Forecasting the Overall Popular Vote of the 2020 American Federal Election Using Multilevel Regression with Post-Stratification

Xinyu Tan, Boxiang Tang, Shiyun Tang, Muxin Zhang

November 2nd, 2020

The code and data that support this analysis can be found at: https://github.com/SHIYUN-TANG/STA304-A3

Model

Given this analysis's objective is to predict the overall popular vote of the 2020 American federal election, we chose to use multilevel regression with post-stratification (MRP) for the most accurate forecasting. MRP is a common technique when it comes to forecasting issues concerning politics. It is used to obtain accurate micro-level subgroups (individual states, demographic subgroups, etc.) estimates, which are then combined with subgroups' weight in the population (calculated from census data) by certain mathematical formula to get a macro-level estimate (Ghitza & Gelman, 2013; Lax & Phillips, 2009; Park et al., 2004). Not only is MRP applicable for representative sampling data, but it can also help researchers generate quite accurate results for non-representative sampling data. (Wang et al., 2014). Hence by employing MRP, the statistical strength of the conclusion of our analysis can be guaranteed. MRP can be divided into two parts, which are multilevel regression (MR) and post-stratification (P). In the following sections, we will elaborate on how each of these two parts works, respectively.

General

Choosing appropriate predictors is crucial to modelling. Several political scientists have observed that certain demographic characteristics, such as age, sex, race, and state, can significantly influence people's voting preferences (Kaufmann & Petrocik, 1999; Penney et al., 2016; Godek, 2018). In addition to the aforementioned demographic characteristics, economic factors, such as household income and employment status, also play an essential role in affecting people's voting preferences. (Arunachalam & Watan, 2018; Grafstein, 2005). Furthermore, education level can also have an impact on one's way of thinking so that it may be correlated with people's voting preferences as well (Marshall, 2016). In conclusion, we chose the predictors (explanatory variables) of our model to be age, sex, race, state, household income, employment status and education level. The response variable of our model is the proportion of voters in favour of Trump. Moreover, considering the American federal election procedures, we also included citizenship and registration variables, which are used to filter out invalid voter samples.

Model Specifics

Our next step is to set up a multilevel logistic regression model and run it in R. Multilevel logistic regression is a statistical modelling approach best for analyzing grouped or clustered data where parameters from both individual level and group level are involved. Multilevel logistic regression model is unlike single-level logistic model where we get the target log odds and convert them into desired probabilities by simply assuming that all individuals share the same estimators, i.e., $y_i = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ such that β_i's are constants for $i = 1, ..., k$. Instead, with multilevel logistic regression, we first partition our sample data into different cells (level 2) based on two demographic characteristics, age and race. Within each cell, members all share similar demographic characteristics, and thus we can reasonably assume that they have the same estimator. The formula of our model is $y_{ij} = \alpha + (\beta + b_j)\ x_{ij} + a_j + \epsilon_{ij}$, where i stands for the individual cell member, j stands for the cell and α and β are coefficient baselines which do not vary across the cells. Notice that b_j and a_j are random variables that make our estimator's intercept and slope vary across different cells; ϵ_{ij} is the residual term, which also varies across different cells. In this way, the data of each cell can be explained by our model, which makes our results much more precise.

Post-Stratification

After cell division and modelling, we will perform post-stratification in order to obtain the proportion of voters who would vote for Trump. The key idea of this technique is to get the population estimate by calculating the weighted average of all cell-level estimates obtained from our built multilevel logistic model with the following mathematical formula: $\hat{y}^{PS} = \frac{\sum_j N_j \cdot \hat{y_j}}{\sum_j N_j}$, where $\hat{y}^{PS}$ is the population estimate which stands for the proportion of voters in the total population who would vote for Trump, $\hat{y_j}$ is the cell-level estimate which stands for the proportion of voters in the jth cell who would vote for Trump and N_j is the size of the jth cell in the population (Wang et al., 2014). Notice that since we are using a logistic model for this study, the values that we directly get from the model are the log odds rather than the proportion of voters. Therefore, we need to mathematically transform each log odds value directly obtained from the model into the corresponding proportion value and plug those proportion values in the aforementioned mathematical formula for calculating $\hat{y}^{PS}$.

Results

There are some particular states that have consistently supported one certain party in the elections over the past decades. Those states are usually referred to as the "safe states" of that party. Generally, the safe states of the Deomcratic Party are Oregon, Massachusetts, Maryland and Michigan; the safe states of the Republican Party are Alabama, Mississippi, Kansas and Idaho. From this graph, we can see that unsurprisingly, both of the two parties will still win in their own safe states. However, note that in Michigan, Biden only got a quite narrow victory, which may indicate an overall change in the preference of voters in Michigan.

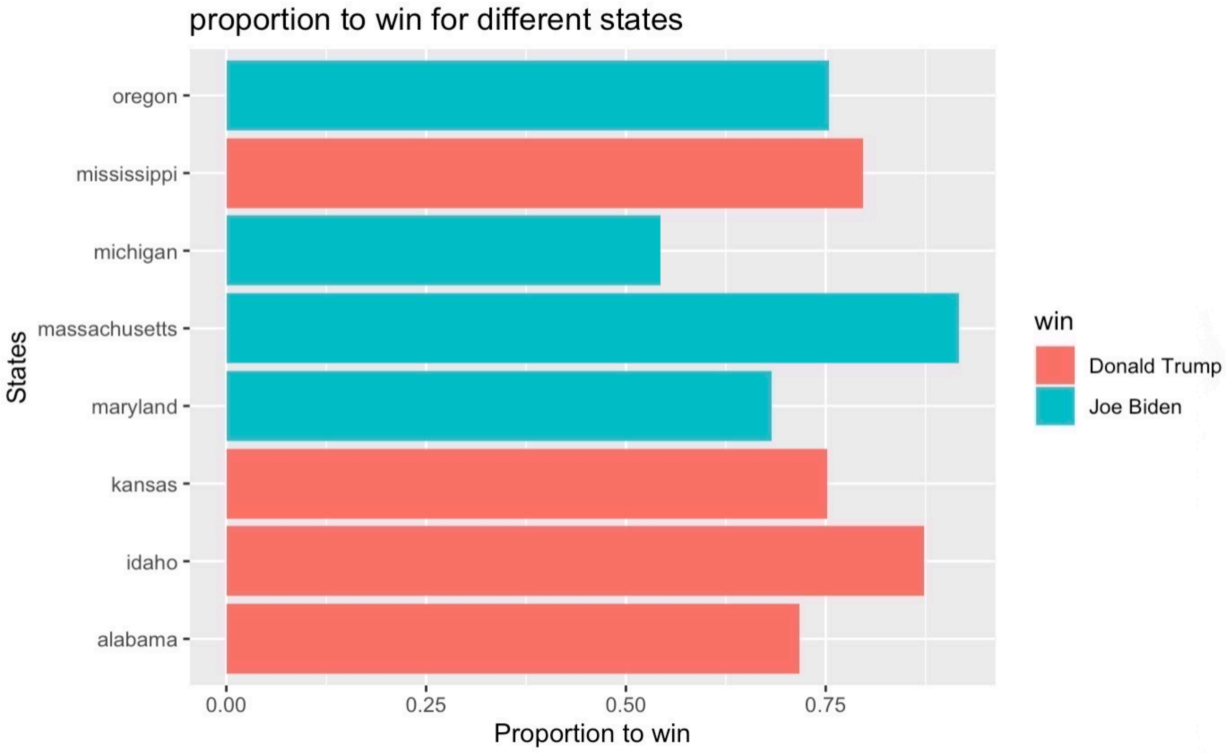
The $\hat{y}^{PS}$ value is 0.442, and this means that the expected proportion of voters who will vote for Trump is 0.442 according to our post-stratification analysis of the proportion of voters in favour of Trump modelled by a multilevel logistic regression model, where the predictors are age, gender, race, state, education, employment status and household income of voters.

In total, Trump gained about 234 votes, and Biden gained about 304 votes. There are several states with large populations, such as California, Florida, Illinois, New York, Pennsylvania and Texas, which hold the majority amount of the electoral college votes. As we can see from the table, Trump and Biden each won half of the numbers of votes of those states (Trump won in Florida, Pennsylvania and Texas; Biden won in California, Illinois and New York). This indicates that the election battle between the two parties this year should be very tense.

Table

States	Trump	Biden	Winner	Electoral Votes
California	1261380	3575503	Biden	55
Florida	1239437	986024	Trump	29

States	Trump	Biden	Winner	Electoral Votes
Illinois	312166	510164	Biden	20
New York	753330	1536186	Biden	29
Pennsylvania	253575	212091	Trump	20
Texas	1295213	437978	Trump	38



Discussion

Summary

Our analysis aims to predict the proportion of voters who will vote for Trump in the 2020 American federal election. To yield results as accurately as possible, we chose to perform a multilevel logistic regression with post-stratification. We got the individual-level survey data from Democracy Fund and UCLA Nationscape. We got the census data from the 2018 five-year American Community Surveys (ACS). While our survey data represents the population, its size is not large enough, which can potentially lower the accuracy of our analysis. We partitioned the sample data into different cells based on age and race variables. We built a multilevel logistic regression with the proportion of voters in favour of Trump being the response variable and age, gender, race, state, education, employment status and household income of voters being the explanatory variables. We then carried out the post-stratification in which we applied the model to the census data. Finally, we predicted the proportion of voters who will vote for Trump to be 0.442.

Conclusions

As mentioned in the Results section, we estimate the proportion of voters who will vote for Trump to be 0.442. Therefore, we expect Biden to win in the 2020 American Federal Election.

Weaknesses

There are a couple of weaknesses in our analysis. Firstly, although our study is based on an unbiased dataset, the dataset's size is not large enough. This means that our conclusions can be inaccurate. Such a small sample size has also hindered our cell division. We could only use two essential demographic features (age and race) to group the data; otherwise, the number of cells would be too large compared to the sample size. However, grouping the data by only two variables could lead to bias in our cell-level estimates. This would cause the population estimate to be imprecise. Secondly, our model could only account for up to a maximum of seven explanatory variables due to limited computing power. Some variables could not be included yet may potentially affect people's voting preferences. In this sense, the strength of our model has been negatively affected. Last but not least, the census data and the survey data were collected from several months ago, and part of it might have already become outdated at the moment. In particular, the current numbers and preferences of voters, which might be different from when the data was collected, are not captured by our study, this can also make our conclusions inaccurate.

Next Steps

We can take further steps to improve our study. We may first compare our prediction to the actual result of the election and evaluate the performance of our model. We may then repeat the study with a larger dataset and partition it into a larger number of cells to see if the result is improved. If there is no essential improvement, we may consider applying MRP by the Bayesian approach and see whether a proper prior distribution assumption yields a more accurate result. We may run our model on more advanced devices that have higher computing power so that more explanatory variables can be included in our model.

References

Arunachalam, R., & Watson, S. (2018). Height, income and voting. British Journal of Political Science, 48(4), 1027-1051.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48.

Democracy Fund and UCLA Nationscape. (2020, October 29). Nationscape data set. https://www.voterstudygroup.org/publication/nationscape-data-set

Encyclopaedia Britannica. (2020, November 2). United States electoral college votes by state. https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: election turnout and voting patterns among small electoral subgroups. American Journal of Political Science, 57(3), 762–776.

Godek, P. E. (2018). Determining state preferences for the electoral college: 1788-2016. The Cato Journal, 38(3), 631+.

Grafstein, R. (2005). The impact of employment status on voting behavior. The Journal of Politics, 67(3), 804-824.

Integrated Public Use Microdata Series. (2020, October 29). American community surveys. https://usa.ipums.org/usa/index.shtml

Kaufmann, K. M., & Petrocik, J. R. (1999). The changing politics of American men: understanding the sources of the gender gap. American Journal of Political Science, 43(3), 864–887.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? American Journal of Political Science, 53(1), 107–121.

Marshall, J. (2016). Education and voting conservative: evidence from a major schooling reform in Great Britain. Journal of Politics, 78(2), 382-395.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Political Analysis, 12(4), 375–385.

Penney, J., Tolley, E., & Goodyear-Grant, E. (2016). Race and gender affinities in voting: experimental evidence. Queen's University. https://www.econ.queensu.ca/research/working-papers

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.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics, 21(20), 7881.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2014). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991. Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: a grammar of data manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Wickham, H. et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.