The Likelihood of Donald Trump winning the 2020 Presidential Election? A Data Analysis on American Survey and Census Data

root4nothing

2020/1/6

Introduction

Between Republican candidate Donald Trump and Democrat candidate Joe Biden, the presidential election in the United States in 2020 appears to be heating up until November 3rd. It draws attention to the US's recognized economic supremacy and super-powerful political position around the world. In this study, we focus on two datasets that contain data from a poll survey and a demographic census. Hierarchical logistic regression models are used to evaluate any factors that are statistically connected with the chance of voting for Donald Trump based on survey data. We then create a credible election prediction using the census dataset and a trained model.

Data

Democracy Fund and UCLA Nationscape provided the survey dataset for modeling. We processed data cleansing at the beginning in order to model with tidy and ordered data. For example, we only look at those who have registered to vote; we create an age group so that age is a categorized rather than a continuous variable; and we divide family income into three levels: lower, middle, and upper. We have restructured the race ethnicity and educational backgrounds to provide a clearer explanation. The survey dataset contains 3,879 observations and 8 variables after missing values are removed, including both geographic variables like state and demographic variables like gender and age.

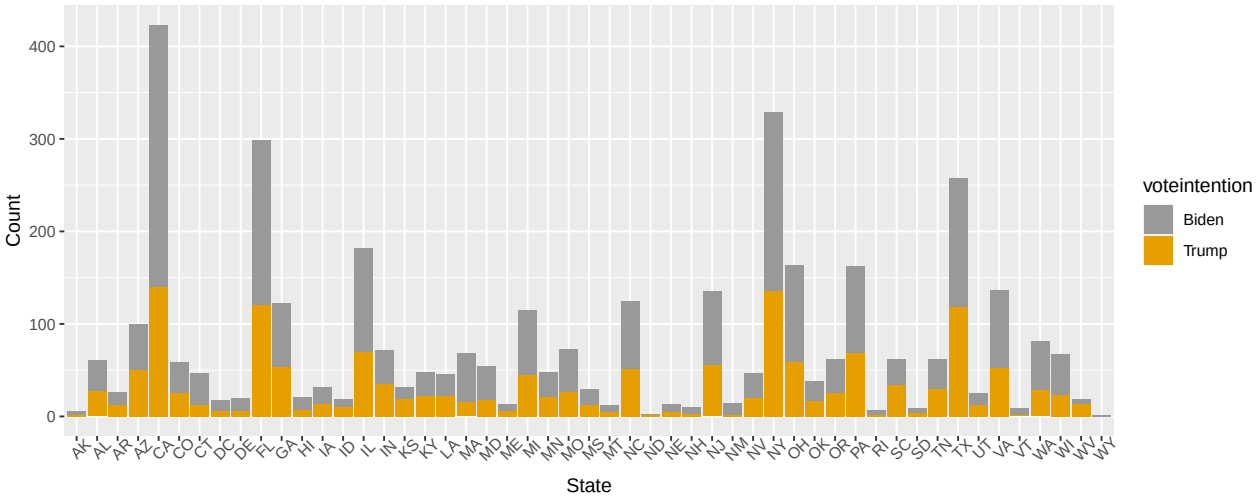

Figure 1: Proportions of Trump and Biden supports across U.S.

Method

Because the response variable we were looking for is binary: either support for Trump or support for Biden. A generalized linear model with a logit link is the rational option for representing this type of data. The Electoral College, on the other hand, determines the presidential election system in the United States, which implies that a candidate who receives the most votes may not win the election. In the 2016 Presidential Election, for example, Hillary Clinton received over 3 million more votes in the final vote tally than Trump, but Trump won the election president. Figure 1 clearly demonstrates that Trump and Biden's support varies substantially across the country; hence, geographic variance should be taken into account for more reasonable and accurate modeling refers to a rate of supporting.

Variance Component Model

The variance components model is a multilevel model without an explanatory variable that can be used to explain variation in supporting for Trump that can be attributed to states. It has the following basic form for a bianry response:

$$Y_{ij} \sim Binomial(\pi_{ij}) logit(\pi_{ij}) = \mathbf{X}_i\beta + \mu_j + \epsilon_{ij}$$

where π_{ij} is the probability of ith individual voting Trump living in jth state, jth is the deviation in probability of voting Trumpof at jth states from average, ϵ_{ij} is the error term.

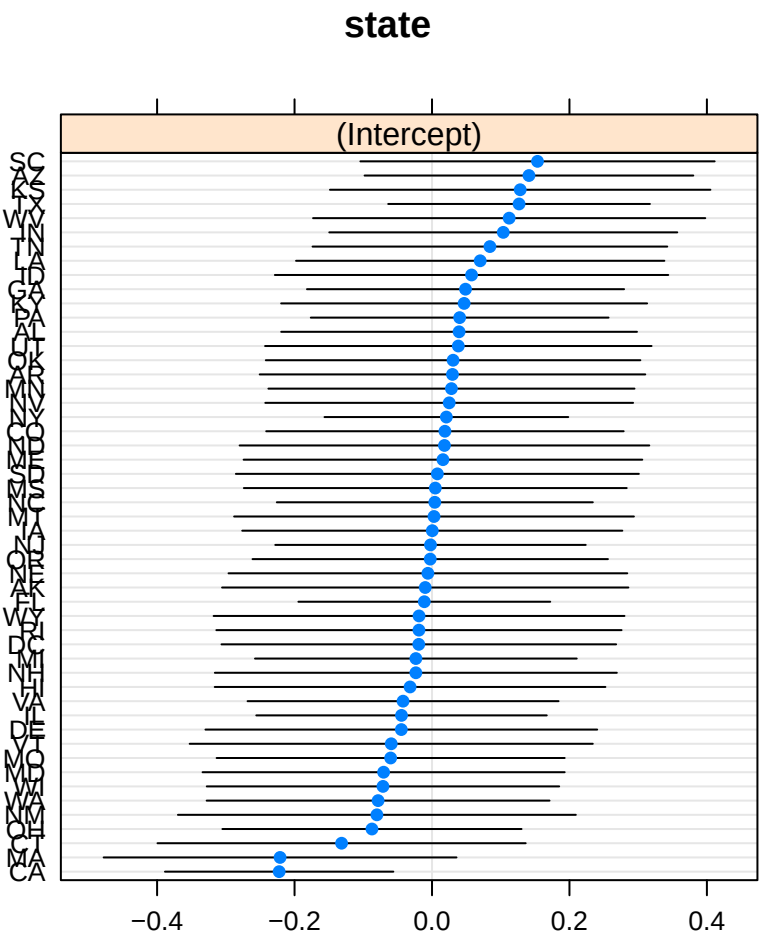

Figure 2: Random effect of states by variance component model

Figure 2 depicts the estimated random effect generated by the variance component model. The graphic evidence does show that the proportion of Trump supporters differs by state. However, we see that the 95% confidence interval overlaps the horizontal line at zero in a large number of states, indicating that the likelihood of backing Trump in these states is neither significantly above nor below normal. A likelihood ratio test comparing the logistic variance component model to a null logistic model, on the other hand, shows that the multilevel model fits the data better than a single level model.

Variable selection

We use Efroymson's method of forward and backward stepwise regression to allocate an optimal combination of explanatory variables for fitting our model (1960). A set of criteria, including as AIC, BIC, and Mallows' Cp, are used to simplify the algorithms used in this selection process. We have a stepwise selected model with variables of gender, age group, race ethnicity, educational background, income level, and working status according to R programming.

Post-Stratification

MRP (multilevel regression with post-stratification) is a prominent and commonly used technique for forecasting election results from polling. The basic idea behind MRP is to estimate the targeted population from a sample population using multilevel regression, which is a convenient way to estimate public opinion across geographic units using individual-level survey data. For the survey and census data we used in this study, all of the factors in the census dataset that we would use for prediction were restructured to match the variables we chose for regression.

Results

Aside from geographic variations in Trump support, we also find that there are changes across multiple levels of variables, such as random effects on gender, ethnicity, age group, and even income level, as seen in Figure 3.

In addition, the GLMM model results in Tables 1 and 2 reveal that gender, age group, race, income level, and working status are all significantly connected to the chance of voting for Trump. For example, the gender coefficient is calculated negatively, meaning that male voters are more excited about Trump than female voters. The likelihood of male voters supporting Trump is equal to the probability of female voters multiplied by a factor of. Similarly, Trump supporters are more likely to be older persons with a higher income (household income $>$ \$129,999). Furthermore, there are differences in support favor among race ethnicities: When compared to Black Americans, the likelihood of voting for Trump in White Americans are multiples of that of Black Americans.

As a result, we apply this GLMM regression to the cleaning census dataset as a method of post-stratification and find that the proportion of voters in favor of voting for Donald Trump is 35.78 percent, implying that Trump's chances of winning this presidential election are not promising if only the polling results on the survey and census data are taken into account. However, as previously stated, the number of electoral districts, rather than the overall proportion of supported voters, is the key to winning an election in the United States.

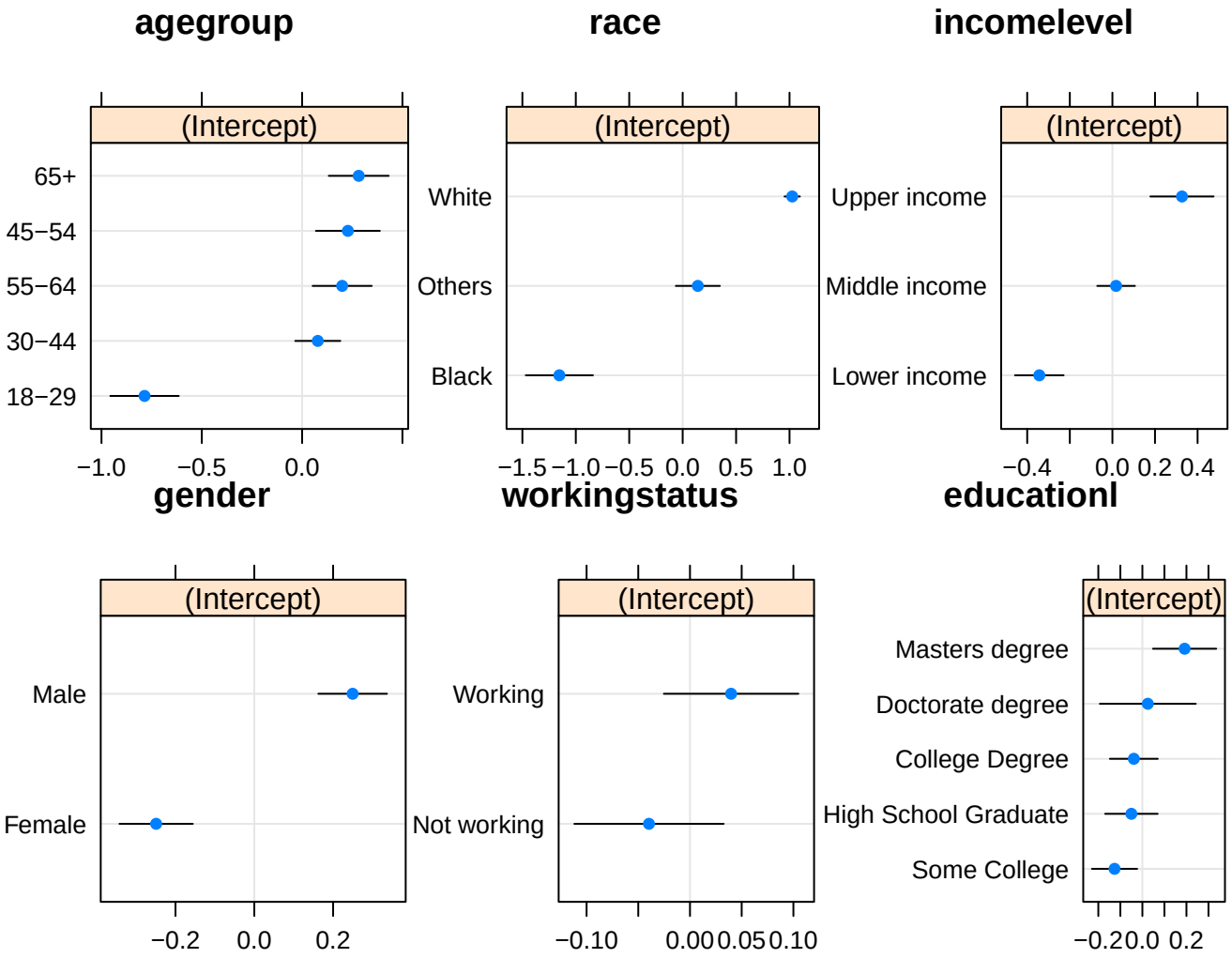

Table 1: Coefficient estimation of the GLMM

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.527	0.222	-15.915	0.000
genderMale	0.333	0.072	4.608	0.000
agegroup30-44	0.656	0.116	5.659	0.000
agegroup45-54	0.856	0.131	6.543	0.000
agegroup55-64	0.822	0.127	6.493	0.000
agegroup65+	0.896	0.135	6.651	0.000
raceOthers	1.427	0.202	7.062	0.000
raceWhite	2.100	0.174	12.036	0.000
educationlDoctorate degree	0.094	0.240	0.391	0.696
educationlHigh School Graduate	0.456	0.108	4.223	0.000
educationlMasters degree	0.086	0.115	0.741	0.458
educationlSome College	0.099	0.092	1.072	0.284
incomelevelMiddle income	0.202	0.088	2.302	0.021
incomelevelUpper income	0.459	0.120	3.830	0.000
workingstatusWorking	0.215	0.084	2.559	0.011

Table 2: MLE's of baseline odds and odds ratios with 95% confidence intervals

	est	2.5	97.5
Baseline	0.029	0.019	0.045
genderMale	1.395	1.211	1.606
agegroup30-44	1.926	1.535	2.418
agegroup45-54	2.353	1.821	3.040
agegroup55-64	2.276	1.776	2.917
agegroup65+	2.450	1.882	3.191
raceOthers	4.168	2.805	6.194
raceWhite	8.168	5.802	11.499
educationlDoctorate degree	1.098	0.686	1.758
educationlHigh School Graduate	1.578	1.277	1.950
educationlMasters degree	1.089	0.869	1.366
educationlSome College	1.104	0.921	1.323
incomelevelMiddle income	1.224	1.030	1.453
incomelevelUpper income	1.582	1.251	2.001
workingstatusWorking	1.240	1.052	1.463

Discussion

As a result, we apply this GLMM regression to the cleaning census dataset as a method of post-stratification and find that the proportion of voters in favor of voting for Donald Trump is 35.78 percent, implying that Trump's chances of winning this presidential election are not promising if only the polling results on the survey and census data are taken into account. However, as previously stated, the number of electoral districts, rather than the overall proportion of supported voters, is the key to winning an election in the United States. However, given to the election laws, this does not rule out the possibility of Trump winning.

Although we discovered several fascinating findings, as we described above, there are numerous flaws in our study that should be noted and rectified for future research. Because GLMM is a straightforward frequentist model, which is estimated by the likelihood function on fixed supplied parameters, the first issue we evaluate is the efficiency and correctness of our hierarchical logistic regression model. However, because the willingness to support a candidate is a personal emotional experience that is difficult to model, a Bayesian study may be better appropriate for thoroughly evaluating the possibility of Trump's support. Furthermore, because the census data we collected did not include regional household income information, a post-stratification analysis was used to forecast the model without these two components. We need to make a precise prediction based on a more specific census data.

In order to conduct future research, we plan to summarize our expected fraction of Trump voters by state and compare our predictions after the results of the 2020 US Presidential Election are known.

Bibliography

[1] Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(3), 659-687.

[2] Harrell, F. E. (2001) "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis," Springer-Verlag, New York.

[3] Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification without population level information on the poststratifying variable with application to political polling. Journal of the American Statistical Association, 96(453), 1-1

[4] Buttice, M. K., & Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys?. Political analysis, 449-467.

[5] Anuta, D., Churchin, J., & Luo, J. (2017). Election bias: Comparing polls and twitter in the 2016 us election. arXiv preprint arXiv:1701.06232.