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

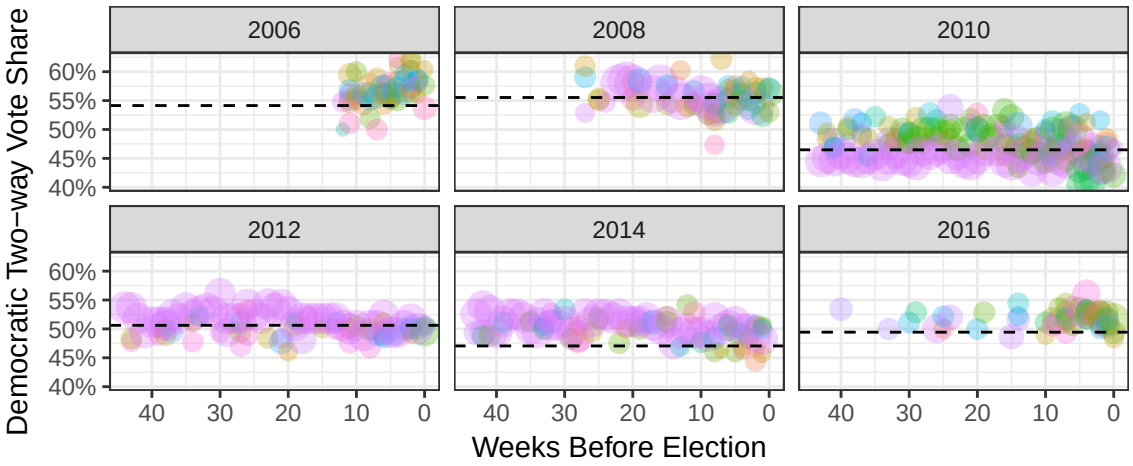
For this project, I have two primary sources of data: past polls and election results. The poll response that I use is the 'generic Congressional ballot.' Each pollster has a slightly different wording (and hence why we measure pollster bias), but they are all similar to: 'If the elections for the U.S. 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.

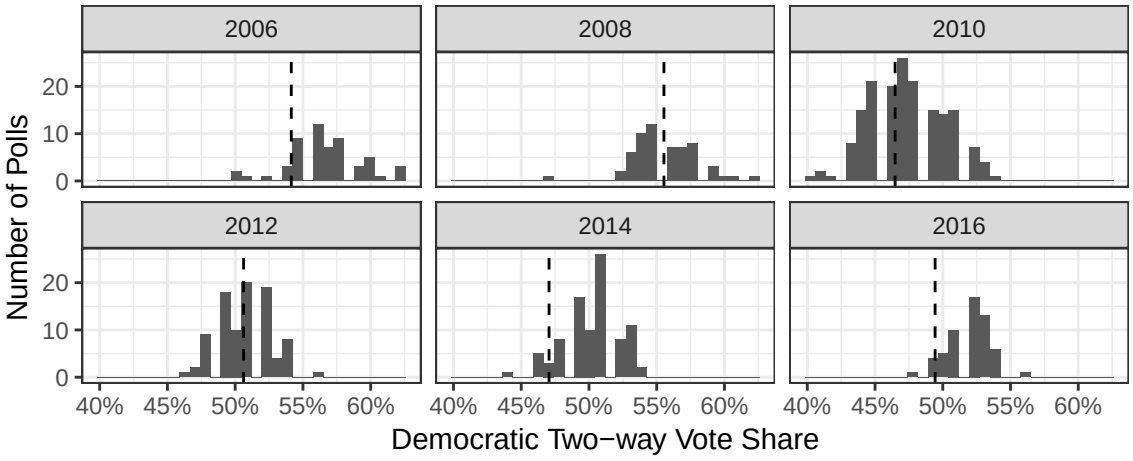
For election results, I use both the popular vote share and the seats won. These were taken from Wikipedia: 2006, 2008, 2010, 2012, 2014, and 2016. Again, I use Democrats' two-way vote share of the popular vote to mimic their two-way support in the polling data, and their percentage share of seats in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