Bayesian Modeling of Polling Data

Nick Ahamed

Summary

I am interested in forecasting the number of seats Democrats will win in the 2018 elections. I use random walk bayesian models of past elections to generate estimates of bias for pollsters. Likewise, I estimate the relationship between past election forecasts and Democrats' seats. Using these estimates, I model a current true level of support for Democrats (54%). Combined with the second model, I estimate Democrats will win 225 seats.

Introduction

To answer the above problem, I break it into three component modeling 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 to generate a support forecast, and then use (2) to predict the number of seats.

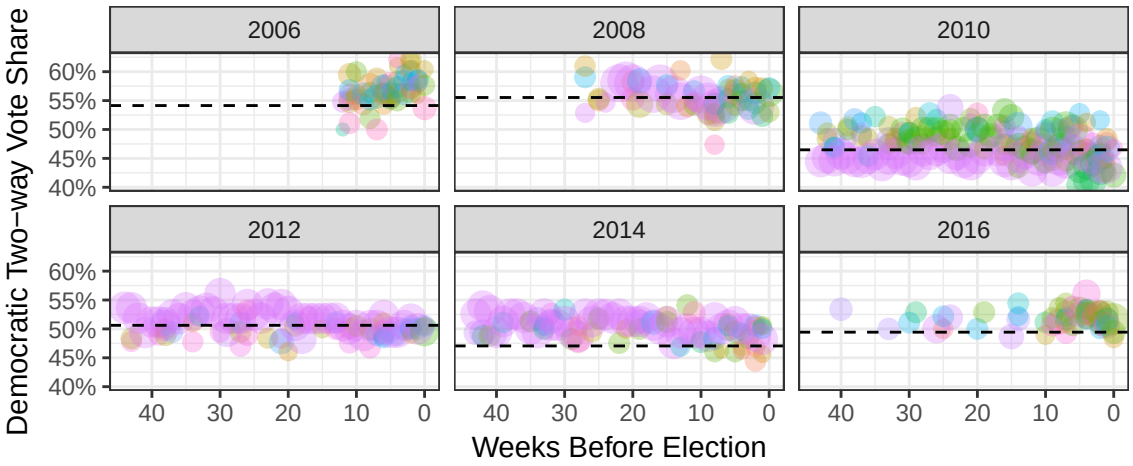
The Data

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?' Since not all candidates are known for 2018 yet, this is the only current question being polled, and so for comparability, I will use the same question for past elections.

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)): it is a proportion between 0 and 1. 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 resources.

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.

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. We 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. Further investigation shows that poll results are not normally distributed around the result **across time**, suggesting we will need a time-dependent model.



It's also worth exploring the relationship between polls and two-way seats won. 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. The correlation between this and two-way seat share is 0.82 suggesting a strong positive relationship.

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, 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 around support at time $t-1$.

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

By anchoring the model in the final election results, and by using a random walk, I will be able to estimate the consistent bias, δ, of each pollster and the effect, θ, of different sampling universes.

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 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 100% bias; 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, as 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:

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

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

My priors for this model are:

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

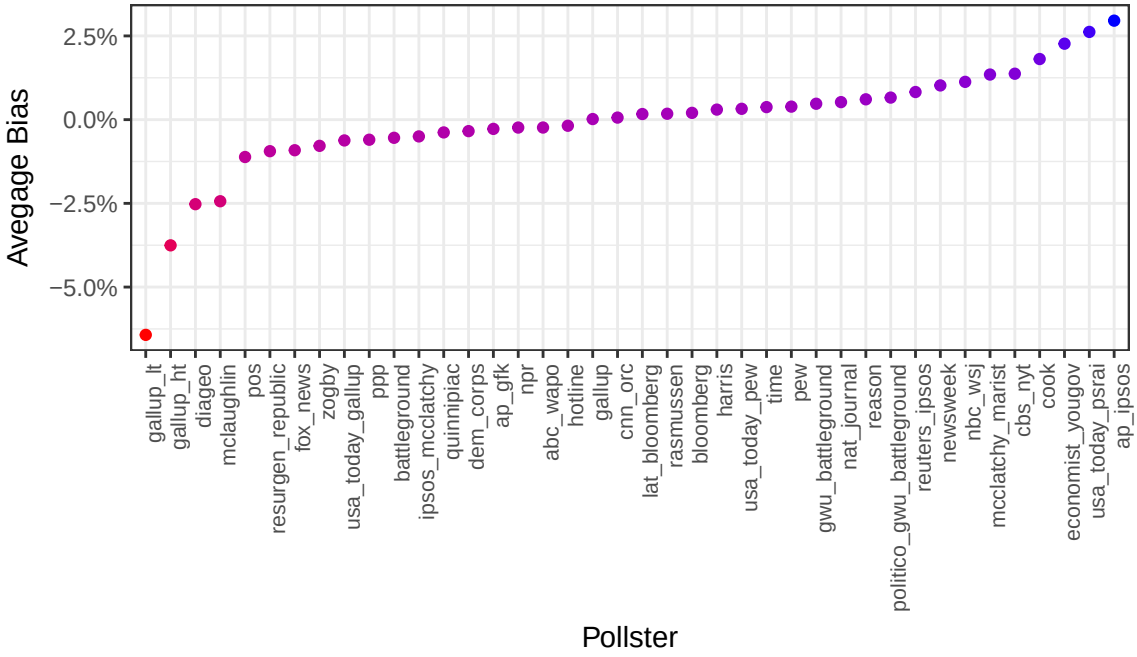
β_0 here has a prior of 0 seats in the House of Representatives with a standard deviation 1. β_1 has a prior that says a 1 unit increase in $\alpha_{E_{cycle}}$ (a 100 perctange point increase in the Democrats' modeled vote share) is associated with a 100 percentage point increase in the share of seats 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 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.

Model Evaluation & Results

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.

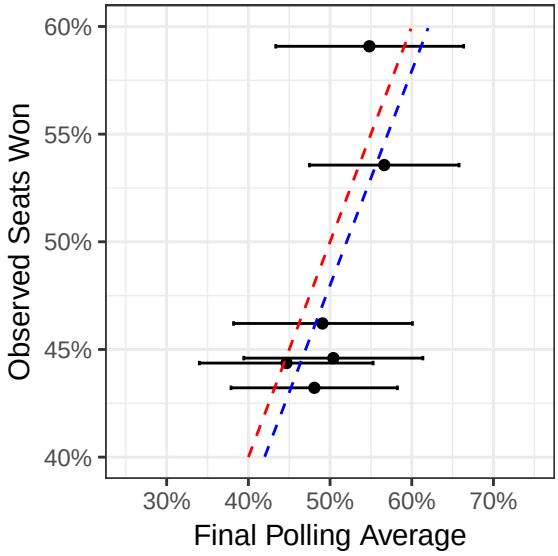
Below I report the bias averaged across estimates for each election the pollster was active in. We see that across elections only a few pollsters are very biased one way or the other. 'Gallup - Low Turnout' consistently unestimates Democratic support and 'Cook' consistently overestimates it.



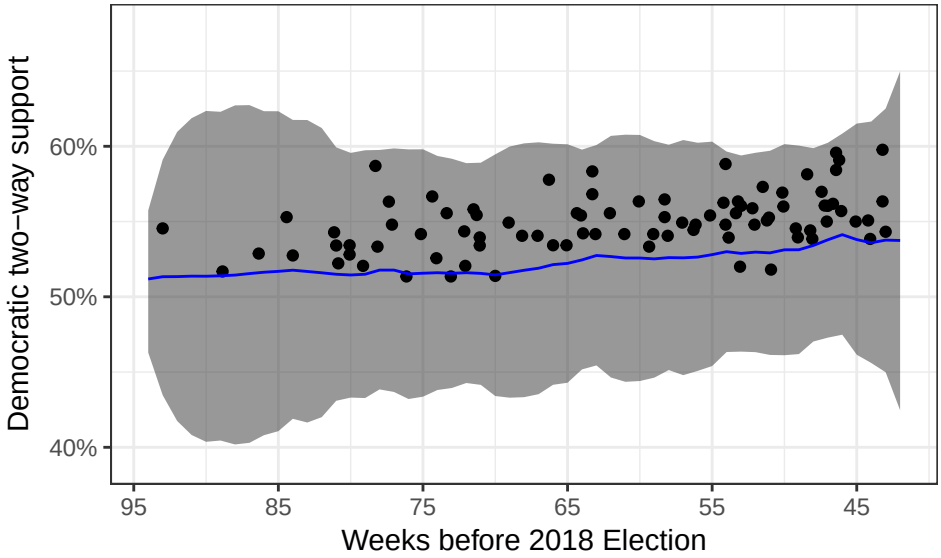
To answer question 2, I used the average bias estimates from above to generate a polling average **without** anchoring the data to the final election result. I then used the previously described model to generate estimates for the relationship between share of seats won and final polling average. Again, I used 250,000 iterations and 3 chains with a burn-in period of 1,000 iterations, and thinned by 10. Gelman and Rubin diagnostics, autocorrelation and residual checks were all satisfactory.

Below I plot the final polling averages with 95% credible intervals along with the estimated linear relationship in blue and a 1:1 relationship in red. I did not plot the 95% credible interval for the regression because it

fell outside the normal range of values. Given more time and space, I would explore a logistic model so it is bounded by [0,1].



Finally, I use a final random walk model on 2018 polling data. I only used pollsters and sampling universes that had polled in previous elections so I would have average bias estimates already. The blue line below represents the model's assesment of the true value of support; it is lower than most polls due to the fact that most pollsters overestimate democratic support, as seen above. However, there is a wide 95% credible interval, especially during times where there were few polls.



Conclusions

Using the estimate for the true current level of support, about 54%, and the parameter estimates from the regression model previously fit, I predict democrats will win about 52% of the seats, or 225 seats, with a 2.5% lower bound of 177 seats and a 97.5% upper bound of 273 seats. This estimate is similar to other's. For example, one respected **author** finds an 8pp advantage in the generic ballot for Democrats will yield 224 Democratic seats.