Who Will be the Winner of the American Federal Election in 2020

Mo wanchen, Zhongfan Sun, Zilong Yi, Xiaoya Li

Due: Novmenber 2, 2020

Code and data supporting this analysis is available at: https://github.com/Jonassun144/STA304A3-2-

Model

Model Specifics

In this analysis, we used logistic regression models to simulate the prediction of which candidate that each observation would vote for, specifically between Donald Trump and Joe Biden. Logistic regression should be applied when investigating the relationship between a binary dependent variable and other independent predictor variables. We have simulated two models, and by comparing both, we will select the model that could better represent the population. For model #1, we applied a simple logistic regression model and for model #2, we used multilevel regression to group observations by cells.

We looked at two values, AIC and AUC, to test and compare both models. AIC is a value that could represent the accuracy of the model under the consideration of complexity. In other words, the smaller the AIC, the more accurate and less complex the model would be. The AIC for model #1 we obtained is 5609.4 while that for model #2 is 5317.1 which is less than model #1, so model #2 performs better in terms of being more accurate and less complex. AUC score ranges from 0 to 1, and it represents the probability of the correctness of the model prediction, and if AUC is 1, then this model is said to be perfect because its prediction is 100% correct. In our simulation, the AUC for model #1 is 0.638 while the AUC for model #2 is 0.721, which indicates that model #2 has a better predictor over the population. Therefore, by comparing both models, model #2 which is the multilevel regression model would be a better choice. After cleaning and organizing both datasets(survey and census), running the multilevel regression model in r studio, we obtained the following formula:

Level 1: Individual Level:

$$log(\frac{y_j}{1-y_j}) = \beta_{0j} + \beta_{edu} * x_{edu} + \beta_{state} * x_{state} + \beta_{age} * x_{age} + \epsilon$$

Where, for the individual level, y_j represents the proportion or probability of voters in the j^{th} group who will vote for Joe Biden. Similarly, β_{0j} represents the j^{th} group's intercept or random effect of the model, and is the log-probability of voting for Joe Biden. For instance, American Indian or Alaska Native Female would have 0.311 probability or -0.796 log-probability less to vote for Joe Biden. Additionally, β_i represents the coefficients of corresponding explanatory variables of the model. So, for everyone one unit increase in variable i, we expect a β_i increase in the log-probability of voting for Joe Biden.

Level 2: Group Level:

$$\beta_{0j} = r_{00} + r_{0j} * W_j + u_{0j}$$

For group level, r_{0j} represents the log-probability of j^{th} group of voting for Joe Biden. r00 represents the non-random term, the intercept of the intercept β_{0j}, in the model, which will be given by the output of fixed effects. Also, it represents the log-probability of people with age less than 25 and education 3rd grade or less and currently lives in AK. Wj would be each single cell group. Finally, for both individual and group level, ϵ and u_{0j} are errors in each model, assumed to follow normal distribution with mean 0 and constant variance.

Post-Stratification

A post-stratification, in general, involves adjusting estimates' weight of each cell specified in the stage of defining the trained model, and then sum those values and divide that by the entire population size. By re-weighting and post-stratifying, the biases, to some extent, can be corrected for non-probability sampling, as in survey data, which provides a less expensive way of collecting data and hence more practical to operate. Also, post-stratification will allow us to extrapolate results of population from what we get from non-representative sample data.

By performing a post-stratification analysis, we could estimate the proportion of the population who would vote for Donald Trump and Joe Biden respectively. We have created cells based on different genders and races. Some examples for cells in our analysis would be Female White, Male Chinese, and Female Black or African American. After grouping observations by cells, we then used the model described in the model specific section to estimate the proportion of voters in each single cell. Then, we weighted each proportion estimate for each cell group by their corresponding population size, summing those percentages up, and finally dividing the summation by the whole population size in the census dataset. The reasons why we chose gender and race as group level variables lie in the fact that different gender or races present different vote intentions. "When no black congressional candidate is on the ballot, the general-election turnout for black voters is, on average, 40 percent in a district where black people make up 10 percent of the citizen voting-age population"(Fraga, Bernard). Furthermore, research conducted by Hatemi, P. et al, indicates gender exerts influence on the decision of voting.

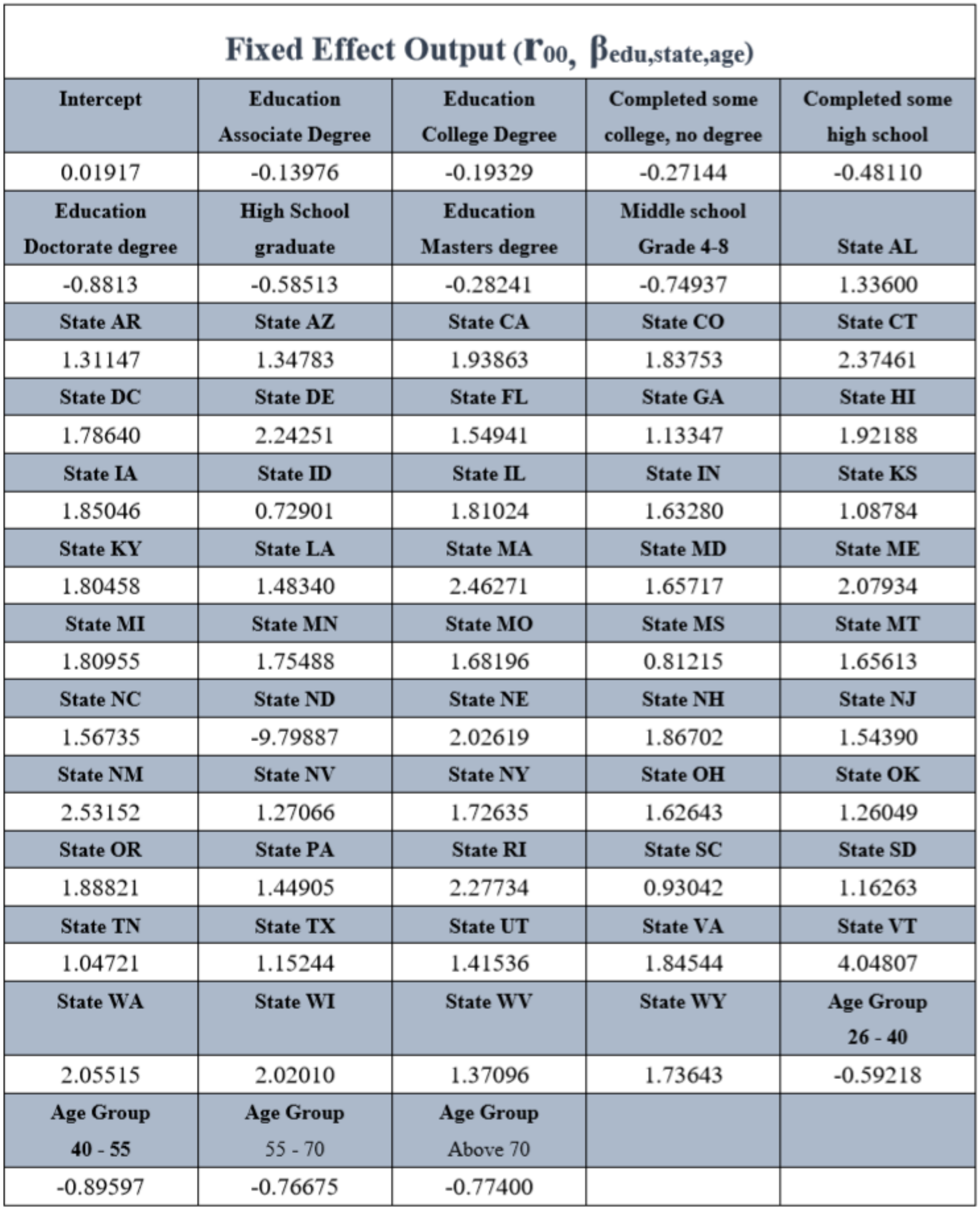
The Electoral College is used in the US presidential election. According to the Electoral College, Citizens of each U.S. state first elect the electors of the state, and then the electors vote who will be the president on behalf of the states. There are currently 538 electors in the U.S., and each state has different numbers of electors, and all electors in a state will vote for one person to be president. For example, California has 55 electors. If Biden is the winner in California, all 55 electors will vote for Biden. Although we are not able to build a model that uses states as a group level, we post a stratified base on the Electoral College to get more accurate and realistic results. We used our model with new data in the census to make predictions. The personal weights for the votes in the census are different, so, we calculate the winner for each state combined with the personal weight. Finally, we add the electors' votes from each state together to predict the winner in the U.S. However, Maine and Nebraska have different Electoral College policies. However, there is no variable in the census dataset that allows us to distinguish the two states. So, we assume the policy for the two states is the same with others.

Results

Results from Model

The table below displays the output of fixed effect. r00 =0.01917 is the non-random intercept in the equation of group level. The rest would be the value for β_i, where i is variable including education, state and age group.

Table 1 - Fixed Effect Output

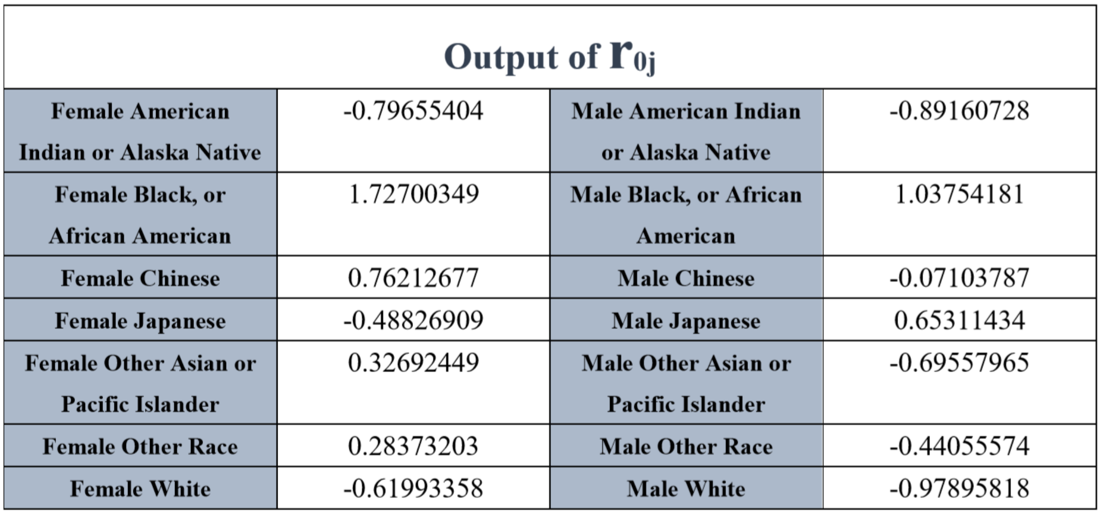


Fixed Effect Output (r_{00}, $\beta_{edu,state,age}$)

Intercept	**Education Associate Degree**	**Education College Degree**	**Completed some college, no degree**	**Completed some high school**
0.01917	-0.13976	-0.19329	-0.27144	-0.48110
Education Doctorate degree	**High School graduate**	**Education Masters degree**	**Middle school Grade 4-8**	**State AL**
-0.8813	-0.58513	-0.28241	-0.74937	1.33600
State AR	**State AZ**	**State CA**	**State CO**	**State CT**
1.31147	1.34783	1.93863	1.83753	2.37461
State DC	**State DE**	**State FL**	**State GA**	**State HI**
1.78640	2.24251	1.54941	1.13347	1.92188
State IA	**State ID**	**State IL**	**State IN**	**State KS**
1.85046	0.72901	1.81024	1.63280	1.08784
State KY	**State LA**	**State MA**	**State MD**	**State ME**
1.80458	1.48340	2.46271	1.65717	2.07934
State MI	**State MN**	**State MO**	**State MS**	**State MT**
1.80955	1.75488	1.68196	0.81215	1.65613
State NC	**State ND**	**State NE**	**State NH**	**State NJ**
1.56735	-9.79887	2.02619	1.86702	1.54390
State NM	**State NV**	**State NY**	**State OH**	**State OK**
2.53152	1.27066	1.72635	1.62643	1.26049
State OR	**State PA**	**State RI**	**State SC**	**State SD**
1.88821	1.44905	2.27734	0.93042	1.16263
State TN	**State TX**	**State UT**	**State VA**	**State VT**
1.04721	1.15244	1.41536	1.84544	4.04807
State WA	**State WI**	**State WV**	**State WY**	**Age Group 26 - 40**
2.05515	2.02010	1.37096	1.73643	-0.59218
Age Group 40 - 55	**Age Group** 55 - 70	**Age Group** Above 70		
-0.89597	-0.76675	-0.77400		

The table below shows the output of r_{0j}, and they represent the slope variables of each cell group in the equation.

Table 2 - Output of r0j



Output of $\mathbf{r}_{0j}$

Female American Indian or Alaska Native	-0.79655404	**Male American Indian or Alaska Native**	-0.89160728
Female Black, or African American	1.72700349	**Male Black, or African American**	1.03754181
Female Chinese	0.76212677	**Male Chinese**	-0.07103787
Female Japanese	-0.48826909	**Male Japanese**	0.65311434
Female Other Asian or Pacific Islander	0.32692449	**Male Other Asian or Pacific Islander**	-0.69557965
Female Other Race	0.28373203	**Male Other Race**	-0.44055574
Female White	-0.61993358	**Male White**	-0.97895818

Results from Post-Stratification(race and gender) Analysis:

Based on our post-stratification analysis which aims to estimate the proportion of observations who will vote for Joe Biden, we can make a prediction about who will win the election. We simulated a multilevel logistic regression model based on education, state and age groups and using gender and race to model the intercepts. Finally, by using the formula to calculate the value yhat^ps, our estimation of the proportion of the population who will vote for Joe Biden is 0.573. In other words, our data analysis concludes that Joe Biden has a probability of 0.573 to win the election while Donald Trump might lose since the proportion of his supporting voters only account for 0.427 which is under 50 percent.

Results from Post-Stratification (states) Analysis:

The state's post-stratification analysis is based on the same logistic model that we used above. By using the formula we calculate the value for yhat^ps. We predict that the total votes for Donald Trump is 234 out of 538 (0.565), and for Joe Biden is 304 out of 538 (0.435). Therefore, similar to the result that we get from the race and gender post-stratification, Joe Biden will be the winner. As we mentioned above, we assume Maine and Nebraska have the same policies as others in our post-stratification analysis, which may cause some errors. The effect on the final result is not significant. Since there are a total of 11 votes for Maine and Nebraska, the difference in final votes for Donald Trump and Joe Biden is 70. The results are the same, that is Joe Biden is the winner.

Discussion

2020 is a dramatic year for the world meanwhile an important event is also happening in the United States. It's almost the 2020 presidential election and the people of the United States cannot wait to see who the next president of America will be. Therefore, we made two models to predict the result of the 2020 presidential election.

In this analysis, we used two modeling methods: logistic regression model and Logistic regression, to select the best representative data for the population and to group observations into cells. From the information above we can see that by using re-weighting and post-stratifying analysis, we were able to estimate the proportion of the American people who would vote for Donald Trump and Joe Biden and predict the result of the 2020 presidential election.

Furthermore, there are a couple of points worth mentioning, from the model section above we can see that model #2 has a higher AUC so we chose model #2 as our predictor over the population. We separate those observations into two levels for our modeling, individual level, and group level. The individual level is the probability of each voter to vote for Biden or Trump. And we group cells with the same characteristics together into group level and from the function we can get the probability of that group voting for Trump and Biden. Finally, from the analysis above we came up with two results from different approaches. The first result is based on race and gender, and we get a result of 57.3% of the population will vote for Biden and 43.7% of the population will vote for Trump. Biden wins. From the second perspective calculated based on states, and we get 234 out of 538 for Trump, 304 out of 538 for Biden. Biden still wins. From the reasons and modeling above we predict Joe Biden will win this 2020 presidential election and become the president of the United States.

Weaknesses and Next Steps

There are some limitations to our study. Firstly, the number of observations in the survey data set is small, and we are not able to use it to build a relatively more accurate model. For example, build the model use states at a group level would be more realistic since America has the Electoral College. Also, we reduce observations of the survey data set, the survey data is too big and our computer can not run it. This result may not represent all Americans. Next, two states have a different election system from other states. They are Maine and Nebraska within these two states; other small areas get their votes and can vote separately from their state. However, the census data did not distinguish those two states, so their data just count as the same as other states. Lastly, this dataset is time-sensitive, our dataset was from August 2020, but the election took place in November 2020, so there is a time gap. During that time some people may change their opinion about Trump or Biden this may cause a difference in our result. To increase the accuracy of our result, we are trying to pick a larger sample size next time and do more modeling. Then we will distinguish the two different states and count them separately. In the future, we will use the most recent dataset in case of the time gap.

References

1. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

2. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=86992484-2979-42c5-8b58-268eeb2b51f3]

3. Fraga, Bernard L. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout" American Journal of Political Science, February 2015. doi: 10.1111/ajps.12172.

4. Hatemi, P., McDermott, R., Bailey, J., & Martin, N. (2012). The Different Effects of Gender and Sex on Vote Choice. Political Research Quarterly, 65(1), 76-92. Retrieved November 2, 2020, from http://www.jstor.org/stable/23209561

5. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=4425de0e-a05d-4d51-8607-9330ae2b8c17

6. Johnny Deer. 2016. predict() and newdata - How does this work. Stack Overflow. Retrieved from https://stackoverflow.com/questions/38036874/predict-and-newdata-how-does-this-work

7. Neale, Thomas H. (October 6, 2017). "Electoral College Reform: Contemporary Issues for Congress" (PDF). Washington, D.C.: Congressional Research Service. Retrieved October 24, 2020.