Project Report

Kai Jeffreys, Benjamin Jia, Rachael Ren, Kobe Sarausad

2023-06-07

Introduction

Research Questions

Voting Bootstrap

We first must cover a disclaimer over our use of bootstrap on this data. One of the main bootstrap is that each observation of data has equal chance of occurring. However, this data set is weighted, indicating that each observation has a different chance of occurring. To fix this, under the professor's instructions, we sampled the data the way you would using a bootstrap (meaning each observation is the same) and then applied the weights afterward. This allows us to obtain a more valid estimate using the data, while still conducting bootstrap.

Our initial move was just to do a bootstrap just of the popular vote. We wanted to get a good estimate about the variance of outcomes of the election. Below, we have a histogram with simulated percentage of vote going to Biden on the horizontal axis.

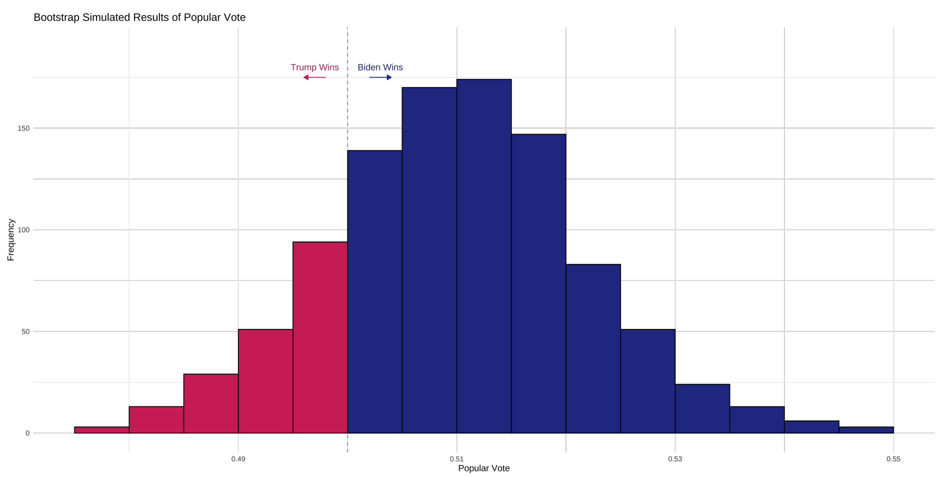

Figure 1: Popular Vote Histogram

As is evident, Biden gets a majority of the votes in most of the bootstrap simulations, around eighty percent of them. However, it is not quite a done deal, as there are still twenty percent of the simulations in which Biden does not get the majority of the votes

We also conducted a bootstrap of each of the states. We wanted to be able to see whether we could predict the 2020 election using bootstrap. Below, we have a simulated electoral map where blue indicates states in which Biden had over fifty percent of the vote in the majority of the bootstrap simulations. Meanwhile red

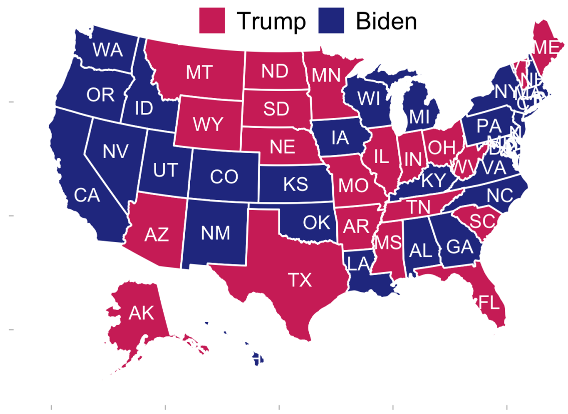

Figure 2: Simulated Electoral Map

indicates states where Biden received less than fifty percent of the vote in most of the simulation, implying that Trump is more likely to win.

This simulation clearly indicates that Biden will win most states. However, this prediction is misguided, as many of the states that Biden wins a majority of the time here he did not in fact win in the actual election. This includes states that are clear Republican strongholds such as Alabama and Idaho, which Biden was not ever going to end up winning in 2020. This is a finding that is evident in real life, as seen in this NYT article, democratic voters are often oversampled in polls which gives a bad representation of the whole country. Thus, we've shown how this is in effect in our analysis.

One final bootstrap we did was a simulation of the popular vote by age. This allows us to get an idea of how each age group will vote, as well as its certainty. Below we have a violin plot, sub grouped by the age bracket of the voter.

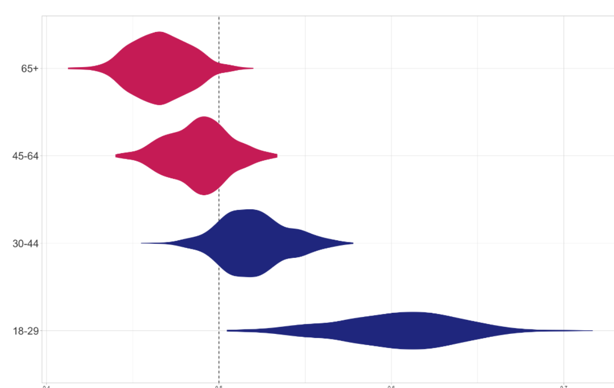

Figure 3: Vote By Age Bracket

We can tell that the older a person gets, the less likely they will vote for Biden. As a result, the vast majority of the simulations show voters over 65 years old giving Biden less than 50 percent of the vote, as well as a majority of voters 45-64 years old doing the same. Meanwhile, Biden gets a majority in most of the 30-44 year old simulations and all of the 18-29 year old ones. However, we can also see that the 18-29 age bracket has by far the most variance. This may be due to how younger voters are less likely to respond to polls, meaning they will be underrepresented and thus are estimates of their voting patterns with be less certain.

Logisitic Model

Bootstrap of Coefficients

After creating our model through the aforementioned model selection process, we wanted to obtain an estimate of the variance of the coefficients. Since bootstrap is a way for us to estimate how much coefficients vary in a model, we decided to conduct a bootstrap on our model. The following chart are violin plots of every coefficient in our model, categorized by question. The colors below indicate whether the median of the coefficient is above or below zero. A blue coefficient means that the median coefficient is above zero, saying that in the majority of simulations responding a question this way means that person is more likely to vote for Biden. Meanwhile, a red coefficient is the reverse.

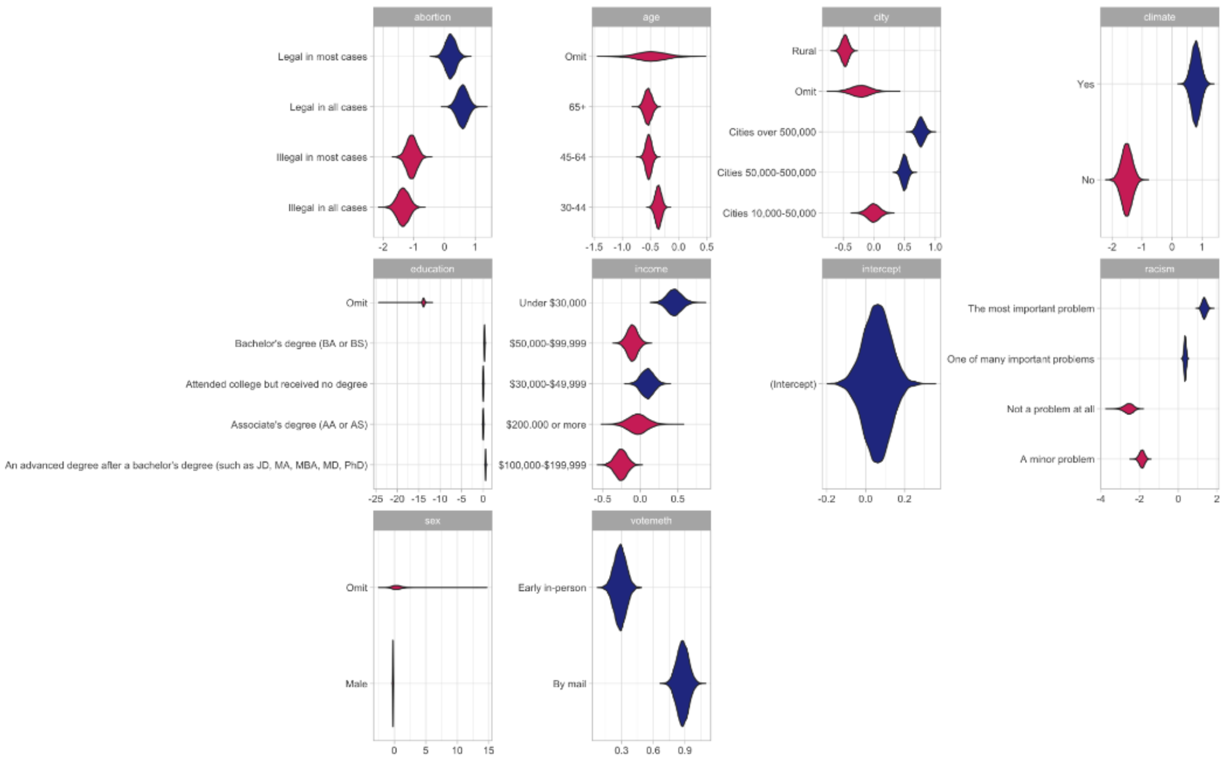

Figure 4: Distribution of Coefficients

A few observations can be made from looking at this chart. One is that the intercept seems to be around zero, with slightly more of the simulations being above, which indicates that our reference levels lead to a voter close to the middle of the political spectrum. However, we can see that some of these coefficients are heavily dependent on the set reference level for that question, such as age and vote method, with all of their coefficients being on only one side of zero.

Other insights from this chart include that the omit response seems to have plenty of variance, especially compared to other responses in the same question. This means that a person omitting answering a question does not tell us much in regards to who they will vote for. We also can see that the climate question has a clear effect, with Yes being completely above zero, and No being completely below zero. Meanwhile, we also can tell that the racism question is similar, with two of the response being clearly above zero and two clearly below zero. Abortion is similar in this respect, but not as extreme, since some of the coefficients appear to cross zero.

Conclusion

In conclusion, this analysis utilized bootstrap techniques to gain insights into the 2020 election, examining the popular vote, state-level predictions, voting patterns by age, and coefficients in a logistic model. The results shed light on important aspects such as oversampling, variable significance, and the predictive capabilities of the models.

When delving into the results, we discovered that certain variables held more significance than others, as indicated by lower p-values. Particularly noteworthy were the variables related to opinions on "Racism" and "Climate." These variables consistently exhibited strong significance across all categories, suggesting their importance in determining whether an individual voted for Biden. This finding adds validity to our model, highlighting the significance of these factors in real-world voting behavior.

It is worth mentioning that our logistic regression model was primarily developed for interpretation purposes rather than accurate prediction. However, recognizing the practical importance of accurate predictions, we are interested in exploring the application of undersampling techniques. By rectifying the issue of democratic overrepresentation, we aim to improve the reliability of our model and assess how it impacts the results.

Additionally, we plan to test alternative tree-based models in comparison to our regression model. This exploration will provide insights into their predictive performance and allow us to evaluate their effectiveness in capturing the complexities of voter behavior. By broadening our analysis and incorporating these modeling approaches, we strive to develop a more robust predictive model for future elections.

In summary, this analysis not only uncovers key findings about oversampling, variable significance, and the impact of certain factors on voting behavior but also emphasizes our commitment to enhancing the predictive capabilities of our model. Through undersampling techniques and the exploration of alternative models, we aim to refine our understanding of the data and create a more accurate and reliable model for predicting election outcomes.