2020 United States Presidential Election Prediction

Yuchen Cong, Jingwen Deng, Ruoxi Guan, Yuwei Sun

Nov.2, 2020

In this report, we are interested in predicting the final result of the 2020 US Presidential Election, we will introduce a multilevel regression model using survey dataset (Tausanovitch, et al., 2020) to predict the final votes for each candidate in the census dataset (Steven, et al., 2020), and we will apply the post-stratification estimate on our model. Detailed information is described in the following sections.

Data Cleaning Process

For the survey data, we only consider the observations that are both registered and have the intention to vote, and we assume people will vote unless they said no explicitly. Thus, we remove the N/A observations, and those we think are invalid, leaving 4152 observations in the survey data. For the census data, only those observations that are eligible for voting are kept, which means that all N/A observations and those with age younger than 18 or do not have citizenships observations are removed, leaving 7664 observations in the census data.

In order to ensure that variables' names and categories in the cleaned survey data can correspond to those in the census data, we performed a further data cleaning process. Noticing the age and gender are in the similar format in both data, so we match them by renaming the survey data column (from gender to sex) and splitting age into five age groups in both datasets (0-20, 21-40,41-60,61-80, and 80+). Moreover, variables of "labforce", "race", and "state" are also able to be matched by regrouping and renaming categories in our survey and census datasets. For education, some changes are made other than direct string conversion, and some assumptions are introduced in order to match two datasets (Appendix #1). For the "vote_2020", we set Donald Trump as a reference and aim to predict the probability of voting for Joe Biden.

Model

Model Specifics

In this project, we do not assume parameters follow any distributions, so we use a frequentist approach, making predictions on the underlying truths of the experiment using only two datasets. Since "vote_2020" is binary, either Donald Trump or Joe Biden, we decide to fit a generalized linear mixed-effects regression model (Using glmer() under package "lme4" in R) with a family of binomials to predict the proportion of voters who will vote for Joe Biden. Here, we set that if the predicted probability is larger than 0.5, Joe Biden wins the election. Before fitting our model, we self-defined a cell with three variables, "sex", "race" and "labforce", and we guess the values of intercept and coefficients of sex and race will change as we change different cells. Thus, the model we have fitted contains both random intercept and random coefficients. The mixed-effects logistic regression model we are using is:

$$P(Y_i = Vote\ for\ Biden \mid cell_j) = logit^{-1}(\alpha+\alpha_j+\beta_{j[i]}^{sex}+\beta_{j[i]}^{race}+\beta_{[i]}^{age\_group}+\beta_{[i]}^{education}+\beta_{[i]}^{state}+\beta_{[i]}^{labforce})$$

Where $P(Y_i = Vote\ for\ Biden \mid cell_j)$ represents the probability that respondents vote for Joe Biden, depending on the cell membership of the i^{th} respondent. α is the intercept baseline, and α_j is a random variable that follows $N(0, \sigma_\alpha^2)$, which can be represented by the difference between baseline and the intercept of each cell of the i^{th} respondent. The terms $\beta_{j[i]}^{sex}$ and $\beta_{j[i]}^{race}$ correspond to the varying coefficients associated with sex and race, which can be interpreted as the difference between the slope baseline and the coefficient of each cell of the i^{th} respondent. Here, the subscript $j[i]$ indicates the cell to which the i^{th} respondent belongs. For example, $\beta_{j[i]}^{sex}$ takes values to form $\{\beta_{male}^{sex},\ \beta_{female}^{sex}\}$ depending on the cell membership of the i^{th} respondent. The random coefficients $\beta_{j[i]}^{sex}$ and $\beta_{j[i]}^{race}$ follow $N(0, \sigma_{sex}^2)$ and $N(0, \sigma_{race}^2)$, respectively. $\beta_{[i]}^{age\_group}$, $\beta_{[i]}^{education}$, $\beta_{[i]}^{state}$ and $\beta_{[i]}^{labforce}$ are the terms with constant slope that will not be affected as we change among j cells. The reference categories are "age under 20", "3rd Grade or less", "AK", and "not in the labour force" for variables "age_group", "education", "state", and "labforce". The probability of an observation that is the i^{th} category to vote for Joe Biden is $\beta_{[i]}^{age\_group}$, $\beta_{[i]}^{education}$, $\beta_{[i]}^{state}$ or $\beta_{[i]}^{labforce}$ times the probability of observation in the corresponding reference category to vote for Joe Biden,controlling for other covariates.

Model Comparison

Besides the complexity, we are also concerned about the accuracy of our model, so we build another model, model1, with "sex" has a random coefficient, and "age_group" has a constant coefficient, keeping other variables the same as the previous model. Comparing the AIC of both models, we find out that the previous model has lower AIC, which means that model fits the data better in the sense of having fewer variables and higher accuracy. Moreover, we have checked the AUC (see Appendix #3, Figure 2) of our chosen model (0.7034), which indicates that our chosen model can discriminate between voting for Donald Trump and voting for Joe Biden 70.34% of the time. As AUC becomes closer to 1, we can say that the model has a better discrimination ability. Thus, we can conclude that the original model fits the data better by comparing AIC and AUC of both models.

Post-Stratification

After discussing the multilevel regression step, we now turn to post-stratification, where the cell-level estimates are weighted by the proportion of the electorate in each cell and aggregated to the appropriate level. It is difficult for us to use survey data to predict the probability of all American citizens who will vote for Joe Biden since there are only around four thousand observations. Since the survey data is biased and not representative enough, we need to apply the technique of post-stratification. In general, multilevel regression with post-stratification (MRP) is a statistical technique to correct estimates when there are known differences between the target population and study population.

First of all, we self-defined a cell with three variables: "race", "sex" and "labforce", in which race has seven categories, "sex" and "labforce" have two categories. After partitioning the population into 28 cells, we can use the model that we built based on the survey data to estimate the response variable per cell of the census data. We choose "labforce" because policies made by the president will have a significant influence on unemployed people. In real life, race and sex are always sensitive topics for the American Presidential Election since female and the black people always get unfair treatment since antiquity. We are interested in how these variables will impact the results of votes and how the coefficients of race and sex will change among different cells. Some variables are not included since they cannot be matched between two data.

Results

Data

Table 1: Summary of Voting Status in each State

State	Total Votes	Trump Supported Rates	Biden Supported Rates	State Winner
AK	8	0.500	0.500	Tie
AL	89	0.438	0.562	Joe Biden
AR	50	0.520	0.480	Donald Trump
AZ	162	0.432	0.568	Joe Biden
CA	717	0.303	0.697	Joe Biden
CO	98	0.418	0.582	Joe Biden

- Table 1 shows the first six observations of the summary table, which summarizes the votes according to each state. This table gives the supported rates for each candidate and shows the winner in the specific state.

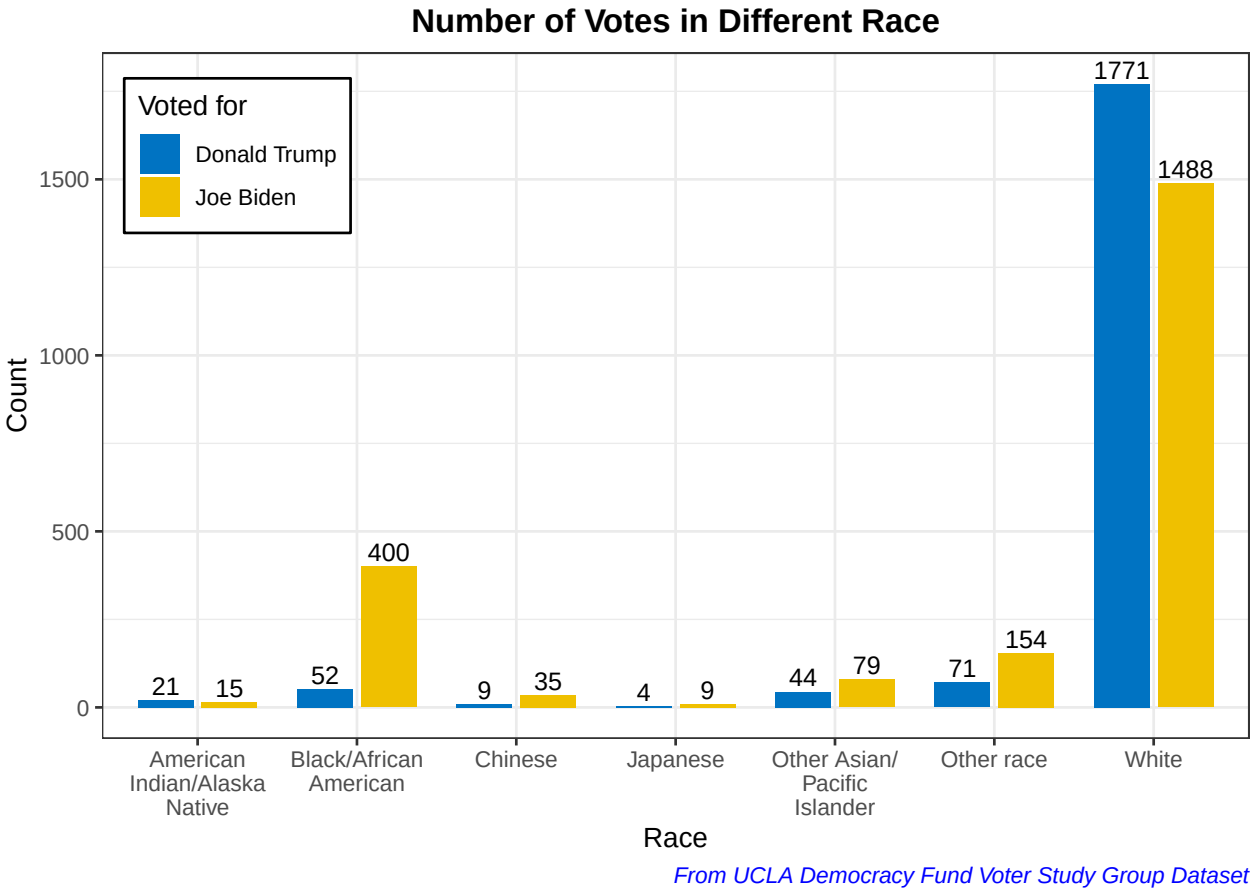

Figure 1: Number of Votes in Different Race

- Figure 1 shows the result of votes based on different types of race. Most voters are White people, and a large portion of them would vote for Donald Trump. Joe Biden gets more votes from other types of the race except for American Indian/Alaska Native and White people. However, the total number of votes for Joe Biden is larger than the total number of votes for Donald Trump in our survey dataset.

Model

Table 2: Summary of Voting Status in each State

term	estimate	std.error	statistic	p.value	group
(Intercept)	-0.600	1.555	-0.386	0.700	fixed
age_groupage 21 to 40	-0.680	0.202	-3.367	0.001	fixed
age_groupage 41 to 60	-1.055	0.203	-5.206	0.000	fixed
age_groupage 61 to 80	-0.983	0.207	-4.753	0.000	fixed
age_groupage above 80	-1.126	0.406	-2.775	0.006	fixed
educationAssociate Degree	-0.019	0.920	-0.020	0.984	fixed

- Table 2 shows the first six observations of the summary table, which summarizes our model results. From this table, we will have all the intercepts and coefficients for our chosen variables. We can also check the p-values for each category in different variables, p-values would tell us if this variable has statistical significance so that it would be influential to our outcome.

Post-Stratification

- Since we use our multilevel regression model to predict the proportion of voters who are willing to vote for Joe Biden in our post-stratification analysis. If we get the proportion is more significant than 50%, then we predict Joe Biden wins the election. Then, we calculate the post-stratification estimate $\hat{y}^{PS} = 0.563$, as $0.563 > 0.5$, which means the winner would be Joe Biden.

Table 3: Predicted Election Result

Presidential Electors	Total Votes
Donald Trump	252
Joe Biden	286

- Table 3 shows the total votes of each candidate. During the process of calculating the total number of votes, we adjust the census dataset by "perwt" variable. It means that each observation is differently weighted; one observation does not represent only one voter. Also, we calculate the total votes in each state, since in reality, the American Election is according to the electoral college (Appendix #2). After we apply these two adjustments, we get a final result of total votes for each candidate. Donald Trump has 252 votes, and Joe Biden has 286 votes.

Discussion

Summary

The very first step of our prediction about the 2020 US election is cleaning both survey dataset (Tausanovitch, et al., 2020) and census dataset (Steven, et al., 2020), since we want the chosen variables to be matched up in each dataset. Then we create two multilevel regression models based on survey data. By comparing AIC and ROC curve as described in Model Comparison section, we finally choose the model with sex and race in the cell variable to predict the election result. We calculate the post-stratification estimate to predict the winner primarily. Moreover, we want to prove our result by the predicted total votes for the two candidates based off census dataset, so we create a table showing the result. By combining all information we gain from the data, we make our final prediction.

Conclusion

The post-stratification estimate shows Joe Biden has a proportion of 56.1% votes. Then we primarily predict that Joe Biden will win. Table 3 shows that Joe Biden receives more votes than Donald Trump, so our prediction has been proved. Our final prediction is that Joe Biden will win the 2020 US Presidential Election.

Weaknesses

As described in Appendix #2, we have an assumption about the State of Maine and Nebraska, so that our result may be affected if we apply the actual electoral policy. However, the difference between total votes of Joe Biden and Donald Trump is 34 votes, and Maine and Nebraska have a total of 9 electoral votes. Even if Donald Trump gets all electoral votes from these two states, he will still lose the election in our model. In addition, most people in these two states support Joe Biden based on our data. Therefore, this assumption is valid and will not affect our final prediction result.

The size of the survey dataset we used is not very large, so the dataset may not be representative of all voters. There may exist some special cases that are not covered in our survey dataset, which can influence our model and final prediction. Additionally, the survey dataset has been done for a few months; thus, the dataset does not have very strong timeliness.

Next Steps

In our next steps, we could consider seeing if the intercept and coefficient of state will change as we use a different cell. According to the policy of the US election (Appendix #2), the final result strongly relies on the winner in each state, which means that the state can be a good factor of predicting the election winner. We can also find a larger survey dataset to support our model; for example, after the election, we collect all the actual votes as our dataset so that our prediction would be more realistic.

References

- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=6bf51fe4-3093-4076-8d2a-92a826a2a9bd.
- 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
- Cite R : R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Cite "tidyverse": Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Cite "ggplot2" : H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Cite "dplyr" : Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Cite "haven": Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1.https://CRAN.R-project.org/package=haven
- Cite "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.
- Cite "pROC": Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/
- Cite "knitr": Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.
- Cite "kableExtra": Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
- Cite "devtools": Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. R package version 2.3.2. https://CRAN.R-project.org/package=devtools
- Cite "jtools": Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data*. R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.
- Cite "stringr": Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. Rpackage version 1.4.0. https://CRAN.R-project.org/package=stringr
- "Voter Registration Age Requirements: USAGov." Voter Registration Age Requirements | USAGov, www.usa.gov/voter-registration-age-requirements.
- "List of State Electoral Votes For 2020." List of State Electoral Votes For The 2020 Election, state.1keydata.com/state-electoral-votes.php.
- Alexander, Rohan, and Sam Caetano. 2 Nov. 2020. "01-data_cleaning-post-strat1.R". Census Dataset cleaning process code.
- Alexander, Rohan, and Sam Caetano. 2 Nov. 2020. "01-data_cleaning-survey1.R". Survey Dataset cleaning process code.
- Birkett, B., Alex Birkett Alex Birkett is a former content and growth marketer at CXL. Currently, Phillip, Snehal, Centenaro, L., Seva, R., . . . Birkett, A. (2020, September 24). Bayesian vs Frequentist A/B

Testing (and Does it Even Matter?). Retrieved November 02, 2020, from https://cxl.com/blog/bayesian-frequentist-ab-testing/

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

Appendix

1. For education:

- Changes made other than direct string conversion

a) Other post high school vocational training = High school graduate (survey data)
b) Completed some graduate, but no degree = College degree
c) professional degree beyond a bachelor's degree = College degree

- Some assumptions

a) Assume Other post high school vocational training as High school graduate (Survey)
b) Assume Completed some graduate, but no degree as College degree (Survey)
c) We assume professional degree beyond a bachelor's degree as College degree (such as B.S, B.A)
d) Assume "ged or alternative credential" and "regular high school diploma" as "High school graduate"
e) Assume "some college, but less than 1 year" and "1 or more years of college credit, no degree" as "Completed some college, but no degree"

2. Rules of America Election

- The electoral college has 538 electors, and the final result comes from these electors. Each state has different numbers of electors, and each elector will vote for the candidate that owns more votes in the specific state.
- Among all states, only Maine and Nebraska choose to use a different method other than the method that the winner in a state would have all electoral votes. Instead, they have several congressional districts, and they will allocate electoral votes to each district. So the winner in each district will get the corresponding number of electoral votes. Also, the winner of the whole state will have two electoral votes as a bonus.
- Since in the census dataset, we can not find any variable related to congressional districts, then we are not able to adjust our dataset to fit the rule. Thus, we assume that Maine and Nebraska will follow the same method as other states.

3. AUC, ROC

```
## Area under the curve: 0.7034
```

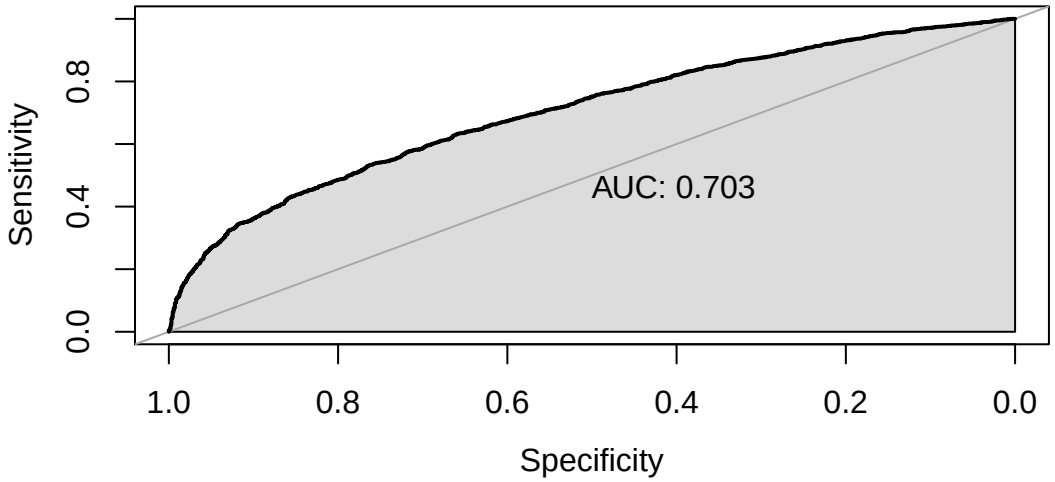

Figure 2: AUC of the Chosen Model

4. Code and data supporting this analysis is available at: https://github.com/jingwennnn/Prediction-of-2020-United-States-Presidential-Election-Result