House Forecast Model

Harry Bendekgey

July 6, 2018

Introduction

Bafumi, Erikson, and Wlezien propose a model for forecasting US house midterm elections based on data available by early July. In this report I will recreate their methodology to predict the results of the 2018 midterm elections in the house of representatives. My goal is not only to produce forecasts but to bring attention to pitfalls in their methodology and where different choices can be made which would yield different results than the ones I obtained.

Estimating National Vote

The first step in Bafumi et al's model is to predict the national house vote for the upcoming midterm election. This is done using only two predictor variables: the party of the current president, and the average Demcoratic share of generic ballot results between 180 and 121 days before the election.

Both national vote and generic ballot averages are measured in percentage point deviations from an even split; that is, a national vote of "10" corresponds to the Democrats winning 60% of voters who voted for either Democrats or Republican.

Professor Bafumi was extremely generous in sharing the historic generic congressional ballot data with me. But there are a couple of choices that must be made in how the regression is done.

In their 2014 paper, Bafumi et al propose decrementing the results of registered voter polls by 1.42 so that they reflect likely voter populations. Extensive research has been done on the difference between thse two polls, and the 1.42 value seems roughly consistent with what others have found; Fivethirtyeight decrements the margin by 2.7 points, roughly equivalent to a 1.35 decrement in Democratic share.[1]

One concern with this adjustment is that they are regressing on data that goes back as far as 1946. The first likely voter polls appeared in the 21st century. Thus we are adjusting every twentieth century poll, when we don't even have likely voter polls to compare to. In particular, it is worth considering that this margin has changed over time.

It is also concerning that we are regressing on generic ballot averages across this time period, where in early years this corresponds to a single poll and in recent years corresponds to a huge body of work conducted over the 60 day period. The variability of this predictor should therefore be going down across elections and thus the variability of the response.

Let's start by regressing with the 1.42 percentage point adjustment to registered vote and adult polls:
@

```
fit0 <- lm(vote ~ adj_genpoll + president_party, data=model)
coef(summary(fit0))
```

[1] Bafumi et al's model only talks about share of two-party vote, thus ignoring undecided respondents, while Fivethirtyeight measures poll results in the margin between Democratic and Republican respondents, meaning Fivethirtyeight's adjustment should be smaller than Bafumi et al's, proportional to the number of undecided voters encountered.

```
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      -0.5628    0.56600 -0.9944 3.358e-01
## adj_genpoll       0.4983    0.08274  6.0224 2.336e-05
## president_party  -2.1484    0.41095 -5.2278 1.022e-04
```

We note that the intercept is not statistically significant here. We want that to be true, because it supports the claim that polls, on average, are not biased. If we want, we can fix the intercept. We must be careful doing this, though; because the vast majority of polls are not likely voter polls, shifting a huge number of polls a set amount and then insisting there is no bias could cause a lot of problems. We can investigate this 1.42 value for ourselves using regression:

```
shift_reg <- lm(vote ~ 0 + genpoll + president_party + pct_rv, data=model)
coef(summary(shift_reg))

##                 Estimate Std. Error t value  Pr(>|t|)
## genpoll           0.4989    0.08359   5.968 2.576e-05
## president_party  -2.1541    0.41072  -5.245 9.893e-05
## pct_rv           -1.2772    0.66646  -1.916 7.457e-02
```

This suggests that registered voter polls should be shifted by -1.277, although it is not significant at the $p < 0.5$ level. However, with so few data points (only 15, almost all of which have 100% registered voter polls) I believe this value to be important.

If we run regression on the new adjustment:

```
fit1 <- lm(vote ~ adj_genpoll + president_party, data=model)
coef(summary(fit1))

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      -0.6327    0.57407  -1.102 2.878e-01
## adj_genpoll       0.4982    0.08271   6.023 2.334e-05
## president_party  -2.1477    0.41094  -5.226 1.025e-04
```

We get basically the same result but with a different intercept, which makes sense given the ubiquity of registered voter polls. If we force the intercept:

```
fit2 <- lm(vote ~ 0 + adj_genpoll + president_party, data=model)
coef(summary(fit2))

##                 Estimate Std. Error t value  Pr(>|t|)
## adj_genpoll       0.4339    0.05904   7.349 1.639e-06
## president_party  -2.1914    0.41176  -5.322 6.878e-05
```

We get different results, which causes the generic ballot lead to translate at a lower rate into a popular vote lead. If we do no transformation at all of registered voter polls:

```
fit3 <- lm(vote ~ genpoll + president_party, data=model)
coef(summary(fit3))

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       -1.255    0.65055  -1.929 7.294e-02
## genpoll            0.497    0.08245   6.028 2.314e-05
## president_party   -2.142    0.41085  -5.213 1.052e-04
```

Then we get a pretty big intercept which is verging on significant. Forcing the intercept to 0 under the claim that all polls, registered or otherwise, are unbiased, we get:

```
fit4 <- lm(vote ~ 0 + genpoll + president_party, data=model)
coef(summary(fit4))

##                 Estimate Std. Error t value  Pr(>|t|)
## genpoll           0.3729    0.05577   6.686 5.229e-06
## president_party  -2.2277    0.44176  -5.043 1.201e-04
```

```
c((summary(fit1))$adj.r.squared, (summary(fit2))$adj.r.squared,
  (summary(fit3))$adj.r.squared, (summary(fit4))$adj.r.squared)

## [1] 0.8081 0.8315 0.8083 0.8057
```

They all have about the same adjusted R^2 except for the 2nd model. This makes sense if we believe likely voter polls to be unbiased, because we regressed to find the shift size that results in registered voter polls that reflect likely voter populations, then fixed the intercept to claim the resulting polls we unbiased.

Now we calculate the mean Democratic share on the congressional ballot in 2018. Fivethirtyeight's dataset is online, and they use a much more sophisticated weighting and adjustment process to account for house effects. In the age of robotic polls this is becoming incredibly important. A few polling houses can dominate the averages by pumping out dozens of polls, all of which may be biased in the same way.

```
genpoll2018

## [1] 3.976
```

Thus we claim that Democrats have on average 54% of generic ballot respondents' support. With this average, we create a prediction interval for 2018's democratic national vote share:

```
interval <- predict.lm(fit2, params18, interval = "prediction")
interval

##     fit    lwr   upr
## 1 3.917 0.1074 7.726
```

Thus I predict that Democrats will win 53.92% of the popular vote. Further, I am 95% confident that the Democrats will win between 50.11% and 57.73% of the vote.

At this point it's worth noting how important the decisions we made before were. Consider using one of the other models:

```
c(interval1[1], interval[1], interval3[1], interval4[1])

## [1] 3.496 3.917 2.863 3.710
```

These are the point estimates for the national vote based on the four models used. It's worth acknowledging that the one I chose is the most friendly to Democrats, and that these two choices outlined could result in a different prediction of 1% of all voters in the country, a large margin that will have a significant effect on predictions.

```
se

## [1] 1.787
```

Mathematically, I am saying that the democratic share of the two-party vote, measured in percentage points away from 50, is distributed t(3.92, 1.79, df=15). This gives us our prediction interval that we are 95% confident the value will be between 0.11 and 7.73. This means I am all but certain the Democrats will win the national house vote.

The 2016 popular vote was:

```
dem_share16

## [1] -0.559
```

Thus if we measure 2016-2018 national vote swing, we get points estimate and prediction interval:

```
interval - dem_share16

##     fit    lwr   upr
## 1 4.476 0.6664 8.285
```

Mathematically, we say that the swing in national democratic share of two-party vote is distributed t(4.476, 1.79, df=15) with the prediction interval shown above.

Mean District Swing

An important question to address at this point is why we care about the national vote. In order to model the covariance of house races, Bafumi et al simulate the election by first picking a value for the national swing from the distribution found above. Then, they define district-by-district predictions such that the mean district is shifted from the previous election results by that national swing. Then each district's uncertainty is simulated. This is done repeatedly, simulating thousands of elections. The proportion of these elections in which a certain event occurs is taken to be the probability of that event occuring.

They key fact here is that the parameter of interest is not the swing of national vote from 2016-2018, but the swing of mean district vote in contested districts. The national vote and the mean contested district vote differ in two important ways.

The first way is outlined in Bafumi et al's 2014 paper. If there is a negative correlation between percentage democratic vote in a district and the total number of people that vote, we would expect that summing all the votes in those districts and calculating percentage democratic vote would look worse for democrats than just taking the mean democratic share across those districts, because the latter weighs all districts evenly.

Because we are only interested in mean district vote swing, we don't necessarily care about the size of this discrepency. We only care if we are reason to believe its size will change election-to-election. Bafumi et al argue that it does, that in midterm election years Democrats are particularly bad at turnout and the Democratic mean district advantage grows in size. To estimate this value, they use a single data point: 2008-2010, and use that to value in their prediction for 2012-2014.

They estimate this discrepency by comparing the mean Democratic district share in 2008 to the share of total votes in those districts, and then doing the same for 2010 to show that the discrepency grew. This is concerning, however, because they find the change in discrepency size for a single transition between elections, and use that value without any uncertainty attached to it. Let's try to recreate what they did:

```
c(
  mdist12-sumcont12,
  mdist14-sumcont14,
```

```
  mdist16-sumcont16
)

## [1] 0.4940 0.5059 1.1582
```

Recall that a large value indicates a stronger correlation between how blue a district is and how few people vote. In this case, contrary to Bafumi et al's findings, 2016 was the worst year for Democratic turnout, and 2012 and 2014 were not substantially different. In fact, 2014 is the year where this effect is the smallest.

But there's an even larger problem with this method: we are investigating the wrong parameters. Bafumi et al's regression was run on national vote, not sum of vote in contested districts. This is important, because not all districts are contested at all. Some have only one candidate running, and some are only contested by third party candidates, meaning that only one of the two major parties is represented in the race. This is especially true of states like Louisiana or California, where a party can be locked out of the general election. Let's take a look:

```
c(nrow(share12), nrow(share14), nrow(share16))

## [1] 389 357 371
```

Of the 435 seats in the house of representatives, a large portion of them, 10-20% of them are not being contested depending on the year. Let's see what happens when we compare these discrepencies:

```
c(
  mdist12-popvote12,
  mdist14-popvote14,
  mdist16-popvote16
)

## [1]  1.03933  0.22008 -0.01597
```

This is behaving highly unpredictably. In 2016 the mean district vote was almost exactly the popular vote. The reason why, in the past two elections, this discrepency is smaller than the previous one because there were more districts uncontested by the Republican party than districts uncontested by the Democratic party.

If we look at the mean district vote swing versus the national vote swing for the last midterm election in 2014:

```
mswing <- mdist14 - mdist12
pswing <- popvote14 - popvote12
mswing

## [1] -4.364

pswing

## [1] -3.545
```

The final problem with this estimator is that the model looks at districts that are contested in both last election and the upcoming one, and adjusts the model such that the mean result for those districts will shift by the mean district swing. The difference between the mean share in 2014 and 2016 for races that are contested in both elections is a different value than the difference between the mean share for races contested in 2014 and the mean share for races contested in 2016.

The reason for this is that if a race is contested in one election and not the other, it is likely because it is so partisan as to potentially not warrant a candidate. Thus, in the election where this district is included, it will have large sway on the mean district vote. To believe that this doesn't have an effect on the mean district swing is to assume that the amount of districts that are uncontested on each side of the aisle remains constant year to year, which is untrue. We can see that between 2012 and 2014:

```
(share12 %>% filter(!(district %in% share14$district)))$dem_share %>% mean()

## [1] 7.042

(share14 %>% filter(!(district %in% share12$district)))$dem_share %>% mean()

## [1] -1.098
```

A lot of very blue seats in 2012 were not contested in 2014, which messes up the estimate for mean district swing. If we only consider districts contested in both elections:

```
nrow(pred14)

## [1] 328

mswing <- mean(pred14$dem_share14) - mean(pred14$dem_share12)
mswing

## [1] -3.503
```

We see that these 328 districts shifted by an average of -3.50 points. Ultimately, this is the value we are interested in, and it is incredibly close to the national vote swing of -3.55.

Because the national vote swing tracked this swing so well in 2014, and because 2016's mean district vote is so close to the national vote. I will ignore these effects, and treat the national vote swing as if it were the mean district vote swing. It is highly possible, perhaps likely, that a drop in turnout will benefit the Democrats mean district vote compared to the national vote in 2018. However, the districts I've discovered are uncontested in 2018 but were contested in 2016 are California 5, 6, 8, 13, 19, and 20, all of which are held by Democrats. This will hurt them in that metric. Moving forward, I am ignoring both of these effects.

Finding Forecast Error

There is one last step before we run our model on 2018. Bafumi et al propose only dividing races up into those with incumbents and those with open seats. Any race that was uncontested in the previous election is conceded immediately. They propose the following models.

For open seats:

$$\text{DemVote\%2014}_k = \beta_0 + 0.95\text{Obama\%2012}_k + u_k$$

For contested seats:

$$\text{DemVote\%2014}_k = \beta_0 + 0.63\text{DemVote\%2012}_k + 0.46\text{Obama\%2012}_k + 2.03\text{frosh}_k + u_k$$

Where frosh is a dummy variable set to 1 if the candidate is a freshman Democrat and -1 if they are a freshman Republican, to simulate incumbency advantage. The intercepts are set so that the mean open district vote in 2014 is shifted from the 2012 vote in those seats according to the national vote swing from 2012, and the mean incumbent district vote in 2014 is shifted from the 2012 vote in those seats according to the national vote swing from 2012.

We want estimates for u_k. Namely, we want to know how much we expect individual races to vary from what we expect from how they've voted in the past combined with national swing. To do this, we run this model on 2014 data.

```
nrow(open14)

## [1] 42
```

We have 42 open seats. They vary from our prediction of them (using the template above):

```
openvar <- var(open14$dem_share14 - open14$pred) %>% sqrt()
openvar

## [1] 6.133
```

Thus we say that districts with open seats will vary with a standard deviation of 6.1 percentage points from their expected center given voting history and national swing.

What happens if we make the linear model ourselves, using this data? The template above is calculated using 2008's prediction of 2010. But we want to use 2012's prediction of 2014 and apply that to 2018:

```
sum_open <- lm(dem_share14 ~ Oshare, data=open14) %>% summary()
sum_open

##
## Call:
## lm(formula = dem_share14 ~ Oshare, data = open14)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.597  -3.076  -0.902   3.283  17.403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.5072     0.9659    -2.6    0.013 *
## Oshare        0.9231     0.0769    12.0  7.6e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.2 on 40 degrees of freedom
## Multiple R-squared:  0.783,Adjusted R-squared:  0.777
## F-statistic:  144 on 1 and 40 DF,  p-value: 7.62e-15
```

This is very close to their findings.

Now let's take a look at seats with incumbents:

```
nrow(inc14)

## [1] 286
```

Now we're looking at the remaining 286 seats. They vary from our prediction (using the incumbent template):

```
incvar <- var(inc14$dem_share14 - inc14$pred) %>% sqrt()
incvar

## [1] 4.053
```

Thus we say that districts with incumbents will vary with a standard deviation of 4.1 percentage points from their expected center given voting history and national swing. It is expected that seats with incumbents should vary less from our expectations than open seats.

```
sum_inc <- lm(dem_share14 ~ Oshare + frosh + dem_share12, data=inc14) %>% summary()
sum_inc
```

```
##
## Call:
## lm(formula = dem_share14 ~ Oshare + frosh + dem_share12, data = inc14)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.490  -2.336   0.221   2.326  15.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.8196     0.2242  -17.04  < 2e-16 ***
## Oshare        0.1282     0.0474    2.70   0.0073 **
## frosh         2.2576     0.4795    4.71  3.9e-06 ***
## dem_share12   0.8726     0.0419   20.84  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.73 on 282 degrees of freedom
## Multiple R-squared:  0.948,Adjusted R-squared:  0.947
## F-statistic: 1.71e+03 on 3 and 282 DF,  p-value: <2e-16
```

This is interesting. If we run our own linear model on 2014's prediction from 2012, we see that a lot more weight it put on how the incumbent preformed in the last election(0.87 in my model vs 0.63 in Bafumi et al's), and much less weight is put on how it voted for President (0.13 in my model compared to 0.46 in Bafumi et al's). These are highly correlated, so it is unsurprising that this would shift year-to-year. Picking which model to use in 2018 is largely ideological: how much, in your opinion, is the GOP the party of Trump? To what degree is it still the party of the incumbents?

Finally, it's worth noting that the model assumes both open seats and incumbent seats shift by the mean district shift. We can see that isn't true:

```
c(mean(open14$dem_share12) + mswing, mean(open14$dem_share14))
```

```
## [1] -1.742 -0.902
```

We overestimate the mean district swing in open seats, and...

```
c(mean(inc14$dem_share12) + mswing, mean(inc14$dem_share14))
```

```
## [1] -3.033 -3.156
```

...underestimate the mean district swing in incumbent seats. One reason for this might be incumbency advantage. Because a majority of the vacated seats were vacated by Republicans, the loss of incumbency advantage caused the open seats to shift by less in the Republican's favor.

Predicting 2018

Now we move on to predict the 2018 election results with what we have. We have two

We start by dividing the races into those which are conceded to democrats, those which are conceded to Republicans, and those which are contested in both 2016 and 2018, which we further divide into open seats and incumbent seats:

```
c(Dconcede,
  Rconcede,
  nrow(open18),
  nrow(inc18))

## [1]  41  27  68 299
```

We see that there are 41 races conceded to Democrats, 27 races conceded to Republicans, 68 open seats and 299 contested incumbent races.

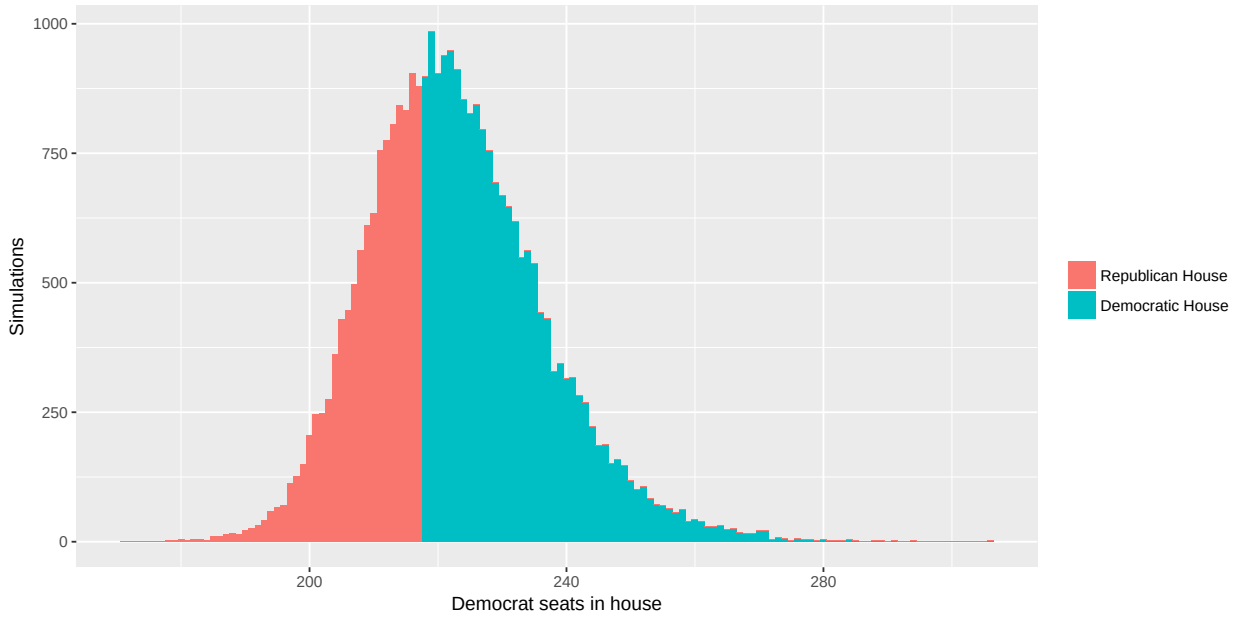
```
dvacate <- open18 %>% filter(incumbent16 == 1) %>% nrow()
rvacate <- open18 %>% filter(incumbent16 == -1) %>% nrow()
dvacate

## [1] 16

rvacate

## [1] 43
```

We note that the open seats overwhelmingly were vacated by Republicans. Thus we probably underestimate Democratic performance in these seats and thus slightly overestimate Democratic performance in incumbent seats.

We run the model using Bafumi et al's template.

We get the following distribution of seats:

```
ggplot(aes(x=dseats, fill=(dseats > 217)), data=posterior) +
  geom_histogram(binwidth=1) + xlab("Democrat seats in house") +
  ylab("Simulations") + scale_fill_discrete(name=element_blank(),
                        breaks=c("FALSE", "TRUE"),
                        labels=c("Republican House", "Democratic House"))
```



We calculate the probability of a blue house:

```
sum(posterior$dseats > 217) / nrow(posterior)

## [1] 0.6295
```

About a two-thirds chance of Democrats winning the house in November!

```
c(safeD, solidD, likelyD, leanD, tossup, leanR, likelyR, solidR, safeR)

## [1] 168  19  11   9  14  19  39  46 110
```

We can also see the distribution of how "safe" seats are according to this forecast. I've written the point estimates and win percentages to a csv file which can be viewed independently.

Choices

A final important part of this forecast, which I have investigated at great length above, is the number of arbitrary decisions one has to make in dealing with data. This is an open-source project. Anyone can edit this, and if you do, I encourage you to play around with some or all of the following components:

- Is the estimated shift of 1.42 for registered voter polls as compared to likely voter polls appropriate? Is it consistent for all elections since World War 2? Should we be using a more sophisitcated shifting mechanism?
- Would the results of the initial regression change if you included poll respondents who "leaned towards the Democrats" as a Democratic respondent, and those who "leaned towards the Republicans" as a Republican respondent?
- If an incumbent runs but loses renomination, I count that as an open seat, not considering incumbency advantage. Is that appropriate? Can a more sophisticated mechanism be used?
- I spent a good deal of time investigating how turnout affects mean district swing. How can you estimate the size of this effect in 2018?

- How would you estimate how the effect on mean district swing caused by the change in which districts are contested between 2016 and 2018, as explored above?
- Currently we concede all races in 2018 that went uncontested in 2016. This poses a potential danger. For example, PA 18 was not contested by Democrats in 2016, but Democrats actually won it in the early 2018 special election (although due to redistricting that district no longer exists). What is a more sophisticated model we could use, given that we don't have a precedent for how the district votes?
- To what degree do you trust the way a district voted for President last cycle and how they voted for representative, given an incumbent is in power? Which of the two templates (Bafumi et al's 2010 one, or my 2014 one) do you trust more?
- Given the investigation above, open seats and incumbent seats might not shift by the same amount, based on who is giving up incumbency advantage. Can you quantify how off our underestimates of open seats and overestimates of incumbent seats would be, to decrease bias?
- Count Pennsylvanian incumbents as incumbents? Currently I'm not because of redistricting.

TODO

- Talk about CA 30, CA 44, IA 03, LA 03, and OH 16 in 2012 (multiple incumbents)