Use a frequentis multilevel logistic regression model (random intercept) to predict 2020 American Federal Election result and post-stratification to verify the model

Problem Set#3

Junming Zhang, Hairong Sun, Xiaoxi Bai, Yangyang Liu

Monday, November 02, 2020

Abstract

This report focused on using a frequentis multilevel regression model (random intercept) to predict if Donald Trump or Joe Biden can be selected as 2020 USA president. In order to build the model, we use a survey dataset (Tausanovitch, Chris and Lynn Vavreck, 2019)(Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek, 2020) to build the model and a census dataset (PUMS USA, 2020) to predict who will be elected for post stratification purpose. Our model predicts that Joe Biden will be elected with significant superiority with respect to electoral votes, and we discussed the result. However, since there are some drawbacks in our model, we also discuss the weakness and how we can improve it.

key words: USA 2020 election, Donald Trump, Joe Biden, multilevel logistic regression model, post-stratification, prediction

Please click **"here"** to access the GitHub repository for all work.

Model

In this model, we are predicting the vote outcome of the 2020 American federal election by employing random intercept model and post-stratification technique. In the following sub-sections we will describe the logistic model regression specifics and the post-stratification calculation.

Model Specifics

We will be using a random intercept model to model the voters who will vote for Donald Trump. We will be using sex, age_group, race, hispan, education, state, and vote_trump to model the probability of voting for Donald Trump. The logistic regression model we are using is:

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 * sexmale + \beta_2 * age\_group18-20 + ... + \beta_{18} * educationtertiary\ \ (not\ \ bachelor) + \epsilon$$

Where p is the probability of Donald Trump got selected. (p/1-p) is the odds of trump winning the election. β_0 represents intercept of the model which is 2.10152. Additionally, β_1 to β_{18} represent the slope of the model, and relate to each variables. β_1 = 0.41511, β_2 = -1.62472, ..., β_{18} = -0.45434 respectively according to the summary data. So, for example, for everyone one unit increase in sexmale, we expect a β_1 increase in the probability of voting for Donald Trump.

```
##       sex             age_group              race               hispan
##   Length:3467        Length:3467        Length:3467        Length:3467
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   education            state             vote_trump
##  Length:3467        Length:3467        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :1.0000
##                                        Mean   :0.5097
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
```

```
##      sex         age_group                                race
##  female:1591   >= 70  :381   american indian or alaska native:  24
##  male  :1876   18 ~ 20: 23   asian or pacific                : 136
##                20 ~ 30:328   black                           : 366
##                30 ~ 40:697   other                           : 174
##                40 ~ 50:701   white                           :2767
##                50 ~ 60:590
##                60 ~ 70:747
##          hispan                         education        state
##  cuban       :  19   at most high school    : 584   CA     : 376
##  mexican     : 238   bachelor               :1620   NY     : 314
##  not hispanic:3059   graduate               : 757   FL     : 292
##  other       : 148   tertiary (not bachelor): 506   TX     : 226
##  puerto rican:   3                                  IL     : 157
##                                                     OH     : 155
##                                                     (Other):1947
##    vote_trump
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5097
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

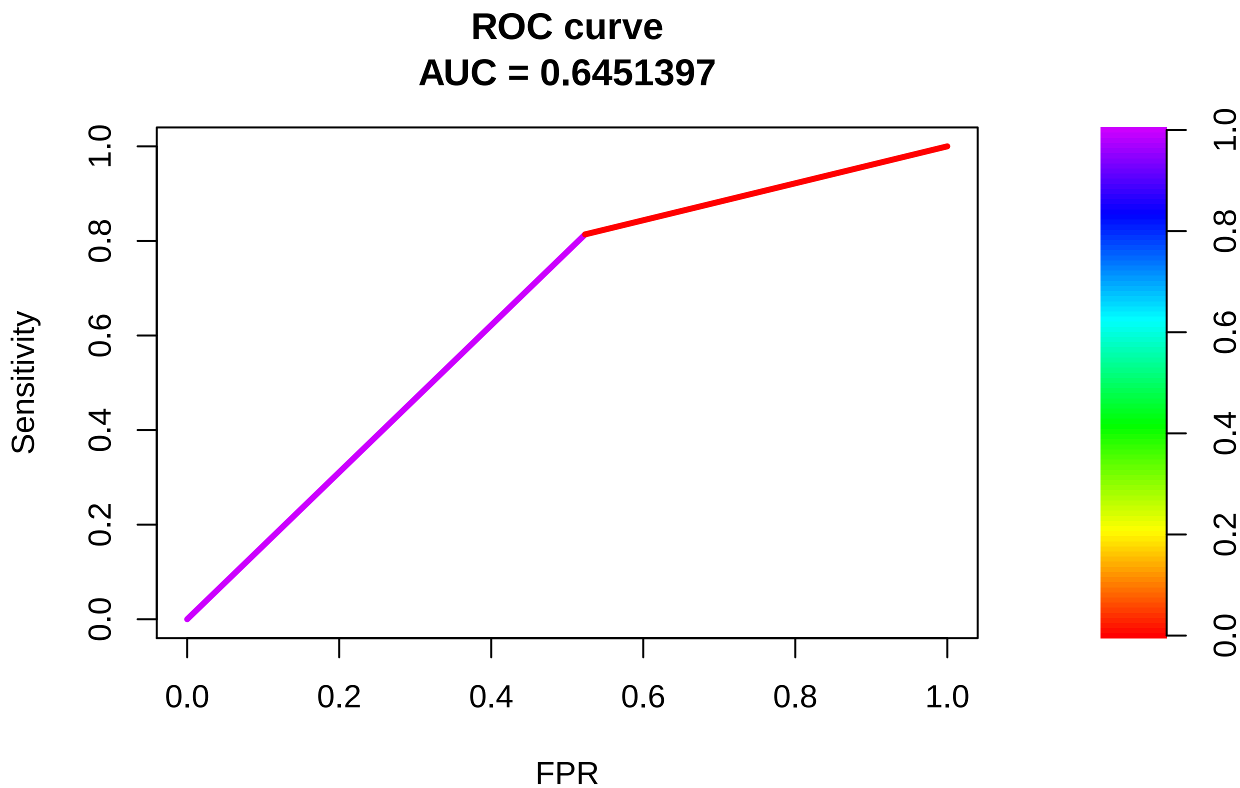
```
## # A tibble: 1 x 1
##   prop_vote_trump
##             <dbl>
## 1           0.510
```

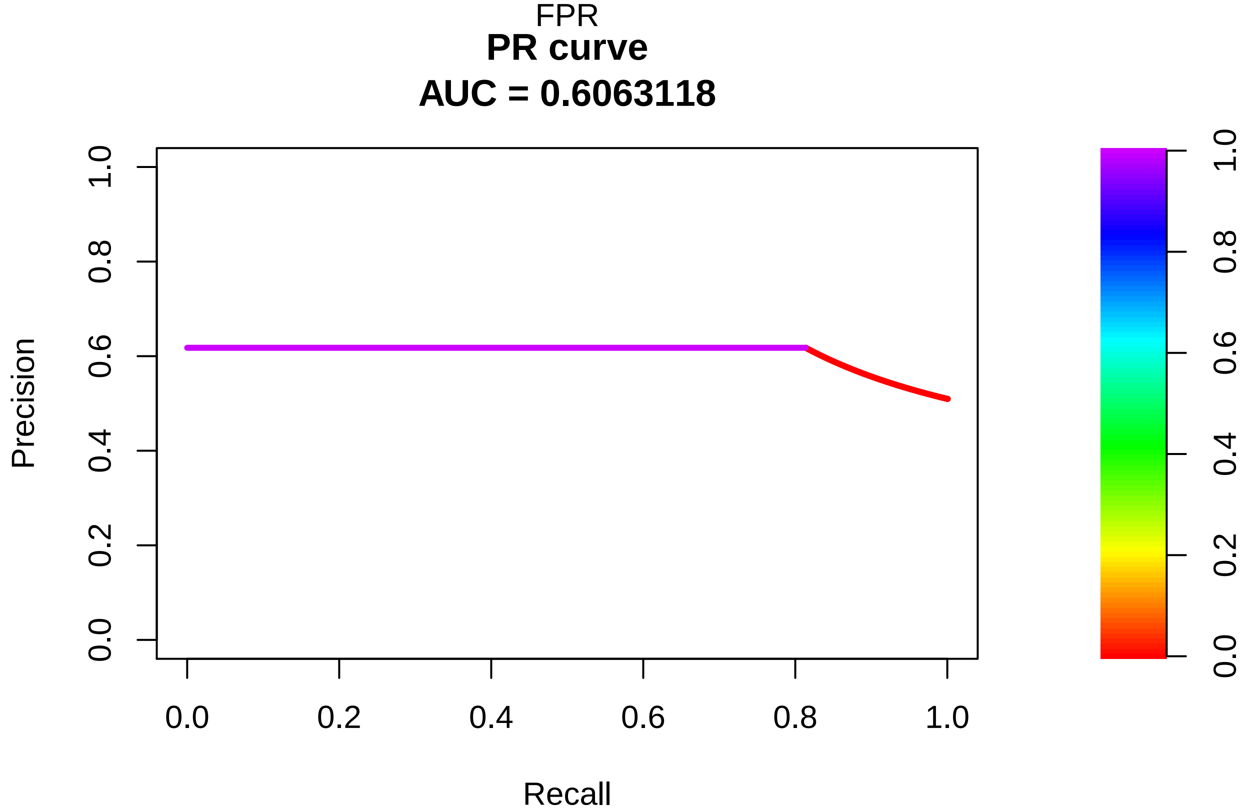
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_trump ~ sex + age_group + race + hispan + education + (1 |
##     state)
##    Data: survey_set
##
##      AIC      BIC   logLik deviance df.resid
##   4335.9   4458.9  -2147.9   4295.9     3447
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.2688 -1.0089  0.5580  0.8129  4.6751
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  state  (Intercept) 0.0565   0.2377
## Number of obs: 3467, groups:  state, 51
##
## Fixed effects:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       2.10152    0.73666   2.853 0.004334 **
## sexmale                           0.41511    0.07578   5.478 4.30e-08 ***
## age_group18 ~ 20                 -1.62472    0.66708  -2.436 0.014868 *
## age_group20 ~ 30                 -0.23053    0.17010  -1.355 0.175340
## age_group30 ~ 40                 -0.01900    0.13696  -0.139 0.889689
## age_group40 ~ 50                  0.20558    0.13742   1.496 0.134634
## age_group50 ~ 60                  0.15819    0.14086   1.123 0.261423
## age_group60 ~ 70                  0.02292    0.13392   0.171 0.864082
## raceasian or pacific             -1.50903    0.49078  -3.075 0.002107 **
## raceblack                        -3.21613    0.49016  -6.561 5.33e-11 ***
## raceother                        -1.00184    0.48213  -2.078 0.037715 *
## racewhite                        -0.48299    0.45368  -1.065 0.287054
## hispanmexican                    -1.74771    0.58520  -2.987 0.002822 **
## hispannot hispanic               -1.11824    0.56529  -1.978 0.047908 *
## hispanother                      -1.30600    0.59093  -2.210 0.027100 *
## hispanpuerto rican               -1.54304    1.45366  -1.061 0.288470
## educationbachelor                -0.49200    0.10938  -4.498 6.86e-06 ***
## educationgraduate                -0.46929    0.12589  -3.728 0.000193 ***
## educationtertiary (not bachelor) -0.45434    0.13576  -3.347 0.000818 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 19 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00781382 (tol = 0.002, component 1)
```






```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
```

```
##          0  810  329
##          1  890 1438
##
##                Accuracy : 0.6484
##                  95% CI : (0.6322, 0.6643)
##     No Information Rate : 0.5097
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2921
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4765
##             Specificity : 0.8138
##          Pos Pred Value : 0.7112
##          Neg Pred Value : 0.6177
##              Prevalence : 0.4903
##          Detection Rate : 0.2336
##    Detection Prevalence : 0.3285
##       Balanced Accuracy : 0.6451
##
##        'Positive' Class : 0
##
```

Post-Stratification

Post-stratification refers to the random sampling of a population and sampling of a sample n. After the survey, n units are divided into several layers according to certain stratification factors. Then stratified sampling estimation is carried out. In this project, post stratification is collecting the census data and applying the MR model to the census data to predict each individuals' voting result. It is useful because it is difficult to stratify the whole in a certain way beforehand. And the operation is simple, low cost, in the case of incomplete information, can be applied. In order to estimate the proportion of voters who will vote for Donald Trump we need to perform a post-stratification analysis. Here we create cells based on different states. Using the model described in the previous sub-section we will estimate the proportion of voters in each state. We will then weight each proportion estimate (within each state) by the respective population size of that state and sum those values and divide that by the entire population size. The reason why we choose "state" is because each state have electoral votes (total 538 votes this year) and who won more than 270 electoral votes wins the election [16]. Therefore, it is essential to predict according to states.

```
##      sex          age_group                                 race
##  female:23559   >= 70  :8211   american indian or alaska native:  117
##  male  :19888   18 ~ 20: 596   asian or pacific                :13784
##                 20 ~ 30:3203   black                           : 3706
##                 30 ~ 40:5419   other                           : 5017
##                 40 ~ 50:8260   white                           :20823
##                 50 ~ 60:9541
##                 60 ~ 70:8217
##           hispan                          education         state
##  cuban       : 1474   at most high school    :21528   CA     :11347
##  mexican     : 6264   bachelor               :11340   NY     : 4836
##  not hispanic:30272   graduate               : 5746   FL     : 4710
##  other       : 5277   tertiary (not bachelor): 4833   TX     : 3600
##  puerto rican:  160                                   NJ     : 2153
##                                                       IL     : 1636
##                                                       (Other):15165
```

```
##      perwt
##  Min.   :   5.23
##  1st Qu.: 308.57
##  Median : 444.55
##  Mean   : 568.17
##  3rd Qu.: 679.90
##  Max.   :9607.51
##
```

```
## # A tibble: 51 x 2
##    state trump_predict
##    <fct>         <dbl>
##  1 AK            0.468
##  2 AL            0.500
##  3 AR            0.560
##  4 AZ            0.484
##  5 CA            0.402
##  6 CO            0.487
##  7 CT            0.381
##  8 DC            0.370
##  9 DE            0.379
## 10 FL            0.545
## # ... with 41 more rows
```

Results

In this section, we will predict whether Donald Trump or Joe Biden will win the final federal election. The total number of electoral votes is 538, and the proportion of people who are willing to vote for Trump is predicted to be around 51% by using survey data. However this result has no personal weight which means that it cannot be used as the final result. It can only be used as a reference for future predictions. Therefore, we make predictions about the outcome of the general election by considering the electoral college. This model is accounted for variables of "sex" "age_group", "race", "hispan", "state" and "education". There are 51 states in America based on the dataset. We will predict the voting situation of these state voters. The prediction results show that there are 27 states where the percentage of voting Trump is between 0.4 and 0.5. For example, the state of Alaska (AK) has a 0.468 (46.8%) predicted voting rate of Trump. In the state of Alabama (AL), the number is 0.500 (50.0%). It is very balanced. Trump and Biden would do more campaigns on this kind of states to increase the possibility of winning. Moreover, the result demonstrates that in the state of California (CA) and Connecticut (CT), Trump's support is slightly weaker, with only 0.402 (40.2%) and 0.381 (38.1%) under the estimation. However, in the state of Arkansas (AR) and Montana (MT) the support rate for Trump is relatively high with rate of 0.560 (56.0%) and 0.583 (58.3%). In the state of District of Minnesota (MN) and Maryland (MD), we predict that voters' support for Trump would be the lowest, 0.353 (35.3%) and 0.345 (34.5%) respectively. According to our model's prediction result, Joe Biden will receive a total of 422 electoral votes, while Donald Trump will only receive 116 electoral votes. Joe Biden has more states in favor of voting for him, so we expect democratic presidential candidate, Joe Biden, to win the election.

```
## # A tibble: 2 x 2
##   elected      total_electoral_votes
##   <chr>                        <dbl>
## 1 Donald Trump                   116
## 2 Joe Biden                      422
```

Discussion

Conclusion

In conclusion, we predict that Joe Biden who is the presidential nominee of democratic party would win in the election. Based on our model, Joe Biden would get 422 electoral votes in total while Donald Trump would only get 116 electoral votes. The difference is quite big because Joe Biden has much more states that in favour of voting him based on our model.

Weaknesses

Based on the above analysis, there are a few weaknesses of this model. First, the number of variables is quite small in both dataset and model. In the model, we only use 6 variables and we mainly focus on the demographic variables. We neglect variables that keep abreast of times. For example, in the 2016 US election, the Facebook marketing plays a key factor in Trump's winning [11]. However, the datasets of census and survey do not provide this kind of campaign variables. Thus, the model would be out of date for election of this year. Second, the dataset of census does not contain the information on people's party preference on democratic and republican. People who like democratic better would be more likely to vote for Joe Biden while people who like republican more would vote for Donald Trump. With this information, the model would be more precise. We could have a basic estimate on the voting of Joe Biden and Donald Trump. However, this information is not available. Third, when we build the model, we do not consider income factor. Typically, voters for Donald Trump are people who have a lower income while people who has a higher income tend to vote for Joe Biden. Although the income variable in census represents personal total income and the one in survey represents household income, we should still take income into this model. Fourth, when we clean the raw data of census, we are too general on some specific variables. For example, when we clean the variable of race, we combine Asian and Pacific Islander together for convenience. There are a lot of groups under Asian like Chinese, Korean, Japanese and so on. These groups would have some trends on voting based on their background and this may influence the result of election a lot. Nonetheless, we ignore these features in the model and this would cause a weak prediction. Furthermore, it is worth noting that education variable plays a very important role in the election. When we design the survey of election preference, we should include more people that have lower education level. In the survey dataset, it includes more people who have higher education level, and this would make the prediction in favor of Joe Biden [12]. People with higher education level would be more likely to vote for Joe Biden [12].

Next Steps

In order to make the estimation of model more accurate, we should find ways to eliminate the above weaknesses. For instance, we could clean the data in a more detailed way and include more variables in the analysis. We should also include survey that contains more people that has lower education level. Since the election result would be out on November 3rd, we could compare the actual result of election with our prediction. We need to do a post-hoc analysis. First, we could figure out if our prediction result match the real result. And then, we should know how big the difference is between the predicted total electoral votes and the actual one. If the difference is very big, we may need to switch to another model. For example, we could use Principal Component Analysis (PCA) which is an algorism that decreases dimensional space to start the model [13]. This would eliminate some unnecessary variables from the data and mainly focus on the key principles to predict the election. It is also important to figure out which factor is the key success factor on the election. If we do not include that in the model, we definitely need to add that. After, the analysis, we could redesign the model and check if the new one provides a result that is closer to the actual result. Moreover, some machine learning technology could help us to build a better model. The use of artificial neural networks, which is a brain-spired system, would help the model a lot [14]. It would decrease the uncertainty of the model. All of these would better improve the estimation of this model in future elections.

References

1. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200131). Retrieved from [URL].
2. 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
3. PUMS USA, University of Minnesota, www.ipums.org.
4. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
5. Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr
6. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
7. Jens Keilwagen, Ivo Grosse and Jan Grau (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. PLOS ONE (9) 3.
8. Jan Grau, Ivo Grosse, and Jens Keilwagen (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics (31) 15, pp. 2595-2597.R package version 1.3.1.
9. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
10. United States Electoral College Votes by State. (n.d.). Retrieved November 01, 2020, from https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124
11. Bump, P. (2019, March 29). All the ways Trump's campaign was aided by Facebook, ranked by importance. Retrieved November 02, 2020, from https://www.washingtonpost.com/news/politics/wp/2018/03/22/all-the-ways-trumps-campaign-was-aided-by-facebook-ranked-by-importance/ 12.How Race and Educational Attainment Factor Into Biden's 2020 Lead. (2020, September 17). Retrieved November 02, 2020, from https://morningconsult.com/2020/09/17/trump-biden-race-education-voters/
12. Jaadi, Z. (n.d.). A Step by Step Explanation of Principal Component Analysis. Retrieved November 02, 2020, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis
13. Dormehl, L. (2019, January 06). What is an artificial neural network? Here's everything you need to know. Retrieved November 02, 2020, from https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/
14. Who Can and Can't Vote in U.S. Elections. (n.d.). Retrieved November 02, 2020, from https://www.usa.gov/who-can-vote
15. Presidential Election Process. (n.d.). Retrieved November 02, 2020, from https://www.usa.gov/election