Result

According to our calculations, we estimated the proportion of voters for Donald Trump to be 0.4875436, meaning approximately 48.75% of people will vote for Donald Trump. By means of post-stratification, we calculated this number as the proportion of voters in favor of Donald J. Trump, modeled by a multilevel logistic regression. Our model which accounted for gender, age, race, education, and employment status gives us a clear answer in what we are searching for. Similarly, we predict that the estimated proportion of voting Joe Biden is 0.5124565. See Table 1 for the regression result, all the varaibles are statistically significant as the level $\alpha = 0.05$.

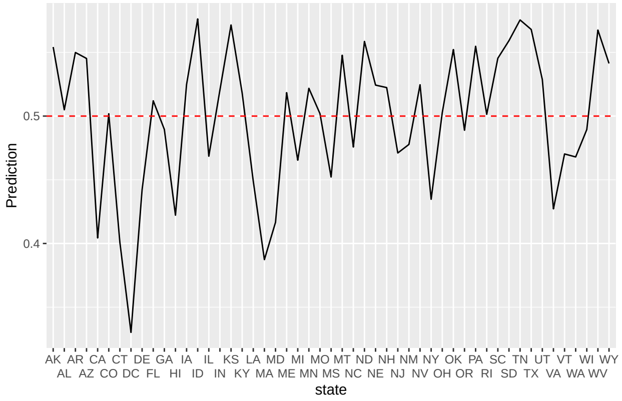
Descrptive Analysis about Prediction Reuslt

We are also interested in predicting how different states behave in voting. By setting the population as each state, we can use the same formula mentioned before to do prediction, result is presented in the following line plot. Among the 51 provinces (including District of Columbia), there are 31 out of 51 provinces having estimated proportion greater than 50% to vote for Donald Trump. Among those, TN and ID have highest estimated ratio of 57.5% and 57.6%. For those provinces with more proportion to vote Joe Biden, we can see that DC has really the lowest proportion of 33.2% and MA with 38.73%.

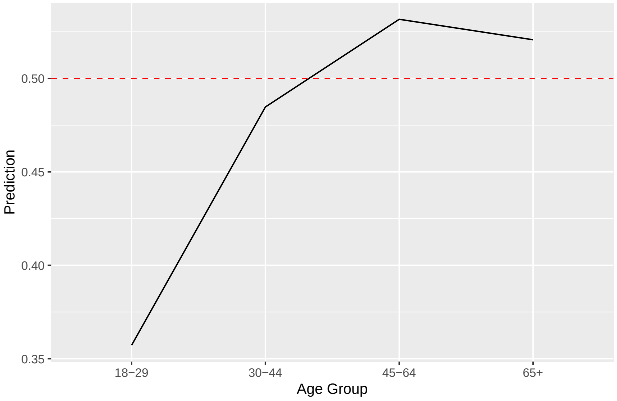
Table 1: Coefficients based on multilevel regression model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.258	0.320	-3.925	0.000
Gender: Male (refer to Female)	0.442	0.064	6.926	0.000
Age Group: 30-44 (Refer to 18-29)	0.574	0.100	5.762	0.000
Age Group: 45-64 (Refer to 18-29)	0.756	0.099	7.634	0.000
Age Group: 65+ (Refer to 18-29)	0.821	0.120	6.813	0.000
Employment Status: Others (Refer to Employed)	-0.257	0.282	-0.911	0.362
Employment Status: Not in Labor Force (Refer to Employed)	-0.360	0.085	-4.209	0.000
Employment Status: Unemployed (Refer to Employed)	-0.206	0.112	-1.849	0.064

Graph 1: Predicted Trump Vote on Different State



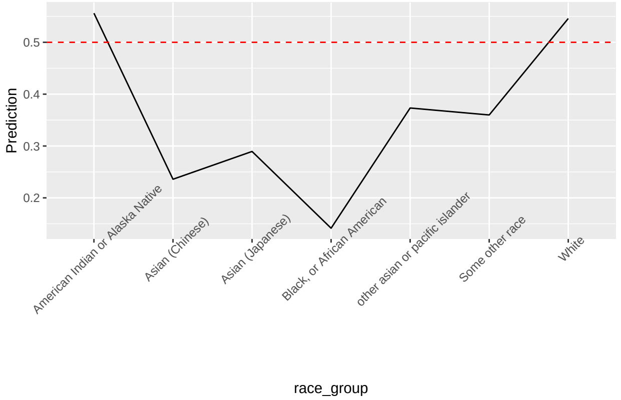
Graph 2: Predicted Trump Vote on Different Age Groups



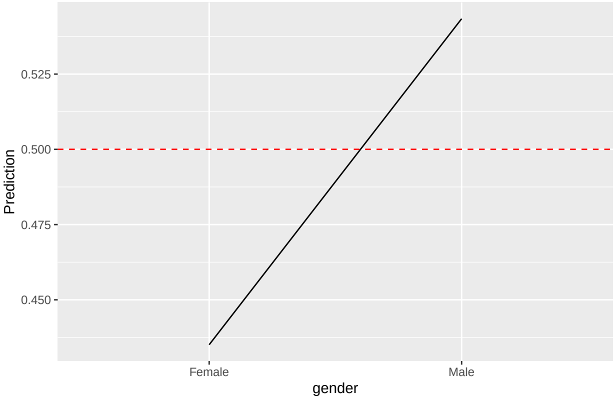
Similarly, we perfom analysis to predict voting in different age groups (18-25, 30-44, 46-54, 65+), results shown in the above line plot(see Figure 2). The graph indicates that most of the Trump supporters concentrate on age greater than 46. It is impressed that only 35% youth(18-25) are estimated to support Trump. A more detailed result in logistic regression analysis also shows this (see Table 1) and we will interpret it in our next section.

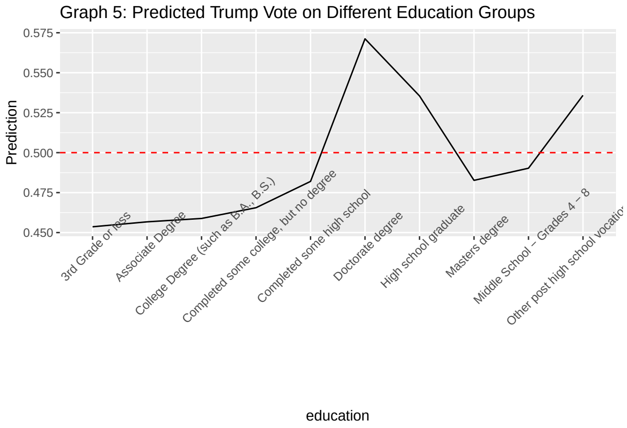
Just like these two variables, we can compare estimated voting results for other categorical variables as well. Plots are attached here.

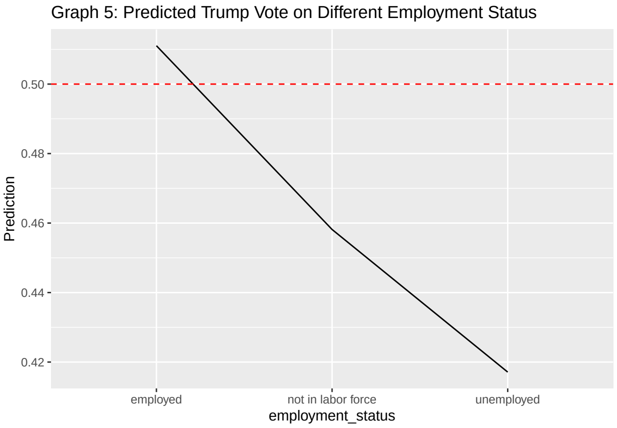
Graph 3: Predicted Trump Vote on Different Races



Graph 4: Predicted Trump Vote on Different Genders




Graph 5: Predicted Trump Vote on Different Education Groups


Graph 5: Predicted Trump Vote on Different Employment Status

Discussion

To be able to predict the outcome of the 2020 American federal election, our group performed a post-stratification calculation using census data, based on a multilevel logistic model from the Democracy Fund + UCLA Nationscape sample data. By eliminating results in the "vote 2020" category from the survey, we created a binomial response using either "Donald Trump" or "Joe Biden". From the model we treated gender, age, employment status as individual level 1 variable; race, state and education as level 2 predictor variable and build the multilevel logistic regression model with random intercept term as stated in the model section.

After splitting the data into cells and taking weighted average of estimates within each cell, post-stratification using the data from the census survey provided us with an estimate for the proportion of all people that will be voting for Donald Trump. We found this number to be 48.75%. Similarly, we also predict that the proportion of voting Biden is 51.2457%, which is slightly higher than that for Trump.

Other than that, we also did analysis for predicting the election based on state, and age (Graphs 1 & 2 respectively). From the first graph, we can see that individual have very contrasting views due to the variance of results. We can say the highest probability to vote for Donald Trump are Idaho, Kansas, and Tennessee. While the states with highest probability to vote Joe Biden are Washington DC, California, and Massachusetts (as stated in the result section). The reason these states have such a large probability for their respective candidates could be due to state influence (depending on which is Republican or Democrat leaning).

When looking at Graph 2, analysis for prediction based on age shows that most young adults are likely to be in favor of Joe Biden. However there is a strongly increasing trend for Donald Trump in relation to increasing age, peaking at ages 45-60, then dropping off thereafter. It is estimated in the regression that there is 0.574 increase in proportion of voting Trump for age group 30-44 and 0.756 increase for age group 45-64 compared to the age group 18-29, with all estimates significant at the level $\alpha = 0.05$.

Another one that is impressing is when we investigate different race groups. It can be seen in Graph 3 that white people and American Indian are major sources for Trump's supporters but most of the minority, especially black and Asians, tend to support Biden, that may be highly related to the social events in recent months.

Conclusion

In conclusion we can say that 48.75% of people will be voting for Donald Trump, therefore Joe Biden representing **Democratic Party** is expected to win the election by a narrow margin (51.25%). Given age, and state, we can also see trends supporting the hypothesis that Donald Trump will lose. Both graphs show favor for Joe Biden, along with more supportive evidence for other observations that can be seen in graphs 3-6 in the appendix.

Weaknesses

Even though MRP analysis has already decreased non response bias and bias from underrepresented groups, there are still some weakness that brings up some inaccuracy of this prediction. The most important one is due to the asymmetry of individual level survey data with the census data. For example, the "race" variable in census data has classification "Two major races" and "Three Major Races" while the "race" variable collected in the individual level data only has single race classification. Although we try to match two data sets, it is still hard to do here.For simplicity, we just deleted those who have multiple races in the data cleaning process and that is a very huge source of error for our prediction. Similar asymmetry also occurs in other variables like employment status and education.

There is also prediction error due to the inadequate population level statistics. The multilevel logistic regression has smaller AIC if we include voting in last election of respondent, while this data is not available from our census data. Also, noticebly, our census data is given in 2018, demographic structure may change in these two years.

Lastly, when we built the model, we choose multilevel logistic regression model with a binary outcome "Vote Trump" and "Vote Biden". This assumption is not accurate because even though Trump and Biden are two primary candidates, people have right to give up voting or vote others. In our analysis, we just ignore this part, which will also leads to inaccuracy.

Next Steps

Based on the weakness discussed above, we can do some adjustment to make our prediction better by resolving those problems. A multinomial logistic regression maybe apply to this context by assigning the third variable as "vote others or give up voting". Also, people can choose latest census data set (2020 dataset) that matches the individual level survey data most (it is better to include variables about political ideology or "voting last time" information).

References

tidyverse:

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

knitr:

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

lme4:

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.

datasets:

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/.

Survey Data:

Press, Courier &, et al. "New: Second Nationscape Data Set Release." Democracy Fund Voter Study Group, Democracy Fund Voter Study Group, 30 Oct. 2020, www.voterstudygroup.org/publication/nationscape-data-set.

Census Data:

Team, MPC UX/UI. "U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH." IPUMS USA, University of Minnesota, Minnesota Population Center, usa.ipums.org/usa/index.shtml.

"Poststratification —Poststratification for Survey Data." Www.stata.com, 1996–2020 StataCorp LLC, www.stata.com/manuals13/svypoststratification.pdf.

Paper:

Wang, W., et al., Forecasting elections with non-representative polls. International Journal of Forecasting(2014), http://dx.doi.org/10.1016/j.iiforecast.2014.06.001

Buttice, M., & Highton, B. (2013). How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? Political Analysis, 21(4), 449-467. Retrieved November 2, 2020, from http://www.jstor.org/stable/24572674

Appendix

Code for data analysis as well as data cleaning can be found on GitHub Repository: https://github.com/ranli123/STA304-Problem-Set-3