An Analysis of Errors in the 2016 Presidential Election Polls

David (Daiwei) Zhang

April 26, 2017

1 Introduction

The 2016 presidential election generated many interesting datasets and statistical problems. Most polls have underestimated Trump's performance, including state polls, national polls, forecasts, and exit polls (Bialik & Enten (2016)). Exports have suggested different survey factors that caused this result, especially the nonresponse bias and low turnout rate (Stein (2016)). In this report, we present an analysis of the election poll and result datasets in order to better understand the errors in the polls.

2 Datasets

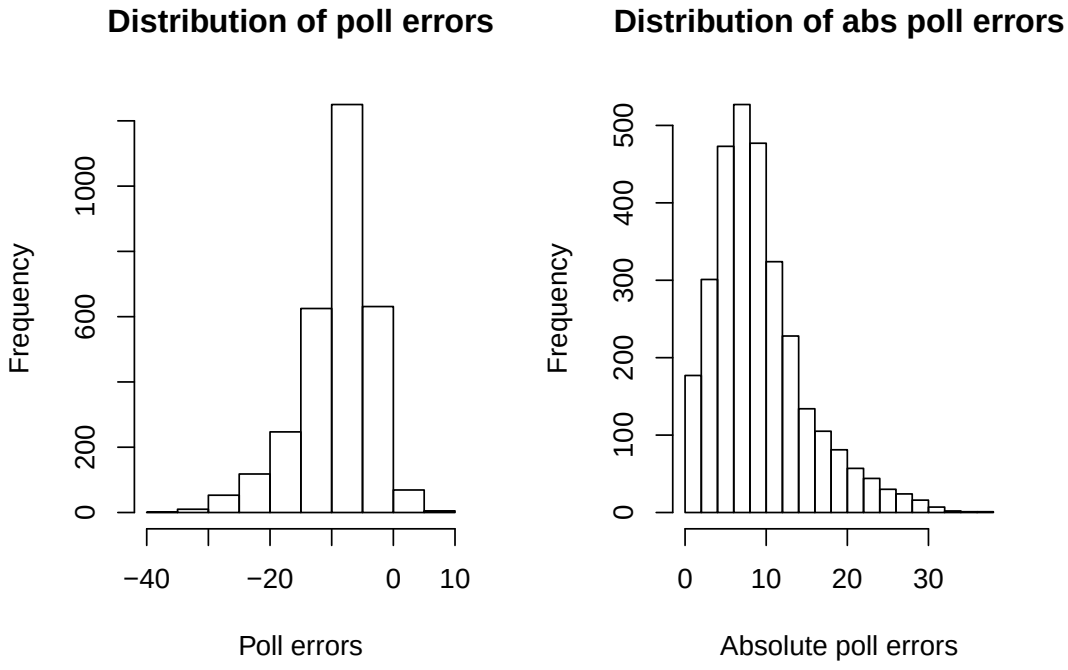
The analysis uses mainly two datasets, one for the polls and the other for the election results. The poll dataset is obtained from FiveThirtyEight (Silver (2016)). It consists of state level and national poll results conducted by media, universities, and other institutes. Beside the proportion of votes won by each candidate, the dataset also contains information such as the sample size and the date of the poll. The election result is contained in an unpublished dataset compiled by Emil O. W. Kirkegaard from the New York Times and a study of

inequality in the U. S. (Kirkegaard (2016)). It contains county-level election result as well as more than a hundred variables about each county's demography, economy, and other information. Both datasets are available to the public.

3 Analysis

3.1 Distribution of Errors

We are interested in how the poll's deviation from the election result is associated with other factors. In this analysis, we use state level data only. Thus we need to aggregate the county level election results and demographic variables into state level. After that, we compare the fraction of people that support Trump according to the poll with the fraction of voters that actually voted for Trump in each state. The distributions of the poll errors and their absolute values are shown in the histograms below:



About 97.5% polls underestimated Trump's performance, and the mean absolute error is 9.2 (percentage) points.

3.2 Linear Regression

Next, we conduct a linear regression of the errorrs. Since we are are interested in what makes the polls to be more wrong, the regression is done on the absolute errors. For the covariates, we choose to include the sample size of the poll, the number of days between the polling date and the voting date, the proportion of voters that support Trump in the poll, the region of the state (Northeast, South, North Central, and West), the proportion of voters in that state for Romney in 2012, and some other state level demographic variables. The result of the linear regression is shown below:

```
##
## Call:
## lm(formula = as.formula(paste0("poll.abserr~", paste(covariates,
##     collapse = "+"))), data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3965 -1.3934  0.0271  1.1494 10.6918
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                -6.836e+01  2.210e+00  -30.935  < 2e-16 ***
## poll.size                  -4.764e-04  6.062e-05   -7.859 5.35e-15 ***
## poll.rep.frac              -8.529e-01  6.587e-03 -129.490  < 2e-16 ***
## daystillvote                7.805e-04  6.366e-04    1.226   0.2202
## BachelorOrAbove            -4.307e-01  1.756e-02  -24.531  < 2e-16 ***
## MedianEarnings              7.773e-04  2.888e-05   26.911  < 2e-16 ***
## White                       1.390e-02  6.272e-03    2.217   0.0267 *
## Farming                     1.100e+00  1.068e-01   10.297  < 2e-16 ***
## PovertyRate                 7.742e-01  3.034e-02   25.515  < 2e-16 ***
## MedianAge                   1.349e+00  3.252e-02   41.470  < 2e-16 ***
## Unemployment               -2.658e+01  3.432e+00   -7.744 1.31e-14 ***
## ViolentCrime               -2.473e-03  4.438e-04   -5.573 2.72e-08 ***
## rep12.frac                  8.179e-01  1.045e-02   78.293  < 2e-16 ***
## state.regionSouth           2.702e-01  1.550e-01    1.744   0.0813 .
## state.regionNorth Central   3.758e+00  1.418e-01   26.494  < 2e-16 ***
## state.regionWest            3.408e-01  1.778e-01    1.917   0.0553 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

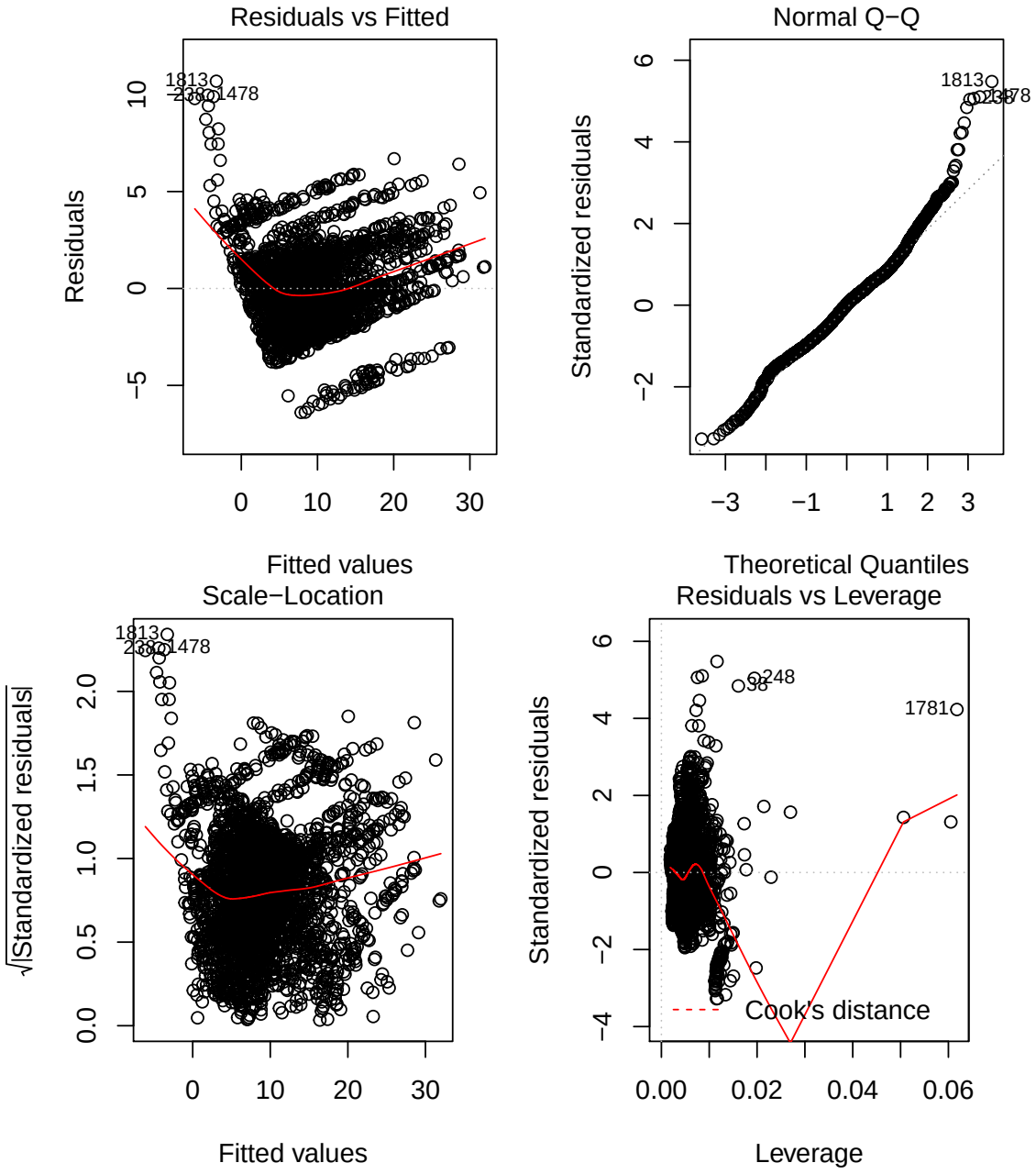
```
## Residual standard error: 1.963 on 2993 degrees of freedom
## Multiple R-squared:  0.8852, Adjusted R-squared:  0.8846
## F-statistic:  1539 on 15 and 2993 DF,  p-value: < 2.2e-16
```

The regression's r-squared value is 0.88, so our covariates contribute substantially to the variation in the absolute poll errors. Moreover, most covariates' effects are significant at $\alpha = 0.05$, alghough some of them have quite small magnitute of effect. The sample size of the poll has a negative effect on the absolute error, which is expected. The poll's fraction of Trump supporters also has a negative effect, but in Section 3.4 we will show some complication for this. The number of days between the poll and the election day, however, does not have a significant effect, so once the other covariates are held at constant, changing in the date of the poll does not chane the absolute poll error much.

As for the characteristics of the states, compared to the North, the poll's performance is not significantly different in the South, though it is significantly worse in the West and the worst in the Midwest (absolute error is 3.7 points higher). The states that had more voters for Romeny in 2012 have more inccorrect polls in 2016. In addition, the higher percentage of college-educated people, unemployment rate, and violent crime are associated with lower absolute error, while higher median income, median age, percentage of white people, percentage of people taking farming, fishing, and forresting occupations, and poverty rate are associated with greater absolute error.

3.3 Model Diaognistic

We now have a look of the residuals. The model diagnostic plots are shown below:



In the residuals vs fitted values plot, we can see that the residuals tend to fall on some "layers". We will discuss this further later. Moreover, the residuals have a parabolic shape, which suggests that the linear model might need some quadratic terms. Moreover, there is a group of outliers on the top left corner.

They also appear around the tail of the Q-Q plot. The Q-Q plot behaves quite well except for this area. Thus we look into these outlier polls in more details:

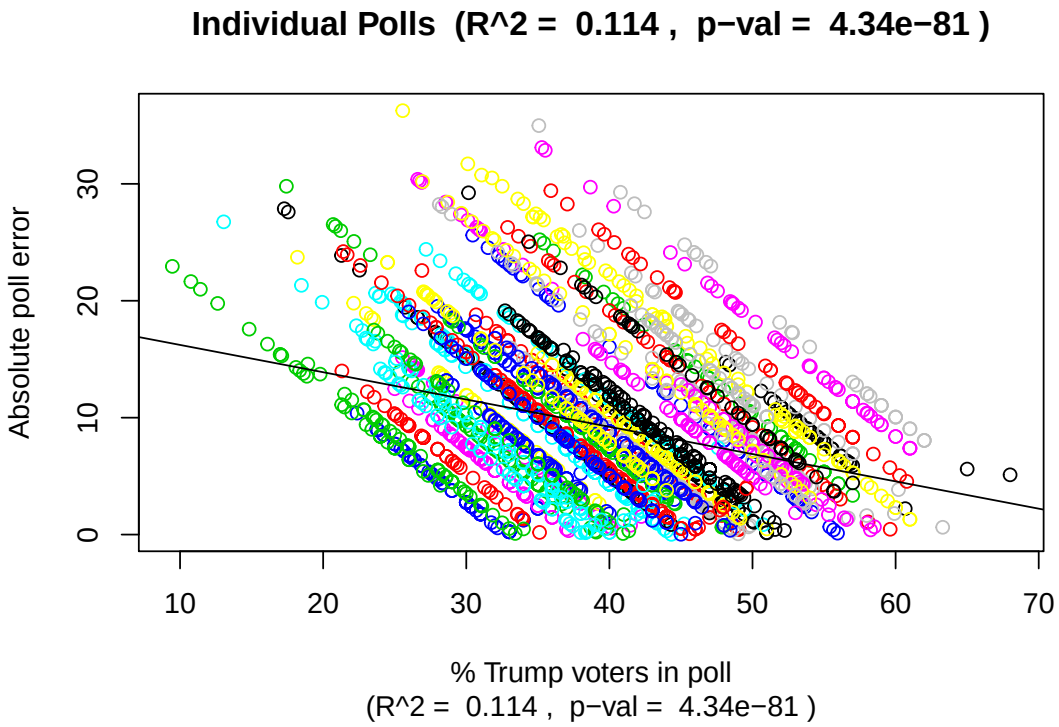
```
##      rstudent unadjusted p-value Bonferonni p
## 1813 5.505076         4.0022e-08   0.00012043
## 1478 5.124276         3.1765e-07   0.00095582
## 238  5.082003         3.9643e-07   0.00119290
## 248  5.060222         4.4407e-07   0.00133620
## 38   4.860400         1.2319e-06   0.00370670
## 1526 4.477626         7.8300e-06   0.02356100
```

```
##           state                                pollster daystillvote
## 1813 New Mexico                                   Ipsos           71
## 1478   Nebraska                         Emerson College           44
## 238  California Public Policy Institute of California          175
## 248    Colorado                  Quinnipiac University          361
## 38      Alabama                      Strategy Research          301
## 1526     Nevada                        Trafalgar Group            5
##      poll.rep.frac poll.err
## 1813         47.60 7.416133
## 1478         65.00 5.604726
## 238          39.00 6.240211
## 248          48.00 3.691615
## 38           68.00 5.111586
## 1526         49.61 4.078072
```

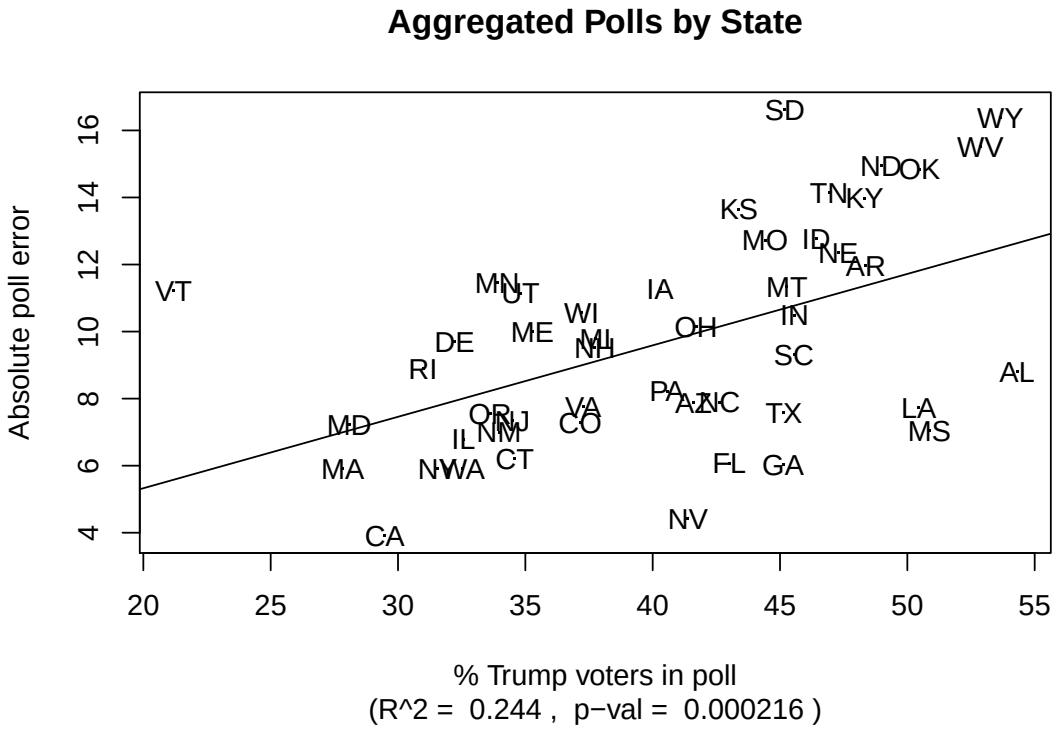
Here we see that the outliers are quite spread out in the covariates listed above. No obvious pattern is observed. Thus the reason they exist is inconclusive from the analysis.

3.4 Aggregated Polls

From the linear model, we observed that the percentage of Trump voters in the poll has a negative effect on the absolute poll error. To further investigate this relation, we plot these two variables.



Here the data points are colored by state. We see that each state has its distinct "layer". If we aggregate the individual polls by state, we get the following plot.



Here we can clearly see the pattern. On the national level, states with higher aggregated polled fraction of Trump voters have greater absolute poll error (and the signs of the errors are most likely to be negative), so the more Trump supporters a state has, the more the polls underestimate his performance. However, within each state, polls that favor Trump more give more accurate prediction. This motivated us to conduct a regression of the absolute poll errors only over the poll's percentage of Trump voters and the state in which the poll is carried out, as shown below:

```
##
## Call:
## lm(formula = poll.abserr ~ state + poll.rep.frac, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9919 -0.1865 -0.0749  0.0533 13.5817
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         61.277741   0.177622  344.990  < 2e-16 ***
## stateArizona       -13.204787   0.140171  -94.205  < 2e-16 ***
## stateArkansas       -2.573899   0.154150  -16.697  < 2e-16 ***
## stateCalifornia    -28.898113   0.152965 -188.920  < 2e-16 ***
## stateColorado      -18.104638   0.142987 -126.617  < 2e-16 ***
```

```
## stateConnecticut    -21.664333   0.162373 -133.423  < 2e-16 ***
## stateDelaware       -20.451780   0.167912 -121.801  < 2e-16 ***
## stateFlorida        -13.658141   0.127956 -106.741  < 2e-16 ***
## stateGeorgia        -11.656365   0.138049  -84.437  < 2e-16 ***
## stateIdaho           -3.673118   0.152387  -24.104  < 2e-16 ***
## stateIllinois       -23.020417   0.154685 -148.822  < 2e-16 ***
## stateIndiana         -6.766245   0.147566  -45.852  < 2e-16 ***
## stateIowa           -11.062111   0.143951  -76.846  < 2e-16 ***
## stateKansas          -5.714842   0.151544  -37.711  < 2e-16 ***
## stateKentucky        -0.616748   0.151087   -4.082 4.58e-05 ***
## stateLouisiana       -4.813519   0.149925  -32.106  < 2e-16 ***
## stateMaine          -17.204293   0.145126 -118.547  < 2e-16 ***
## stateMaryland       -26.928162   0.166107 -162.113  < 2e-16 ***
## stateMassachusetts  -28.473850   0.163756 -173.879  < 2e-16 ***
## stateMichigan       -15.013675   0.140930 -106.533  < 2e-16 ***
## stateMinnesota      -17.071086   0.159898 -106.762  < 2e-16 ***
## stateMississippi     -5.109488   0.155377  -32.884  < 2e-16 ***
## stateMissouri        -5.647096   0.142437  -39.646  < 2e-16 ***
## stateMontana         -6.198236   0.158716  -39.052  < 2e-16 ***
## stateNebraska        -3.248351   0.153639  -21.143  < 2e-16 ***
## stateNevada         -16.875767   0.136749 -123.407  < 2e-16 ***
## stateNew Hampshire  -15.300495   0.135903 -112.584  < 2e-16 ***
## stateNew Jersey     -20.622900   0.155597 -132.541  < 2e-16 ***
## stateNew Mexico     -21.463428   0.157843 -135.980  < 2e-16 ***
## stateNew York       -24.876766   0.153843 -161.702  < 2e-16 ***
## stateNorth Carolina -12.202956   0.130667  -93.390  < 2e-16 ***
## stateNorth Dakota     0.969549   0.170028    5.702 1.30e-08 ***
## stateOhio           -10.840717   0.132597  -81.757  < 2e-16 ***
## stateOklahoma         2.331338   0.154555   15.084  < 2e-16 ***
## stateOregon         -21.259881   0.156354 -135.973  < 2e-16 ***
## statePennsylvania   -13.866305   0.131987 -105.058  < 2e-16 ***
## stateRhode Island   -22.447946   0.182954 -122.697  < 2e-16 ***
## stateSouth Carolina  -7.964006   0.149962  -53.107  < 2e-16 ***
## stateSouth Dakota    -0.994885   0.162986   -6.104 1.17e-09 ***
## stateTennessee       -1.805066   0.153750  -11.740  < 2e-16 ***
## stateTexas          -10.081881   0.146652  -68.747  < 2e-16 ***
## stateUtah           -16.501398   0.149616 -110.291  < 2e-16 ***
## stateVermont        -29.596730   0.184592 -160.336  < 2e-16 ***
## stateVirginia       -17.499092   0.140129 -124.878  < 2e-16 ***
## stateWashington     -23.939425   0.159103 -150.465  < 2e-16 ***
## stateWest Virginia    5.310973   0.151833   34.979  < 2e-16 ***
## stateWisconsin      -14.762013   0.142946 -103.270  < 2e-16 ***
## stateWyoming          6.947645   0.172843   40.196  < 2e-16 ***
## poll.rep.frac        -0.965964   0.002572 -375.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7193 on 2960 degrees of freedom
## Multiple R-squared:  0.9848, Adjusted R-squared:  0.9845
## F-statistic:  3984 on 48 and 2960 DF,  p-value: < 2.2e-16
```

Here all the covariates have extremely low p-values, and the r-squred value is as high as 0.98. This is a convincing evidence for the relation between what the poll predicts and how accurate the poll predicts.

4 Discussion

In this analysis, we observed that by using the attributes of the poll and the characteristics of the state, we can gain substatial information about the absolute poll error. Some covariates make the poll more accurate as they increase, while others make the poll more accurate as they decrease. The question is what determines whether a covariate is a booster or a inhibitor of poll accuracy?

From a survey perspective, this relation may be strongly associated with the sample nonresponse bias. If we see each demographic covariate as a strata, then the sample's representativeness of the population highly depends on how willing people within this strata are to response to the poll, provided that poll is reaching different people groups uniformly. (If the poll is not well-designed and its reaching of people is biased, which is also possible, then the sample's bias will be even greater.) For example, as proposed in Mercer, Deane, & Mcgeeney (2016), many Trump supporters have anti-institutional feelings, and this factor may cause them to be more annoyed by phone calls from pollsters and less likely to respond to polls. This will make it hard for the polls to reach these Trump supporters and thus underestimate Trump's popularity in a state. In our analysis, the percentage of white people has a positive effect on the absolute poll error. Since from exit polls we know that 58% white people voted for Trump (New York Times (2017)), white people might be underrepresented by poll surveys. On the other hand, the percentage of population with at least a bachelor's degree has a negative effect on the absolute poll error. This suggests that college graduates may be more willing to share their opnions to pollsters.

On the other hand, the regression of absolute poll errors over states and

the predicted Trump's performance is very informative. The analysis shows that the better Trump performs in a state based on aggregated polls, the more he will outperforms the polls in the actual election. Moreover, the effect of Trump's predicted performance on the absolute poll error is surprisingly close within each state, as shown in the plot in Section 3.4. This suggests that there may be some systematic bias in the polls. Further investigation on finer-level poll datasets is needed to reveal the deeper problems in the polling system for the 2016 presidential election.

References

Bialik, C. and Enten, L. The Polls Missed Trump. We Asked Pollsters Why. FiveThirtyEight (2016).

Kirkegaard, E. O. USA County Data. (2016). Unpublished dataset. Retrieved from https://github.com/Deleetdk/USA.county.data.

Kirkegaard, E. O. Inequality across US counties: an S factor analysis. (2016).

Mercer, A., Deane, C., and Mcgeeney, K. Why 2016 election polls missed their mark. Pew Research Center (2016).

New York Times. Presidential Election Results: Donald J. Trump Wins. NYTIMES.COM (2017).

Silver, N. Presidential General Polls, 2016. FiveThirtyEight (2016). Retrieved from https://projects.fivethirtyeight.com/2016-election-forecast.

Stein, J.. "7 experts try to explain how the polls missed Donald Trump's victory". Vox (2016).

A R codes

Below is the R codes that are used for analyzing the datasets.

```
library(MASS)
library(car)


# vote result dataset
load("~/data/presidential_2016/USA.county.data/data/USA_county_data.RData")
# Selection variables to be included
to.be.weighted <- c(
                     "At.Least.Bachelor.s.Degree",
                     "Median.Earnings.2010.dollars",
                     "White",
                     "Farming.fishing.and.forestry.occupations",
                     "Poverty.Rate.below.federal.poverty.threshold",
                     "median_age",
                     "Unemployment",
                     "Violent.crime"
                     )
to.be.summed <- c("Total.Population",
                   "votes16_trumpd",  "total16",
                   "rep12", "total12")
vote <- USA_county_data[, c("State", to.be.summed, to.be.weighted)]


to.be.weighted <- c("BachelorOrAbove", "MedianEarnings",
                     "White", "Farming", "PovertyRate",
                     "MedianAge", "Unemployment", "ViolentCrime")
```

```r
colnames(vote)[c(7:(7+length(to.be.weighted)-1))] <- to.be.weighted


# Clean out unwanted observations
vote <- vote[complete.cases(vote),]
vote <- vote[(vote$State != "Hawaii" &
                        vote$State != "Alaska" &
                        vote$State != "District of Columbia"),]
vote$State <- as.factor(vote$State)




# Add state population to each county
stpop <- aggregate(vote$Total.Population,
                   by = list(State = vote$State), FUN = sum)
colnames(stpop)[2] = "st.pop"
vote <- merge(vote, stpop, by = "State")
# Weight the variables
vote$wt <- vote$Total.Population / vote$st.pop
vote.wt <- vote
vote.wt[,to.be.weighted] <- vote.wt[,to.be.weighted] * vote.wt$wt




# Aggregate by state
vote.agg <- aggregate(
  vote.wt[,!(names(vote.wt) %in% c("State", "st.pop", "wt"))],
                      by = list(State=vote.wt$State), FUN = sum)
```

```r
vote.agg$rep16.frac <- vote.agg$votes16_trumpd / vote.agg$total16 * 100
vote.agg$rep12.frac <- vote.agg$rep12 / vote.agg$total12 * 100
vote.agg$turnout16 <- vote.agg$total16 / vote.agg$Total.Population
# Make vote.agg compatible to poll.agg
colnames(vote.agg)[1] = "state"
vote.agg$state <- as.character(vote.agg$state)



# poll dataset
poll <- read.csv("~/data/presidential_2016/silver_poll.csv")



# Convert poll date to days before the voting day
poll$middledate <- (as.Date(poll$enddate, "%m/%d/%Y")
                    - as.Date(poll$startdate, "%m/%d/%Y")) / 2
+ as.Date(poll$startdate, "%m/%d/%Y")
poll$daystillvote <- as.numeric(as.Date(poll$forecastdate, "%m/%d/%y")
                                - as.Date(poll$middledate, "%m/%d/%Y"))
poll$poll.rep.frac <- poll$rawpoll_trump
# Clean out the unwanted observations
poll <- poll[poll$type == "polls-only",]
poll <- poll[,c("state", "samplesize", "poll.rep.frac",
                "daystillvote", "pollster")]
poll <- poll[complete.cases(poll),]
poll <- poll[(poll$state != "Hawaii" &
                        poll$state != "Alaska" &
                        poll$state != "U.S." &
```

```
                        poll$state != "District of Columbia"),]
poll[(poll$state=="Nebraska CD-1"|
          poll$state=="Nebraska CD-2"|
          poll$state=="Nebraska CD-3"),]$state <- "Nebraska"
poll[(poll$state=="Maine CD-1"|
          poll$state=="Maine CD-2"|
          poll$state=="Maine CD-3"),]$state <- "Maine"
poll$state <- as.character(poll$state)
colnames(poll)[2] <- "poll.size"



# Combine poll result and vote result
merged <- merge(poll, vote.agg, by = "state")
merged$poll.err <- merged$poll.rep.frac - merged$rep16.frac
merged$poll.abserr <- abs(merged$poll.err)

# Add state region and state abbreviations to dataset
data(state)
state <- state.name
stregion <- data.frame(state, state.region, state.abb)
merged <- merge(merged, stregion, by = "state")

# Linear regression
covariates <- colnames(merged)[!(colnames(merged) %in% c("state",
                                      "Total.Population", "votes16_trumpd",
                                      "total16", "rep12", "total12",
```

```
                                        "rep16.frac", "poll.abserr", "pollster",
                                        "state.abb", "poll.err",
                                        "turnout16"
                                        ))]
fit <- lm(as.formula(paste0("poll.abserr~",paste(covariates, collapse = "+"))),
          data = merged)
poll.abserr.st <- aggregate(merged$poll.abserr,
                            by = list(state.abb=merged$state.abb), FUN = mean)
poll.repfrac.st <- aggregate(merged$poll.rep.frac,
                             by = list(state.abb=merged$state.abb), FUN = mean)
poll.abserr.st <- merge(poll.abserr.st,
                        poll.repfrac.st, by = "state.abb")

# Distribution of erros
par(mfrow = c(1,2))
hist(merged$poll.err,
     main = "Distribution of poll errors", xlab = "Poll errors")
hist(merged$poll.abserr,
     main = "Distribution of abs poll errors", xlab = "Absolute poll errors")

# Linear regression of absolute error
# over poll and state attributes
print(summary(fit))

# Model diagnostic
par(mfrow = c(1,2))
plot(fit)
```

```r
# Show outliers
outlierTest(fit)
idx <- as.numeric(names(outlierTest(fit)$rstudent))
merged[idx, c("state", "pollster", "daystillvote",
              "poll.rep.frac", "poll.err")]



# Plot abs error vs percentage of Trump supports in the poll
fit.ind <- lm(merged$poll.abserr ~ merged$poll.rep.frac)
plot(merged$poll.rep.frac, merged$poll.abserr,
     main =  paste(
       "Individual Polls ",
       "(R^2 = ",
       format(summary(fit.ind)$adj.r.squared, digits=3),
       ", ",
       "p-val = ",
       format(summary(fit.ind)$coefficients[2,4], digits=3),
       ")"
       ),
     xlab = "% Trump voters in poll",
     ylab = "Absolute poll error",
     col = as.numeric(as.factor(merged$state))
)
title(sub = paste(
       "(R^2 = ",
       format(summary(fit.ind)$adj.r.squared, digits=3),
```

```
        ", ",
        "p-val = ",
        format(summary(fit.ind)$coefficients[2,4], digits=3),
        ")"
        )
      )
abline(fit.ind)



# Linear regression of error over
# state and percentage of Trump supporters in the poll
fit.simple <- lm(poll.abserr ~ state + poll.rep.frac, data = merged)
summary(fit.simple)
```