STAT540 - Project 2

Jesse Claiborne III

2023-10-22

Code

Hide

```r
# load the packages for graphing and data wrangling
library(ggplot2)
library(dplyr)
```

Note: If you `Rmd` file submission knits you will receive total of **(10 points)**.

For the data wrangling use function from the `dplyr` package

Project Objectives?

Leading up to the 2016 presidential election, many pollsters predicted that the Democratic candidate, Hillary Clinton, would win a "decisive victory.". However, as we all know, the election was won by the Republican candidate, and current president, Donald Trump. In general biases, not accounted for by prediction models, often affect many pollsters. In this project, you are going to further investigate these biases through comparisons across both national and state-level races.

The project requires an **.RData** file, `election_polls.RData`, containing a `data.frame` (`polls`) with several years worth of polling data (2008, 2010, 2012, 2014 and 2016). The polls cover federal elections for house representatives, senators and the president, and includes polling data from up to a year before the election date.

Hide

```r
library(tidyverse)
load("elections_polls.RData")
str(polls)
```

```
'data.frame':   6847 obs. of  16 variables:
 $ race            : chr  "2016_Pres_NM" "2016_Pres_VA" "2016_Pres_IA" "2016_Pres_WI" ...
 $ race_state      : chr  "2016_Pres_NM" "2016_Pres_VA" "2016_Pres_IA" "2016_Pres_WI" ...
 $ state           : chr  "NM" "VA" "IA" "WI" ...
 $ state_long      : chr  "new mexico" "virginia" "iowa" "wisconsin" ...
 $ type            : chr  "Pres" "Pres" "Pres" "Pres" ...
 $ year            : num  2016 2016 2016 2016 2016 ...
 $ pollster        : Factor w/ 636 levels "ABC News/Washington Post",..: 195 130 148 95 149 87 132 132 1 65 ...
 $ samplesize      : num  8439 1238 800 1255 800 ...
 $ startdate       : Date, format: "2016-11-06" "2016-11-03" "2016-11-01" "2016-10-26" ...
 $ enddate         : Date, format: "2016-11-06" "2016-11-04" "2016-11-04" "2016-10-31" ...
 $ democrat_name   : chr  "clinton" "clinton" "clinton" "clinton" ...
 $ democrat_poll   : num  46 48 39 46 44 46 46 47 48 44 ...
 $ democrat_result : num  48.3 49.8 41.7 46.5 46.2 45.9 47.8 46.2 49.8 45.9 ...
 $ republican_name : chr  "trump" "trump" "trump" "trump" ...
 $ republican_poll : num  44 43 46 40 44 49 45 45 42 48 ...
 $ republican_result: num  40 44.4 51.1 47.2 49.8 51 49 49.8 44.4 51 ...
```

The `polls` `data.frame` contains the following columns:

- `race` : race identifier year_electiontype_location.
- `race_state` : race identifier year_electiontype_state. In contrast to the previous column, this identifier ignores information about counties and only contains information at the state level.
- `state` : abbreviation of state of the election
- `state_long` : full name of the state
- `type` : type of race. Could be either presidential (Pres), senatorial election (Sen-G) or house representative election (House-G).
- `year` : election year
- `pollster` : name of the pollster
- `samplesize` : size of the sample used in the poll
- `startdate` : start date of the pole. If this date was not available, this will be the same as enddate
- `enddate` : end date of the pole
- `democrat_name` : name of the democratic candidate
- `democrat_poll` : percentage of people from the poll saying they would vote for the democratic candidate
- `democrat_result` : actual percentage of people voting for the democratic candidate in the election
- `republican_name` : name of the republican candidate
- `republican_poll` : percentage of people from the poll saying they would vote for the republican candidate
- `republican_result` : actual percentage of people voting for the republican candidate in the election

Part 1 (10 pts)

Subset the `polls` `data.frame` to only keep polls which ended within approximately 6 weeks preceding any [Election Day (i.e. in October or November, 10th and 11th months).

Hint: you might need to extract the month from the `enddate`. The `strftime` function might be useful for this.

Solution:

Hide

```r
polls <- polls %>%
mutate( endmonth = strftime( polls$enddate, "%m") ) %>%
filter( endmonth %in% c("10", "11") )
polls
```

	race	race_state	state	state_long	type	year
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
9	2016_Pres_NM	2016_Pres_NM	NM	new mexico	Pres	2016
14	2016_Pres_VA	2016_Pres_VA	VA	virginia	Pres	2016
16	2016_Pres_IA	2016_Pres_IA	IA	iowa	Pres	2016
18	2016_Pres_WI	2016_Pres_WI	WI	wisconsin	Pres	2016
19	2016_Pres_NC	2016_Pres_NC	NC	north carolina	Pres	2016
20	2016_Pres_GA	2016_Pres_GA	GA	georgia	Pres	2016
21	2016_Pres_FL	2016_Pres_FL	FL	florida	Pres	2016
22	2016_Pres_NC	2016_Pres_NC	NC	north carolina	Pres	2016
23	2016_Pres_VA	2016_Pres_VA	VA	virginia	Pres	2016
25	2016_Pres_GA	2016_Pres_GA	GA	georgia	Pres	2016

1-10 of 4,330 rows | 1-7 of 17 columns Previous 1 2 3 4 5 6 ... 100 Next

Part 2 (10 pts)

For each poll, calculate the difference between the fraction of people saying they would vote for the Republican Party and the fraction of people saying they would vote for the Democratic Party. Add these values to your `data.frame` as a new column, `spread`. Similarly, calculate the true (actual) difference between the fraction of people who ended up voting for the Republican Party and the fraction of people who ended up voting for the Democratic Party. Create new variable `spread_act` by adding the true (actual) difference, to your `data.frame`.

Solution:

Hide

```r
polls <- polls %>%
  mutate( spread = republican_poll/100 - democrat_poll/100,
          spread_act = republican_result/100 - democrat_result/100 )
polls
```

	race	race_state	state	state_long	type	year
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
9	2016_Pres_NM	2016_Pres_NM	NM	new mexico	Pres	2016
14	2016_Pres_VA	2016_Pres_VA	VA	virginia	Pres	2016
16	2016_Pres_IA	2016_Pres_IA	IA	iowa	Pres	2016
18	2016_Pres_WI	2016_Pres_WI	WI	wisconsin	Pres	2016
19	2016_Pres_NC	2016_Pres_NC	NC	north carolina	Pres	2016
20	2016_Pres_GA	2016_Pres_GA	GA	georgia	Pres	2016
21	2016_Pres_FL	2016_Pres_FL	FL	florida	Pres	2016
22	2016_Pres_NC	2016_Pres_NC	NC	north carolina	Pres	2016
23	2016_Pres_VA	2016_Pres_VA	VA	virginia	Pres	2016
25	2016_Pres_GA	2016_Pres_GA	GA	georgia	Pres	2016

1-10 of 4,330 rows | 1-7 of 19 columns Previous 1 2 3 4 5 6 ... 100 Next

Part 3 (10 pts)

Now collapse polls for each race. For this, group polls by the type, year, and state of the corresponding election. There are several polls for each race, and each one provides an approximation of the real θ value. Generate a point estimate for each race, $\hat{\theta}$, that summarizes the polls for that race using the following steps: [1] use the column `race_state` to group polls by type, year, and state, and [2] use the `summarize` function to generate a new `data.frame` called `reduced_polls` with the following columns:

1. the mean `spread`,
2. the standard deviation of the `spread`,
3. the mean `spread_act`, and
4. the number of polls per race.

Make sure you also keep information about the `year` and `state` of each race in this new `data.frame`.

Solution:

Hide

```r
reduced_polls <- polls %>%
  group_by( race_state ) %>%
  summarize(avg = mean( spread ),
            act = mean( spread_act ),
            sd = sd( spread ),
            year = unique( year ),
            state = unique( state ),
            state_long = unique( state_long ),
            type = unique( type ),
            n=n() )
reduced_polls
```

race_state	avg	act	sd	year	state	state_long	type	n
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<int>
2008_House-G_AK	-0.0725000000	0.051700000	1.500000e-02	2008	AK	alaska	House-G	4
2008_House-G_AL	-0.0250000000	-0.006200000	6.363961e-02	2008	AL	alabama	House-G	2
2008_House-G_AZ	0.1000000000	0.120100000	NA	2008	AZ	arizona	House-G	1
2008_House-G_CA	-0.0150000000	-0.011325000	8.812869e-02	2008	CA	california	House-G	4
2008_House-G_CT	-0.0300000000	-0.008100000	NA	2008	CT	connecticut	House-G	1
2008_House-G_FL	0.0350000000	0.089450000	1.513747e-01	2008	FL	florida	House-G	8
2008_House-G_GA	-0.1050000000	-0.144800000	9.192388e-02	2008	GA	georgia	House-G	2
2008_House-G_IA	0.1350000000	0.211400000	1.202082e-01	2008	IA	iowa	House-G	2
2008_House-G_ID	0.1033333333	0.131833333	2.916048e-01	2008	ID	idaho	House-G	3
2008_House-G_IL	-0.0850000000	-0.094000000	1.144552e-01	2008	IL	illinois	House-G	4

1-10 of 423 rows Previous 1 2 3 4 5 6 ... 43 Next

Part 4 (10 pts)

Note that the previous question merges different congressional elections held in the same year across districts in a state. Thus, using the collapsed `data.frame` from the previous question, filter out races from congressional elections. Also, **filter** out races that had less than 3 polls. For each remaining races, build a 95% confidence interval for $\hat{\theta}$. Include the boundaries of these confidence intervals in the `reduced_polls` `data.frame`.

Hint: C.I has the form `avg +/- 1.96*sd/sqrt(n)`

Solution:

Hide

```r
reduced_polls <- reduced_polls %>%
  filter( n >= 3, type != "House-G") %>%
  mutate( se = sd/sqrt(n) ) %>%
  mutate(start = avg - 1.96*se, end = avg + 1.96*se)
reduced_polls
```

race_state	avg	act	sd	year	state	state_long	type	n	se
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<int>	<dbl>
2008_Pres_AK	0.1450000000	0.2150	0.036968455	2008	AK	alaska	Pres	4	0.018484228
2008_Pres_AL	0.2300000000	0.2160	0.035590261	2008	AL	alabama	Pres	4	0.017795130
2008_Pres_AR	0.0933333333	0.1980	0.020816660	2008	AR	arkansas	Pres	3	0.012018504
2008_Pres_AZ	0.0350000000	0.0850	0.017320508	2008	AZ	arizona	Pres	4	0.008660254
2008_Pres_CA	-0.2171428571	-0.2400	0.041918288	2008	CA	california	Pres	7	0.015843624
2008_Pres_CO	-0.0633333333	-0.0900	0.027628488	2008	CO	colorado	Pres	21	0.006029030
2008_Pres_DE	-0.2100000000	-0.2490	0.079372539	2008	DE	delaware	Pres	3	0.045825757
2008_Pres_FL	-0.0267647059	-0.0280	0.027930714	2008	FL	florida	Pres	34	0.004790078
2008_Pres_GA	0.0400000000	0.0520	0.025819889	2008	GA	georgia	Pres	4	0.012909944
2008_Pres_IA	-0.1288888889	-0.0950	0.028037673	2008	IA	iowa	Pres	9	0.009345891

1-10 of 204 rows | 1-10 of 12 columns Previous 1 2 3 4 5 6 ... 21 Next

Part 5 (10 pts)

For each election type in each year, calculate the fraction of states where the actual result was **outside** of the 95% confidence interval. Which race was the most unpredictable, (i.e. for which race was the polling data most inaccurate compared to the actual result)?

Solution:

Hide

```r
reduced_polls %>%
  mutate( in_range = act > start & act < end ) %>%
  group_by(year, type) %>%
  summarize( elections_in_range = sum( in_range ),
             n=n(),
             percentage_in_range = sum( in_range )/length(in_range) ) %>%
  arrange(percentage_in_range)
```

```
`summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

year	type	elections_in_range	n	percentage_in_range
<dbl>	<chr>	<int>	<int>	<dbl>
2012	Sen-G	5	21	0.2380952
2016	Pres	13	51	0.2549020
2010	Sen-G	7	24	0.2916667
2012	Pres	8	24	0.3333333
2008	Sen-G	10	26	0.3846154
2008	Pres	22	38	0.5789474
2014	Sen-G	12	20	0.6000000

7 rows

The 2012 Senatorial polling data was the most inaccurate, followed closely by the 2016 presidential election.

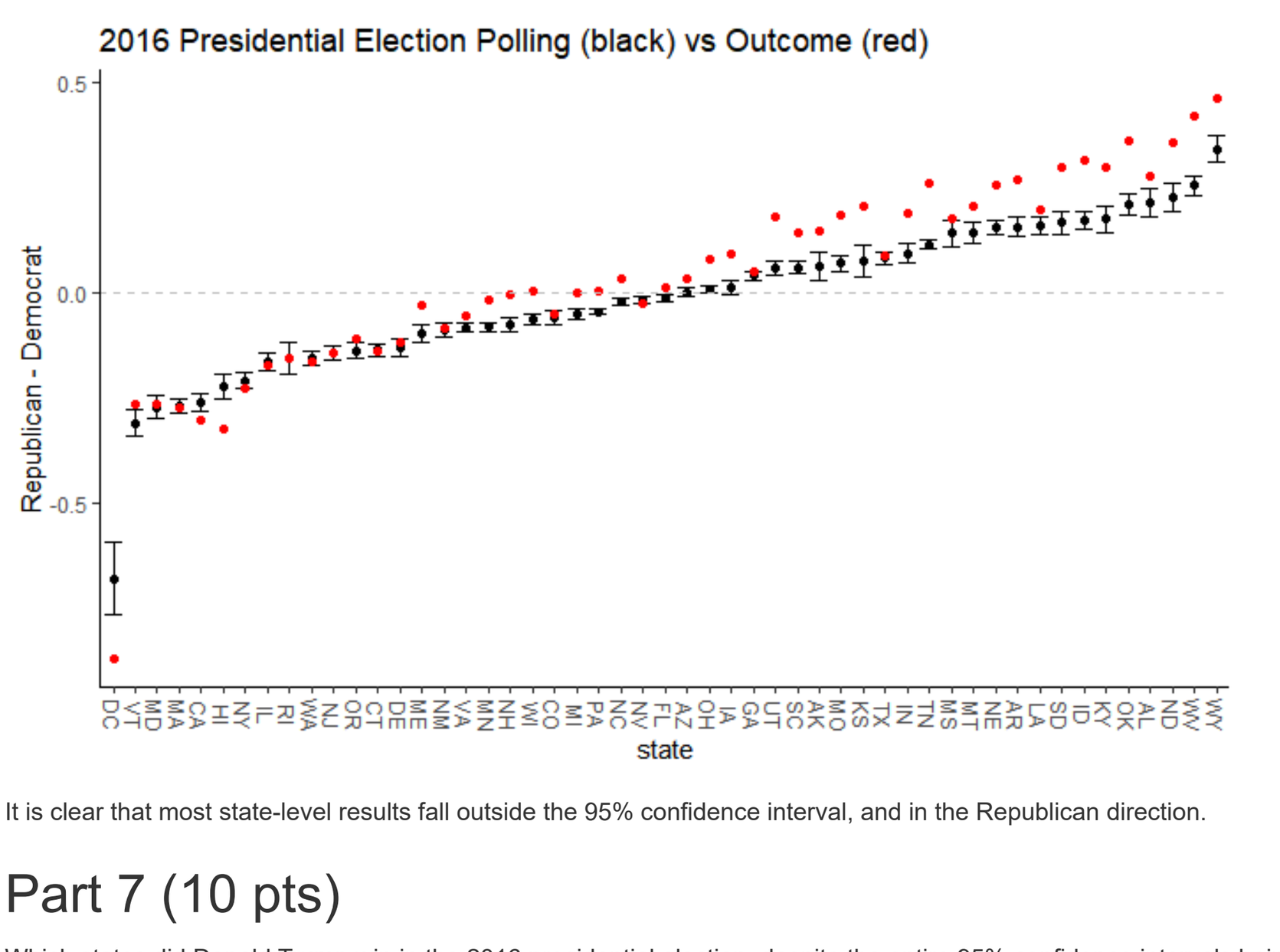
Part 6 (10 pts)

Using data from *only* the 2016 presidential election, make a plot of states (x-axis) and $\hat{\theta}$ estimates (y-axis). Using the `gg_errorbar` function, include the 95% confidence intervals of $\hat{\theta}$ for each state. Finally, using a different color, include the actual results for each state. Describe the resulting plot.

Solution:

Hide

```r
reduced_polls %>%
  filter( type=="Pres", year %in% 2016 ) %>%
  mutate( state=reorder(state, avg, mean) ) %>%
  ggplot(aes(state, avg) ) + geom_point() +
  geom_errorbar(aes(ymin=start, ymax=end), colour="black", width=.8) +
  geom_point( aes( state, act ), col="red" ) +
  theme_classic() +
  theme(axis.text.x=element_text(angle=-90, vjust=0.5)) +
  ylab("Republican - Democrat") +
  ggtitle("2016 Presidential Election Polling (black) vs Outcome (red)") +
  geom_hline(yintercept=0, linetype="dashed", color="grey")
```



It is clear that most state-level results fall outside the 95% confidence interval, and in the Republican direction.

Part 7 (10 pts)

Which states did Donald Trump win in the 2016 presidential election, despite the entire 95% confidence intervals being in favor of his opponent, Hillary Clinton?

Solution:

Hide

```r
reduced_polls %>%
  filter( year==2016 & type == "Pres" & end < 0 & act > 0 ) %>%
  select( state_long )
```

state_long
<chr>
florida
michigan
north carolina
pennsylvania
wisconsin

5 rows

Donald Trump won Florida, Michigan, North Carolina, Pennsylvania, and Wisconsin, despite the entire 95% confidence intervals from polling data predicting a win for Hillary Clinton.

Part 8 (10 pts)

Looking again at all races, calculate the the difference between θ and $\hat{\theta}$ (Hint: use the data for all races in the `reduced_polls` object created in Part 4). We call this the bias term. Add these values as a column to `reduced_polls`.

Solution:

Hide

```r
reduced_polls <- reduced_polls %>%
  mutate( bias = act - avg )
reduced_polls
```

race_state	avg	act	sd	year	state	state_long	type	n	se
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<int>	<dbl>
2008_Pres_AK	0.1450000000	0.2150	0.036968455	2008	AK	alaska	Pres	4	0.018484228
2008_Pres_AL	0.2300000000	0.2160	0.035590261	2008	AL	alabama	Pres	4	0.017795130
2008_Pres_AR	0.0933333333	0.1980	0.020816660	2008	AR	arkansas	Pres	3	0.012018504
2008_Pres_AZ	0.0350000000	0.0850	0.017320508	2008	AZ	arizona	Pres	4	0.008660254
2008_Pres_CA	-0.2171428571	-0.2400	0.041918288	2008	CA	california	Pres	7	0.015843624
2008_Pres_CO	-0.0633333333	-0.0900	0.027628488	2008	CO	colorado	Pres	21	0.006029030
2008_Pres_DE	-0.2100000000	-0.2490	0.079372539	2008	DE	delaware	Pres	3	0.045825757
2008_Pres_FL	-0.0267647059	-0.0280	0.027930714	2008	FL	florida	Pres	34	0.004790078
2008_Pres_GA	0.0400000000	0.0520	0.025819889	2008	GA	georgia	Pres	4	0.012909944
2008_Pres_IA	-0.1288888889	-0.0950	0.028037673	2008	IA	iowa	Pres	9	0.009345891

1-10 of 204 rows | 1-10 of 13 columns Previous 1 2 3 4 5 6 ... 21 Next

Part 9 (10 pts)

Plot and compare the distribution of bias terms for races in each year. Describe the bias patterns. Are these centered around zero? Give possible explanations.

Solution:

Hide

```r
reduced_polls %>%
  ggplot( aes( as.factor(year), bias, fill=type) ) +
  geom_hline(yintercept=0, linetype="dashed", color="grey") +
  geom_boxplot( ) +
  xlab("Electoral year") +
  ylab("Bias (in Republican - Democrat difference)") +
  theme_classic()
```