Moses Glickman / mag23
Elliot Reisman-Tremonte / err6
STAT 410
16 Nov 2020

Who's John Boozman? Examining Determinants of U.S. Senatorial Public Prominence

Introduction

The Senate, the upper house of the United States' bicameral federal legislature, has 100 members, apportioned two to a state and elected to staggered six-year terms. Some members of the Senate, such as Vice President-elect Kamala Harris (D-CA,) Bernie Sanders (I/D-VT,) and Mitch McConnell (R-KY,) are both widely recognized and discussed on the national stage. Other members of the Senate are much less prominent, creating a gap in public stature that exists despite the equal weighting of each senator's vote. We set out to identify factors that explain variation in public prominence, using the 2-year Google Trends scores of each Senate member, which measure the volume of content and recent searches that relate to each senator, as a proxy for the otherwise difficult-to-quantify factor of public prominence.

Data Collection

Among the factors that we surveyed were the competitiveness of their seat, the partisanship of their state relative to their nation, the partisanship of their voting record in office, the difference between their own partisanship and their state's partisanship, their time in office, whether they were up for reelection in 2020, whether they had run for higher office (i.e., the presidency or vice-presidency) since 2012, the population of their state, whether they chaired or served as ranking member of a large Senate committee, whether they were the leader of their caucus, and, of course, their party affiliation (or, in the case of nominally independent Senators Angus King (I/D-ME) and Bernie Sanders (I/D-VT,) their caucus affiliation.) Many of these factors were chosen based on commonalities we noted among currently prominent senators.

Due to the difficulty in directly quantifying such factors as voting record partisanship and state partnership, we used various proxies. We also had to generate estimates for some variables for some Senators due to quirks in their voting history. Each regressor, coupled with an

explanation of our sources for that regressor and any estimates we made in computing that portion of the dataset, is listed below in the interests of transparency.

Because two senators, Angus King of Maine and Bernie Sanders of Vermont, are not members of the Democratic Party but nonetheless caucus with the Democrats, we shied away from direct reliance on party affiliation and relied instead on caucus affiliation, classing all Democratic Senators and Democratic-caucusing independents as "D" and all Republican Senators as "R." (There are currently no Republican-caucusing independent Senators, and every current Senator caucuses with either the Republicans or the Democrats.)

To estimate state-level partisanship, we relied on the Cook Political Report's Partisan Voting Index, an index that averages the two previous presidential election returns in each state and congressional district to produce analyses of how partisan, on average, each state and district are relative to the nation. A score of R+2, for example, indicates that in a 50/50 national environment, a given state can be expected, on average, to vote for the Republican candidate by 2 percentage points, and can be expected, on average, to be a true toss-up in a 51/49 national environment that leans toward the Democratic candidate by 2 percentage points. As 2020 ballots have not finished being counted and new PVI scores have not been calculated, the scores we use are generated from the 2016 and 2012 presidential elections. Although probably unavoidable, given the natural unreliability of incomplete returns, it is worth noting that this presents a slight source of error, as it is possible that states that only moved to the right under Trump, such as Pennsylvania, Wisconsin, and Michigan, have current PVI scores that bias their current partisanship toward the Democrats through the incorporation of 2012 results (and vice versa for states like California, Texas, and Arizona that have become more Democratic since 2016.)

To estimate voting record partisanship, we used FiveThirtyEight's "Trump Score." The Trump administration frequently issues statements in support of, or against, a bill, nomination, or resolution put before the Senate. The Trump Score calculates the percentage of the time a Senator's vote is aligned with the stated preferences of the Trump administration. These scores run from 12.1% (Kirsten Gillibrand, D-NY,) to 100% (Kelly Loeffler, R-GA.)

To estimate state population, we used US Census Bureau 2019 population estimates, as full 2020 census data is not yet available.

We coded retiring senators whose seats were up for election in 2020 as "up in 2020," even though they were not running, because we determined that this would be the best way to

account for the increased buzz around senators such as Lamar Alexander (R-TN) and Pat Roberts (R-KS) generated by their retirement announcements, the efforts of primary candidates to obtain their endorsements, the heightened media interest in their (newly unpressured) voting records during their lame-duck periods following their retirement announcements, and other such factors. Although we believed that this coding would be error-minimizing, it is still a potential source of error, as it is far from clear that retirement buzz generates the same degree of public prominence as an active re-election campaign. Although some safe-seat senators faced only nominal opposition, and Tom Cotton (R-AR) faced no major-party opposition after his Democratic challenger dropped out after the filing deadline, we coded these unopposed or scarcely opposed Senators as "up in 2020" as well, in order to avoid making the inherently subjective determination of what constitutes an "active" re-election campaign.

We coded all Senators who had filed for President as having sought higher office irrespective of their fate in the primary. As such, Lindsey Graham (R-SC,) Michael Bennet (D-CO,) Elizabeth Warren (D-MA,) and several others were coded as having sought higher office, although, having lost in the primary, none ran in the general election. Only three current senators, Tim Kaine (D-VA,) Mitt Romney (R-UT,) and (of course) Vice President-elect Kamala Harris (D-CA,) have run in the general election for the post of President or Vice President.

To determine whether a given Senator currently serves as the chair or ranking member of a large Senate committee, we used a committee listing from the senate.gov site. To avoid equalizing between, for example, the 22-member Senate Foreign Relations Committee and the five-member Joint Committee on the Library, we only coded the chairs and ranking members of committees with 15 or more members as "Committee Chair or Ranking Member."

We created a binary "PVI Gap" variable to identify senators whose caucus affiliation differs from their state's partisanship- Democratic senators from red states, for example, or vice versa.

We also created a binary "Caucus Leader" variable to code the leaders of the Senate Democratic and Republican Caucuses, Mitch McConnell (R-KY) and Chuck Schumer (D-NY.)

We used CNN election returns from 2018, 2016, and 2014 to determine most recent election margins, as 2020 results are not yet finalized. These margins were calculated by taking their share of the popular vote and subtracting the share of the popular vote of their closest competitor- a Senator who won 52-48, for example, would have a margin of 4. For Senators

most recently elected in special elections, such as Doug Jones (D-AL,) we used their special election margins, also calculated from CNN data. Martha McSally (R-AZ) and Kelly Loeffler (R-GA) were appointed by the governors of their states after the retirement of their predecessors and therefore had no margins of victory. Because McSally had run statewide in the past, losing a 2018 Arizona US Senate election to now-Sen. Kyrsten Sinema (D-AZ) by 2.4 percentage points, we assigned the negative of Sinema's margin of victory (-2.4) to McSally. Loeffler had never previously run in a statewide election. In the 2020 election, she came in second place in a "jungle primary" with dozens of Democratic and Republican candidates and progressed to an as-yet lightly polled January runoff election against the first place finisher, Democratic Rev. Raphael Warnock, to retain her Senate seat. We used a post-election Remington Research poll that put her 1 percentage point ahead of Warnock, the only post-election poll conducted in that race as of Nov. 16, as a proxy for her victory margin.

To determine years in office, we measured total longevity of service. As no current member of the Senate has left and returned to the Senate- i.e., all current members' times of service have been continuous- we subtracted the day they took office from the present day (Nov. 16) and rounded down. (As such, all senators elected in the 2018 election who took office in Jan 2019 are said to have 1 year of service, all elected in 2014 who took office in Jan 2015 are said to have 5 years of service, and so on.)

To obtain data for the public prominence of each senator, we looked at data provided by Google Trends. Since some senators have only held office since the beginning of 2019, we looked at each senator's Google searches from January 20th, 2019 until November 15th, 2020. Unfortunately, Trends does not provide a standardized score for every search topic over a given period of time. Instead, Trends allows for direct comparisons between two or more search terms. By standardizing Doug Jones (D-AL) as an 8, we were able to proportionally allocate scores of each senate member. For example, a direct Trends comparison of Doug Jones and Dianne Feinstein (D-CA) gave Jones a score of 4 and Feinstein 8. Because each score was relative to the Jones 8, Feinstein received a prominence score of 16. Trends' presentation of data only to the nearest whole number may have introduced a possible source of error, but no further precision was available for the Trends data and no equally reliable analogue was available to measure search interest. No cross-senator alterations in the Trends methodology were necessary, with the exception of Sen. Michael Bennet (D-CO). Bennet's Trends level was difficult to obtain, as

Bennet's results were intermingled with spillover results from Google searches of Seattle Seahawks player Michael Bennett. As a result, Google did not allow us to evaluate Trends data for the US Senator. Instead, we used the search term "michael bennet" rather than the search topic "Michael Bennet," giving Bennet a derived Trends score of 7. The timeline of the Google Trends data we evaluated, from the date of the 2019 inauguration of the senators elected in 2018 to the present day, presents a possible source of error as well for one data point, as one senator, Kelly Loeffler (R-GA,) has assumed office since that date following the retirement of her predecessor.

Data Analysis

After generating the dataset, we checked for high-leverage observations. Due to the low number of observations, we discovered that the conventional 2-times-the-average test for high-leverage observations flagged 8 observations- 8% of our dataset. We determined that a higher 3-times-the-average threshold was more reasonable. This filtered out only one entry- Susan Collins (R-ME,) likely because of her idiosyncratic voting record, long time in office, and enormous most recent margin of victory compared to her state's partisanship. (Maine is 3 points more Democratic than the nation as a whole, as per its PVI, and Collins won re-election there in 2014 by 37 percentage points.)

Using our newly filtered dataset and all regressors, we discovered that our AICc value was minimized through the incorporation of only two regressors: whether a Senator had run for president or vice president since 2012 ('PVP2012'), and whether a Senator was a caucus leader- i.e., if they chaired either the Senate Democratic or Republican caucuses ('Cleader'.) Although the coefficients of both regressors were positive, the p-value of our caucus leader variable was only 0.129, meaning that we cannot reject the null hypothesis at the standard alpha=0.05 level that state population has no effect on the public prominence of individual senators. Running for president or vice president, meanwhile, had a p-value of 3.73e-08, permitting us to conclusively reject the null hypothesis that running for president or vice president has no effect on a senator's public prominence at the 0.05 level.

Although, as mentioned, the inclusion of additional regressors brought down our AICc value, we still examined the coefficients of each. Being up for election in 2020 ('Up2020',) representing a state with the opposite partisanship ('PVIGap',) having an additional year of

tenure ('YrsServed',) and chairing or serving as ranking member of a large Senate committee ('ComChairRM') each had p-values higher than 0.44 when regressed on Trends data alongside PVP2012 and Cleader, indicating that each had relatively low predictive value.

Ultimately, when added to the regression of PVP2012 and Cleader on Trends, only one regressor was significant at the 0.05 level, and only just: PVI (p=0.0498.) All else held equal, a senator from one state could be expected to, on average, have a Google Trends score 1.38 points higher than a senator from another state 1 percentage point more Republican. We can then reject the null hypothesis at 0.05 and conclude that the more Republican a state leans, the lower the public prominence, on average, of its senators. However, as we will note in Diagnostics, the significance of this particular regressor is dependent on the inclusion of two outliers within the regression.

Diagnostics

The first diagnostic test we provide plots the fitted values of the model that uses the regressors of 'PVP2012' and 'Cleader' against the residuals, indicating that this data does not closely follow a normal distribution. The second diagnostic plot, a Q-Q, demonstrates that the residuals of the model are not normally distributed; the Normal Q-Q points do not generally fall near the line $y = x$. Again, having non-normally distributed residuals poses a threat to the linear regression assumption of equal variance.

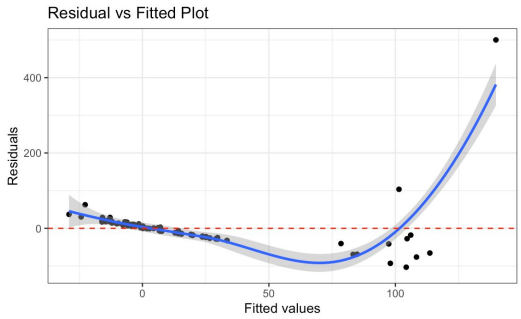
Scale-Location plots help to evaluate the homoscedasticity of a dataset. In this particular Scale-Location plot, as the fitted values increase, the variance of the residuals increases as well. Thus, these residuals are not spread independently of the fitted values or actual Trends data.

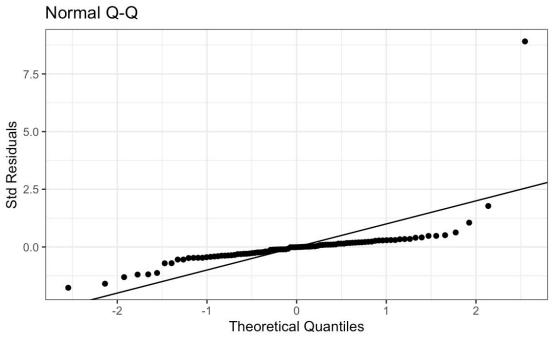
The clear outlier on the Residual vs. Leverage Plot is Sen. Bernie Sanders (I/D-VT). Although not high-leverage points, Sanders, and Vice President-elect Kamala Harris (D-CA,) remain extreme outliers; Harris and Sanders' Trends scores, in fact, are higher when put together than those of the other 98 Senators combined. When Harris and Sanders are excluded from the dataset, 'Cleader' becomes significant at alpha=0.05, and PVI loses its significance altogether, indicative of these two observations' heavy effect on the data. As shown in plots 7-8, although some derivations from normality still occur, the errors, with their exclusion, are more in line with the coalescense of the regression to two binomial variables (in a dataset with a naturally high rate

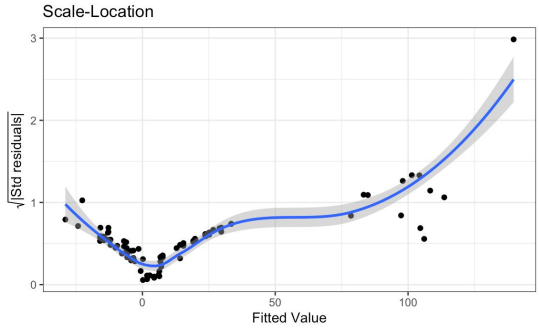
of variation.) In sum, the diagnostics indicate a much lessened threat to our normality assumptions.

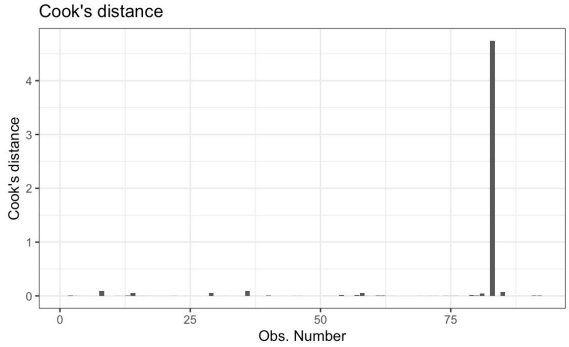
As shown in Table 1 and Table 2, our adjusted R2 rises from 0.2793 to 0.4769 with the exclusion of Harris and Sanders. As shown in Table 3, the inclusion of six additional regressors- the six next most predictive regressors (being up for re-election, state partisanship, caucus affiliation, gaps between party and state partisanship, voting record, and the log of state population)- into the sans-Harris/Sanders model only edges multiple R^2 up from 0.4769 to 0.4944.

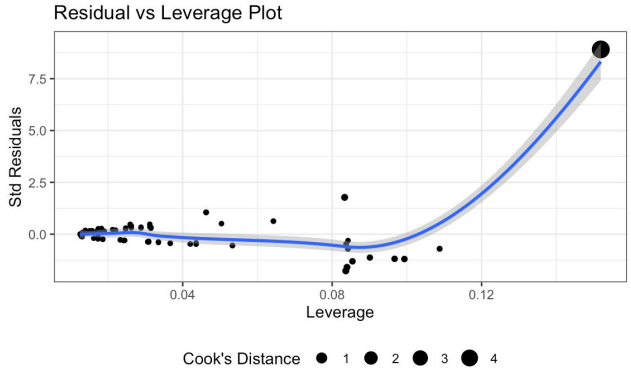
Plots 1-6: Sanders/Harris Plots

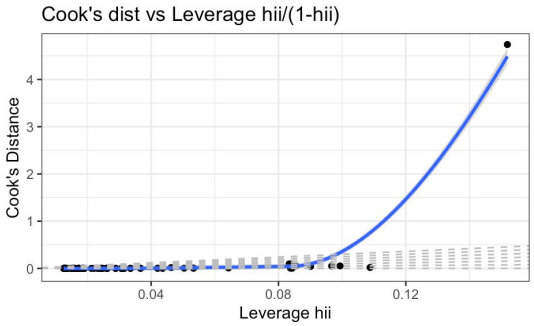




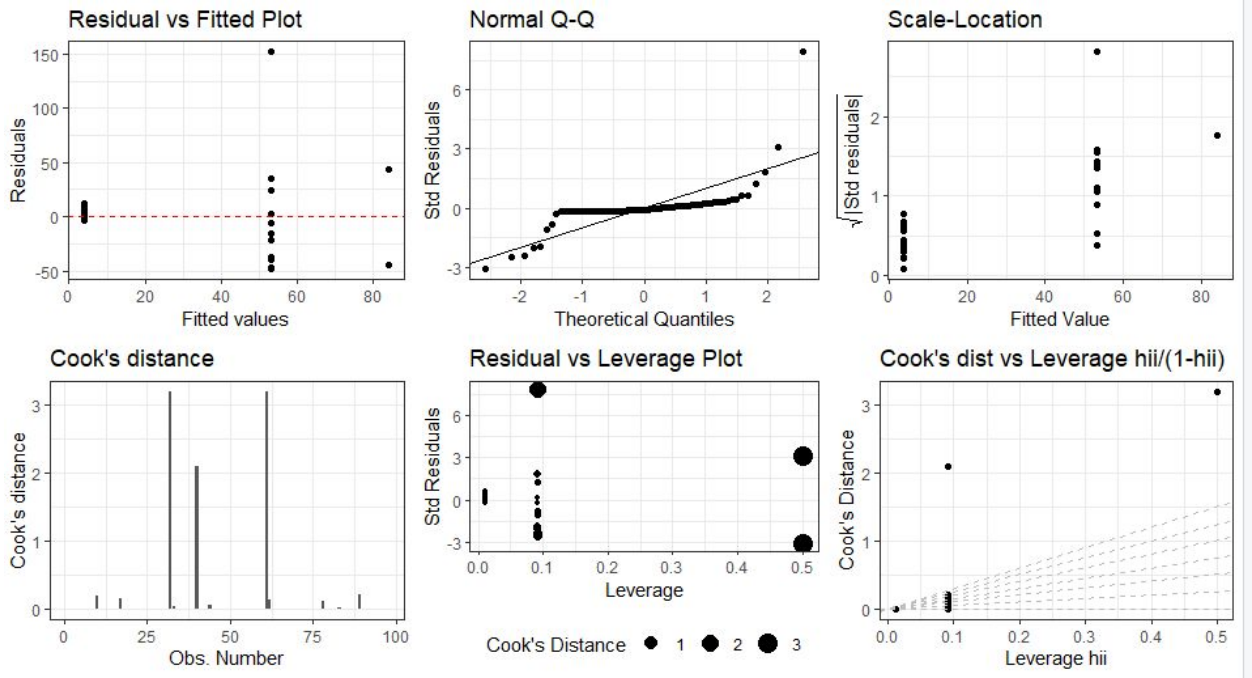








Plot 7,8: Sans-Sanders/Harris Plots



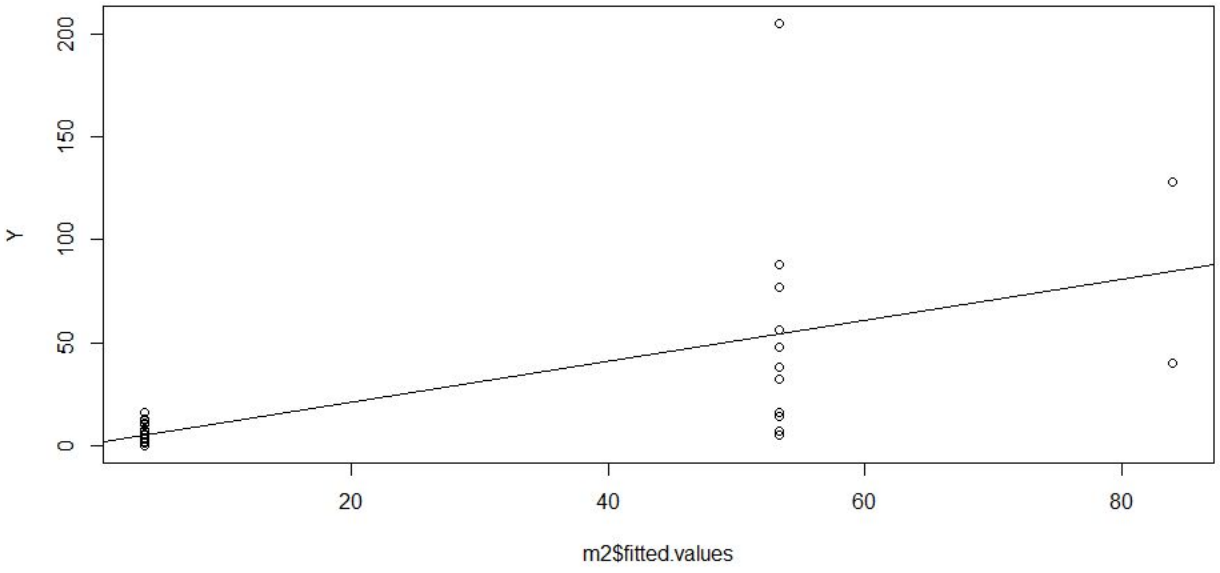


Table 1: Coefficients and P-Values of AIC-minimizing regression, incl. Sanders/Harris

```
Residuals:
    Min      1Q  Median      3Q     Max
-128.69   -3.07   -2.07    1.93  506.31

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.071      7.856   0.518    0.606
PVP2012      129.622     21.569   6.010 3.26e-08 ***
Cleader       79.929     51.812   1.543    0.126
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.43 on 97 degrees of freedom
Multiple R-squared:  0.2793,    Adjusted R-squared:  0.2644
F-statistic: 18.79 on 2 and 97 DF,  p-value: 1.264e-07
```

Table 2: Coefficients and P-Values of AIC-minimizing regression, sans-Sanders/Harris

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.273  -2.869  -1.869   2.131 151.727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.869      2.192   1.765   0.0808 .
PVP2012       49.404      6.441   7.670 1.56e-11 ***
Cleader       80.131     14.372   5.575 2.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.09 on 94 degrees of freedom
Multiple R-squared:  0.4769,    Adjusted R-squared:  0.4658
F-statistic: 42.86 on 2 and 94 DF,  p-value: 5.917e-14
```

Table 3: Introduction of 6 additional regressors into sans-Sanders/Harris model

```
Residuals:
    Min      1Q  Median      3Q     Max
-43.382  -3.338  -0.655   2.449 148.357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9661    36.2931   0.605    0.547
PVP2012      49.6957     7.2907   6.816 1.12e-09 ***
Cleader      80.5062    14.8020   5.439 4.75e-07 ***
Up2020        4.8066     4.4435   1.082    0.282
PVI          -0.3355     0.3652  -0.919    0.361
Caucus       23.2422    23.2830   0.998    0.321
PVIGap        8.9125    10.1218   0.881    0.381
Tscore       -0.2753     0.3640  -0.756    0.451
logpop       -0.9734     2.4127  -0.403    0.688
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.41 on 88 degrees of freedom
Multiple R-squared:  0.4944,    Adjusted R-squared:  0.4485
F-statistic: 10.76 on 8 and 88 DF,  p-value: 1.943e-10
```

Conclusion

Due to severe (although not unusually high-leverage) outliers, our normality assumptions were threatened, and the elimination of Bernie Sanders (I/D-VT) and Vice President-elect Kamala Harris (D-CA) from the dataset brought it far closer to our typical normality assumptions. Removing these values changes the coefficients of our regressors, and which regressors are considered significant. The elimination from the dataset of two very prominent senators in very Democratic states removes the significance of PVI, and the elimination from the dataset of two very prominent senators who are not caucus leaders lowers the p-value of 'Cleader' as well, rendering it significant. Besides the tradeoff between PVI and Cleader's significance, however, no other regressors became significant, and PVP2012 maintained its significance.

These results make implicit sense. It is perfectly believable that running for higher office significantly increases the public prominence of a given member of the Senate. Although the 'Cleader' variable's significance is dependent on whether Harris and Sanders are included in the dataset, its inclusion in the AICc-minimizing model is hardly surprising, as being chosen to lead a caucus can be expected to naturally increase a senator's sway and national profile. Although the linkage between Democratic slant and higher senatorial prominence, which exists only in the model that includes Harris and Sanders, is less clear-cut, there are many plausible narratives there as well. Democratic senators may disproportionately invest in building power through public prominence because they, as members of the minority, are unable to utilize the Senate's institutional power. Red-state Democrats, meanwhile, may keep lower national profiles than blue-state Democrats, making PVI more reliable than caucus identification; red-state Democratic senators, far more numerous than blue-state Republicans, may fear that wide coverage risks nationalizing their Senate races, which in turn could jeopardize the high levels of cross-over support from Republicans that they need to win re-election. In an alternate, sans-Harris/Sanders model, where PVI has lost its significance, these factors may be weaker and countered by the increased prominence that results from the passage of legislation and the confirmation of nominees, both of which only the majority can effectively and routinely accomplish.

Ultimately, our conclusions are that a small handful of regressors account for a large proportion of variation in public prominence, but that the factors behind senatorial prominence are far more idiosyncratic once we reach beyond the two most powerful predictive factors.

Although the multiple R^2 of a two-regressor model based only on running for higher office and being a caucus leader, both binomial variables, accounts for a staggering 46.98% of total variation as per un-adjusted R^2, the inclusion of half a dozen additional regressors, many that one would intuitively expect to have an outsize role on the process, only increases R^2 by less than 2.5pp, to 0.4944.

It is worth reiterating that these half-dozen regressors include many that would be expected to have an outsize role in determining public prominence. Caucus affiliation and voting record have very little predictive power, possibly undercutting the idea that the media gives disproportionate coverage to Democratic senators. Representing a large state population, being up for re-election, and representing a state with a different partisan lean than one's own, each of which would be expected to increase the number of people hearing a senator's name on a day-to-day basis, in fact have little predictive power.

Other factors did not even make the top eight regressors. Margin of victory, a proxy for seat competitiveness, could be expected to negatively correlate with senatorial prominence, by increasing the probability of a tough (and nationally watched) re-election if especially narrow. It is fairly clear that it does not. Being the committee chair or ranking member of a critical Senate committee gives members of the Senate enormous power when it comes to assuring their legislative priorities, and it is also a sign of deep respect within one's Senate caucus. The power these offices afford, and the clout they signify, could be expected to increase senatorial prominence, all else held equal. This effect appears to be minimal, to the extent that it exists at all. Length of time in office might be seen to give senators more time to accrue power, rack up legislative achievements, and increase public prominence. This does not seem to be the case.

Ultimately, our conclusions are threefold. The first is that running for higher office, and leading a caucus, together explain almost half of all variation in public prominence. The second is that, counterintuitively, many other key factors are essentially uncorrelated with public prominence. The third, arising from the second, is that a large portion of the variation in Senatorial prominence arises from unanticipated, idiosyncratic, and possibly Senator-specific factors. Further research to identify potential regressors to explain remaining variation could be justified, as well as research to identify whether these factors exist (or whether Senator-specific factors indeed dominate.)

Bibliography

"Annual Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019 ." Census.gov. US Census Bureau, December 30, 2019. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx.

Baquet, Dean, ed. "Senate Election Results." The New York Times. The New York Times, November 4, 2014. https://www.nytimes.com/elections/2014/results/senate.

Bluestein, Greg. "An Early Poll of Georgia's Twin Senate Runoffs Shows Tight Races." AJC. The Atlanta Journal-Constitution, November 11, 2020. https://www.ajc.com/politics/politics-blog/an-early-poll-of-georgias-twin-senate-runoffs-shows-tight-races/DUDDDLIHCJH4LORTT57HAOUG64/.

Bycoffe, Aaron, and Nate Silver. "Tracking Congress In The Age Of Trump." FiveThirtyEight. ABC News, October 27, 2020. https://projects.fivethirtyeight.com/congress-trump-score/.

Cook, Charles. "State PVIs." The Cook Political Report. Cook Political Report, 2019. https://cookpolitical.com/state-pvis.

Klein, Richard, ed. "2017 Elections: Alabama Senate." CNN. Turner Broadcasting Corp, December 12, 2017. https://www.cnn.com/election/2017/results/alabama-senate.

Klein, Richard, ed. "Election Results 2016." CNN. Turner Broadcasting Corp, November 8, 2016. https://www.cnn.com/election/2016/results.

Klein, Richard, ed. "Full Senate Election Results." CNN. Turner Broadcasting Corp, November 6, 2018. https://www.cnn.com/election/2018/results.

"List of Current United States Senators." Wikipedia. Wikimedia Foundation, November 13, 2020. https://en.wikipedia.org/wiki/List_of_current_United_States_senators.

Preston, Mark. "U.S. Senate Results -- 2014 Election Center -- Elections and Politics from CNN.com." CNN. Cable News Network, November 4, 2014. https://www.cnn.com/election/2014/results/race/senate/.

Remington Research Group. (2020). *Georgia: 2020 Runoff Election. Retrieved from https://www.ajc.com/politics/politics-blog/an-early-poll-of-georgias-twin-senate-runoffs-shows-tight-races/DUDDDLIHCJH4LORTT57HAOUG64/*

Appendix I: Building a Model that Includes Outliers

```
#Setting up the Dataset by reading a file
library(readxl)
library(ggplot2)
library(dplyr)
s410 <- read.csv("Senate_Datasheet.csv")
s410$Trends[11] <- 7
###Replacing Caucus R/D with Bernoulli 1/0 values
for (j in 1:nrow(s410)) {
  if (s410$Caucus[j] == "R") {
    s410$NumericCaucus[j] = 1
  } else {
    s410$NumericCaucus[j] = 0
  }
}
s410$Caucus <- s410$NumericCaucus
##Putting in stock vals for leverage before calculating
s410$leverage <- rep(0, nrow(s410))

s410$Caucus <- as.numeric(s410$Caucus)

s410$logpop <- log(s410$Population)

#Setting up the y-matrix, x-matrix and the hat-matrix
x_sampmat <- cbind(1, s410$PVI, s410$Tscore, s410$Population, s410$PVP2012,
                   s410$Up2020, s410$ComChairRM, s410$PVIGap, s410$YrsServed,
                   s410$Margin)

###as Caucus is not numeric/possibly multcoll. with PVI/PVIGap, it is excluded
kval <- 9
```

```
y_sampmat <- cbind(s410$Trends)
hat_sampmat <- x_sampmat %*% solve(t(x_sampmat) %*% x_sampmat) %*% t(x_sampmat)

#Eliminating High Leverage Points

#Calculating leverage for each observation by finding h_ii

for(i in 1:nrow(s410)){
  s410$leverage[i]<-hat_sampmat[i,i]
}
#Filtering out entries with leverage higher than 3 times the average
s410_filtered <- filter(s410, leverage <= 3*(kval+1)/100)
#Running initial linear regression
m6 <- lm(Trends~PVP2012+logpop+Caucus+Cleader+ComChairRM+Up2020,
data=s410_filtered)

X <- cbind(s410_filtered$PVP2012, s410_filtered$logpop, s410_filtered$Caucus,
s410_filtered$Cleader, s410_filtered$ComChairRM, s410_filtered$Up2020)

Y <- as.matrix(s410_filtered$Trends, ncol = 1)

library(leaps)



#With regsubsets, we are forcing the model to consider a set number of regressors.
#Regsubsets will choose the best
#These results will be invariant between AIC, AICc, Adj R2, BIC, etc, as we are forcing
#the consideration of a set number of regressors



b <- regsubsets(as.matrix(X), Y)
```

```
rs <- summary(b)

rs

rs$adjr2

par(mfrow = c(1, 2))
plot(1:6,rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")

## Creating and Choosing the Models

#Create a model for each subset size using optimal variables
m1 <- lm(Trends~PVP2012, data = s410_filtered)
m2 <- lm(Trends~PVP2012+Cleader, data = s410_filtered)
m3 <- lm(Trends~PVP2012+Cleader+Caucus, data = s410_filtered)
m4 <- lm(Trends~PVP2012+logpop+Caucus+Cleader, data = s410_filtered)
m5 <- lm(Trends~PVP2012+logpop+ComChairRM+Cleader+Caucus, data = s410_filtered)

# Subset Size 1

n <- length(m1$residuals)
npar <- length(m1$coefficients) +1
npar

#AIC and AICc for 1 regressor
extractAIC(m1,k=2)
extractAIC(m1,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 2
npar <- length(m2$coefficients) +1
```

```
#AIC and AICc for 2 regressors
extractAIC(m2,k=2)
extractAIC(m2,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 3

npar <- length(m3$coefficients) +1

#AIC and AICc for 3 regressors
extractAIC(m3,k=2)
extractAIC(m3,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 4
npar <- length(m4$coefficients) +1

#AIC and AICc for 4 regressors
extractAIC(m4,k=2)
extractAIC(m4,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 5

npar <- length(m5$coefficients) +1

#AIC and AICc for 5 regressors
extractAIC(m5,k=2)
extractAIC(m5,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 6
npar <- length(m6$coefficients) +1
```

```
#AIC and AICc for 6 regressors
extractAIC(m6,k=2)
extractAIC(m6,k=2)+2*npar*(npar+1)/(n-npar-1)

# Note that the lowest AIC and AICc occur on the model with 2 variables
# These two variables are PVP2012 and State Population

## Standard Diagnostic Tests

diagPlot<-function(model){
#Generating our diagnostic plots
#Fitted Values vs. Residuals
p1<- ggplot(model) + aes(.fitted, .resid) + geom_point()
p1<- p1 + stat_smooth(method="loess") + geom_hline(yintercept=0,
col="red",linetype="dashed")
p1<- p1 + xlab("Fitted values") + ylab("Residuals")
p1<- p1 + ggtitle("Residual vs Fitted Plot") + theme_bw()

#Normal Q-Q Plot
p2<- ggplot(model) + stat_qq(aes(sample = .stdresid)) + geom_abline()
p2<- p2 + xlab("Theoretical Quantiles")+ ylab("Std Residuals")
p2<- p2 + ggtitle("Normal Q-Q") + theme_bw()

#Scale-Location
p3<- ggplot(model, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(na.rm=TRUE)
p3<- p3 + stat_smooth(method="loess", na.rm = TRUE) + xlab("Fitted Value")
p3<- p3 + ylab(expression(sqrt("|Std residuals|")))
p3<- p3 + ggtitle("Scale-Location") + theme_bw()

#Cook's Distance
```

```
p4<- ggplot(model, aes(seq_along(.cooksd), .cooksd)) + geom_bar(stat="identity",
position="identity")
p4<- p4 + xlab("Obs. Number") + ylab("Cook's distance")
p4<- p4 + ggtitle("Cook's distance") + theme_bw()

#Residuals vs. Leverage
p5<- ggplot(model, aes(.hat, .stdresid)) + geom_point(aes(size=.cooksd), na.rm=TRUE)
p5<- p5 + stat_smooth(method="loess", na.rm=TRUE)
p5<- p5 + xlab("Leverage") + ylab("Std Residuals")
p5<- p5 + ggtitle("Residual vs Leverage Plot")
p5<- p5 + scale_size_continuous("Cook's Distance", range=c(1,5))
p5<- p5 + theme_bw() + theme(legend.position="bottom")

#Leverage and Cook’s Distance
p6<- ggplot(model, aes(.hat, .cooksd))+geom_point(na.rm=TRUE) +
stat_smooth(method="loess", na.rm=TRUE)
p6<- p6 + xlab("Leverage hii") + ylab("Cook's Distance")
p6<- p6 + ggtitle("Cook's dist vs Leverage hii/(1-hii)")
p6<- p6 + geom_abline(slope=seq(0,3,0.5), color="gray", linetype="dashed")
p6<- p6 + theme_bw()

return(list(rvfPlot=p1, qqPlot=p2, sclLocPlot=p3, cdPlot=p4, rvlevPlot=p5, cvlPlot=p6))}

diagPlot(m2)

plot(m2$fitted.values, Y)
abline(1, 1)
```

Appendix II: Building a Sans-Harris/Sanders Model

```
#Setting up the Dataset by reading a file
library(readxl)
library(ggplot2)
library(dplyr)
library(gridExtra)
s410 <- read.csv("Senate_Datasheet.csv")
s410$Trends[11] <- 7
###Replacing Caucus R/D with Bernoulli 1/0 values
for (j in 1:nrow(s410)) {
  if (s410$Caucus[j] == "R") {
    s410$NumericCaucus[j] = 1
  } else {
    s410$NumericCaucus[j] = 0
  }
}
s410$Caucus <- s410$NumericCaucus
##Putting in stock vals for leverage before calculating
s410$leverage <- rep(0, nrow(s410))

s410$Caucus <- as.numeric(s410$Caucus)

s410$logpop <- log(s410$Population)

#Setting up the y-matrix, x-matrix and the hat-matrix
x_sampmat <- cbind(1, s410$PVI, s410$Tscore, s410$logpop, s410$PVP2012,
                   s410$Up2020, s410$ComChairRM, s410$PVIGap, s410$YrsServed,
                   s410$Margin)

###as Caucus is not numeric/possibly multcoll. with PVI/PVIGap, it is excluded
kval <- 9
y_sampmat <- cbind(s410$Trends)
hat_sampmat <- x_sampmat %*% solve(t(x_sampmat) %*% x_sampmat) %*% t(x_sampmat)

#Eliminating High Leverage Points

#Calculating leverage for each observation by finding h_ii

for(i in 1:nrow(s410)){
  s410$leverage[i]<-hat_sampmat[i,i]
}
#Filtering out entries with leverage higher than 3 times the average
s410_filtered <- filter(s410, leverage <= 3*(kval+1)/100, Trends < 500)
```

```
#Running initial linear regression
m11 <- lm(Trends~PVP2012+logpop+Caucus+Cleader+ComChairRM+Up2020
        + PVI + Tscore + Margin + YrsServed + PVIGap, data=s410_filtered)

X <- cbind(s410_filtered$PVP2012, s410_filtered$logpop, s410_filtered$Caucus,
s410_filtered$Cleader, s410_filtered$ComChairRM, s410_filtered$Up2020,
          s410_filtered$PVI, s410_filtered$Tscore, s410_filtered$Margin, s410_filtered$YrsServed,
s410_filtered$PVIGap)

Y <- as.matrix(s410_filtered$Trends, ncol = 1)

library(leaps)


#With regsubsets, we are forcing the model to consider a set number of regressors.
#Regsubsets will choose the best
#These results will be invariant between AIC, AICc, Adj R2, BIC, etc, as we are forcing
#the consideration of a set number of regressors


b <- regsubsets(as.matrix(X), Y)

rs <- summary(b)

rs

rs$adjr2

par(mfrow = c(1, 2))
plot(rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")

## Creating and Choosing the Models

#Create a model for each subset size using optimal variables
m1 <- lm(Trends~PVP2012, data = s410_filtered)
m2 <- lm(Trends~PVP2012+Cleader, data = s410_filtered)
m3 <- lm(Trends~PVP2012+Cleader+Up2020, data = s410_filtered)
m4 <- lm(Trends~PVP2012+Cleader+Up2020+PVI, data = s410_filtered)
m5 <- lm(Trends~PVP2012+Cleader+Up2020+PVI+Caucus, data = s410_filtered)
m6 <- lm(Trends~PVP2012+Cleader+Up2020+PVI+Caucus+PVIGap, data = s410_filtered)
m7 <- lm(Trends~PVP2012+Cleader+Up2020+PVI+Caucus+PVIGap+Tscore, data =
s410_filtered)
m8 <- lm(Trends~PVP2012+Cleader+Up2020+PVI+Caucus+PVIGap+Tscore+logpop, data =
s410_filtered)
```

```
# Subset Size 1

n <- length(m1$residuals)
npar <- length(m1$coefficients) +1
npar

#AIC and AICc for 1 regressor
extractAIC(m1,k=2)
extractAIC(m1,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 2
npar <- length(m2$coefficients) +1

#AIC and AICc for 2 regressors
extractAIC(m2,k=2)
extractAIC(m2,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 3

npar <- length(m3$coefficients) +1

#AIC and AICc for 3 regressors
extractAIC(m3,k=2)
extractAIC(m3,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 4
npar <- length(m4$coefficients) +1

#AIC and AICc for 4 regressors
extractAIC(m4,k=2)
extractAIC(m4,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 5

npar <- length(m5$coefficients) +1

#AIC and AICc for 5 regressors
extractAIC(m5,k=2)
extractAIC(m5,k=2)+2*npar*(npar+1)/(n-npar-1)

# Subset Size 6
npar <- length(m6$coefficients) +1
```

```
#AIC and AICc for 6 regressors
extractAIC(m6,k=2)
extractAIC(m6,k=2)+2*npar*(npar+1)/(n-npar-1)

# Note that the lowest AIC and AICc occur on the model with 2 variables
# These two variables are PVP2012 and Caucus Leadership

## Standard Diagnostic Tests

diagPlot<-function(model){
  #Generating our diagnostic plots
  #Fitted Values vs. Residuals
  p1<- ggplot(model) + aes(.fitted, .resid) + geom_point()
  p1<- p1 + geom_hline(yintercept=0, col="red",linetype="dashed")
  p1<- p1 + xlab("Fitted values") + ylab("Residuals")
  p1<- p1 + ggtitle("Residual vs Fitted Plot") + theme_bw()

  #Normal Q-Q Plot
  p2<- ggplot(model) + stat_qq(aes(sample = .stdresid)) + geom_abline()
  p2<- p2 + xlab("Theoretical Quantiles")+ ylab("Std Residuals")
  p2<- p2 + ggtitle("Normal Q-Q") + theme_bw()

  #Scale-Location
  p3<- ggplot(model, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(na.rm=TRUE)
  p3<- p3 + xlab("Fitted Value")
  p3<- p3 + ylab(expression(sqrt("|Std residuals|")))
  p3<- p3 + ggtitle("Scale-Location") + theme_bw()

  #Cook's Distance
  p4<- ggplot(model, aes(seq_along(.cooksd), .cooksd)) + geom_bar(stat="identity",
position="identity")
  p4<- p4 + xlab("Obs. Number") + ylab("Cook's distance")
  p4<- p4 + ggtitle("Cook's distance") + theme_bw()

  #Residuals vs. Leverage
  p5<- ggplot(model, aes(.hat, .stdresid)) + geom_point(aes(size=.cooksd), na.rm=TRUE)
  p5<- p5 + xlab("Leverage") + ylab("Std Residuals")
  p5<- p5 + ggtitle("Residual vs Leverage Plot")
  p5<- p5 + scale_size_continuous("Cook's Distance", range=c(1,5))
  p5<- p5 + theme_bw() + theme(legend.position="bottom")

  #Leverage and Cook's Distance
  p6<- ggplot(model, aes(.hat, .cooksd))+geom_point(na.rm=TRUE)
  p6<- p6 + xlab("Leverage hii") + ylab("Cook's Distance")
```

```
  p6<- p6 + ggtitle("Cook's dist vs Leverage hii/(1-hii)")
  p6<- p6 + geom_abline(slope=seq(0,3,0.5), color="gray", linetype="dashed")
  p6<- p6 + theme_bw()

  grid.arrange(p1,p2,p3,p4,p5,p6, ncol=3)}

diagPlot(m2)

plot(m2$fitted.values, Y)
abline(1, 1)
summary(m2)
```

Self-Reflection

Moses

This project was fairly time-consuming over the course of this week, maybe to the tune of about 15 hours, although data collection and write-ups were more time consuming than the actual coding. Part of this was that some of the data we needed was difficult to quantify and difficult to obtain, so we put a lot of effort into identifying good instruments and proxies, collecting reliable data, and making the right judgment calls- determining estimated margins for an unelected senator, determining whether to code a senator without major party opposition as up for re-election, and similar complications. I learned a lot about the difficulty of assembling reliable datasets in a world where edge cases like the ones we came across often proliferate.

In light of that time breakdown, where the brunt of the work was done in data collection, our problem of normality assumptions falling under the weight of non-high-leverage outliers is somewhat ironic. If I did this over again, I would have run the regression and generated sans-outlier diagnostics before the presentation, not after, so we could make our conclusions with greater confidence during the presentation. For future students, I would urge that when doing breakdowns of the project's time to completion, especially if assembling the dataset is the primary difficulty of your project, you should give yourself a day's wiggle room to ensure that unexpected, late-breaking errors that might only come up in the coding and analysis phase have adequate time to be resolved.

Elliot