Precept 5

Summary of R Topics Covered in this Precept

We will use the data from today's precept to illustrate the use of these functions.

```
polls <- read.csv("data/polls2016States.csv")
pres <- read.csv("data/pres2016States.csv")
```

Loops!

- Loops are a programming construct that allow you to repeatedly execute similar code chunks in a compact manner.
- Here is a simple example.

```
for (i.count in 1:5) {
  print(i.count)
}
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

This loop has a for() command, and the counter is 5. R interprets the code as doing the following:

1. Set `i.count` to 1.
2. Do whatever is in the brackets.
3. Set `i.count` to the next value.
4. If `i.count` is less than or equal to 5, go back to (2).

Error

- In general, the prediction error is defined as: prediction error = actual outcome - predicted outcome
- Let's calculate the error in Clinton's vote share in Pennsylvania

```
pennsylvania <- polls[polls$state == "PA",]
pennsylvania.latest <- sort(pennsylvania$days_to_election, index.return = TRUE) # sort by date
PA.recent <- pennsylvania[pennsylvania.latest$ix[1], ] # use sorted index to get latest polls
error <- pres$Clinton[pres$state == "PA"] - PA.recent$Clinton
error
```

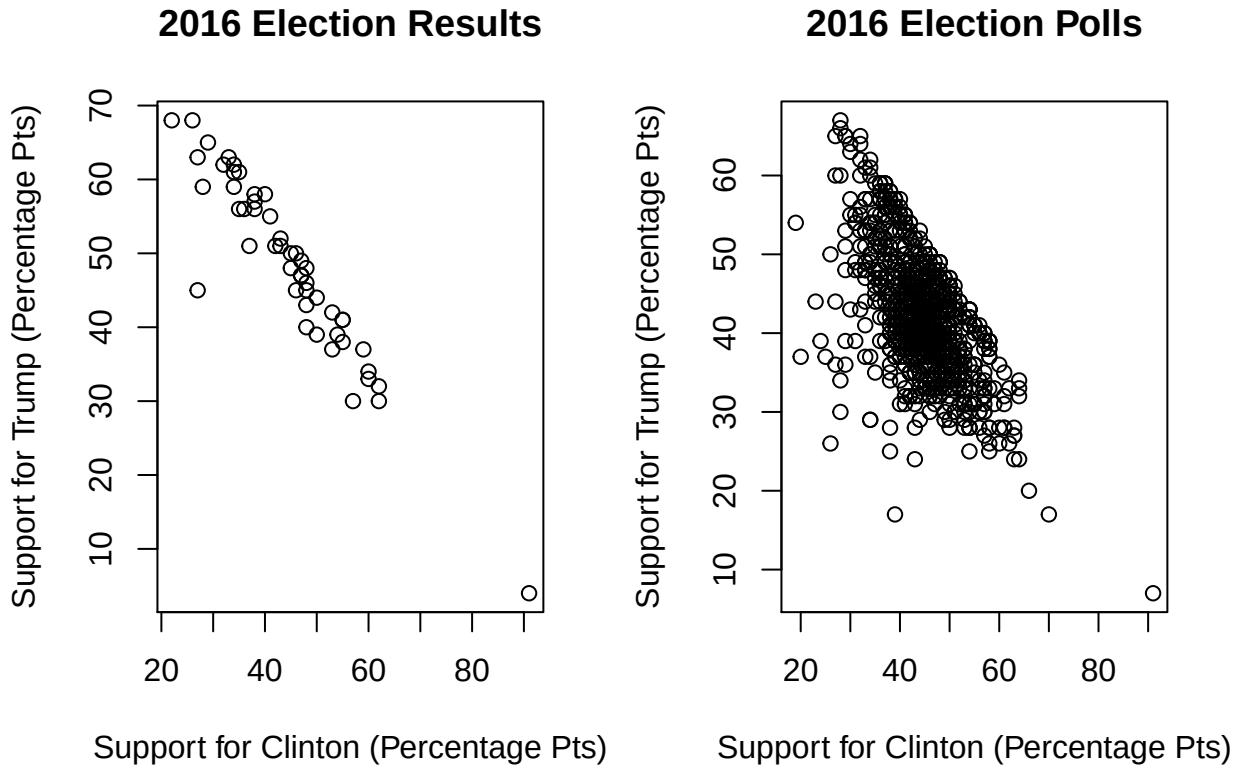
```
[1] 1
```

Note that to do this we are using the `sort` command to sort the polls from the latest to the oldest. When the `index.return` argument is set to `TRUE`, this function will also return a vector of the original indicies of each value. This variable is called `ix`. If, for an observation, `ix` is 5, this indicates that the observation was originally the fifth observation in the dataset before it was sorted.

Side by side plots

- Often, it can be useful to visual our data side-by-side
- To do this, we can use the function `par()` as in `par(mfrow = c(X,Y))` before we start making plots. This will create an X by Y grid of "sub-plots". Our multiple plots will fill this grid in row by row.

```
par(mfrow = c(1, 2))
plot(pres$Clinton, pres$Trump,
     xlab = "Support for Clinton (Percentage Pts)",
     ylab = "Support for Trump (Percentage Pts)",
     main = "2016 Election Results")
plot(polls$Clinton, polls$Trump,
     xlab = "Support for Clinton (Percentage Pts)",
     ylab = "Support for Trump (Percentage Pts)",
     main = "2016 Election Polls")
```




```
# or we can fill the grid column by column
# using syntax such as par(mfcol = c(2,1))
```

Precept Questions

You might remember the 2016 election, when "the polls got it wrong." In today's precept we will analyze state-level polls downloaded from the Huffington Post's Pollster and 3 additional polls for Washington D.C. available at (http://www.electoral-vote.com/evp2016/Pres/pres_polls.txt) that were conducted BEFORE the 2016 election to predict the outcomes of the 2016 presidential election. We will analyze these results and compare poll predictions to the *actual* 2016 election outcomes to see where the predictions and actual results diverged.

The first dataset we will be using this week, `data/polls2016States.csv`, has 905 observations, each representing a different poll, and includes the following 7 variables:

Name	Description
`id`	Poll ID
`state`	U.S. state where poll was fielded
`Clinton`	The poll's estimated level of support for Hillary Clinton (in percentage points)
`Trump`	The poll's estimated level of support for Donald Trump (in percentage points)
`days_to_election`	Number of days before November 4, 2016.
`electoral_votes`	Number of electoral votes allocated to the state where the poll was fielded (a state-level variable)
`population`	The poll's target population, which may be `Adults`, `Registered Voters`, or `Likely Voters`

The second dataset (`pres2016States.csv`) contains the actual results from the election, as reported by the New York Times. It has 905 observations (one per state and DC) and includes the following 7 variables:

Name	Description
`state.name`	U.S. state name where results were recorded
`state`	U.S. state abbreviation where results were recorded
`Clinton`	The state's level of support for Hillary Clinton (in percentage points)
`Trump`	The state's level of support for Donald Trump (in percentage points)

Hint: To do this assignment, you will have to sort the polls by the `days_to_election` variable within each state. Use the `sort` function to sort the polls from the latest to the oldest. When the `index.return` argument is set to `TRUE`, this function will return the ordering index vector, which can be used to extract the 3 most recent polls for each state.

Question 1

Load in the data from `polls2016States.csv`. Then, let's go through the preliminaries of a loop. In this first loop, we are going to calculate the mean of the last three polls for candidates Clinton and Trump.

First, create a vector which has all the unique state names. There should be 51 (verify this!). We also need to create two containers for the output. Name them `Clinton.support` and `Trump.support`, with each a vector of 51 `NA`'s (one for each state).

Answer 1

Question 2

Start with the code from above. Create a variable named `i`, which we will use as a counter through the loop. Set the value of `i` to one. Then, extract a subset of the data for the state `i`.

Answer 2

Question 3

Start with the code from above. From this subset, extract the three most recent polls for Clinton and for Trump. Save the mean of the three most recent Clinton polls and most recent Trump polls in the i^{th} position of the containers `Clinton.support` and `Trump.support`.

Answer 3

Question 4

Start with the code from above. Place it inside a `for` loop. Remember to initialize the containers outside the loop. When you are done, check `Clinton.support` and `Trump.support`. Which states are the most pro-Clinton according to the polls? Which are the most pro-Trump?

Answer 4

```
## load data:
polls <- read.csv("data/polls2016States.csv")
## state names
state.names <- unique(polls$state)
## initialize prediction vectors
Clinton.support <- Trump.support <- rep(NA, 51)
names(Clinton.support) <- names(Trump.support) <- state.names
for (i in 1:51) {
  state.data <- subset(polls, state == state.names[i]) # subset each state
  idx <- sort(state.data$days_to_election, index.return = TRUE) # sort by date
  state.recent <- state.data[idx$ix[1:3], ]  # use sorted index to get the 3 latest polls
  Clinton.support[i] <- mean(state.recent$Clinton)
  Trump.support[i] <- mean(state.recent$Trump)
}
# Most pro-Clinton
sort(Clinton.support, decreasing = T)[1:5]
```

```
      DC       HI       VT       MD       CA
75.66667 64.00000 61.00000 58.00000 57.66667
```

```
# Most pro-Trump
sort(Trump.support, decreasing = T)[1:5]
```

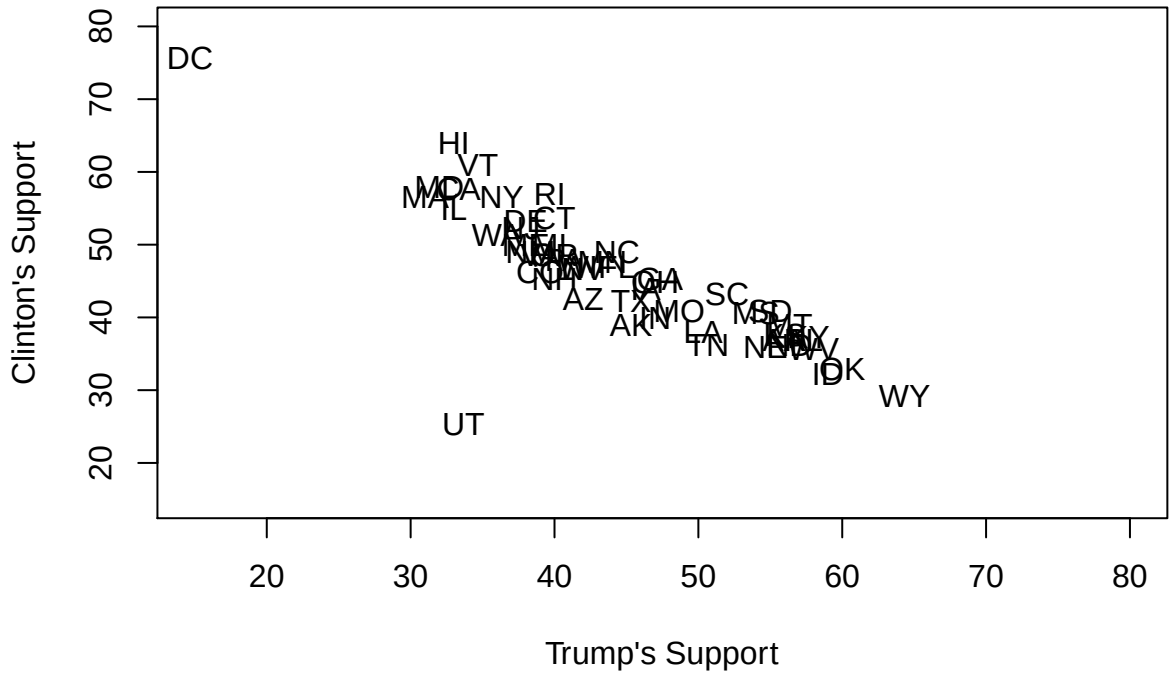
```
      WY       OK       ID       WV       KY
64.33333 60.00000 59.00000 58.00000 57.66667
```

Question 5

Create a scatterplot showing support for Clinton vs. support for Trump. Use state abbreviations to plot the results. Briefly interpret the results.

Answer 5

```
## scatterplot
plot(Trump.support, Clinton.support, type = 'n',
     xlim = c(15, 80), ylim = c(15, 80),
     xlab = "Trump's Support", ylab = "Clinton's Support")
## add state abbreviations instead of points
text(Trump.support, Clinton.support, labels = state.names)
```



The scatterplot shows support for Clinton on the y-axis and Support for Trump on the x-axis. States where Trump was expected to win appear below the 45 degree line, which is indicated by a dashed line, and states where Clinton is expected to win appear above this line. States where both candidates had nearly the same level of support (see Arizona) appear along the 45 degree line. These were the potentially contested states. Nearly perpendicular to the 45 degree line we see the expected negative relationship between support for Clinton and Support for Trump. Utah is a notable outlier because Trump was only a few points ahead of Clinton there and neither candidate was broadly supported. Instead, an independent candidate, Evan McMullin, was doing well.

Question 6

Now let's compare the *predicted* to the *actual* election results. Load the dataset (`pres2016States.csv`). Plot the poll predictions for Clinton support against Clinton's actual support. Instead of a point for each state, plot the state name. Provide a brief interpretation.

Answer 6

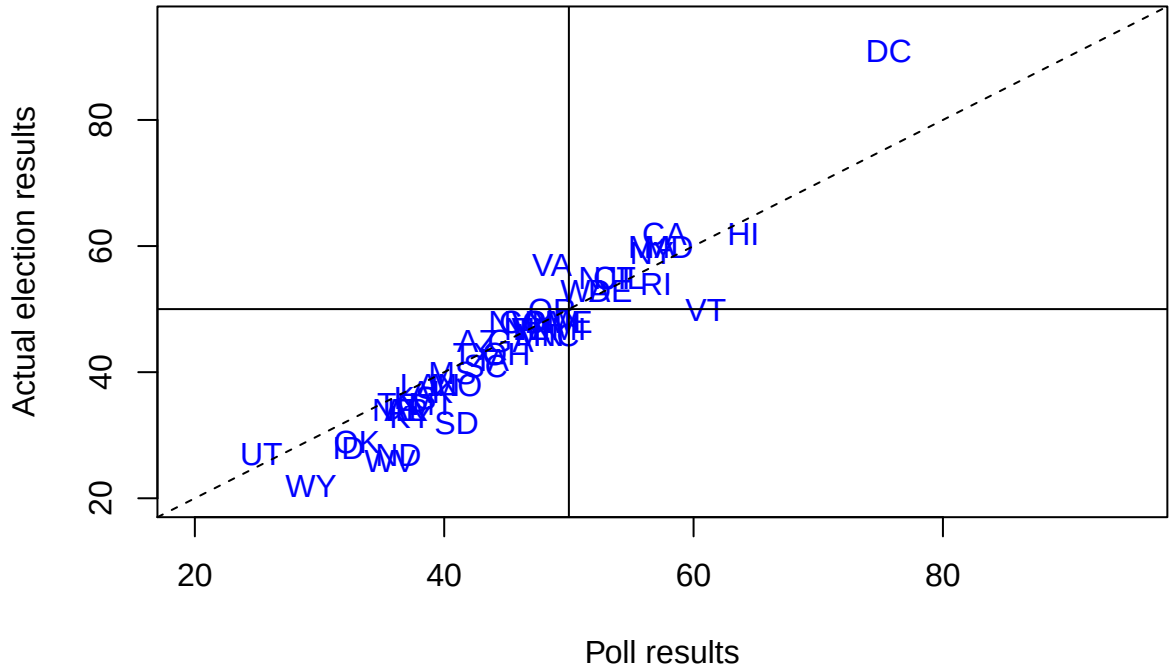
```
## load data:
pres <- read.csv("data/pres2016States.csv")
```

```r
## plot

plot(Clinton.support, pres$Clinton,
     type = "n", main = "", xlab = "Poll results",
     xlim = c(20,95), ylim = c(20,95),
     ylab = "Actual election results")

## add state abbreviations
text(Clinton.support, pres$Clinton, pres$state,
     col = "blue")

## lines
abline(a = 0,b = 1, lty = "dashed")
# 45 degree line
abline(v = 50)
# vertical line at 0
abline(h = 50) # horizontal line at 0
```



The plot shows states that Clinton was incorrectly predicted to lose in the upper left quadrant. In the lower right, we see places were Clinton was predicted to win but actually lost.

Question 7

Using the results from question 1, compare the predicted Clinton vote margin to the actual Clinton vote margin within each state, which can be calculated with (`pres2016States.csv`). Were the polls biased? Were they accurate? Which state polls were wrong?

Answer 7

```
# Calculate predicted and actual margin
pred.margin <- Clinton.support - Trump.support
pres$margin <- pres$Clinton - pres$Trump
names(pres$margin) <- pres$state

## error of latest polls
errors <- pres$margin - pred.margin
names(errors) <- pres$state
mean(errors) # mean prediction error
```

```
[1] -4.810458
```

```
## Wrong polls
pres$state[sign(pred.margin) != sign(pres$margin)]
```

```
[1] AZ FL MI NH NC PA WI
51 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

```
pres$margin[sign(pred.margin) != sign(pres$margin)]
```

```
[1] -3 -2  0  0 -4  0  0
```

When we look at the errors by state, we see that for some states, the poll predictions were higher than the actual vote share won by Clinton and for others, the poll predictions were lower. Therefore, it seems as though the results are unbiased. But the mean error of -4.81 suggests that the predicted Clinton margin was substantially higher than the actual margin. As we all know by now, the polls were *not* accurate in certain consequential states. We see that a number of swing states ended up going Trump or that areas where Clinton was supposed to win ended up in a tie (Michigan, Pennsylvania, and Wisconsin ended up going Trump).