Predict American Election Popular Vote Outcome with MLR

Jo-Yen Lin

November 2nd, 2020

Predict American Election Popular Vote Outcome with MLR

Jo-Yen Lin

November 2nd, 2020

Code and data supporting this analysis is available at: https://github.com/jlin213/Predict-American-Election-With-MLR

Model

In this study, we are interested in the popular vote outcome of the 2020 American federal election (Dassonneville et al.). By using post-stratification with survey data from Democracy Fund + UCLA Nationscape and census data from IPUMS USA, we will predict who will win the presidential election.

Model Specifics

The models that will be used are two multilevel logistic linear regression (MLR) models, where one will be modeling proportion of voters who will vote for Donald Trump and another will be modeling the proportion of voters who will vote for Joe Biden. The predictors are race and education, with census region to model the intercept. Mathematically, the model for proportion of voters who will vote for Donald Trump is:

$$y_{Trump} = \beta_{race} + \beta_{education} x_{education} + u_{0j}$$

$$\beta_{race} = r_{00} + r_{01} x_{region}$$

The model for proportion of voters who will vote for Joe Biden is:

$$y_{Biden} = \beta_{race} + \beta_{education} x_{education} + u_{0j}$$

$$\beta_{race} = r_{00} + r_{01} x_{region}$$

The first equation is simplified since there are 7 races and 11 different education levels. y_{Trump} represents the proportion of voters who will vote for Donald Trump and y_{Biden} represents the proportion of voters who will vote for Joe Biden. For β_{race}, r_{00} and r_{01} are the intercept and slope of the random effects term, in which we use different census region to model the intercept. u_{0j} is the random error component for the deviation of the intercept of different census regions from the overall intercept.

Post-Stratification

With post-stratification analysis, we can estimate the different proporitions of voters who will vote for Donald Trump and Joe Biden through with the census data. The cells are divided based on race, education, and census regions. Therefore, each bin is different in race, education, and census regions. Then, by estimating the proportion of voters in each bin, we will weight each proportion estimate based on the population size given by the census data. Then, we will sum all these values and divide by the eniter population size.

Results

Figure 1: Count of different races in the survey data

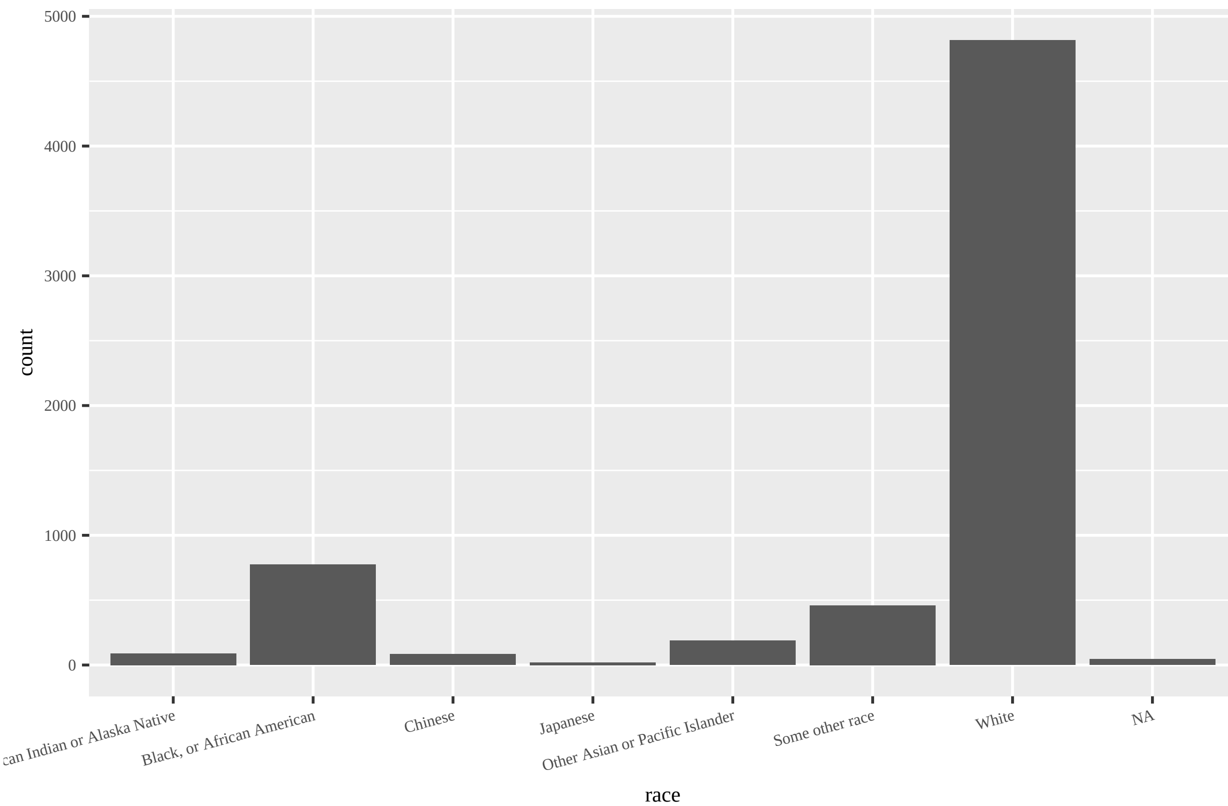

Figure 2: Count of different education levels in the survey data

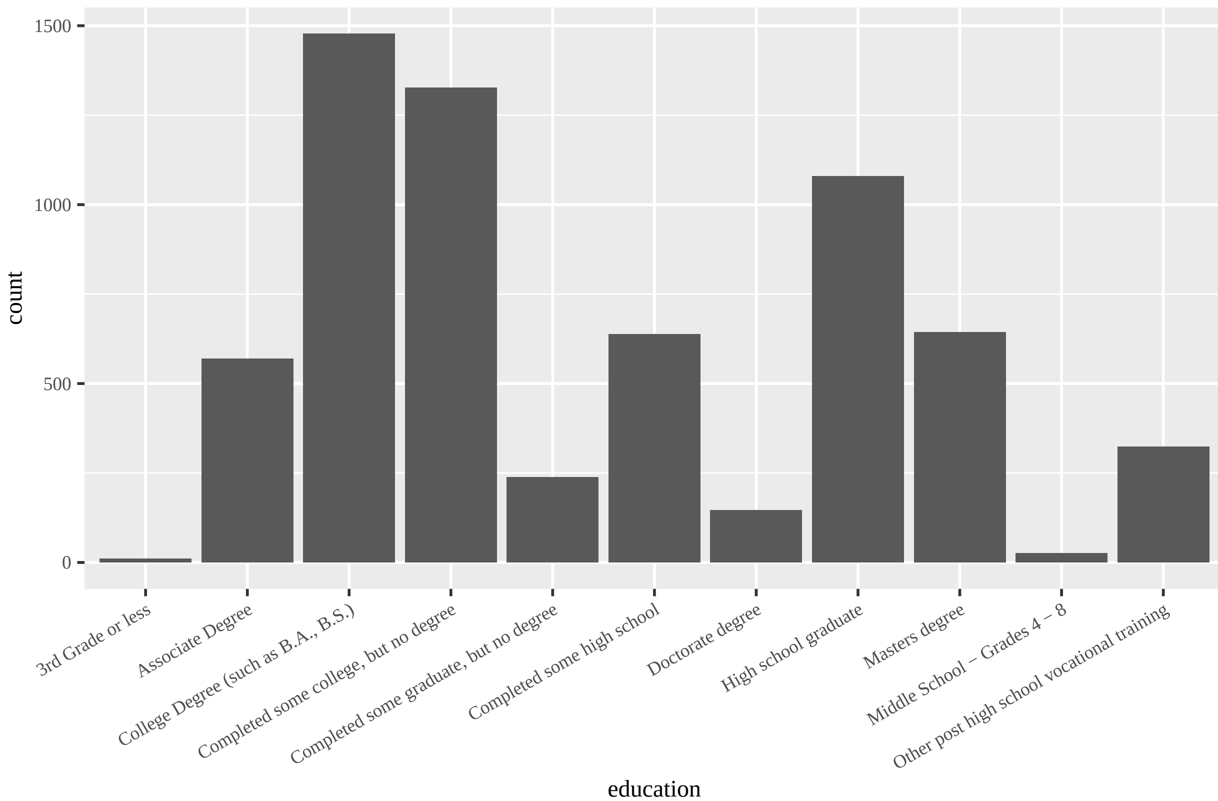

Table 1: Slope and intercept values of the MLR model - Donald Trump

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1141	0.6333	-0.1802	0.8570
raceBlack, or African American	-1.8371	0.2547	-7.2125	0.0000
raceChinese	-1.0972	0.3695	-2.9697	0.0030
raceJapanese	-0.8542	0.5995	-1.4250	0.1542
raceOther Asian or Pacific Islander	-0.6512	0.2817	-2.3115	0.0208
raceSome other race	-0.6283	0.2468	-2.5455	0.0109
raceWhite	0.3413	0.2232	1.5292	0.1262
educationAssociate Degree	-0.5817	0.6124	-0.9499	0.3422
educationCollege Degree (such as B.A., B.S.)	-0.4521	0.6086	-0.7429	0.4575
educationCompleted some college, but no degree	-0.5224	0.6087	-0.8582	0.3908
educationCompleted some graduate, but no degree	-0.3438	0.6213	-0.5533	0.5801
educationCompleted some high school	-0.5595	0.6117	-0.9146	0.3604
educationDoctorate degree	0.1645	0.6304	0.2609	0.7941
educationHigh school graduate	-0.4936	0.6092	-0.8103	0.4178
educationMasters degree	-0.2288	0.6118	-0.3739	0.7085
educationMiddle School - Grades 4 - 8	-0.8242	0.7581	-1.0872	0.2769
educationOther post high school vocational training	-0.2517	0.6166	-0.4082	0.6831

Table 2: Slope and intercept values of the MLR model - Joe Biden

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5942	0.6659	-0.8923	0.3722
raceBlack, or African American	1.5885	0.2486	6.3907	0.0000
raceChinese	1.0055	0.3264	3.0806	0.0021
raceJapanese	1.5630	0.5444	2.8713	0.0041
raceOther Asian or Pacific Islander	0.7682	0.2788	2.7554	0.0059
raceSome other race	0.7282	0.2543	2.8641	0.0042
raceWhite	0.2586	0.2384	1.0847	0.2781
educationAssociate Degree	0.0121	0.6413	0.0188	0.9850
educationCollege Degree (such as B.A., B.S.)	0.0356	0.6379	0.0558	0.9555
educationCompleted some college, but no degree	-0.1705	0.6380	-0.2672	0.7893
educationCompleted some graduate, but no degree	-0.0925	0.6501	-0.1423	0.8868
educationCompleted some high school	-0.5249	0.6408	-0.8191	0.4127
educationDoctorate degree	-0.3826	0.6596	-0.5801	0.5618
educationHigh school graduate	-0.5369	0.6387	-0.8406	0.4006
educationMasters degree	0.0349	0.6411	0.0544	0.9566
educationMiddle School - Grades 4 - 8	-0.2986	0.7579	-0.3940	0.6936
educationOther post high school vocational training	-0.3164	0.6460	-0.4897	0.6244

Table 3: Proportion of voters who will vote for Donald Trump

alp_predict_trump
0.3899

Table 4: Proportion of voters who will vote for Joe Biden

alp_predict_biden
0.4093

From Table 3, we predicted that the proportion of voters that will vote Donald Trump is 0.3899, and the proportion of voters that will vote for Joe Biden is 0.4093. This is based on the the post-stratification analysis mentioned in previous section with the multilevel logistic regression models with race and education as independent variables and census region to model the intercept.

Discussion

From Table 3 and Table 4, we can see that we estimate about 39% of the population will vote for Donald Trump, while about 41% of the population will vote for Joe Biden. That is, from our mathematical model in Model section, y_{Trump} is equal to 0.3899 and y_{Biden} is equal to 0.4093. After our post-stratification analysis, we predict that Joe Biden will win the popular vote of the 2022 American Federal election.

Weaknesses

A significant weakness of the models are how education have high p-value across all levels. From Table 1 and Table 2, each factor level of education seems to have high p-value, that is, they are not as significant factor. In contrast, most of the levels in race are lower than 0.05 significant level. This suggests that choosing education as the independent variable may not have been the right choice. Initially education was chosen due to both data sets have very similar levels, which requires little cleaning. Also, another reason that we considered education as a variable, we are interested if higher education will vote for certain candidate. Another weakness that we can observe is from Figure 2, there is significant population that are white, while most of the other races are covering very little percentage in the survey. While post-stratification takes into account this issue with dividing the census data into cells, and use the proportion to predict the results, we still cannot ignore the fact that the imbalance of population in different races could mean we have many underrepresented population in almost all the races in the data. This could suggests race is not a good independent variable to use in the models.

Next Steps

Next steps of this study include exploring different independent variables in both survey and census data. For example, we can include income as an independent variable. Biden has proposed heavy taxes on those who are high income (400k annual income), as well as imposes higher corporate income tax (Watson et al.). It would be interesting to see if this tax plan influences those who have higher income to vote for Trump. Also, a subsequent survey to collect newer data on would be beneficial, since the surveys are collected on June 25th, 2020. Getting survey data that is closer to the election, would likely to predict the more plausible candidate who will win the election with popular vote, since people are more likely to think through who they are going to vote closer to the election instead of five months before the election.

References

Rohan Alexander. Telling Stories With Data, May 17, 2020. Retrieved from https://www.tellingstorieswithdata.com/

Dassonneville, Ruth, and Charles Tien. "Introduction to Forecasting the 2020 US Elections." PS: Political Science & Politics, 2020, pp. 1–5., doi:10.1017/S104909652000147X.

Garrett Watson, Huaquan Li, Taylor LaJoie, "Details and Analysis of Democratic Presidential Nominee Joe Biden's Tax Plan", Tax Foundation, October 22, 2020.

Marnie Downes, Lyle C Gurrin, Dallas R English, Jane Pirkis, Dianne Currier, Matthew J Spittal, John B Carlin, Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples, American Journal of Epidemiology, Volume 187, Issue 8, August 2018, Pages 1780–1790, https://doi.org/10.1093/aje/kwy070

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [AMERICAN COMMUNITY SURVEY 2014-2018 5-YEAR SAMPLE]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved fromhttps://www.voterstudygroup.org/publication/nationscape-data-set.