Challenge 1 — Iowa Caucus Predictions

Jeffrey Barrera & Jacob Fenton

```
##   candidate vote_share
## 1  huckabee        NaN
## 2    romney        NaN
## 3  thompson        NaN
## 4    mccain        NaN
## 5      paul        NaN
## 6  giuliani        NaN
```

```
##          lm_1_iowa lm_2_iowa lm_1_natl   ep_iowa   ep_natl  avg_iowa
## huckabee 28.345420 28.801950 25.010638 28.869357 21.291409 29.200000
## romney   27.549618 26.669361 15.000000 29.256954 15.484202 27.800000
## thompson 10.982824 11.977731  0.000000 11.375056  0.000000 10.600000
## mccain   13.112595 11.617325 16.095745 12.158672 14.271585 12.800000
## paul      6.902672  8.639657  3.095745  6.616153  3.515044  7.000000
## giuliani  5.895038  5.816867 21.648936  5.217361 21.349279  6.400000
## santorum 17.402961 18.837681  2.083558 19.504119  2.686278 17.000000
## romney   23.071206 22.911543 26.514825 23.485236 24.990564 22.000000
## paul     19.230738 19.240705 12.846361 19.512301 14.163681 21.333333
## gingrich 13.555444 12.959580 20.566038 12.022257 19.522251 14.000000
## perry    10.430960  9.362463  5.149596 10.616829  6.809332 10.333333
## bachmann  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
## huntsman  2.772485  3.217364  1.537736  2.727386  1.716597  2.666667
##          lm_1_iowa lm_2_iowa lm_1_natl   ep_iowa   ep_natl avg_iowa
## trump    29.164063  9.628023 35.583463 26.805189 33.543468     23.5
## cruz     24.717722 22.798427 15.629718 24.516817 16.733275     24.5
## rubio    15.260691 11.939531 10.335291 14.843658 10.618634     13.5
## bush      3.899054  4.188242  5.491312  3.401568  5.426284      3.0
## carson    9.054276 17.232851  8.887358 10.142133  9.751513     11.5
## christie  1.604235  1.658057  2.080647  1.826542  2.094138      2.0
```

```
## paul      4.442434  8.311828  0.000000  4.839894  0.000000      6.0
## huckabee  2.624383  4.597916  1.731096  2.864345  1.604311      3.0
## kasich    1.807360  1.286787  2.557819  1.842860  2.653244      1.5
## [1] "############# LINEAR #############"
##
## Call:
## lm(formula = results_train$percentage ~ ., data = as.data.frame(preds_train))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.236 -1.251  0.204  1.267  3.026
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9742     2.1331   0.926    0.390
## lm_1_iowa    -1.6210     2.5673  -0.631    0.551
## lm_2_iowa     1.1982     1.1205   1.069    0.326
## lm_1_natl     1.1548     0.8804   1.312    0.238
## ep_iowa       0.7487     1.7804   0.421    0.689
## ep_natl      -1.3281     0.9244  -1.437    0.201
## avg_iowa      0.7360     1.3005   0.566    0.592
##
## Residual standard error: 2.953 on 6 degrees of freedom
## Multiple R-squared:  0.9565, Adjusted R-squared:  0.9129
## F-statistic: 21.97 on 6 and 6 DF,  p-value: 0.000772
##
##      trump       cruz      rubio       bush     carson   christie
##  0.1430955 21.4375905 10.4247062  4.5617452 21.3155593  3.8215069
##       paul   huckabee     kasich
## 12.7720697  7.4503349  2.5000774
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```

```
## [1] "Training RMSE:"

## [1] 2.329601

## [1] "############# LASSO BETAS #############"

##  lm_1_iowa  lm_2_iowa  lm_1_natl    ep_iowa    ep_natl   avg_iowa
##  0.0000000  1.1442176  0.0000000  0.0000000 -0.1195629  0.0000000

##          sorted_preds
## cruz        27.417292
## carson      21.321157
## rubio       14.534358
## paul        11.360091
## trump        8.600876
## huckabee     6.467138
## bush         5.447284
## christie     2.696731
## kasich       2.155073
```

Predictions

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

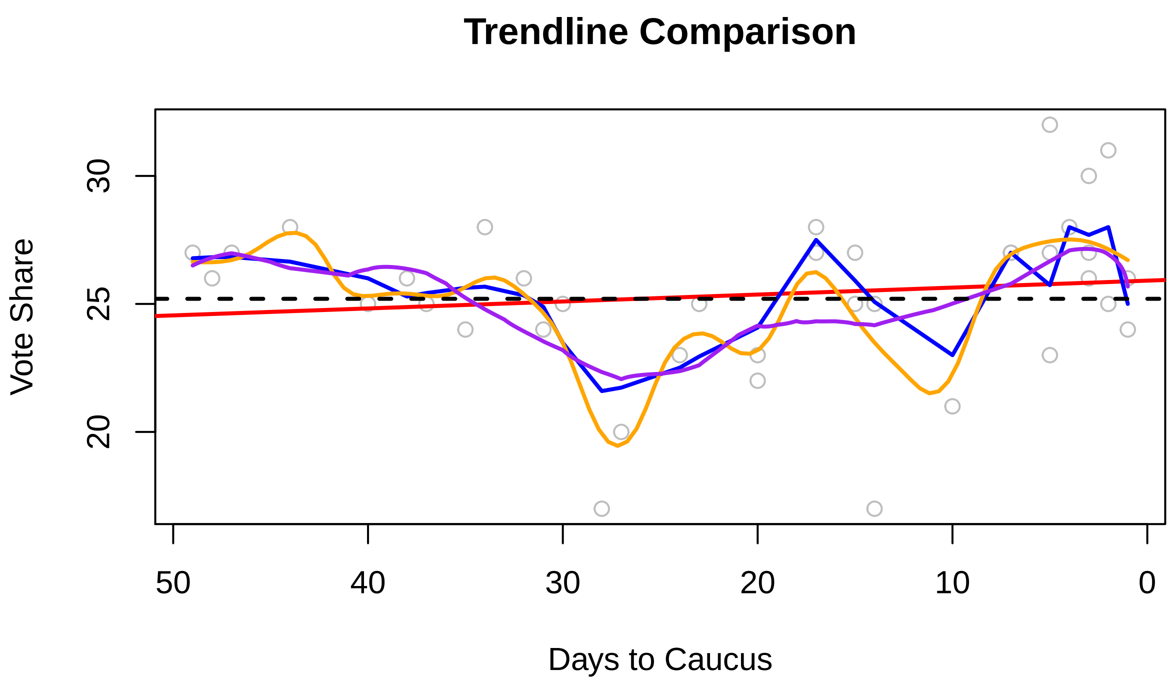
Candidate	Vote Share
Donald Trump	0
Ted Cruz	0
Marco Rubio	0
Jeb Bush	0
Ben Carson	0
Chris Christie	0
Rand Paul	0
Mike Huckabee	0
John Kasich	0

Methodology

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 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 epanechnekov kernels.



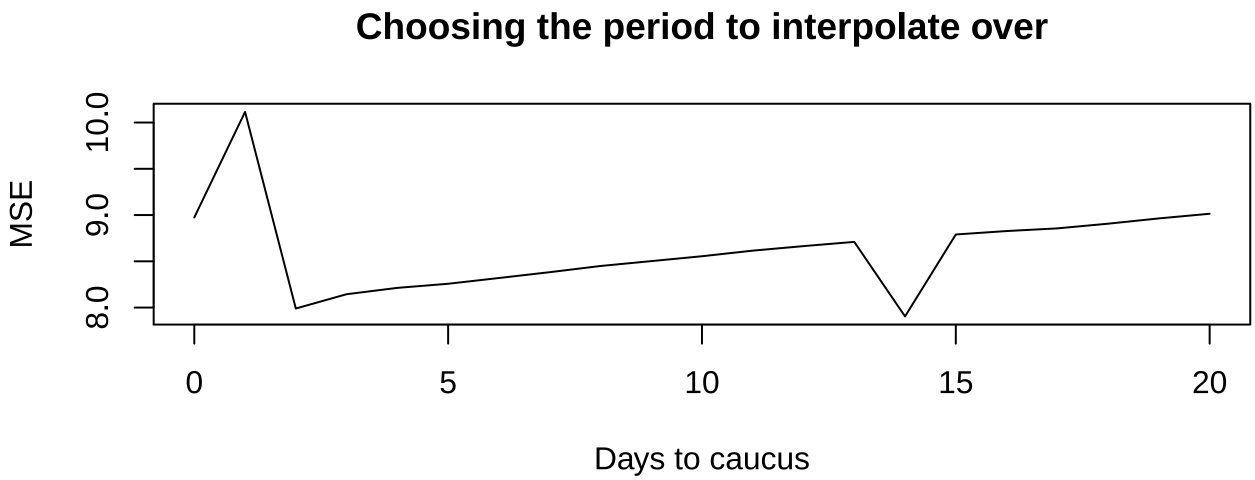
Comparison of different extrapolation techniques applied to Mitt Romney in 2012: actual vote share in black, simple linear regression in red, lowess in blue, gaussian in orange, and epanechnekov in purple.

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	11.58	8.97	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.



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

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:

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 tried a similar approach, applying the epanechnikov technique we used to estimate statewide polling trends to national polling for each candidate. However, we ultimately decided that since our training data didn't have a comparable case of a celebrity frontrunner, including this predictor skewed our results – Trump was likely being over-penalized for his high national poll numbers.

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 targetting 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.

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.

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. Ultimately, this approach led us to use a single predictor: .94 times the epanechnikov trendline estimate for election day.

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.