American federal election prediction research

Groups member: Jingying Liu, Zhijian Zhu, Mingze Xu, siyuan shen

Due Date: Nov. 2, 2020

Gitgub: https://github.com/liujin59/sta304ps3

Model

we want to predict the popular vote outcome of the 2020 American federal election. We will use a multivariable logistic (logit) regression model to model this outcome since it is a binary outcome. The variables used in the model to predict the outcome are age, sex, state, race, education level.To do this, we use post-stratification technique to analyze and calculate the data.

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{state} + \beta_4 x_{race} + \beta_5 x_{education}$$

where

p_i *represents the probability of voters who will vote for Donald Trump*

and

β_1*is the coefficient for the age variable,*

β_2 *is the coefficient for the sex variable,*

β_3 *is the coefficient for the state variable,*

β_4 *is the coefficient for the race variable,*

β_5 *is the coefficient for the education variable, for every male*

and

$\mathbf{X_i}$ is the design matrix for the logit model

Post-Stratification

The formula of Postratification is:

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

Results

We calculated from the poststratification conducted the 95% confidence interval around it, are presented in Table below. The code is in the r-code file.

Point estimate and margin of error of probabiliy of Donald Trump winning

Point.estimate	Margin.of.error
0.317	0.328

Discussion

Summary

I analyzed survey data at the individual level from Democracy Fund + UCLA Nationscape and poststratification census data from the American Community Surveys (ACS) to o predict Donald Trump's chances of winning the United States presidential election in 2020. I established a multivariate Logistic regression model and used it to calculate post-stratified estimates which my result shows that Donald Trump has an estimated probability of winning the election of 3.17%. Which means Donald Trump will beat Joe Biden by just over 40%.

Conclusion

In this report, I am interested in predicting the popular vote for the 2020 U.S. federal election which I have used a logistic regression model and a simple linear regression model to analysis and processing data. Also, I have used Post-Stratification method to estimated percentage of people who will vote for Donald Trump and my analysis shows that the Democrats have a good chance of winning the election. At the same time, my calculation also has some shortcomings, since the model only has two variables, I did not take into account the voters of the legal voting age eligibility. Also, the popular vote may not be quite the same as the electoral vote. I should wait for the actual results in the future and then compare the actual data with what I have calculated.

Weaknesses

The power of our model is suboptimal since I only used five variables in our logistic regression model. It should be lots of factors that influence whether a vote would vote for Donald Trump. These are farrago that should be accounted for in the model. Also, I think we should build model for each state individually, since the voting law

will have difference among different state, such as the age can voting. Also, the technique for calculating the margin of error can be more robust takes into the account of post-stratification weighting.

Next steps

I want to try some more complicated model for example, I would try to add randomness (fix effect or random effect) the model and use a multilevel model. And try to search for factors that could increase model prediction power. We could also try to compare models based on determine whether overfit and nested model comparison statistical tests and try A more robust technique of calculating the margin of error.

References

The data is coming from:

Press, C., Finance, Y., & Newsweek. (2020, October 30). New: Second Nationscape Data Set Release. Retrieved November 03, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 03, 2020, from https://usa.ipums.org/usa/index.shtml

Cite for R:

```
@Manual{,
    title = {R: A Language and Environment for Statistical
Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2020},
    url = {https://www.R-project.org/},
  }
```