Prediction of Popular Vote of the 2020 American Federal Election Using Logistic Regression with Post-stratification

STA304

By
Haichuan Xue (1004070346)
Jiafei Lyu (1005004727)
Wenzhuo Zeng (1005227050)

2020.11.02

Model

We are interested in predicting the outcome of the overall popular vote of the 2020 American federal election. A logistic regression model with post-stratification is introduced to estimate the proportion of voters who are likely to vote for Donald Trump.

Model Specifics

In our model, demographic information such as race, age, sex, educational attainment, household income, and state are used as predictors. A binary indicator variable, vote trump, is used as our response variable to indicate a voter's voting result. Our model is presented as the following :

$$log\left(\frac{P(Y_i|X_i)}{1-P(Y_i|X_i)}\right) = \mu + X_i\beta_i$$

Where:

- Y_i represents the result of whether a subject votes for Donald Trump or not. If the i^th subject votes for Donald Trump, then $Y_i = 1$. Otherwise, $Y_i = 0$.
- $P(Y_i|X_i)$ is the probability of voting for Donald Trump for the i^th subject, given the conditions of X_i.
- μ is the intercept. It is the probability of voting for Donald Trump for an American Indian or Alaska Native female at age 18-20, having a household income at $100,000 to $124,999, with associate degree in Alaska.
- β's are the coefficients of each predictors to show the effectiveness. For example, we expected that the probability is decreased when the state characteristic is changed into Alabama from Alaska.
- The full summary of the model can be viewed in the appendix.

Post-stratifications

Post-stratification is a method to adjust sampling weights and the difference between sample and population. It balances the underestimation caused by the underrepresentation of a group. Its concept is to divide the population into cells. Based on the sample, the proposed model is used to estimate the interest within each cell. The weight in each cell is calculated by its proportion in the population. We create cells based on a few demographic variables that we believe to be important in predicting the popular vote. The variables are race, educational attainment, and house income. Race is chosen because some races, especially certain minority races, might have a preference of one candidate over the other. Educational attainment and household income may also be crucial since the two candidates' campaign styles were different. They also hold different political views on certain topics. People with the dissimilar educational background may resonate with different candidates.

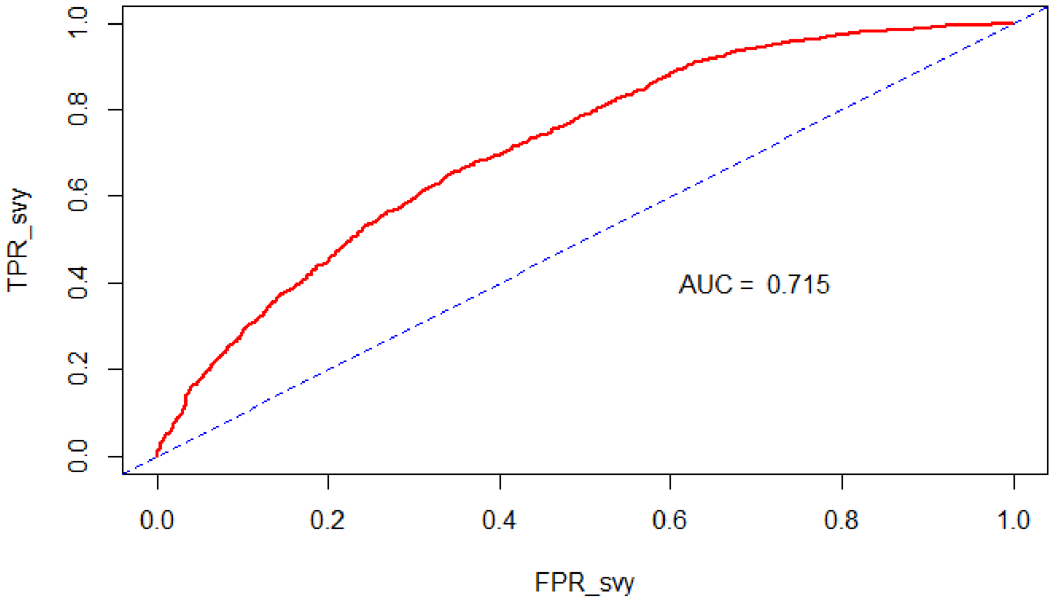
Model analysis

We propose to use and the area under the receiver operating characteristics curve (AUC-ROC curve) to perform the model diagnosis. The statistic value of AUC is calculated for the model using the survey data. With a high value of AUC, it indicates that the model has a very good discriminate ability to distinguish the results between voting for Donald Trump and not voting for Donald Trump.

Results

From Figure 1, based on the survey data, the AUC of our model is 0.715. It indicates that the model can distinguish the results between voting for Donald Trump and not voting for Donald Trump at 71.5% of the times.

Figure 1: AUC-ROC curve using survey data



Through the post-stratification analysis, we propose a logistic regression model, containing six variables, state, age, gender, education level, race, and income of the household, to estimate the proportion of voters who are likely to vote for Donald Trump. The estimated value is 0.39.

Discussion

Summary

Prediction of the popular vote of the 2020 American Federal Election is performed using the logistic regression model we propose included race, age, sex, educational attainment, household income, and state as predictors. People who are under age 18, non-American citizens, and have no intention to vote are excluded from the sample. The binary response variable is vote_trump, meaning the variable has value 1 if the person intends to vote for Trump and 0 for Biden. After the model is constructed, a post-stratification modification is used to make reasonable predictions to the final result of the election in real life. The sample population is divided into 840 groups by creating cells containing variables, race, educational attainment, and household income. By calculating the estimated voting probability for Donald Trump of each cell, a weighted average of each cell can be used to make the inference of the population based on its proportion.

Conclusions

Based on the estimated proportion of voters in favour of voting for the Republicans, which is 34.24% via the proposed model with the AUC value equaling to 0.715, we predict that the Democrats will win the election. Also, the newest election poll for the Republicans is 43.1%. We can conclude that Biden will be more likely to win the 2020 American Federal Election.

Weakness

One weakness of the prediction could be that not enough variables were introduced to the model. Also, some observations of the sample were dropped because it contains missing values at certain variable columns. Responses that are not certain about their intention were filtered out from the prediction, but they may have an influence on the election. The linear regression model may not be the best fit because it can distinguish between the voting outcomes at 71.5% of the cases. It may not be accurate enough. There are still many unknown confounding variables potentially influencing the final election result. The final prediction was calculated based on the weighted average of the total sample estimate. The voting policy of America is to count the total votes of the electoral college, which we did not use to approach the final result.

Next Steps

Future steps could be done to improve the accuracy of the prediction. A multilevel regression model can be applied to make the prediction via better computational methods. Analyzing by states and then calculating by votes of the electoral college could be done to improve the predicted results. Cells used in the model can be more specific by having combinations of more factors to make a more accurate prediction.

Appendix

GitHub link

https://github.com/HaichuanXue/STA304-Problem-Set-3.git

Summary of Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.01	1.21	0.84	0.40
stateAL	-1.35	1.17	-1.15	0.25
stateAR	-0.94	1.22	-0.77	0.44
stateAZ	-1.46	1.16	-1.27	0.21
stateCA	-2.07	1.14	-1.80	0.07
stateCO	-1.82	1.17	-1.55	0.12
stateCT	-2.53	1.18	-2.13	0.03
stateDC	-2.19	1.26	-1.74	0.08
stateDE	-2.52	1.25	-2.01	0.04
stateFL	-1.64	1.15	-1.43	0.15
stateGA	-1.21	1.16	-1.04	0.30
stateHI	-1.88	1.25	-1.51	0.13
stateIA	-1.88	1.19	-1.57	0.12
stateID	-0.93	1.25	-0.74	0.46
stateIL	-1.94	1.15	-1.69	0.09
stateIN	-1.70	1.17	-1.46	0.14
stateKS	-1.19	1.20	-0.99	0.32
stateKY	-1.84	1.17	-1.57	0.12
stateLA	-1.52	1.18	-1.29	0.20
stateMA	-2.53	1.17	-2.16	0.03
stateMD	-1.91	1.17	-1.63	0.10
stateME	-2.11	1.25	-1.69	0.09
stateMI	-1.93	1.15	-1.67	0.10
stateMN	-1.76	1.18	-1.49	0.14
stateMO	-1.72	1.16	-1.48	0.14
stateMS	-0.94	1.22	-0.77	0.44
stateMT	-1.63	1.27	-1.28	0.20
stateNC	-1.66	1.15	-1.44	0.15
stateND	10.16	228.55	0.04	0.96

stateNE	-2.05	1.27	-1.62	0.11
stateNH	-1.91	1.26	-1.52	0.13
stateNJ	-1.69	1.15	-1.47	0.14
stateNM	-2.59	1.27	-2.04	0.04
stateNV	-1.23	1.18	-1.04	0.30
stateNY	-1.86	1.15	-1.63	0.10
stateOH	-1.73	1.15	-1.50	0.13
stateOK	-1.38	1.20	-1.15	0.25
stateOR	-2.11	1.17	-1.81	0.07
statePA	-1.57	1.15	-1.36	0.17
stateRI	-2.58	1.45	-1.79	0.07
stateSC	-1.06	1.18	-0.90	0.37
stateSD	-1.22	1.30	-0.94	0.35
stateTN	-1.01	1.17	-0.86	0.39
stateTX	-1.26	1.15	-1.10	0.27
stateUT	-1.43	1.22	-1.17	0.24
stateVA	-2.03	1.16	-1.75	0.08
stateVT	-3.89	1.56	-2.50	0.01
stateWA	-2.19	1.16	-1.88	0.06
stateWI	-2.14	1.17	-1.83	0.07
stateWV	-1.13	1.21	-0.93	0.35
stateWY	-1.81	1.83	-0.99	0.32
age_int21-30	0.54	0.23	2.38	0.02
age_int31-40	0.96	0.22	4.45	0.00
age_int41-50	1.10	0.22	4.97	0.00
age_int51-60	1.16	0.22	5.28	0.00
age_int61-70	1.01	0.22	4.63	0.00
age_int71-80	1.02	0.24	4.32	0.00
age_int80up	1.21	0.42	2.91	0.00
sexMale	0.41	0.07	5.86	0.00
educationCollege Degree	-0.03	0.13	-0.24	0.81
educationGrade12&under	0.14	0.63	0.23	0.82
educationHighschool	0.31	0.12	2.47	0.01
educationMasters&Doc	0.17	0.14	1.28	0.20
raceBlack/African American	-2.39	0.38	-6.27	0.00
raceChinese	-1.45	0.52	-2.78	0.01
raceJapanese	-1.00	0.74	-1.36	0.17

raceother asian or pacific islander	-0.66	0.40	-1.64	0.10
raceOther race	-0.89	0.38	-2.36	0.02
raceWhite	-0.16	0.35	-0.45	0.65
household_income$125,000 to $149,999	-0.11	0.17	-0.63	0.53
household_income$15,000 to $19,999	-0.57	0.20	-2.86	0.00
household_income$150,000 to $174,999	-0.28	0.21	-1.35	0.18
household_income$175,000 to $199,999	0.24	0.25	0.98	0.33
household_income$20,000 to $24,999	-0.33	0.19	-1.71	0.09
household_income$200,000 to $249,999	0.57	0.23	2.43	0.01
household_income$25,000 to $29,999	-0.32	0.19	-1.66	0.10
household_income$250,000 and above	0.14	0.24	0.60	0.55
household_income$30,000 to $34,999	-0.32	0.19	-1.69	0.09
household_income$35,000 to $39,999	-0.36	0.19	-1.87	0.06
household_income$40,000 to $44,999	-0.57	0.20	-2.80	0.01
household_income$45,000 to $49,999	-0.28	0.19	-1.47	0.14
household_income$50,000 to $54,999	-0.25	0.18	-1.38	0.17
household_income$55,000 to $59,999	-0.09	0.23	-0.39	0.69
household_income$60,000 to $64,999	-0.38	0.22	-1.70	0.09
household_income$65,000 to $69,999	-0.28	0.24	-1.17	0.24
household_income$70,000 to $74,999	-0.50	0.21	-2.36	0.02
household_income$75,000 to $79,999	-0.12	0.21	-0.58	0.56
household_income$80,000 to $84,999	-0.55	0.26	-2.14	0.03
household_income$85,000 to $89,999	-0.45	0.26	-1.71	0.09
household_income$90,000 to $94,999	-0.32	0.29	-1.08	0.28
household_income$95,000 to $99,999	-0.59	0.23	-2.62	0.01
household_incomeLess than $14,999	-0.53	0.16	-3.26	0.00

References

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.

Kirk, A., Gutiérrez, P., Hulley-Jones, F., &amp; Adolphe, J. (2020, November 01). US election polls tracker: Who is leading in swing states, Trump or Biden? Retrieved November 02, 2020, from https://www.theguardian.com/us-news/2020/nov/01/us-election-polls-tracker-who-is-leading-in-swing-states-trump-or-biden

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686,https://doi.org/10.21105/joss.01686

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77<http://www.biomedcentral.com/1471-2105/12/77/>