Election Pulse Checker Methodology

Alex Bass

9/29/2020

Welcome!

(TL;DR: I modeled survey methodology and population and made adjustments accordingly. I used a one-sample t-test to calculate a probability for each state. I then used each state's probability directly to simulate the election.)

A few disclaimers before we begin: calculating polling standard errors are a bit controversial since there is always a lot of bias that is really hard to account for (non-response bias, social desirability bias, etc.). I will try my best to be detailed as I walk you throught the process.

Probability Calculation

First, I loaded the 538 presidential polling data set found on their website: fivethirtyeight.com

Next, I filtered out polls that 538 ranked as C or below and filtered for state specific results. This way I will only have higher quality polls in my analysis.

```
TvB <- read.csv("https://projects.fivethirtyeight.com/polls-page/president_polls.csv")

new <- TvB %>%
  select(question_id,
         answer,
         pct,
         pollster,
         pollster_id,
         fte_grade,
         end_date,
         methodology,
         state,
         population,
         candidate_party,
         sample_size) %>%
  filter(fte_grade %in% c("A+", "A", "A-", "A/B","B+", "B", "B-", "B/C"), state!="")

head(new)
```

```
##   question_id    answer pct                                pollster pollster_id
## 1      130405     Biden  48 Siena College/The New York Times Upshot        1424
## 2      130405     Trump  41 Siena College/The New York Times Upshot        1424
## 3      130405 Jorgensen   4 Siena College/The New York Times Upshot        1424
## 4      130426     Biden  49 Siena College/The New York Times Upshot        1424
## 5      130426     Trump  40 Siena College/The New York Times Upshot        1424
```

```
## 6      130401     Biden  48                  Public Policy Polling         383
##   fte_grade end_date methodology         state population candidate_party
## 1        A+  9/27/20  Live Phone Nebraska CD-2         lv             DEM
## 2        A+  9/27/20  Live Phone Nebraska CD-2         lv             REP
## 3        A+  9/27/20  Live Phone Nebraska CD-2         lv             LIB
## 4        A+  9/27/20  Live Phone  Pennsylvania         lv             DEM
## 5        A+  9/27/20  Live Phone  Pennsylvania         lv             REP
## 6         B  9/26/20    IVR/Text         Texas         lv             DEM
##   sample_size
## 1         420
## 2         420
## 3         420
## 4         711
## 5         711
## 6         612
```

My goal was to try to look at every poll on the same scale, but there are a few important differences between polls that I had to consider at this point.

1. Methodology bias: Would different methodologies reach different groups and be biased?
2. House bias: How would I deal with systematic differences from pollsters?
3. 3rd party/Unsure bias: Some of the questions are "full ballot" questions meaning they include third party candidates while some do not. Some questions include an unsure option while some do not. Support for third party candidates may be different in different states.
4. Bias from different sample populations: Different polls sample from different populations which obviously creates systematic differences in the results.

I chose to address 1 and 4 by adjustments using model estimates. I chose to address 3 by weighting. I found I did not have enough power for estimating house effects (3), so will simply acknowledge house effects as a possible bias and limitation of my design. However, I still included pollsters in my model in an effect to control for house effects for the sake of the other variables.

The regression models I used to estimate the effects are included below.

```
trumplm <- plm(pct ~ meth + state + pollster + population + sample_size +
                 third_option + fb,
               data = wFB %>% filter(answer=="Trump"), model = "within", index = c("state"))

bidenlm <- plm(pct ~ meth + state + pollster + population + sample_size +
                 third_option + fb,
               data = wFB %>% filter(answer=="Biden"), model = "within", index = c("state"))

TRUMP <- coeftest(trumplm, vcov=vcovHC(trumplm, type="sss", cluster="group"))
BIDEN <- coeftest(bidenlm, vcov=vcovHC(bidenlm, type="sss", cluster="group"))

head(TRUMP)
```

```
##                                      Estimate Std. Error    t value
## methLive Phone                      0.2930141  1.3684404  0.2141227
## methOnline                         -1.6013054  1.3812190 -1.1593422
## methIVR/Online                     -1.4709712  0.6103007 -2.4102402
## methIVR/Text                        2.6455016  1.2774384  2.0709426
## pollsterABC News/The Washington Post -3.9552045  1.1888137 -3.3270181
```

```
## pollsterAlaska Survey Research       -5.9082344  1.8210397 -3.2444292
##                                          Pr(>|t|)
## methLive Phone                       0.8305143613
## methOnline                           0.2467160280
## methIVR/Online                       0.0162014757
## methIVR/Text                         0.0387337981
## pollsterABC News/The Washington Post 0.0009241563
## pollsterAlaska Survey Research       0.0012333856
```

```
head(BIDEN)
```

```
##                                          Estimate Std. Error      t value
## methLive Phone                       -0.884212277  1.1404833 -0.775296121
## methOnline                           -0.001655723  1.0363463 -0.001597654
## methIVR/Online                        1.826450968  0.7644304  2.389296537
## methIVR/Text                         -2.204552459  1.6794785 -1.312641105
## pollsterABC News/The Washington Post  3.255503688  1.2683838  2.566654965
## pollsterAlaska Survey Research        3.335159002  1.2933004  2.578796841
##                                            Pr(>|t|)
## methLive Phone                       0.43842917
## methOnline                           0.99872572
## methIVR/Online                       0.01714710
## methIVR/Text                         0.18973845
## pollsterABC News/The Washington Post 0.01047735
## pollsterAlaska Survey Research       0.01011962
```

A few notes about the models above.

1. I estimated Trump and Biden in separate models.
2. I used state fixed effects to control for differences found at the state level.
3. I estimated the effects of methodologies, pollsters, populations, etc.

I then filtered the coefficients for only those which showed statistical significance (those values statistically different from zero) to use for adjustments for each poll.

From this point, I will just show what happens from one state's perspective for simplification.

Here is an example question of an Ohio poll:

```
example <- wFB %>% filter(question_id == min(as.numeric(question_id)))
example
```

```
## # A tibble: 2 x 16
##   question_id answer   pct pollster pollster_id fte_grade end_date methodology
##         <int> <chr>  <dbl> <chr>          <int> <chr>     <chr>    <chr>
## 1       92080 Biden     48 Public ~         383 B         11/28/18 Automated ~
## 2       92080 Trump     44 Public ~         383 B         11/28/18 Automated ~
## # ... with 8 more variables: state <chr>, population <fct>,
## #   candidate_party <chr>, sample_size <int>, unsure <dbl>, third_option <chr>,
## #   meth <fct>, fb <chr>
```

I would first adjust the candidate according to the modeled biases.

So, for this case, population is voters and methodology is automated phone. Below, I create a table for each candidates model coefficients *only significant ones*, show a series of t/f statements, and adjust the population for trump because it was a statistically significant effect in my model for trump.

```
#creates a vector of significant Trump coeficients
TCoef <- TRUMP[which(TRUMP[,4]<0.05),]
TCoef <- TCoef[,1]

#cleaning the names of the vector
names(TCoef) <- gsub("pollster|population|third_option|fb|meth", "", names(TCoef))

#creates a vector of significant Trump coeficients
BCoef <- BIDEN[which(BIDEN[,4]<0.05),]
BCoef <- BCoef[,1]

#cleaning the names of the vector
names(BCoef) <- gsub("pollster|population|third_option|fb|meth", "", names(BCoef))

#checking to see if elements in the survey are part of the significant coefficient list
example$population[1] %in% names(TCoef) ##this is true
```

```
## [1] TRUE
```

```
example$population[1] %in% names(BCoef)
```

```
## [1] FALSE
```

```
example$methodology[1] %in% names(TCoef)
```

```
## [1] FALSE
```

```
example$methodology[1] %in% names(BCoef)
```

```
## [1] FALSE
```

```
## In this case, only the pollster was found to be one of the two lists of coeficients.
#adjusting the percent according to the coefficient
example$pct
```

```
## [1] 48 44
```

```
example$pct[example$answer=="Trump"] <-
  example$pct[example$answer=="Trump"] - TCoef[example$population[1]]

example$pct
```

```
## [1] 48.00000 49.90823
```

After we adjusted the poll for the bias, we come up with Biden at 48% and Trump at 49.9%.

We then weight to 100.

```
weight_to_100 <- function(data){
  weight <- 100/sum(data[["pct"]])
  data <- data %>% filter(answer == "Trump")
  data[["pct"]] <- data[["pct"]]*weight
  return(data)
}

example <- weight_to_100(example)
example
```

```
## # A tibble: 1 x 16
##   question_id answer   pct pollster pollster_id fte_grade end_date methodology
##         <int> <chr>  <dbl> <chr>          <int> <chr>     <chr>    <chr>
## 1       92080 Trump   51.0 Public ~         383 B         11/28/18 Automated ~
## # ... with 8 more variables: state <chr>, population <fct>,
## #   candidate_party <chr>, sample_size <int>, unsure <dbl>, third_option <chr>,
## #   meth <fct>, fb <chr>
```

Above I have trump's polled vote share of Ohio of 50.97%. Now I do a one sample t-test using .5 as the null hypothesis to generate a probability of Trump winning.

```
one_sample_ttest <- function(data){
  data[["pct"]] <- as.numeric(data[["pct"]])/100
  SE <- sqrt(data[["pct"]]*(1-data[["pct"]])/data[["sample_size"]])
  z_score <- (data[["pct"]]-.5)/SE
  return(pnorm(z_score))
}

one_sample_ttest(example)
```

```
## [1] 0.6901337
```

Our output is then that according to a rough standard error calculated from this poll, 69% of the time the true proportion will be above 50% in Trump's favor.

(I calculate the standard error using the Bernoulli equation because once i weight to 100 third parties are out of the picture)

To get the final probability, I repeat this process with 4 more polls. Then, I take an average of the probabilities and that is what is displayed on the map for each state.

Simulation

To generate the simulation, I use the probability for each state to predict a trump win. I add up all the electoral votes for trump, and then determine the win, loss, or tie. This function is found in my predict_elections.R file.

(There has been some talk of how these simulations are biased because state wins are correlated with each other. Therein is another limitation to be accounted for. I am only trying to get a rough idea of what polling is looking like at the moment).