Group 6
Alexander Sahn
POLI 381: Data in Politics II
9 December 2022

Data-Driven Red Teaming

Introduction

President Joseph Biden came into office during one of the most unprecedented times in United States history. Over the past two years, we witnessed monumental historic events that have undoubtedly changed the trajectory of the country. Consequently, these events held significant influence over the 2022 midterm elections, the results of which we have been investigating. This year, 34 seats in the US Senate were up for taking. From international events like the COVID-19 pandemic and the war in Ukraine to national crises like sky-rocketing inflation and political upheaval regarding women's rights and gun legislation, the American public had much to consider at the polls. The outcome of this election is vital, as the livelihoods of countless Americans depend on the ballot. Whether it be the right to receive reproductive and gender-affirming healthcare or the freedom to teach our youth unbiased history, there is a lot at stake. That being said, we have created a model to project the results within each of the 34 states holding senate elections, determining whether we expect the Democratic or Republican party to take the lead. Furthermore, we compare our projection to the actual results of the election.

To approach this research question, our group developed a regression model. The model incorporates multiple variables to account for the many factors that dictate an election. Predicting an election can be difficult because there are countless determinants that impact voters. The question required us to select the most relevant of these. We considered: the year and state of the election; whether there is an incumbent and if there is, the party of the incumbent; the percentage of unemployment within the state; the percentage of inflation across the country; the proportion of the urban or rural population in a state and the party of the President. We regressed the vote share of Democrats in the senate against these variables to predict the election outcome in each of the states. From this, we also created visualizations to represent the party we projected to win across the country. As for uncertainty, we computed the confidence intervals for the vote share percentages and a residual plot against the true election outcomes; we also interpreted our correlation coefficient. Once we calculated these results, we created residuals from the real election results to observe our accuracy. We concluded that our results were fairly true to the actual election results. Our model calculated that Democrats would win 13 states and Republicans would win 18 states, with three races too close to decisively predict. The actual outcome, Democrats won in 13 states and Republicans won in 18 states. Of those states, our specific states' victories aligned with the true victories for the most part. We also conducted simulations for the toss-up state's elections to see who is more likely to win. We determined that

the error within our procedure was a consequence of unpredictability within each race; because this was merely a prediction, it limited our analytical capabilities and overall accuracy. However, this provided us with follow-up research questions regarding electoral politics. We brainstormed other variables that could be integrated into the regression to further control for potential confounders and increase the model's precision.

Analysis

We were charged with the task of predicting an election forecast for the 2022 Midterm Election. To construct the model for forecasting the expected vote share, we built a linear regression (since vote share is additive) of the fraction of the major party vote received by the Democrats in historical elections on several other economic and political indicators that were available months and years before the elections. This detailed regression included data from Senate races from 1976 to 2020 for various variables like incumbents, the popularity of the president, unemployment data, inflation, population, and etc. The model created from this regression was then used to generate predicted values for the 2022 races, which were then plotted among various visual mediums along with the level of confidence for each value.

Building the Regression Model: Compiling Data

First, we will break down why we decided to include the data we did in our model and where we sourced it from. The first set of data we had to collect was the previous results of the Senate election from 1976 to 2020, which were easily available at the MIT election lab. This data set included information on the candidate's name, vote return, and party. This data set is highly crucial as it gives us our dependent variable, vote share in a specific state at a specific time, and independent variables, vote share from previous time periods. Along with this data set, we had to determine if the candidate was an incumbent or not, which we then merged from a Harvard database of Hainmueller, Snyder, and Hall (2015). This variable is incredibly important as it has repeatedly been shown that being an incumbent in an election gives candidates insurmountable advantages and they are more than likely to win their election than be unseated (Ansolabehere 2002). Next, we wanted to have the measure of partisanship of each state from the last presidential election because if a state voted for a Democrat president, they might be more likely to vote for a Democrat senator. This is an adequate measure of ideology for each state that will influence their senatorial vote. With this, we also needed approval statistics for the president. It is not only important to know who the president was at the time, but also how well-liked he was prior to the elections. Despite some exceptions, historically, a president's approval is at its highest during their inauguration with regular increases and decreases during their tenure. We surmise that how well a president is doing will vitally impact the chances of candidates in lower races. A Democratic candidate running during the presidency of a popular Democratic president will perform better, while the same candidate running under a popular Republican president will do worse. The approval data is acquired from the American Presidency Project, and it includes

the average of Presidential approval polls in the year to date of the election as well as the party of the President.

Next, we acquired a series of economic measure variables because historically and intrinsically, economic conditions are critical factors voters consider when voting for candidates. An average voter may have to research or seek out information on where a candidate stands on policies and laws, but they already know the prices of regular household items and gas (a good indicator of inflation), or they know how the job market is going around them (if they or their friends have lost jobs in the past few years) (Wright 2012). They can make an instinctive choice about which party they want to support for their Senator based on the economy they are experiencing (Palmer 1999). We elected to find several unemployment statistics. One is the snapshot measure, which is the unemployment rate one month out from the election, another is an average unemployment measure in the year before the election, and the third one is a change in unemployment from October the year before the election until October the year of the election. All of this data is collected from the Current Population Survey data set compiled by the St. Louis Federal Reserve. Another economic measure we chose to focus on was inflation. Inflation is a key variable because it can be used as a predictive and a control variable in our model. If inflation is high and the incumbent is a Democrat, then they might be blamed and perform worse in the election; vice versa, if there have been steady prices and the incumbent is a Republican, they might perform better in their election. This may have a confounding effect, so inflation could be a key variable that needs to be controlled for. This data was collected from the World Bank, and it included year-to-year inflation change from 1970 to 2021 in the US. We had to manually insert the 2022 figures, but we will expand upon that later.

Finally, we wanted to include another potential control variable that will account for the proportion of urban versus rural voters in a state. It is common knowledge that urban areas or cities with high populations tend to lean Democratic, whereas rural areas with lower population densities tend to lean Republican (University 2019). The National Historic Geographical Information System (NHGIS) stores census data from 1970 to 2010 that gives the population breakdown for each state. Now that we had gathered all of the individual data we wanted to use in our model, we had to compile them into one data set with some expert manipulation.

Building the Regression Model: Coding of Variables

In this section, we will break down and expand upon each relevant variable in our dataset, explaining how we coded it and why.

Year. This is a categorical identifying the year of each row of observations. The years identify only the years there was a Senatorial election between 1976 to 2022. Used to identify/organize each observation.

State. This is another categorical variable identifying the state in each of the observations. They are abbreviated and used to identify/label the observations.

Democratic senate vote share. This is our dependent and one of the most crucial variables. This is the proportion of votes in a given race that the Democratic candidate received,

and it is a numeric, continuous variable. We used this variable to determine the influence of our independent variables and construct a model as accurately as possible.

Open seat. This is a dummy variable taking a value of 1 if the seat is open, meaning there is no incumbent, or a value of 0, meaning the seat has an incumbent running. This variable provides useful context and influences another variable in the data set that will be used, inc_party.

Incumbent party or inc_party. This is a categorical variable that's identifying the party of the incumbent *if there is one*. There's only value in it if open_seat = 0. The values are either "R" or "D" indicating the party of the incumbent. This variable is used to code a dummy variable that will be used in the regression, inc_d.

Incumbent democrat or inc_d. This is the categorical dummy variable for the incumbent party variable. The democratic incumbent is coded as 1, the republican as 0, and 0.5 if there is no incumbent. This variable was a bit difficult to code, as we wanted to represent all possibilities (Democratic incumbent, Republican incumbent, or open seat) so there would be no NA values. Consequently, an open seat has a "value" of 0.5, which could lead to some complications as the computer interprets there being an "effect" of going between an open seat and an incumbent seat, or vice versa. This could potentially have some minor effects, but given there was no viable alternative to account for incumbency without having NA values we felt this was the best option available.

Unemployment change. This is a continuous variable containing the change in a state's unemployment from October the year before the election to October the year of the election. So, if unemployment was 10% in October 1981, and 5% in October 1982, this value would be -5%. This is useful to determine how the economic picture in a state is changing. Are things getting better, worse, or staying the same? This change is another economic indicator that voters will likely notice so it's important to include it in our regression model.

Democratic presidential vote share. This continuous variable consists of the percentage of the vote that the Democratic candidate got in the previous election. This is a critical measure to have as it speaks to the ideology and partisanship of a state.

Party president. This is a categorical variable taking the value of "R" if the president is Republican and "D" if he is a Democrat. This particular variable is not used, but it is re-coded into a dummy variable, which is used in the regression model.

President republican or pres_r. This is the dummy variable for the Party president that is coded as 1 if the president is Republican and 0 if the Democrat. This variable is interacted with inflation and approval variables. There is a tendency for the President's party to lose seats in a midterm election, so this is one of the key variables in our prediction model.

Approval. This is the average approval of the President in the year leading up to the election.

Inflation. This is a continuous variable of inflation in the United States in the year of the election. This is another relevant economic indicator that shows the overall picture of the economy and is felt by the average voter, and especially in the 2022 election, it is highly relevant

as it has been reaching peak highs. We also choose to interact this variable with the dummy variable for the Party president.

Proportion of urban or propurb. This is the proportion of a state's population that is urban based on 10-year period census data, so it's the same for each decade. This is a control variable based on the assumption that the more urban a state is the more Democrat-leaning they are.

The Final Model and Justifications

In these paragraphs, we will explain what the final model looks like, and then justify why we used each of the variables we used. Following the justifications, we will explain other models that were considered and why they were ultimately rejected. After carefully considering the relevant variables, and creating different models, we determined which control and interaction variables we wanted to use. We finally reached a final model that we feel gives the most accurate and robust predictions for the senate elections. The model is as follows:

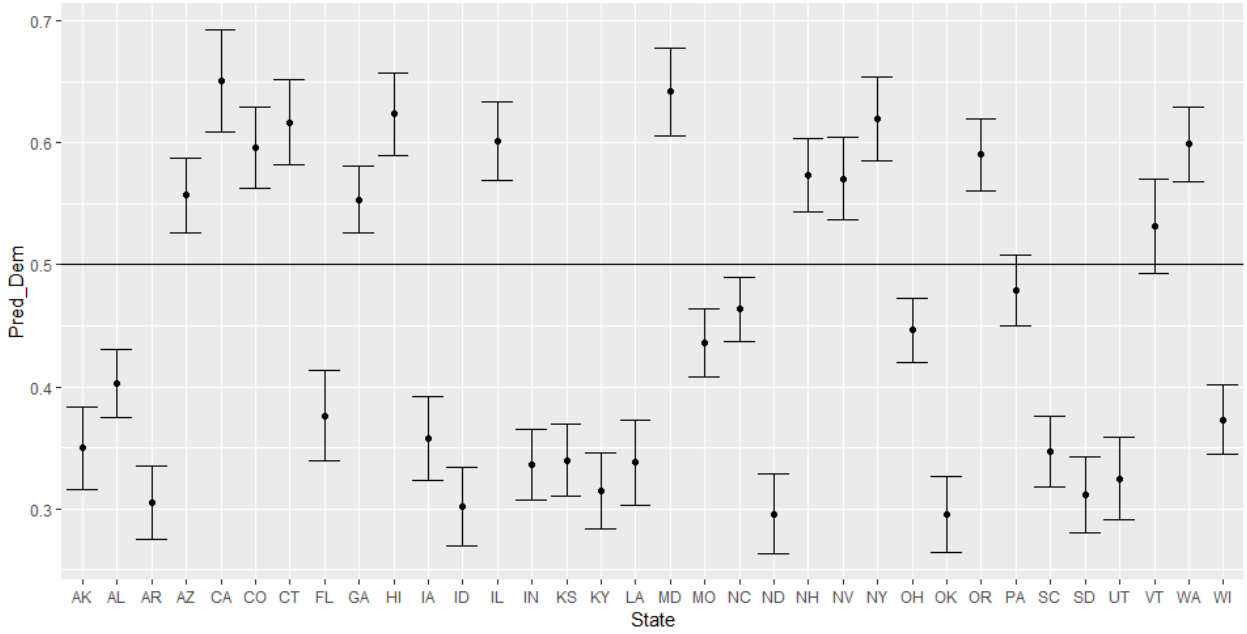
> Democratic vote share = $\boldsymbol{\beta}_0$ + $\boldsymbol{\beta}_1$ Democratic presidential vote share + $\boldsymbol{\beta}_2$ Urban proportion + $\boldsymbol{\beta}_3$ (Unemployment change x Incumbent Democrat) + $\boldsymbol{\beta}_4$ Incumbent Democrat + $\boldsymbol{\beta}_5$ (Inflation x President Republican) + $\boldsymbol{\beta}_6$ (President Republican x Approval) + $\boldsymbol{\beta}_7$ President Republican

We included presidential vote share to gauge what partisanship looked like in each state, as we believed this was highly predictive of Democratic senate vote share. We included the urban proportion of the population as a control variable because we know that urban areas tend to be skewed toward Democratic candidates. We created an interaction term of unemployment change and the indicator variable incumbent democrat. We thought these variables went hand in hand, where if there was a Democrat incumbent and unemployment increased, the person in power would be blamed for the change in unemployment and face the electoral consequences. Likewise, we predicted the opposite to be true if unemployment decreased, or if the democrat was not the incumbent. We also included the incumbent Democrat indicator variable on its own, knowing that having an incumbent seat has a strong correlation with winning an election. We used the interaction between inflation and the indicator variable of a Republican president. This interaction is complementary to the "unemployment change x incumbent democrat" interaction term explained previously. We included inflation because it has been a significant talking point going into the midterm elections, and we interacted with the presidential party to provide data that should be important for making this prediction, particularly given current relevance. We also interacted with approval with the presidential Republican term to see how much the popularity of a president impacts the vote share in senate races, and how much that impact changes depending on the party of the president. Finally, we included the indicator variable president Republican on its own. This is based on the idea that historically the president's party does not do well in midterm elections.

When creating the regression model, we experimented with a few other models. We strongly considered running the regression model with only recent data (post-2000 elections), thinking that information from 1976 would not be as predictive of modern politics. We tried running our regression analysis on data from only the past 20 years, however, this data was flawed in numerous ways, so we ultimately decided against it. We also considered adding a quadratic variable to our regression, but after examining our data we did not think this would improve our projections. Also, we figured theoretically, a quadratic variable wouldn't make sense in our model as we believe there to be mostly linear relationships between all the variables. Finally, we considered adding more robust economic data but ultimately decided unemployment and inflation statistics would be satisfactory. We did not want to crowd the model with too many variables that could make prediction and interpretation too complicated.

Discussing Results

Our model has Democrats winning a majority of the senate vote share in the following states: Arizona, California, Colorado, Connecticut, Georgia, Hawaii, Illinois, Maryland, New Hampshire, Nevada, New York, Oregon, and Washington. It shows Republicans winning: Alaska, Alabama, Arkansas, Florida, Iowa, Idaho, Indiana, Kansas, Kentucky, Louisiana, Montana, North Carolina, North Dakota, Oklahoma, South Carolina, South Dakota, Utah, and Wisconsin. The confidence intervals for the remaining states; Pennsylvania, and Vermont go through 0.5, meaning we cannot confidently make a prediction for either of these elections. Figure 1 shows the predicted vote shares and their confidence intervals.[1]



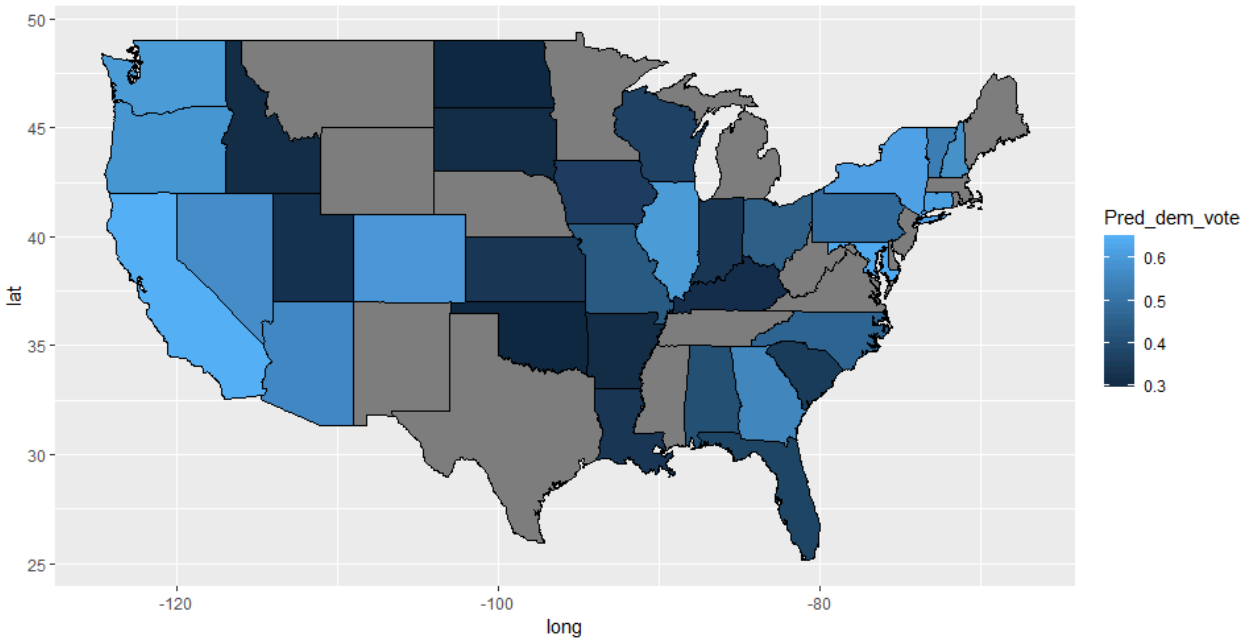
[1] In the confidence interval plot, you can notice that many of our predictions our quite strong, with only two confidence intervals crossing 50%. For a discussion of the implications of this and what might be causing this see "Response to Comments"

Our final regression ends up as displayed below. The best indicators of what the Democratic senate vote share from our analysis is the Democratic presidential vote share which has a coefficient of 0.444, whether the incumbent is a Democrat, which has a coefficient of 0.168, and whether the president is a Republican, which has a coefficient of 0.287. Each of these three variables is in line with our thinking; for the presidential vote share, we thought that national partisan politics would have a strong effect. This effect is also contained within the indicator variable that shows if the president is a Republican. This follows our logic that the party in control of the presidency tends to do poorly in midterm elections. Finally, the incumbent Democrat indicator variable shows that having an incumbent holding a senate seat has a strong impact on their share of the vote in the following election.

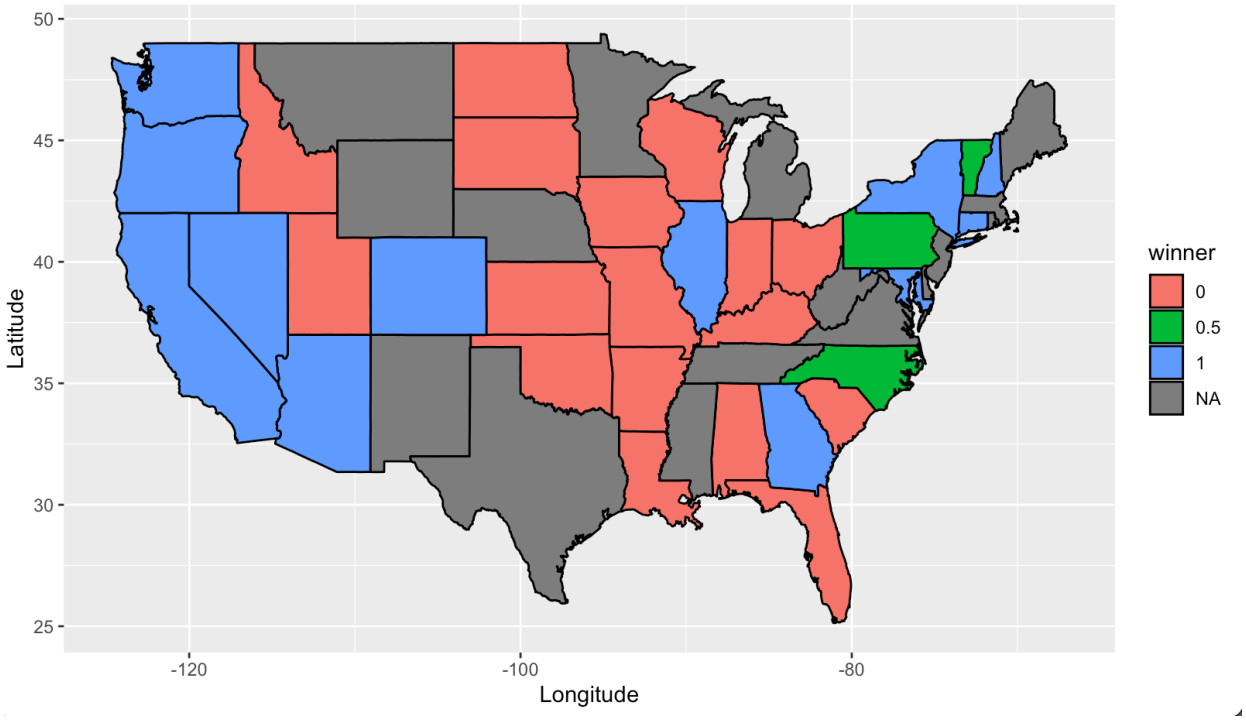
Coefficients				
	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-0.122	0.067	-1.776	0.0761
dem_pres_vote_sharre	0.444	0.056	7.968	6.17e-15***
propurb	0.017	0.032	0.537	0.591
unemp_change	-0.006	0.006	-0.951	0.342
inc_d	0.168	0.011	15.338	<2e-16***
inflation	0.286	0.002	4.789	2.03e-06***
pres_r	0.286	0.069	4.171	3.40e-05***
approval	0.005	0.001	4.054	5.58e-05***
unemp_change : inc_d	-0.010	0.008	-1.308	0.191
inflation : pres_r	0.011	0.005	2.313	0.021*
pres_r : approval	-0.005	0.001	-4.105	4.50e-05***

We made two different maps, one that shows the proportion of the Democratic senate vote share for every state on a gradient and one that shows whether or not Democrats won the

senate race in each state. The maps only display the mainland United States, however, from our data we predict Democrats to win Hawaii, and Republicans to win Alaska.

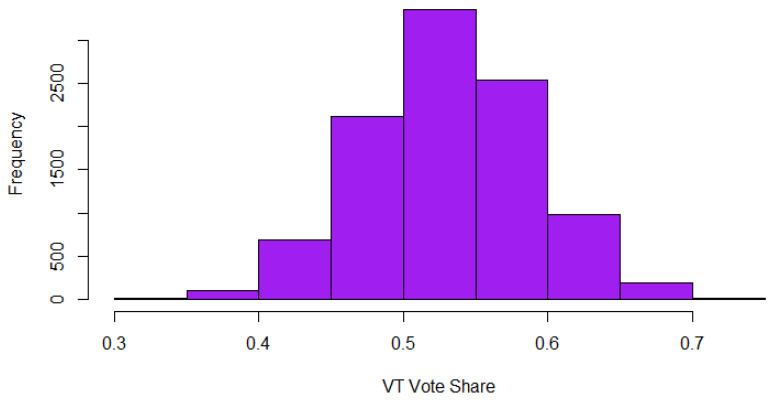


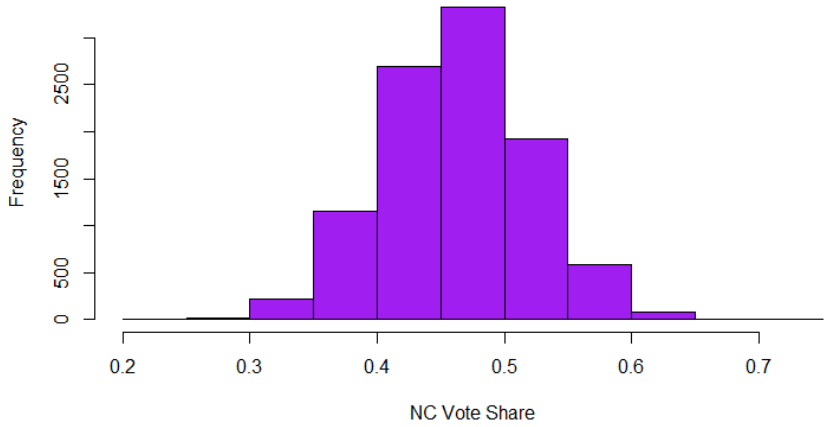
Map 1 shows the results on a map displaying the proportion of votes Democrats won on a gradient. This map shows states like Maryland and California in very light blues, indicating that we predict Democrats will win a large proportion of the vote in those states. In contrast, it shows states like North Dakota and Oklahoma in very dark blues, which indicates that we are confident Democrats will win a much smaller proportion of the votes in those states.

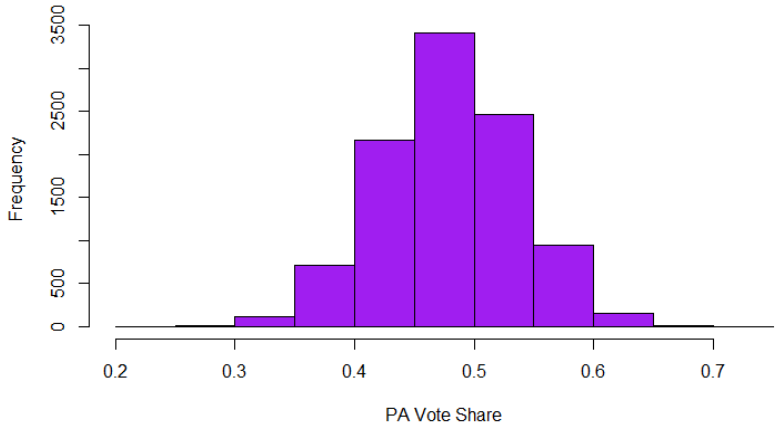


Map 2 displays the election results color-coded according to the party that we expect to win. Red is a Republican state, blue is a Democratic state, and green is a toss-up state. From this map we can see Democrats winning 12 states, Republicans winning 17 states, and 3 states being toss-ups. This second map is coded to include North Carolina as a toss-up state because the confidence interval for it was close to passing 0.5, so we did not feel confident enough to leave it as a Republican state.

We created three different sampling distributions, one for each of the three toss-up states. We did this to get a better idea of we could expect to see happen in those states, given that they were the closest races according to our model.

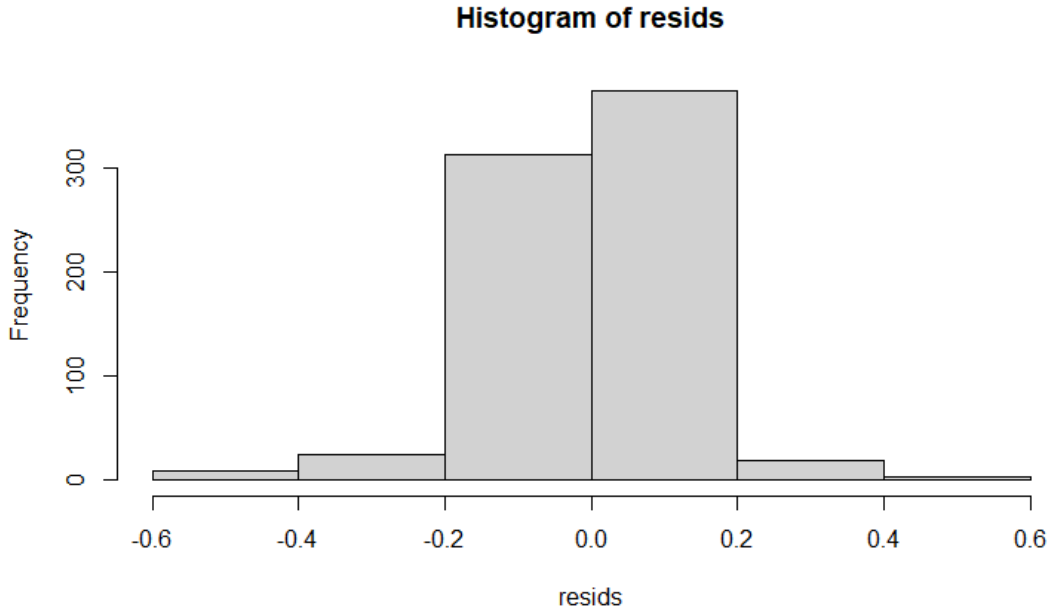

Frequency
0
500
1500
2500
0.3
0.4
0.5
0.6
0.7
VT Vote Share


Frequency
0
500
1500
2500
0.2
0.3
0.4
0.5
0.6
0.7
NC Vote Share

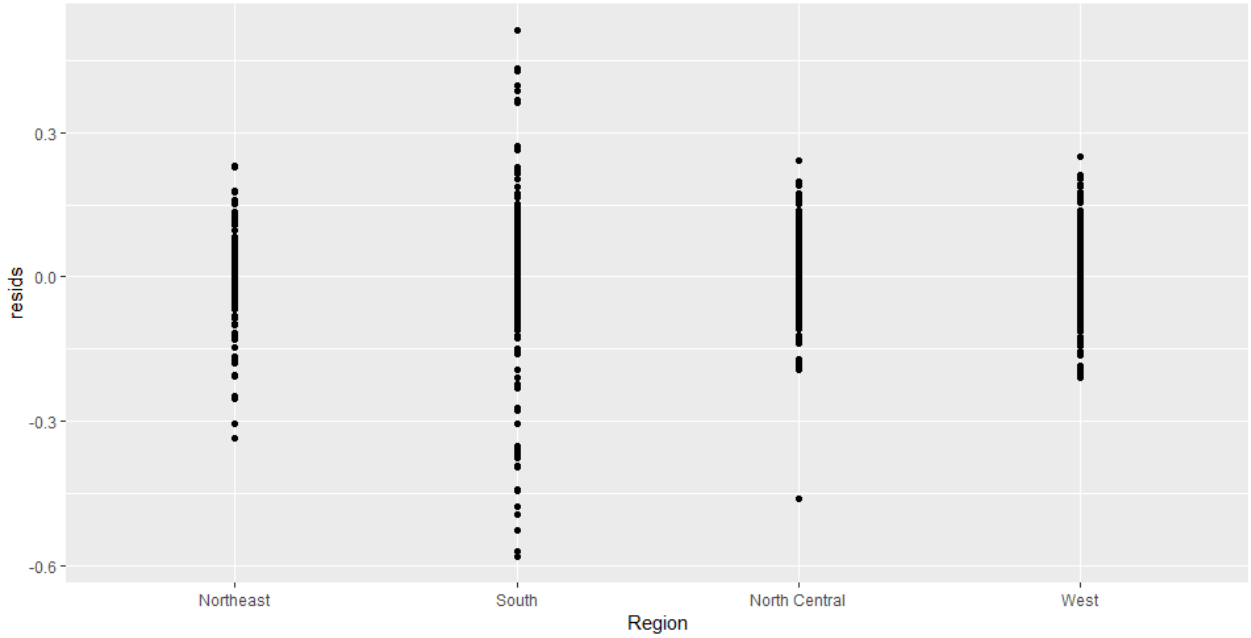

Frequency
0
500
1500
2500
3500
0.2
0.3
0.4
0.5
0.6
0.7
PA Vote Share

The distributions show that we can expect Democrats to win the race in Vermont and that we expect Republicans to win in Pennsylvania and North Carolina. A majority of the samples in the Vermont distribution show most of the samples having more than 50% of the Democratic vote share, while the North Carolina distribution shows Democrats winning less than 50% in most of the samples. The Pennsylvania sampling distribution shows a slightly more even distribution, however, Democrats still lose in a majority of the samples.

Finally, we built a few residual plots to characterize how confident we are in our model. The first of which plots a histogram of the residuals from every election in the dataset.



This model shows that for the majority of elections the residuals are close to 0. This means that our model closely predicts the vote share most of the time. Additionally, the histogram displays a fairly normal distribution, meaning there are not many obvious issues.



The next plot shows residuals and how they vary by region. The most noteworthy thing we find here is that there are clearly larger residuals in the south than in any other region. Most of the outliers are from this region. We decided to add a dummy variable for the region south in our regression, however, this new model did not differ significantly from our current one, so we made the decision to preserve our initial model instead.

Comparing Results with Reality

Our model predicted results that were fairly similar to what we actually observed on election day. In the table below, our predicted results are compared side-by-side with the actual results the elections yielded.

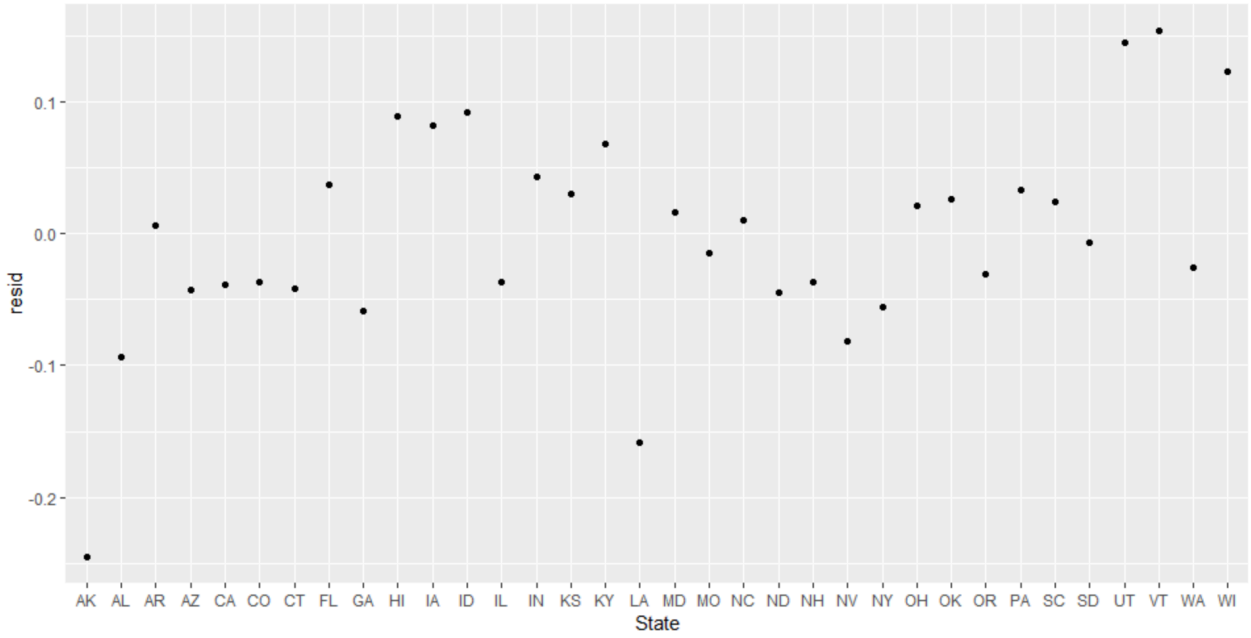
State	Predicted Outcome	Actual Outcome	State	Predicted Outcome	Actual Outcome
Alaska	Rep.	Rep.	Maryland	Dem.	Dem.
Alabama	Rep.	Rep.	Missouri	Rep.	Rep.
Arkansas	Rep.	Rep.	North Carolina	Toss-Up[2]	Rep.
Arizona	Dem.	Dem.	North Dakota	Rep.	Rep.

[2] Confidence intervals did not cross 50% but came very close to doing so, which is why we decided to code as a Toss Up. However, if we had not called a toss-up it would have been a predicted Republican win.

California	Dem.	Dem.	New Hampshire	Dem.	Dem.
Colorado	Dem.	Dem.	Nevada	Dem.	Dem.
Connecticut	Dem.	Dem.	New York	Dem.	Dem.
Florida	Rep.	Rep.	Ohio	Rep.	Rep.
Georgia	Dem.	Run-off	Oklahoma	Rep.	Rep.
Hawaii	Dem.	Dem.	Oregon	Dem.	Dem.
Iowa	Rep.	Rep.	Pennsylvania	Toss Up	Dem.
Idaho	Rep.	Rep.	South Carolina	Rep.	Rep.
Illinois	Dem.	Dem.	South Dakota	Rep.	Rep.
Indiana	Rep.	Rep.	Utah	Rep.	Rep.
Kansas	Rep.	Rep.	Vermont	Toss Up	Dem.
Kentucky	Rep.	Rep.	Washington	Dem.	Dem.
Louisiana	Rep.	Rep.	Wisconsin	Rep.	Rep.

Our predictions were very accurate for the most part. The only prediction we got "wrong" was Georgia. However, Georgia is a special case; since neither candidate got more than 50% of the vote, the top-two candidates, Raphael Warnock and Herschel Walker, are heading to a run-off election later this month. Thus, it is possible that our predicted result (Democratic win) can still happen, with the caveat that we did not predict the election to head to a run-off. For our three toss-up states, North Carolina went Red, while Vermont and Pennsylvania each went Blue. Overall, our model's basic predictions were correct and thus, this was a strong model for predicting the election.

While our basic predictions (which party would win) were very strong, our more precise predictions (the actual percentage of the vote Democrats would receive) were off in several cases. The residuals derived from subtracting our predicted results from the real results are plotted below.



As seen above, the majority of the residuals fall between -0.1 and +0.1. A few notable outliers can be explained. Alaska had no Democratic candidate, which is why we predicted a much higher Democrat vote than there was in reality. Louisiana had several Democratic candidates who split the Blue vote; we only included the percentage of the vote that the candidate who did the best received. In Utah, there was no Democratic candidate, so the percentage of the Democratic vote was coded as 1-percentage of vote the Republican received. Thus, it looks like Democrats did better than expected when there actually was no true Democratic candidate. Finally, there is no good excuse for Vermont and Wisconsin. The Democratic candidates just did much better than expected in these states.

Response to Comments

One concern noted by Professor Sahn was that the majority of our predictions were very confident, with only two states having confidence intervals crossing 50%. While we did not feel we needed to change anything to address this, as our model was still remarkably accurate, we did want to offer an explanation of why we believe our model was so confident. We included two variables that were very predictive, but also skewed results to go on one side or the other. The first is the percentage of the vote the Democratic candidate for president received in the previous election. This is highly predictive of the percentage of the vote share the Democratic candidate for senate will receive in the election, but because it is so predictive it tends to skew results to go on one side or the other. The second relevant variable is the incumbent Democrat. This has the same effect, it is highly predictive but because of this, it tends to skew results to be highly confident one way or the other. If the incumbent running for re-election is a Democrat, our model will almost always predict a confident victory for the Democrat. The effect of this variable is the

smallest when there is an open seat, otherwise, it will often predict the incumbent to win confidently. Since this reflects reality, we do not feel it needs to be addressed. Ultimately, these two variables mean that our predictions are often confident, but it also means that our predictions are often right.

Other comments we received from Professor Sahn related to our variable coding and interaction choices. We coded our incumbent variable as Democrat is 1, Republican is 0, and 0.5 if there is no incumbent. Doing it like this means there is linearity between the 3 categories, so going from Democrat incumbent to open seat is the same "effect" as going from open seat to Republican incumbent. Essentially, if we coded no incumbent as 0 instead of 0.5, it would effectively be the same as having a Republican incumbent (since Republican incumbents are coded as 0). This is a problem that we thought about as it could lead to biased results, but since there was no other viable alternative that was known to us, we decided this was the best way to account for incumbency without having to worry about NA values. When we got our accurate predictions, we decided that any problems that might have arisen from this coding choice probably were not significant enough to derail our model, and thus, we decided to stick with it. We were also questioned about why we chose to interact unemployment change with the incumbent, but not unemployment means. Our other interactions variables were inflation with republican president and approval with republican president. So then our model already had three interactions, and two of the pairs were economic indicators. Since unemployment change was something that is capturing a change in time, a change that is most likely seen and felt by constituents on a daily, we decided that it was the more superior variable of the two to interact. We did not want to add more interactions and make our model too complicated to interpret, so we decided these three interactions were the best, and unemployment change was a better variable to interact with incumbent than unemployment means. We did then tried a model without the unemployment mean altogether and it had no effect on our predictions and such, so we decided to leave it out. Other comments that we received from Professor Sahn were about our formatting and editing which we fixed. This included making a more clear chart of our coefficients rather than a screenshot of the coefficient screen straight from R, and making it clear that our simulations were of the states that we deemed were toss ups and not a random selection of states.

Comments that we received from Group 1 were more on the side of edits on grammar, elaboration, and formatting. They suggested revisions on clarity, conciseness and diction to make understanding easier for people who are not as versed in statistical language. This was enlightening as we can become blind to using jargon to explain. The people in our class are probably familiar with and used these terms, but many others are not, and to make our report more accessible, it's important that we write in terms that are comprehensible to everyone.

They did have one issue with our variable choices that we did not agree with. They point out we addressed numerous issues that we thought were hot topics for this election such as the Ukraine war, abortion, gun control, the Pandemic, and rampant inflation. We controlled for inflation, but while the other issues are important, they are not super measurable as they are more

ideological. Also, these issues did not exist in the elections before 2020 at the latest, we cannot build a regression model for viewpoints on the Ukraine war as there was no war before this year.. So, we do not agree that we should have looked into other hot topic issues to discuss or test in our model, or even justify why we didn't look into it.

Conclusion

Our group was tasked with predicting the results of the midterm election for the open 34 US senate seats and comparing these results to the true outcomes. We developed a model, regressing Democratic vote share in the senate against the year and state of the election, whether there is an incumbent and if there is, the party of the incumbent, the percentage of unemployment within the state, the percentage of inflation across the country, proportion of the urban or rural population in a state and the party of the President. Each of these variables was carefully selected, justified by their relevance to politics. We determined that Democrats won in 12 states and Republicans won in 18 states. We also made confidence intervals to account for uncertainty, finding that there were 3 "toss-up" elections. For those remaining states, we simulated trials to find which party was most likely to win in said states. Finally, for our comparison, we created residuals with the true outcomes to evaluate our accuracy. Most of our residuals were between -0.2 and 0.2, supporting our discernment that our model was mostly accurate. We visualize our research through the diagrams pictured in Figures 1 through 9.

However, we must ask: what could we have done better? We immediately returned to our variables to find what we should reconsider. We knew that the regression model must not have been the issue, given the theoretical reasons we previously explained that led to its selection. As previously stated, there are a multitude of factors that influence an election. Demographic information of both constituents and candidates is highly significant when it comes to determining whether a constituent is to vote Democrat or Republican and if a candidate of either party is to win. Characteristics such as race, sex, income, and education often sway voters towards certain politicians that they relate to; incorporating this data into our model could have potentially been beneficial. Perhaps, if such state-level census data were available, the data would be equipped in the future to improve upon our work. These could be coded as indicator (sex), discrete (race and education), and continuous (income) variables. Moreover, we initially intended to use campaign funding as a variable. A candidate's spending could be an indicator of how much effort was put into outreach and voter awareness; the costs of making themselves recognizable through video and print ads, door-to-door canvassing, and in-person and virtual events add up. Although this would have been a great variable to include in our regression, we found that politicians are often elusive and we cannot be certain of how truthful public statistics are. And it would have been even more difficult to find reliable campaign funding data from the 1970s and 80s, which is how far the rest of our data goes. If we could have found credible and trustworthy data, it could be coded as an indicator variable, based on if it goes above a specific amount. Lastly, following the completion of our study, we discussed the increasing weight of presidential endorsements. For this midterm election, former president Donald Trump endorsed

many candidates running for a variety of government positions. His approval was supposed to garner support from the right-wing masses to ensure their wins, but many political pundits after the election surmised that the endorsement could have actually been more detrimental than helpful. This would have been a very interesting variable to explore and it could be coded as an indicator variable, based on if or not a candidate was endorsed.

Ultimately, our findings were nearly fully accurate to the real election outcomes, indicating that our regression model is effective. With only a difference in the victory of a single Republican state, we believe that our model can be used in further research regarding election results. All elections have a degree of unpredictability that no model can control. That being said, our model has proven to be highly successful in working within these limits.

Works Cited

Ansolabehere, Stephen and James M. Snyder Jr. 2002. “The Incumbency Advantage in U.S. Elections: An Analysis of State and Federal Offices, 1942-2000.” Election Law Journal 1, 315-339.

Ha, Jongrim, M. Ayhan Kose, and Franziska Ohnsorge (2021). "One-Stop Source: A Global Database of Inflation." Policy Research Working Paper 9737. World Bank, Washington DC.

Palmer, H. D., & Whitten, G. D. (1999). The Electoral Impact of Unexpected Inflation and Economic Growth. British Journal of Political Science, 29(4), 623–639. http://www.jstor.org/stable/194241

Wright, J. (2012). Unemployment and the Democratic Electoral Advantage. American Political Science Review, 106(4), 685-702. doi:10.1017/S0003055412000330

University, S. (2019, June 3). How the urban-rural divide shapes elections. Stanford News. https://news.stanford.edu/2019/06/03/urban-rural-divide-shapes-elections/