Statistics for Engineers (MAT2001)- Lab Experiment-II: Correlation and Regression

1 Karl Pearson's Coefficient of Correlation

1.1 Scatter Diagram

Problem:
Illustrate the relationship between the average age versus the time spent in the library, by using scatterplot.

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
10-19	15	302.38
20-29	25	193.63
30-39	35	185.46
40-49	45	198.49
50-59	55	224.30
60-69	65	288.71

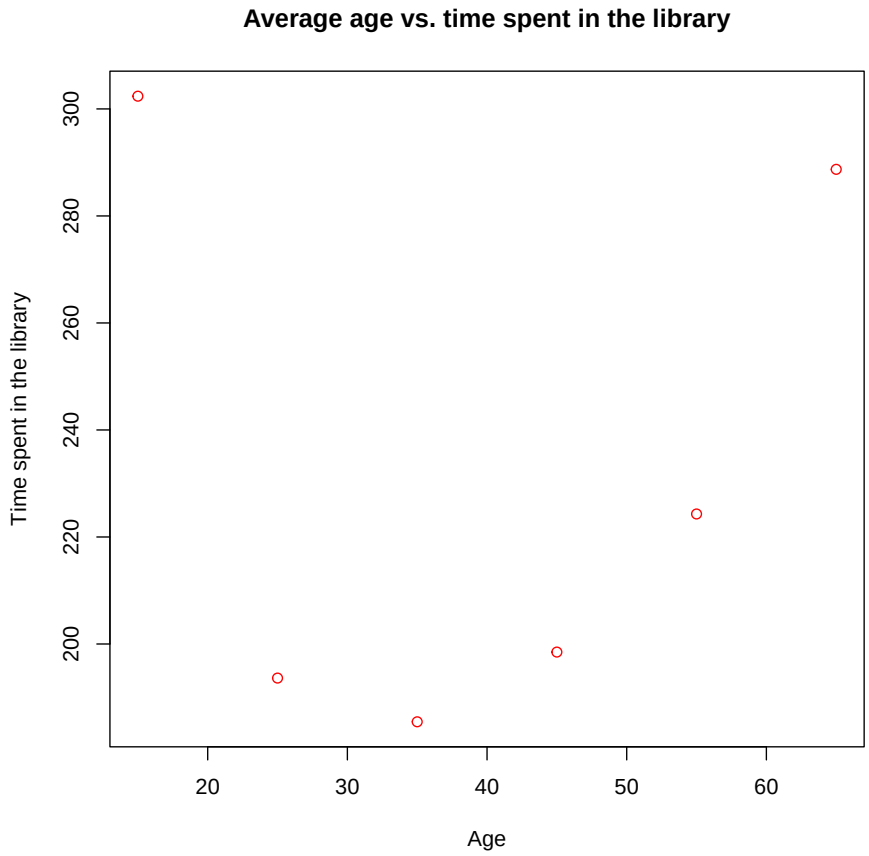
```
x <- c(15,25,35,45,55,65)
x

## [1] 15 25 35 45 55 65

y <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)
y

## [1] 302.38 193.63 185.46 198.49 224.30 288.71

# Scatter plot
plot(x,y, main="Average age vs. time spent in the library",
     xlab="Age", ylab="Time spent in the library",col="red")
```



1.2 Manual method

Problem:
Find the correlation coefficient for the given data: x=(23,27,28,28,29,30,31,32, 33,35), y=(18,20,22,27,21,29,27,29,28,29). Give a scatter plot to the x and y

```
x=c(23,27,28,28,29,30,31,32, 33,35)
x
```

```
##  [1] 23 27 28 28 29 30 31 32 33 35
```

```
length(x)
```

```
## [1] 10
```

```
y=c(18,20,22,27,21,29,27,29,28,29)
y
```

```
##  [1] 18 20 22 27 21 29 27 29 28 29
```

```
length(y)
```

```
## [1] 10
```

```
var(x)
```

```
## [1] 11.6
```

```
var(y)
```

```
## [1] 18.22222
```

```
var(x, y)
```

```
## [1] 12.11111
```

```
r=var(x,y)/sqrt(var(x)*var(y))
r
```

```
## [1] 0.8330179
```

Therfore there is positive correlation between the x and y.

1.3 Using R code

1.3.1 Method 1

```
cor(x,y, method ="pearson" )
```

```
## [1] 0.8330179
```

1.3.2 Method 2

```
cor.test(x,y,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 4.2587, df = 8, p-value = 0.002766
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4277335 0.9594318
```

```
## sample estimates:
##       cor
## 0.8330179
```

2 Spearman's Rank Correlation Coefficient

Problem :
Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below

Recruit	1	2	3	4	5	6	7	8	9
Selection Test Score	44	49	52	54	47	76	65	60	63
Proficiency Test Scrore	48	55	45	60	43	80	58	50	77

```
selection =c(44,49,52,54,47,76,65,60,63,58,50,67)
proficiency =c(48,55,45,60,43,80,58,50,77,46,47,65)
cor(selection,proficiency,method ="spearman")

## [1] 0.7202797

# (or)
cor.test(selection,proficiency,method ="spearman")

##
##  Spearman's rank correlation rho
##
## data:  selection and proficiency
## S = 80, p-value = 0.01102
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7202797
```

There is a positive correlation between selection and proficiency.

Problem :
Import the iris data set.
i) Find the Pearson correlation coefficient using the manual method for the features sepal length and sepal width when the Species is setosa. Give a plot visualization. Describe the relationship.
ii) Find the rank correlation coefficient for the features petal length and petal width when the Species is versicolor and describe the relationship.

iii) Find the Pearson correlation coefficient for the features petal length and petal width when the Species is virginica and describe the relationship.

```
data <- iris
head(data)
```

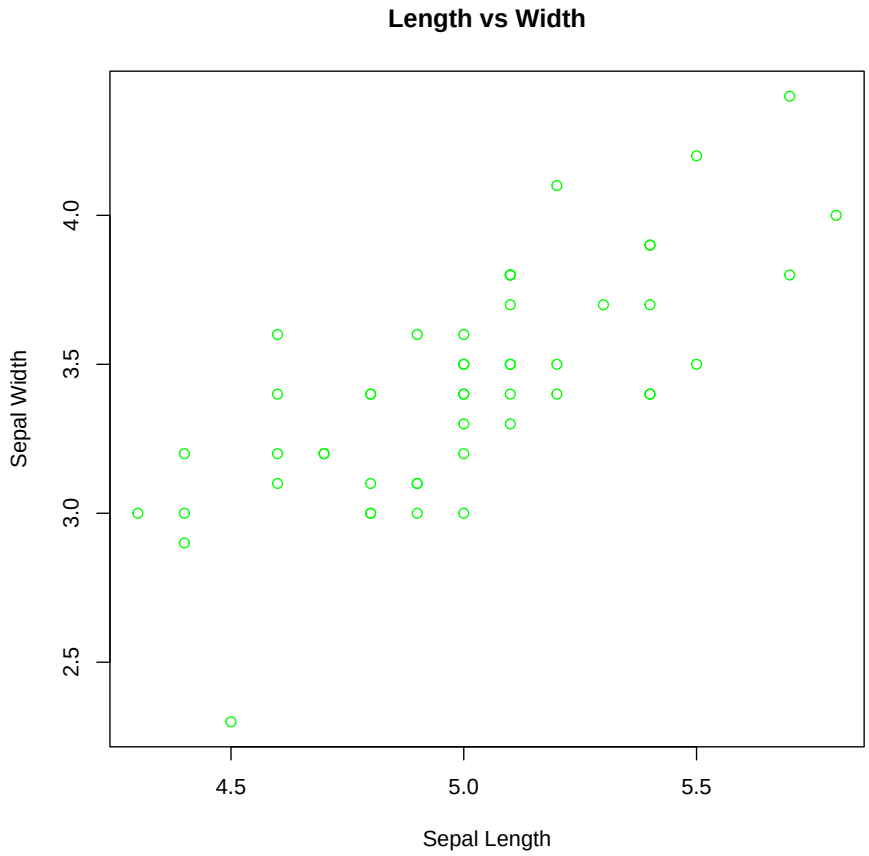
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
setosa_data <- subset(data, Species == "setosa")

r=var(setosa_data$Sepal.Length, setosa_data$Sepal.Width
      )/sqrt(var(setosa_data$Sepal.Length)*
               var(setosa_data$Sepal.Width))
r
```

```
## [1] 0.7425467
```

```
plot(setosa_data$Sepal.Length, setosa_data$Sepal.Width,
     main = "Length vs Width",  xlab ="Sepal Length",
     ylab = "Sepal Width", col= "green")
```



There is a positive correlation between sepal length and sepal width.

```
versi_data <- subset(data, Species == "versicolor")
cor(versi_data$Petal.Length, versi_data$Petal.Width,
         method = "spearman")

## [1] 0.7870096
```

There is a positive correlation between petal length and petal width.

```
virginica_data <- subset(data, Species == "virginica")
cor(virginica_data$Petal.Length, virginica_data$Petal.Width,
         method = "pearson")

## [1] 0.3221082
```

There is a positive correlation between petal length and petal width.

Experiment

1. Import the Orange data set. Find the Pearson correlation coefficient for each type of tree age and circumference. Also draw the scatter plot.
2. Import the trees data set.
i) Find the rank correlation for Girth and Volume. Draw a plot.
ii) Find the rank correlation for Height and Volume. Draw a plot.