Bayesian Modeling of Polling Data

Nick Ahamed

10 Jan 2017

The Problem

I am interested in using Bayesian modeling to predict how many seats Democrats will win in the 2018 House of Representative elections. This broad problem can be broken down into three composite tasks:

1. Use past election results and polling data to generate distributions for pollster and universe bias.
2. Estimate the relationship between past polling data and number of seats.
3. Apply the bias estimates from (1) to current polling for the 2018 election to generate a polling average, and then use (2) to predict the number of seats.

The Data

Description

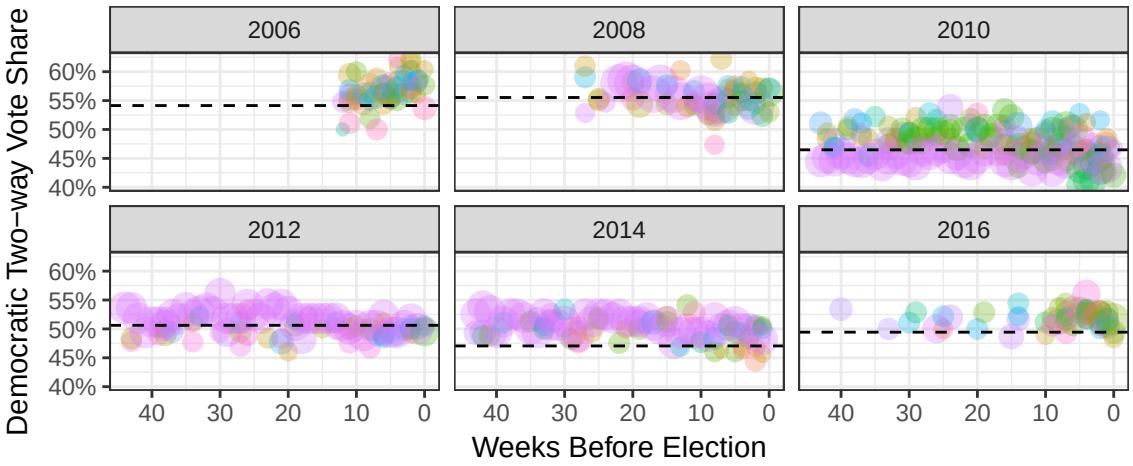
For this project, I have two primary sources of data: past polls and election results. The poll response that I use is the 'generic Congressional ballot.' Each pollster has a slightly different wording (and hence why we measure pollster bias), but they are all similar to: 'If the elections for the U.S. House of Representatives were being held today, which party's candidate would you vote for in your congressional district: The Democratic candidate or the Republcian candidate?' Asking specific candidate names would be preferrable because it would test voters' preference for actual people; voters may know the name of their incumbent and feel favorable to them even though they might otherwise support the other party, for example. However, I am interested in generating a prediction for 2018 and most candidates that will be on the ballot have not been decided (and may not be until close to the election). Thus, only generic congressional ballot polls are available for 2018 and I will only use these for past elections as well.

The past polls were taken from Real Clear Politics' database across 6 election cycles: 2006, 2008, 2010, 2012, 2014 and 2016. Only polls where the year, date range, pollster, sampling universe and sample size are all known were included. Additionally, the polls' results were transformed to reflect the two-way share for Democrats (Dem/(Dem+Rep)). Time is transformed to be the rounded number of weeks between the middle day of the poll and election day. A daily model would be more precise, but would take more computation time.

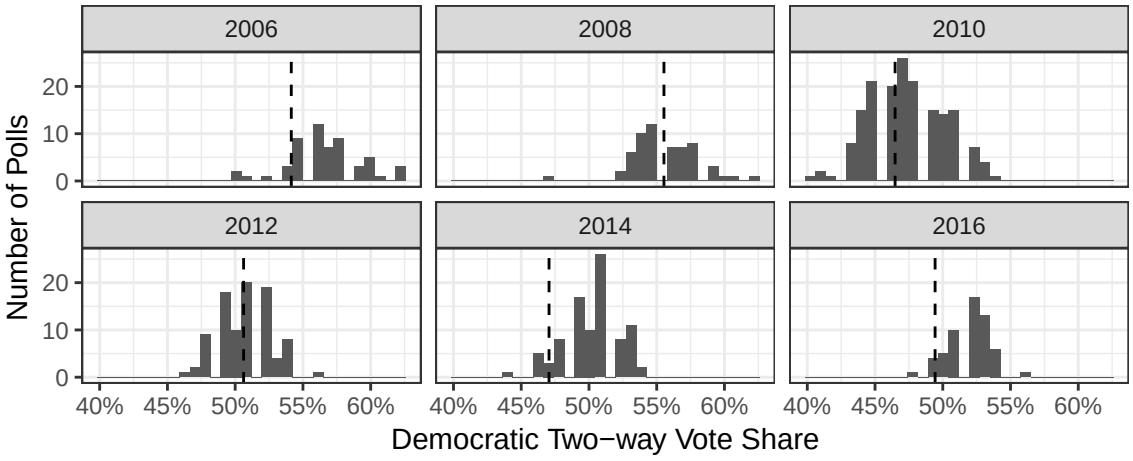
For election results, I use both the popular vote share and the seats won. These were taken from Wikipedia: 2006, 2008, 2010, 2012, 2014, and 2016. Again, I use Democrats' two-way vote share of the popular vote to mimic their two-way support in the polling data, and their percentage share of seats in the Congress.

Exploration

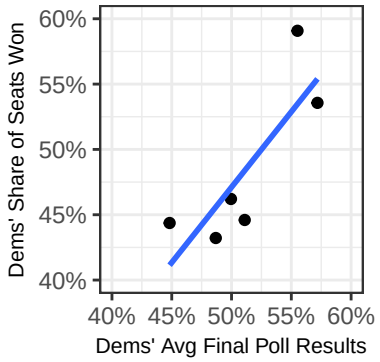
First, let's explore the trends over time in each cycle. Here, each point is a poll; it's size relfects the sample size and color represents the pollster. The dashed line represents the final two-way popular vote share of Democrats.



A couple of observations from this are clear. Polls are not conducted uniformly over time between elections. For example, 2006 had a lot of polls just before the election, whereas in 2010 many were conducted nearly a year in advance. We also see that by election, some pollsters are systematically off. For example, the pink pollster in 2010 was consistently below the final election result, suggesting bias. Last, we see that there are trends in results over time. For example, in 2014 the polls got closer and closer to the true result over time. Investigating this further, we see that poll results are not normally distributed around the result **across time**, suggesting we will need a time-dependent model.



Let's also investigate the relationship between polls and two-way seats. While I later improve upon this through modeling, a crude measure is the average poll result within 1 week of election day, weighted by sample size. Below I plot that on the x-axis and share of seats won on the y-axis with a linear line of best fit. We see that the greater the support in final polls, the greater share of seats won.



The Models

To answer question 1 above, I follow Jackman (2005) to specify my model to estimate biases, but with an added term for sampling universe. A given poll is assumed to be normally distruted with support as the mean and the standard deviation a function of y_i and sample size. This would be specified as:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

That poll is centered around mean μ_i, which itself is a function of α_t, the true value of support at the time the poll was taken t_i, δ_j, the bias of pollster j, and θ_k, the bias of sampling universe k. Fully specified, this is:

$$\mu_i = \alpha_{t_i} + \delta_{j_i} + \theta_{k_i}$$

Due to the trends we see in our initial data exploration, a random walk model is appropriate. In such a model, support at time t is normally distributed at time $t-1$.

$$\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \omega^2)$$

By including the final election results as the 'reference' for *delta*, and by using a random walk, I will be able to estimate the consistent bias of each pollster and the effect of different sampling universe.

For these given specifications, we have the following priors:

$$\sigma_i^2 = \sqrt{\frac{y_i(1-y_i)}{n_i}}, \quad \delta_j \sim \mathcal{N}(0,1), \quad \theta_k \sim \mathcal{N}(0,1), \quad \alpha_1 \sim \mathcal{U}(0.46, 0.56), \quad \omega^2 \sim IG(1/2, 1/2)$$

σ_i^2 just follows the standard formula for standard deviation of a sample. For pollster biases (δ), my prior is that there is no bias with a standard deviation large enough to capture total bias (awarding Democrats 100% when they actually had 0% support); my prior for bias from sampling universe (θ) is the same. As a prior for the starting true value of support (α_1), I use a uniform distribution over the minimum and maximum actual vote share of Democrats in the six elections analyzed. Lastly, a prior for the true standard deviation of support (ω), I use the inverse gamma distribution with an effective sample size of 1 and a prior guess of 1 like the standard deviation for δ and θ.

To answer question 2 above, I will use the pollster and universe biases estimated above, and the same random walk algorithm to generate a final polling average at the time of the election, α_E. I will then use the following model to estimate number of seats:

$$S_{cycle} \sim \mathcal{N}(\mu_{cycle}, \sigma^2)$$

$$\mu_{cycle} = \beta_0 + \beta_1 * \alpha_{E_{cycle}}, \quad cycle = 2006, ..., 2016$$

My priors for this model are:

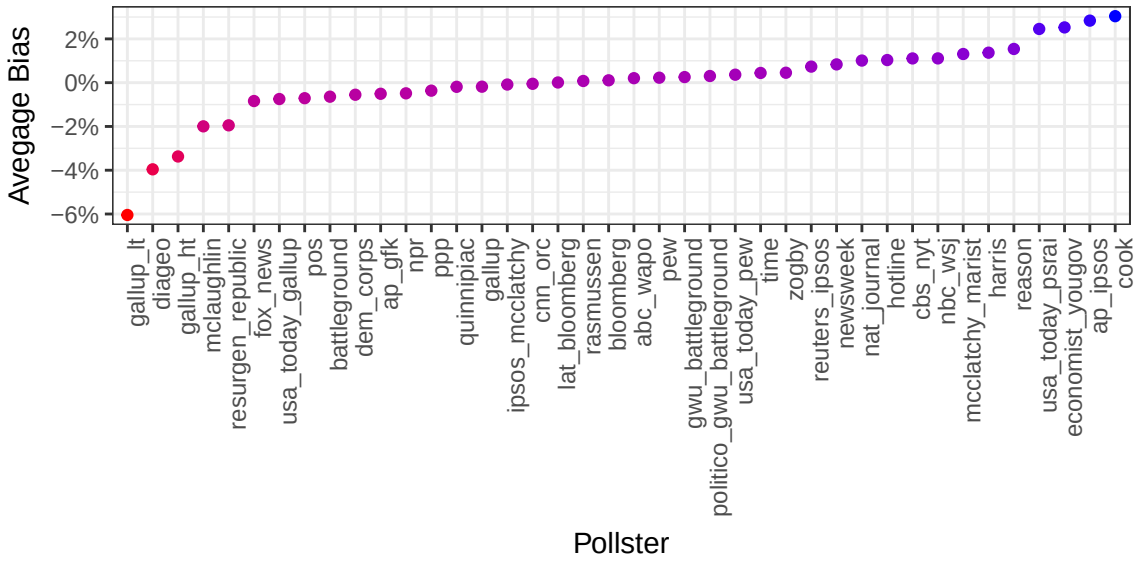
$$\beta_0 \sim \mathcal{N}(0,1), \quad \beta_1 \sim \mathcal{N}(1,1), \quad \sigma^2 \sim IG(1/2, 1/2)$$

β_0 here has a prior of none the seats in the House of Representatives with a standard deviation 1. β_1 has a prior that says a 1 unit increase in $\alpha_{E_{cycle}}$ (a 100 perctange point increase in the Democrats' modeled vote share) is associated with a 100 percentage point increase in the share of seats awarded to Democrats, with a standard deviation of the same. Lastly, I use an inverse gamma distribution with a prior guess of 1 and effective sample size of 1 for the standard deviation.

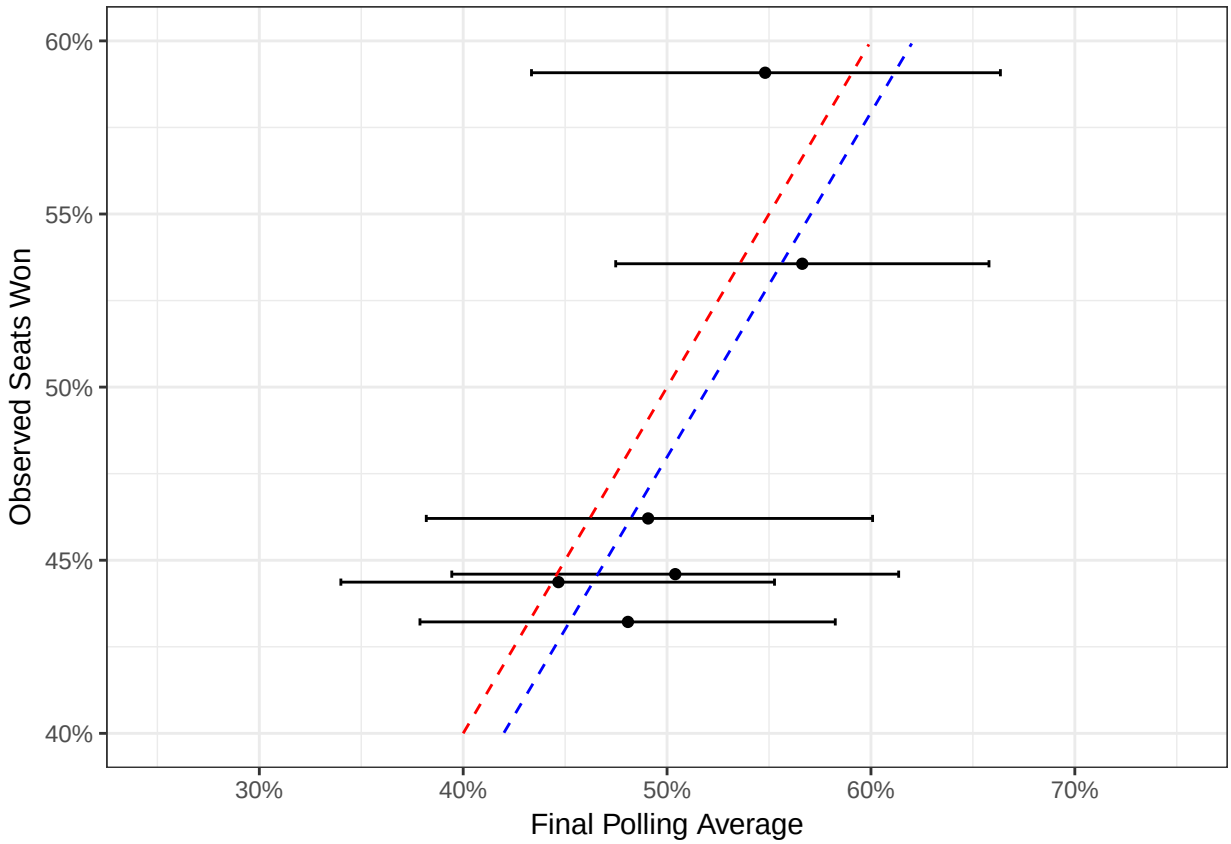
To answer question 3, I will use the same random walk algorithm already mentioned, along with the pollster and universe biases and distributions to generate a polling average for today. I will then use this α with the coefficients estimated in the second model to predict the number of seats Democrats will win in 2018.

For models to answer question 1, I used 250,000 iterations with 3 chains and a burn-in of 1000 iterations. Convergance was quick so this is all that was needed. Gelman and Rubin diagnostics for the model for each election are close to 1. Autocorrelation for some pollsters' bias was high, so I thined the chains, using only 1 in 10 samples. Residuals for poll results look normal.

Below I report the bias averaged across estimates for each election the pollster was active in. We see that across elections only a few pollsters are very biased one way or the other. With 'Gallup - Low Turnout' most consistently unestimating Democratic support and 'Cook' being the most consistently overestimating Democratic support.



To answer question 2, I used the average bias estimates from above to generate a polling average without anchoring the data with the final election result. I then used the previously described model to generate estimates for the relationship between share of seats won and final polling average. Below I plot the final polling averages with 95% credible intervals along with the estimated relationship in blue and a 1:1 relationship in red. I did not plot the 95% credible interval because it fell outside the normal range of values. Given more time and space, I would explore a logistic model so the SE is constrained within [0,1].


60%
55%
50%
45%
40%
Observed Seats Won
30%
40%
50%
60%
70%
Final Polling Average