Electoral Polls and Prediction

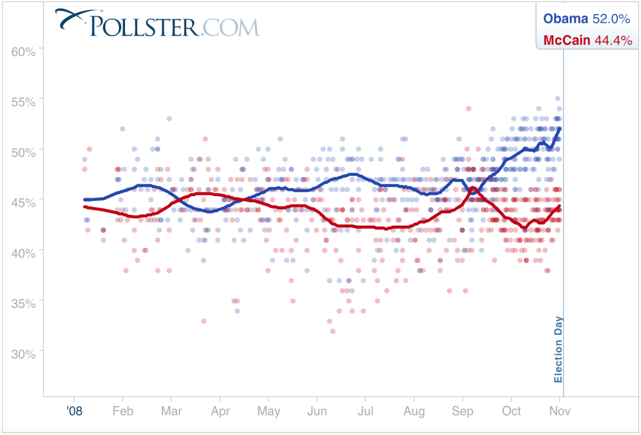
Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

2008 US Presidential Election

- A historic election ⇝ first African-American president
- Barack Obama won 52.9% of the national votes while McCain won 45.7%

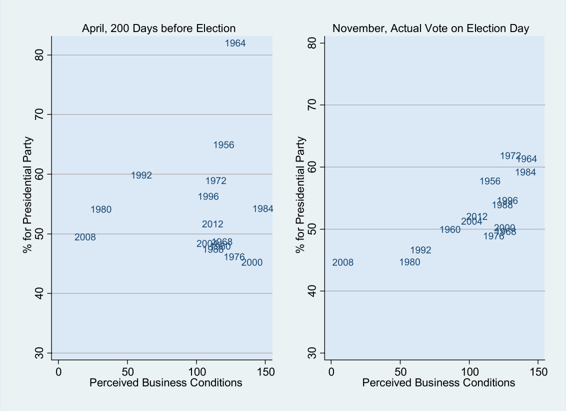


- Polls fluctuate early

How Should We “Forecast” the Election Results?

- Macro political and economic fundamentals for early forecasting

Predicting the Polls in April and the Vote in November from Economic Perceptions at Each Point in Time

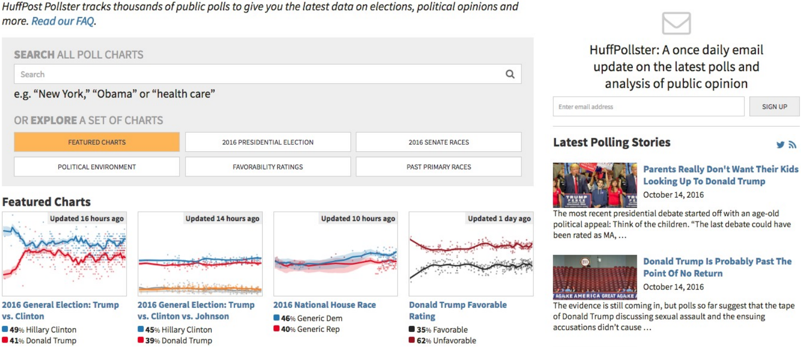


Economic perceptions increasingly predict the vote over the campaign timeline.

- Recent method: combine them with polls

Let's Analyze Some Polls

- R package **pollstR** scrapes the data from Huffington Post:



HuffPost Pollster tracks thousands of public polls to give you the latest data on elections, political opinions and more. Read our FAQ.

SEARCH ALL POLL CHARTS

Search

e.g. "New York," "Obama" or "health care"

OR EXPLORE A SET OF CHARTS

FEATURED CHARTS | 2016 PRESIDENTIAL ELECTION | 2016 SENATE RACES

POLITICAL ENVIRONMENT | FAVORABILITY RATINGS | PAST PRIMARY RACES

Featured Charts

Updated 16 hours ago
2016 General Election: Trump vs. Clinton
■ 49% Hillary Clinton
■ 41% Donald Trump

Updated 14 hours ago
2016 General Election: Trump vs. Clinton vs. Johnson
■ 45% Hillary Clinton
■ 39% Donald Trump

Updated 10 hours ago
2016 National House Race
■ 46% Generic Dem
■ 40% Generic Rep

Updated 1 day ago
Donald Trump Favorable Rating
■ 35% Favorable
■ 62% Unfavorable

HuffPollster: A once daily email update on the latest polls and analysis of public opinion

Enter email address | SIGN UP

Latest Polling Stories

Parents Really Don't Want Their Kids Looking Up To Donald Trump
October 14, 2016
The most recent presidential debate started off with an age-old political appeal: Think of the children. "The last debate could have been rated as MA, …

Donald Trump Is Probably Past The Point Of No Return
October 14, 2016
The evidence is still coming in, but polls so far suggest that the tape of Donald Trump discussing sexual assault and the ensuing accusations didn't cause …

```r
library(pollstR)
chart_name <- "2016-general-election-trump-vs-clinton"
polls2016 <-
    pollster_charts_polls(chart_name)[["content"]]

## Warning:  replacing previous import
```

```
polls2016 <- as.data.frame(polls2016)
names(polls2016)
```

```
##  [1] "Trump"                  "Clinton"
##  [3] "Other"                  "Undecided"
##  [5] "poll_slug"              "survey_house"
##  [7] "start_date"             "end_date"
##  [9] "question_text"          "sample_subpopulation"
## [11] "observations"           "margin_of_error"
## [13] "mode"                   "partisanship"
## [15] "partisan_affiliation"
```

```
polls2016[1:3, c("Trump", "Clinton", "start_date", "end_date")]
```

```
##   Trump Clinton start_date   end_date
## 1    43      46 2016-11-04 2016-11-06
## 2    39      44 2016-11-02 2016-11-06
## 3    43      47 2016-11-02 2016-11-06
```

Plotting Polls over Time

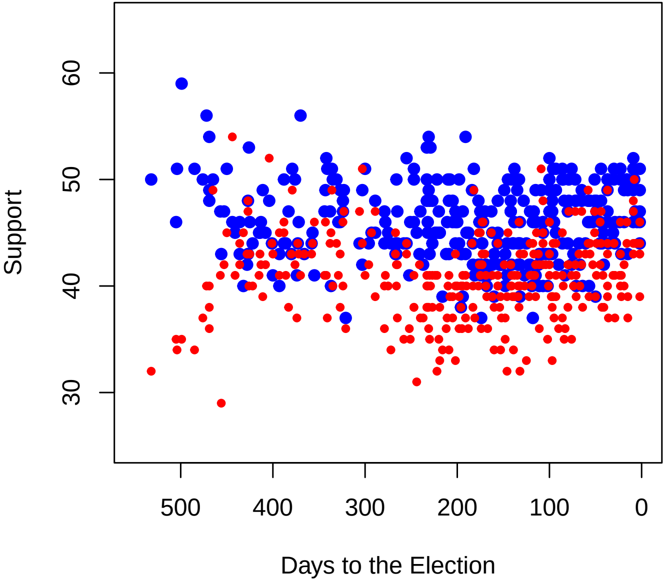
- Compute the days to the election variable:

```
class(polls2016$end_date)

## [1] "Date"

polls2016$DaysToElection <-
    as.Date("2016-11-8") - polls2016$end_date
```

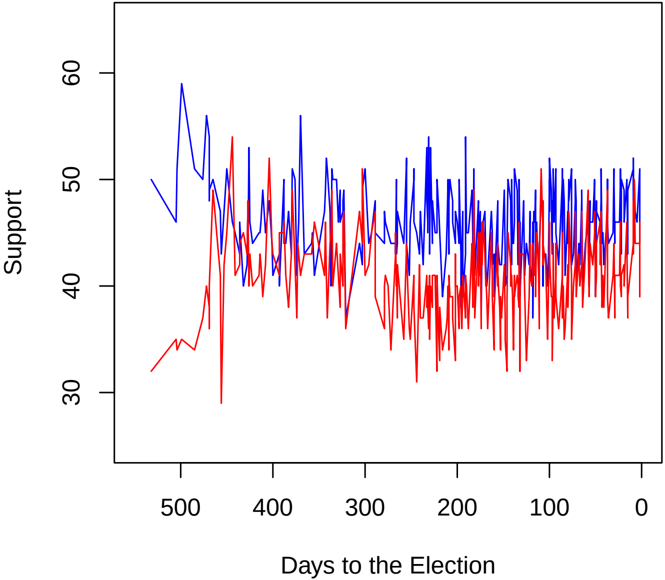
- Plot polling results:

```
plot(polls2016$DaysToElection, polls2016$Clinton,
     xlab = "Days to the Election", ylab = "Support",
     xlim = c(550, 0), ylim = c(25, 65), pch = 19,
     col = "blue")
points(polls2016$DaysToElection, polls2016$Trump,
       pch = 20, col = "red")
```

What's Wrong with this Plot?



Time-Series Plot Looks Even Worse



Smoothing over Time

- Moving average: average polls within a one-week period
- For example, on October 17, we will take all polls conducted within the past week
- Window size: amount of smoothing

- Coding strategy: for each day, we subset the relevant polls and compute the average
- Range of the `DaysToElection` variable:

```r
range(polls2016$DaysToElection)
```

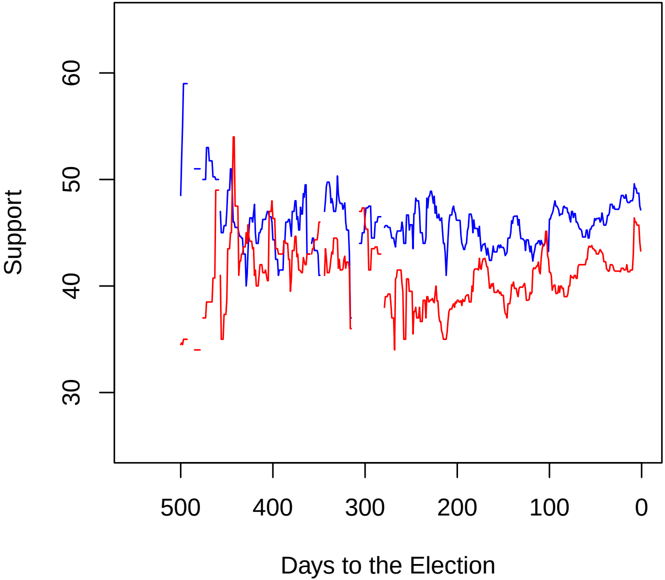
```
## Time differences in days
## [1]   2 532
```

Plotting US Presidential Election Polls over Time

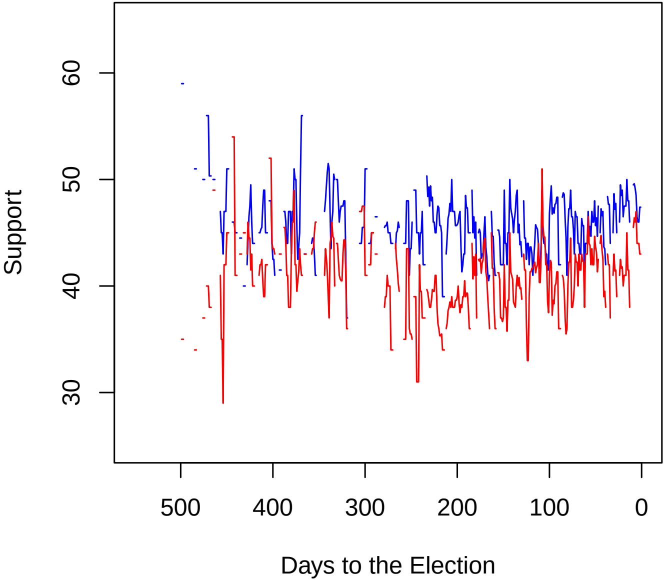
```r
window <- 7
days <- 500:1
```

```r
Clinton.pred <- Trump.pred <- rep(NA, length(days))
for (i in 1:length(days)) {
    week.data <-
        subset(polls2016,
               subset = ((DaysToElection < (days[i] + window))
                   & (DaysToElection >= days[i])))
    Clinton.pred[i] <- mean(week.data$Clinton)
    Trump.pred[i] <- mean(week.data$Trump)
}
plot(days, Clinton.pred, type = "l", col = "blue",
     xlab = "Days to the Election", ylab = "Support",
     xlim = c(550, 0), ylim = c(25, 65))
lines(days, Trump.pred, col = "red")
```

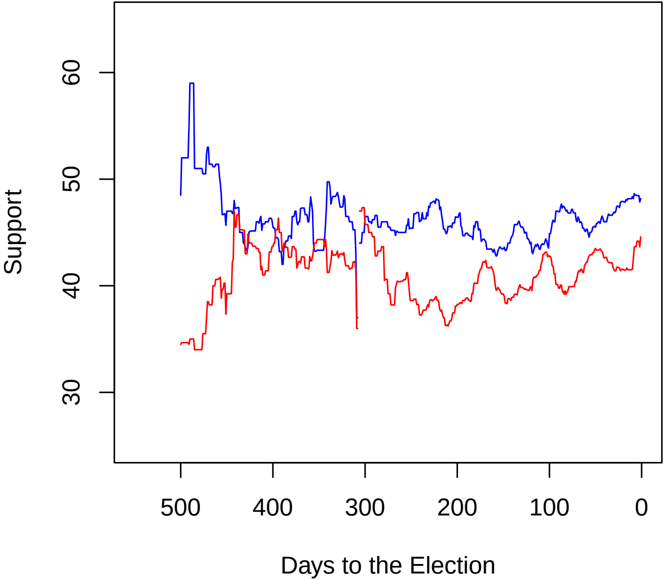
1-Week Moving Average



3-Day Moving Average



2-Week Moving Average



Let's Add Some Informative Labels

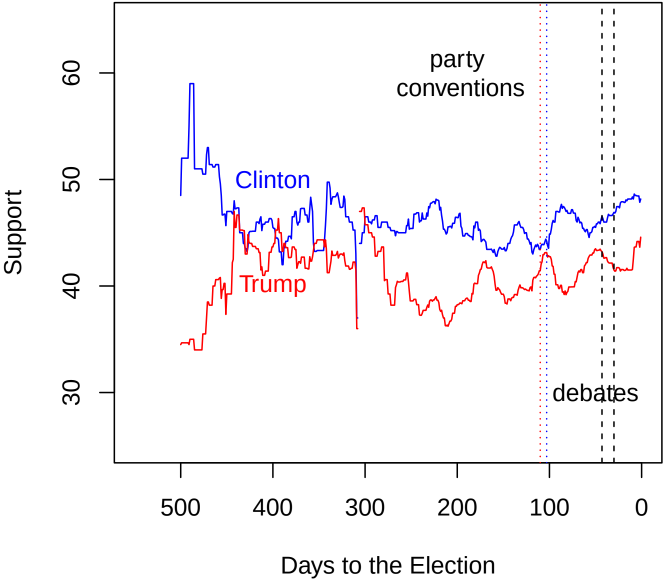
- Candidate names:

```r
text(400, 50, "Clinton", col = "blue")
text(400, 40, "Trump", col = "red")
```

- Events:

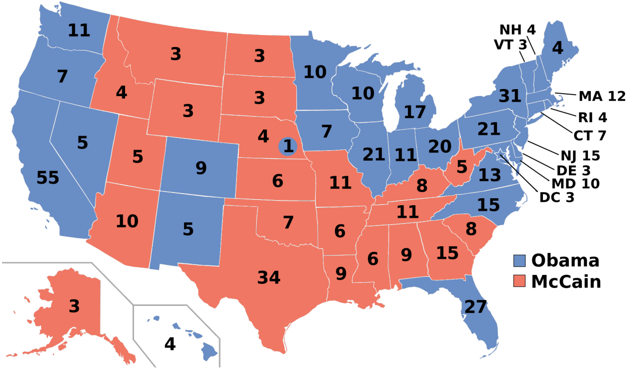
```r
text(200, 60, "party\n conventions")
abline(v = as.Date("2016-11-8") - as.Date("2016-7-28"),
       lty = "dotted", col = "blue")
abline(v = as.Date("2016-11-8") - as.Date("2016-7-21"),
       lty = "dotted", col = "red")
text(50, 30, "debates")
abline(v = as.Date("2016-11-8") - as.Date("2016-9-26"),
       lty = "dashed")
abline(v = as.Date("2016-11-8") - as.Date("2016-10-9"),
       lty = "dashed")
```

The Final Graph: 1-week Moving Average



Predicting US Presidential Election

- Electoral college system
  - must win an absolute majority of 538 electoral votes
  - 538 = 435 (House of Representatives) + 100 (Senators) + 3 (DC)
  - must win at least 270 votes
  - nobody wins an absolute majority ⇝ House of representatives
- Must predict the winner of each state



Poll Prediction for the 2008 Election

- Election data: pres08.csv

Name	Description
state	abbreviated name of state
state.name	unabbreviated name of state
Obama	Obama's vote share (percentage)
McCain	McCain's vote share (percentage)
EV	number of electoral college votes for the state

- Polling data: polls08.csv

Name	Description
state	abbreviated name of state in which poll was conducted
Obama	predicted support for Obama (percentage)
McCain	predicted support for McCain (percentage)
Pollster	name of organization conducting poll
middate	middate of the period when poll was conducted

- Predict the state-level support for each candidate using polls
- Allocate the electoral college votes of that state to its predicted winner
- Aggregate the electoral college votes across states to determine the predicted winner
- Repeat this on each date

- Coding strategy: for any given date, do the following
  1. For each state, subset the polls conducted within it
  2. Further subset the latest polls (there may be multiple polls conducted on the same day)
  3. Aaverage the latest polls to estimate the support for each candidate
  4. Allocate the electoral votes to the candidate who has greater support
  5. Repeat this for all states and aggregate the electoral votes

Some Preprocessing

```r
## election results, by state
pres08 <- read.csv("data/pres08.csv")
## polling data
polls08 <- read.csv("data/polls08.csv")
## Obama's margin
polls08$margin <- polls08$Obama - polls08$McCain
pres08$margin <- pres08$Obama - pres08$McCain
## convert to a Date object
polls08$middate <- as.Date(polls08$middate)
## number of days to the election day
polls08$DaysToElection <- as.Date("2008-11-04") -
    polls08$middate
```

Poll Prediction for Each State

```r
poll.pred <- rep(NA, 51) # initialize a vector place holder
## state names which the loop will iterate through
st.names <- unique(polls08$state)
## add labels for easy interpretation later on
names(poll.pred) <- as.character(st.names)
## loop across 50 states plus DC
for (i in 1:51){
    ## subset the ith state
    state.data <- subset(polls08,
                         subset = (state == st.names[i]))
    ## subset the latest polls within the state
    latest <- state.data$DaysToElection ==
        min(state.data$DaysToElection)
    ## compute the mean of latest polls and store it
    poll.pred[i] <- mean(state.data$margin[latest])
}
```

- prediction error = actual outcome − predicted outcome

```r
errors <- pres08$margin - poll.pred
names(errors) <- st.names # add state names
```

- Mean prediction error

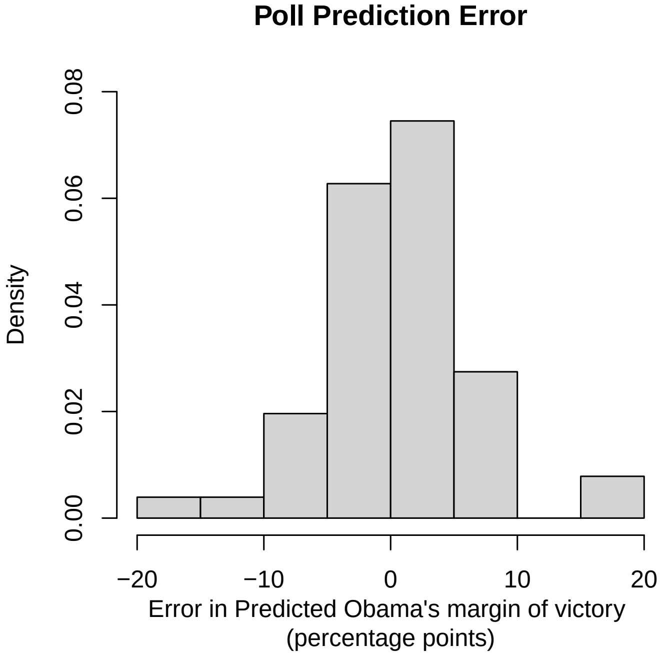
```r
mean(errors) # mean prediction error

## [1] 1.06
```

- Root mean squared error

```r
sqrt(mean(errors^2))

## [1] 5.91
```

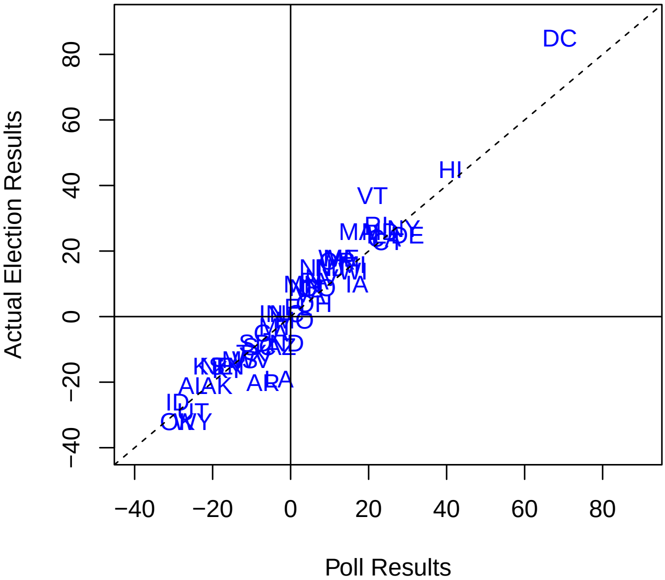
- Histogram

```r
hist(errors, freq = FALSE, ylim = c(0, 0.08),
     main = "Poll Prediction Error",
     xlab = "Error in Predicted Obama's margin of victory
(percentage points)")
```


Poll Prediction Error
Density
0.00
0.02
0.04
0.06
0.08
−20
−10
0
10
20
Error in Predicted Obama's margin of victory
(percentage points)

State by State Prediction Error

```r
## type = "n" generates "empty" plot
plot(poll.pred, pres08$margin, type = "n", main = "",
     xlim = c(-40, 90), ylim = c(-40, 90),
     xlab = "Poll Results", ylab = "Actual Election Results")
## add state abbreviations
text(poll.pred, pres08$margin, pres08$state, col = "blue")
## lines
abline(a = 0, b = 1, lty = "dashed") # 45 degree line
abline(v = 0)  # vertical line at 0
abline(h = 0)  # horizontal line at 0
```


DC
HI
VT
RI
NY
MA
DE
CT
NM
WI
IA
OH
IN
MO
SD
AZ
KS
WV
AL
AK
AR
LA
ID
UT
OK
WY
80
60
40
20
0
−20
−40
Actual Election Results
−40
−20
0
20
40
60
80
Poll Results