Prediction on Overall Popular Vote of 2020 American Federal Election by Multilevel Regression with Post-Stratification

Yuika Cho, Mengyu Lei, Yimeng Ma, Qiyun Wang

2020-10-31

Model

Here we are going to predict the popular vote outcome of the 2020 American federal election (include citation). To do this, we access the survey data (Tausanovitch and Vavreck 2020) and census data(Steven Ruggles and Sobek 2020) and then employ the post-stratification technique(Lohr 2009). In the following sub-sections we will describe the data reprocessing, the model specifics and the post-stratification calculation.

Data Preprocessing

After comparison with survey data and census data, we find out the co-variables common to these two datasets, which are describing `citizen`, `gender`, `census_region`, `hispanic`, `race`, `income`, `education`, `state` and `age`. However, the levels or the ways to classification of variables in these two datasets are not the same. Thus, we unified the grouping of co-variables. The details of regrouping variables will be shown in Appendix. The response variable is `vote_trump`,which is a binary variable of voting Donald Trump.

Model Specifics

Since the response variable is binary. Firstly, we apply Generalized linear model(GLM) (Nelder and Wedderburn 1972), specifically a logistic linear regression, with using all valid covariates introduced above and their two-way interaction terms (`Model1`).

From the summary report of the `Model1`, we could find most of the variables are not significant. This may because the variables are correlated and redundant. The Bayesian information criterion (BIC)(Schwarz and others 1978), usually results in more parsimonious model than the Akaike information criterion (AIC). So we use the BIC stepwise function to select a better model. Let p represent the probability of voting Trump. As a result, the final model(`Model2`) is:

$$\begin{aligned} Model2: \hat{\eta} = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) &= \hat{\beta}_0 + \hat{\beta}_1\text{age } + \hat{\beta}_2\text{citizen} + \hat{\beta}_3\text{census region} + \hat{\beta}_4\text{income}+ \\ &= \hat{\beta}_5\text{race} + \hat{\beta}_6\text{gender} + \hat{\beta}_7\text{income:age} \end{aligned}$$

As for model diagnostics, we use QQ-Plot for checking Normality of residuals, fitted value v.s. residuals plot for checking Homoscedasticity and Ljung-Box test for Independence. The model can be acceptable after validation and the details are shown in Appendix.

Post-Stratification

After we decided the variables in the model above, we can apply this model on the census data to have a overall prediction on the vote rates. However, the covariates have seriously imbalanced levels in the census data. For example, only 238805 observations are citizen while 2940341 are non-citizen. Post-stratification involves adjusting the sampling weights so that they sum to the population sizes within each post-stratum. This usually results in decreasing bias because of non-response and underrepresented groups in the population. Thus, we create 4 cells based on the most 4 imbalanced variable, which are `citizen`, `census_region`, `race` and `income`. Since every variable has several levels, the combinations of these 4 cells will generate $2 \times 4 \times 5 \times 3 = 120$ groups. In each group, the features are similar so that we can get more accurate prediction.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Results

The summary of final model is:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.072	0.320	-0.224	0.823
age	-0.012	0.006	-2.051	0.040
citizenNon_citizen	-0.426	0.122	-3.489	0.000
genderMale	0.418	0.055	7.591	0.000
raceAmerican Indian	0.530	0.266	1.992	0.046
raceBlack	-1.399	0.193	-7.255	0.000
raceOthers	-0.054	0.180	-0.303	0.762
racePacific	-0.477	0.582	-0.820	0.412
raceWhite	0.628	0.147	4.270	0.000
incomeLow Income	-1.897	0.319	-5.955	0.000
incomeMedian Income	-1.729	0.303	-5.709	0.000
census_regionNortheast	-0.141	0.086	-1.643	0.100
census_regionSouth	0.298	0.074	4.034	0.000
census_regionWest	-0.107	0.083	-1.284	0.199
age:incomeLow Income	0.027	0.007	4.068	0.000
age:incomeMedian Income	0.028	0.006	4.327	0.000

As for survey data, the vote rates between Trump and Biden for each state can be shown in `Figure 1`. From the plot, we can find almost all the states have higher probability to vote Biden, except Arkansas and West Virginia. Again, the red and blue horizontal lines represent the mean of rates, where the higher support rates for Biden is more obvious.

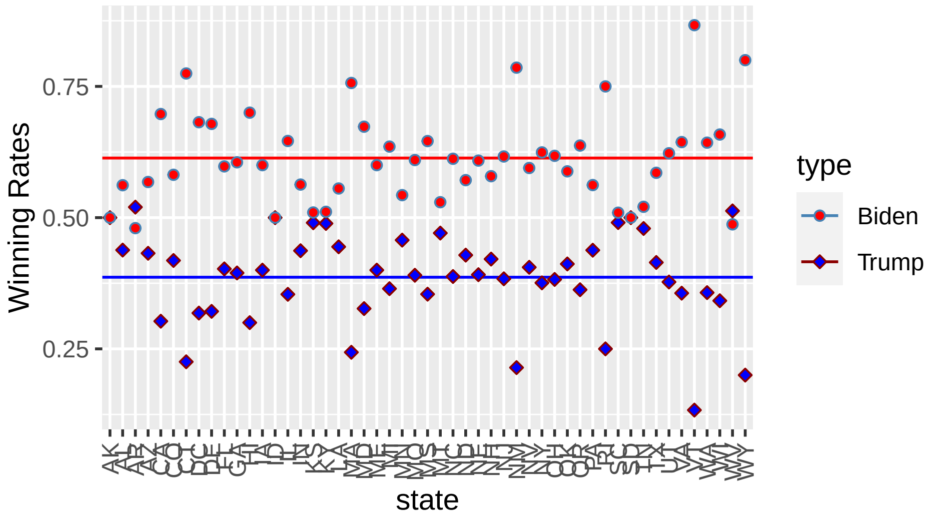
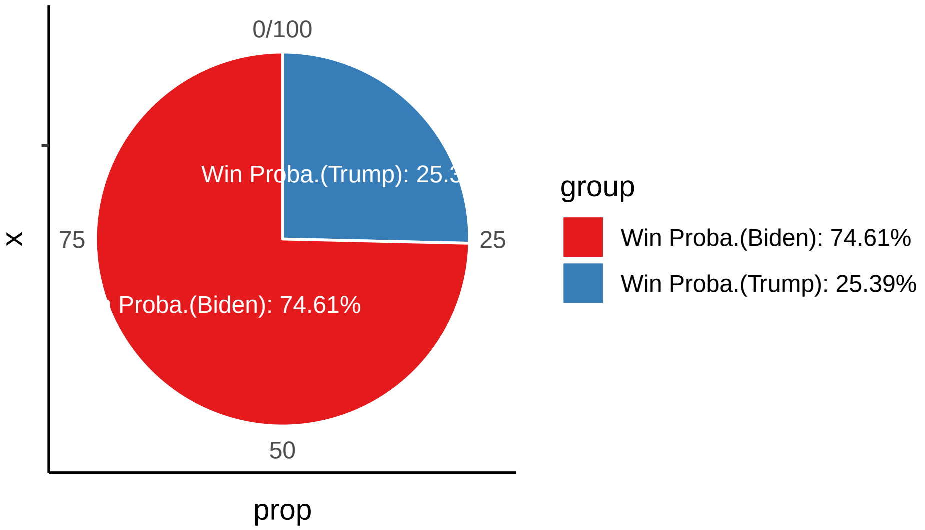


Figure 1: Vote Rates for 2020 Survey Data

Turning to the census data, we use `Model2` with post-stratification to predict the overall vote rates for Trump and Biden. Firstly, we calculate the probability using $\hat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}$ and then split the probability to the binary predicted value. Since p represents the probability of voting Trump, the predicted value will be 1 if $\hat{p} \geq 0.5$, otherwise 0.

Later, we apply post-stratification by $\hat{y}^{PS} = \frac{\sum N_j \widehat{y_j}}{\sum N_j}$, where $\hat{y}_j$ is the estimate in each cell and N_j is the population size of the j^{th} cell based off demographics.

`Figure 2` illustrates the prediction of overall popular vote in the 2020 American federal election, which tells us Biden will have absolutely advantage of 74.61% probability to be the president, while Donald Trump will have only 25.39% to win the election. This prediction result may be surprised to some people since there are some issues about the sample data and model, which will be talked in the next part.



e 2: The Pie Chart of Overall Winning Probabilities

We use `State` as cell to do Post-stratification again, the creation results are still show Biden will have 74.64% probability to win the election.

```
## [1] 0.7464093 0.2535907
```

Discussion

Conclusions

Prediction Results: By calculating the vote rates from survey data, we find that almost all the states are more likely to vote Biden. The simple mean vote rates for Trump and Biden are 38.65% and 61.35% separately. After considering the effect of imbalanced data and applying post-stratified technique, the overall vote rates for Trump and Biden are 25.39% and 74.61%. Again, after applying `State` as cell to run the Post-stratification, the results are still similar to the original post-stratification ways. Although the final election result is not affected by the specific prediction rates, the probabilities changed a lot when applying post-stratification. We may conclude post-stratification helps deal with the issues caused by imbalanced data.

Model Coefficients: From the summary of `Model2`, we can conclude:

(1) Voters who are citizens, male, white, American Indian with high income and from South region are more likely to vote Trump;

(2) Black with low or median income voters are more willing to supporting Biden.

Weakness and Future Work

Weakness:

(1) Although the final prediction results show Biden will have absolute advantage to win the election, there still exist a lot of uncertainty before the official results are announced. Because there may be black swan incidents in recent days, for example, Trump has unfavorable information about Biden, which may directly lead to the defeat of Biden. Thus, these uncertainty but vital factors are those we can not expect in our prediction.

(2) In the United States, candidate wins majority of votes will win all votes allocated by the state. However, in our prediction, we only consider the prediction for each voter without taking care of the overall situation in each state. This may lead to some forecast bias.

(3) The response bias and selection bias of the survey data might also be the reasons to prediction bias. Under COVID-19, the survey data are mostly collected by Internet or telephone. It is difficult to get the true responses from those who are old, poor, indifferent to politics and not familiar with the electronic equipment.

Future Work:

(1) We will continue to pay attention to any behavioral information about the candidates before the results of the general election are announced to avoid black swan incidents, which can provide corrections to our predictions.

(2) In next step, it will be a big direction to apply the "Winner take all" rule in our prediction to make correction. Also, the method of collecting data should be more refined so that the data can represent the intentions of more even all voters.

Appendix

Modified variables in two datasets

As for `survry_data` and `census_data`, we just group each level and rename them to be consistent without adding more levels for `citizen`, `gender`, `census_region`, `hispanic`, `race`, `state`, and `age`. For `education` and `income` , we integrate some levels to one new level, which can help simplify our model. The `education` variable is split to `Below Bachelor`, `Bachelor` and `Above Bachelor`, and `income` is classified to `Low Income`, `Median Income` and `High Income`.

Model Diagnostics

1. Normality of residuals

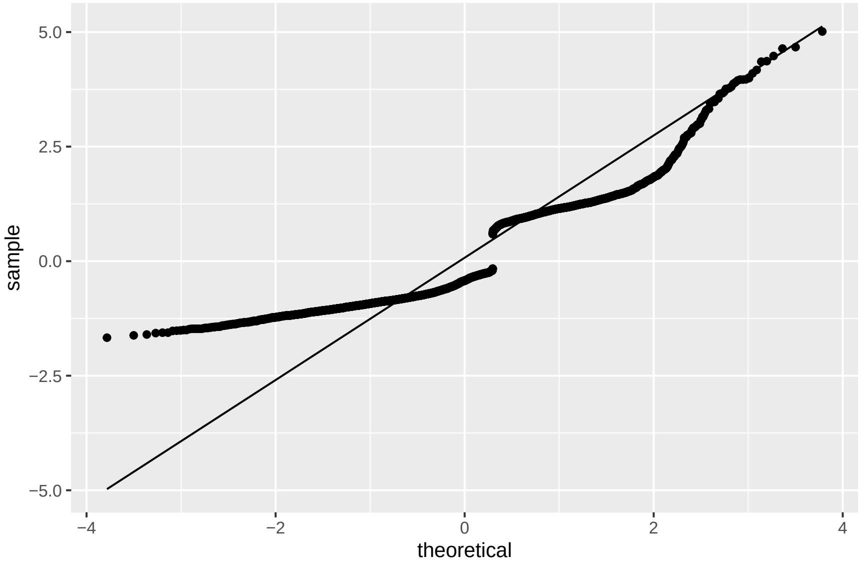


Figure 3: QQ−Plot

In statistics, a QQ (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. From the QQ plot above `Figure 3`, we can find although some points are on the straight line, there still large part of points are not on the straight line, which means the residuals doesn't fit normal distribution very well.

2. Homoscedasticity

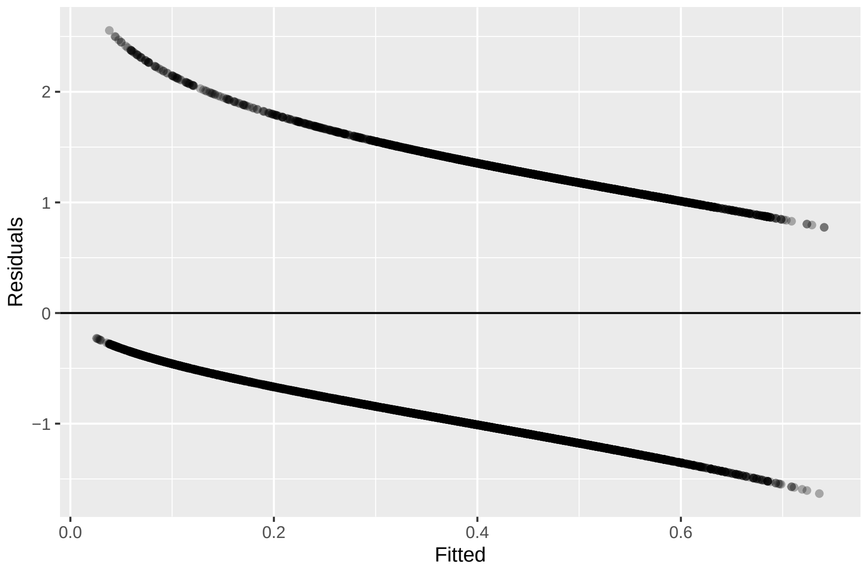


Figure 4: Fitted values v.s. Residules

From the figure above `Figure 4`, we can find though the points have some trends, they are on the two sides around zero. Thus, it is not very confident to state the variance is constant, which also means there may not exist homoscedasticity.

3. Independence

From the R output, we can find the p-value of Ljung-Box test (Ljung and Box 1978) is 0.6644, which means we have no enough evidence to reject null hypothesis and then all of the variables are independent.

```
# 3. independence
Box.test(residual)
```

```
##
##  Box-Pierce test
##
## data:  residual
## X-squared = 0.18799, df = 1, p-value = 0.6646
```

References

Ljung, Greta M, and George EP Box. 1978. "On a Measure of Lack of Fit in Time Series Models." *Biometrika* 65 (2): 297–303.

Lohr, Sharon L. 2009. *Sampling: Design and Analysis*. Nelson Education.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–84.

Schwarz, Gideon, and others. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS Usa: Version 10.0 [Dataset]. Minneapolis, Mn: IPUMS." https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "Democracy Fund + Ucla Nationscape, October 10-17, 2019 (Version 20200814)." https://www.voterstudygroup.org/publication/nationscape-data-set.