The Predictive Power and Accuracy of State-Level Presidential Polls

Brittany Alexander
Twitter: @balexanderstats
[*Github with Code and Data*](#)

Outline

- Literature Review & Data Source
- Definition of Accuracy and Predictiveness
- The State of State-Level Polls
- 2016 Mythbusting
- Conclusion

Data Source

Huffington Post Pollster has polling data for 2012-2016. There was a site to get 2008 data before Pollster merged with Huffington Post, but that link is broken. There are 5756 state level polls across the three elections. Most states have multiple polls for each election year. I combined all the data into a single file and added new variables to model polling across years. New variables includes:
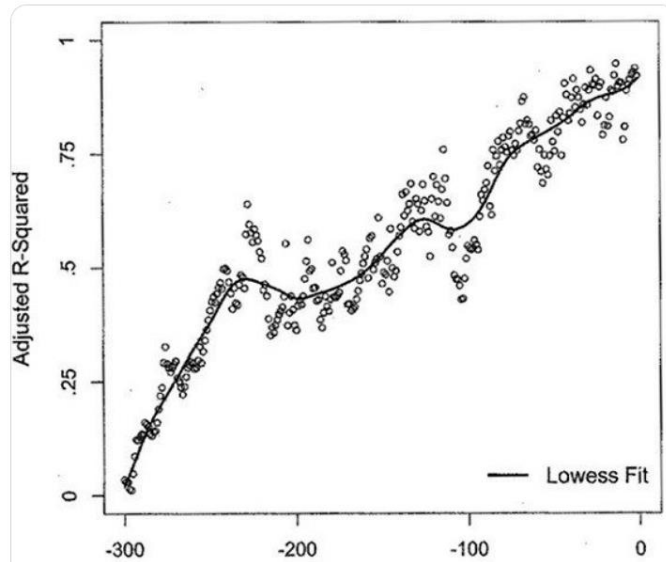
- The election results as both the margin and the two party vote
- Days until the election at the start, end, and middle of a poll
- Various versions of polling error

Inspiration

- [Twitter Thread](#)
- The goal of this paper is to analyze the accuracy and predictiveness of state-level presidential polling across the entire cycle

- Hillygus, D. S. (2011) describes a history of election polling but doesn't include much on state level polls

- Bon et.al (2019) focused on the effects of undecided voters and

- polling bias

- Shirani-Mehr (2018) built a model to decompose bias and variance in polls but focused on the last two weeks of the election

- Alexander (2019) built a model and looks at the accuracy of

- averaging all the polls

- None of these studies focuses on individual polls during a broad range of time

Accuracy in polling has two components: percent called correctly, and distance between a polls results and what happens on election day. Additionally accuracy can be viewed in terms of margin, and in terms of vote. Margin is defined as the difference between the Democratic vote (or poll support) and the Republican vote (or poll support). Accuracy in terms of vote is measured by first applying the formula: $d_{new} = \frac{d}{d+r}, r_{new} = \frac{r}{d+r}$ to polls and vote results so that the Republican and Democratic support sums to 1. This standardizes results to deal with different levels of undecideds, and the inclusion of minor candidates.

Predictiveness is defined by the strength of a correlation between two variables and R^2 of fit regressions. Predictiveness matters because it tells us if and when we can consider polls to have predictive value in the election.

Accuracy over Time

- This plot is a lowess smooth fit with 95% confidence bands on the absolute error of the margin of the poll versus the final election day margin. Date is the midpoint of the poll

- A state is competitive if it's margin on election day is between -5 and 5 points and otherwise is noncompetitive.

- 200 days is about late January and 100 days before is the end of July

- Before 100 days until the election polls could have a higher error than the margin suggesting they could not accurately predict the winner
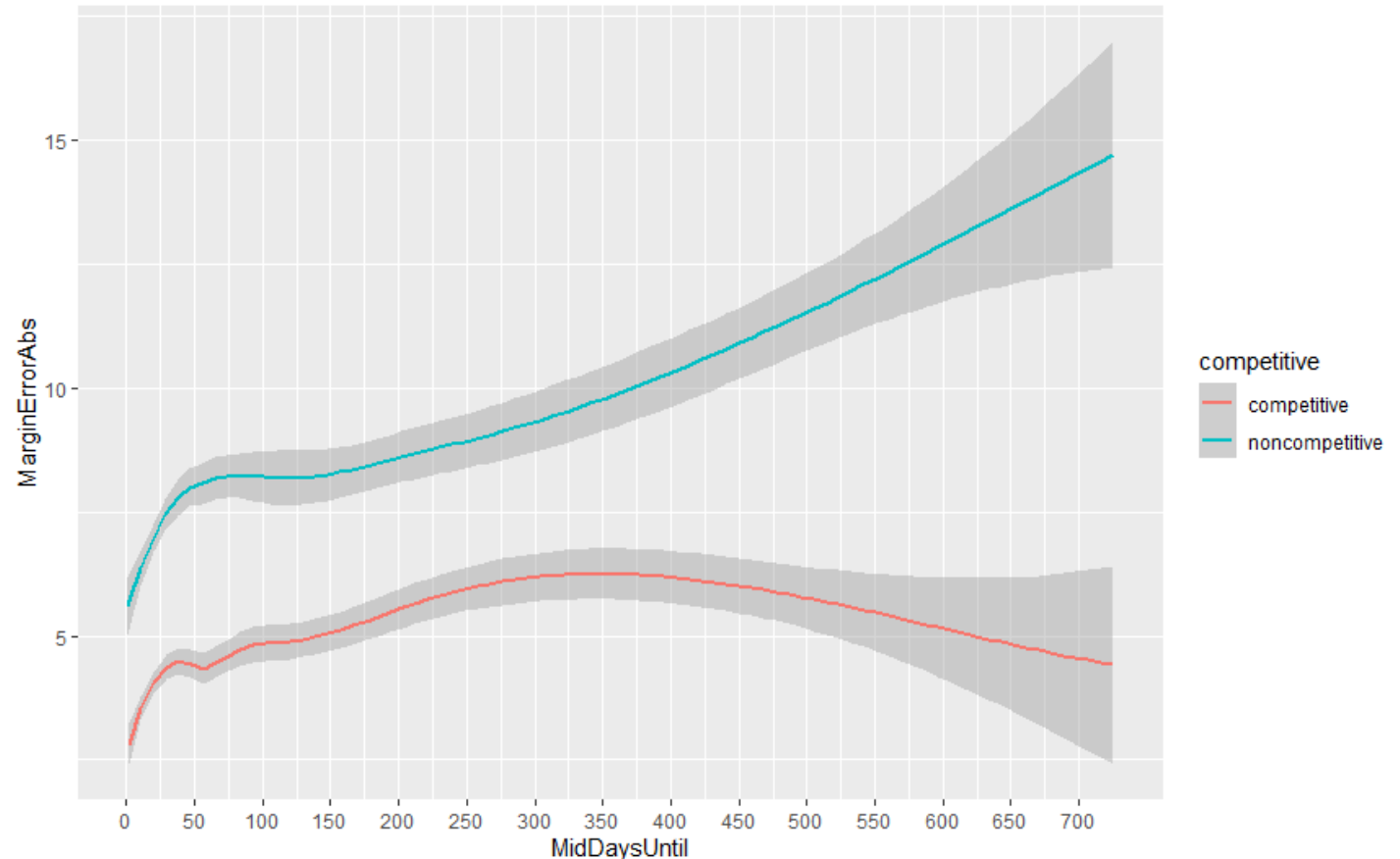
Average Error by State in Last 60 days


Average Margin Polling Error
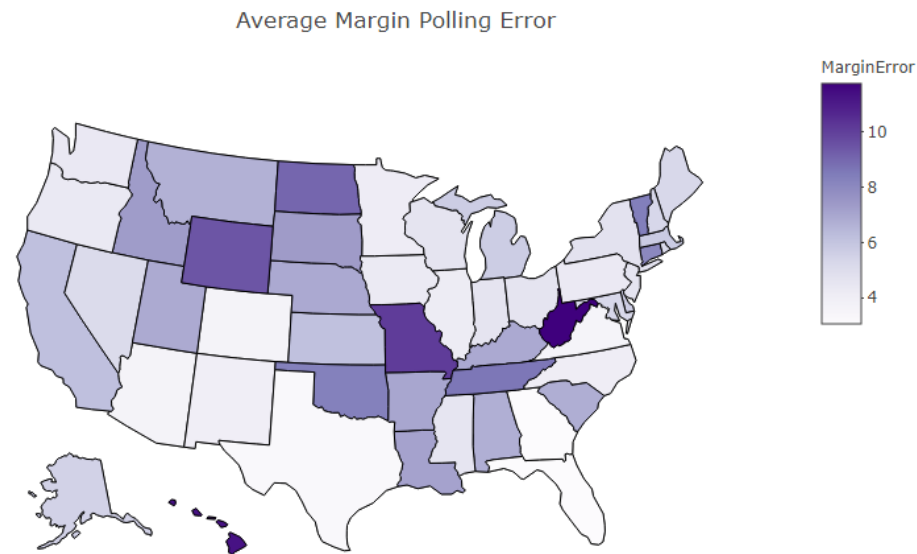
- We see that states that are more partisan have higher errors
- Higher population states have more accurate polls
- In most cases the average error is less than the average margin

Partisanship Paradox

- Right is a plot of the Average Margin Error of polls from the last 60 days for a state by it's actual margin on election day.

- Recall that proportions are most variable when they are closest to .5 suggesting that competitive states should have more sampling error.

- Possible explanation: Nonsampling factors such as poll quality, frequency polled, etc. explain this paradox.
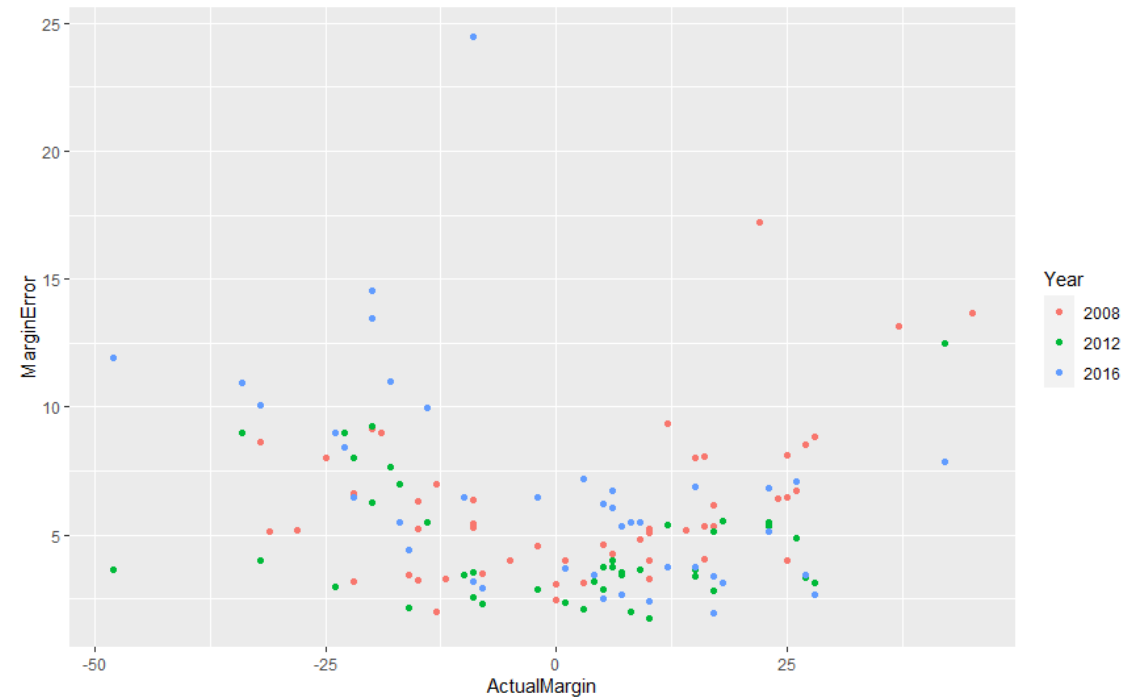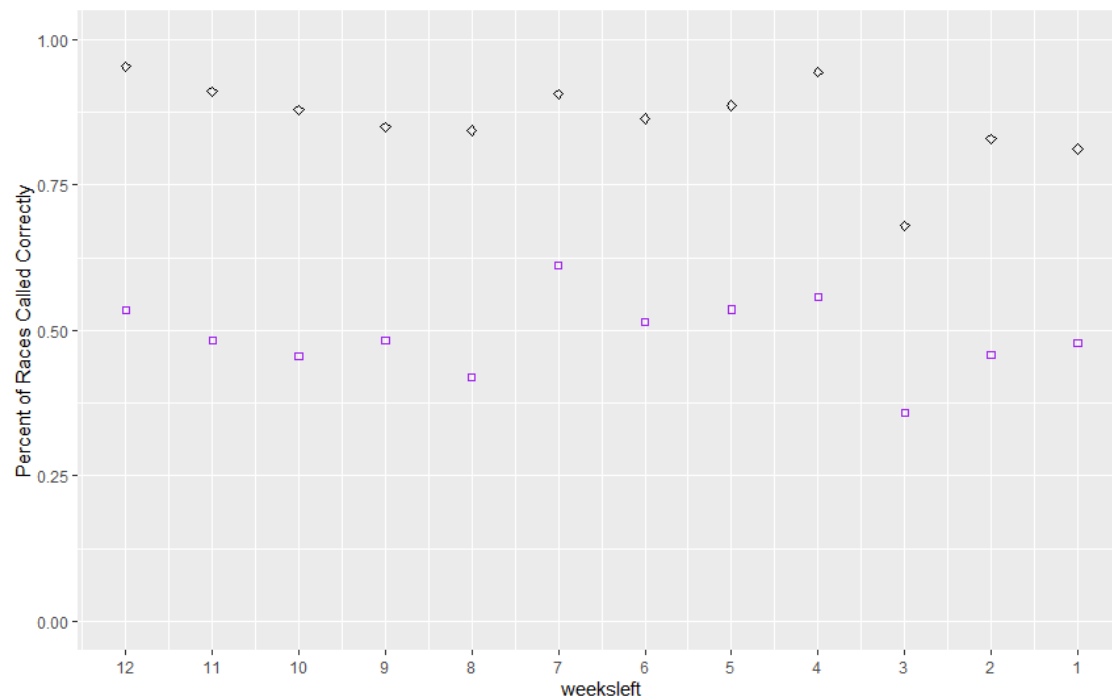
# Margin Error in Competitive and Non-Competitive States
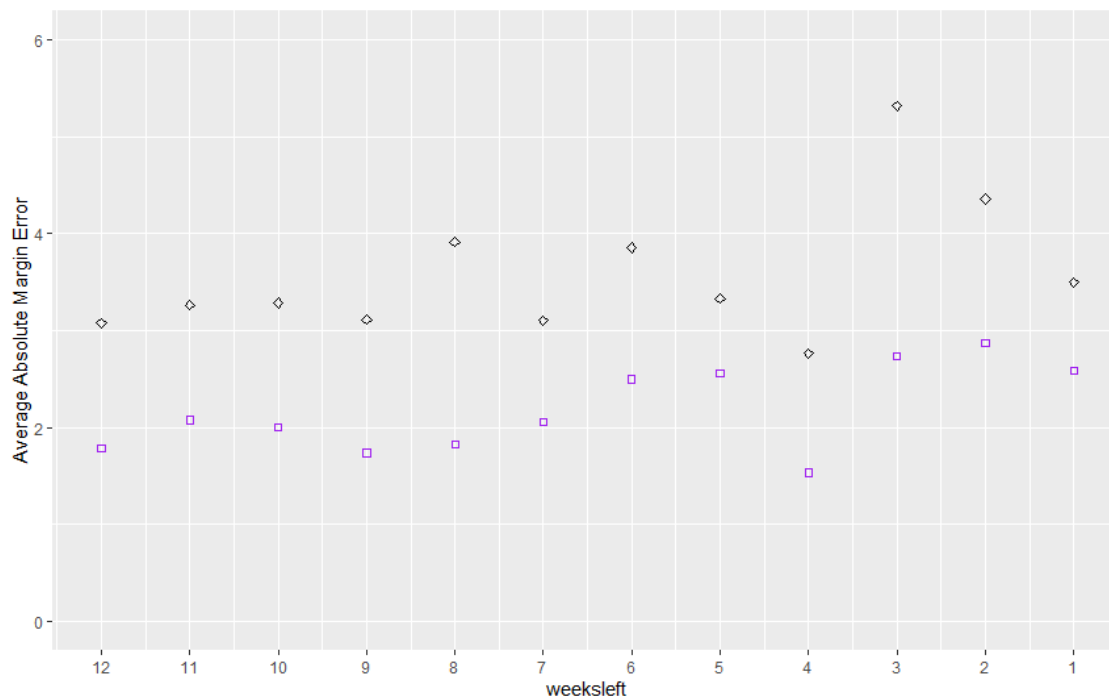
- This is a plot showing the average Margin error in competitive and noncompetitive states, broken up by week for the last 12 weeks of the election. Purple Square is competitive states, and Black Rhomus is noncompetitive states.

Percent of Races Called Correctly

- This is a plot of the percent of races called correctly. Purple Square is competitive states, and Black Rhomus is noncompetitive states.

R-square Estimates
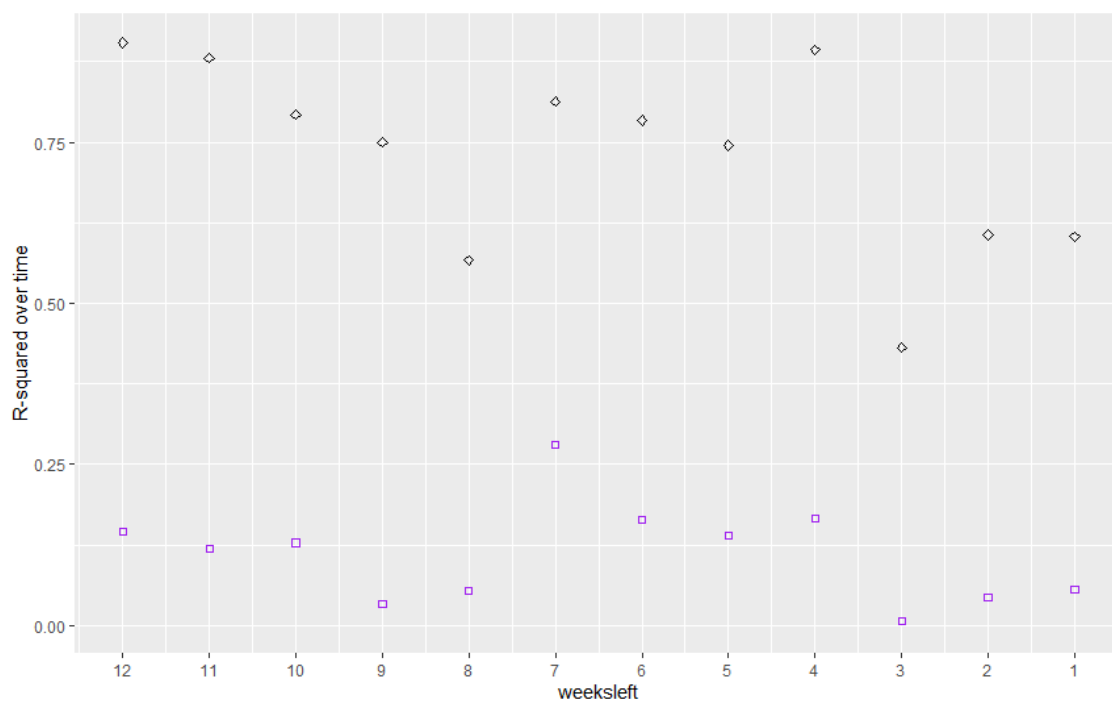
- This is an estimated R-square of a OLS model just predicting Democratic two-party vote share with the poll two-party vote share
- Purple Square is competitive states, and Black Rhomus is noncompetitive states.
- These values are low in competitive states but this is complicated by state and year level random effects

2016 wasn't abnormal
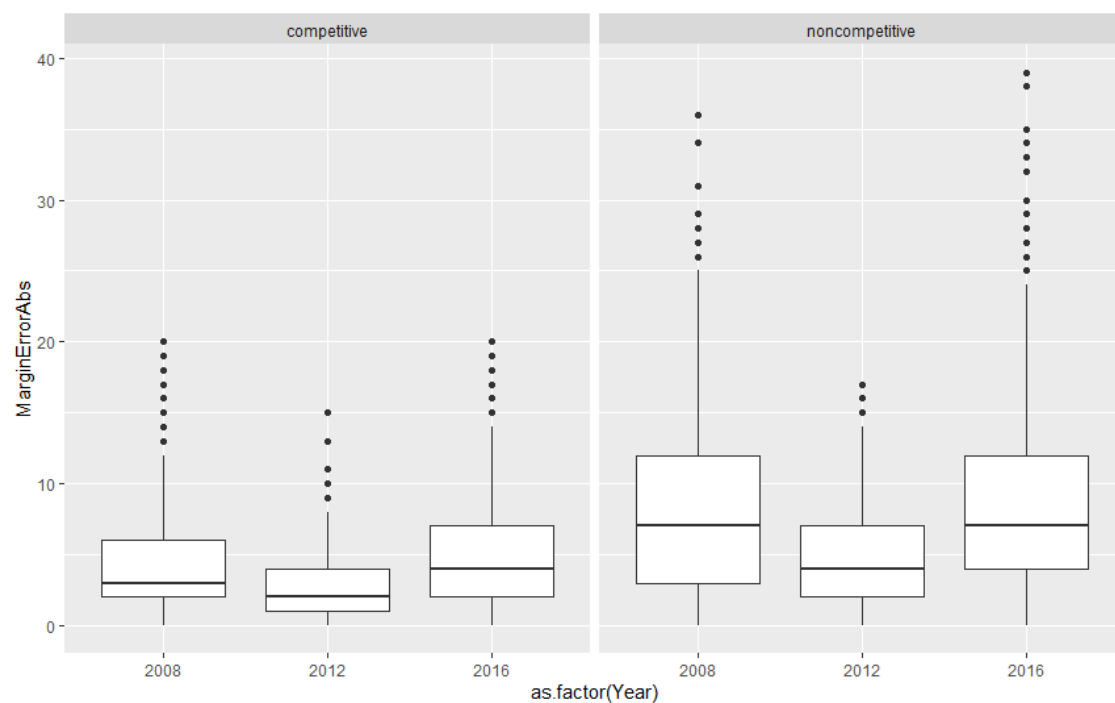
- Overall, the absolute errors in 2016 where not any higher than in 2008

- However a much closer race made these errors matter more

2016 wasn't abnormal in FL, MI, NC, PA, WI
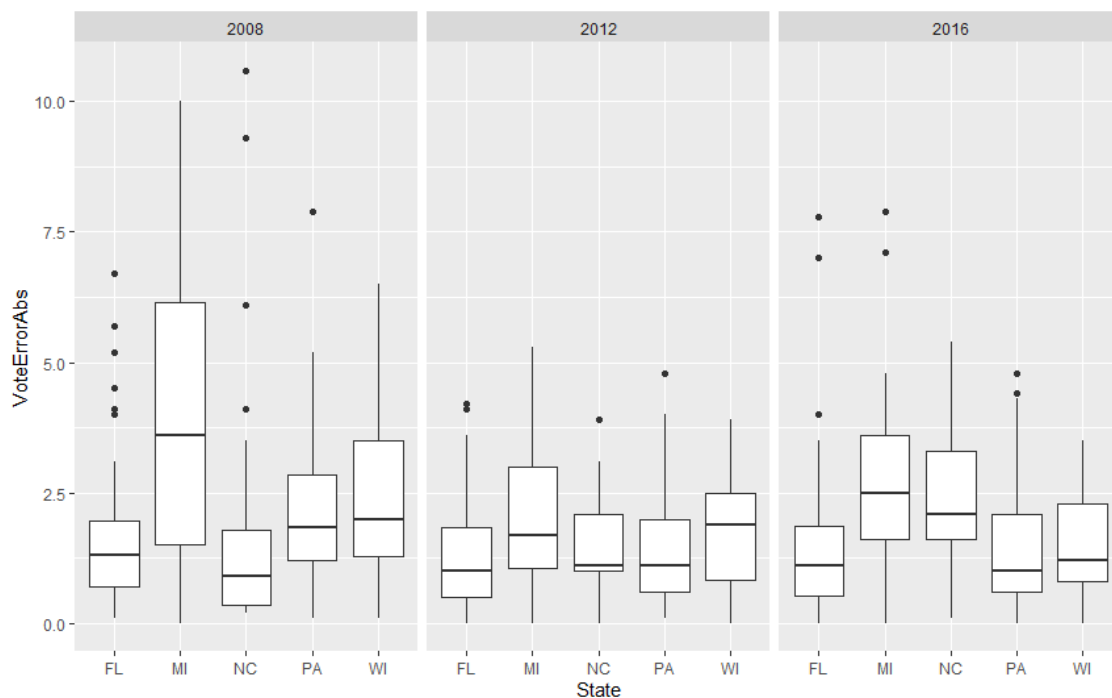
- Overall, the absolute errors in 2016 where not any higher than in 2008 or 2012 in the five states (FL, MI, NC, PA, WI) the polls generally failed to predict the winner on average
- This is based on last 60 days of polling
- The polls weren't more wrong, these races were just closer
- Given this data and how close the races were on election day it is not surprising Trump won
- These historical average errors should be used when interpreting 2020 polls

Conclusion

- Claims that in 2016 the polls were inaccurate or biased have little evidence to support them and don't look at the whole picture
- There is some non-sampling error that exists and should be accounted for
- Methods to predict non-sampling error in elections should be studied
- Pollsters who didn't weight for education should update their methodology
- Polls are incapable of distinguishing electability during a primary because polls at that time point are not predictive and accurate enough to do so

References

- Alexander, Brittany (2019), "A Bayesian Model for the Prediction of United States Presidential Elections," *SIAM Undergraduate Research Online,* **12**

- Bon, J. J., Ballard, T., & Baffour, B. (2019), "Polling bias and undecided voter allocations: US presidential elections, 2004–2016," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **182(2)**, 467-493.

- Hillygus, D. S. (2011). "The evolution of election polling in the United States," Public opinion quarterly, **75(5)**, 962-981

- Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018), "Disentangling bias and variance in election polls," *Journal of the American Statistical Association,* **113(522)**, 607-614.