Predict the popular vote outcome of the 2020 American federal election

Zijun Ye, Yunhan Zhao, Zuoyu Wang

2020.11.02

Code and data supporting this analysis is available at: https://github.com/zuoyuwang/STA304-PB3.git

Model

The major focus of this study is to predict the popular vote outcome of both Donald Trump and Joe Biden since they are considered the top two candidates in the 2020 American Federal Election. To ensure valid predictions, a post-stratification technique will be employed. Firstly, a model will be produced by using the sample dataset, then the census data will be applied for estimating the vote outcome. The detailed model specifics and the post-stratification calculation will be demonstrated in the following subsections.

Model Specifics

The vote outcome of each candidate is either yes or no, which is a binary outcome. Hence, a logistic regression model is suitable for predictions. The formula of this logistic regression model is shown below:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Where p represents the probability of voters to vote for Donald Trump or Joe Biden. The β coefficient is the degree of change in the outcome for every 1-unit of change in a continuous predictor variable. Similarly, the β coefficient for a categorical predictor variable is the degree of change when switching categories.

Initially, there are nine chosen predictor variables x_i involving a voter's race, gender, household income, language speaking etc..., and the detailed information of the variables selected is listed in Table 1. Since the variables appear in both dataset, they will be helpful for applying post-stratification calculation. Most of the variables' categories have been reduced and variables are all transformed into categorical which would make further interpretation of the model coefficients easier and make each partitioning featured demographic cell of people more clearly.Table 2 shows all the reformatted variable. For example, the variable "age" in the original dataset has been reformatted into variable "age_group" which contains only 3 categories. The further variable selection from these nine variables involves stepwise selection method by Bayesian Information Criterion(BIC). Since there are 2^9 possible models, the basic idea for such a method is to iteratively remove the non-important predictors or add useful predictors from the full model according to some criterion and choose the best model. The model after being stepwise BIC selection will become the final model. The model's estimated coefficients are required to be exponential which could be interpreted as the odds ratio of the interest, and the 95% confidence intervals are used to assess the significance of the estimates. The variance inflation factor(VIF) is introduced to check whether there is multicollinearity between variables and the value of VIF under 4 means there is nothing that needs to be considered. The AUC-ROC curve is used to check the final model's goodness of fit. This curve tells how much a model is

capable of distinguishing between classes. The higher the AUC, the better the model is at predicting for the true value. All of the variables, model selection and model diagnostic would be done by the software programming language R.

Post-Stratification

Post-stratification is a common technique in survey analysis, it involves weighing samples within different demographic cells based on each cell's population. Then the estimation on the total population can be calculated according to each cell's estimation and weights. This usually produced better predictions since larger cells should be more representative. The formula for post-stratification calculation is demonstrated below:

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

Where each N_j and y_j represents the population and estimation within each cell, and $\hat{y}^{ps}$ yields the post-stratification prediction over the total population.

The number of cells is produced by the logistic model mentioned above, and the cells used for Trump and Biden might be different based on two constructed models. The first step is to group people into corresponding cells and find the population within each cell. Since the variables are chosen to ensure their appearance in the census data, all the predictors variables can be applied to create the cells. However, some variables, such as "education", may have different named categories in sample dataset and census dataset, then it is important to make sure every variable is formatted well to create valid cells. Another noticeable issue is that this census dataset was conducted on 2018, so the variable 'age' should increase two in order to make predictions on 2020. Finally, the predictions on vote outcomes for both Trump and Biden can be calculated by using the formula above

Results

The final model for estimating the probability (p_1) whether people will vote for Donald Trump can be expressed as:

$$log(\frac{p_1}{1-p_1}) = \beta_0 + \beta_1 income + \beta_2 gender + \beta_3 age\_group + \beta_4 race + \beta_5 Hispanic + \beta_6 foreign\_born + \beta_7 degree$$

The final model for estimating the probability (p_2) whether people will vote for Donald Trump can be demonstrated as:

$$log(\frac{p_2}{1-p_2}) = \beta_0 + \beta_1 gender + \beta_2 race + \beta_3 degree$$

Table 3 and 4 indicate the summary of the odds ratio($\frac{p}{1-p}$) of each of the predictor variables and their lower and upper bound of the 95% confidence intervals. If the confidence interval does not contain 1, it indicates this variable is significant. In general, each variables appears to be significant. The VIF value in Table 5 and 6 indicates there are no multicollinearity between variables in both models. The ROC curves in Figure 1 and 2 show the AUC are 69% and 62% which means that these two logistic regression is able to distinguish whether people will vote for Donald Trump or Joe Biden at 69 and 62 percent of the time respectively.

The calculated post-stratification estimation indicates the probability to vote for Donald Trump is 0.418, and that for Joe Biden is 0.406, which indicates minor difference between these two probabilities. Moreover, the addition of their probabilities does not equal to one because people may not vote for either.

Discussion

The sample data is drawn from the public opinion survey projects done by Democracy Fund + UCLA Nationscape. The census data is drawn from the American Community Survey(2018 sample) from IPUMS USA. Nine potential predictor variables in both dataset are selected, and stepwise BIC method constructs logistic models for both Donald Trump and Joe Biden. Finally, census data are grouped into corresponding cells according to the models and post-stratification analysis are performed for both candidates. Based on the calculated estimation, the probability for voting Trump and Biden appears to tie with 0.418 percent for Trump and 0.406 percent for Biden. As a result, Donald Trump should have minor chance than Joe Biden to win the 2020 American federal election. However, though this study has utilized recent survey projects and census data, there may exist some potential weaknesses during estimation procedures. Following subsection will discuss some weaknesses, and raise the next steps for consideration.

Weaknesses

For the survey and datasets, first of all, the datasets have an imperfect population. It is noticed that some remote states only have a few responses. Also, the sample dataset and the census dataset may not share the sampling population while the frame population is the same in these two data. Besides, the census was done in 2018, though the variable 'age' has been increased by 2 to accommodate the year 2020, it is not practical and will influence the accuracy of the predictions. For the logistic model, there are seven predictor variables selected for Donald Trump but only three for Joe Biden. Relatively speaking, the resulting model for Biden may not be as accurate as Trump's. Another limitation is the variables are selected so that they appear in both census and sample dataset, thus some of the significant variables may not be considered. For example, the census does not contain the variable 'vote_2016'(who you voted for in 2016). While there may exist potential confounding variables. For instance, some respondents' decisions may be influenced by their families, friends or other personal reasons, and this factor is not considered in this study. Last but not least, this logistic model may not fit the real situation perfectly in voting. In this study, only individual voters are considered, but some states may have different ways to vote, such as voting by the electoral college instead of each voter. The best example is the 2016 U.S. presidential election, Hillary Clinton not only lost seven swing states but 100 electoral votes that Barack Obama had won four years earlier even though she did so despite winning the popular vote(Silver, 2017). Therefore, the current estimations may not be completely accurate.

Next Steps

After making the model and discussing the results, any subsequent work should be considered, which may help build more reasonable processes in selecting related variables and predict more accurate results by making the model more effective and practical in the future. At first, the method of the survey could be improved. As mentioned above, these surveys might contain non-sampling errors, such as imperfect sampling population and non-response problems. Hence investigating organizations are supposed to offer more channels to do the survey. For instance, introducing various ways for data collection, including telephone surveys, in-person surveys, and online surveys. Study groups can also launch some incentive policies to gain people's attention for the survey. Moreover, both census dataset and sample survey dataset contain a lot of valuable and significant variables, but this study only analyzes some parts of them due to technical and time constraints. Therefore, for the next step, other variables can be focused to analyze their correlation with the estimation. Furthermore, after the U.S presidential election, a post-study survey can be designed to compare the real result and the result predicted in this study. As a result, it may reveal potential mistakes of the current model and indicate any further adjustment to make it more useful in the future predictions.

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=89989917-05a1-4b08-8c54-9d0f2b969625.

[3] Silver, N. (2017, February 13). Clinton's Ground Game Didn't Cost Her The Election. Retrieved November 01, 2020, from https://fivethirtyeight.com/features/clintons-ground-game-didnt-cost-her-the-election/

[4] Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, | 2020. 3-52.

[5] Alan Agresti "An Introduction to Categorical Data Analysis" Second Edition. John Wiley & Sons, Inc. 2007. 99-163.

Appendix

Table 1: Original Variables Summary Table

Names	Types	Levels
vote_2020	Categorical	[Donald Trump, Joe Biden, . . .]
household_income	Categorical	[$75,000 to $79,999,. . .]
gender	Binary	[Female,Male]
age	Categorical	[25, 34,. . .]
race_ethnicity	Categorical	[Black, or African American, White,. . .]
hispanic	Binary	[Yes, No]
foreign_born	Binary	[The United States, Another country]
education	Categorical	[Associate Degree,. . .]

Table 2: Reformatted Variables Summary Table

Names	Types	Levels
vote_trump	Binary	[1,0]
income	Categorical	[15,000-49,999,. . .]
gender	Binary	[Female,Male]
age_group	Categorical	[18-34,35-59,Above 60]
race	Categorical	[Black, White, Other]
hispanic	Binary	[Yes, No]
foreign_born	Binary	[The United States, Another country]
degree	Categorical	[Less than Associate, Associate and Above. . .]

Table 3: Odds Ratio and 95% Confidence Interval of the Coefficients for Trump Model

	est	2.5	97.5
(Intercept)	0.0320811	0.0217193	0.0473862
income100,000-149,999	1.7411383	1.3932463	2.1758986

	est	2.5	97.5
income15,000-49,999	1.2316465	1.0298263	1.4730185
income50,000-99,999	1.3618081	1.1251893	1.6481861
incomeabove 150,000	2.1310559	1.6700690	2.7192883
genderMale	1.4443232	1.2921510	1.6144162
age_group35-59	1.5984297	1.3925769	1.8347118
age_groupAbove 60	1.6537358	1.4134554	1.9348625
raceOther	4.0918596	3.0052088	5.5714314
raceWhite	6.9358509	5.3306885	9.0243555
hispanicYes	0.7583050	0.6381742	0.9010495
foreign_bornThe United States	1.4713703	1.1574191	1.8704810
degreeLess than Associate	1.2199394	1.0795937	1.3785297

Table 4: Odds Ratio and 95% Confidence Interval of the Coefficients for Biden Model

	est	2.5	97.5
(Intercept)	3.0774447	2.5740355	3.6793066
genderMale	0.7148493	0.6430643	0.7946477
raceOther	0.4093992	0.3321705	0.5045832
raceWhite	0.2794098	0.2361008	0.3306632
degreeLess than Associate	0.6711442	0.6036924	0.7461325

Table 5: VIF Table of Trump Model

	GVIF	Df	GVIF^(1/(2*Df))
income	1.312648	4	1.034591
gender	1.045091	1	1.022297
age_group	1.126443	2	1.030214
race	1.196123	2	1.045789
hispanic	1.134917	1	1.065325
foreign_born	1.079522	1	1.039001
degree	1.262759	1	1.123726

Table 6: VIF Table of Biden Model

	GVIF	Df	GVIF^(1/(2*Df))
gender	1.027225	1	1.013521
race	1.025741	2	1.006374
degree	1.029669	1	1.014726

Figure 1. ROC curve of Trump Model

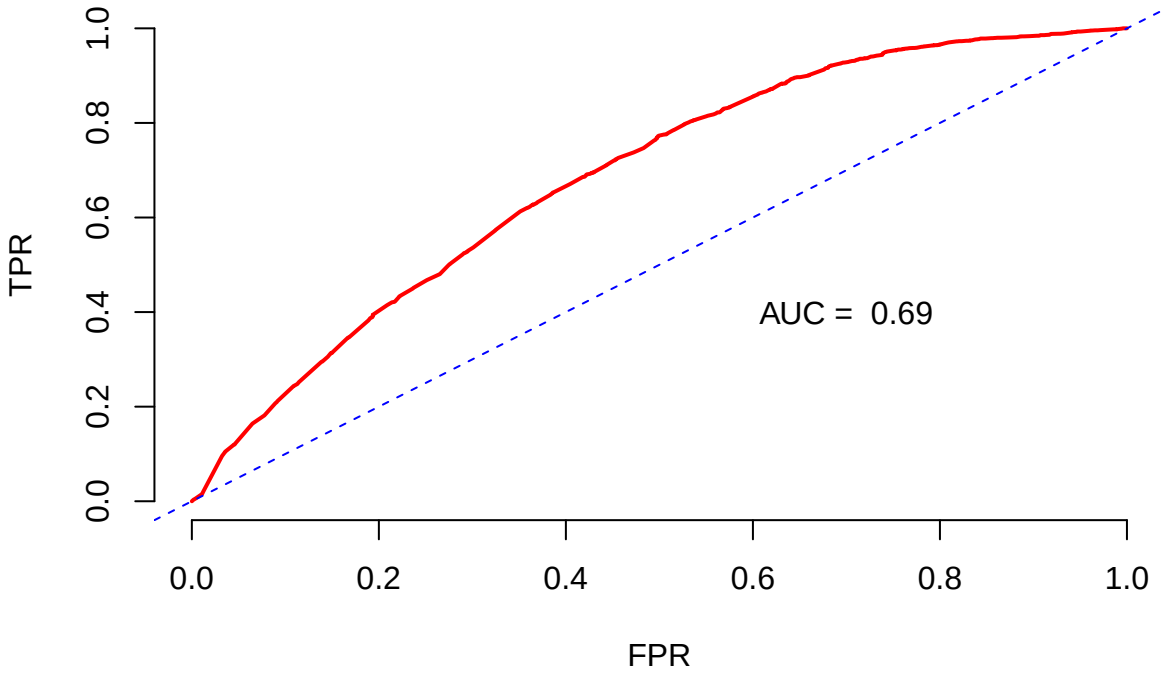
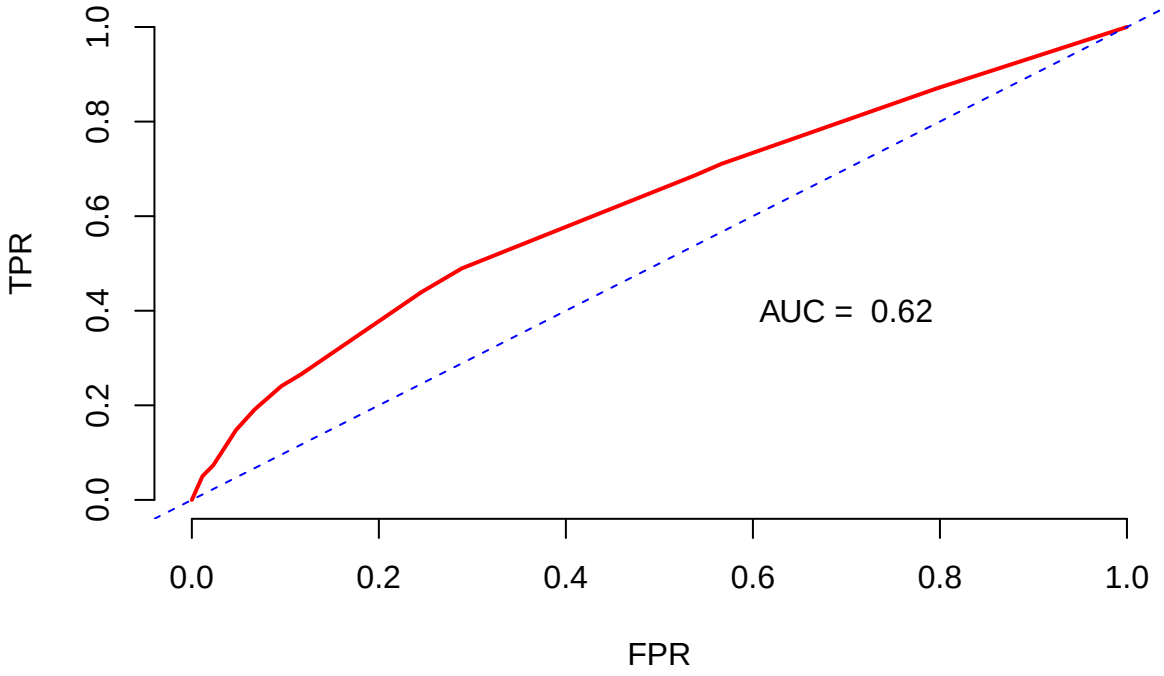


Figure 2. ROC curve of Biden Model



Code and data supporting this analysis is available at: https://github.com/zuoyuwang/STA304-PB3.git