Forecasts the next president for the United States by Age, Gender, State, Race, Education and Employment

Jiawei Du, Lin Zhu, Siri Huang, Wang Xinyu

2020-11-02

Abstract

The 2020 United States Presidential election is scheduled to be held on Tuesday, November 3, 2020. We used a logistic regression model with poststratification to estimate the probability of Donald Trump winning this election. We estimated the probability to be 0.411. Limitations of our analysis include assuming all voters were over 18 and only using six variables in the prediction model which resulted in the model having suboptimal prediction power. Next steps include increasing model prediction and quantifying the margin of error with more robust techniques.

Keywords: election, politics, United States, Donald Trump, Joe Biden, United States Presidential election

Introduction

The 2020 US presidential election is scheduled to be held on Tuesday, November 3, 2020. This will be the 59th quadrennial presidential election. Voters will choose presidential electors who then will vote on December 14, 2020, to either elect a new president and vice president or re-elect the incumbents Donald Trump and Mike Pence, respectively.The series of presidential primary elections and caucuses took place from February to August 2020. This nominating procedure is an indirect election, where voters cast ballots choosing a slate of delegates to a nominating convention of a political party, who subsequently elect their parties' nominees for president and vice president. The major two-party candidates are Republican incumbent President Donald Trump and Democratic former Vice President Joe Biden, considered a referendum on the Trump presidency.

The primary goal of this paper is to predict the overall popular vote of the 2020 American presidential election using multilevel regression with post-stratification. The outcome of interest was whether a person would vote for Donald Trump, which was binary. We first used a logistic regression model to model this outcome variable using socio-demographic and politically relevant variables. Then, we post-stratified the sample using the variables in the model and assigned individuals into different groups based on variable combinations. Then, we used the model estimates to predict the probability of voting for Donald Trump for each variable combination. Finally, we combined the estimated probabilities to calculate the probability of Donald Trump wininng the election.

Data

We used two datasets in this analysis. The first dataset is the Democracy Fund + UCLA Nationscape 'Full Data Set' (Tausanovitch et al. 2019). Which is a weekly political poll on conducted by UCLA Democracy Fund Voter Study Group. We used Nationscape Wave 50 which was held from June 25 - July 01, 2020. This dataset has Individual level survey data. We pre-processed the dataset and selected demographic and politics-related variables including voting preferences, education level, geographic areas, ethniicty and employment status.

The second dataset has post-stratification data. We used the U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH from IPUMS USA. "IPUMS USA collects, preserves and harmonizes U.S. census microdata and provides easy access to this data with enhanced documentation. Data includes decennial censuses from 1790 to 2010 and American Community Surveys (ACS) from 2000 to the present" (USA 2018). The dataset we chose was the 2018 1-year American Community Surveys (ACS). "The American Community Survey (ACS) helps local officials, community leaders, and businesses understand the changes taking place in their communities. It is the premier source for detailed population and housing information about our nation." (USA 2018) We selected demographic and politics-related variables including voting preferences, education level, geographic areas, ethniicty and income level.

Model

The outcome of interest is whether a person would vote for Donald Trump in the 2020 American Presidential Election. This is an binary outcome. We will use a multivariable logistic (logit) regression model to model this outcome. The variables used in the model to predict the outcome are age, sex, state, race, education level and work status (currently working or not). These are important demographic variables that are typically found in many surveys and political polls. We will use these variables to predict whether a person would vote for Donald Trump. The model formulation is as follows:

$$\log(\frac{p_i}{1-p_i}) = \mathbf{X_i}\boldsymbol{\beta}$$

where

$$p_i = \text{the probability of individual i voting for Donald Trump}$$

and

$$i = 1 \ldots n$$

and

$$\boldsymbol{\beta} \text{ is a vector of regression coefficients}$$

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

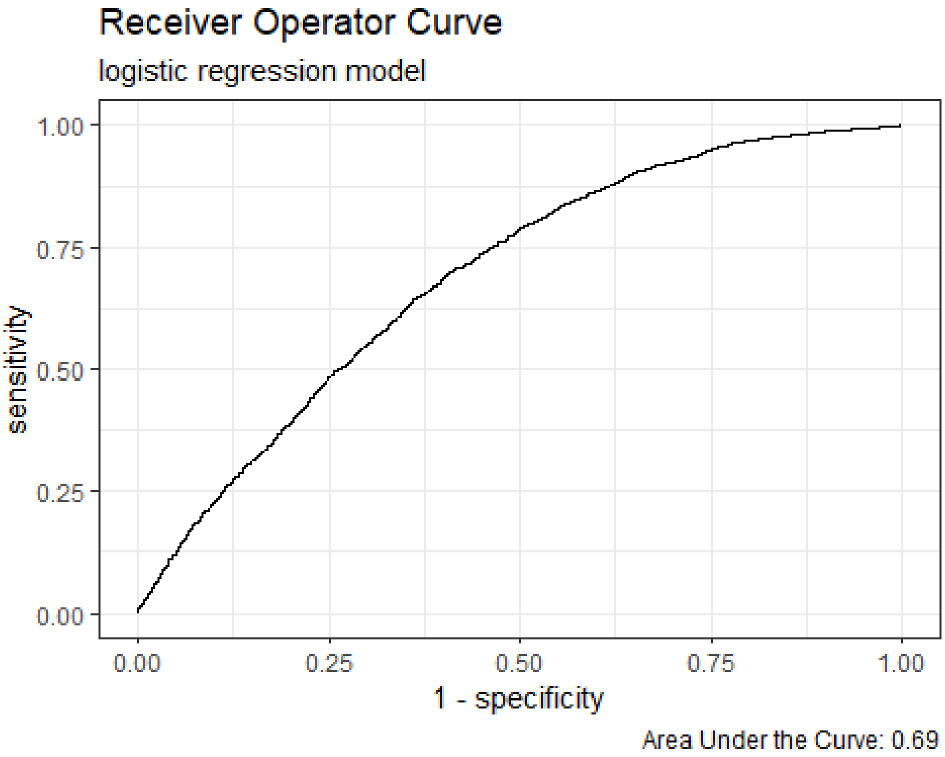
We were interested in the prediction power of our logistic regression model. We assessed this by using the Receiver Operator Curve and the Area Under the Curve. The Receiver Operating Characteristic curve, or ROC, is a visualization techinque that illustrates the diagnostic ability of a binary classifier model such as the logistic regression model as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR), or sensitivity, against the false positive rate (FPR), or specificity, at various threshold settings.

Next, we conducted poststratification on the sample using the multivariable logistic regression model we built. Multilevel regression with poststratification is a statistical technique for adjusting model estimates for known differences within a sample population, and a target population. The poststratification refers to the process of adjusting the estimates, basically a weighted average of estimates from all possible combinations of variables/ characteristics such as age, sex and education level. Each combination is sometimes called a “cell”. The multilevel regression is used to smooth out random noisy estimates in the cells with too little data by using overall or close-by averages. In this case, we are using the variables we built our logit model with to construct the cells for poststratification.

The point estimate from poststratification is the estimated probability of Donald Trump winning the 2020 United States Presidential Election. The margin of error is the standard deviation of the estimated probability times the 97.5% quantile of the normal distribution, so that the margin of error corresponds to 95% level of confidence.

Results

The ROC curve for the logistic regression model is shown in Figure @ref(fig:fig1) The AUC is 0.69 indicating sufficiently good model prediction power.



ROC curve of the fitted logistic regression model

The point estimate of the probability of Donald Trump winning the election, the margin of error around it, and the 95% confidence interval around it, calculated from the poststratification conducted are presented in Table @ref(tab:tab1) below.

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

Point.estimate	Margin.of.error
0.411	0.35

Discussion

We analyzed the individual level survey data from Democracy Fund + UCLA Nationscape and poststratification cnesus data from the American Community Surveys (ACS) to predict the probability of Donald Trump winning the United States 2020 Presidential Election. We built a multivarible logistic regression model and used it to compute poststratification estimates. Our results indicate that the probability of Trump winning the election was estimated to be 0.411. This means Donald Trump has just a little over 40% of winning against Joe Biden.

Weaknesses

Since we only used six variables in our logistic regression model, the predicting power of our model is suboptimal, only at 69%. This indicates that there are other factors that influence whether a vote would vote for Donald Trump. These are confounding factors that should be accounted for in the model. Also, we made the assumption that the eligible voting age is 18 years old. In fact in the United States the eligible voting age is different for each state. Also, the technique for calculating the margin of error is naive, a more robust foruma takes into the account of postratification weighting.

Next steps

We would try to incorporate randomness in different levels of the sampling such as geographic variation and use a multilevel model. In addition, we could try to search for factors that could influence our outcome of interest to reduce the level of confounding and increase model prediction power. We could also try to compare models based on model comparison criteria based on overall model fit and nested model comparison statistical tests. A more robust technique of calculating the margin of error could be applied.

Link

https://github.com/christy723/Forecasts-the-next-president-for-the-United-States-by-Age-Gender-State-Race-Education-and-Work.git

References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frederique Lisacek, Jean-Charles Sanchez, and Markus Muller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

Tausanovitch, Chris, Lynn Vavreck, Tyler Reny, Alex Rossell Hayes, and Aaron Rudkin. 2019. "Democracy Fund+ Ucla Nationscape Methodology and Representativeness Assessment."

USA, IPUMS. 2018. *2018 American Community Survey 1-Year 2018 1-Year Acs*. https://usa.ipums.org/usa/index.shtml.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Francois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.