Big Data Analytics using R - Coursework 1

JP Sainsbury

15 October 2017

## 1. (a) Suppose you ask a group of 10 students at Birkbeck College how many brothers and sisters they have. The number obtained are as follows:2 3 0 5 2 1 1 0 3 3 Find the following measures of central tendency: (i) the mean, (ii) the median and (iii) the mode.

Sample

n = 10  
x = c(2, 3, 0, 5, 2, 1, 1, 0, 3, 3)

1. Mean

mean(x)

## [1] 2

(ii)Meadian

median(x)

## [1] 2

(iii)Mode

names(sort(-table(x)))[1]

## [1] "3"

## Find the following measures of spread:

1. the variance

var(x)

## [1] 2.444444

1. the standard divation

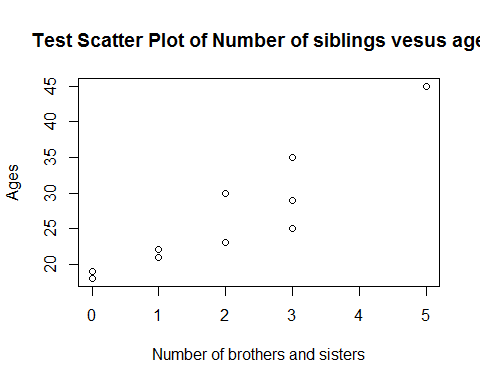
sd(x)

## [1] 1.563472

## (b) Suppose these 10 students have the following age: 23 25 18 45 30 21 22 19 29 35

1. Find the covariance and correlation between the number of siblings and their age.
2. Is there a positive or negative or no correlation between the two?
3. Is there causation between the two? Justify your answers.

y = c(23, 25, 18, 45, 30, 21, 22, 19, 29, 35)  
plot(x,y, xlab="Number of brothers and sisters", ylab="Ages", main="Test Scatter Plot of Number of siblings vesus age")



1. Covariance

cov(x,y)

## [1] 11.88889

& Correlation

cor(x,y)

## [1] 0.9116971

1. There is a positive correlation between the two.
2. The older generation are more likely to have more brothers and sisters. The plot clearly shows the positive correlation between the two.

## 2 Getting familiar with R

1. Load library

library(MASS)

Number of rows and columns.

The rows represent the cases, the cases here are suburbs of Boston. The columns represent the variables.

nrow(Boston)

## [1] 506

ncol(Boston)

## [1] 14

Columns also known as the variables contain 14 atributes.They are:

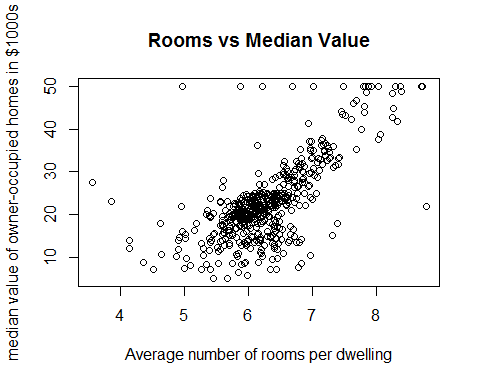
names(Boston)

## [1] "crim" "zn" "indus" "chas" "nox" "rm" "age"   
## [8] "dis" "rad" "tax" "ptratio" "black" "lstat" "medv"

## Some pairwise scatterplots

1. The below scatterplot displays a positive correlation between the number of rooms per dwelling and the median value. The more rooms a property has positivley influences the value of the house. The more rooms the dwelling has the higher the value of the house.

plot(Boston$rm , Boston$medv, xlab="Average number of rooms per dwelling", ylab="median value of owner-occupied homes in $1000s", main = "Rooms vs Median Value")

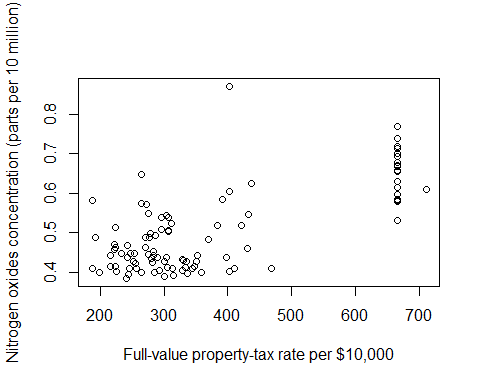


cor(Boston$rm , Boston$medv, )

## [1] 0.6953599

The below scatterplot displays a positive correlation between the property tax rate and Nitrogen oxides concentration. This idicates the higher the nitrogen oxides concentraction the more property tax paid. This leads me to believe house prices and nitrogen oxides positivly influence each other. It is known property prices are higher in the innercitys, demand for houses, population, traffic and polution are also concentrated in the inner citys. The scotterplot below supports this hypothesis.

plot(Boston$tax , Boston$nox, xlab="Full-value property-tax rate per $10,000", ylab="Nitrogen oxides concentration (parts per 10 million)" )



cor(Boston$tax , Boston$nox, )

## [1] 0.6680232

1. Are any of the predictors associated with per capita crime rate? If so, explain the relationship

library(xtable)  
newobject<-xtable(cor(Boston$crim,Boston))  
print.xtable(newobject, type="latex")

## % latex table generated in R 3.4.2 by xtable 1.8-2 package  
## % Thu Oct 26 21:54:21 2017  
## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrrrrrrrrrrrr}  
## \hline  
## & crim & zn & indus & chas & nox & rm & age & dis & rad & tax & ptratio & black & lstat & medv \\   
## \hline  
## 1 & 1.00 & -0.20 & 0.41 & -0.06 & 0.42 & -0.22 & 0.35 & -0.38 & 0.63 & 0.58 & 0.29 & -0.39 & 0.46 & -0.39 \\   
## \hline  
## \end{tabular}  
## \end{table}

table(cor(Boston$crim,Boston))

##   
## -0.388304608586812 -0.385063941994224 -0.379670086951024   
## 1 1 1   
## -0.219246702862514 -0.200469219662547 -0.0558915822222415   
## 1 1 1   
## 0.28994557927952 0.352734250901364 0.406583411406259   
## 1 1 1   
## 0.420971711392456 0.455621479447946 0.582764312032585   
## 1 1 1   
## 0.625505145262602 1   
## 1 1

Lets order the results

sort(cor(Boston$crim,Boston),decreasing = TRUE)

## [1] 1.00000000 0.62550515 0.58276431 0.45562148 0.42097171  
## [6] 0.40658341 0.35273425 0.28994558 -0.05589158 -0.20046922  
## [11] -0.21924670 -0.37967009 -0.38506394 -0.38830461

Lets pull out the strongest relationship

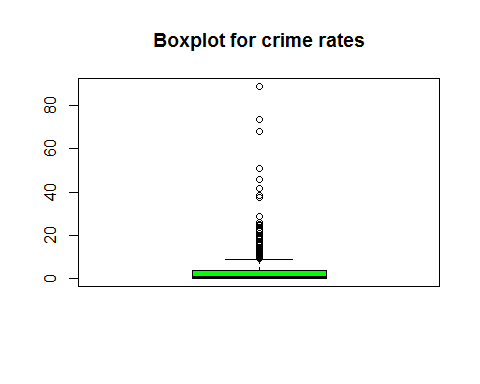
sort(cor(Boston$crim,Boston),decreasing = TRUE)[2]

## [1] 0.6255051

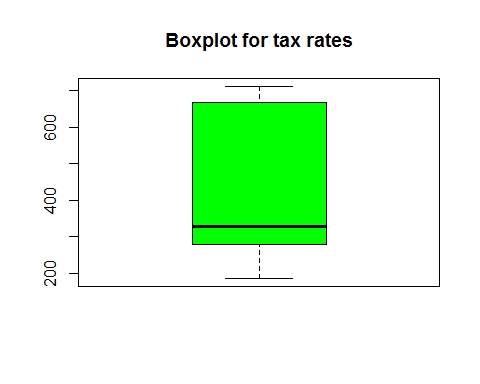
Index of accessibility has the strongest relationship with per capita crime rate. The easier the access to a particular area has a positive correlation on crime rate. The easier it is to access an area, the more lickly it is to have criminal activities take place there.

Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-?-teacher ratios? Comment on the range of each predictor.

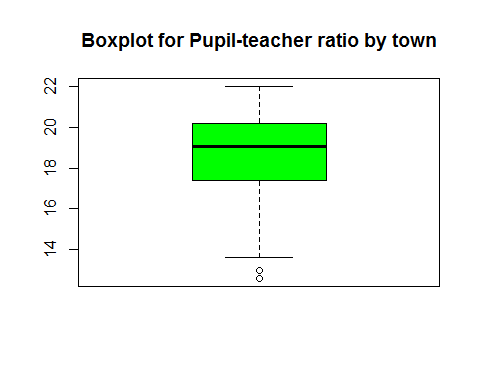
boxplot(Boston$crim,col=c("green"), names=c("Crime Boxplot"), main="Boxplot for crime rates")



boxplot(Boston$tax,col=c("green"), names=c("Tax Boxplot"), main="Boxplot for tax rates")



boxplot(Boston$ptratio,col=c("green"), names=c("Pupil-teacher ratio by town"), main="Boxplot for Pupil-teacher ratio by town")



1. How many of the suburbs in this data set bound the Charles river?

table(Boston$chas)

##   
## 0 1   
## 471 35

sum(Boston$chas)

## [1] 35

1. What is the median pupil-?-teacher ratio among the towns in this data set?

median(Boston$ptratio[Boston$chas=="1"])

## [1] 17.6

1. Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

#Row 399 has the lowest median value of owner occupied homes  
#399 its other metrics are:  
head(Boston[order(Boston$medv),],1)

## crim zn indus chas nox rm age dis rad tax ptratio black  
## 399 38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 396.9  
## lstat medv  
## 399 30.59 5

## 3. Linear Regression (2%) [Textbook 3.13 (a-?-f)] In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results. (a) Using the rnorm() function, create a vector, ???? , containing 100 observations drawn from a N (0, 1) distribution, i.e., a normal distribution with mean 0 and variance 1. This represents a feature, ????.

set.seed(1)  
x = rnorm(100,0,1)

1. Using the rnorm() function, create a vector, ????????????, containing 100 observations drawn from a N (0, 0. 25) distribution i.e. a normal distribution with mean 0 and variance 0. 25.

eps = rnorm(100,0,0.5)

1. Using ???? and ????????????, generate a vector ???? according to the model ???? = ???1 + 0.5???? + ??. What is the length of the vector ????? What are the values of ????! and ????! in this linear model?

MAking the prediction of Y on the basis X=x

y= -1 + (0.5 \* x) + eps  
Y=y

Making the prediction of y on the basis Y=y

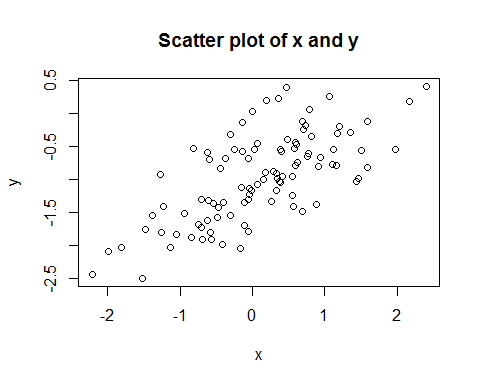
length(y)

## [1] 100

Beta zero is equal to the intercept -1. Beta 1 is equal to the slope 0.5. Yes this is a linear model.

1. Create a scatterplot displaying the relationship between ???? and ????. Comment on what you observe.

plot(x,y,main = "Scatter plot of x and y", xlab = "x", ylab= "y")



There is a positive corelation between x and y.