To Predict the Result of 2020 American Federal Election

Yifei Zhang, Ziyi Qu, Peilin Chen, Ziru Nie

11/02/2020

To Predict the Result of 2020 American Federal Election

Yifei Zhang, Ziyi Qu, Peilin Chen, Ziru Nie

11/02/2020

Code and data supporting this analysis is available at: https://github.com/EmiChen/STA304-A3-Group58.git

Model

The goal of our study is to predict overall popular vote of the 2020 American federal election. We construct a multilevel regression with the technique of post-stratification. We choose vote_trump(binary) to be our response variable, and thus choose to run logistic regression. We also select age(numerical), gender(binary), race(binary), and education(binary) to be our level 1 explanatory variables, with laborforce (binary) being the level 2 group variable. After building up a multilevel regression using survey data from Voter Study Group (Tausanovitch et al, 2020), we perform post-stratification using census data collected from IPUMS (Ruggles et al, 2020): partitioning the population into cells, then using model from sample to estimate y value per cell, and lastly aggregating cell-level values by weighting each cell by relative proportion in population. In order to predict the final result of the election, we also perform a similar MRP analysis with response variable being vote_biden, after which we will compare two $\hat{y}^{PS}$ values and thus accomplish the goal of our study. In the following sessions, we will describe, in details, the model selection, model specifics, and the post-stratification technique.

Model Specifics

We use a multilevel regression model, with logistic regression being the main regression model and random coefficient being the multilevel technique, to predict the proportion of voters who will vote Donald Trump in the upcoming 2020 U.S election. The software that we use to perform statistical analysis is R. The multilevel regression model that we build looks like:

Individual level & Group Level:

$$log(\frac{p}{1-p}) = \beta_{0,laborforce} + \beta_1 x_{age,laborforce} + \beta_2 x_{gender,laborforce} + \beta_3 x_{race,laborforce} + \beta_4 x_{education,laborforce} + \epsilon$$

$$\beta_{0,laborforce} = \alpha + a_{laborforce}$$

The first explanatory variable that we choose is age, which yields numerical response about the voter's age. The reason why we choose age is that scholars (Holland, Jenny Lynn, 2013) found young voters prefer more liberal candidates than their elder counterparts do and young voters prefer candidates with challenging and extreme ideologies while elder voters prefer conservatism and traditions. The second explanatory variable that we choose is gender, which yields binary response about the voter's gender. The reason why we select this variable is that researchers found more than 40% of women voted for Trump in 2016, although the majority of Trump voters were male (Setzler & Yanus, 2018), and such fact indicates that gender may be highly correlated with voters' decisions to support Donald Trump. The third variable that we choose is race, which yields binary response about whether the voter is a white. The reasons why we choose whether a vote is a white instead of whether a voter is black or yellow are that white is the racial majority in the U.S. and that we hypothesize white people tend to have more shared feelings with Donald Trump and thus have greater possibility to support him. The fourth explanatory variable that we choose is education, which yields binary response on whether the voter has completed college education. We select college instead of high school because researchers have found that whether having a college degree is the watershed that decides the voting tendency for Trump (Tamari et al, 2020).

We choose laborforce to be our level 2 group factor, and choose to perform random intercept. We count Full-time employed, Unemployed or temporarily on layoff, Part-time employed, Self-employed as part of the laborforce, while we count Homemaker, Retired, Permanently disabled, Student and Other as not in the laborforce. It means that if voters move between two groups(in the laborforce vs not in the laborforce), the intercept of the estimated regression line will be different. It is the technique that we use to deal with clustered or grouped data.

Our response variable is vote_2020 and yields binary response indicating whether this voter will vote for Donald Trump, which is the reason why we select logistic regression. β_0 represents the possibility of voting for Donald Trump if this voter is at age 0, a female, not a white, and has a college degree or above, whose value will be changed if the voter moves between two groups of laborforce. Additionally, β_1 means that a voter will be β_1 higher in log odds of voting for Donald Trump than a voter who is one year younger than him while holding other factors the same. $\beta_2 \sim \beta_4$ have similar meanings. For example, β_2 represents the additional log odds of voting for Donald Trump for male than female while holding other characteristics of the two voters the same.

```r
# Creating the Model
model <- glmer(vote_trump ~ age + gender + race + education
               + (1|laborforce), family = binomial, data=survey_data)
```

To see if our original model, "model", is a good choice, our group will do model diagnostics on model with random intercepts and an alternative model, model_alt, with random coefficient. Specifically, our group create an alternative model called model_alt where the coefficient of age varies among laborforce. The model diagnostics methods our group use are AIC, BIC and AUC.

```r
# The alternative model used to do diagnostics
model_alt <- glmer(vote_trump ~ gender + race + education + (age |laborforce),
               family = binomial, data=survey_data)
```

```
## boundary (singular) fit: see ?isSingular
```

Basically, AIC (Akaike information criterion) and BIC (Bayesian information criterion) are penalized-likelihood criteria, which add a penalty for including more predictors. The lower AIC and BIC indicates the closeness of our logit multilevel model to the true model. AUC (Area Under the ROC Curve) serves as a global measure of diagnostic accuracy, helping to estimate how high is the discriminate power of a test. Ideally, we would want a AUC above 0.5 to show the accuracy of our model.

By conducting AIC, BIC and AUC on both "model" and "model_alt", we find out that "model" has a lower AIC(6691.841<6701.175), a lower BIC(6731.246<6747.148) and a higher AUC(0.6527 > 0.6507), according to which we infer that our original model with random intercept is better and more suitable.

model	AIC	BIC	AUC
model	6691.841	6731.246	0.6527
model_alt	6701.175	6747.148	0.6507

Post-Stratification

After building up multilevel regression models, we need to perform a post-stratification analysis in order to estimate the proportion of voters who will vote for Donald Trump (Note: We also did the same analysis on vote_biden, but it is only for comparison purpose. We will focus on vote_trump for this part). Here we create cells based off the combination of five factors: age, sex, education, race, and laborforce. There are 80 subgroups for age (from 18 to 97), 2 subgroups for gender, 2 subgroups for education, 2 subgroups for race and 2 subgroups for laborforce. Thus, there are total $80 * 2 * 2 * 2 * 2 = 1280$ cells(groups) from our selected data. We will then calculate the proportion of voters voting for Trump in each cell, and aggregate cell-level values by weighting each cell by relative proportion in population.

To do so, we first need to do some cleaning in order to qualify the data we use in building our models. For instance, we filter the survey data by "registration" to only include people who are eligible to vote. In order to match this criteria, we used filter by "age >= 18" in census data cleaning, assuming only people above 17 are registered to vote. we also need to match the variable names in census_data and survey_data by mutating the corresponding data. For example, we rename variable "sex" in census data to "gender" and change "male" and "female" to "Male" and "Female" respectively in order to match the variable name in survey data.

Results

From the regression table, we find that $\hat{\beta}_1$ is 0.0126, $\hat{\beta}_2$ is 0.3945, $\hat{\beta}_3$ is 1.1676, and $\hat{\beta}_4$ is 0.1478, which are all significant at 5% significance level.

```
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
## # A tibble: 6 x 7
##   effect   group      term                estimate std.error statistic   p.value
##   <chr>    <chr>      <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 fixed    <NA>       (Intercept)          -2.13     0.142      -15.0   1.00e-50
## 2 fixed    <NA>       age                   0.0126   0.00204      6.20  5.58e-10
## 3 fixed    <NA>       genderMale            0.395    0.0597       6.61  3.92e-11
## 4 fixed    <NA>       raceWhite             1.17     0.0762      15.3   5.74e-53
## 5 fixed    <NA>       educationunder col~   0.148    0.0684       2.16  3.07e- 2
## 6 ran_pars laborforce sd__(Intercept)       0.102   NA           NA    NA
```

We get a $\hat{y}^{PS}$ value of 0.4328 and thus estimate that the proportion of voters in favor of voting for Donald Trump is 0.4328. We get this value from the post-stratification analysis of the proportion of voters in favor of Donald Trump modeled by a multilevel logistic regression model with laborforce being group factor, which accounts for age, gender, race and education.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.433
```

Discussion

The conclusion that we draw from the study is that Donald Trump stands a good chance to be elected as the next president, and the elder, males, white people, and people with education level lower than college are more likely to vote for Trump than base groups in 2020 election.

From model regression table, we discover that the elder, males, white people, and people with education level below college level are more likely to vote for Donald Trump in 2020 election. For example, $\hat{\beta_1}$ being 0.0126 means that a voter will be 0.0126 higher in log odds of voting for Donald Trump than a voter who is one year younger than him while holding other factors the same. Another example is that $\hat{\beta_3}$ being 1.1676 means a white will have 1.1676 more in log odds of voting for Donald Trump than a non-white voter while holding other factors the same. This discovery fits our hypothesis that white people have larger tendency to vote for Donald Trump

The overall probability that Donald Trump get elected is 43.28%. For comparison, after doing the above regressions again for Joe Biden, we find that the probability that Biden get elected is 43.05%, which is lower than that for Trump. Therefore, we may draw the conclusion that the result of the 2020 American federal election is that Donald Trump will continue as president.

Weaknesses

The first weakness is associated with our data cleaning process. Among the options under race variable, we assume that "two major races" and "three or more major races" do not count toward "white". We have to make such biased assumption since we lack access to more detailed data behind those two subgroups. In addition, our model would be more accurate and representative if we could involve more other relevant variables such as vote intention.

Second, the steps we follow to predict the probability of voting is not as accurate and specific as the real voting process in the United States, where we ignore the impact of states. Precisely, the actual procedure is where the final voting decision is based on a group called "electoral college" consisting of 538 representative electors from fifty states and Washington, D.C, instead of on all individuals in the US, .

Moreover, the model could include the geographical categories because the number of swing states are not particularly low (Glaeser et al, 2006). According to Skelley et al, the forces that drove Trump to win in 2016 are still prevalent in 2020. We witnessed Hillary Clinton seemed to have a solid lead in 2016, however, Donald Trump eventually won.

Another drawback is that we have relatively high AIC and BIC. We assume that a better regression models can be developed by better techniques with acquiring more knowledge on statistics analytical tools, and by constructing a more specific model.

Next Steps

The first suggestion is to involve more potentially relevant and statistically significant demographic variables in our model. The aspects of a registered and eligible voter that could impact its voting decision in our model now are education, work, and some basic demographic measures. Therefore, our model could be more comprehensive and thorough if we have included more common variables regarding a voter's other aspects, for instance, marital status, religious beliefs, occupation and so on.

Second suggestion is to include more advanced diagnosis methods, and confusion matrix is one of them. Confusion matrix is a metric measuring the performance of a classification model, which compares the actual results with the predicted results and visualizes the comparison in a table. One of the biggest advantages of this diagnosis is that it not only gives you insights into the errors being made but also identify the type of errors made by you.

Last, since we do not consider the "state" effect and the real and complex president voting procedure in the United States, our original model can be further improved adding other relevant variables such as "state". We can also research more on the specific voting data collection methodology in each region of the United States to include more complex calculations, and to make our predicting process on the winning probability of candidates more realistic.

References

Glaeser, E. L., & Ward, B. A. (2006). Myths and realities of american political geography. Journal of Economic Perspectives, 20(2), 119–144. https://doi.org/10.1257/jep.20.2.119

Halligan, S., Altman, D. G., & Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. European Radiology, 25(4), 932–939. https://doi.org/10.1007/s00330-014-3487-0

Holland, Jenny Lynn(2013). Age Gap? The Influence of Age on Voting Behavior and Political Preferences in the American Electorate.https://research.libraries.wsu.edu/xmlui/handle/2376/4982

Tamari, J. (n.d.). How Biden's lead is different from Clinton's—And why the polls are different this time. Https://Www.Inquirer.Com. Retrieved October 30, 2020, from https://www.inquirer.com/politics/election/trump-biden-2020-pennsylvania-polls-clinton-2016-20201028.html

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=b6ec5a2c-8755-4293-9095-306f369fee3e].

Setzler, M., & Yanus, A. B. (2018). Why did women vote for donald trump? PS: Political Science & Politics, 51(3), 523–527. https://doi.org/10.1017/S104909651800035.

Skelley, E. M., Geoffrey. (2020, August 26). Is the electoral map changing? FiveThirtyEight. http://projects.fivethirtyeight.com/swing-states-2020-election/

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

McCarthy, Devin (November 22, 2014). "How the Electoral College Became Winner-Take-All". https://en.wikipedia.org/wiki/United_States_Electoral_College