Lara Schuman

Professor Jack Santucci

PSCI T280

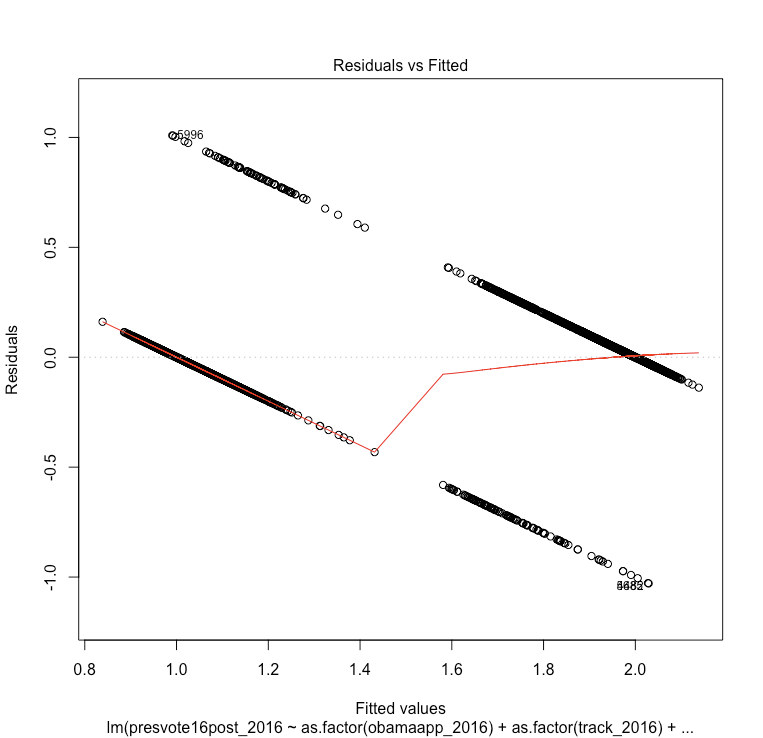
June 9, 2019

How can we use voter survey data?

For my final project I decided to study the voter survey data. I wanted to determine which factors led people to make their decisions as to who they were going to cast their vote for the 2012 and 2016 general elections. I wanted to determine if the survey data was a good predictor in determining who someone voted for. I used a linear model, decision trees, and predictors to help me guide my findings.

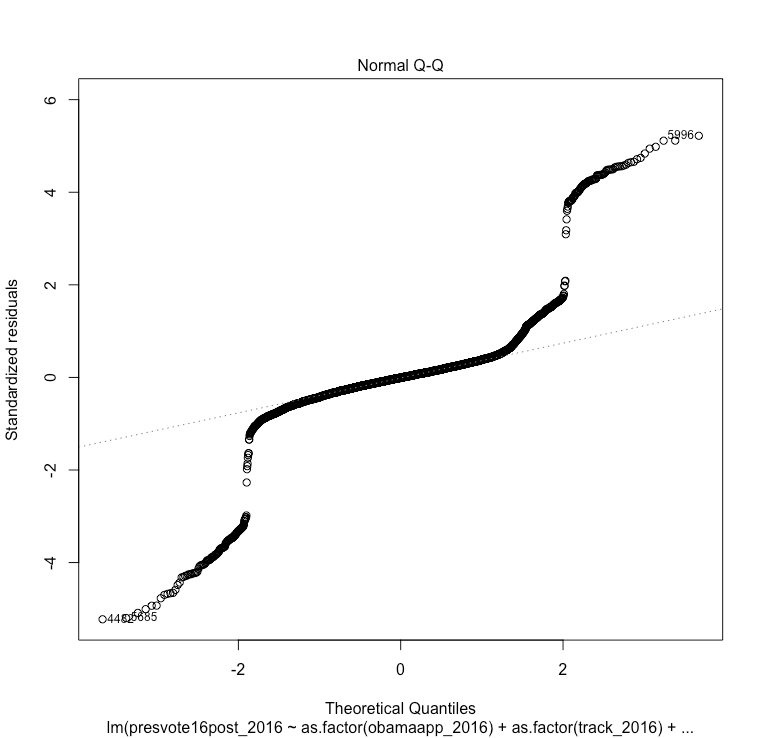
With this data set many questions can be answered. My question is what factors led people to make their decisions in voting in the 2012 and 2016 election. I’m also looking at which factors were more prevalent in their respective election years. I used decision trees and linear modeling to determine the most relevant factors of the survey from a few samples. I chose 19 factors for each model. Each of the factors were the same in both models, to get better accuracy. I also used predictors to determine the accuracy and effectiveness of the models. I compared these models to determine which one was more accurate in predicting who people voted for.

In my findings I had some surprises. A few of the factors I chose for the models were not as statistically significant as I had assumed. For example, in 2016 the importance of how much Democrats or Republicans focused on the distribution of taxes was not as important in 2016 as it was in 2012. This came as a surprise to me because one of Trump’s main points of his campaign was lowering the taxes on the rich. He did this because he knew it would appeal to business people like himself. One of the few bills that passed during Trump’s first two years was the tax bill. It was surprising to see that it had little statistical significance in the 2016 election. Some of the factors more relevant in 2012 than in 2016 are the importance of crime, immigration, health care, taxes, and abortion. The factors that were more relevant in 2016 were the importance of budget deficit, terrorism, economy, and personal finance. Most of these factors were important in both elections, but there were some variations between the elections.

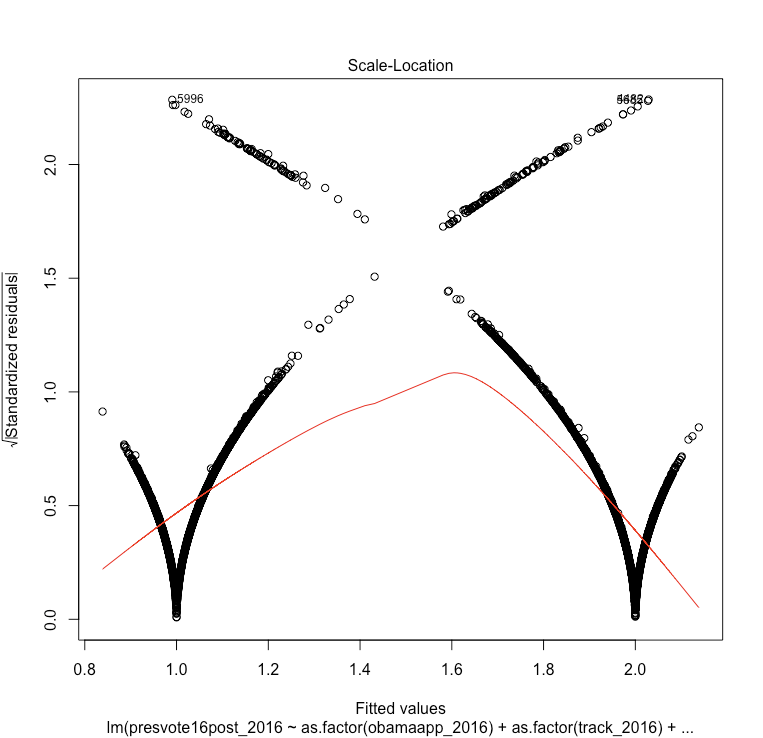
The model also predicts how people vote based on each particular answer; it gives an estimated score, and the score is either closer to one or two. If it is closer to one it means that based on that answer the person probably voted for Clinton and if the number is closer to two they probably voted for Trump. Or in the case of the 2012 election one would represent Obama and a two would represent Romney. These numbers are a great indicator, but they are nothing without testing the accuracy of the predictions. I used a predictor function to test the accuracy of the model. For the 2016 election the linear model was 96 percent accurate, and the decision tree model was 95 percent accurate. By adding different predictors the accuracy fluctuates. If all of the factors could be added to the model the model would probably be closer to 99 percent accurate. (This particular model wouldn’t allow me to input more than 19 factors.) The linear model was a little more accurate in predicting who people would vote for. For the 2012 election the linear model was 93 percent accurate and the decision tree model was 91 percent. Overall the linear model proved to be slightly more accurate than the decision trees. The 2016 election predictor model was a little more accurate than the 2012 predictor model. This may be because of the drastic difference between the elections. In the 2016 elections there was a greater disparity between the parties than in 2012. The voters in the two parties were a lot more polarized in 2016, because of a candidate unlike any other we’ve ever had. This election created a greater divide across opinion in the two parties. With the greater disparity it was probably easier to determine who voted for whom based on a survey. 

Looking at the residuals from the linear model, I was able to get some insight into the reliably of the linear fit analysis. Residuals come from the difference between the observed values and theoretical linear fit values. Residuals can reveal unexplained patterns in the data and act as a check to make sure our linear model is actually linear. There are 4 plots for the residuals: Residuals vs Fitted, Residuals vs Leverage, Normal Q-Q, and Scale-Location.

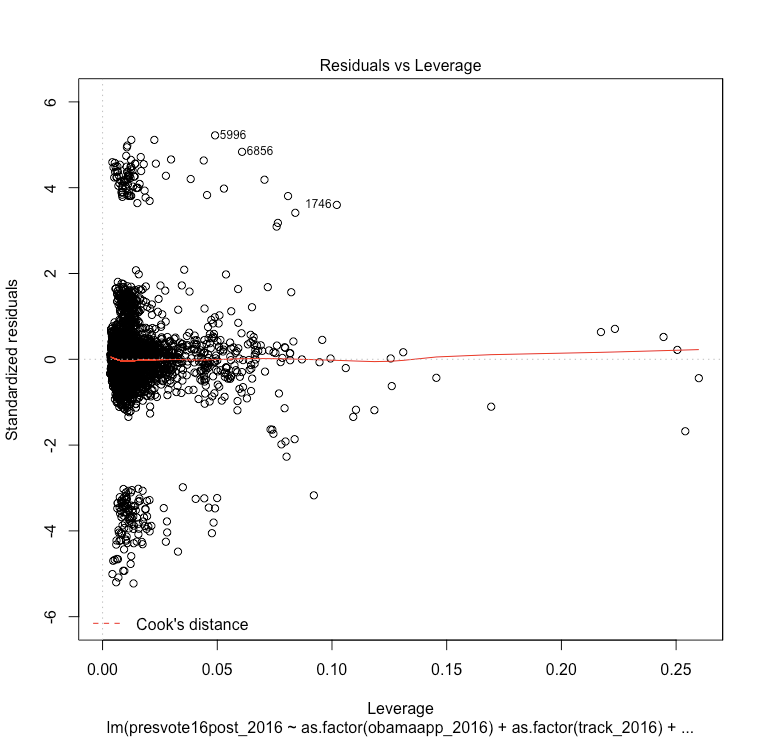
The Residuals vs Fitted plot indicates whether there are non-linear patterns hidden in the data. The data should be equally spread residuals around a straight line. The data is densely matching a straight line for the lower values and then matches another straight line at a value of 1.4. There does not appear to be non-linear relationship in the data, but it is not straightforward to interpret.



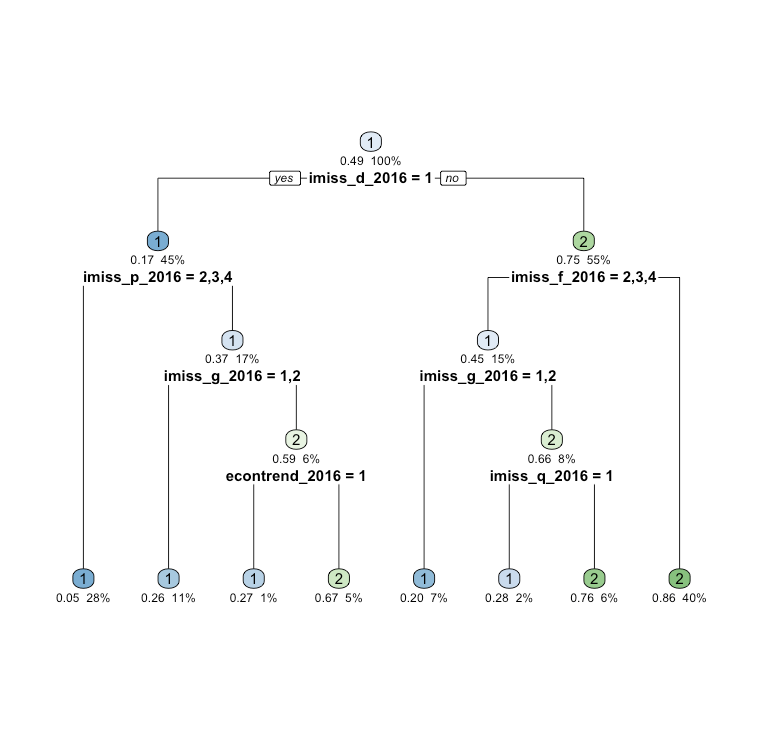
The Normal Q-Q plot shows the data follows the normal distribution as the residues essentially follow a straight line with the standard deviations of +/- 2.



The Scale-Location plot is difficult to interpret. This plot reveals whether the assumption of equal variance is true. The plot indicates there is a possibility of type 1 errors, meaning the data might reach a false positive conclusion (i.e., rejection of the null hypothesis).



The Residuals vs Leverage plot indicates the data is not sensitive to outliers since the data is within the Cook’s distance.

There were some challenges with the data. One was reaching a large group of people who are willing to take this survey. However, a positive thing about the surveys is that each person is weighted according to their location and age. For example, younger people are less likely to take a survey, so they are weighted more heavily than an age group that is more likely to take a survey. Another issue with the data is that it is hard to find an unbiased group of people because the surveys given on the internet probably appeal a subset of the population that may not represent the general population. Without knowing if the survey population represents the general population, it is difficult to know if the results apply to the general population.

If these surveys are conducted before the 2020 election, these models can be a good tool for identifying the most prevalent election issues for the respective parties. This data would be useful in determining what issues people care about most and inform candidates about the issues that are most important to their constituents. By seeing how a voter answers the survey questions, we can determine if they are more likely to vote for the Republican or Democratic candidate.

Decision Tree

