

Prediction of the 2020 Presidential Election*

Spoiler Alert: Biden Wins

Anees Shaikh, Jaffa Romain, Lu Mu, and Cameron Fryer

31 October 2020

Abstract

In this paper, we first consider the Democracy Fund + UCLA Nationscape Wave 50 dataset, which contains the results of a survey (conducted June 25-July 1, 2020) on American voter attitudes. Thereafter, the survey data is used to train a model relating voter intent to a few explanatory variables. The model is then applied to the post-stratification dataset; namely, the results of the 2018 1-year American Community Survey (ACS). Since the ACS data pertains to individual persons and their characteristics, the associated use of our model allows us to conclude that Joe Biden will be the next President of the United States.

Keywords: forecasting; US 2020 election; Trump; Biden; multilevel regression with post-stratification

1 Introduction

To accurately predict the winner of the United States 2020 presidential election is difficult, as many different factors are involved.

2 Data

Our data is of penguins (Figure ??).

Talk more about it.

Also bills and their average (Figure ??). (Notice how you can change the height and width so they don't take the whole page?)

Talk way more about it.

3 Model

In order to forecast the popular vote of the U.S 2020 elections, multilevel logistic regression with post-stratification (MRP) was used. The individual survey gives sample data on voters of the general U.S population, and post-stratification allows us to reweight estimates, adjusting bias between our sample and the target population, so that we have a representative sample of likely voters Wang et al. [2014]. To achieve this, cells are constructed using variables in the cleaned Nationscape survey such as age, household

*Code and data are available at: https://github.com/aneesshaikh/elections_prediction

income, and race. The model is then trained on the survey using the proportions given by the ACS post-stratification data set. This approach gives us an advantage when attempted to forecast voting, as we can use a broad survey to speak to subsets in the population, and also tends to be less expensive to collect than non-probability samples ?. However, using this approach places limitations on how much we can interpret from the estimates. Although we can estimate voting intention in different demographic groups, it doesn't tell us any qualitative information on voting patterns. For example, we might be able to see how different age categories vote, but we won't have a measure of the policy preferences, or their views on the candidates.

Using the Nationscape, multilevel logistic regression model can be fitted to the survey data set to model the proportion of voters who will vote for Trump. Logistic regression can be used to model a binary dependent variable, which is the choice between the Republic candidate, Donald Trump, and the Democratic candidate, Joe Biden in our case. The variable 'vote_2020' in the data set was used as dependent variable, with the variable return 1 for voters intending to vote for Trump, and 0.

Our model uses the variables pertaining to age, race, sex, household income, and whether a voter is Hispanic to predict our dependent variable. The model takes the form: *** TO DO: Regression and Other Stories to cite equation **

$$Pr(y_i = 1) = \text{logit}^{-1}(X_i * \beta)$$

where $y_i = 1$ represents a voter voting for Trump in the election. $\Pr(y_i = 1)$ is the probability of a voter choosing to vote for Trump. The $X_i\beta$ are the linear predictors, where β represents the fitted coefficients for each independent variable X_i in the model.

The logit function $\text{logit}(x) = \log(x/1 - x)$ maps the range (0, 1) to $(-\infty, \infty)$

Using the 'glm()' function in [R],

4 Results

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

6 References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [URL].

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>

References

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. *Forecasting Elections With Non-Representative Polls*, 2014. URL <https://dx.doi.org/10.1016/j.ijforecast.2014.06.001>.