2020 American federal election voting prediction

using multilevel regression with post-stratification

Z.Zhu

02 November 2020

Abstract

|Use logistic regression to model the history votings for Donald Trump using the survey_data.Then perform the post-stratification analysis based on this model to predict the future votings for Donald Trump. | | **Keywords:** logistic regression; post-stratification; forecast

Contents

1 Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation. Also, we use R Core Team [2020] to perform some statistical analysis in R.

1.1 Model Specifics

I will be using a logistic regression model to model the proportion of voters who will vote for Donald Trump. I will be using age, which is recorded as a numeric variable, sex, race and education, which are recorded as categorical variables, to model the probability of voting for Donald Trump. The logistic regression model I am using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{male} + \beta_3 x_{white} + \beta_4 x_{college} + \epsilon$$

Where p represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, and is the probability of non-white female without a college degree voting for Donald Trump at age 0. Additionally, β_1,β_2,β_3 and β_4 represent the coefficients of each corresponding variables. So, for everyone one unit increase in age, we expect a e^{β_1} increase in the odds $(\frac{p}{1-p})$ of voting for Donald Trump.

Similarly for the categorical variables, one unit increase, we expect a e^{β_1} increase in the odds ($\frac{p}{1-p}$) of voting for Donald Trump.

However, after seeing the summary output of the above model, we removed education from our model. We will discuss it in details later in the results section. So, here is our final model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{male} + \beta_3 x_{white} \epsilon$$

1.2 Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Post-stratification involves adjusting the sampling weights so that they sum to the population sizes within each post-stratum. Here I create cells based off different ages, genders and races. Using the model described in the previous sub-section I will estimate the proportion of voters in each age bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size. In the end, I should get an estimated prediction of the percentage of people voting for Donald Trump with decreased bias.

```r
census_data <- census_data %>%
  rename(gender = sex) %>%
  mutate(gender = ifelse(gender=="male","Male","Female"))

census_data <- census_data %>%
  rename(race_ethnicity = race) %>%
  mutate(race_ethnicity = ifelse(race_ethnicity=="white","White","Some other race"))


# Here I will perform the post-stratification calculation
census_data$logodds_estimate <-
  model2 %>%
  predict(newdata = census_data,type = "response")

census_data$estimate <-
  exp(census_data$logodds_estimate)/(1+exp(census_data$logodds_estimate))

census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict = sum(alp_predict_prop)/sum(n))
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.598
```

2 Results

```
## # A tibble: 27 x 5
##    term                                     estimate std.error statistic p.value
##    <chr>                                       <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)                                -0.674     0.654    -1.03    0.302
##  2 educationAssociate Degree                  -0.737     0.633    -1.16    0.244
##  3 educationCollege Degree (such as B.A., ~   -0.594     0.629    -0.944   0.345
##  4 educationCompleted some college, but no~  -0.620      0.629   -0.986    0.324
##  5 educationCompleted some graduate, but n~  -0.577      0.642   -0.900    0.368
##  6 educationCompleted some high school       -0.472      0.632   -0.747    0.455
##  7 educationDoctorate degree                 -0.0483     0.651   -0.0742   0.941
##  8 educationHigh school graduate             -0.557      0.630   -0.885    0.376
##  9 educationMasters degree                   -0.417      0.632   -0.659    0.510
## 10 educationMiddle School - Grades 4 - 8     -0.833      0.781   -1.07     0.286
## # ... with 17 more rows
```

The above is the summary output of the first model. From the output, we can see that all the p-values of t-test for the variable education is above 0.05. This means we do not have enough evidence to include education as a significant variable in our model. Therefore, we could remove the education from our prediction model.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.598
```

From the above, I estimate that the proportion of voters in favor of voting for Donald Trump to be 0.598. I used a logistic model with age, sex and race as variables to perform a post-stratification analysis of the proportion of voters in favor of Donald Trump. The mathematical theory behind this post-stratification analysis is demonstrated as the equation below:

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

where $\hat{y}$ is the proportion of voters who will vote for Donald Trump.

3 Discussion

So, I have fitted a logistic regression to model the voting for Donald Trump. Also, I have successfully predicted the voting for Donald Trump in the upcoming 2020 election based on this model.

Therefore, it is time to give my conclusion. Based off the estimated proportion of voters in favor of voting for Donald Trump being 0.598, I predict that Donald Trump will win the election.

3.1 Weaknesses

However, the above analysis exists certain problems. Firstly, the number of predicting variables in the logistic regression seems to be too small. We could add more variables in to stabilize the model. Secondly, since there are too many race categories in the race variable, I simplified the procedure and only chose to re-categorize the race into two categories (white and some other race).

3.2 Next Steps

For the first problem, we could definitely add all the variables available in the dataset to logistic regression model. Then we could perform a model selection using AIC to choose the best reduced model. This could generate a better model than the one we used. For the second question, if time permitted, we should re-categorized race into every race categories in the dataset. And this should gave us a more precise result.

References

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.