Will Donald Trump win the 2020 Presidential Election? A statistical analysis on American survey and census data.

Siyi Ma, Heye Liu, Yan Wang, Yuanzhe Yang

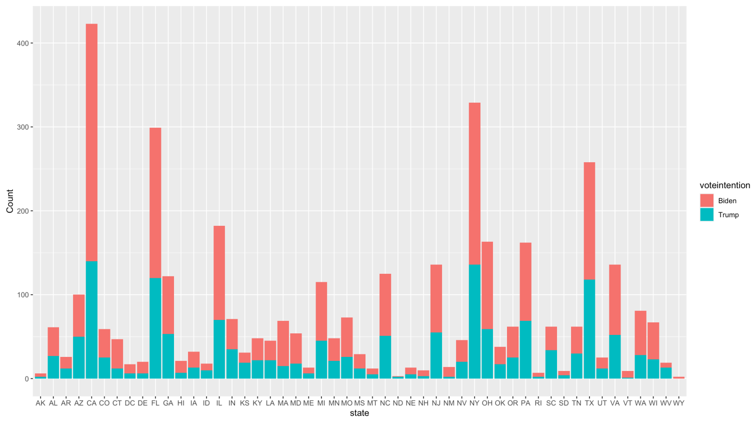
11/01/2020

Introducion

The 2020 U.S. Presidential race between Republican candidate Donald Trump and Democrat candidate Joe Biden appears to be heating up until Election day on November 3. It attracts all the attentions of the world regarding the esteemed economic strength and super-powered political position of U.S. In this study, we focus on two datasets containing information from a poll survey and demographic census. GLMMs are conducted on survey data for investigating factors that are statistically related with the probability of voting for Donald Trump. Through the method of post-stratification, we find that the proportion of voters in favor of supporting for Donald Trump is only 35.78%, however, it does not mean that Donald Trump has no chance for serving his consecutive terms since the key of winning the U.S. election is the number of electoral district rather than the overall percentage of supported voters

The survey dataset for modeling is original from Democracy Fund + UCLA Nationscape. For modeling with neat and organized data, we have processed data cleaning at initial. For instance, we only focus on individuals who have registered for voting; we create an age group such that age is a categorized variable instead of a continuous one; the household income is classified with three levels, lower, middle and upper. Additionally, we also reformat the race ethnicities and education background for more explicit explanation. After removing missing values, the survey dataset contains 3,879 observations and 8 variables, including both geographic variables like state, and demographic variables, gender and age.

Figure 1. Support proportions of Trump and Biden across U.S.



Method

Since the response variable we targeted is distichous: either supports Trump or supports Biden. The intuition for modeling this typical data is a generalized linear model with a logit link. However, the system of Presidential Election in U.S. is determined by the Electoral College, which means that a candidate who obtains the most votes may not win the election. For instance, In the 2016 Presidential Election, Hillary Clinton had nearly 3 million more votes than Trump in the final vote count, but Trump became the president. In Figure 1, it also shows that the supporting proportion of Trump and Biden do vary dramatically across the whole country, therefore, the geographic variation should be taken in to consideration for more reasonable and accurate modeling refers to a rate of supporting.

Variance Component Model

Variance components model is a multilevel model without explanatory variable which could be used to explain the variation in the tendency of supporting Trump that can be attributed to the states, for a bianry response, it has a basic form as follows:

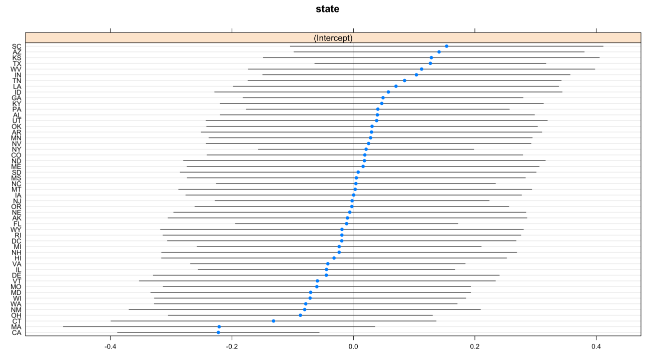
$$Y_{ij} \sim Bernoulli(\pi_{ij})$$

$$\text{logit}\left(\pi_{ij}\right) = \boldsymbol{X}_{ij}\beta + u_j + \varepsilon_{ij}$$

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.

The estimated random effect generated by variance component model are shown in Figure 2. The graphic evidence does indicate that the proportion of supporting Trump varies from state to state. However, we notice that, for a substantial number of states, the 95% confidence interval overlap the horizontal line at zero, indicating that likelihood of supporting Trump in these states are not significantly above average or below average. On the other hand, a likelihood ratio test between the logistic variance component model and a null logistic model does show that the multilevel model does show a significant improvement in model fitting than a single level model.

Figure 2. Random effect of states by variance component model



Variable selection

In order to allocate an appropriate combination of explanatory variables for fitting our model, we apply the method of forward and backward stepwise regression proposed by Efroymson (1960). The algorithms within this selection process is straightforward by a series of criteria, such as AIC, BIC and Mallows's Cp. By applying R programming, we have a stepwise selected model with variables of gender, age group, race ethnicities, educational background, income level and working status.

Post-Stratification

The multilevel regression with post-stratification (MRP) is a technique that is prevalent and frequently applied in predicting election results from polling. The fundamental idea of MRP is estimate the targeted population from the multilevel regression on a sample population, which is a convenient method to estimate public opinion across geographic units from individual-level survey data. For survey and census data we performed in this study, all the variables in census dataset we would utilize for prediction are reorganized to keep consistency with the ones we selected for regression.

Results

Other than geographic variation in the rate of supporting Trump, we also find that there are variations across different level of variables, i.e., random effects on gender, race, age group and even income level are substantially detected from Figure 3.

Furthermore, the regression results of GLMM in Table 1 and Table 2 show that, gender, age group, race, income level and working status are statistically related with the likelihood of voting for Trump. For instance, the coefficient of gender is negatively estimated, implying that men voters are more enthusiastic for Trump than females. Specifically, the probability of supporting Trump for men voters is equal to the odds of female voters multiple a factor by $\exp(0.333) = 1.395$. Likewise, older people with upper income (household income > $129,999) are more likely to be a fan of Trump. Additionally, there are divergence of supporting favor within race ethnicities: Comparing with Black American, the odds of voting for Trump in White American is overwhelming by multiple a factor of $\exp(2.100) = 8.168$ on that odds of Black American.

Consequently, we apply this GLMM regression to the cleaning census dataset as the method of post-stratification, find that the proportion of voters in favor of voting for Donald Trump is 35.78%, implying that the chance of winning this presidential election for Trump is not promising if we barely consider the polling results on the survey and census data. However, as we mentioned above, the key of winning the U.S. election is the number of electoral district rather than the overall percentage of supported voters

Figure 3. Heterogeneous effects of variables on the probability of supporting Trump

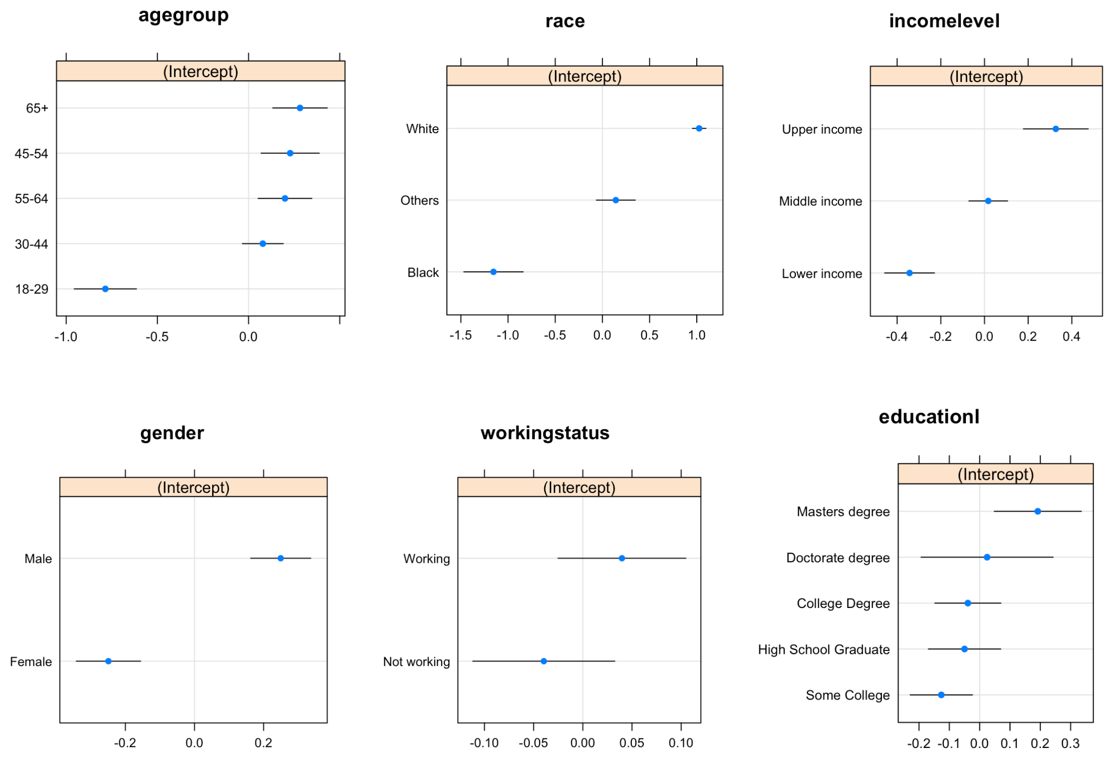


Table 1. Coefficient estimation of the GLMM

	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

In this study, we focus on two datasets containing information from a poll survey and demographic census. Variance component models and General linear multilevel models are conducted for detecting geographic variations and factors that are associated with the probability of voting for Donald Trump. The graphic and numeric evidence suggest that the variation in probability of voting for Trump changes across states, gender, age, income and education level, working status and race ethnicities are significantly related with the proportion of supporting Trump. Specifically, a white male American who have higher education and income level tend to vote for Trump rather than Biden. The post-stratification analysis on census dataset based on the trained model we applied provide a predicted overall proportion of voting for Trump is only 35.78%. Nevertheless, it does not mean that there is no chance of winning for Trump due to the electoral rules.

Although some interesting facts we have found as we mentioned above, several drawbacks of this study should be notified and modified for further study. The first issue we consider about is the efficiency and accuracy of our hierarchical logistic regression model, Since GLMM is a straightforward frequentist model, which is estimated by the likelihood function on fixed given parameters. However, the willingness of supporting is a personal feeling and way to complicated on modeling, a Bayesian analysis might be more appropriate for thoroughly investigating the likelihood of supporting Trump. For further study, we consider to summarize our predicted proportion of voting Trump by states and compare our prediction once the outcome of the 2020 U.S. Presidential Election is revealed.

Reference

[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), 1-11.

[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.