The prediction of the outcome of 2020 American federal election

Linxia Li (1005715488), Leyi Wang (1006318682), Xingnuo Zhang (1006145306), Yanlin Li (1003770305)

November 2, 2020

Model

We are quite interested in the 2020 American federal election and would like to predict the popular vote outcome. To accomplish the goal, we fit a logistic regression model with post-stratification technique. In the following sub-sections, we will describe the model specifics and the post-stratification calculation.

Model Specifics

Model selection:

We applied a logistic regression model to estimate the proportion of voters who vote for Donald Trump. Logistic regression is a statistical model that in its basic form, uses a logistic function to model a binary response variable. Our response variable "vote_trump" is a binary response variable where 1 represents voting for Donald Trump and 0 represents voting for Joe Biden, thus a logistic model can be appropriate here.

We selected five predictors which is relevant to the probability of voting for Donald Trump (the outcome of the election): sex, age, education, state and race. In order to select the best model, we used Bayesian Information Criterion and identified sex, age and race as three most significant variables for our model. (See more information in Appendix)

Model identification:

The logistic regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Male} + \beta_2 x_{age} + \beta_3 x_{AsianorPacificIslander} + \beta_4 x_{Black,orAfricanAmerican}$$
$$+\beta_5 x_{Otherrace} + \beta_6 x_{White} + \epsilon$$

Where:

- p: the proportion of voters who will vote for Donald Trump.
- β_0: the intercept of the model.
- β_1: for a one-unit increase in male (in other words, going from female to male), we expect a β_1 increase in the log-odds of the response variable vote_trump.
- β_2: change in log odds for every unit increase in age. In other words, for every one unit increase in age, we expect a β_2 increase in the log odds.
- β_3, β_4, β_5 and β_6: similar to β_1, but corresponding to different variables

Now we fit the model.

Table 1: Logistic Model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.2378727	0.2750658	-0.8647849	0.3871568
sexMale	0.4359793	0.0587300	7.4234521	0.0000000
age	0.0075144	0.0018028	4.1681603	0.0000307
raceAsian or Pacific Islander	-1.0595698	0.2971767	-3.5654538	0.0003632
raceBlack, or African American	-2.2598234	0.2937543	-7.6929040	0.0000000
raceOther race	-0.8946695	0.2895885	-3.0894510	0.0020053
raceWhite	-0.1560129	0.2664891	-0.5854382	0.5582531

From the summary of the model we have fitted above (Table 1), the logistic regression is:

$$log(\frac{p}{1-p}) = -0.237873 + 0.435979x_{Male} + 0.007514x_{age} - 1.059570x_{AsianorPacificIslander}$$
$$-2.259823x_{Black,orAfricanAmerican} - 0.894669x_{Otherrace} - 0.156013x_{White}$$

Post-Stratification

The post-stratification analysis refers to dividing sample census data into post-strata by different estimators and adjusting weight within each post-stratum. It provides an effective way to decrease bias by eliminating non-response and underrepresented groups. We classified the sample data into cells by five weighting factors, which includes: age, sex, education, state and race. The feature selection criterion are as follows:

1. The variable is related to one's voting decision
2. The variable can be found both in survey data and census data. (Example: Previous voting record is not)
3. No too many non-responses. (Example: income is not)

For our selected variables:

1. Age: Age tends to influence the voter outcome as older people are more likely to participate in voting than young people.
2. Gender: Female voter turnout surpassed male in the past presidential election.
3. Education: the education that the voter received makes an impact on the outcome. A higher education level leads to a higher probability of voting.
4. State: Different results of presidential election will influence the states differently.
5. Race: Candidates for presidents gave different promises to different race group, thus the race group one belongs will have an impact on voting behavior.

Additional Information - Hypothesis testing

We set our null hypothesis to be: the factor is not relative to our response variable. Here is a table for hypothesis testing (Table 2):

Table 2: Hypothesis Testing

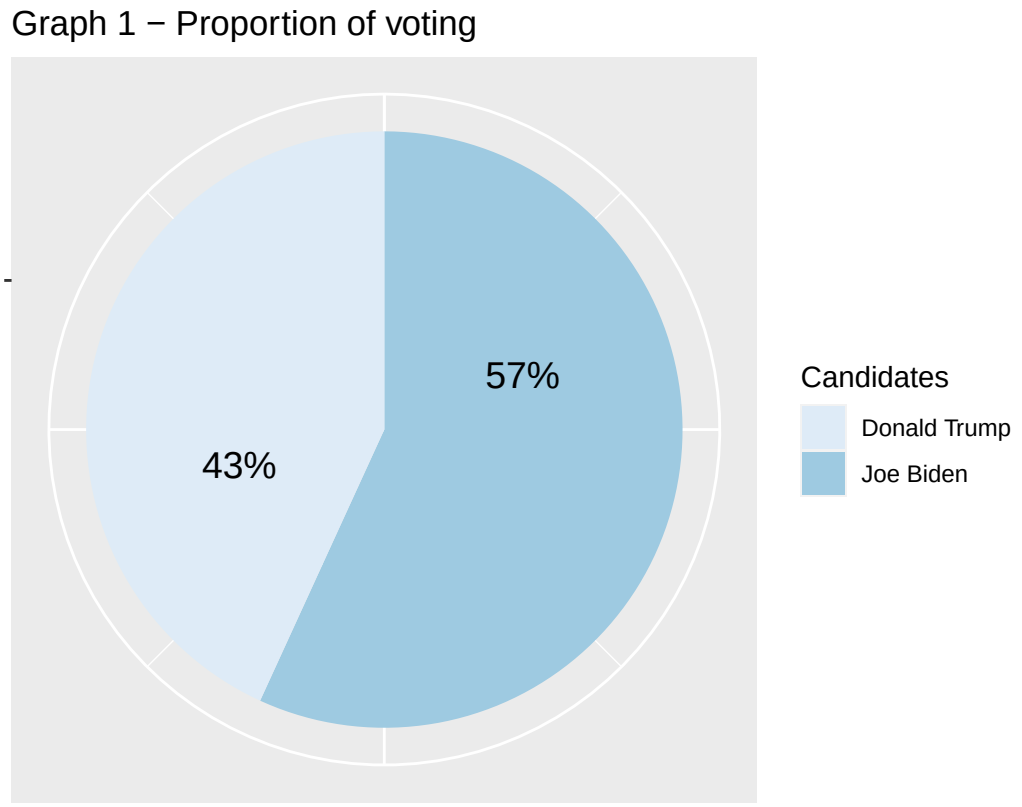
term	probability	threshold	significant
(Intercept)	0.3871568	0.05	FALSE
sexMale	0.0000000	0.05	TRUE
age	0.0000307	0.05	TRUE

term	probability	threshold	significant
raceAsian or Pacific Islander	0.0003632	0.05	TRUE
raceBlack, or African American	0.0000000	0.05	TRUE
raceOther race	0.0020053	0.05	TRUE
raceWhite	0.5582531	0.05	FALSE

The column "probability" is the probability of the null hypothesis to be true. Setting the benchmark significance level to be 0.05, we can see that there is only one insignificant factor in our model: Race-White. We can conclude from this discovery that the factors we have chosen are largely related to whether one will vote trump or not.

Results

Basing on our post-stratification analysis (introduced above) of the proportion of voters in favour of Donald Trump, and modelling by a logistic regression model with variables age, sex and race, the proportion of voters in favour of voting for Donald Trump is 43.16%. Here is a pie chart (Graph 1) that can illustrate this result:



Proportion of voting for presidential candidates

Discussion

Summary

We contribute to forecast the outcome of the American federal election and eager to know whether Donald Trump or Joe Biden will win the election. Therefore, we use two data sets, Democracy Fund + UCLA Nationscape 'Full Data Set' and American Community Surveys, to assist our prediction. The methods

of analysis are the logistic regression model and the Post-stratification technique. We selected sex, race, education, state and age to be the major factors of cells in post-stratification. By using model selection (BIC), we included sex, race and age in our logistic model. Finally, by applying our model to the post-stratification cells, we conclude that the estimated percentage of voters voting for Donald Trump will be around 43%, which means he is less likely to win the election than his rival Joe Biden.

Conclusions

According to our estimation by model, Joe Biden will win the primary vote. To illustrate, the estimated proportion of voters in favor of voting for Joe Biden being 0.569, which is greater than a half. Therefore, the result of national polling shows that Joe Biden at a significant advantage. For Trump, he needs at least 50% polling to win in the 2020 presidential election; while our analysis predicts that only 43.16% of voters votes for Trump. Comparing with Mr.Trump wins the 2016 election, which implies that there exists some disapproval towards his management.

Weaknesses

We cannot ignore the weaknesses of our analysis. Here is a list of them:

1. Our model cannot include time effect of election process. The data we used is collected once at a different time than the election. The result may change due to new speeches or political changes.
2. There are tons of other factors which can affect the outcome of the 2020 American federal election. Due to the limitation of the dataset, we can only include five of them, which may lead to underfitting.
3. National polls we used in our analysis may not be representative to the result of the election. For instance, in 2016, Hillary Clinton led in the polls and won nearly three million more votes than Donald Trump, but she still lost. This is partly because the US used an electoral college system to measure the election results, which is more complicated than national polls.

Next Steps

Here are the analysis that we can do in the future to improve our accuracy:

1. We can get access to more survey and census data which are complete and tidy. In that case, we can introduce more highly correlated factors to our model to improve accuracy.
2. We can try more models, such as time series model to analyze time effect of election.
3. After the final election outcome is released, we can compare our conclusions with the actual election results and do a post-hoc analysis to improve our estimation method.

References

1. Wikipedia, Bayesian information criterion, https://en.wikipedia.org/wiki/Bayesian_information_criterion
2. USA.gov, Who Can and Can't Vote in U.S. Elections, https://www.usa.gov/who-can-vote
3. "Poststratification-Poststratification for Survey Data", https://www.stata.com/manuals13/svypoststratification.pdf
4. Royal, Kenneth D, "Survey Research Methods: a Guide for Creating Post-Stratification Weights to Correct for Sample Bias", vol.2, 2019, pp.48-50, https://www.ehpjournal.com/article.asp?issn=2590-1761;year=2019;volume=2;issue=1;spage=48;epage=50;aulast=Royal
5. Feess, Simon, "Does Education Influence Voter Turnout?", 2001, https://www.grin.com/document/101356

6. "What Affects Voter Turnout Rates", Fairvote, https://www.fairvote.org/what_affects_voter_turnout_rates

7. "Biden vs Trump: Who Is Leading the 2020 US Election Polls", Financial Times, https://ig.ft.com/us-election-2020/

8. "Logistic Regression Analysis | State Annotated Output", https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/

9. The Visual and Data Journalism Team. "US Election 2020 Polls: Who Is Ahead - Trump or Biden?" BBC News, BBC, 1 Nov. 2020, www.bbc.com/news/election-us-2020-53657174.

10. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.

11. Kirill Müller (2017). here: A Simpler Way to Find Your Files. https://github.com/krlmlr/here, http://krlmlr.github.io/here.

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

13. 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.

14. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

15. Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. https://scales.r-lib.org, https://github.com/r-lib/scales.

16. "New: Second Nationscape Data Set Release", Democracy Fund Voter Study Group, 30 Oct. 2020, https://www.voterstudygroup.org/publication/nationscape-data-set

17. "U.S. Census Data For Social Economic, and Health Research", IPUMS USA, https://usa.ipums.org/usa/index.shtml

Appendix

Model selection - Bayesian Information Criterion (BIC)

In order to select the most significant factors that can influence the outcome of the election, we used Bayesian Information Criterion (BIC). BIC is a criterion for model selection preferring the model with the lowest BIC. It adds a penalty term to parameters in the model to avoid overfitting. The BIC is defined as:

$$BIC = kln(n) - 2ln(\hat{L})$$

where:

- $\hat{L}$: the maximized value of the likelihood function of the model
- n: the number of observed data in the data set
- k: the number of parameters in the model

In our case, we use R to minimize BIC and get an optimal model. Here is a Table (Table 3) that depicts the process of model selection:

Table 3: BIC

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	5137	6463.963	7003.017

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
- state	50	133.04092	5187	6597.004	6708.237
- education	6	37.52995	5193	6634.534	6694.428

From the table, we can see that variables state and education are removed from the model. We can then use the remaining variables: sex, age and race in our model.

Repo

Code and data supporting this analysis is available at https://github.com/Hiraethwly/Forcasting_US_Election.git