Logistic Regression & Post-Stratification Analysis to Predict the Results of the 2020 American Election

Xiaoxuan Han, Yicheng Ding, April Ding, Amber Kao

2020/11/02

Code and data supporting this analysis is available at: https://github.com/YichengDing0106/PS3-Group14.git

Model

We are interested in predicting the popular vote outcome of the 2020 American federal election with individual-level public opinion survey data from the Democracy Fund + UCLA Nationscape and the 2018 5-year American Community Survey (ACS) census data from IPUMS USA (Ruggles et al., 2020; Tausanovitch & Vavreck, 2020). We will conduct regression analysis on the survey data to estimate the voter response, and select demographic variables from the census data to conduct a post-stratification analysis and estimate the voting results of the American population. In the following analysis, we will construct two logistic regression models–one to estimate the vote response for Donald Trump, the other to estimate the vote response for Joe Biden–and perform the post-stratification calculations.

Model Specifics

We will be using multinomial logistic regression models to model the proportion of voters who will vote for Donald Trump and the proportion of voters who will vote for Joe Biden. The reason we chose this type of model is that we are predicting a binary response based on two or more independent variables, where the log odds of the outcome is modeled as a linear combination of the predictor variables. We will be using age group, race ethnicity, gender, education level, and census region, which are all recorded as categorical variables, to perform the modeling. We will run the two multinomial logistic regression models by using the function `glm()`, and then save the descriptive statistics into a summary table.

The multinomial logistic regression model for modeling the proportion of voters who will vote for Donald Trump is:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{21} x_{21} + \epsilon$$

p_{Trump} represents the proportion of voters who will vote for Donald Trump. y_1 represents the response variable in the multinomial logistic regression. It represents the log odds of the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, which is the fixed baseline intercept representing the proportion of voters between ages 18-29, who are female American Indians or Alaska Natives with 3rd Grade or less education, are from the Midwest, and will vote for Donald Trump.

β_1 to β_3 are the estimates of the age group categories "age30-44", "age45-64", and "age65+", respectively, with "age18-29" as the reference category; β_4 to β_9 are the estimates of the race ethnicity categories "Black or African American", "Chinese", "Japanese", "Other Asian or Pacific Islander", "Other race, or two or more races", and "White", respectively, with "American Indian or Alaska Native" as the reference category; β_{10} is the multinomial logit estimate of proportions comparing females to males given the other variables in the

model are held constant; β_{11} to β_{18} are the estimates of the education level categories "4th to 8th Grade", "Associate's degree", "Bachelor's degree", "College, no degree", "Doctorate or professional degree beyond bachelor's", "High school diploma or GED", "High school, no degree", and "Master's degree", respectively, with "3rd Grade or less" as the reference category; β_{19} to β_{21} are the estimates of the census region categories "Northeast", "South", and "West", respectively, with "Midwest" as the reference category.

The multinomial logistic regression model for modeling the proportion of voters who will vote for Joe Biden is:

$$y_2 = \beta_{22} + \beta_{23}x_{22} + ... + \beta_{43}x_{42} + \epsilon$$

The interpretation of multinomial logistic regression model for modeling the proportion of voters who will vote for Joe Biden is similar to the interpretation of multinomial logistic regression model for modeling the proportion of voters who will vote for Trump. p_{Biden} represents the proportion of voters who will vote for Joe Biden. y_2 represents the response variable in the multinomial logistic regression. It represents the log odds of the proportion of voters who will vote for Joe Biden. β_{22} represents the intercept of the model, which is the fixed baseline intercept representing the proportion of voters in the same reference demographic as described above for β_0 who will vote for Joe Biden. The predictor variables are the same variables as mentioned above for the Trump model, and in the same order as well.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump, we need to perform a post-stratification analysis. Post-stratification analysis is performed with the aim of conducting inference on a large data sample that may not necessarily be representative of the target population. In order to conduct post-stratification analysis, we need to collect large amounts of demographic data along with the sample. We partition the data by cells with all the possible combinations of demographic characteristics, estimate a response for each cell, and estimate the response of the target population by weighting each cell by the number of data points in each cell. Here we create cells based on the chosen demographic characteristics, which are gender, age, census region, race ethnicity, and education. The data is partitioned into a total of 2016 cells, of which 2005 cells have at least one person with the described characteristics. The weighting of each cell is calculated by multiplying the response estimate of each cell by the respective population size of that cell, then adding together all of these values, and finally dividing the sum by the entire population size.

The formula for the post-stratification calculation is as follows:

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

Where $\hat{y}_j$ is the estimated response for the j^{th} cell, N_j is the number of samples per cell, and $\hat{y}^{PS}$ is the estimated response for the target population.

The estimation of the proportion of voters who will vote for Joe Biden is performed with the same process and calculations.

Results

Model Results

Figure 1: Descriptive statistics of the Trump model regression estimates

and Figure 2: Descriptive statistics of the Biden model regression estimates can be found in the Appendix section.

The interpretation of each estimate will follow the same principles. For example, with significance of variables in this analysis determined by the p-value threshold of $p < 0.05$, it can be seen in Figure 1 that the age30-44 group is significant and has an estimate of $\beta_1 = 0.6562082$. The regression estimates are the log odds of the voting estimate, therefore the interpretation of this result is that the likelihood of voting for Trump in the age30-44 group is $\exp(\beta) = 1.9274699$ times that of voting for Trump in the age18-29 group, when controlling for all other variables.

For the model estimating the proportion of voters voting for Trump, the following variable categories are significant: age30-44, age45-64, age65+, Black or African American, Chinese, Other race or two or more races, Male, and South.

For the model estimating the proportion of voters voting for Biden, the following variable categories are significant: age30-44, age45-64, Black or African American, Chinese, Japanese, Other Asian or Pacific Islander, Other race or two or more races, Male, and South.

Interestingly, none of the education level categories were significant compared to 3rd Grade or less education in both of the models.

Table 1: Selected variable categories for comparison between the two models

Category	Likelihood of voting for Trump	Likelihood of voting for Biden
age30-44	1.9274699	0.7673829
age45-64	2.1099462	0.7040927
Black or African American	0.1357733	5.4653772
Chinese	0.2875775	3.9639735
Other race, or two or more races	0.5032533	2.3653255
Male	1.5191956	0.7161594
South	1.3702654	0.8083042

The selected variable categories in Table 1 are the ones that are significant for both of the models. The voting likelihood values were calculated from the regression estimates with the $\exp(\beta)$ formula, therefore the values are in comparison with the reference category. Interesting findings to note include the greatly increased likelihood of Black or African American voters to vote for Biden compared to their greatly decreased likelihood of voting for Trump, the trend of older voters being more likely to vote for Trump, and the trend of ethnic minorities being more likely to vote for Biden. Male voters and Southern voters are also both more likely to vote for Trump than for Biden. The education level of an individual does not have a significant correlation with whether the individual will vote for Trump or for Biden.

Post-Stratification Results

$$\hat{p}_{Trump} = 0.4181068$$

$$\hat{p}_{Biden} = 0.4441705$$

We corrected the log odds from the logistic estimation in the post-stratification calculations, and calculated the proportions of votes that the two candidates are each likely to receive. y_1 for Trump is 0.4181068, and y_2 for Biden is 0.4441705. These results indicate that ~41.81% of voters are willing to vote for Trump, while another ~44.42% of voters prefer to vote for Biden. The two values do not add up to 100% because as mentioned above, some respondents either indicated that they would not vote, or did not yet know how they would vote. Our results predict that Joe Biden will gain more votes in total compared to Trump.

Discussion

Our survey data set is from Democracy Fund + UCLA Nationscape and census data set is from IPUMS USA. In our study, we used two logistic regression models on the survey data to perform estimation of voter response for both candidates. Following that, we conducted a post-stratification analysis on the census data focusing on the selected variables. The logistic regression models showed that certain demographic characteristics are indeed correlated with voting preference. Our results after post-stratification correction of the data showed that Biden is likely to have a higher proportion of votes, ~44.42%, compared to that of Trump, which was ~41.81%. Therefore, we predict that Joe Biden will win the popular vote of the 2020 American federal election.

Weaknesses

We selected gender, age, census region, race ethnicity, and education as our post-stratification variables. While these partitioned our data into cells with distinguishable demographic characteristics, in real life, there are many more factors (i.e. income, employment status, etc.) and variables that we need to consider when predicting the voting outcome. In terms of our data cleaning process, we took out the "respondent skipped", or the NA values from the original data set. This means that our analysis does not account for the opinions of those who did not choose to reveal their personal information, and it may cause the non-sampling error. Furthermore, for the convenience of calculation, we divided the age category into four groups, spanning from 18 years old to 65 and older, and each group covers different age gaps. This can be improved by shrinking the size of each group. Since age is a crucial factor for any average voter, and people's opinions change drastically as they age, doing so allows us to look into the specific trend of voting behaviors by a certain group of people differed by age. The same applies to other categorical variable like race ethnicity, where the criteria of each option could be more narrowed down.

In terms of modelling, our results would be improved if we used a multilevel regression post-stratification (MRP) model instead. An MRP analysis would require fitting two nested multilevel logistic regressions for estimating candidate support in each cell, and then weighing cell-level estimates by the proportion of the electorate in each cell and aggregated to the appropriate level.

Next Steps

One way to make our study more practical would be to conduct post-hoc analysis and compare the results from the actual election outcome. Follow-up surveys or exit polls can be conducted to either confirm our analysis or provide insight for ways to improve our analytical methods. We could use a least squares regression model to contrast and compare the actual observed data and our predicted data. Otherwise, we could design a survey to collect people's opinions on the determinant factors when they make decisions on whom to vote for.

Appendix

Data Cleaning

Both the survey data and the census data were cleaned with R. For the survey data, the variables `interest`, `registration`, `vote_2016`, `vote_intention`, `vote_2020`, `ideo5`, `gender`, `age`, `census_region`, `race_ethnicity`, and `education` were selected for the preliminary dataset. Rows that contained any response other than "Registered" in the `registered` variable were filtered out, since one must register to vote in order to be eligible to vote. Then, the rows containing NA values were omitted for more convenient analysis. Next, the `age`, `race_ethnicity`, and `education` variables were either split into groups, regrouped, or renamed in order to match the variables in the census data. `age` was originally recorded as a numerical variable, and it

was split into 4 age groups. The census data contained fewer `race_ethnicity` categories, so some of the categories were consolidated and renamed to match the census data.

When constructing the two logistic regression models, `gender`, `age`, `census_region`, `race_ethnicity`, `education`, and `vote_trump` or `vote_biden` were chosen for analysis. `vote_trump` and `vote_biden` are the binary response variables constructed from the `vote_2020` variable, which was not originally binary and contained "Dont' know" and "I would not vote" responses as well. In the Trump model, "Donald Trump" was recoded to 1 and all other responses were recoded to 0, and in the Biden model, "Joe Biden" was recoded to 1 and all other responses were recoded to 0.

For the census data, `SEX`, `AGE`, `REGION`, `RACE`, and `EDUCD` were selected for the preliminary dataset. Since the census data records individuals of all ages, data for people under 18 years old were removed, since one needs to be at least 18 years old to vote. `AGE` was also originally a factor variable, and was converted to integers. Next, the `AGE`, `REGION`, `RACE`, and `EDUCD` variables were either split into groups, regrouped, or renamed in order to match the variables in the survey data. `AGE` was originally recorded as a numerical variable, and it was split into the same 4 age groups as those in the survey data. `REGION` was originally divided into more specific categories compared to the survey data, and categories in the same survey region were consolidated. `RACE` was regrouped and renamed to match the categories in the survey data. The survey data contained fewer `EDUCD` categories, so some of the categories were consolidated and renamed to match the survey data. To create the demographic cells, the census data was grouped by each of the five variables, and the number of data points in each cell was counted and recorded as variable `n`. Lastly, the census data variables were renamed to match the names of the variables in the survey data for more convenient analysis.

Summary tables

Figure 1: Descriptive statistics of the Trump model regression estimates

term	estimate	std.error	statistic	p.value
(Intercept)	-1.0472542	0.7972549	-1.3135751	0.1889892
age30-44	0.6562082	0.0977051	6.7162112	0.0000000
age45-64	0.7466625	0.0961617	7.7646562	0.0000000
age65+	0.7423845	0.1065371	6.9683171	0.0000000
race_ethnicityBlack or African American	-1.9967686	0.3163854	-6.3111908	0.0000000
race_ethnicityChinese	-1.2462630	0.4409127	-2.8265528	0.0047052
race_ethnicityJapanese	-0.8750185	0.6401215	-1.3669568	0.1716388
race_ethnicityOther Asian or Pacific Islander	-0.3811987	0.3324607	-1.1465977	0.2515479
race_ethnicityOther race, or two or more races	-0.6866616	0.3121295	-2.1999250	0.0278122
race_ethnicityWhite	0.1009529	0.2860807	0.3528827	0.7241764
genderMale	0.4181810	0.0605194	6.9098646	0.0000000
education4th to 8th Grade	-0.1505323	1.0199477	-0.1475883	0.8826677
educationAssociate's degree	-0.2660056	0.7586491	-0.3506306	0.7258655
educationBachelor's degree	-0.1644920	0.7545074	-0.2180124	0.8274194
educationCollege, no degree	-0.1520190	0.7552161	-0.2012920	0.8404703
educationDoctorate or professional degree beyond bachelor's	0.3337828	0.7750782	0.4306441	0.6667272
educationHigh school diploma or GED	0.0732293	0.7553939	0.0969419	0.9227725

term	estimate	std.error	statistic	p.value
educationHigh school, no degree	0.0394930	0.7606088	0.0519229	0.9585901
educationMaster's degree	-0.0400684	0.7576753	-0.0528833	0.9578249
census_regionNortheast	-0.0272193	0.0925083	-0.2942358	0.7685777
census_regionSouth	0.3150044	0.0807863	3.8992308	0.0000965
census_regionWest	-0.0395491	0.0906559	-0.4362549	0.6626518

Figure 2: Descriptive statistics of the Biden model regression estimates

term	estimate	std.error	statistic	p.value
(Intercept)	0.4486893	0.7899222	0.5680171	0.5700234
age30-44	-0.2647693	0.0881415	-3.0039113	0.0026653
age45-64	-0.3508453	0.0870682	-4.0295470	0.0000559
age65+	-0.1846063	0.0986350	-1.8716101	0.0612606
race_ethnicityBlack or African American	1.6984331	0.3084324	5.5066624	0.0000000
race_ethnicityChinese	1.3772469	0.4050035	3.4005807	0.0006724
race_ethnicityJapanese	1.1712090	0.5849287	2.0023108	0.0452513
race_ethnicityOther Asian or Pacific Islander	0.7141866	0.3339322	2.1387174	0.0324586
race_ethnicityOther race, or two or more races	0.8609157	0.3149739	2.7332917	0.0062705
race_ethnicityWhite	0.2746155	0.2961888	0.9271638	0.3538415
genderMale	-0.3338525	0.0586088	-5.6962871	0.0000000
education4th to 8th Grade	-0.3362178	0.9731256	-0.3455030	0.7297163
educationAssociate's degree	-0.5585868	0.7486288	-0.7461466	0.4555789
educationBachelor's degree	-0.5155856	0.7447849	-0.6922611	0.4887734
educationCollege, no degree	-0.7173126	0.7454329	-0.9622765	0.3359107
educationDoctorate or professional degree beyond bachelor's	-0.9685347	0.7663998	-1.2637460	0.2063212
educationHigh school diploma or GED	-0.9197077	0.7457687	-1.2332345	0.2174883
educationHigh school, no degree	-0.8402077	0.7503774	-1.1197135	0.2628359
educationMaster's degree	-0.5307612	0.7480319	-0.7095434	0.4779873
census_regionNortheast	0.0379953	0.0898670	0.4227945	0.6724452
census_regionSouth	-0.2128169	0.0787175	-2.7035529	0.0068603
census_regionWest	0.0303012	0.0875508	0.3460982	0.7292689

References

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, C. & Vavreck, L. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814) [dataset]. https://www.voterstudygroup.org/downloads?key=d51ab17d-81fc-4d5b-a912-92945015c722

Therneau, T. (2020). *A Package for Survival Analysis in R*. R package version 3.2-7, https://CRAN.R-project.org/package=survival

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN 0-387-98784-3.

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