Election Prediction Report

Ashley Murray, Hunter Gregory, Matty Pahren, Nathan O'Hara, Scott Heng

10/29/2020

Introduction

Forecasting elections is of great interest, as politicians might use election outcome models to adjust campaign strategies, economists might use them to predict how markets will react, and citizens might use them to decide whether to vote. The 2020 election is one of great importance, and it has seemingly gained even more attention than past elections.

In this report, we seek to model predicting the outcomes of several key races. Specifically, we aim to predict and provide uncertainty predictions for the following: 1. the outcome of the presidential election 2. whether the US Senate remains in Republican control 3. the electoral college vote 4. the outcomes of all NC Congressional elections (the 13 federal Representatives to Congress) 5. the outcome of the NC Senate election

First, we will provide background on the elections and our data. Then, we will describe our methods for all of the models, including a model of who votes in North Carolina, which aims to better the predictive power of our NC senate and House models. Next, we will discuss the results of our models. After that, we will walk through some of the limitations of our models. Finally, the appendices will include more detailed information about our modeling procedures and the data sources used.

Background and Data

Election Overview

On November 3, 2020, all 50 states will hold congressional elections for their respective Representatives as well as the presidential election between Donald Trump and Joe Biden. The senate currently has 53 Republican seats of 100 total seats, and senatorial elections across many states will determine 35 seats, 23 of which are currently under Republican control [1]. Overall, this election is clearly unprecedented due to the current pandemic since the "most severe pandemic in recent history" - the 1918 influenza - was over a century ago and not during a presidential election year [2]. It will be interesting to see if the dramatic increase in mail-in voting will favor different candidates or political parties.

Polling Data

We use polls extensively in nearly all our models. Specifically, we use FiveThirtyEight's polling data for the presidential and US Senate races. We did not end up using these data for North Carolina's elections for the House of Representatives since there were only three of thirteen districts with polls, and each of these districts only had two to four polls.

By considering the presidential polls data, as seen in Fig. 1, we can observe how the probabilities of being elected for both candidates change over time. We see that Biden has consistently had higher probabilities of being elected than Trump since April 2020, while both candidates experience similar degrees of increase in

percentage during their 'Convention Bounce'. Biden appears to consistently have a mean percentage of being elected around 50% and up, and has had a recent surge during October. On the contrary, Trump's percentage estimates have largely been below 45% with a significant drop towards October. From the exploratory data analysis on national polls we can already see that Biden is projected to have an advantage over President Trump.

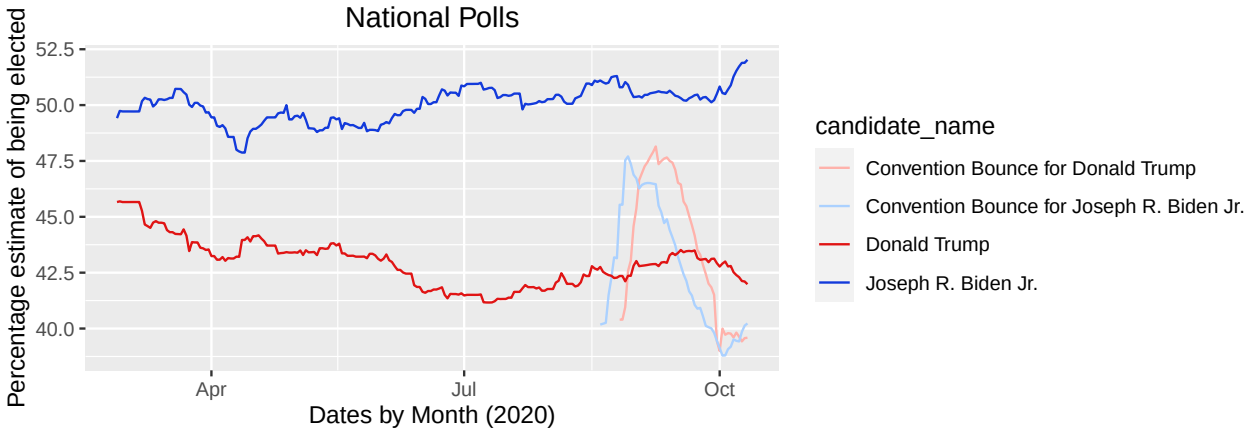

Fig 1.1 National Polls over time for Biden and Trump

However, this does not tell the full picture, as the United States Presidential Elections adopt an Electoral College system, which sometimes can allow a candidate to become President if that candidate did not receive the popular vote nationally, but secured support in key states that contain a large number of electoral votes. These states have a larger influence on the course of the elections and therefore are highly contested between the two candidates. Figure 1.2 shows the polls for both candidates by state over time, revealing which candidates have advantages over the other in certain states and which states require more attention in order to gain the people's support. Fig 1.3 shows the polls in swing states- States that have relatively equal support for either party and therefore can heavily influence the outcome of the election. For the majority of swing states, Biden similarly is favored statistically to be elected, with the exception of Iowa and Ohio.

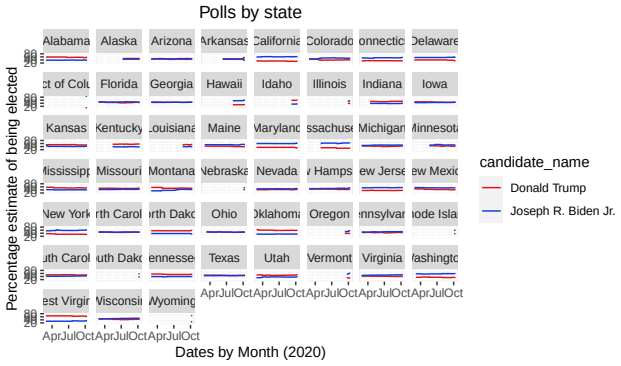

Fig 1.2 State Polls over time for Biden and Trump

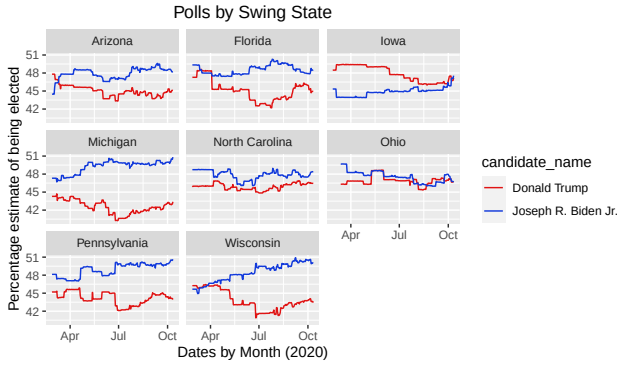

Fig 1.3 Swing State Polls over time for Biden and Trump specifically in the swing states (Democratic or Republican

For senate data, we look at the senate races and their probabilities for being elected (Democrat or Republican). Over the years, we can observe that there are many close senate races as seen in Fig 1.4. In Fig 1.5, we can also see the percentages for being elected into the senate for individual candidates based on their party affiliation. We can observe that overall Democratic candidates tend to have high percentages for being elected into the senate over Republican candidates, and this distinction becomes more prominent closer to the election date.

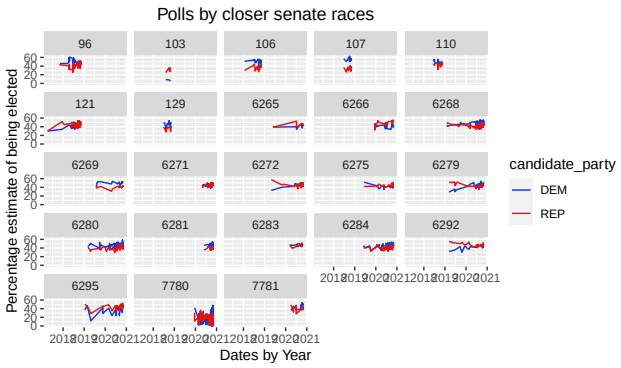

Fig 1.4 Polls by closer senate races

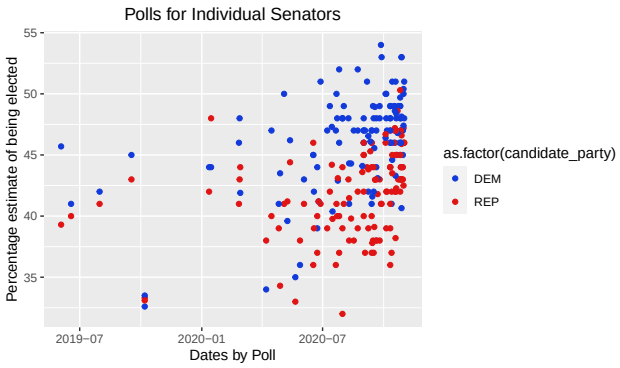

Fig 1.5 Polls for individual Senate races labelled by part

Electoral College Model

The Linzer model is a Bayesian forecast model that specifies how preferences evolve over time and how events are noisy measurements of the underlying preferences.

The purpose of this model is to predict the outcome of the presidential election by looking at electoral college votes. The probability of Trump winning is calculated by looking at the percentage of times in our 1500 simulations Trump gets 270 or more electoral college votes. In modeling the overall outcome of the presidential election, this model also predicts how each state's electoral college vote will go. Each state's electoral college outcome is modeled by looking at polling data from FiveThirtyEight for that state in order to predict what percent of the popular vote both Trump and Biden receive. Whichever candidate receives a larger percentage of the popular vote receives all of the electoral votes for that state. In reality, Maine and Nebraska can have split electoral college votes instead of the all-or-nothing approach other states take, but we chose to also model these states as all-or-nothing for simplicity sake due to the nature of the polling data that was collected.

US Senate Model

Similar to our electoral college model, this model looks at senate race polling data from fivethirtyeight in order to predict the outcome of all current US senate races. The data was filtered to only include general election polling data, so in other words all jungle primary or runoff election polls were removed. Additionally, the data was filtered to only include polls less than a year out from the election. After filtering, we were left with data for all states where senate elections were being held except for Arkansas, Rhode Island, South Dakota, Louisiana, and the second Georgia Race. However, most of these races were predicted to be blowout races, with Arkansas, South Dakota, and Louisiana deemed as safely Republican and Rhode Island deemed safe Democrat. Georgia was the only race we were missing data for that seemed close, but we chose to use the Georgia predictions generated from the other Georgia race we did have data on, since senators are elected by the whole state and it is likely that the same people are voting for both senators. Lastly, polling data for third-parties was filtered out, and the remaining percentages were standardized. In other words, we took the final Democrat and Republican share for each poll, and created a final variable equal to Republican share divided by the total share occupied by Republican and Democrat percentages, to figure out what percentage of the two-party share Republicans were expected to win. Since this is the case, we counted these states as a corresponding Republican or Democratic win in each of our simulations. After making predictions for each individual race, we aggregate the results and look at the percentage of times where Republicans win more than 21 seats, where 21 is the number of seats Republicans need to win in order to have a 50/50 Republican/Democrat split in the senate.

North Carolina Senate Model

The North Carolina Senate Model we constructed is a multi-level model. We include a random effect for each county in North Carolina, and the response variable we used is the percentage of registered Republicans in 2018. We chose to use the year 2018 here since this was the year we had most recent data for. Additional predictors are North Carolina Census variables aggregated at the county level, including the percentage of people who live 15 minutes away from where they were, the percentage of people who are foreign-born, the percentage of single parents, the percentage of people who have at least some college education, the percentage of people living below the poverty line, the average rent of a two-bedroom apartment in the county, the average annual job growth from 2004-2013, and then the share of the population that is white, black, asian, and hispanic. All of these variables are taken from the 2010 Census, with the exception of annual average job growth and the rent of a two-bedroom apartment, which was measured in 2015. After the model was used to calculate the percentage of Republicans per county, this data was aggregated and weighted using the total number of people that voted in the 2016 election to come up with a final estimate of the percentage of people who would vote for Republican Thom Tillis.

North Carolina House Model

The North Carolina House Model is almost identical to our senate model. We constructed a multi-level model with a random effect for each county in North Carolina, and the response variable we used is the percentage of registered Republicans in 2018. We chose to use the year 2018 here since this was the year we had most recent data for. Additional predictors are North Carolina Census variables aggregated at the county level, including the percentage of people who live 15 minutes away from where they work, the percentage of people who are foreign-born, the percentage of single parents, the percentage of people who have at least some college education, the percentage of people living below the poverty line, the average rent of a two-bedroom apartment in the county, the average annual job growth from 2004-2013, and then the share of the population that is white, black, asian, and hispanic. All of these variables are taken from the 2010 Census, with the exception of annual average job growth and the rent of a two-bedroom apartment, which was measured in 2015. Additionally, we included a categorical variable which says what party the incumbent belongs to in a particular district. All of the incumbents were Republican or Democrat with the exception of District 11, which had a vacant seat. After the model was used to calculate the percentage of Republicans per county, this data was aggregated and weighted using the total number of people that voted in the 2016 election to come up with a final estimate of the percentage of people who voted Republican in each district. For districts that were split between counties, we divided said counties in half and assigned half of the population to one district and half to the other. We attempted to come up with a better division than this arbitrary half/half split, but we were unable to collect data telling us which zip codes, Census tracts, or other smaller regions belonged to which Congressional district.

Results

Electoral College Model Results

Our model predicts that Trump has a 0% chance of winning. This number seems very low, but we think this is primarily due to the fact that more recent polls are weighted more, and in the most recent polls Biden edges out Trump in many key states. The state predicted to have the highest share of Trump voters is West Virginia, where Trump is predicted to have 65.5% of the two-party share vote, with 95% credible interval (61.7%, 68.4%). The region predicted to have the lowest share of Trump voters is Washington D.C., and our model predicts that Trump will only get 8.9% of the two-party share vote, with 95% credible interval (2.9%, 16.9%).

Electoral College Model Validation

For model validation, we plot traceplots for all the beta coefficients. From the traceplots, we can observe that there is randomness across iterations, signifying that convergence has been reached with 1500 iterations. Modifying our initial burn-in rate (250, 500,1000) did not seem to affect model convergence, and thus decided to stick to the original burn-in rate of 500. We also tried different quantities of iterations to see if it would improve the model. We concluded that 1500 simulations produced Rhat values close to 1, signifying that the chains have mixed well. Similar good results are shown in the lag-1 scatterplots where there seems to be a lot of randomness which means that the model sampler is sampling the entire space in an uncorrelated manner. The ACF plots show little correlation between new samples of beta with previous samples which is also desirable. To see these plots please refer to the appendix.

On top of standard model validation tools, we also performed out-of sample validation by implementing the same model on 2016 data to predict the results of the Clinton-Trump When validating on 2016 election data, we predicted that Clinton had a 87.1% chance of winning. Clinton didn't end up winning the election, however many election prediction models forecasted that she would have a landslide victory. Thus, our model is seemingly consistent with what experts were predicting, which is another aspect of good model fit for the study.

US Senate Model Results

We predict that the Republicans have a 0.2% chance of keeping control of the senate. The Republican party wins between 10 and 18 seats most of the time in our simulations, which is shy of the 22 seats needed to win a majority.

US Senate Model Validation

Traceplots for this model indicate that convergence was also reached for this model. Additionally, lag-1 scatterplots and autocorrelation plots show that there is no need to be concerned about non-random data or correlations. All of our Rhat values are close to 1, again indicating that the chains have mixed well.

When using our model to predict the chance of Republicans controlling the senate after the 2018 midterm elections, we get a probability of 71.4%. This estimate seems to be close to what many other well-known political models forecasted, and the Republicans did indeed keep control of the senate after this election.

North Carolina Senate Model Results

We predict that Republican Thom Tillis will win 45.6% of the two-party vote-share, and we are 95% confident that this prediction will fall between 40.2% and 50.9%. Since our interval contains 50%, we can not be sure that he will lose the race to Cal Cunningham.

North Carolina Senate Model Validation

The residual plots for this model show a random scatter around zero, and the qq plot of residuals shows that they are also approximately normally distributed.

However, when using the 2016 election as an out-of-sample validation, we predict that the Republican candidate would win 45.6% of the popular vote, with a 95% prediction interval from 40.4% to 50.7%. However, in reality, Republican candidate Richard Burr won the race with 51.1% of the popular vote, and Democratic candidate Deborah K. Ross fell with only 45.4% of the vote [4]. This meant that he won 53.0% of the two-party share vote. Thus, our model doesn't hold up perfectly in this out of sample validation.

For additional sensitivity analysis, we can compare our senate estimates from this model to the ones that we used to factor into our US Senate model. In the Linzer model, North Carolina was predicted to have a 47.9% chance of going Republican, with interval 46.0% to 50.0%. This interval is contained within the interval we calculated from our NC Senate model, indicating that the two model predictions converge, albeit the variance is higher for the NC Senate model.

North Carolina House Model Results

We predict that the Republican running in District 1 will receive 29.7% of the two-party share (24.7%, 34.6%), and 40.4% in District 2 (33.1%, 48.3%). In District 3, we expect them to win 49.4% of the two-party share (43.6%, 55.2%), 30.4% in District 4 (23.4%, 38.0%), and 59.1% in District 5 (54.8%, 63.3%). For District 6, our prediction is 40.0% (35.4%, 45.1%), followed by 52.2% in District 7 (47.3%, 57.1%), and 40.6% in District 8 (42.6%, 51.9%). In District 9, we expect the Republican to win 40.6% of the two-party share (34.1%, 47.2%) and 60.8% in District 10 (57.7%, 64.8%). Finally, in District 11, we expect this to be a 49.8% Republican share (45.4%, 54.2%), 0% Republican share in District 12, and a 60.6% Republican share in District 13 (57.0%, 64.1%). District 12 is 0% because the Democrat is running unopposed, and there appears to be no write-in campaign. Districts 3, 7, 8, and 11 appear to be the ones without clear winners, as their prediction intervals all contain 50%.

North Carolina House Model Validation

Again, the residual plots for this model show a random scatter around zero, and the qq plot of residuals shows that they are also approximately normally distributed.

For the 2016 Congressional elections, our model predicted District 1 would receive 29.4% of the vote (24.7%, 34.0%), District 2 would receive 42.7% (35.4%, 50.1%), District 3 would receive 48.3% (42.8%, 53.9%), District 4 would receive 32.7% (25.7%, 40.0%), District 5 would receive 57.9% (53.7%, 62.2%), District 6 would receive 39.9% (35.2%, 44.6%), District 7 would receive 51.6% (46.9%, 56.1%), District 8 would receive 46.9% (42.4%, 51.3%), District 9 would receive 41.1% (34.8%, 47.5%), District 10 would receive 59.3% (55.4%, 63.1%), District 11 would receive 50.4% (46.0%, 54.5%), District 12 would receive 36.9% (30.5%, 43.5%), and finally District 13 would receive 58.9% (55.5%, 62.4%). In reality, Republicans won Districts 2, 3, 5, 6, 7, 8, 9, 10, 11, and 13, while Democrats won Districts 1, 4, and 12 [5]. We would have predicted Democrat wins in Districts 2, 3, 6, 8, 9 where this did not actually happen. Part of the reason for this is that North Carolina's Congressional Districts were re-drawn after a gerrymandering lawsuit, so 2020 is the first election where these new districts will be used. Since our model groups districts according to the new lines, it makes sense that our model wouldn't have great predictive power on old data.

Who Votes in North Carolina Model

An additional model was created in order to better understand who votes in North Carolina, incorporating information from the North Carolina voter registration database. Ultimately, we modeled the proportion of registered voters who voted in the 2016 and 2018 elections, validating on data from the 2012 and 2014 elections. The predictors included in the voter registration database included race, gender, ethnicity, age group, county, party affiliation, and whether the election was a presidential or midterm election.

We implemented a binomial mixed-effects model modeling the proportion of registered voter turnout using the aforementioned predictors as fixed effects, as well as random effects on the same predictors as well as an intercept to quantify heterogeneity across congressional districts. Using this model, we were able to understand the trends in voter turnout both across North Carolina broadly, and within competitive congressional districts. For more information about the model and analysis of its results, please consult our separate paper on Who Votes in North Carolina.

Additinal Discussion

Limitations

One common limitation across all models was the quality of the data. As polling data often has a disproportionate amount of observations for groups and variables, there will be a degree of uncertainty in our predictions as our sample size does not truly reflect the true opinions of people and certain demographics. There can be underrepresented or misrepresented groups of identities that can dramatically affect the effectiveness of the model.

Another prominent limitation in our Linzer models is that is a consequence of the quality of the data was the unrealistic weightage of polls closer to the election. There is a significant increase in influence that recent polls have on the predictions of the linzer model, which is perhaps why our results for the senate and electoral college are near unanimous as the results of the most recent poll before the election would have a heavy influence on the predictions made by the Linzer model.

Bibliography

[1] 2020 Senate Election Interactive Map. (n.d.). Retrieved from https://www.270towin.com/2020-senate-election/ [2] 1918 Pandemic (H1N1 virus). (2019, March 20). Retrieved from https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html#:~:text=The 1918 influenza pandemic was,spread worldwide during 1918-1919 [3] 2016 United States Senate election in North Carolina. (2020, July 31). Retrieved from https://en.wikipedia.org/wiki/2016_United_States_Senate_election_in_North_Carolina [4] 2016 United States House of Representatives elections in North Carolina. (2020, August 16). Retrieved from https://en.wikipedia.org/wiki/2016_United_States_House_of_Representatives_elections_in_North_Carolina

Appendix A: Models Used

Who Votes Model

Model Purpose

This model aims to predict which groups of people in North Carolina are most likely to vote based on demographic characteristics like race, gender, ethnicity, age, and county of residence. This model also factors in election year, and it can determine whether groups were more likely to vote in the 2018 midterm elections or the 2016 election.

Model Structure

Estimates From Model

Electoral College Model

Model Purpose

The purpose of this model is to predict the outcome of the 2020 presidential race. This is accomplished by simulating the outcome of electoral college votes for each state based on polling data and then calculating the total number of times where Trump gets 270 or more electoral votes. This percentage is our probability that Trump wins the election.

Model Structure

We model the percent Republican support y_k for each poll k under the following Bayesian model:

FIXME!!!!

$$\begin{aligned} y_k &\sim Binom(\beta_{i,s[k]}, \sigma^2_{yj}) \\ logit(\pi_{ij}) &= \beta_{ij} + \delta_j \\ for \;\; j > 1 : \beta_{ij} &\sim N(\beta_{i,j-1}, \sigma^2_\beta) \\ \delta_j &\sim N(\delta_{j-1}, \sigma^2_\delta) \\ for \;\; j = 1 : \beta_{i1} &\sim N(logit(h_i), s_i^2) \\ \delta_1 &= 0 \end{aligned}$$

Estimates From Model

add Beta estimates/ JAGS output for all 50 states

US Senate Model

Model Purpose

This model predicts the probability that the senate will remain in Republican control. This is accomplished by simulating the outcomes of all 35 senate races based on polling data and calculating the total number of times where Republicans win more than 21 of these seats, as 21 is the number of seats where the senate remains in a 50/50 Democrat/Republican split. The percentage of our simulations that have Republicans winning over 21 seats is the probability that they will keep control of the senate.

Model Structure

Estimates From Model

add Beta estimates/ JAGS output

North Carolina Senate Model

Model Purpose

The purpose of this model is to predict the outcome of the 13 races in North Carolina that will determine who gets elected to the US House of Representatives. In order to achieve this prediction, we use the percentage of Republican votes (out of all Democratic and Republic votes in a given county) as our response variable as a proxy for the number of people we expect to vote Republican, and predictor variables include economic, race, and education data from the North Carolina Census.

Model Structure

We used a linear model with a random effect for county.

$$\begin{aligned} \textit{Republican Share}_i = \alpha_i + \beta_{\textit{travel time}_i} + \beta_{\textit{foreign share}_i} + \beta_{\textit{single parent share}_i} + \\ \beta_{\textit{fraction college}_i} + \beta_{\textit{poor share}_i} + \beta_{\textit{two bed rent}_i} + \beta_{\textit{job growth}_i} + \\ \beta_{\textit{share white}_i} + \beta_{\textit{share black}_i} + \beta_{\textit{share hispanic}_i} + \beta_{\textit{share asian}_i} + \\ \beta_{\textit{incumbent party Republican}_i} + \beta_{\textit{incumbent party None}_i} + \epsilon_i \\ \textit{where } \alpha_i \sim N(\gamma_0, \tau^2) \\ \textit{and } \epsilon_i \sim N(0, \sigma^2) \end{aligned}$$

where i represents a given county. The republican share for a district j is then calculated as

$$\textit{Republican District}_j = \sum_{i \in \textit{counties}_j} t_i^*$$

where $\textit{counties}_j$ is the set of counties in district j, and $t_i^* = \frac{t_i}{2I(\textit{split}_i)}$, where t_i is the total number of voters in county i, I is the indicator function, and $\textit{split}_i$ is a boolean of whether county i is divided into two districts. Each split county was examined individually and looked geographically split in half for each district, so dividing the total by 2 is a reasonable approximation we can make with our given data. The Republican was considered to win the district if $\textit{Republican District}_j > 0.5$.

Estimates From Model

print out prediction table

North Carolina Senate Model

Model Purpose

The purpose of this model is to predict the outcome of the North Carolina senate race between Democrat Cal Cunningham and Republican Thom Tillis. Predictor variables are the same as those in the model above, and similarly, we use the percentage of Republican votes (out of all Democratic and Republic votes in a given county) as our response variable as a proxy for the number of people we expect to vote for Thom Tillis.

Model Structure

This model is identical to the House model above, except for each $\textit{Republican Share}_i$, we do not include predictors based on incumbent party. That is, we have

$$\begin{aligned} \textit{Republican Share}_i = \alpha_i + \beta_{\textit{travel time}_i} + \beta_{\textit{foreign share}_i} + \beta_{\textit{single parent share}_i} + \\ \beta_{\textit{fraction college}_i} + \beta_{\textit{poor share}_i} + \beta_{\textit{two bed rent}_i} + \beta_{\textit{job growth}_i} + \\ \beta_{\textit{share white}_i} + \beta_{\textit{share black}_i} + \beta_{\textit{share hispanic}_i} + \beta_{\textit{share asian}_i} \\ \textit{where } \alpha_i \sim N(\gamma_0, \tau^2) \\ \textit{and } \epsilon_i \sim N(0, \sigma^2) \end{aligned}$$

We calculated the overall Republican share for all districts simply as

$$\textit{Tillis} = \sum_{j=1}^{13} \textit{Republican District}_j$$

. Tillis was considered to win the election if $\textit{Tillis} > 0.5$.

Estimates From Model

print out prediction table

Appendix B: Data Sources