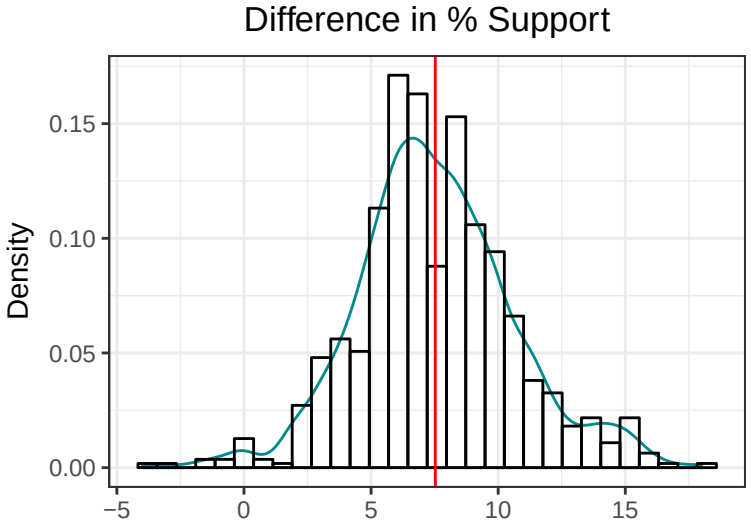
Final Project: Modeling Midterm Congressional Elections

Ricardo Alexandro Aguilar

May 3, 2018

Data Management

We started by merging congressional generic ballet polls and Trumps' approval datasets from the FiveThirtyEight website. The dataset was then subsetted to only contain the variables of interest which also resulted in not missing any data. The dataset's observations are the results and details of national polls from January 1, 2017 to April 19, 2018. A new measure was created for the difference between the percent support for Democrats and Republicans such that a positive value means there is more support for Democrats. This variable was used to determine whether Democrats will have the majority (or super majority) in both the House of Representatives and the Senate as a result of the upcoming midterm elections. This was done by finding the probability that a Democrat will win a seat and multiplying it by the available seats to find the predicted total amount of seats Democrats will have in the House and Senate. These totals are used to predict whether Democrats will have a majority or super majority.



The density plot shows us that the distribution of the difference between the percent support for Democrats and Republicans is normal, with a mean of 7.52, and a standard deviation of 3.12.

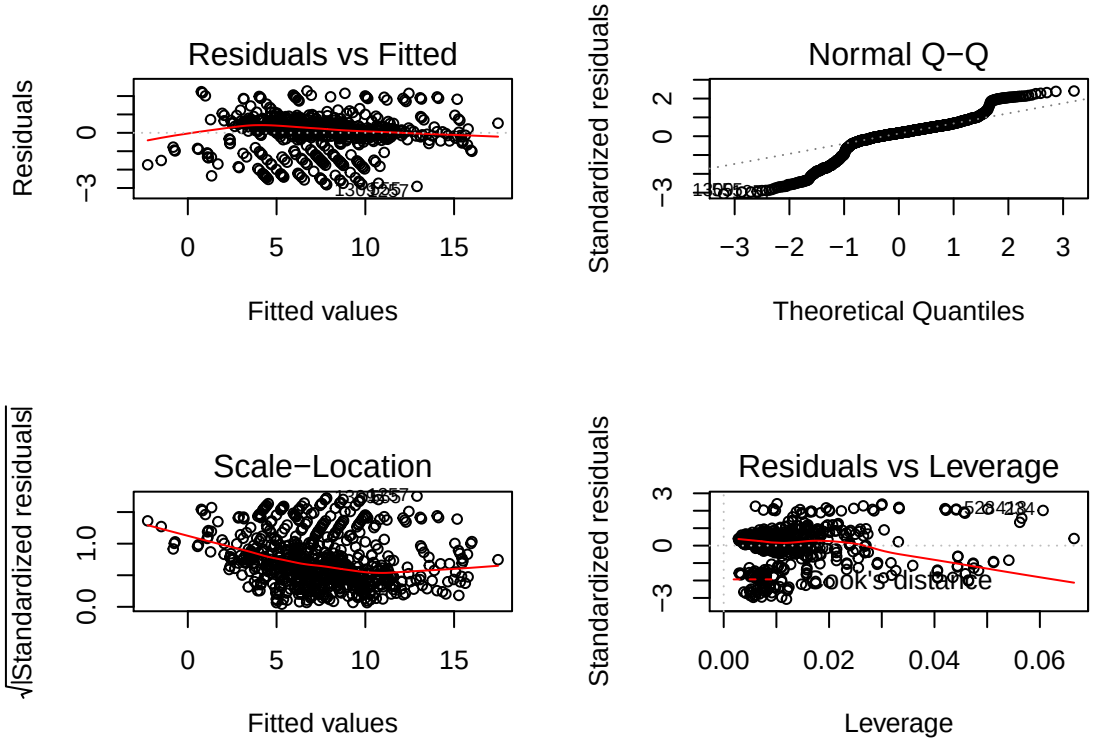
Linear Models

Table 1: RMSE for Both Linear Models

W/ Adjusted Dem/Rep	W/O Adjusted Dem/Rep
0.9476	2.717

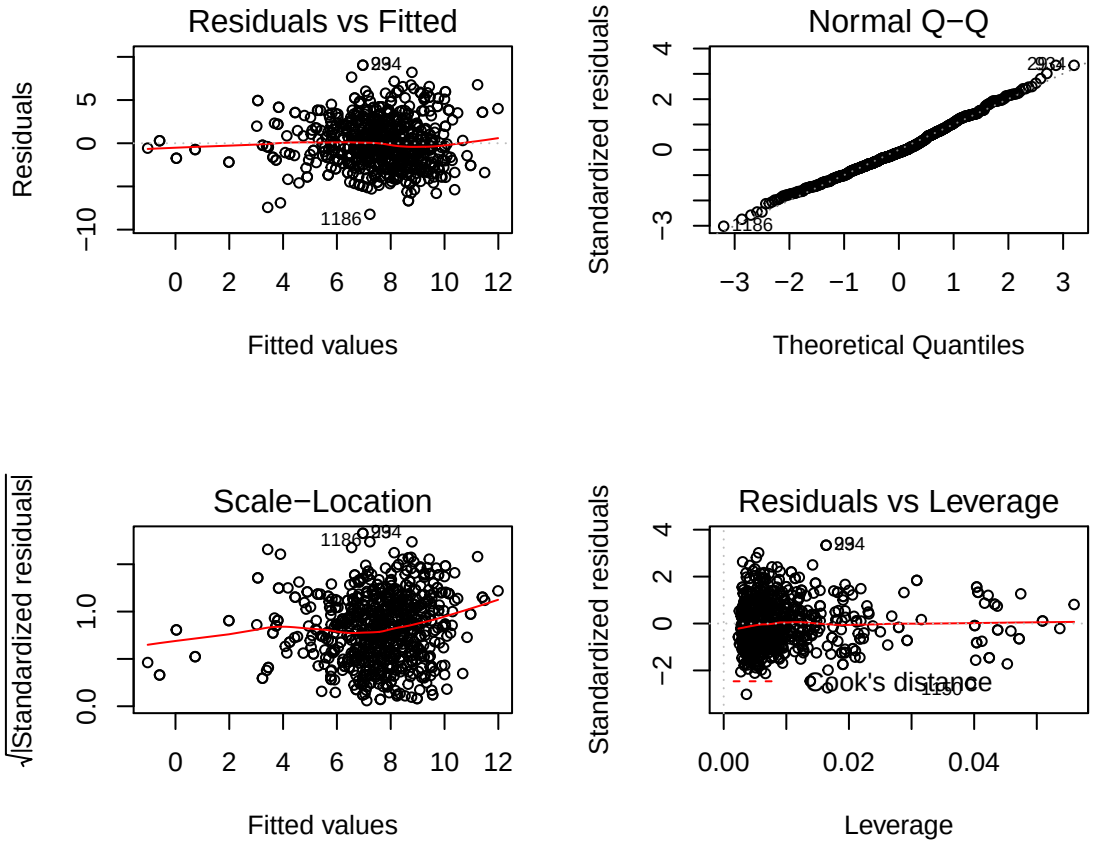
The two linear models used a training set. Both models used population, weight, percent who approve of Trump, the end date of the poll, and the poll's sample size as predictors. However, one of the models includes adjusted support for Democrats (cor = .62, VIF = 1.466) and Republicans (cor = -0.661, VIF = 1.993) as predictors. Polls are weighted on their sample size, pollsters' historical accuracy, and methodological tests. *Population* categorizes polls into three groups: all adults, registered voters, and likely voters. Looking just at the values for RMSE, the model that includes adjusted support for Democrats and Republicans is superior since its RMSE is lower.

W/ Adjusted Dem/Rep



We can see from the Normal Q-Q plot that the residuals deviate from line at lower theoretical value which raises concerns regarding whether the residuals are normally distributed. The Scale-Location plot also raises concerns since the residuals do not seem to be randomly spread out at lower fitted values. This tells us that the model might not work well with the data.

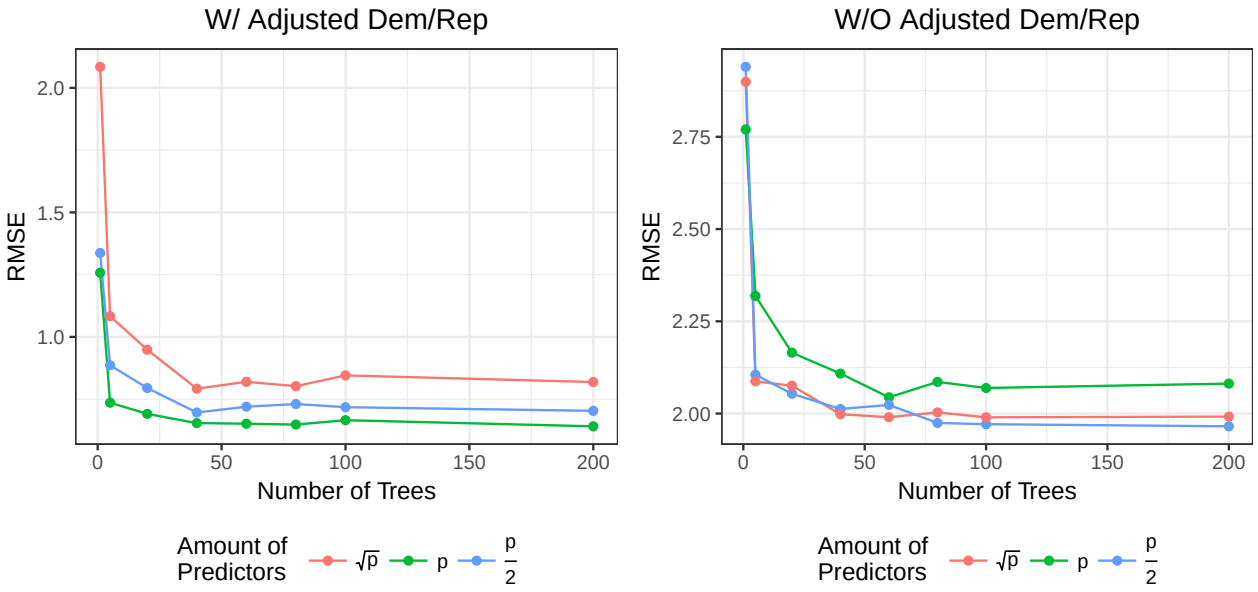
W/O Adjusted Dem/Rep



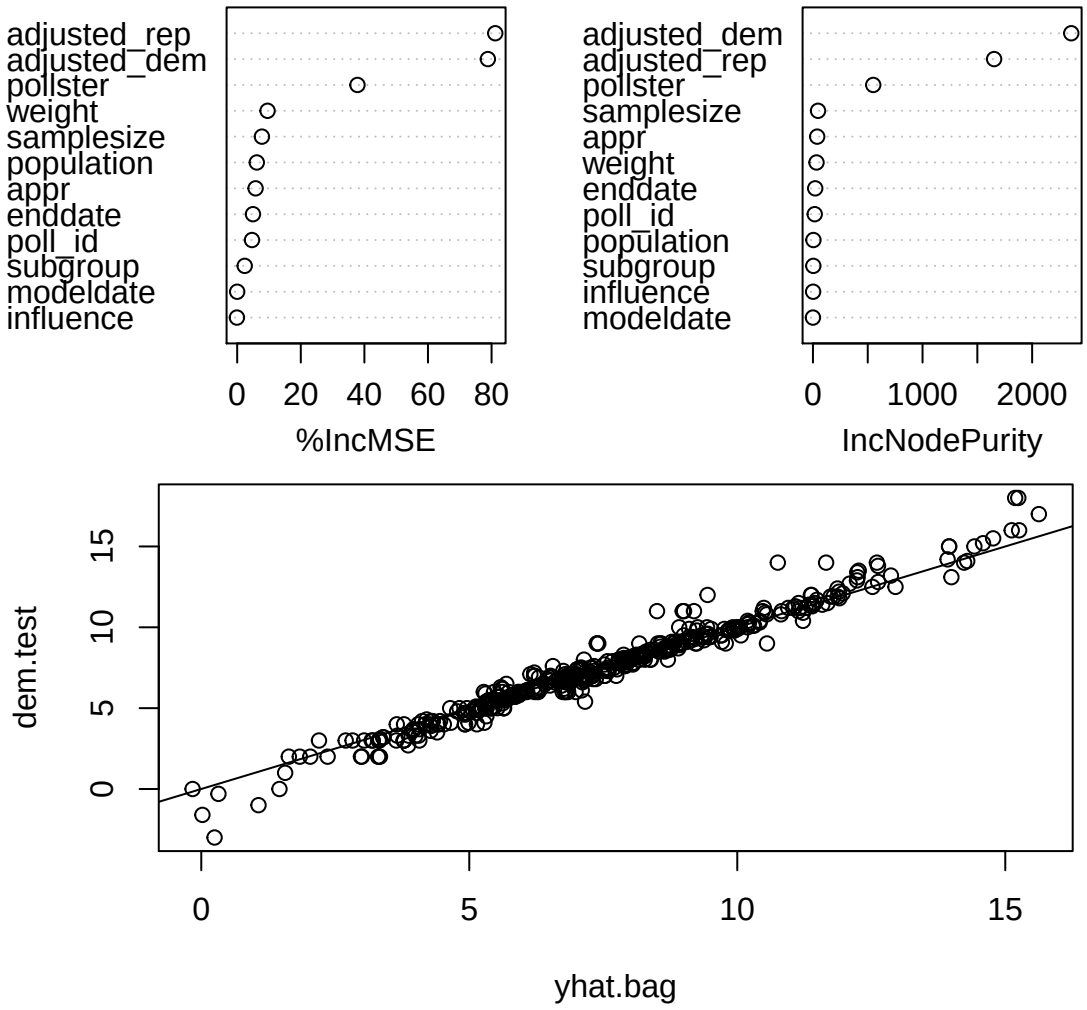
There is no distinctive pattern in the Residuals vs Fitted plot. We can see that the residuals are normally distributed since they do not deviate far from the line. In the Scale-Location plot, the line is horizontal and the residuals seem randomly spread. There are no severe violations in the last plot. The four plots show us that the model works well for the data. However, it is important to note that the residuals are a lot bigger in magnitude than the first model. We will use this model.

Random Forests

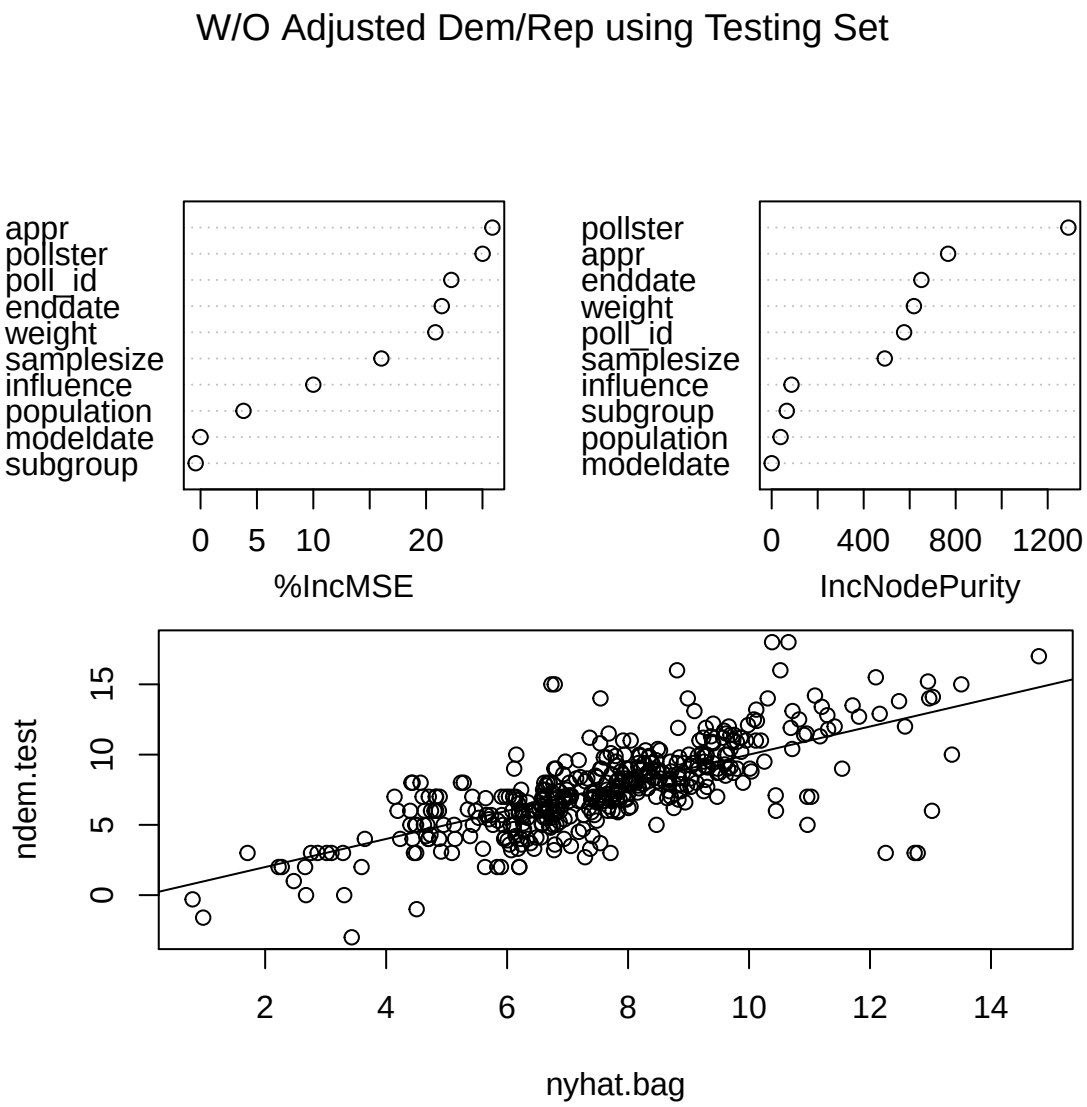
Two models were considered: one with adjusted % support for Democrats and Republicans, and one without.



The *validation set approach* was used as the method for cross-validation by randomly selecting a third of the dataset as the training set, another third as the validation set, and rest as the testing set. Multiple combinations of number of trees (1, 5, 20, 40, 60, 80, 100, and 200) and predictors (p, $\frac{p}{2}$, $\sqrt{p}$ where p is the total possible predictors in the dataset) were considered to attain the lowest RMSE for both models using the validation set. p was set to 14 for the first model and 10 for the second model. In the first plot, we can see the lowest RMSE is attained where 200 trees are used and there are p predictors. In the second plot, we can see the lowest RMSE is attained where 200 trees are used and there are $\frac{p}{2}$ predictors. These are the parameters that will be used in their respective models.



This random forest model is able to explain 94.05% of the variance in the difference between the percent support for Democrats and Republicans. We can see that the variables for the adjusted percent of support for Democrats and Republicans are by far the most important variables in the model while *pollster* is the third most important. In the last plot we can see that the predictions from our model are very close to the "true" values from our testing set.



This random forest model is able to explain 58.69% of the variance in the difference between the percent support for Democrats and Republicans. We can see that the pollster and the percent who approve of Trump (*appr*) are the most important variables in the model. In the last plot we can see that the predictions from our model are not as close to the "true" values from our testing set compared to the previous model.

Table 2: RMSE for Both Random Forest Models

W/ Adjusted Dem/Rep	W/O Adjusted Dem/Rep
0.6193	2.127

The RMSE for both models is calculated using the testing set. The model with adjusted percent support for Democrats and Republicans has a lower RMSE than the one without those variables. However, we will use the second model since the first model has an obvious advantage.

Results

The probability of a Democrat winning a seat was calculated for both the linear model and the random forest model by determining the proportion of predicted values that were greater than the RMSE of their respective models using the testing set. A predicted value of 0 means that a Democrat is predicted to tie with a Republican, and if the value was greater than the RMSE, then it was considered "far enough" from 0 that the Democrat is likely to win. Predictions were only made for populations of likely voters. Since not all adults can vote and not all registered voters will actually vote, a population of likely voters is more representative of the voter turnout. We will define a supermajority in the House of Representatives as having 218 seats and a supermajority in the Senate as having 67 seats. Democrats have 23 seats that will not be up in the 2018 midterm elections, so they must be added to the predicted totals to correctly calculate the amount of seats they will have after the elections. With our definition of a supermajority, Democrats cannot have a supermajority in the Senate since the maximum amount of seats they can possibly have after the elections is 58.

The linear model predicts that the Democrats will have a supermajority (and majority) in the House of Representatives ($.85 * 435 = 370$ seats) and a majority in the Senate ($.85 * 35 = 30$ seats, $30 + 23 = 53$ seats after the elections). The random forest model predicts that the Democrats will have a supermajority (and majority) in the House of Representatives ($.818 * 435 = 356$ seats) and a majority in the Senate ($.818 * 35 = 29$ seats, $29 + 23 = 52$ seats after the elections). It is important to note that the dataset uses national polls, so it does not have any state-level data. This prevents us from creating a much more accurate model that takes into consideration how the elections actually work (the entire nation does not vote for every district's representative and every state's senators). In summary, both the models predicted that Democrats will have a supermajority in the House of Representatives and a majority in the Senate.