Final Prediction

Annelies Quinton

2022-11-07

Intro

This blog includes my final prediction for the 2022 Midterm elections. Over the past 10 weeks, I have looked at various variables, assessing their predictive power, with the goal of creating a prediction model. The outcome of my model is a national vote share prediction.

The Last 10 Weeks

Before I share my model, I think it is important to summarize what variables I have considered before arriving at my final model. I started by looking at economic fundamentals, assessing GDP, unemployment, and inflation. I found that subsetting the data to the last quarter of the election cycle was the most predictive(Healy and Lenz, 2014).. I then looked at polling data, specifically the average generic ballot within the 50 days prior to the election. I then shifted to looking at district level variables. This included incumbency and expert ratings. I also considered the "air war" and advertising trends (cost, tone, purpose) grouped by district. Finally, I considered demographics and turnout at both a district and national level. Each of these variables include different variations, resulting in another group of variables to consider. For example, with incumbent data, I looked at the incumbent House party, the incumbent president party, and when these parties match. The challenge became choosing the variables.

Building My Final Model

When deciding on my final model, I built three variations of a similar model. I will walk through my thought process for choosing which variation I selected.

The final variables I considered were:

1. **Average democrat polling (50 days):** This the average of the democratic generic ballot polls from 50 days prior to the election. The generic ballot is an effective way of understanding the public's view and filtering for the days leading up to the election gets a better sense of the true pulse of the country (Bafumi, Erikson, Wlezien, 2018).
2. **President party:** This is a binary variable with 1 being Democrat and -1 being Republican. Often the party in power is punished during the midterms.
3. **Presidential and party match:** A binary variable with 1 being the parties match and -1 being the parties don't match. This is a continuation of the previous variable, but my hypothesis is that if the party matches there will be a larger punishment.

4. **Change in percent of white voters:** This the white vote share from that election minus the white vote share from the previous election. As seen in lab, white and Hispanic voters had the strongest coefficient toward prediction democratic vote share.

5. **Change in Hispanic voters:** This the Hispanic vote share from that election minus the Hispanic vote share from the previous election.

Below I show the regression outputs for the three models I considered.

```
##
## Regression Results
## ===========================================================================================
##                                                                     Dependent variable:
##                                                     -------------------------------------------
##                                                             Democratic Major Vote Share Percent
##                                                            (1)                  (2)                 (3)
## -------------------------------------------------------------------------------------------
## Change in percent of white voters                      -1.0**               -2.2**
##                                                         (0.3)                (0.7)
##
## Change in percent of Hispanic voters                    1.1***               1.7**
##                                                         (0.2)                (0.6)
##
## Presidential Party                                     -3.1***              -2.4**              -2.2**
##                                                         (0.3)                (0.8)               (0.5)
##
## President and House Party Match                        -2.7***
##                                                         (0.4)
##
## Average Democrat Polling Support (50 days out)          -0.2                  0.1                 0.6**
##                                                         (0.1)                (0.3)               (0.1)
##
## Constant                                               57.0***               42.7*              23.0**
##                                                         (5.5)               (16.7)               (7.1)
##
## -------------------------------------------------------------------------------------------
## Observations                                             10                   10                  14
## R2                                                      1.0                  0.9                 0.8
## Adjusted R2                                             1.0                  0.9                 0.8
## Residual Std. Error                                 0.4 (df = 4)         1.3 (df = 5)        1.9 (df =
## F Statistic                                     187.4*** (df = 5; 4) 20.3*** (df = 4; 5) 21.4*** (df
## ===========================================================================================
## Note:                                                                  *p<0.1; **p<0.05; *
```

It is interesting to see the significance of polling decrease as more variables are added. Polling is the only variable that is not significant at the 95% interval for models 1 and 2. Additionally, models 1 and 2 have the highest adjusted R2 values. However, the near perfect fit with model 1 makes me suspicious about its validity. Also, it is important to note that both model 1 and model 2 have 5 fewer observations than model 3 because they consider demographics, and the data for demographics begins in 1982. Additionally, model 3 is similar to the model built by Bafumi, Erikson, Wlezien in their 2018 forecast. However, they manipulate the polling variable more than my model does (Bafumi, Erikson, Wlezien, 2018).

Model Validation

RMSE

To evaluate the models, I first calculated the root mean square error (rmse) value.

Interestingly, both model 1 and model 2 have the same rmse value of 0.8164966. These are both greater than the rmse value for model 3, 0.9660918.

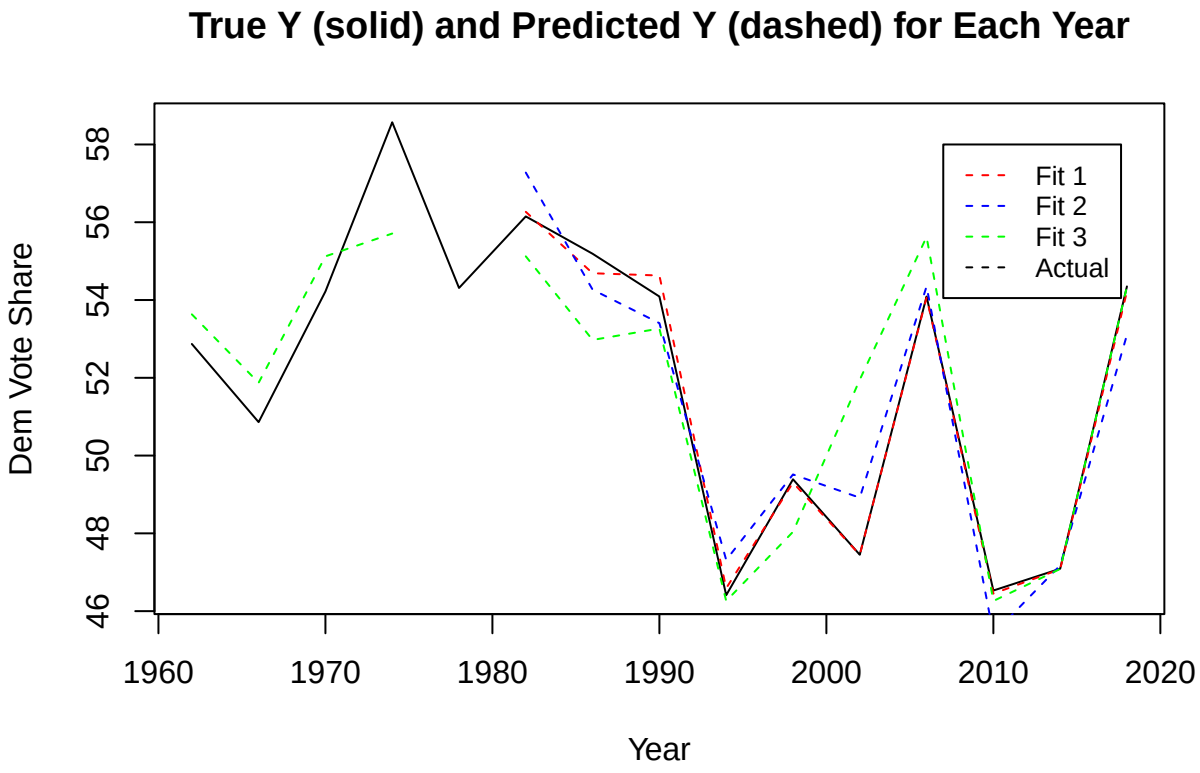
```
## [1] 0.8164966
```

```
## [1] 0.8164966
```

```
## [1] 0.9660918
```

Plotting Residuals

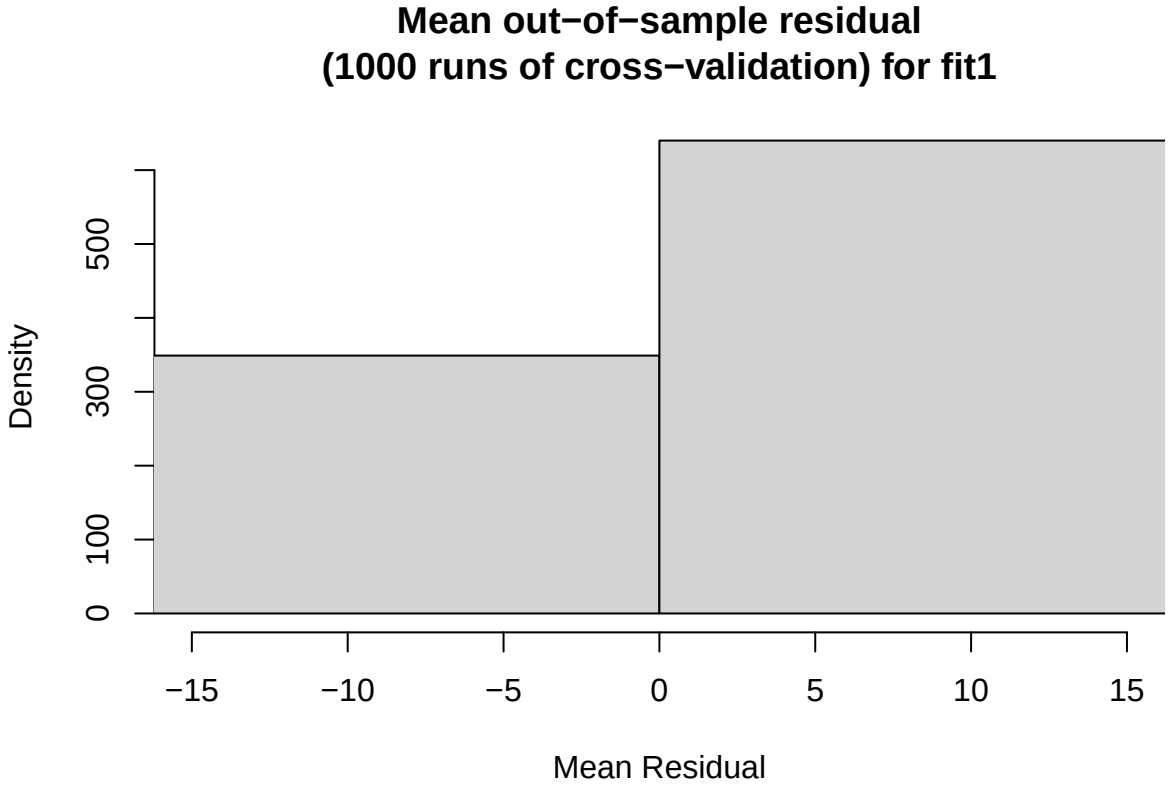
I then plotted the residuals for each model between the predicted and the actual values. The plots for model 1 and model 2 begin after 1982 because of limited demographic data. The plot shows that fit 1 appears to most directly correspond to the actual values.



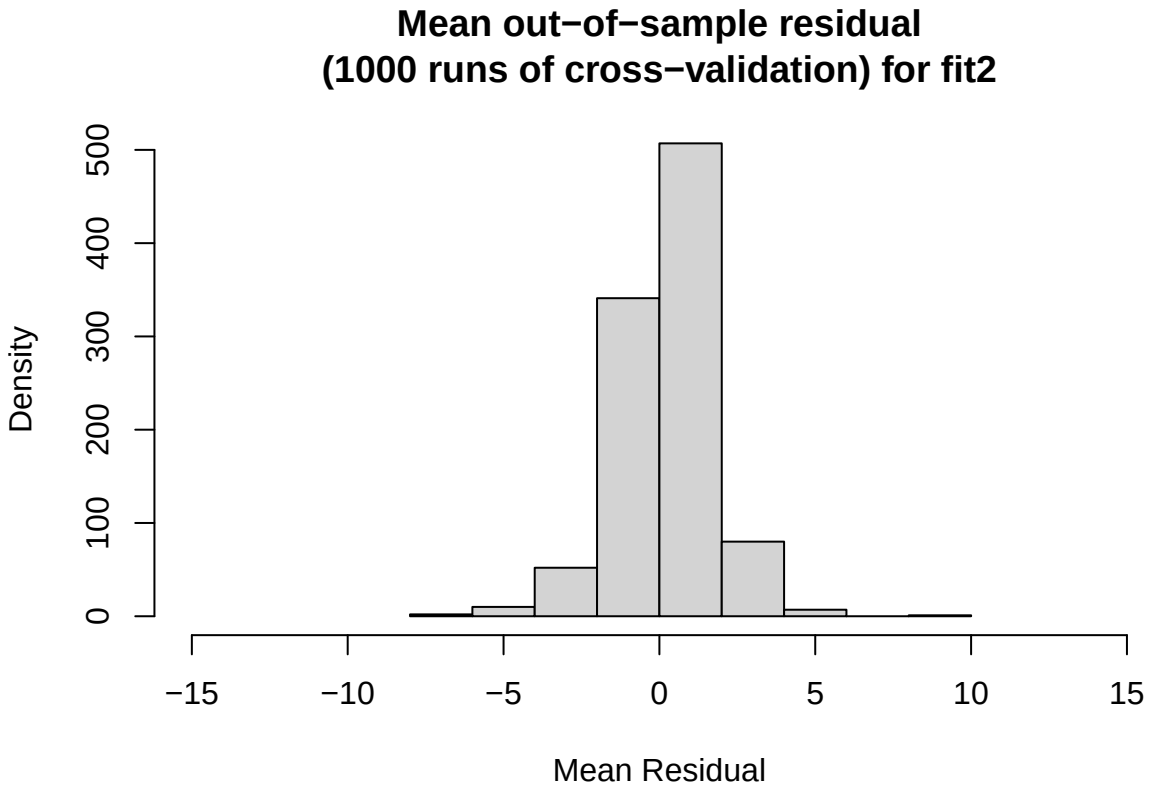
Cross Validation

I then do out-of-sample modeling through cross validation. I randomly 80% to be the train data and the other 20% as the test data. I found the fit 2 had the lowest average mean out-of-sample residual. The

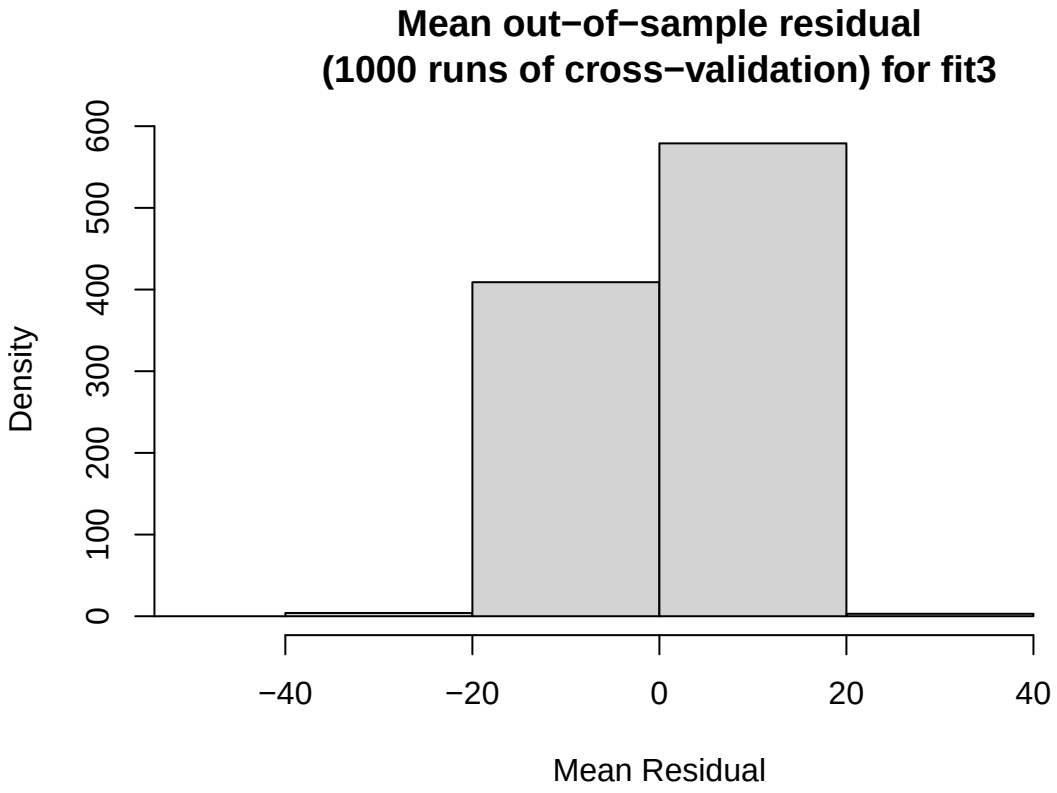
histogram for fit 2 also appears the most uniform between the three. Model 1 performed the worst among the three with, with an average mean out-of-sample residual value of 4.896344.




```
## [1] 3.521279
```


Mean out−of−sample residual
(1000 runs of cross−validation) for fit2
Density
0
100
200
300
400
500
−15
−10
−5
0
5
10
15
Mean Residual

```
## [1] 1.217813
```




```
## [1] 3.252421
```

Prediction Interval

Finally, I consider the prediction interval range at 95% confidence for the 2022 prediction for each of the models. Model 1 had the smallest interval, whereas model 3 had the largest. This could be explained by the fact that the largest coefficient for model 3 is a binary variable, and so there is more variability in the outcome.

Choosing a model

Each model performed well in separate validation testing. To choose a model, I used the chart below to average their performance. From this, model 1 has the highest score.

```
## New names:
## Rows: 5 Columns: 4
## -- Column specification
## ------------------------------------------------------------------------ Delimiter: "," chr
## (1): ...1 dbl (3): Model 1, Model 2, Model 3
## i Use `spec()` to retrieve the full column specification for this data. i Specify the
## column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
## # A tibble: 5 x 4
##   Test                ‘Model 1‘ ‘Model 2‘ ‘Model 3‘
##   <chr>                   <dbl>     <dbl>     <dbl>
## 1 Adjusted R2              1         2         3
## 2 RMSE                     1         1         3
## 3 Out of Sample Error      2         1         3
## 4 Prediction Interval      1         3         2
## 5 Average:                 1.25      1.75      2.75
```

My official prediction will be off of model 1, however, I am interested in evaluating both model 1 and model 3's success after the election. I am interested in model 3 because it has fewer variables and more heavily relies on polls. By comparing both models, I can see to what extent polls matter in predicting.

The equations for my models are:

Model 1:

Democratic Major Vote Share = 57 -1(change in white voters) +1.1(change in Hispanic voters) -3.1(presidential party) -2.7(president house match) -0.2(polling)

Model 3:

Democratic Major Vote Share = 23 - 2.2(president party) + .6(polls)

Discussion of Coefficients

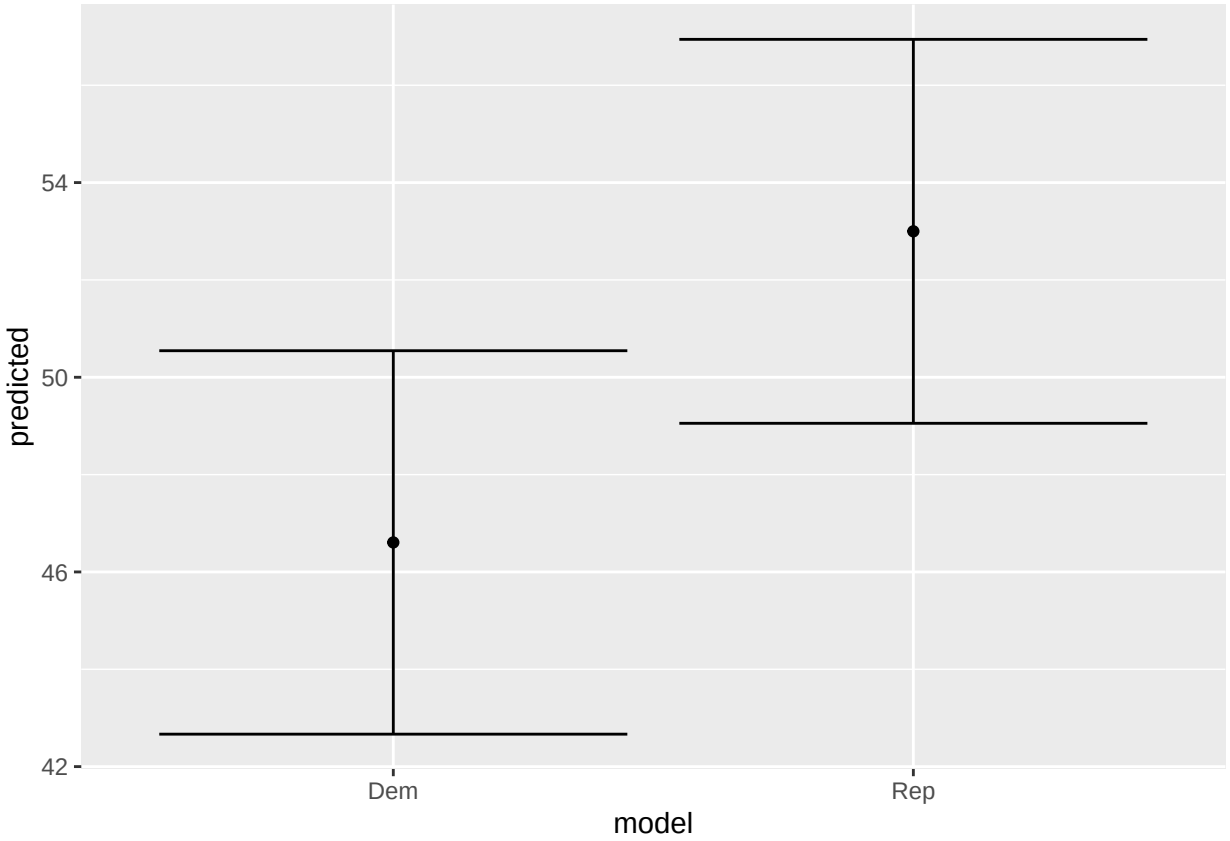
From these models, we see that the presidential party has the greatest magnitude in both models. Conversely, polling has the lowest magnitude. Polling as the lowest is a surprise to me because I would expect this variable to be the closest to the actual outcome. However, polling is often deemed fairly unreliable in predictions(Gelman and King, 1993). With regard to model 1, it is interesting that the presidential house match has a smaller coefficient than the presidential party. I would have thought this match would make people want to punish the incumbent party even more (as seen with the presidential party).

Prediction

Below I show the predicted democratic majority vote share and republican majority vote share and their confidence intervals.

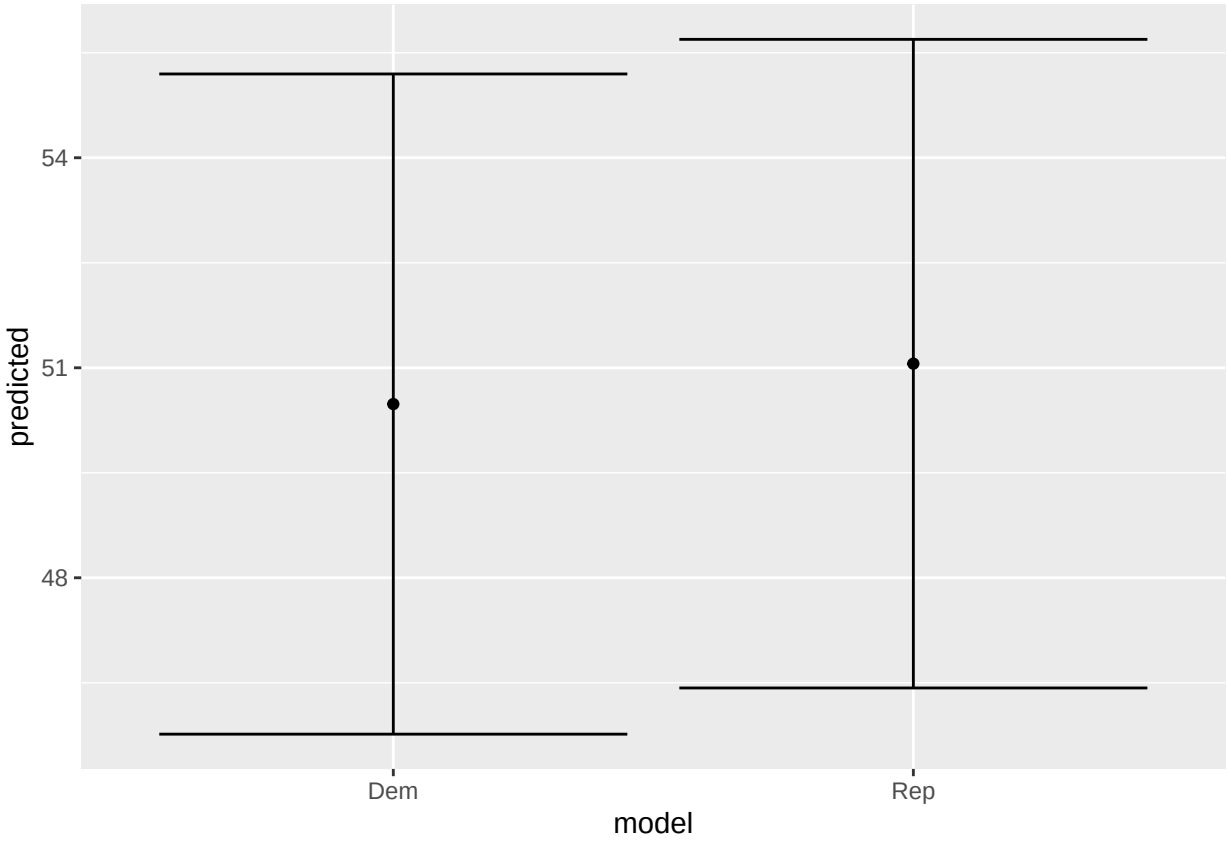
Model 1

The graph shows a republican majority, with fairly small intervals.



Model 3

The graph shows a slight democratic majority, but has very large confidence intervals.



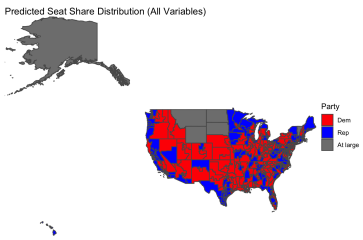
Final Prediction:

My final prediction will be off of model 1. I predict that democrats will have 46.6% percent of the popular national vote and republicans will have 54.4% of the national vote.

Extra

Why not seats?

Over these past weeks, I have attempted to create a model that predicts both vote share and seat share. However, due to a lack of consistent data across all districts, I found it difficult to create a model. The closest I got to creating a prediction for seat share was in my week 6 blog I have included the map of seat distribution below. This model iterates through each district considering average polling, incumbency, and expert ratings. I go into more detail about the specifics in the blog here. This model predicts a seat share distribution of 214 democrat and 221 republican. Although, I do not claim this model has much validity, I am including this prediction to have a value to compare the actual results to. The main focus of this blog and my work has been for national vote share, not seat share.



References

Bafumi, Joseph, Robert S. Erikson, and Christopher Wlezien. "Forecasting the 2018 Midterm Election Using National Polls and District Information." PS: Political Science & Politics 51, no. S1 (2018): 7–11. doi:10.1017/S1049096518001579.

Gelman, & King, G. (1993). Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable? British Journal of Political Science, 23(4), 409–451. https://doi.org/10.1017/S0007123400006682

Geoffrey. "Why the President's Party Almost Always Has a Bad Midterm." FiveThirtyEight. FiveThirtyEight, January 3, 2022. https://fivethirtyeight.com/features/why-the-presidents-party-almost-always-has-a-bad-midterm/.

Healy, & Lenz, G. S. (2014). Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy. American Journal of Political Science, 58(1), 31–47. https://doi.org/10.1111/ajps.12053