Challenge 1 — Iowa Caucus Predictions

Jeffrey Barrera & Jacob Fenton

Predictions

We predict Donald Trump to be the winner of the 2016 Iowa Republican caucus.

Candidate	Vote Share
Donald Trump	26.1%
Ted Cruz	25%
Marco Rubio	15.9%
Jeb Bush	4.3%
Ben Carson	11.1%
Chris Christie	3.4%
Rand Paul	6.8%
Mike Huckabee	4.3%
John Kasich	3.2%

Methodology

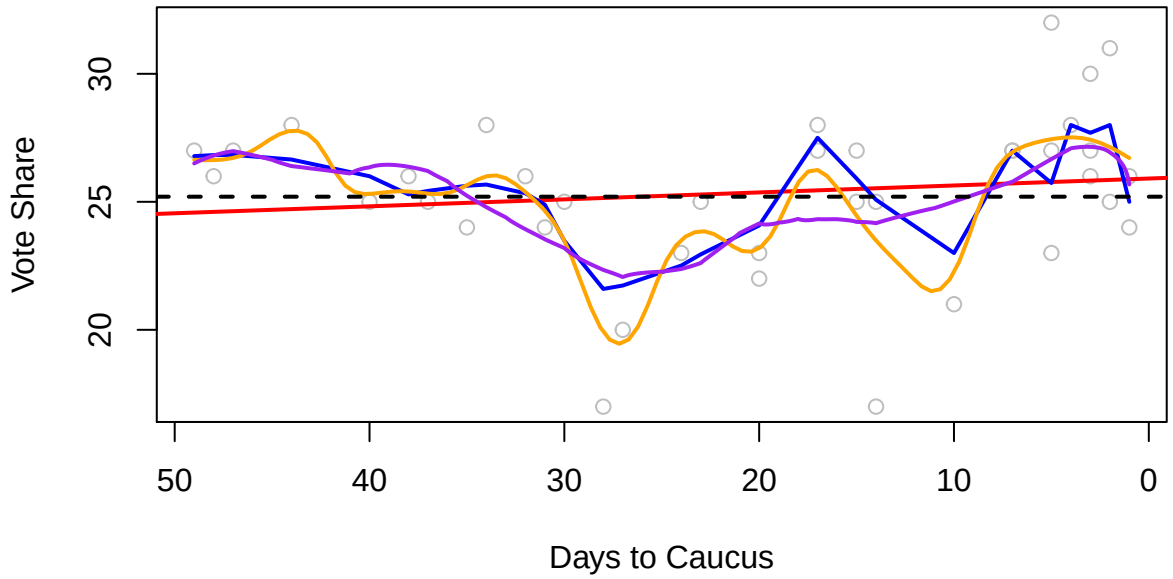
Key Feature: Iowa Polling Trends

We focused our attention on in-state polling leading up to the caucus, since this has historically been the best indicator of a candidates' standing in Iowa. We used polls aggregated for Iowa by pollster.com (which was later acqured by *The Huffington Post*) for 2008 and 2012. (All of the code used in this project is available at: https://github.com/jeffbarrera/iowa-caucus/). We wrote python scripts (2008 ; 2012/16) to standardize the data across years and add one key variable: the number of days before the Iowa Caucus is held.

Our core challenge with this polling data is twofold: First, to calculate a polling average that's accurate as of the present; and second, to use our knowledge of the present to estimate the final caucus vote totals.

We tested a number of approaches for finding a polling average by looking at 2008 and 2012 polling data

(which goes all the way to caucus): linear regression, lowess regression and non-parametric regression with gaussian and epanechnekov kernels.



Comparison of different extrapolation techniques applied to Mitt Romney in 2012: actual vote share in black, linear regression in red, lowess in blue, gaussian in orange, and epanechnekov in purple.

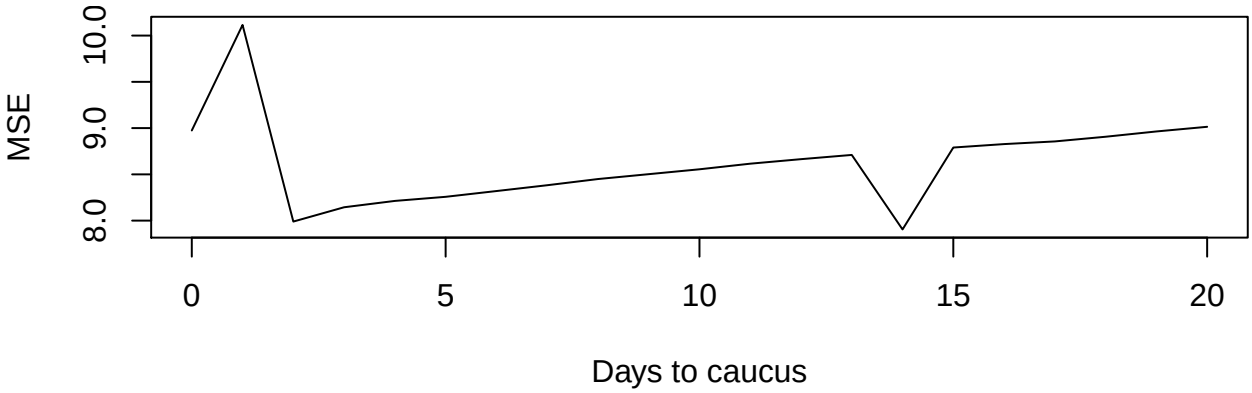
Testing scripts are here for lowess (results); gaussian (results) and epanechnikov (results) kernels. We tested each approach with a variety of bandwidths (or f values in the case of lowess) and compared the final estimated point with the actual polling result. The most accurate estimation result in terms of overall MSE for 2008 and 2012 was obtained using an epanechnikov kernel with a bandwidth of 13 days, which gave us an MSE of 7.29.

Arguably running this test naively–that is, just testing it on 2008 and 2012 without any cross-validation–could lead to a type of overfitting, but we'd expect the size of improvement to be minimal. We noticed that MSE was minimized for the 2008 and 2012 cycle with a bandwith of 13 and 14 days respectively, which suggests there are similar dynamics (despite the races being vary different).

	Linear Model	Lowess	Gaussian	Epanechnikov
Lowest MSE	8.52	7.89	9.63	7.29

To answer the second half of the problem — how to interpolate from a polling average several days ahead of the caucus to a final vote share on caucus day — we assumed it would be best to start with the best polling average result. We tested a number of possible measures of the 'momentum' by drawing a line from

the final average poll result and the result from several days earlier. Our testing suggested 14 days worked best. Again, this wasn't cross-validated in any way, but it sounds good — that's the last point outside of the epanechnikov window. We also assume that interpolation error isn't huge (Santorum 2012 aside).



We then applied this model to generate a predicted vote share for each candidate in the 2008, 2012, and 2016 caucuses.

National Polling Trends

In their "polls-plus" model, 538 uses national polling as a contrarian indicator, based on data suggesting that candidates who poll better in a particular state than they do nationally tend to do better than their statewide polls. We adopted a similar approach, applying the linear and epanechnikov techniques we used to estimate statewide polling trends to national polling for each candidate.

Iowa Polling Average

In case our interpolated projections were worse than the unimproved polling average several days out, we also included as a feature the vote shares for each candidate in the days leading up to the election. We tested from 2 to 21 days out from the caucus, and calculated the RMSE for each interval. The lowest RMSE was at 4 days in 2008 and 3 days in 2012, so we went with the rounded average of 4 days.

Features Not Included

Campaign Finance Data

Reports filed with the Federal Elections Commission give some insight into a candidates' fundraising and spending, but we didn't make analysis of those a priority for several reasons:

- This year is different! One candidate (guess who) has been the beneficiary of millions in "earned media" — coverage that's not paid for.
- Candidates' reports are filed at a significant lag. Quarterly reports covering the fourth quarter of 2015 are due Jan. 31, but do not reflect any spending or fundraising that took place in 2016.
- The 2012 and 2008 Iowa caucuses were held Jan. 2 (two days after the end of a filing period) whereas the 2016 caucuses are held Feb. 1 (a month after the end of the most recently available candidate spending data). Thus a relationship between financial figures for 2008 and 2012 wouldn't necessarily hold true for 2016.
- The way that campaigns spend money is in flux and increasingly money spent is excluded from public accounting.

Endorsements

Fivethirtyeight uses a weighted endorsements system to help predict primary results. What their data show, this year, however, is that there are far fewer endorsements this year than in previous cycles. The outsider nature of several candidates, and their relative paucity of endorsements to date, also makes us skeptical of this measure.

Crosstabs available in polls

Most reputable polls provide results cross-tabulated by various demographic groups. Unfortunately, we were unable to find any easily available aggregation of poll crosstabs (and the inconsistent approach pollsters take would make this a considerable challenge). Nonetheless, we believe this might be a useful indicator. Were this data available in bulk we might be able to make different assumptions about the electorate. Data suggest many potential voters who say they plan to participate in caucuses do not actually do so; we believe voter subgroups' lie to pollsters at a differential rate, introducting a meaningful bias into polls. Careful consideration of prior years voter crosstabs and exit polls (confusing known as entrance polls in Iowa) might help.

Model

Once we had these predictors, we used a LASSO regression to determine how to weight each feature. Treating the actual vote share for each candidate in 2008 and 2012 as our training Y variable, we calculated β

coefficients for each feature at different values of λ. We then used leave-one-out cross-validation to see how each model performed out-of-sample. We selected the model with the lowest out-of-sample MSE, in the hope that this would be predictive in 2016 but would also avoid overfitting to the 2008 and 2012 data.

Our final model used these coefficients:

```
## linear_iowa linear_natl     ep_iowa     ep_natl    avg_iowa
##         0.91        0.05        0.00       -0.13        0.15
```

We then used this model to predict vote shares for each of the 2016 candidates. Finally, since the sum of these predictions likely would not exactly equal 100%, we scaled them proportionally to produce our estimates of vote share.

Limitations & Opportunities for Improvement

The biggest limitation of our model is that it does not include any features intended to measure voter turnout or the "ground game" — the campaigns' efforts to identify supporters and get them to show up to caucus. Historically, this has been a critical aspect of winning in Iowa: since caucuses typically have low turnout rates, effectively mobilizing supporters can have a big impact on a candidate's vote share. However, we were unable to find a good way to measure organizing operations or predict turnout, since we couldn't obtain useful campaign finance or crosstab data.

Ultimately, we had to rely on assumptions about the relationship between polling and turnout. All the polls we're aggregating are of "likely voters," and some of the the polling firms also weight their responses based on estimated turnout models. We thus hope that the relationship between these polls of likely voters and the final vote share is reasonably consistent across elections, and that the coefficients in our LASSO model will capture this relationship.

If this relationship is not consistent, however (perhaps because supporters of "outsider" candidates like Trump may be more likely to lie to pollsters about whether they will caucus), this could throw off our model. Given more time and detailed cross-tabulated data, it may be possible to explore and account for these differences, and produce more nuanced models of the relationship between a candidate's poll numbers, turnout rates, and actual vote share. This could be an interesting avenue to explore in the future.