Joe Biden projected to win Popular Vote in 2020 US Election with 51% of Vote

Alen Mitrovski, Xiaoyan Yang, Matthew Wankiewicz

November 2nd, 2020

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

Keywords: Forecasting, US 2020 Election, Trump, Biden, multilevel regression with post-stratification;

1 Introduction

2 Data

We have used R (R Core Team [2019]), specifically Tidyverse (Wickham et al. [2019]) for data analysis

Data is from ACS (Ruggles et al. [2020]) and from Voter Study Group (Tausanovitch and Vavreck [2020]).

Figure **??** show us the voter demographics from the VSG data (Tausanovitch and Vavreck [2020]) vs the ACS data (Ruggles et al. [2020]).

Figure 2 show us more of the voter demographics from the VSG data (Tausanovitch and Vavreck [2020]) vs the ACS data (Ruggles et al. [2020]).

3 Model

In order to predict whether or not a person plans to vote for Joe Biden or Donald Trump, we plan to use logistic regression. Since logistic regression only works for binary response variables, we used a variable we created called `vote_biden` which returns a 1 is the respondent plan to vote for Joe Biden and a 0 if they don't (in this scenario we assume that they are voting for Donald Trump). The logistic regression model takes the form of:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{agegroup} + \beta_3 x_{race} + \beta_4 x_{state} + \beta_5 x_{income} + \beta_6 x_{hispanic}$$

In the equation above, each β represents a coefficient that the regression model will compute for us. As for our variables, we have chosen to use sex, age, race, income, state, and whether the respondent is hispanic. We decided to use the first 3 because they are generally strong predictors of which candidate a person would support, such as how some states tend to vote republican year after year while some states flip between democratic and republican almost every election. Next, we decided to choose income, because Joe Biden has made claims to increase taxes on the rich, which may influence their support for him. Lastly, we wanted to focus on whether the respondent was hispanic and if so, where they were from. This variable was important for our predictions because we know how poorly Donald Trump has spoken of hispanic people and we believe they could have a strong impact on the election.

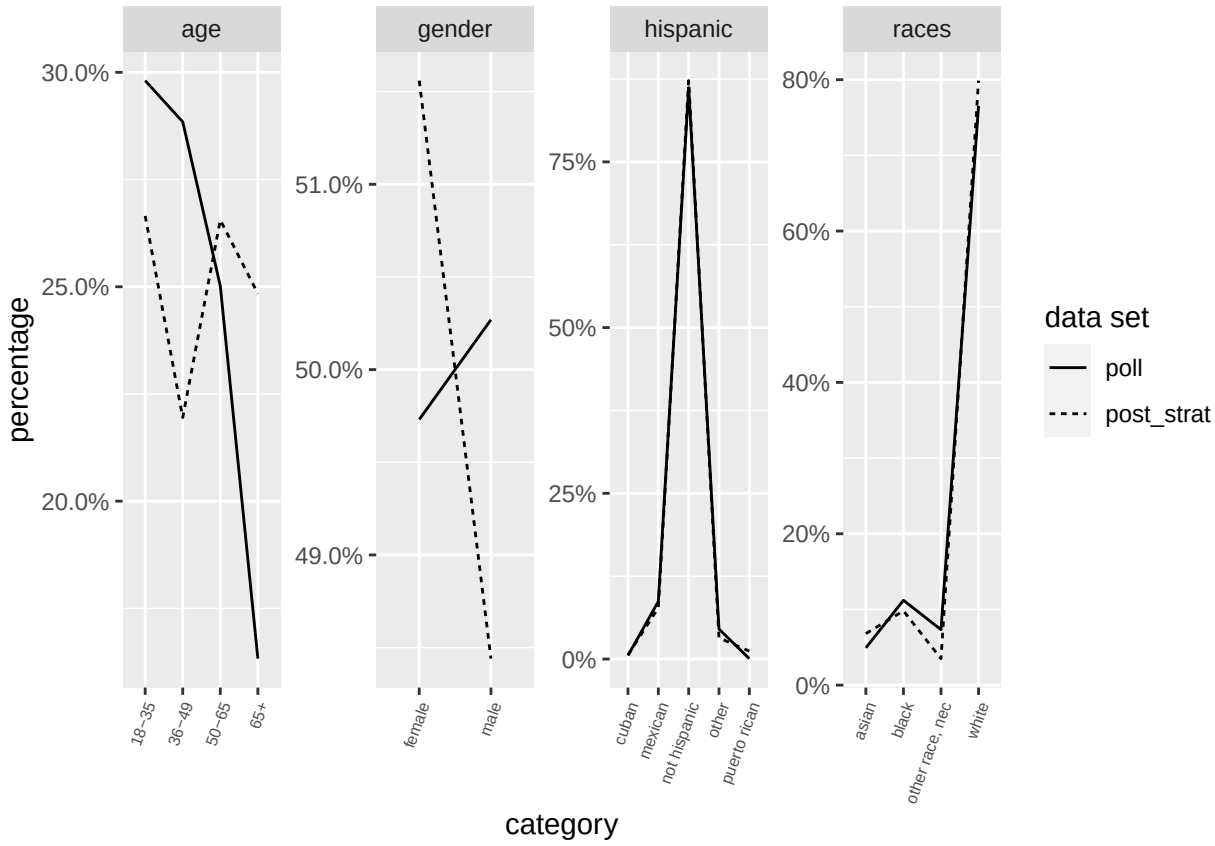

Figure 1: Demographics of Sample and Population

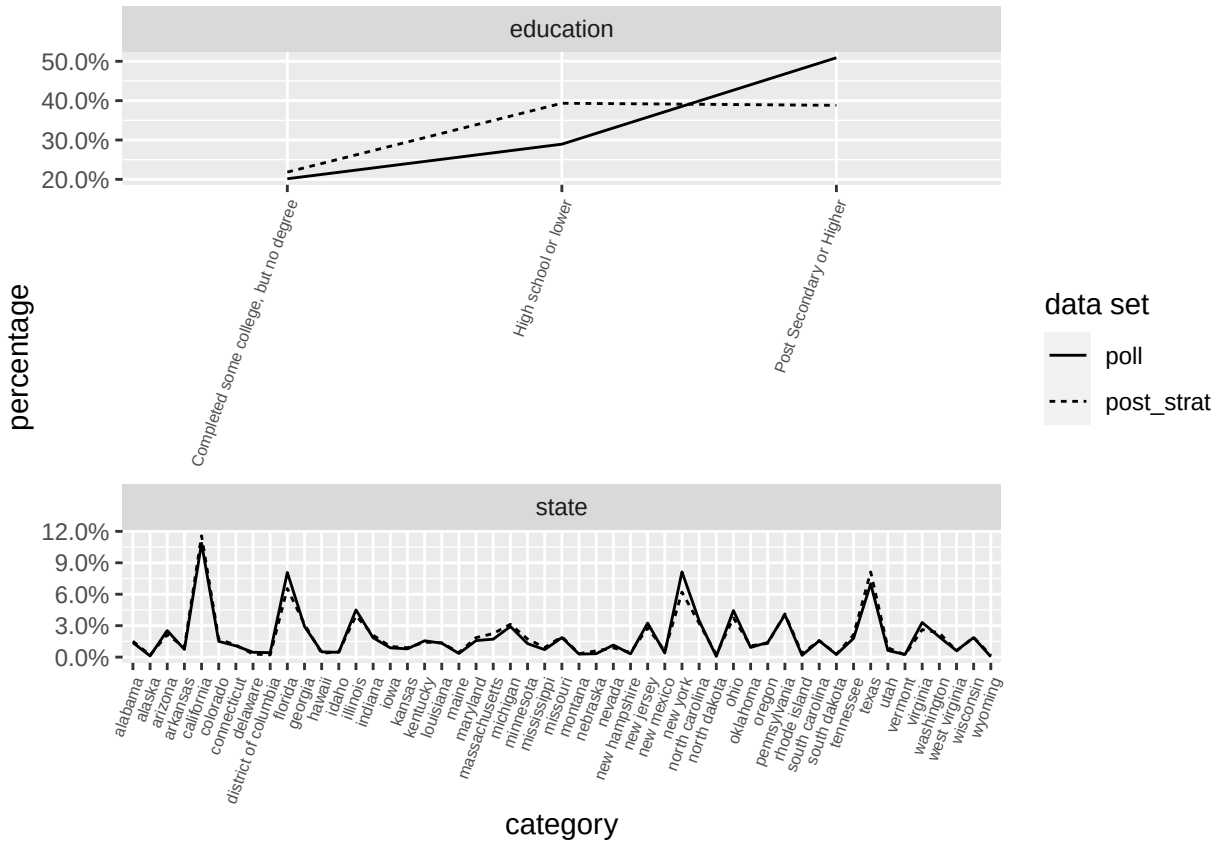

Figure 2: More Demographics of Sample and Population

The output of the logistic regression model will give us a probability of whether or not a voter plans to vote for Joe Biden or not. In order to find this probability, we take the sum of the right side of the equation and plug it into the equation below:

$$\frac{e^{sum}}{1 + e^{sum}}$$

This equation is just a manipulation of the initial equation, where e is the exponential equation ad sum is the sum of the right side of equation 1. We see that as the sum of the right side increases, the probability that a person will vote for Joe Biden increases as well.

We are running our regression model using the glm() function in R (R Core Team [2019]). The decision to run this model over other models like linear regression was made by the fact that we were predicting a binary variable about a voter's decision. Since there are only two possible options our data will likely follow an S shape and a straight line equation will not be helpful to model this relationship.

Our model does have some weaknesses, since the output must be binary, we cannot account for other candidates or a person deciding not to vote. This issue isn't too large because our main goal is to determine which of the two main candidates will be chosen by the people of America. Another weakness our model does encounter is that

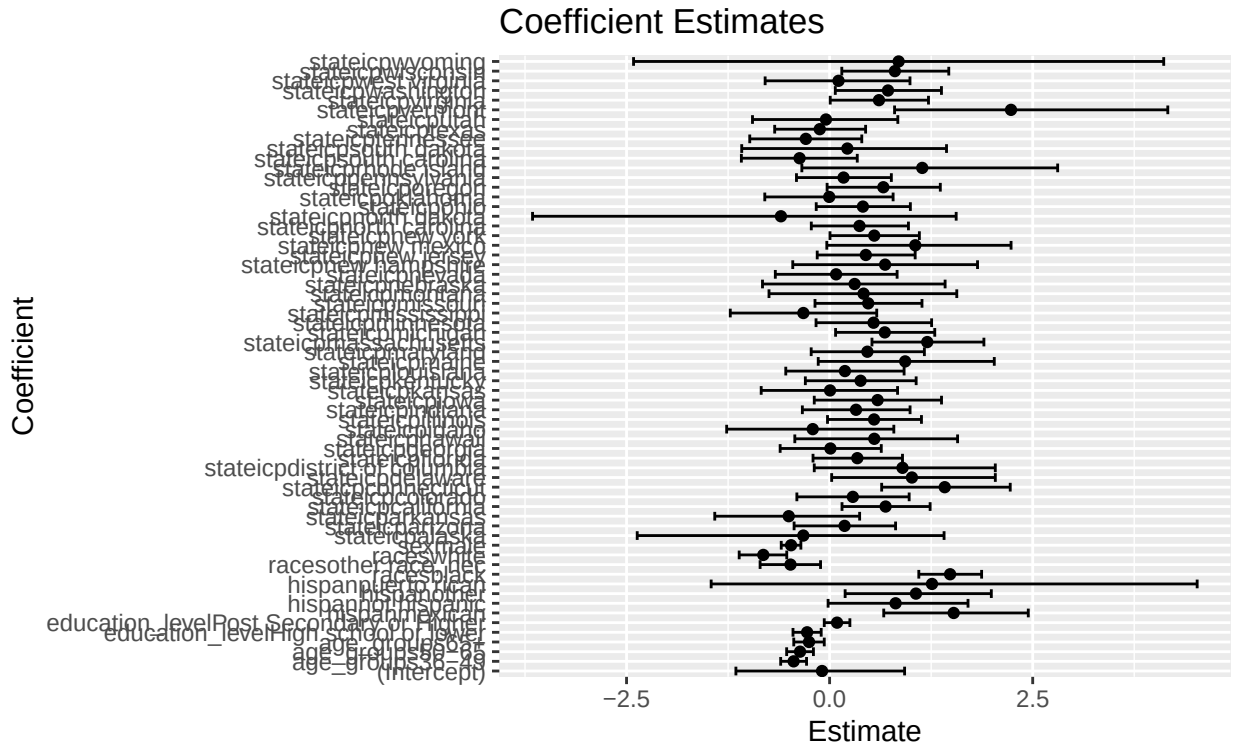

Figure 3: Coefficient Estimates

Figure **??** show us... using the polling data (Tausanovitch and Vavreck [2020]).

4 Results

Figure 4 show us...

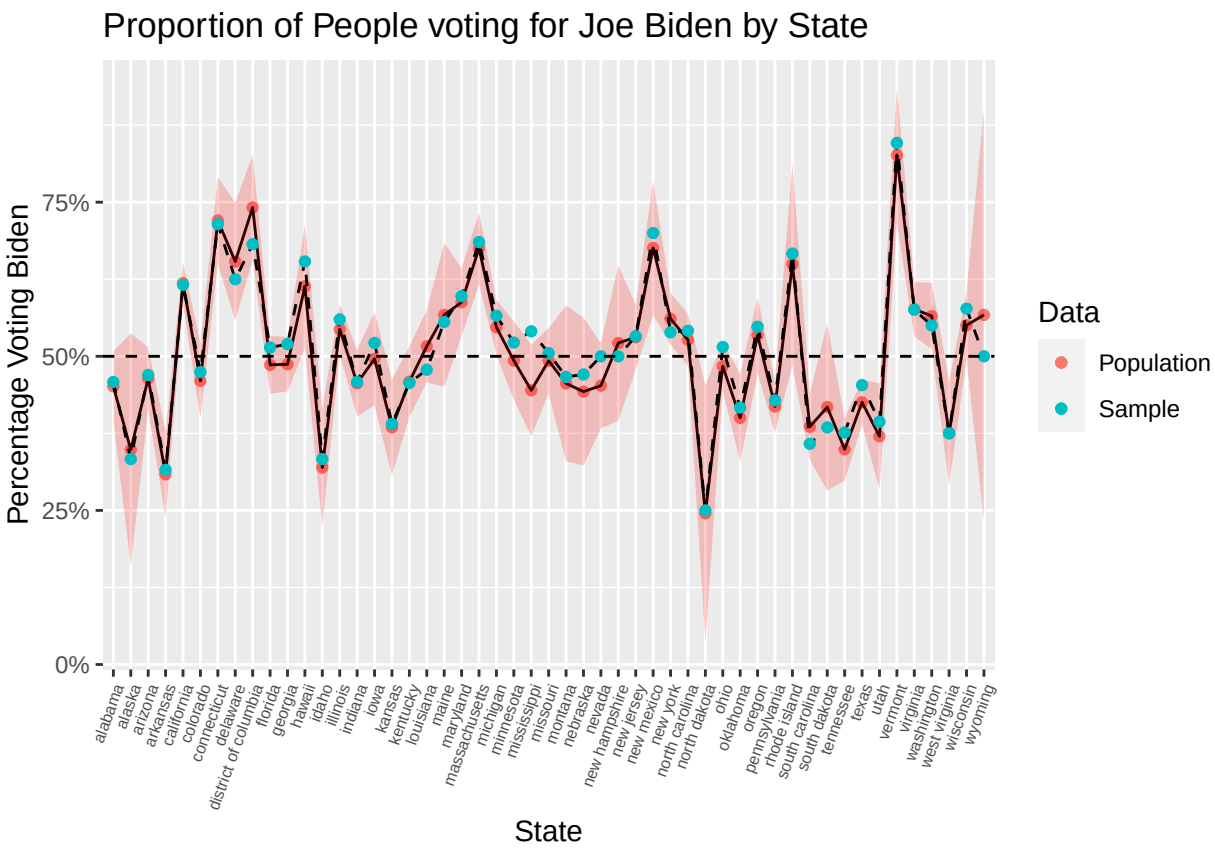

Figure 4: Proportion of each State Voting for Biden

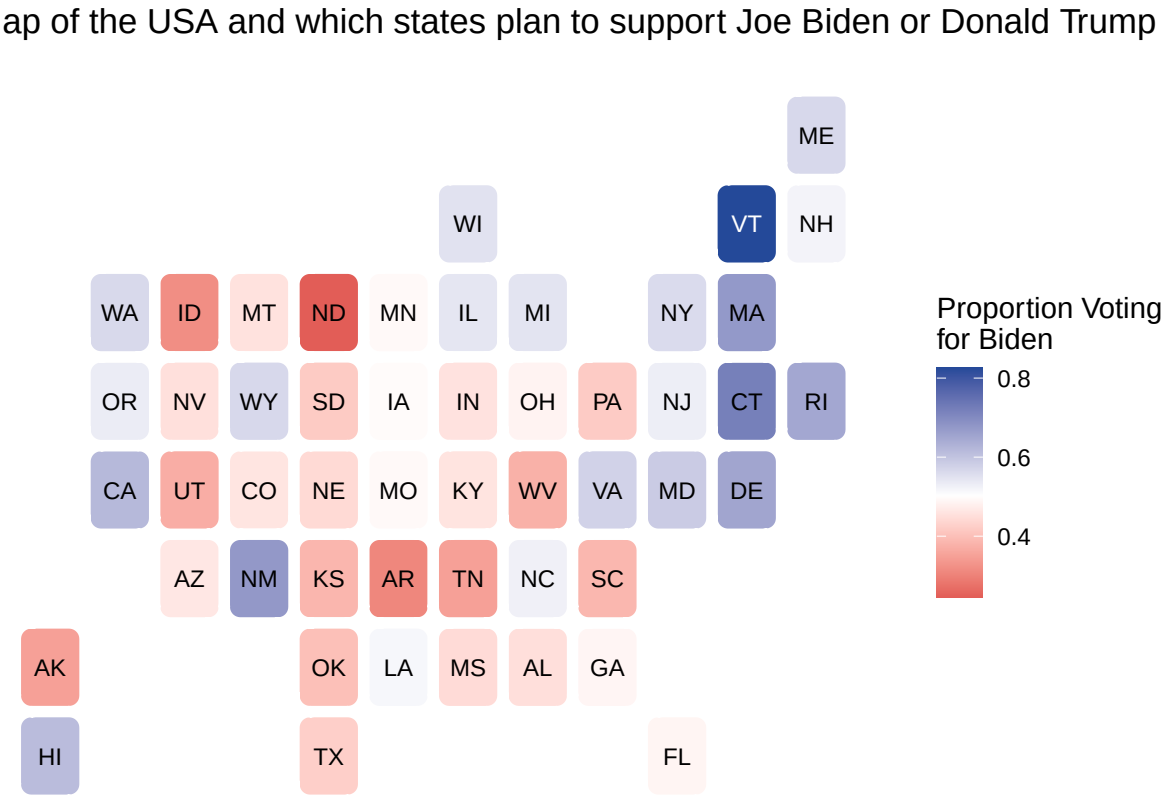

Figure 5: Proportion of Voters from Each State Voting Joe Biden

Table 1: Joe Biden Voting Result Estimates

	Lower Estimate	Mean Estimate	Upper Estimate
Number of Colleges	191.0	260	423.0
Proportion of Vote	45.8	51	56.3

Figure **??** is an amazing view of all the states and which candidate they are leaning towards voting for. We see that states like Vermont, Conneticut and California are some of the more "blue" states, meaning they plan to vote for Joe Biden, while North Dakota, Arkansas and Idaho are the "red" states, planning to vote for Donald Trump. The states that are more white in colour can be the most important for the race when it comes to deciding an actual winner through the electoral college. We see that Florida, Louisiana and New Hampshire are some of the more undecided states, and a switch in these states, can influence the election greatly.

Table 1 shows the lower, mean and upper predictions for the results of the election for Joe Biden. We see that on the lower end, Biden can expect to get 45.8% of the popular vote while only getting 191 electoral colleges. We also see that our middle estimate says Biden will get 51% of the popular vote while still losing the election by getting only 260 colleges. Lastly, on the upper estimate for Biden's results, he can get 56.3% of the popular vote while getting 423 college! This truly shows how close this election is, as we can see in Figure 4, many states are hovering around the 50% mark, which shows the colleges can go either way.

Figure 6 is a different view of Figure 4, this time only focusing on the predictions for each state with the errors included. We see that many states have error bars overlapping the 50% line, showing that many states are a toss up, given the nature of the electoral college.

Figure 7 shows the cooks distance for observations in our polling data set. Cook's distance is useful for

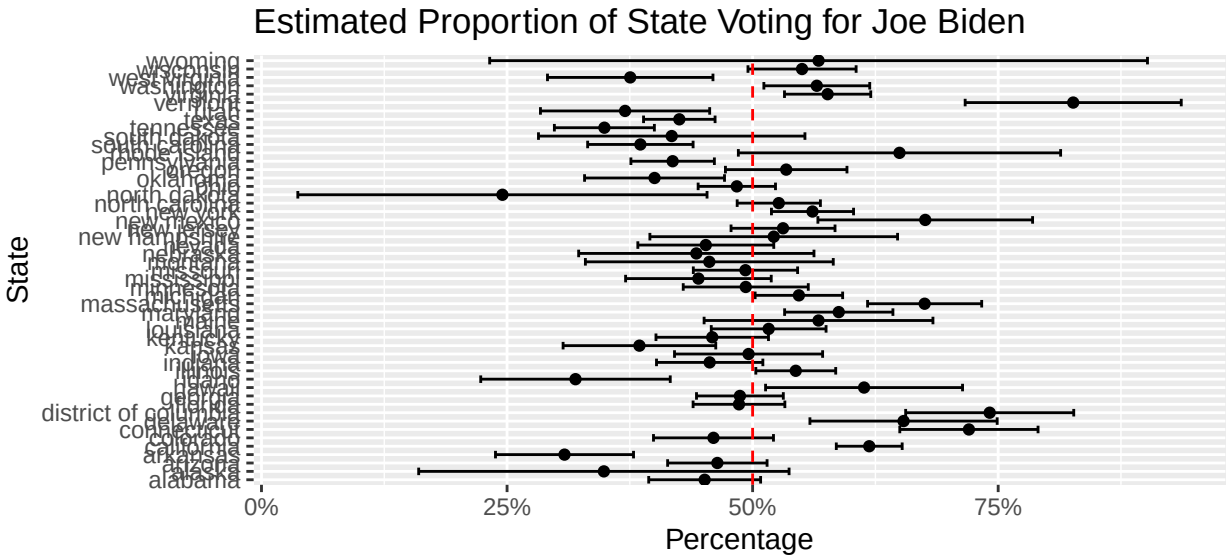

Figure 6: Proportion of Biden votes by state

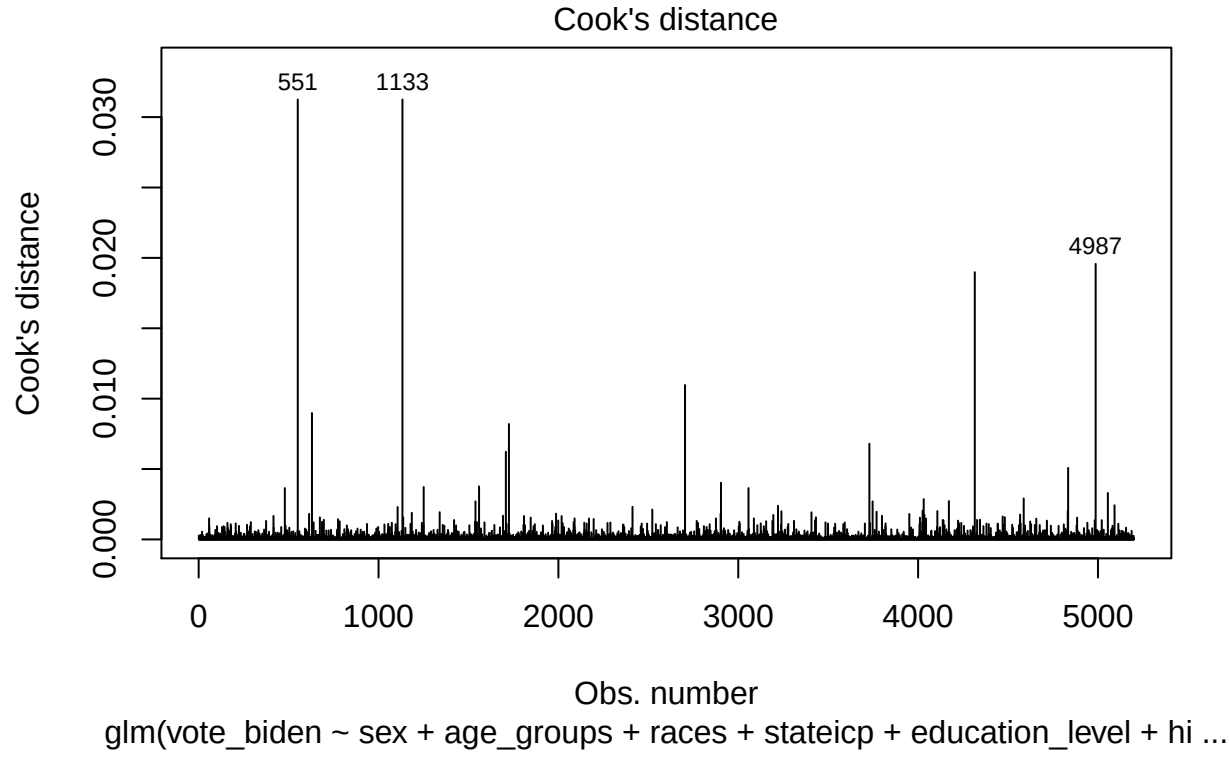

Figure 7: Cook's Distance Plot for Model

checking if a model is working correctly because it tells us how far a point is from the predicted value of it, telling us which points negatively impact our model. Figure 7 takes all observations of our polling data (Tausanovitch and Vavreck [2020]), and calculates the Cook's distance for it, showing which points can hurt the results of our model. We find that out of the 5200 obsevations, there aren't too many points that deviate from what we predict. This makes sense for our model because in the real world, there are some people who will support Donald Trump or Joe Biden, even if they don't "fit" the usual voter demographic for the candidate. Since the number of observations with large Cook's Distances are low, we can conclude that our model is fairly strong.

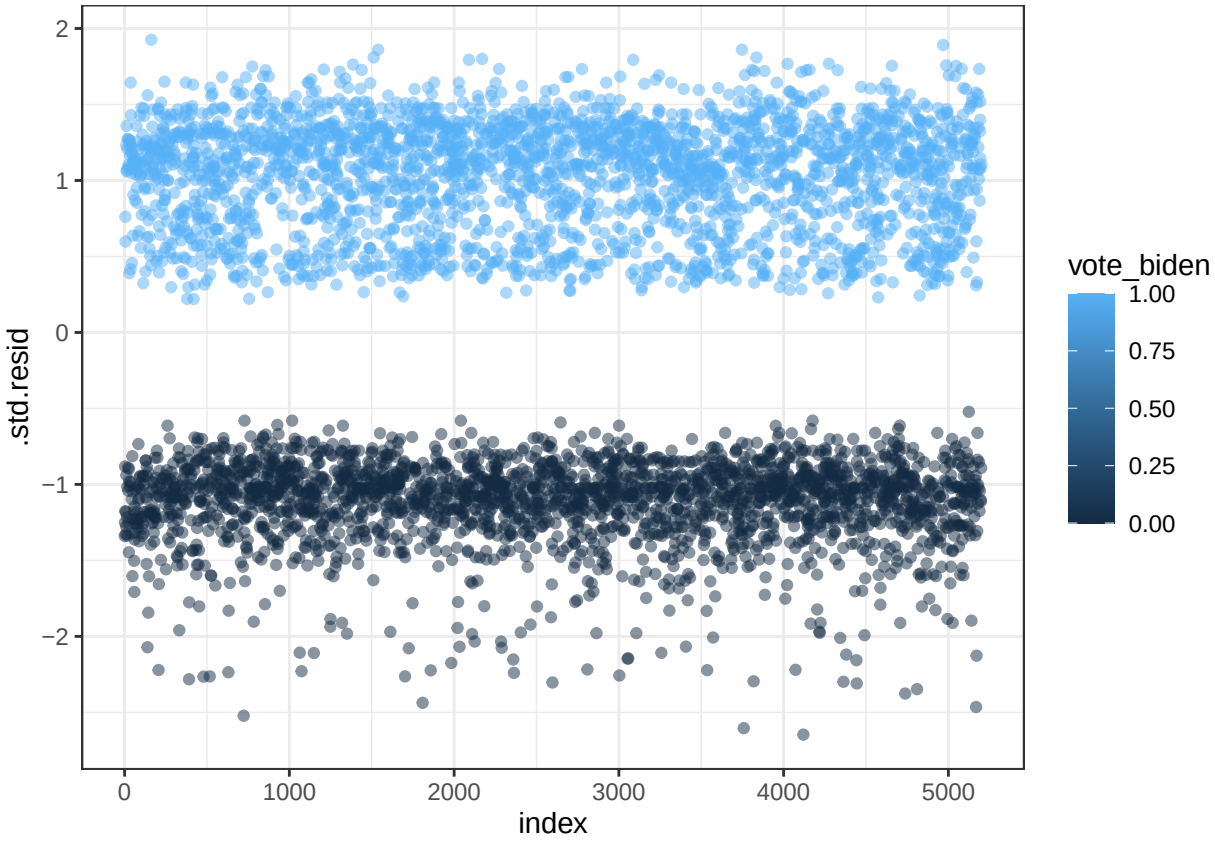

Figure 8: Residual Plot for Model

Figure 8 shows us the...

5 Discussion

References

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

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. URL https://doi.org/10.18128/D010.V10.0.

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

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.