IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

6. Data Exploration

Some Suggestions for Student Projects

- If you have very large dataset, you can use multiple R Markdown files
  - R Markdown File 1 is used for data transformation and manipulation. Save the cleansed data into a local data file. For example:

    ```
    save.image(file="MyData.RData")
    ```

  - In R Markdown File 2 used for data visualization or other analyses, load the local data file. For example,

    ```
    load(file = "MyData.RData")
    ```

  - By doing so, you don't need to run the time consuming data cleansing and transformation process every time when you run the visualization or analyses procedures.

- Not recommend to load all libraries at the beginning of your R markdown. Some packages may have functions with the same name. This could lead to some hard-to-detect bugs in your code. Instead, load the package when you need to use it and unload it when your work is finished.

- Do NOT reinvent the wheel. Always prefer to use functions in the current packages (such as various data manipulation functions in dplyr). Do not write your own complicated logic unless you have to do so.

Reading Assignment

- Hypothesis Testing
  - https://courses.edx.org/c4x/UTAustinX/UT.7.01x/asset/Chapter_12.pdf
- How to Correctly Interpret P Values
  - http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values

Learning Objectives

- Understand methods (listwise deletion, and imputation) used to deal with missing data
- Be able to detect outliers in a dataset by using multiple methods such as boxplot rule, z-score, and density-based local outlier
- Understand the difference between covariance and correlation
- Be able to visualize correlation relationships (scatter plot, scatter plot matrix, correlation matrix)
- Understand the distinction between population and sample and the concept of hypothesis testing
- Understand the meaning of p-value and be able to interpret p-value correctly
- Be able to conduct one-sample t-test and two-sample t-test

Agenda

- Missing Data
- Outliers
- Correlation Analysis
- Hypothesis Testing

Missing Data

Missing Data

- A missing value is indicated as NA (not available). Arithmetic calculation on missing values yield missing values.

- Two Basic Methods:
  - Listwise deletion: remove missing values or cases from analysis.
  - Imputation: replacing missing data with substituted values

- Common R functions:
  - is.na() : Determine which elements are missing
  - na.omit(): Removing missing values or cases (listwise deletion)

- Many R functions provide options for dealing with missing values. Refer to the help file for each function.

R Code: Dealing with Missing Values

```r
## Missing Values ##
x <- c(1,2,3,4,NA)

# Determine which elements are missing
is.na(x)
# Count missing values
table(is.na(x))
# Remove missing values
na.omit(x)

# Arithmetic calculation on missing values yield missing values.
min(x)
max(x)
mean(x)
var(x)

# Exclude missing values from analysis
min(x, na.rm = TRUE)
max(x, na.rm = TRUE)
mean(x, na.rm = TRUE)
var(x, na.rm = TRUE)

# Simply Imputation: Replace by Mean
x
x[is.na(x)] <- mean(x, na.rm=TRUE)
X
```

```r
## Handling missing values in data frames
library(dplyr)

# Create a group variable
g <- c(rep("group_1",5),rep("group_2",5))

# Generate two random variables
a <- rnorm(10, mean = 0, sd = 1)
b <- a*0.5 + rnorm(10,mean = 0, sd = 1)
a[4] <- NA
b[7] <- NA

dat <- data.frame(g,a,b)
dat

cor(a,b)
cor(a,b, use = "pairwise.complete.obs")

dat %>% group_by(g) %>%
summarise(mean(a),mean(b))

dat %>% group_by(g) %>% summarise(mean(a, na.rm =
TRUE),mean(b, na.rm = TRUE))

dat %>% na.omit() %>% group_by(g) %>%
summarise(mean(a),mean(b))
```

Summary of Missing Values Handling

- Listwise deletion
  - Applies when the amount of missing data is relatively small
  - Use parameter `na.rm = TURE` in function calls to exclude missing value from data analysis
  - Use `na.omit()` function to remove missing values from dataset

- Simple imputation: Replace by mean
  - Applies when leaving out available data points deprives the data of some amount of information
  - Usually introduces bias
  - Sample Syntax:
    - `x[is.na(x)] <- mean(x, na.rm=TRUE)`
    - `x <- dplyr::recode(x, .missing = mean(x,na.rm = TRUE))`
- Other advanced imputation methods
  - ……

Outliers

Outliers

- Outliers are unusually small or large values in a dataset.

- Outliers have a huge impact on the results of data analyses.

- Common ways of dealing with outliers
  - Recode: when outliers are incorrectly recorded in the dataset
  - Remove: when outliers are incorrectly included in the dataset
  - Keep: when outliers are correct values in the dataset

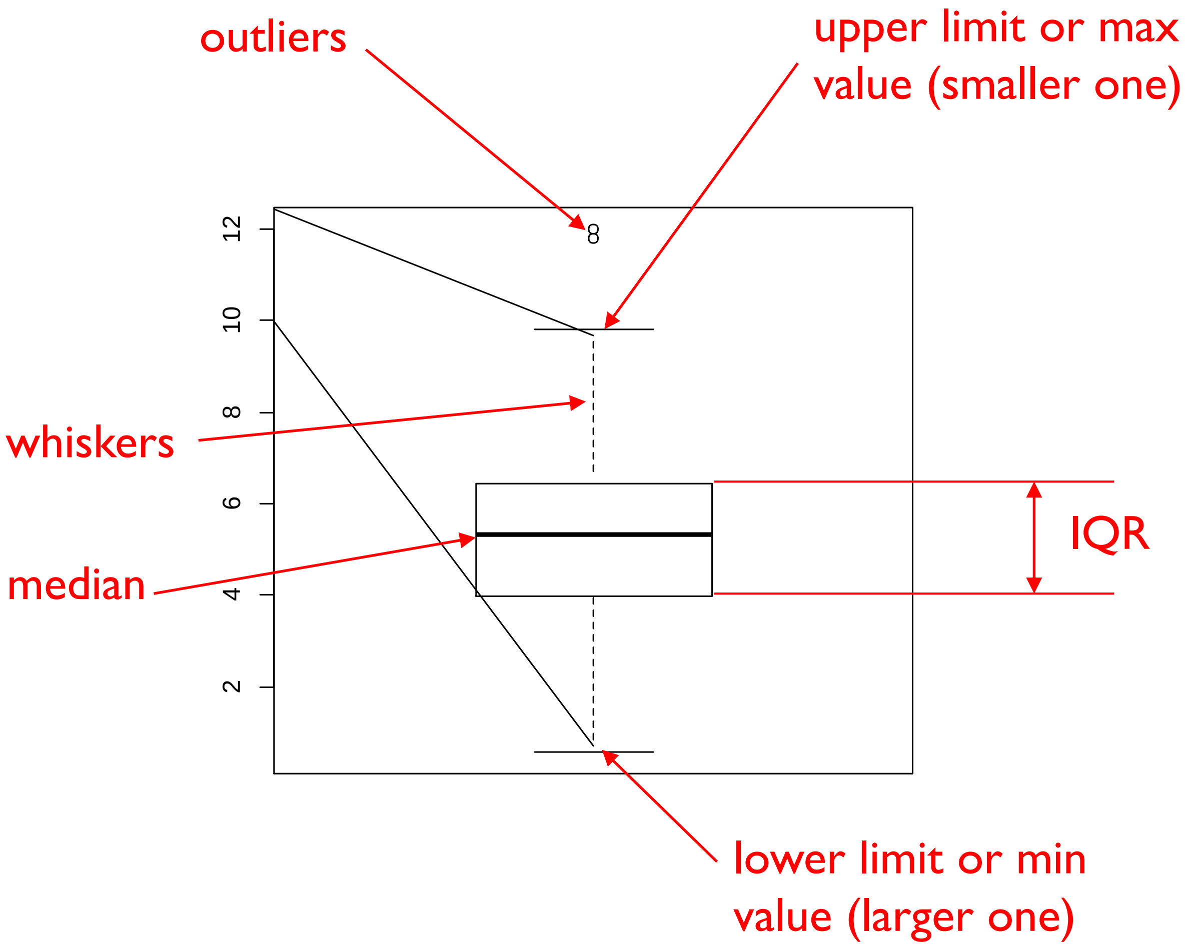
Detecting Potential Outliers

- Basic methods
  - Boxplot rule
  - z-Score
  - Density-based local outlier

- Different methods may lead to different results.

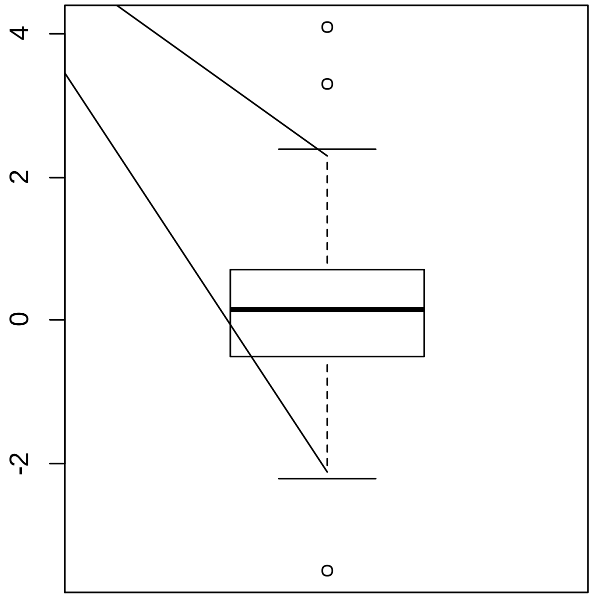
Boxplot Rule

- Values beyond [Q1 – 1.5 * IQR, Q3 + 1.5 * IQR] are outliers.


outliers
upper limit or max value (smaller one)
whiskers
median
IQR
lower limit or min value (larger one)
12
10
8
6
4
2

R Code: Detecting Outliers According to Boxplot Rule

```
# Generate 100 values
set.seed(1)
x <- rnorm(100)
# Intentionally add 3 outliers
x <- c(x,-3.5, 3.3, 4.1)

# Create boxplot
out <- boxplot(x)
out
# Show the structure
str(out)

# Print outliers
out$out
```



z-Score Method

- A z-Score (a.k.a. standard score) measures the relative location of the observation in a data set.

$$z(x_i) = \frac{x_i - \bar{x}}{sd(x)}$$

- An empirical rule tells us:
  - For data with a bell-shaped distribution, almost all values will be within three standard deviations of the mean.

- Thus, the detection rule is simply:
  - Any data value with a z-score less than -3 or greater than 3 can be regarded as an outlier.

R Code: Detecting Outliers Using z-Scores

```
# Generate 100 random values
set.seed(1)
x <- rnorm(100)
# Intentionally add 3 potential outliers
x <- c(x,-3.5, 3.3, 4.1)

# Calculate z-score
z <- (x - mean(x))/sd(x)

# Print outliers
x[which(z > 3 | z < -3)]
# Note: In the 2nd condition, make sure to have a space between "<" and "-".
# Or else R may interpret "<-" as assignment rather than comparison.
```

Only -3.5 and 4.1 are detected as potential outliers. Why?

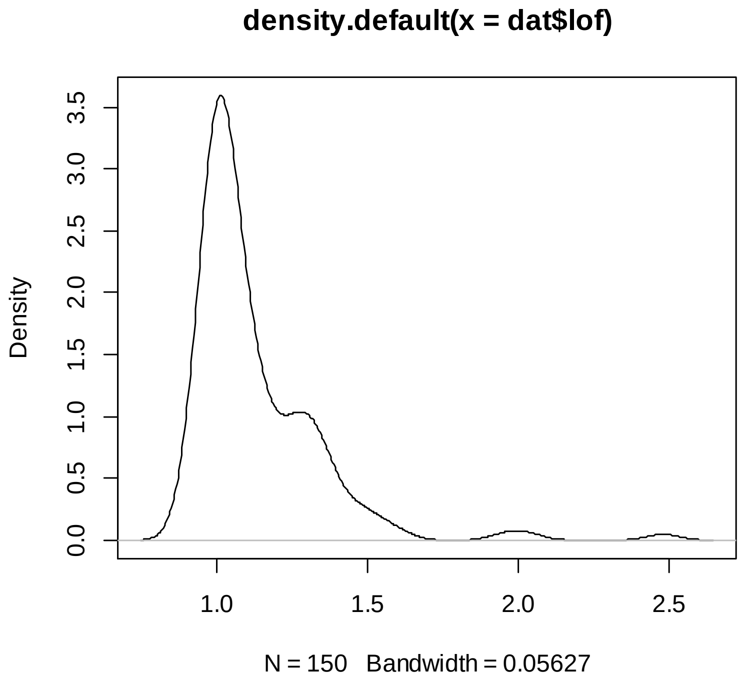
Density-based local outlier detection

- LOF (Local Outlier Factor) Algorithm
  - The local density of a value is compared with that of its neighbors.
  - If the value's density is significantly lower than that of its neighbors (with an LOF value greater than one), the value is in a sparser region than its neighbors.
  - This suggests that the value is an outlier.

For the detail, refer to http://www.rdatamining.com/examples/outlier-detection

R Code: Detecting Outliers Using LOF Algorithm

```
# Detecting outliers using LOF algorithm
library(dprep)

dat <- iris[-5]

# Calculate local outlier factor score for each observation.
# Partition all observations into k = 5 clusters.
dat$lof <- lofactor(data = dat, k = 5)
dat

# Plot the density of the LOF scores
plot(density(dat$lof))

# Pick top 5 as outliers
outliers <- head(dat[order(dat$lof, decreasing=T),],5)

# Print outliers
print(outliers)
```



How to deal with outliers?

- If the outliers detected turn out to be wrong data, we certainly need to correct the data or exclude the outliers from data analysis.

- If the outliers are valid data (for example, income of Bill Gates), we may keep it in our data analysis but with caution of how such outliers could interfere with our data analysis.

Use %in% to filter out outliers

- Syntax: x %in% y
- %in% returns a logical vector indicating if there is a match or not for its left operand

  ```
  > 1:4 %in% 1:2
  [1] TRUE TRUE FALSE FALSE
  > !1:4 %in% 1:2
  [1] FALSE FALSE TRUE TRUE
  ```

- Remove outliers from dataset

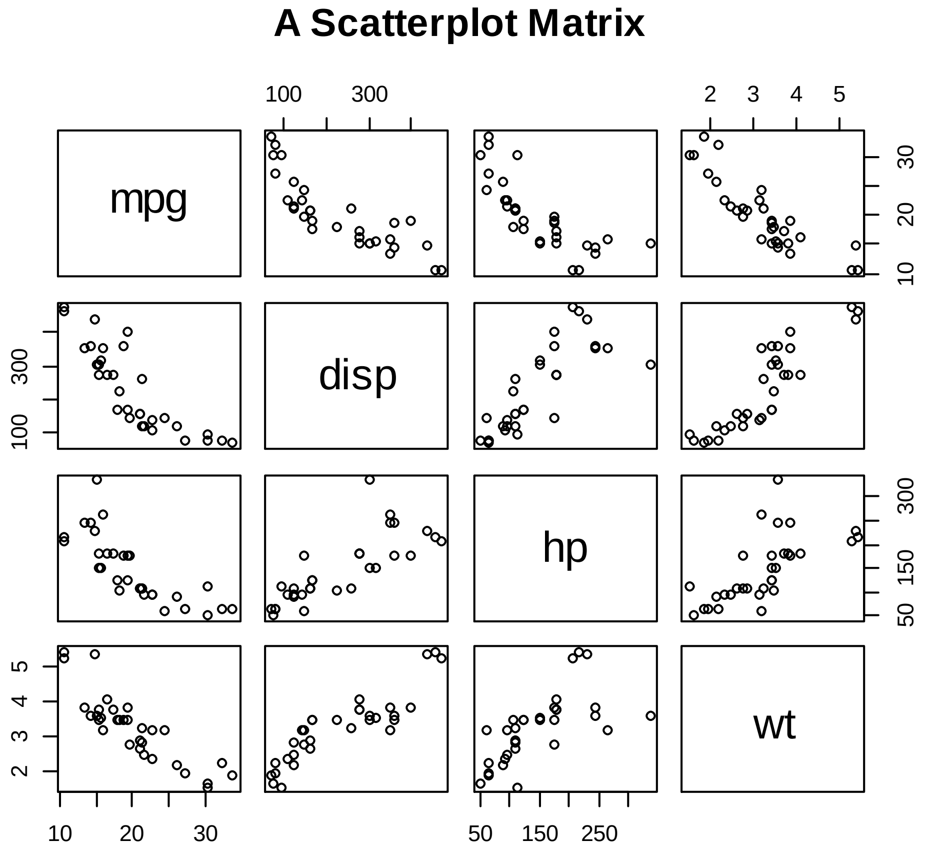
  ```
  dat2 <- dplyr::filter(dat,!dat$lof %in% outliers$lof)
  ```

Correlation Analysis

Recap: Scatter Plot Matrix

- A way to roughly show the relationship between multiple variables

```
pairs(~mpg+disp+hp+wt, data=mtcars, main="A Scatterplot Matrix")
plot(mtcars[c("mpg","disp","hp","wt")], main="A Scatterplot Matrix")
```



Motivation

- We can use various tabular and graphical tools to explore the relationship between variables. For example, cross table, scatter plot, and scatterplot matrix.

- Although such tabular and graphical tools vividly show the relationship, they have some limitations:
  - It's difficult to compare multiple relationships. For example, is the negative relationship between mpg and wt stronger than the one between mpg and hp?
  - It's inefficient when there are many variables.

- Can we numerically measure the relationship between variables?

Correlation Analysis

- Two ways of measuring linear association between two variables
  - Covariance
    - Direction of the linear association
  - Correlation
    - Direction of the linear association
    - Strength of the linear association

SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISHBURNE
© marketoonist.com

Covariance

- Covariance measures of the linear association between two variables.

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Covariance > 0: a positive relationship
- Covariance < 0: a negative relationship

R Code: Covariance

```
# Define a function to calculate covariance
mycov <- function(x,y){
  n <- length(x)
  x_hat <- mean(x)
  y_hat <- mean(y)
  sum((x-x_hat)*(y-y_hat))/(n-1)
}

cat("covariance =", mycov(mtcars$mpg,mtcars$wt))
cat("covariance =", mycov(mtcars$hp,mtcars$wt))

# Use cov() function to calculate variance
cov(mtcars$mpg,mtcars$wt)
cov(mtcars$hp,mtcars$wt)
```

```
cov(mtcars$mpg,mtcars$wt)) = -5.116685

cov(mtcars$hp,mtcars$wt)) = 44.19266
```

Interpretation: The covariance of mpg and wt is -5.116685, which indicates a negative linear relationship between the two variables.

Correlation Coefficient

- Correlation coefficient is a normalized measurement of the linear association between two variables.

$$cor(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

- The correlation coefficient can take on values between -1 and +1.
  - 0 < cor < 1: a positive relationship; values near 1 indicate a strong positive linear relationship.
  - -1 < cor < 0: a negative relationship; values near -1 indicate a strong negative linear relationship.

Interpreting Correlation Coefficient

- Correlation coefficient measures the direction and strength of the linear association between two variables

- Direction
  - Increasing (positive)
  - Decreasing (negative)

- Strength
  - How closely are X and Y related?

R Code: Correlation

```r
# Manually calculate correlation coefficient
cov(mtcars$mpg,mtcars$wt)/(sd(mtcars$mpg)*sd(mtcars$wt))

# Use cor() function to calculate correlation coefficient
cor(mtcars$mpg,mtcars$wt)
cor(mtcars$hp,mtcars$wt)
```

cor(mtcars$mpg,mtcars$wt)) = -0.8676594

cor(mtcars$hp,mtcars$wt)) = 0.6587479

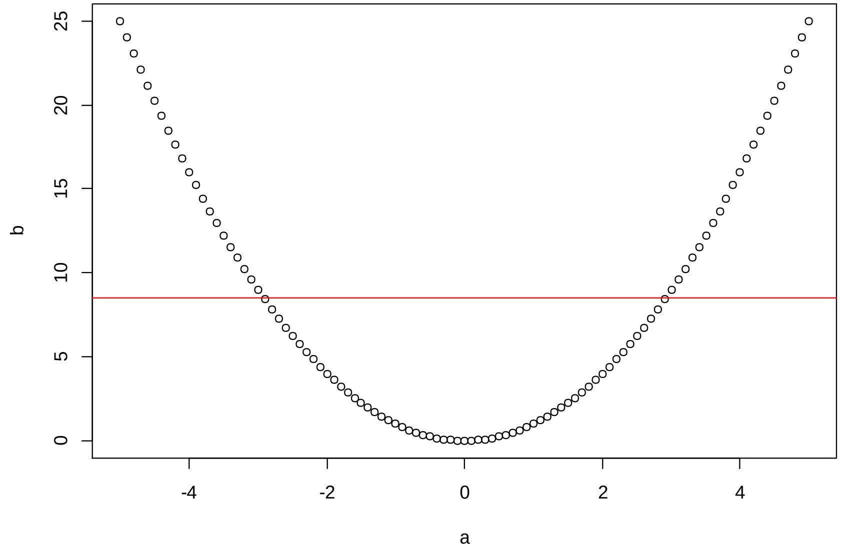
Interpretation:

- The correlation coefficient of mpg and wt is -0.8676594, which indicates a strong negative linear relationship between the two variables.
- The correlation coefficient of hp and wt is 0.6587479, which indicates a moderate positive linear relationship between the two variables.

Note: Correctly Interpret Covariance and Correlation

- Covariance and correlation can only measure linear relationship.
- A zero covariance or correlation does NOT mean that the two variables have no relationship.

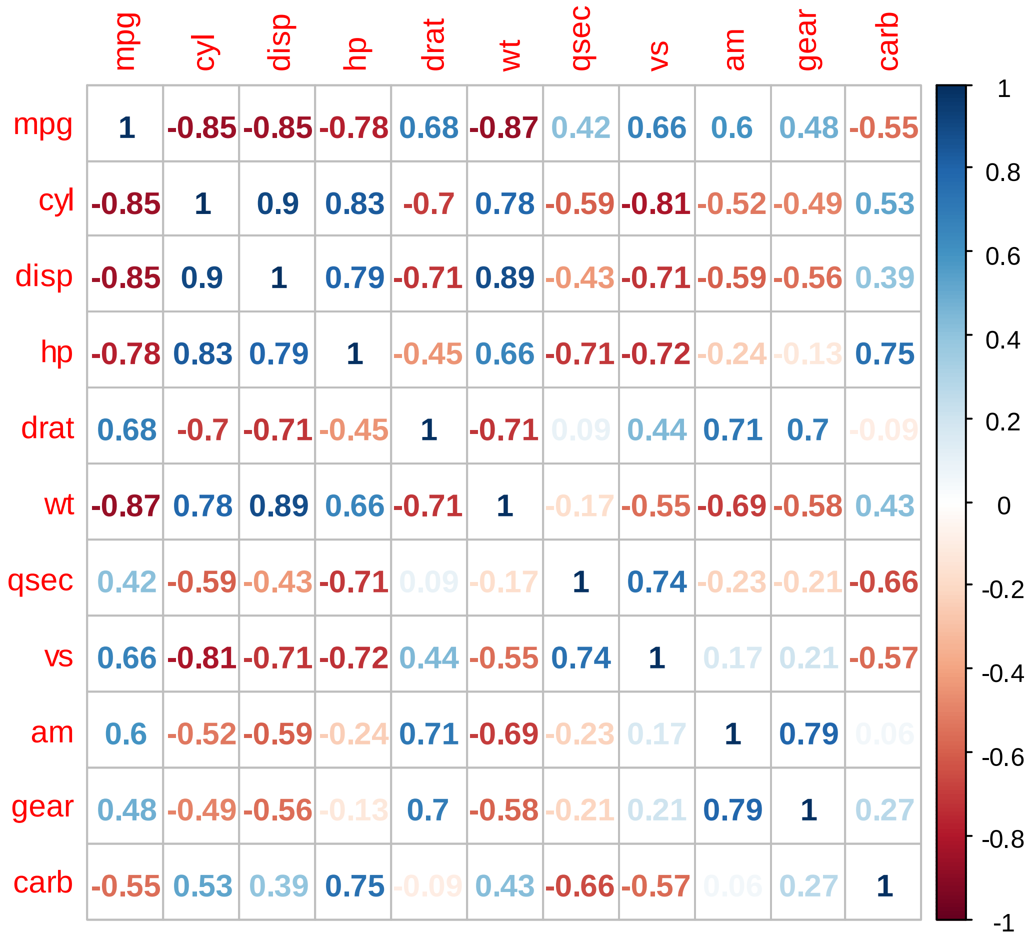
```
a <- seq(from = -5, to = 5, by = 0.1)
b <- a^2

plot(a,b)
# Add regression line
abline(lm(b~a),col = "red")

cov(a,b)
cor(a,b)
```



Note: a and b has a functional relationship (a = b^2). However, their covariance and correlation are very small (<0.0001). In statistics, we usually call such relationship as a non-linear relationship.

Visualization of Correlation Matrix

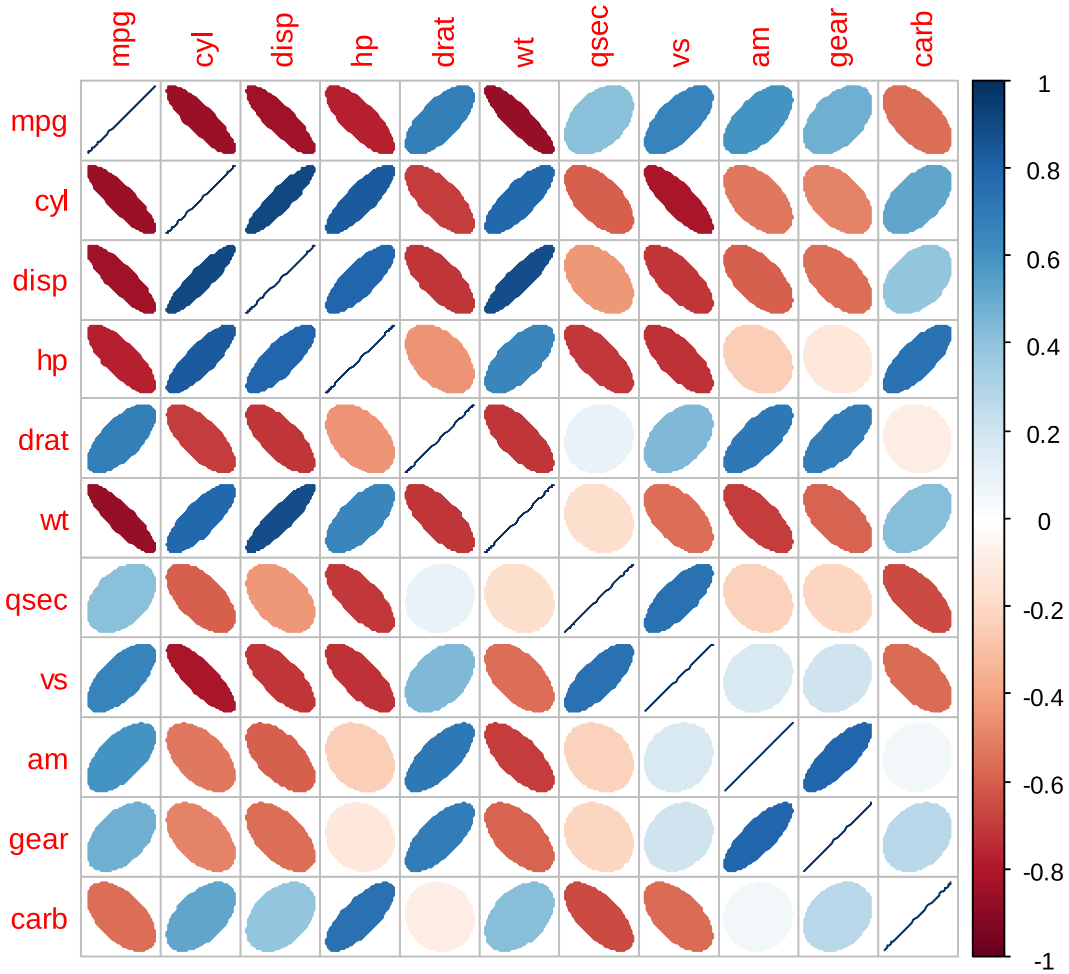
```
# install.packages("corrplot")
library(corrplot)
corrplot(cor(mtcars), method="number", tl.cex=1)
```



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

(cont.)

```
corrplot(cor(mtcars), method="ellipse", tl.cex = 1)
```



Correlation = Causality?

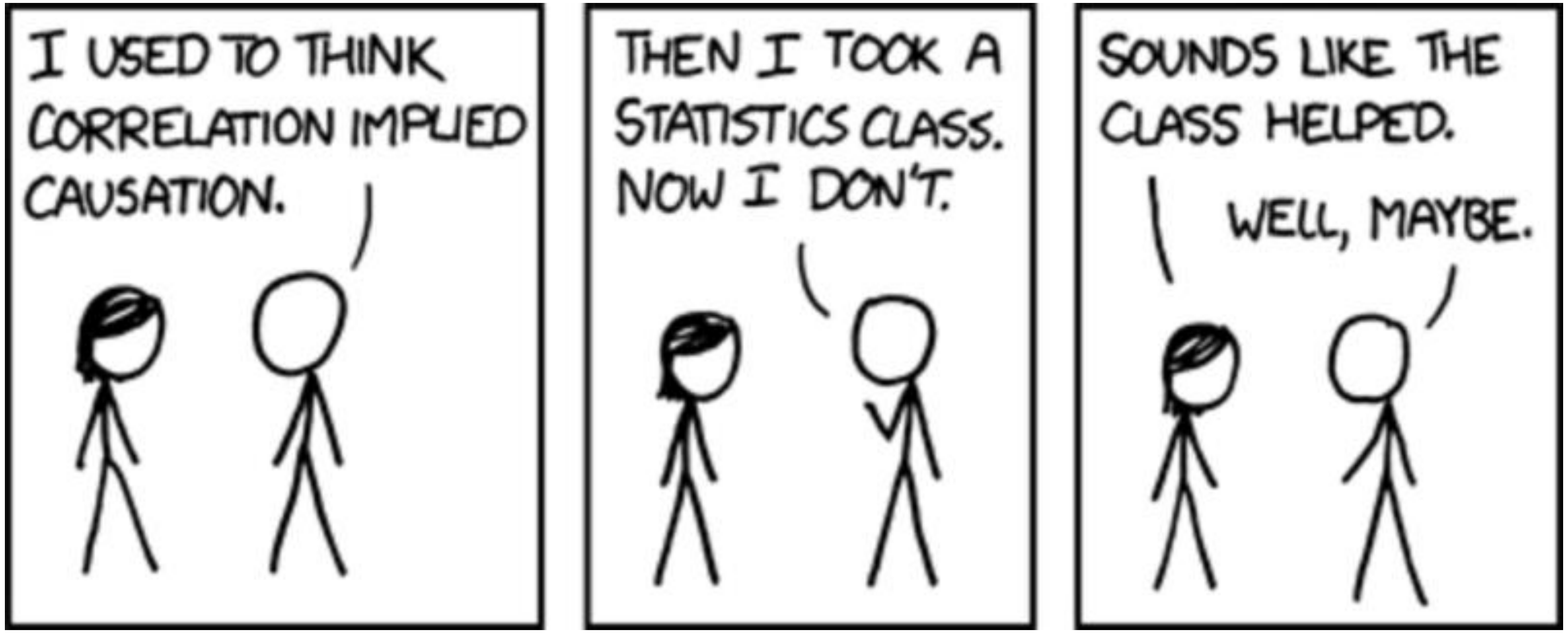
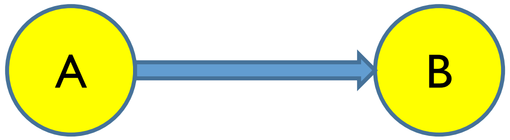

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

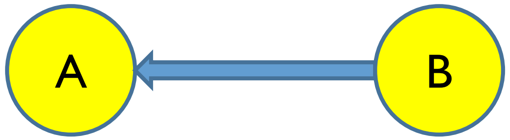
Image source: https://openparachute.wordpress.com/2015/11/15/anti-fluoride-hypothyroidism-paper-slammed-yet-again/correlation/

Correlation Does Not Mean Causality

- Reverse Causality



A causes B



B causes A

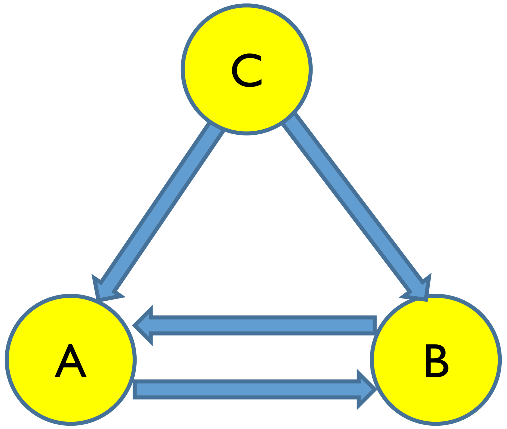
When a country's debt rises above 90% of GDP, growth slows.

Therefore, high debt causes slow growth.

In actuality, low growth causes debt to increase.

Correlation Does Not Mean Causality (cont.)

- Third Factor C Causes both A and B



As ice cream sales increase, the rate of drowning deaths increases sharply.
Therefore, ice cream consumption causes drowning.

Ice cream is sold during the hot summer months at a much greater rate than during colder times, and it is during these hot summer months that people are more likely to engage in activities involving water, such as swimming.
The increased drowning deaths are simply caused by more exposure to water-based activities, not ice cream.

Hypothesis Testing

Population vs. Sample

- Population: the whole body of subjects that are of interest to us in a particular study.

- Sample: a portion of the population from which we collect data.

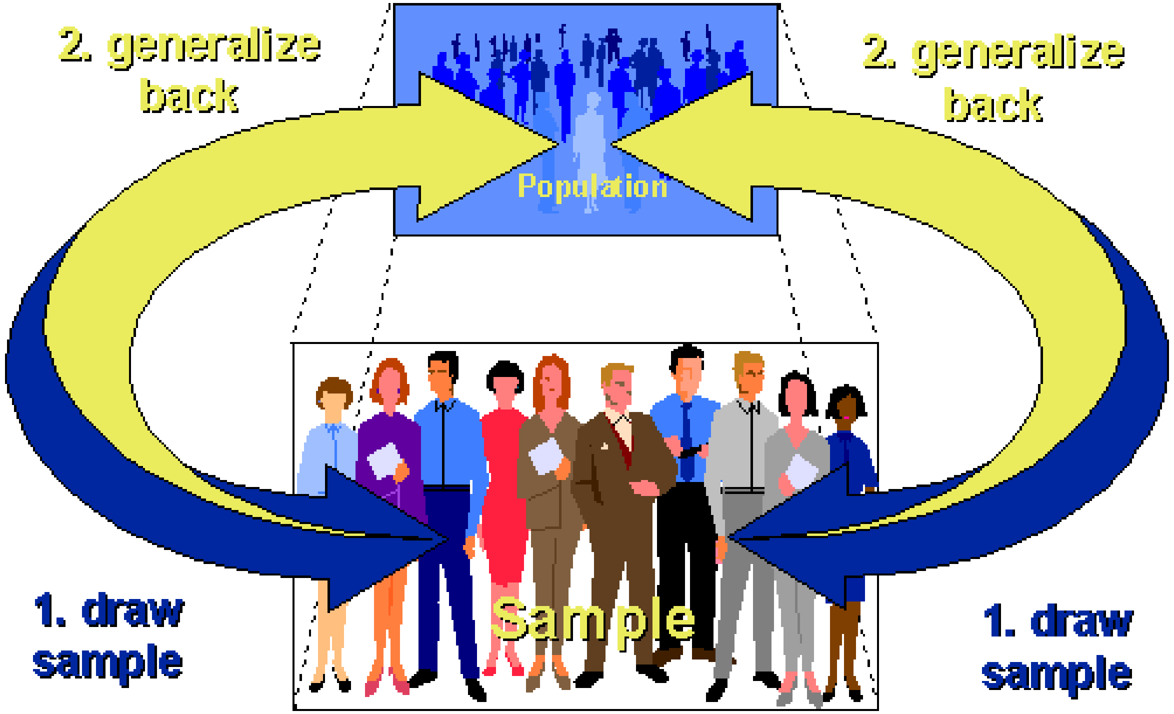
For example, if we want to study college students in U.S.A:
population = all college students in U.S.A.
a sample = 1000 students randomly selected in MST

- In most cases, the dataset set we have in hand does not cover all subjects. However, we still want to know the patterns or relationships for the whole subject.

- In order to make statements on population based on data collected from sample, the sample must be representative of the population. This involves hypothesis testing.

Sampling Model

- Identify the population and then draw a fair sample from that population and do statistical analysis;
- Because the sample is representative of the population, you can automatically generalize your results back to the population.



Source: http://www.socialresearchmethods.net/kb/external.php

Hypothesis Testing

- A hypothesis is a statement about a population parameter.

- The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.
  - Null hypothesis H_0: a tentative assumption about a population parameter
  - Alternative hypothesis H_a: the opposite of the null hypothesis

Forms for Null and Alternative Hypotheses about a Population Mean

- μ is the population mean
- μ_0 is a hypothesized value of the population mean

$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
One-tailed (lower-tail)	One-tailed (upper-tail)	Two-tailed

- The null and alternative hypotheses divide all possibilities into two non-overlapping sets.
- The equality part is always in null hypotheses.

In-Class Exercise

- The school principal wants to test if it is true what teachers say – that high school juniors use the computer an average 3.2 hours a day.
- What are our null and alternative hypotheses?
- If we want to test if the average daily computer use is at least 3.2 hours, what are the null and alternative hypotheses?

What is p-value?

- P value is the probability of obtaining the observed or more extreme effect in your sample data, assuming that the null hypothesis is true.
- The smaller the p-value, the stronger the evidence against the null hypothesis.
- Commonly used threshold for p-value

$P > 0.10$	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
$0.05 < P < 0.10$	Weak evidence against the null hypothesis in favor of the alternative.
$0.01 < P < 0.05$	Moderate evidence against the null hypothesis in favor of the alternative.
$0.001 < P < 0.01$	Strong evidence against the null hypothesis in favor of the alternative.
$P < 0.001$	Very strong evidence against the null hypothesis in favor of the alternative.

Source: What is a p-value? http://www.stat.ualberta.ca/~hooper/teaching/misc/Pvalue.pdf

One and Two Sample Tests

Purpose	Type of Data		
	Gaussian	**Non-Gaussian**	**Binomial**
Compare one group to a hypothetical value	One sample t-test	Wilcoxon Test	Binomial Test
Compare two paired groups	Paired t-test (student's t-test)	Wilcoxon Test	McNemar’s Test
Compare two unpaired groups	Two sample t-test	Wilcoxon-Mann-Whitney Test	Chi-Square or Fisher’s Exact Test

Note: Paired groups have the same subjects in different conditions. For example, same patients before and after some treatment.

R function t.test()

- Syntax:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

  - y:
    - If y is excluded => a one-sample t-test using x
    - If y is included => a two-sample t-test using both x and y
  - mu:
    - one sample test: the true value of the mean
    - two sample test: the difference in means
  - alternative: a character string specifying the alternative hypothesis
    - "two.sided" (default option) => H_0: mean(x)=mu, H_a: mean(x) <> mu
    - "greater" => H_0: mean(x)<=mu, H_a: mean(x) > mu
    - "less" => H_0: mean(x)>=mu, H_a: mean(x) < mu
  - paired: a logical indicating whether you want a paired t-test.
    - paired = FALSE => unpaired t-test
    - paired = TRUE => paired t-test

R function t.test()

- Syntax using formula:

```
t.test(formula, data, subset, na.action, ...)
```

- formula: a formula of the form lhs ~ rhs
  - lhs is a numeric variable giving the data values
  - rhs a factor with two levels giving the corresponding groups

Example 1: One sample t-test

- Suppose a car manufacturer claims that a model gets mpg >= 26. We want to test whether this hypothesis is true by randomly collect mpg information from 10 cars of the model.

- H_0: mean(mpg)>=26, H_a: mean(mpg) < 26

Example 1: R Code

```
# 10 mpg data points we collected
mpg = c(24.2, 25.3, 23.8, 27.1, 28.0, 25.4, 22.7, 24.6, 25.9, 23.1)

# one sample t-test
t.test(mpg, alternative = "less", mu = 26)
```

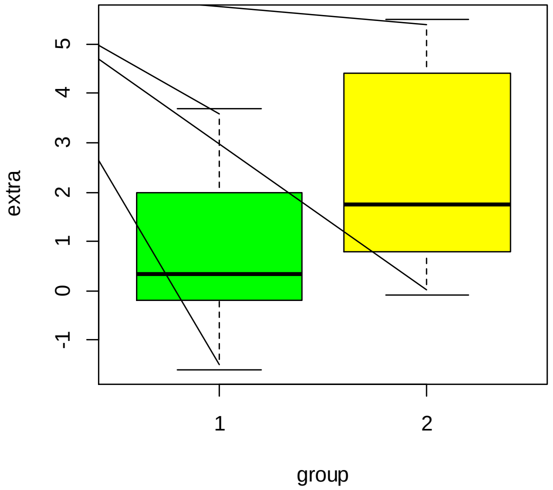
```
> t.test(mpg,alternative = "less",mu = 26)

	One Sample t-test

data:  mpg
t = -1.8559, df = 9, p-value = 0.04822
alternative hypothesis: true mean is less than 26
95 percent confidence interval:
     -Inf 25.98783
sample estimates:
mean of x 
    25.01 
```

The null hypothesis is rejected ($p < 0.05$). Thus, the manufacturer’s claim is suspicious.

Example 2: Two sample t-test

- A dataset shows the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.

	Extra Sleep										**Mean**	**Var**
Soporific Drug 1	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0	2	0.75	3.20
Soporific Drug 2	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4	2.33	4.01



- Do the two soporific drugs have different effects on extra hours of sleep?
  - H_0: mean(extra sleep)$_{\text{drug 1}}$ = mean(extra sleep) $_{\text{drug 2}}$
  - H_a: mean(extra sleep)$_{\text{drug 1}}$ <> mean(extra sleep) $_{\text{drug 2}}$

Example 2: R Code

```
# Two sample t-test
extra1 <- sleep$extra[sleep$group == 1]
extra2 <- sleep$extra[sleep$group == 2]
t.test(extra1, extra2, alternative = "two.sided", mu = 0, paired = TRUE)

# A simpler syntax
t.test(extra1, extra2, paired = TRUE)

# Use formula
t.test(extra ~ group, data = sleep, paired = TRUE)
```

```
> t.test(extra ~ group, data = sleep, paired = TRUE)

	Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences 
                  -1.58 
```

The null hypothesis is rejected ($p < 0.01$). Thus, the two drugs have significantly different effects on extra hours of sleep.

Extension

- T-test is used to compare means of some variable of interest between two groups (two sample t-test) or compare mean of a single variable to a benchmark (one way t-test).

- To compared means of some variable between two or more groups, we can use ANOVA (analysis of variance) or regression (by using dummy variables).
  - Type "?aov" to learn more…

Q & A