Polls, Fundamentals, or Both?

Predicting 2021 Gubernatorial Races in Virginia & New Jersey

Allison Marie Towey

October 28, 2021

Forecasting elections is an enterprise that spans academia, campaigns, and journalism. Crafting accurate predictive models has become an industry of its own, with campaigns and news organizations spending significant funds to predict electoral outcomes. Whereas presidential elections have taken significant attention, gubernatorial predictive modeling has not seen the same widespread info-tainment attention. One potential reason for this discrepancy is the variance of state governors' races. In this report, I evaluate three simple, reusable models that may work across states and that reflect either campaign/state fundamentals, pre-election polling, or both. In my analyses of both states' 2021 gubernatorial races, the model that reflects both fundamentals (campaign finance and state unemployment race) and pre-election polling was shown to be the most predictive. *Please note: full R code available at end of report.*

Selection of Indicators for Predictive Model

To create my gubernatorial prediction model, I built three models based on different data:

1. State Gubernatorial Polling Data
2. State and Campaign Fundamentals
3. Both Polling Data and Fundamentals

State Gubernatorial Polling Data

One increasingly less-trusted by the public, yet still common, method of ascertaining the state of a particular race is to survey prospective voters. In recent high-profile Presidential elections, polling predictions have not perfectly matched observed electoral outcomes, leading to some write off polling as unreliable or unusable. Others, however, claim that polling is still the most reliable method to predict electoral outcomes. (Clinton) Public opinion is undoubtedly difficult to ascertain with any degree of certainty, but polls and surveying methods remain the most accepted methods to understanding electoral behavior. (Hillygus) I chose this particular predictor because of polling's continued prevalence in academic literature and campaigning as the preferred method of choice in understanding and forecasting voting behavior.

Presidential elections have seen a dramatic rise in horse-race polling data. On any given week during a presidential elections, dozens of polls could be conducted. Presidential elections remain the most-talked about (and often most lucrative to talk about) elections in the United States, so it is not surprising that presidential elections receive far more polling attention than do gubernatorial races. Gubernatorial races concern smaller electorates and often do not command the attention of the national media that may have the resources conduct polls. Further, there is a great variance between states. Larger states like California or more politically-center states like Virginia often receive more attention than do small, ideologically homogeneous

states like Hawaii or Wyoming, and therefore less high-quality polling. This likely makes the predictive power of polls less generalizable across state contests.

While some polls are superior to others given pollsters' methods of sampling and representative weighting, they are still commonly used and respected predictor of elections. In crafting my model, I wanted to test pre-election polls' predictive power in New Jersey and Virginia.

State and Campaign Fundamentals

There are infinite factors that may impact a race in any given state in any given year. Predictive modeling, despite how well-researched and considered, cannot account for all noise in the data. There are, however, certain characteristics about a state gubernatorial race that may be correlated with electoral outcomes. The two "fundamentals" that I wanted to look into in this model were state-wide economic factors and individual candidate factors.

For my measure of state-wide economic well-being, I selected the August-of-election-year unemployment rate. Particularly, this model is looking at whether a higher or lower unemployment rate correlates with the state choosing a Democratic candidate instead of a Republican. I chose unemployment rates for practical and theoretical reasons. From a practical perspective, this data was available from a single, reliable source (FRED) for all of the years I wanted to include in the model. From a theoretical perspective, I chose unemployment rates because they impact the voters directly, and voters are more likely to know and understand what it is. Unemployment rates are discussed often in the news, so it is more likely that the voters will know this figure in comparison to other figures such as median household income or state GDP growth rates. Unemployment also impacts individual people, so it may be a more salient indicator of statewide economic well-being that other economic measures. I chose to calculate this figure in August, as voters likely make their preferences for who to vote for when the campaigns are in full-swing, which is around late summer.

For my measure of candidate factors, I selected campaign contribution usage. If a candidate is able to raise more funds, it follows that the candidate has more voters or more enthusiastic voters (or potentially richer voters/interest groups) in their camp. This may portend a higher turn-out for this candidate. Additionally, more campaign funds mean more possible advertisements or campaign events. If the voters are to be exposed to the candidate more, they may feel more inclined to vote for them.

My gubernatorial prediction model seeks to work across states without a new model being built for each state each election cycle. Unemployment rate and campaign fund usage are two factors that largely do not need contextualization, so they are easily generalizable across state races.

A Few Notes on Data Availability and Quality

In order to create these models, I had to rely on publicly-available data. In Virginia, polling data and campaign finance data were only available from the 1993 contest to the present day. As a result, there are only 7 observations in my dataset for Virginia. For New Jersey, there was polling data and campaign finance information from 1981 to the present day, which gave me 10 observations to use in model building. The small number of observations (in large part due to the lack of digitization prior to the 2000's and the four-year election cycles in both states) may portend some lack of reliability, as one outlier observation could impact the data significantly.

The predictions in this report assume the data used is equally reliable and accurate. Due to the difficulty of finding data prior to the internet age circa 2000, the average polling data was captured from different sources including Ballotopedia, Wikipedia, and the University of Virginia's Cooper Center. Because polling averages used may not include every poll from that state's gubernatorial race, I cannot be certain the polling averages used are completely accurate.

Campaign finance data also suffers from this potential inconsistency problem, particularly in Virginia. For races prior to 1997, the Virginia Department of Elections did not have this data available online. The 1993

finance data was taken from the University of Virginia's Cooper Center. This discrepancy in where the data is sourced could also bias my findings.

In order to incorporate enough usable data in my model, I decided to use this data from different sources instead of relying on a smaller sample to use. I opted to use more observations using potentially less-consistent data instead of fewer observations in which one observation may have undue influence on my model.

This model also assumes no significant change in current-year polling data or campaign finance data in the next week. Because this report was produced 10/27/2021, polls and campaign contributions used from October 27 to November 2 (election day) are not factored into the calculation.

Race Predictions

Virginia (Terry McAuliffe and Glenn Youngkin)

For Virginia, I first built the three models as described previously. I collected August unemployment rate data from FRED, polling data from RealClearPolitics, and campaign finance data from Ballotopedia and the Virginia Board of Elections Website. Data was available for these variables from 1993 to the present day. Using this data, I regressed each of the variable groups on Democratic Vote Share.

To determine the predictive ability of each model, I cross-validated the data using a two-fold method. I chose a two-fold method instead of a larger k-value because I only had 7 rows of data (1993 to 2017) to use in my model, and groups smaller than 3 or 4 would have made my data far noisier.

The model containing polls and fundamentals was shown to have the smallest absolute mean testing error, which means it was the most predictive model:

- Polls Only Model: 0.0951284
- Fundamentals Only Model: 0.1230897
- Polls and Fundamentals Model: 0.0646501

I then decided to use this particular model to predict the outcome of the 2021 Virginia gubernatorial race using relevant data from this year.

After regressing my variables for this model (polls, August unemployment rate, and Democratic share of campaign contributions used), I found the intercept and coefficients of this model to be:

- Intercept: 0.0201425
- Democratic Poll Share: 1.2598037
- August Unemployment Rate: -0.0090538
- Democratic Money Share: -0.2124773

My prediction for the 2021 campaign uses the coefficients from the model created and up-to-date 2021 data. *Note: All data calculated 10/27/2021.*

- State Unemployment Rate August 2021 = 4.0
- Democratic Average Poll Share = 0.5189474
- Democratic Money Share = 0.6286291

The model's prediction for the Democratic vote share in the Virginia gubernatorial race is: **0.5041297**. I therefore predict Terry McAuliffe, the Democratic candidate, to win the race narrowly over his Republican counterpart, Glenn Youngkin.

New Jersey (Phil Murphy and Jack Ciattarelli)

I followed the same approach to predict New Jersey's gubernatorial election. I collected August unemployment rate data from FRED, polling data from RealClearPolitics, and campaign finance data from Ballotopedia and the New Jersey Board of Elections Website. Data was available for these variables from 1981 to the present day (significantly more than the data available from Virginia). Using the New Jersey data, I regressed each of the variable groups on Democratic Vote Share. I then cross-validated the data using a three-fold method to determine the predictive ability of each model.

I used a three-fold cross validation method to determine which model was the most accurate in predicting electoral outcomes. I chose a three-fold method instead of a two-fold method as I used for the Virginia prediction because I had 11 rows of data (1981 to 2017) to use in my calculations instead of 7. The mean validation errors I received for each model are as follows:

- Polls Only Model: 0.0526938
- Fundamentals Only Model: 0.0780431
- Both Polls and Fundamentals Model: 0.0267934

The Polls and Fundamentals Model had the smallest mean error of 0.0267934 as a result of the three-fold cross-validation, meaning it is the most accurate. I then used this particular model to predict the outcome of the 2021 New Jersey gubernatorial race using relevant data from this year.

I first ran a regression for the Polls and Fundamentals model and received the following intercept and coefficients:

- Intercept: 0.2303239
- Democratic Poll Share: 0.6184894
- August Unemployment Rate: -0.0110064
- Democratic Money Share: 0.0558206

I then calculated my prediction for the 2021 campaign using the coefficients from the model created and up-to-date 2021 data. *Note: All data calculated 10/27/2021.*

- State Unemployment Rate August 2021 = 6.9
- Democratic Average Poll Share = 0.5189474
- Democratic Money Share = 0.6286291

The model's prediction for the Democratic vote share in the Virginia gubernatorial is: **0.5421135**. I therefore predict Phil Murphy, the incumbent Democratic candidate, to win the race by a few percentage points over his Republican challenger, Jack Ciattarelli.

Discussion

Performing a historical prediction model on these gubernatorial races proved difficult. Data, particularly polling data and campaign finance data, were hard to find and came from different, non-centralized or questionably-trustworthy sources. Services like Ballotpedia or Wikipedia are helpful in finding this data for informational purposes, but using them for academic research is likely problematic.

The rows in my dataset for this predictive modeling exercise comprised solely of prior gubernatorial elections from which I had polling, unemployment, and campaign finance data. I was only able to find reliable data from 1993 to today in Virginia and 1981 to today in New Jersey As a consequence, I only had 7 rows and 10 rows respectively with which to build and test the complete model. This lack of rows in my datasets made data analysis noisy. In order to cross validate for Virginia, I had to split my data into training and testing groups of 3 or 4, which made findings vary due to the small sample size.

Despite the logistical challenges due to the small sample sizes of the observations, the predictions in this report appear feasible. According to news sites and polling analysis sites like FiveThirtyEight, the Virginia race appears to be a toss up/lean-Dem and the New Jersey race appears to be a lean-Dem race. (FiveThirtyEight, Vox, New York Magazine) My vote share predictions of *0.5041297* and *0.5421135* respectively agree.

Appendices

Sources

- Clinton, Joshua. https://www.vanderbilt.edu/unity/2021/01/11/polling-problems-and-why-we-should-still-trust-some-polls/
- Hillygus, D. Sunshine. https://academic.oup.com/poq/article/75/5/962/1830219
- FiveThirtyEight. https://projects.fivethirtyeight.com/polls/governor/virginia/?ex_cid=rrpromo
- Vox Media. https://www.vox.com/22725133/virginia-new-jersey-elections-midterms-biden
- New York Magazine. https://nymag.com/intelligencer/2021/10/new-jersey-governors-race-is-getting-interesting.html
- FRED. https://fred.stlouisfed.org/series/
- Ballotpedia. https://ballotpedia.org/
- RealClearPolitics. https://www.realclearpolitics.com
- New Jersey State Board of Elections. https://www.elec.state.nj.us/
- Virginia State Board of Elections. https://www.elections.virginia.gov/
- Cooper Center Newsletter. https://newsletter.coopercenter.org/

Code

Virginia

```
#Set up
options(warn = -1)
setwd("~/Downloads")
library(dplyr)
library(readxl)
va <- read_excel("Virginia (1).xlsx")
va = va[!is.na(va$rep_money), ] #remove rows with no data
va["dem_pollsh"] <- (va$dem_poll_avg / (va$dem_poll_avg + va$rep_poll_avg))
va["dem_moneysh"] <- (va$dem_money / (va$dem_money + va$rep_money))
va["dem_vtsh"] <- (va$dem_vts / (va$dem_vts+va$rep_vts))
```

```r
# Find predicted values
#polls
reg3 = lm(dem_vtsh ~ dem_pollsh, data = va)
va <- va %>%
  mutate(predict_poll = reg3$coefficients[1] + reg3$coefficients[2]
         *dem_moneysh)

#polls + fundamentals
reg4 = lm(dem_vtsh ~ dem_pollsh + unemploy + dem_moneysh, data = va)
summary(reg4)
va <- va %>%
  mutate(predict_all = reg4$coefficients[1] + reg4$coefficients[2]
         *dem_pollsh + reg4$coefficients[3] *unemploy + reg4$coefficients[4]
         *dem_moneysh)

#fundamentals only
reg5 = lm(dem_vtsh ~ unemploy + dem_moneysh, data = va)
va <- va %>%
  mutate(predict_allnop = reg5$coefficients[1]  + reg5$coefficients[2]
         *unemploy + reg5$coefficients[3] *dem_moneysh)



#Find mean all-sample errors:
#polls
va <- va %>%
  mutate(residual_poll = dem_vtsh - predict_poll)

va <- va %>%
  mutate(abs_error_poll = abs(residual_poll))
mean(va$abs_error_poll, na.rm = TRUE)

#polls + fundamentals
va <- va %>%
  mutate(residual_all = dem_vtsh - predict_all)

va <- va %>%
  mutate(abs_error_all = abs(residual_all))
mean(va$abs_error_all, na.rm = TRUE)

#fundamentals
va <- va %>%
  mutate(residual_allnop = dem_vtsh - predict_allnop)

va <- va %>%
  mutate(abs_error_allnop = abs(residual_allnop))
mean(va$abs_error_allnop, na.rm = TRUE)



# Perform 2-fold cross validation
set.seed(1) #set seed to simulate randomness
va <- va %>%
```

```
  mutate(random = runif(7)) %>%
  arrange(random)

va$partition[va$random<.5] <- 0
va$partition[va$random>=.5] <- 1

va <- va %>%
  mutate(test_error = NA)

# Run validation
for (i in 0:1) {
  #Model 3: Polls
  model3 <- lm(dem_vtsh ~ dem_pollsh, data = va[va$partition != i,])
  va$yhat <- predict(model3, newdata = va)
  va[[paste0('train_error', i)]] <- va$dem_vtsh - va$yhat
  va$test_error <- ifelse(va$partition == i, va[[paste0('train_error', i)]],
                          va$test_error)
  va$abs_test_error3 = abs(va$test_error)
  }

for (i in 0:1) {
  #Model 4: All (Polls, Money, Unemployment)
  model4 <- lm(dem_vtsh ~ (dem_pollsh + dem_moneysh + unemploy), data =
               va[va$partition != i,])
  va$yhat <- predict(model4, newdata = va)
  va[[paste0('train_error', i)]] <- va$dem_vtsh - va$yhat
  va$test_error <- ifelse(va$partition == i, va[[paste0('train_error', i)]],
                          va$test_error)
  va$abs_test_error4 = abs(va$test_error)
}

for (i in 0:1) {
  #Model 5: Money, Unemployment
  model5 <- lm(dem_vtsh ~ (dem_moneysh + unemploy), data =
               va[va$partition != i,])
  va$yhat <- predict(model5, newdata = va)
  va[[paste0('train_error', i)]] <- va$dem_vtsh - va$yhat
  va$test_error <- ifelse(va$partition == i, va[[paste0('train_error', i)]],
                          va$test_error)
  va$abs_test_error5 = abs(va$test_error)
}

#Summarize absolute testing errors
summary(va$abs_test_error3, na.rm = TRUE)
summary(va$abs_test_error4, na.rm = TRUE)
summary(va$abs_test_error5, na.rm = TRUE)
```

New Jersey

```
#Set Up
options(warn = -1)
setwd("~/Downloads")
```

```
library(dplyr)
library(readxl)
nj <- read_excel("NJ (1).xlsx")
nj$rep_money <- as.numeric(gsub(",","", nj$rep_money))
nj$dem_money <- as.numeric(gsub(",","", nj$dem_money))
nj["dem_pollsh"] <- (nj$dem_poll_avg / (nj$dem_poll_avg + nj$rep_poll_avg))
nj["dem_moneysh"] <- (nj$dem_money / (nj$dem_money + nj$rep_money))
nj["dem_vtsh"] <- (nj$dem_vts / (nj$dem_vts+nj$rep_vts))

#Predicted values
reg3 = lm(dem_vtsh ~ dem_pollsh, data = nj)
nj <- nj %>%
  mutate(predict_poll = reg3$coefficients[1] + reg3$coefficients[2]
         *dem_moneysh)

reg4 = lm(dem_vtsh ~ dem_pollsh + unemploy + dem_moneysh, data = nj)
summary(reg4)
nj <- nj %>%
  mutate(predict_all = reg4$coefficients[1] + reg4$coefficients[2] *dem_pollsh
         + reg4$coefficients[3] *unemploy + reg4$coefficients[4] *dem_moneysh)

reg5 = lm(dem_vtsh ~ unemploy + dem_moneysh, data = nj)
nj <- nj %>%
  mutate(predict_allnop = reg5$coefficients[1]  + reg5$coefficients[2] *unemploy
         + reg5$coefficients[3] *dem_moneysh)

#Find mean all-sample errors:
#polls
nj <- nj %>%
  mutate(residual_poll = dem_vtsh - predict_poll)

nj <- nj %>%
  mutate(abs_error_poll = abs(residual_poll))
mean(nj$abs_error_poll, na.rm = TRUE)

#polls + fundamentals
nj <- nj %>%
  mutate(residual_all = dem_vtsh - predict_all)

nj <- nj %>%
  mutate(abs_error_all = abs(residual_all))
mean(nj$abs_error_all, na.rm = TRUE)

#fundamentals
nj <- nj %>%
  mutate(residual_allnop = dem_vtsh - predict_allnop)

nj <- nj %>%
  mutate(abs_error_allnop = abs(residual_allnop))
mean(nj$abs_error_allnop, na.rm = TRUE)

# Perform 3-fold cross validation
nj <- nj %>%
```

```
  mutate(random = runif(11)) %>%
  arrange(random)

nj$partition[nj$random<1/3] <- 0
nj$partition[nj$random>=1/3 & nj$random<2/3] <- 1
nj$partition[nj$random>=2/3] <- 2

nj <- nj %>%
  mutate(test_error = NA)

#Run validation
for (i in 0:2) {
  #Model 3: Polls
  model3 <- lm(dem_vtsh ~ dem_pollsh, data = nj[nj$partition != i,])
  nj$yhat <- predict(model3, newdata = nj)
  nj[[paste0('train_error', i)]] <- nj$dem_vtsh - nj$yhat
  nj$test_error <- ifelse(nj$partition == i, nj[[paste0('train_error', i)]],
                          nj$test_error)
  nj$abs_test_error3 = abs(nj$test_error)
}

for (i in 0:2) {
  #Model 4: All (Polls, Money, Unemployment)
  model4 <- lm(dem_vtsh ~ (dem_pollsh + dem_moneysh + unemploy), data =
                 nj[nj$partition != i,])
  nj$yhat <- predict(model4, newdata = nj)
  nj[[paste0('train_error', i)]] <- nj$dem_vtsh - nj$yhat
  nj$test_error <- ifelse(nj$partition == i, nj[[paste0('train_error', i)]],
                          nj$test_error)
  nj$abs_test_error4 = abs(nj$test_error)
}

for (i in 0:2) {
  #Model 5: Money, Unemployment
  model5 <- lm(dem_vtsh ~ (dem_moneysh + unemploy), data = nj[nj$partition
                                                              != i,])
  nj$yhat <- predict(model5, newdata = nj)
  nj[[paste0('train_error', i)]] <- nj$dem_vtsh - nj$yhat
  nj$test_error <- ifelse(nj$partition == i, nj[[paste0('train_error', i)]],
                          nj$test_error)
  nj$abs_test_error5 = abs(nj$test_error)
}

#Summarize absolute testing errors
mean(nj$abs_test_error3, na.rm = TRUE)
mean(nj$abs_test_error4, na.rm = TRUE)
mean(nj$abs_test_error5, na.rm = TRUE)
```