An Analysis of Polls

Harry Bendekgey

June 14, 2018

Introduction

Methods in election forecasting are highly varied. Some propose historic voting models, where "fundamental" factors like incumbency and economic conditions are used to predict election results. These models have been shown to predict election results with a high amount of accuracy. However, these methods suffer from a few notable deficiencies.

Firstly, these models implicitly convey the notion that candidates and campaigns do not matter. Secondly, with so few data points (elections) to work with, it is difficult to parse the important information from the random noise. This is true of all models that use previous elections as precedent to build a forecast, but is particularly consequential here. "Economic conditions" is a vague enough term that some combination of them can surely be used to predict elections, opening up the door to p-hacking.

For example, the 2000 presidential election seemed inconsistent with the idea that when the economy is doing well, the incumbent party performs well nationally. However, if we consider real disposable income (RDI) growth, instead of gross domestic product (GDP) growth, we get a different picture of the economy that is entirely consistent with the results of the presidential election. Although the correlation between a poor economy and voter sentiment towards change seems intuitive, the ease with which definitions can be changed to retroactively make results consistent with these models should be concerning, at the least.

This leads us to the importance of trial-heat polls in forecasting elections. These polls, taken months, weeks, and days in advance of the election are meant to give an indication of national support for each major party candidate. However, these polls are much more variable than votes. Therefore, it is important to not throw out historic voting entirely, but combine these two data sources into a more robust model.

Campbell proposes a simple linear regression model to predict election results. Not only does this model only attempt to predict net change in congressional seats and presidential popular vote, (note that Presidential elections are decided at the state level) but it suffers from a larger problem, the same problem faced by historic voting models: there are so few data points. In addition, the country has become more partisan, with presidential candidate vote share getting closer to 50% over time. Thus the points with the highest leverage in this regression will be elections far in the past. In addition, if we tried to move this model down to the state level these problems would be more pronounced due to large-scale changes in the way states vote (consider, for example, the Republican stronghold known as California in the decades preceeding 1992).

I argue that due to the paucity of precidents, and inherent bias in the data, no purely machine learning model could "learn" its way to good election prediction without someone having to answer hard questions about what data to consider, and how to consider it.

In the wake of the 2016 presidential election, many pundits and even statisticians such as Drew Linzer asked themselves how much trust we should put in polling accuracy after they failed to predict a Trump victory. Of the notable forecasting models for the 2016 presidential election, Princeton put Trump's chance of winning at less than 1%. ¡List notable forecasts¿. Fivethirtyeight, on the other hand, put Trump's chance of winning at 30%, and have argued extensively that polls are not the ones to blame for pundits' and statisticians' inaccuracy. Indeed, Fivethirtyeight was an outlier which gave Trump a higher chance of victory than anyone else. Linzer argues that this is because they put less trust in the polls than other forecasters, increasing the chance of a surprise in either direction.

This paper is an attempt to answer these fundamental questions: can we trust polls? Were they actually bad in 2016? How can we use them well in a forecasting model? Fivethirtyeight maintains a collection of all polls conducted in the last 21 days of each election since 1998. I have removed every poll by a pollster banned by Fivethirtyeight. “Pollsters that are banned by FiveThirtyEight because we know or suspect that they faked their data or we are otherwise not confident in the legitimacy of their polling operation.” 7 Out of 396 pollsters are banned by Fivethirtyeight.

It is worth noting that polls conducted close to the election are siginificantly more accurate than those made long before, and so my analysis is constrained to end-of-election polls.

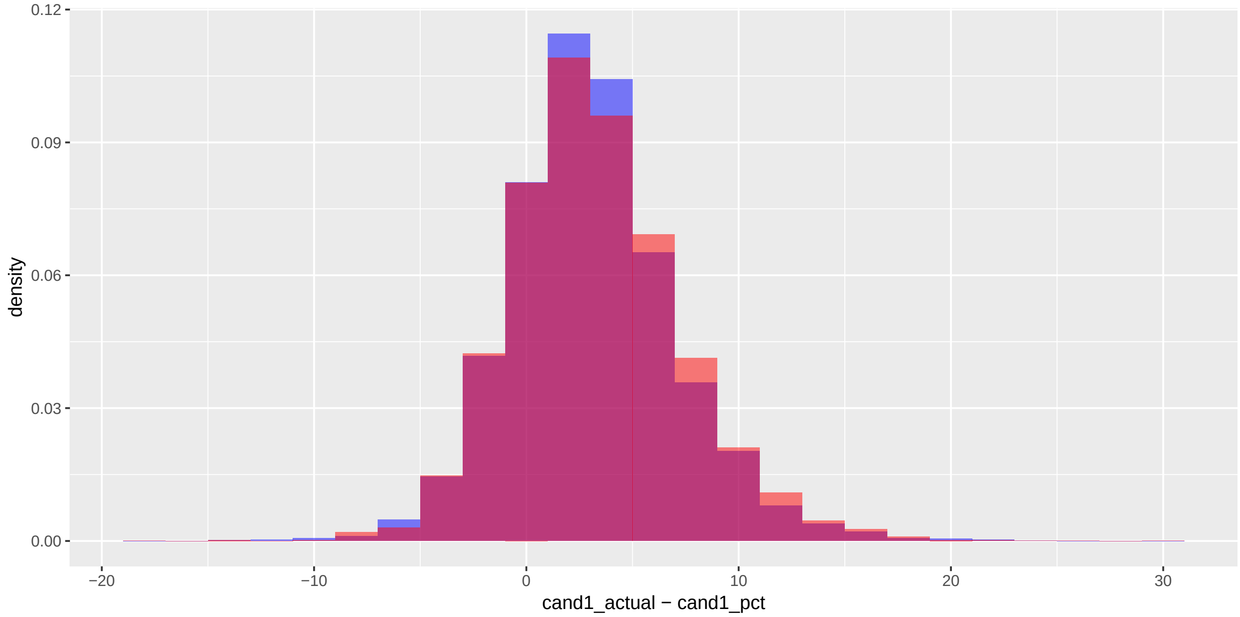
```
all_polls <- read_csv("~/SummerResearch/raw-polls.csv",
                      col_types = cols(cand1_pct = col_double(),
                                       cand2_pct = col_double(),
                                       cand3_pct = col_double(),
                                       margin_poll = col_double(),
                                       samplesize = col_double())) %>%
  filter(year < 2018) %>%
  filter(location!="US" | polldate!="10/17/14")
#remove 2018 polls so my analysis is not sensitive to when in 2018 I pulled the data
#also filtered out duplicate poll found

#further note that samplesize is a double; there are two polls with sample sizes
#ending in .5. I will look into that later.

all_polls <- all_polls %>%
  filter(pollster != "Research 2000") %>%
  filter(pollster != "TCJ Research") %>%
  filter(pollster != "Strategic Vision, LLC") %>%
  filter(pollster != "Pharos Research Group") %>%
  filter(pollster != "Overtime Politics") %>%
  filter(pollster != "People's Pundit Daily") %>%
  filter(pollster != "CSP Polling")
```

The Undecided

One of the first challenges faced in trying to use polls to predict election results is that polls account for undecided voters, who are difficult to measure with regards to the actual election result. Besides, we are trying to estimate a particular candidate’s vote share on election day, which does not take into account those who don’t vote.

With a large number of people identifying as undecided in polls, it would be greatly ineffective to try to use a candidate’s share of poll respondents as an estimator for their share of votes. We would end up systematically underestimating all proportions:

```
partisan_polls <- all_polls %>%
  filter(cand1_name == "Democrat")
#filters to only races between a Republican and a Democrat
#note that in these partisan races, cand1 is the Democrat and cand2 is the Republican

ggplot(partisan_polls) +
      geom_histogram(aes(x=cand1_actual - cand1_pct, y=..density..),
                     fill="blue", alpha=0.5, binwidth=2) +
      geom_histogram(aes(x=cand2_actual - cand2_pct, y=..density..),
                     fill="red", alpha=0.5, binwidth=2)
```



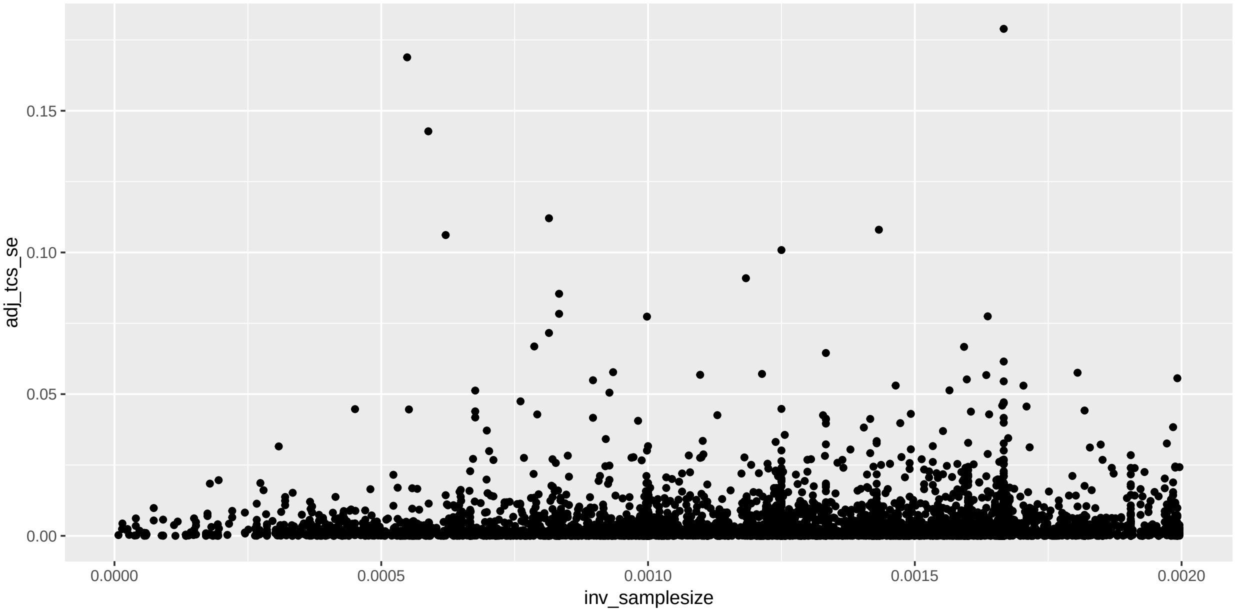
Thus most models attempt to estimate each major candidate's share of the two-party vote. Thus they consider, for each poll, of the number of people who indicated one of the two major candidates what proportion chose each. The other option is to only estimate the margin of victory, and therefore record only the margin in the poll. Both of these are limited in that they are unable to predict third-party performance. These flaws will be explored at greater length later.

For polls with a high number of undecided voters, the share of two-party vote estimator risks overstating the difference between candidates. The disadvantage of the projected victory margin estimator is that we lose access to statistical theory about binomial trials, but I will explore that in detail in the following section.

Binomial Trials and Sample Size

Statisticians have noted that polls tend to be more variable than election results. Let us treat polls as a binomial trial, where the population is the set of people who will vote for one of the two major candidates, and success means voting for the democrat. Theory tells us that the expected value of squared error is given by $\frac{p(1-p)}{n}$. We can test this.

```
polls_se <- partisan_polls %>%
  filter(samplesize > 500) %>%
  mutate(inv_samplesize = 1/samplesize) %>%
  mutate(cand1_tcs_actual = cand1_actual/(cand1_actual + cand2_actual)) %>%
  mutate(tcs_se = (cand1_pct/(cand1_pct + cand2_pct) - cand1_tcs_actual)^2) %>%
  mutate(adj_tcs_se = tcs_se/(cand1_actual/100 * (1-cand1_actual/100)))
# adjusted two-candidate share square error = (two-candidate share error)^2/(p(1-p))
nrow(polls_se)

## [1] 4824

ggplot(polls_se, aes(x=inv_samplesize, y=adj_tcs_se)) + geom_point()
```

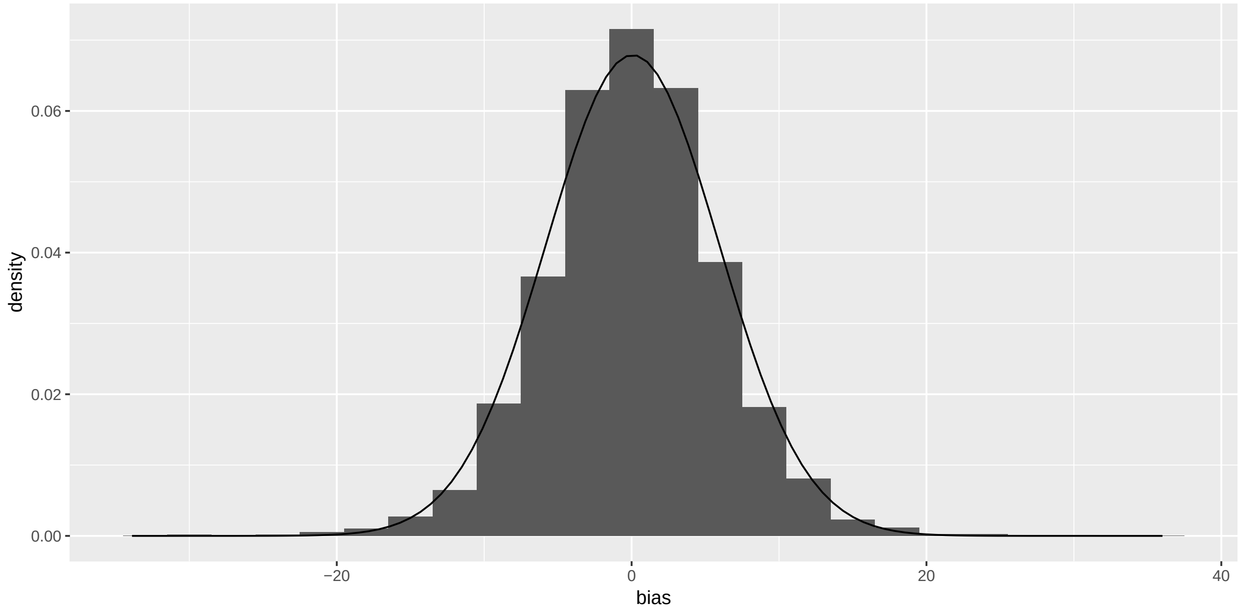


```
lm(tcs_se ~ inv_samplesize, data=polls_se) %>% summary()

##
## Call:
## lm(formula = tcs_se ~ inv_samplesize, data = polls_se)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.00109 -0.00094 -0.00064  0.00012  0.03859
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.20e-04   9.32e-05    9.88   <2e-16 ***
## inv_samplesize 8.61e-02   6.97e-02    1.23     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00206 on 4822 degrees of freedom
## Multiple R-squared:  0.000316,Adjusted R-squared:  0.000109
## F-statistic: 1.52 on 1 and 4822 DF,  p-value: 0.217
```

Fascinatingly, with almost 5000 data points, there is no evidence to suggest that a larger sample size gives you better results (not counting polls with extremely small sample sizes which do suffer problems; in this case the threshhold was set at $n = 500$). Thus comparisons between the expected and actual variance are somewhat missing the point; it isn't that polls are more variable by some additive or multiplicative factor, they just aren't behaving binomially at all. This is a problem for models that assume the data is distributed binomially or multinomially.

Thus, I propose that projected margin of victory is a better (and more intuitive) estimator and target, and will use it from here on. In the dataset I am using, "bias" refers to the difference between the poll-projected Democratic margin of victory and the actual margin. A positive value indicates poll bias towards Democrats, while a negative value indicates poll bias towards Republicans. We note:

```
mu <- mean(polls_se$bias)
sigma <- var(polls_se$bias) %>% sqrt()
ggplot(polls_se, aes(x=bias)) +
  geom_histogram(aes(y=..density..), binwidth = 3) +
  stat_function(fun = function(x) dnorm(x,mean=mu,sigma))
```



```
mu

## [1] 0.06395

sigma

## [1] 5.875
```

This tells us that polls with big enough (greater than 500) sample sizes will give us roughly normally distributed results centered at the true margin of victory with standard deviation of about 6 percentage points. This is independent of sample size.

2016

Now we can look into the idea that polls are to blame for what happened in 2016. Nate Silver has argued that polls were not particularly worse in 2016 than they have been in the past. We can investigate this by looking at a few quantities of interest: mean polling error, which is simply the absolute value of polling bias, mean polling bias, and number of polls conducted.

```
# only Presidential General Elections Polls
pres_polls <- all_polls %>%
  filter(type_simple == "Pres-G")
pres_polls %>%
  group_by(year) %>%
  summarize(merror = mean(error), mbias = mean(bias), polls=n())
```

```
## # A tibble: 5 x 4
##    year merror  mbias polls
##   <int>  <dbl>  <dbl> <int>
## 1  2000   4.55 -2.30    349
## 2  2004   3.22  1.14    366
## 3  2008   3.47  0.102   485
## 4  2012   3.33 -2.36    411
## 5  2016   5.03  3.29    464
```

The values for mean error are very similar to the ones Fivethirtyeight calculated, even though they used a weighting algorithm in which decreased the weight of polls from the most prolific pollsters to combat the risk of individual pollsters having too much leverege on these numbers. Presidential general election polls did indeed have higher error than in previous years, but only by one or two points. The bias, too, was up about one point from 2012. This difference dissappears if we limit ourselves to national polls only:

```
pres_polls %>%
  filter(location == "US") %>%
  group_by(year) %>%
  summarize(merror = mean(error), mbias = mean(bias), polls=n())
```

```
## # A tibble: 5 x 4
##    year merror   mbias polls
##   <int>  <dbl>   <dbl> <int>
## 1  2000   4.02 -3.27      64
## 2  2004   2.16  0.855     60
## 3  2008   2.33  0.0101    77
## 4  2012   3.30 -3.23      79
## 5  2016   2.79  1.82      69
```

In the last 21 days of the election, polls in 2012 were more wrong and more systematically biased than polls in 2016. Thus 2016 suffered more at state-level polling, in places like Pennsylvania, Michigan, and Florida, right? For each Presidential election, I looked at the states where 10 or more polls were conducted in the last 3 weeks of the election and call these the "competitive" races. If we use the average of all polls for a given rate as our estimator, I calculate the mean error across all competitive races, the mean bias, and a sort of mean z-score: how much bias was there relative to how variable the polls were? Values far from 0 indicate that the polls conveyed disproportionate confidence in their prediction for how wrong they ended up being, because of how close they were to each other.

```
pres_polls %>%
  filter(location != "US") %>%
  group_by(race) %>%
  summarize(year = year[1], merror = mean(error), mbias = mean(bias),
            polls=n(), sdbias = sqrt(var(bias))) %>%
  filter(polls >= 10) %>%
  group_by(year) %>%
  summarize(mmerror = mean(merror), mmbias = mean(mbias),
            mzs = mean(mbias/sdbias), competitive = n())
```

```
## # A tibble: 5 x 5
##    year mmerror mmbias     mzs competitive
##   <int>   <dbl>  <dbl>   <dbl>       <int>
## 1  2000    4.74 -2.68  -0.736          10
## 2  2004    3.26  0.509  0.133           12
## 3  2008    3.53 -0.131  0.0144          13
## 4  2012    3.26 -2.53  -0.946           12
## 5  2016    4.75  3.52   0.947           14
```

Here, we see that each competitive state in 2016 was predicted by polls to be about a point further from the truth than in 2012, and on average bias for each state was 1 point greater. Interestingly, because the polls in 2016 were more variable, if we measure how right they were relative to how confident they seemed, we get almost the exact same value as in 2016.

So then why are people saying that 2016 broke everything? I propose a deceptively simple explanation. In the early days of election forecasting, it was proposed that "correctly predicting the winning presidential candidate is ultimately the most important political test of a presidential election forecasting model". However, the field of statistics has moved away from this idea: an ideal model is less about point estimates and more about confidence intervals. As Nate Silver argues, a good model is "Probablistic, not deterministic".

We can see the problem if we look at

```
pres_polls %>%
  filter(location != "US", year==2012 | year==2016) %>%
  group_by(race) %>%
  summarize(year = year[1], mp = mean(margin_poll), ma = margin_actual[1], polls=n(),
            called = !xor(mp > 0, ma > 0)) %>%
  filter(polls >= 10) %>%
  filter(!called)

## # A tibble: 6 x 6
##   race             year     mp    ma polls called
##   <chr>           <int>  <dbl> <dbl> <int> <lgl>
## 1 2012_Pres-G_FL   2012 -0.819  0.87    31 FALSE
## 2 2016_Pres-G_FL   2016  1.10  -1.19    33 FALSE
## 3 2016_Pres-G_MI   2016  4.77  -0.22    19 FALSE
## 4 2016_Pres-G_NC   2016  1.72  -3.66    24 FALSE
## 5 2016_Pres-G_PA   2016  3.74  -0.72    27 FALSE
## 6 2016_Pres-G_WI   2016  5.21  -0.76    13 FALSE
```

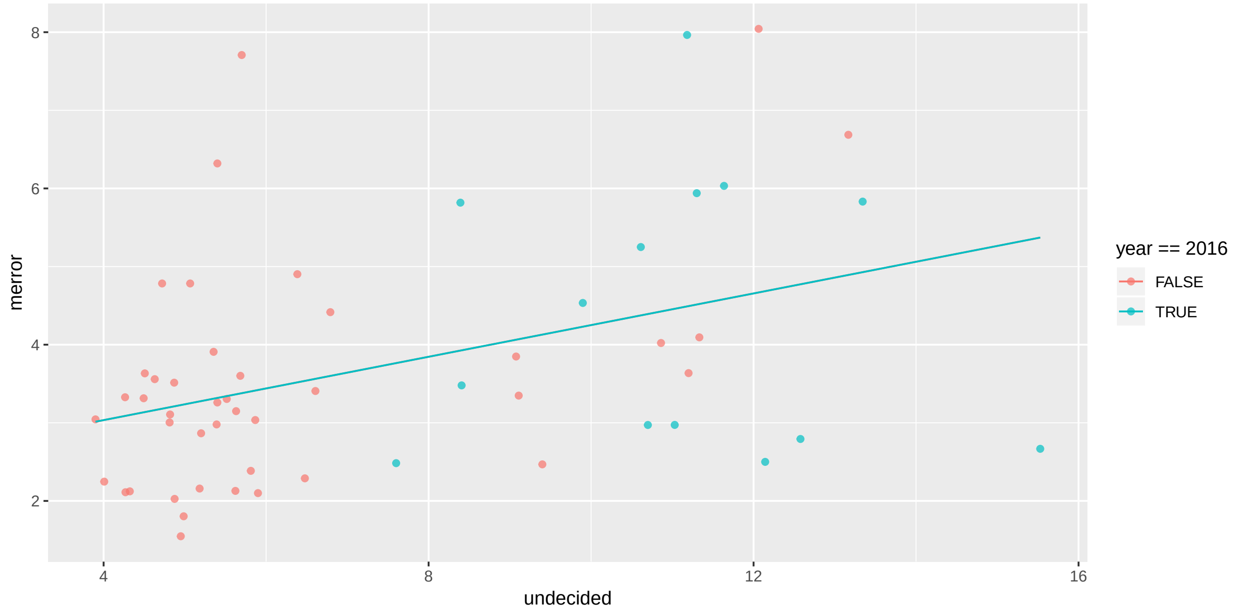
This is the heart of the problem: in 2016, the error hit where it mattered. Five of six races for which the error crosses the 50% threshhold in the past two elections took place in 2016. In 2012, the polls predicted Obama to carry Iowa by 2.30 points. He carried it by 5.81. No one freaked out, but when the error goes in the other direction, turning a 1.1-point lead for Clinton in Floria into a 1.2-point lead for Trump, in a state holding 29 electoral votes, the game changes. The kernel is this: polls were not worse in 2016 than they have been in the past, but statisticians and pundits had not been punished in previous elections for what Silver referred to as a "lack of appreciation for uncertainty," because that uncertainty often manifested in unobstructive ways.

We have established that precedent should have convinced statisticians to not be overconfident in the 2016 election. But the question still remains: why are polls so much more variable than we think they should be? Why are they often biased? And how can we use them better?

The Undecided (cont.)

In defining his election forecasting model, Linzer makes two assumptions. The first is that "Undecided voters are excluded from the analysis because it is not known how they would vote... If undecided voters disproportionately break in favor of one candidate or the other, it will appear as a systematic error... The results I present do not show evidence of such bias." However, that's arguably exactly what happened in 2016: most models which only looked at two-party vote share ignored the unusually high number of third-party and undecided voters responding to many polls. We can investigate this correlation:

```
und_error <- partisan_polls %>%
  filter(samplesize > 500) %>%
  filter(type_simple == "Pres-G") %>%
  group_by(race) %>%
  summarise(undecided = mean(100- cand1_pct - cand2_pct), merror = mean(error), year=year[1], polls=n()
  filter(polls >= 10)

lm(merror ~ undecided, data=und_error) %>% summary()

##
## Call:
## lm(formula = merror ~ undecided, data = und_error)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.704 -1.041 -0.205  0.654  4.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2236     0.5136    4.33  6.5e-05 ***
## undecided     0.2027     0.0636    3.18   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 54 degrees of freedom
## Multiple R-squared:  0.158,Adjusted R-squared:  0.142
## F-statistic: 10.1 on 1 and 54 DF,  p-value: 0.00241

und_error %>%
  ggplot(aes(x=undecided, y=merror, col=year==2016)) + geom_point(alpha=0.7) +
  stat_function(fun = function(x) 2.2236 + 0.2027 * x)
```



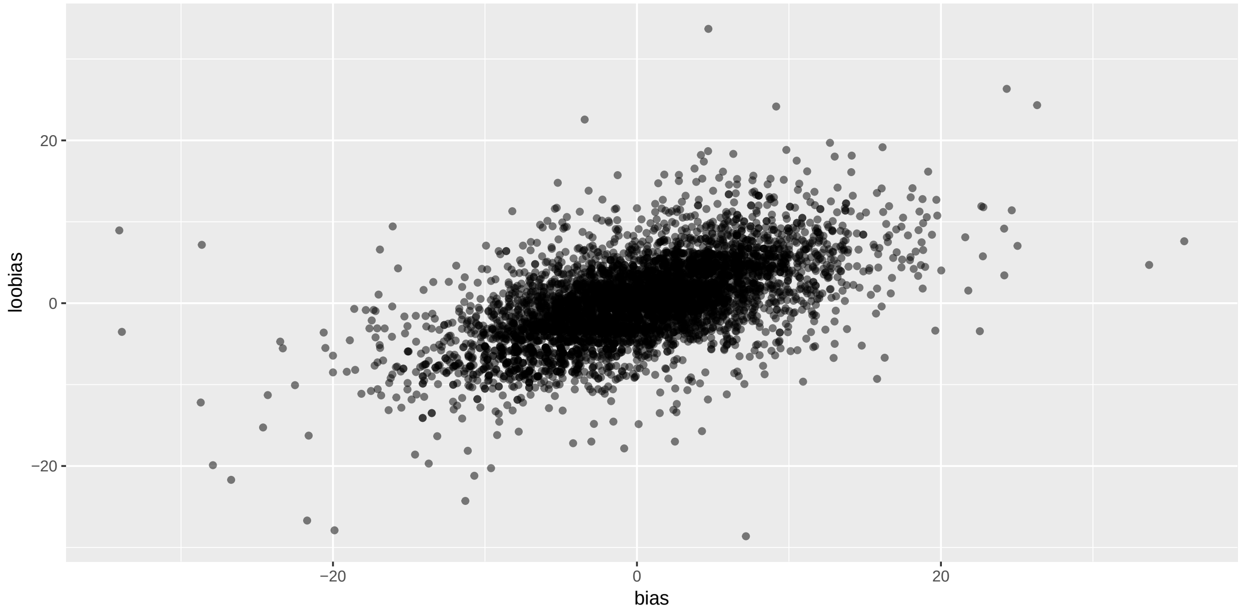
Here we can see that 2016's competitive races featured many more undecided and third-party voters than other elections. Other than Missouri, which has the highest residual (and the highest mean error at roughly 8 points) the 2016 races above the fit line are North Carolina, New Hampshire, Ohio, Michigan, Wisconsin, and Pennsylvania, which are all races of particular interest and many of which won Trump the Presidency. However, a combination between the low R^2, the low level of significance, and that the races of interest all lie above the fit line, we can conclude that taking into account undecided voters cannot explain poll error in 2016 or in previous elections, although it might be one (small) piece of the puzzle.

Linzer's other assumption is that "the bias arising from [house] effects usually cancels out by averaging over multiple concurrent surveys by different pollsters." The term house effects refers to systematic bias an individual pollster might produce, likely due to methodology. We have seen that for a particular election bias does not cancel out, but we can explore that in more depth.

Covariance

I noted earlier that mean error is not the only way to look at poll performance. In particular, it's dangerous when polls are all showing error in the same direction, perhaps exhibiting low variance despite high bias. Here we will consider the covariance structure of the polls: given that a poll is biased in one direction, how likely is it that the average of all other polls for the same race is also biased in the same direction? For us to consider poll results an unbiased estimator for voting intent, we hope to see zero correlation.

```
sbias <- partisan_polls %>%
  group_by(race) %>%
  summarize(sbias = sum(bias), num = n())
corr_polls <- partisan_polls %>%
  merge(sbias, by = "race") %>%
  filter(num > 1) %>%
  mutate(loobias = (sbias - bias)/(num-1))
ggplot(corr_polls, aes(x=bias, y=loobias)) + geom_point(alpha=0.5)
```



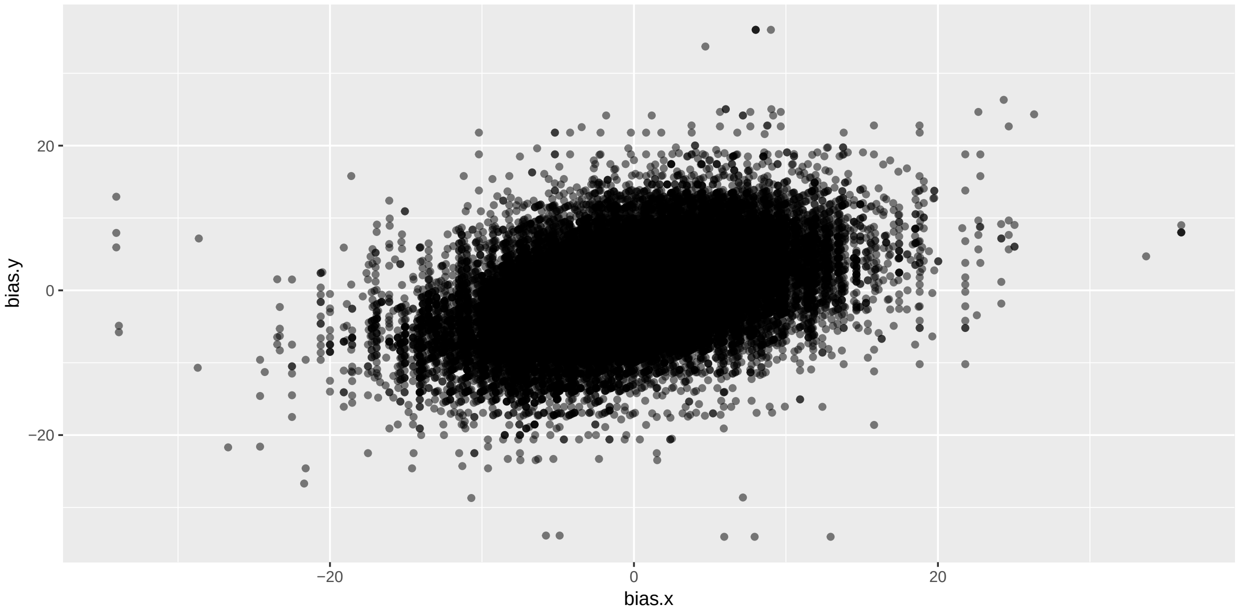
```
lm(loobias ~ bias, data=corr_polls) %>% summary()

## 
## Call:
## lm(formula = loobias ~ bias, data = corr_polls)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -32.10  -2.22  -0.03   2.13  31.39 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.10240    0.05193    1.97    0.049 *  
## bias         0.46986    0.00847   55.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.97 on 5842 degrees of freedom
## Multiple R-squared:  0.345,Adjusted R-squared:  0.345 
## F-statistic: 3.08e+03 on 1 and 5842 DF,  p-value: <2e-16
```

Instead we see a great deal of correlation. The correlation between an individual poll's bias and the "leave one out" bias of the remaining body of polls is 0.59. If we limit ourselves to races with at least 5 polls we get a correlation of 0.66, and the correlation remains the same if we further limit to races with at least 10 polls.

Just to be sure, we can also check the correlation between the bais of any two polls predicting the same race which were conducted by different pollsters, to be sure that this correlation isn't the result of house effects:

```
obias <- corr_polls %>%
  select(race, bias, pollster)
```

```
pcorr_polls <- corr_polls %>%
  merge(obias, by = "race") %>%
  filter(pollster.x != pollster.y)
ggplot(pcorr_polls, aes(x=bias.x, y=bias.y)) + geom_point(alpha=0.5)
```



```
lm(bias.y ~ bias.x, data=pcorr_polls) %>% summary()

##
## Call:
## lm(formula = bias.y ~ bias.x, data = pcorr_polls)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.60  -2.68  -0.08   2.61  32.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.12099    0.01599   -7.57  3.9e-14 ***
## bias.x       0.43740    0.00327  133.72  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.39 on 75588 degrees of freedom
## Multiple R-squared:  0.191,Adjusted R-squared:  0.191
## F-statistic: 1.79e+04 on 1 and 75588 DF,  p-value: <2e-16
```

We see a 0.44 correlation between any two polls from different pollsters predicting the same race. This holds true if we filter down to only races with more than 5 or more than 10 polls, too.

We can run anova, limiting ourselves to races with at least 10 polls, to get a sense of how big this effect is.

```
pollsaov <- corr_polls %>%
  filter(num >= 10)
aov(bias ~ race, data=pollsaov) %>% summary()
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race         176  40806   231.9      17 <2e-16 ***
## Residuals   2898  39493    13.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This demonstrates that just over half of the variance in bias can be explained by an inherent bias factor for each race.

The Polling Gap

Academics have frequently asked why polls are more variable than votes. However, lining up histograms and expected densities does not give us the full image: polls are highly coordinated in their error. There is therefore, perhaps something systematically wrong.

A poll conducted by University of Michigan in 2000 found that two-thirds of respondents did not believe a group of 1,500 or 2,000 respondents could accurately reflect the beliefs of the country. To a statistician, this appears as an opportunity to learn about the central limit theorem. But as I showed above, variance isn't dropping like we would expect it to, so perhaps there's some truth to their belief.

In truth, there is an assumption being made by statistical models that is even bigger than the assumption that undecided voters will break proportionately or house effects cancel. That assumption is that polls act as a draw of election day voter intent, when in fact, they are only a draw of current day claimed voter intent.

Let us therefore define a new population parameter: polling intent, defined as the proportion of registered voters who would indicate a particular candidate if we polled all of them. This definition can be adjusted to refer to "likely voters" or even adults.

Gelman and King find that although voters are generally well informed by election day, poll respondents are often uninformed and not even rational, using a definition of rationality they put forward. Thus I argue that true voting intent and true polling intent are not the same thing. There are a number of obstacles between these two values:

1. Change over time. The polling intent is measured usually on a day other than the election day, and people can change their minds. This point is frequently built into many statistical models.

2. Bias in who responds to polls.

3. Bias in how people respond. For example, the so called "shy Trump" effect, where people are less likely to indicate support for someone they think is losing the race.

   If this were a primary culpit, however, we would observe polls to systematically overstate the winner's margin of victory. However, across all partisan polls, we see a slope of 0.95 between polling margin and actual margin. Perhaps the 5% drop is attributable to this effect

4. who turns out to vote. Polls are drawn from registered voters or "likely voters," but cannot be drawn from voters, because pollsters do not know who will vote on election day. Turnout is a well studied phenomenon, and systematic ideological, religious, or racial changes in turnout can have huge effects on results. Drops in African American turnout for Clinton are considered to have a huge effect on the results of the 2016 presidential election. I find this to be the most cogent explanation for the polling gap, and explains the large correlation in polling bias.

Many models which treat polls as draws from population voting intent attempt to build confidence intervals to demonstrate the level of confidence of the prediction. However, the problem with these confidence intervals is that with enough polling data, the size of the confidence interval will inevitably shrink towards 0, never hitting a minimum.

A lot of convenient mathematical properties don't seem to hold because of the prevalence of bias in the data. Polls weren't particularly bad in 2016; they were just as bad as ever. And you should be careful before trusting them.