

# Challenge 1 — Iowa Caucus Predictions

*Jeffrey Barrera & Jacob Fenton*

## Predictions

We predict **XXX** to be the winner of the **2016 Iowa Republican caucus**.

Candidate	Vote Share
Donald Trump	0
Ted Cruz	0
Marco Rubio	0
Jeb Bush	0
Ben Carson	0
Chris Christie	0
Rand Paul	0
Mike Huckabee	0
John Kasich	0

## Methodology

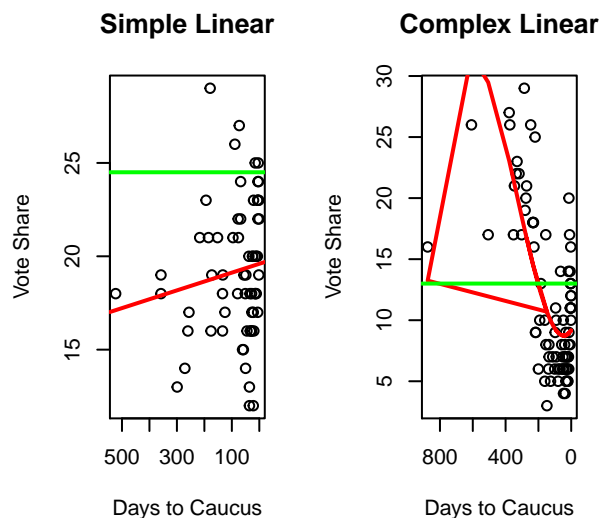
### Features

#### 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 the candidates' standing in Iowa. We used polls aggregated for Iowa by pollster.com (which was later acquired 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 was how to extrapolate a predicted vote share on caucus day from

the multitude of polls conducted days or weeks earlier. To predict out a trend line to caucus day, we tested a number of approaches: linear regression, lowess regression and non-parametric regression with gaussian and epanechnikov kernels.



*Comparison of different extrapolation techniques: model line in red, actual result in green.*

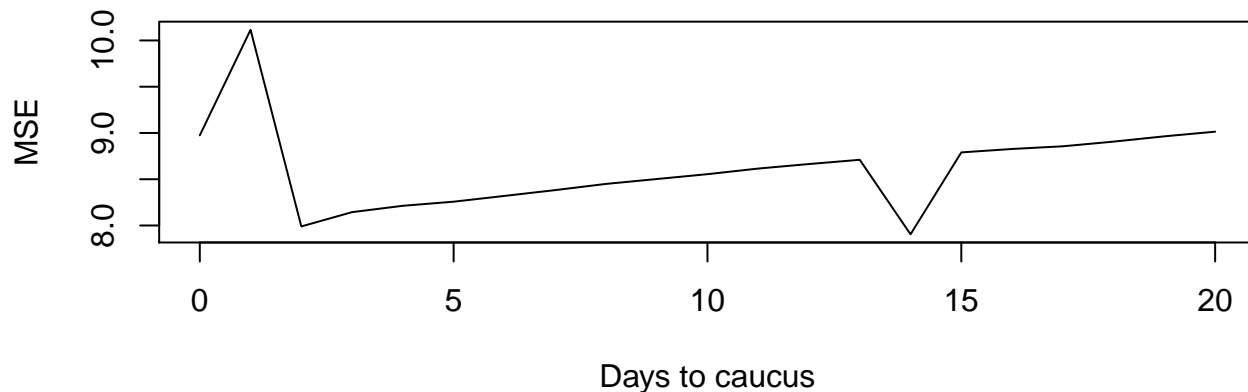
We wrote scripts to test several techniques under a variety of circumstances: [lowess \(results\)](#), and first-order non-parametric regression with [gaussian \(results\)](#) and [epanechnikov \(results\)](#) kernels. We tested each approach with a variety of bandwidths (or  $f$  values in the case 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](#).

	Simple Linear Model	Complex Linear Model	Lowess	Gaussian	Epanechnikov
Lowest MSE	38.18	24.94	7.89	9.63	7.29

## TK TRENDLINE WIDTH

One additional complication is that prior years' polling included results the day of the caucus; the most recent poll results we have access to are several days ahead of the caucus. Thus we had to interpolate from a point several days ahead of the caucus to a final result; we tested several approaches and found the best to be TK TK.

## Choosing the period to interpolate over



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 epanechnikov technique we used to estimate statewide polling trends to national polling for each candidate.

### Iowa Polling Average

In case our trend projections were placing too much weight on the direction of the polling, we also included a simple average of 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

We considered a number of additional features, but chose not to include them for various reasons:

#### Campaign Finance Data

Reports filed with the Federal Elections Commission give some insight into a candidates’ fundraising and spending, but we chose to disregard these as not predictive of vote share 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. Because he’s been so effective in winning earned media, he hasn’t sought contributions in the same manner as other candidates. Thus many of the usual governing assumptions (probably) don’t hold.
- 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. Polling data is generally much more current than that.
- 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. Increasingly spending from a candidates’ principal campaign committee is overshadowed by money spent by independent expenditure only committees (aka super PACs). In 2012, super PACs primarily spent money on media buys, but increasingly these “outside” groups are taking on tasks previously handled by candidate committees, including last-minute voter targeting and mobilization. Other money spent by non-profit groups is not publically reported at all, and anecdotal reports suggest this type of spending is rising.

## **Endorsements**

Fivethirtyeight uses a weighted endorsements system to help predict primary results. What their data show<sup>[footnote]</sup>, however, is that there are far fewer endorsements this year than in previous cycles.

## **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, introducing a meaningful bias into polls.

## 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  $\beta$  coefficients for each feature at different values of  $\lambda$ . 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. The final coefficients in this model were:

```
## lm_1_iowa lm_2_iowa lm_1_natl ep_iowa ep_natl avg_iowa
##      0.00      0.00      0.00      0.94      0.00      0.00
```

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.