Correlation and Regression

Correlation

The term correlation is used by a common man without knowing the meaning of the term correlation. For example, when parents advise their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only one variable such as height, weight, age, marks, wage etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-variate Analysis. Sometimes the variables may be inter-related. We study the relationship between blood pressure and age, consumption level of some nutrients and weight gain, total income and medical expenditure etc. The nature and strength of a relationship may be examined by correlation.

Thus correlation refers to the relationship of two variables or more. For example, relation between height of father and son, yield of crop and rainfall, wage and price index etc. Correlation is statistical Analysis which measures and analyzes the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. If two or more variables are so related that the change of one variable brings a change in the value of another variable, then the variables are said to be correlated. This relationship between the variables is called Correlation. Hence two variables are said to be correlated if change in one variable is accompanied by change in other variables.

For Example: The change in quantity of irrigation brings the changes in production of rice. So irrigation and production are correlated. The measure of Correlation is called the Coefficient of Correlation. It measures the degree and direction of the relationship between the variables but it does not say which is cause and which is the effect.

A simple example is to evaluate whether there is a link between time spent for study and marks obtained in examination.To overview the relationship, first we have to check the scatterplot.

Types of Correlations

Pearson correlation (r), which measures a linear dependence between two variables (x and y). It's also known as a parametric correlation test because it depends to the distribution of the data. It can be used only when x and y are from normal distribution.
The Pearson correlation formula is:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

where m_x , m_y are means of the distributions x and y respectively.
Kendall tau and **Spearman rho**, which are rank-based correlation coefficients (non-parametric). The Spearman correlation method computes the correlation between the rank of x and the rank of y variables.

$$rho = \frac{\sum (x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum (x' - m_{x'})^2 \sum (y' - m_{y'})^2}}$$

where $m_{x'}$, $m_{y'}$ are medians of the distributions x and y respectively.
Kendal Tau is defined as

$$tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}$$

nc is nos of concordant pairs and nd is no of discordant pairs and n is total nos of obs. A pair of observations is concordant if the subject who is higher on one variable is also higher on the other variable. A pair of observations is discordant if the subject who is higher on one variable is lower on the other variable.

R Code: `cor(x, y, method = c("pearson", "kendall", "spearman"))` `cor.test(x, y, method=c("pearson", "kendall", "spearman"))`

```
cor(x, y, method = "pearson", use = "complete.obs")
```

Let us make two vector a and b having 10 numbers each

```
a <- c(70, 78, 62, 87, 84, 79, 61, 74, 83, 85)
b <- c(90, 93, 79, 99, 93, 94, 83, 92, 96, 99)
```

Then find the correlation between a and b

cor(a,b)

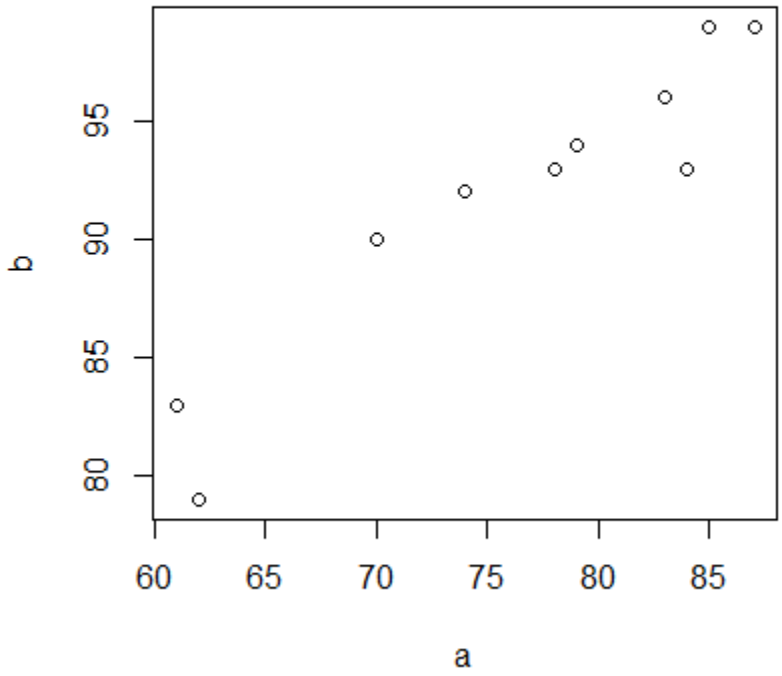
The result may be like this

```
0.9465224
```

You can see this graphically

```
plot(a,b)
```

Gives you the graph like



Similarly lets make two vector c and d having some missing values

```
c <- c(70, 78, 90, 87, 84, 79, 91, NA, 83, 85)
d <- c(90, 93, NA, 86, 84, 83, 88, 92, 76, 75)
```

Here cor(c,d) will not give result

It will give answer

NA

(because some data are missing)

In such case we have to add some characteristics like

```
cor(c,d,use="complete.obs")
```

Now this will give result

```
 -0.7764366
```

Let's make a dataframe (df) with three vectors including missing value

```
df <- data.frame(x=c(70, 78, 90, 87, 84, NA, 91, 74, 83,
     85), y=c(70, NA, 79, 86, 84, 83, 88, 72, 76, 75),
     z=c(57, 57, 58, 59, 60, 78, 81, 83, NA, 90))
```

Now lets find the correlation between variable x and y

```
cor(df$x,df$y,use="complete.obs" )
```

The result may be

```
0.8034962
```

Read data and save in data3

Lets read txt data from github. Go to https://github.com/bijayprad/R and select Happiness.txt and go to raw then copy the addressbar and use the following code to get the data

```
data3=read.delim("https://raw.githubusercontent.com/bijayprad/R/main/Happiness.txt")
```

Read So there is six variables

```
Age
Income
Education
Experience
Saving
Happyness
```

Preliminary Checks before finding pearson correlation

Is the covariation linear?

Do the data from each of the 2 variables (x, y) follow a normal distribution?

First let's attach the data using code

```
attach(data3)
```

use shapiro test for normality

```
shapiro.test(Happyness)
shapiro.test(Education)
```

Check the relation using plot

```
plot(Education,Happyness)  ## basic plot
```

Find correlation using cor code

```
cor(Education,Happyness)
cor(Experience,Happyness)
cor(saving,Happyness)
```

You can check the normality using Graphical Check

```
qqnorm(Happyness)
qqline(Happyness)
```

Now to find out correlation coefficient of all possible variables
We can use cor(filename)
In our example we can write code

```
cor(data3)
```

It gives us the correlation matrix like

```
                Age    Income Education Experience    Saving Happyness
Age        1.0000000 0.9867562 0.9125675  0.8891465 0.9758803 0.9741866
Income     0.9867562 1.0000000 0.9182890  0.8646724 0.9625769 0.9598953
Education  0.9125675 0.9182890 1.0000000  0.8101068 0.9225878 0.9037547
Experience 0.8891465 0.8646724 0.8101068  1.0000000 0.9142875 0.8679587
Saving     0.9758803 0.9625769 0.9225878  0.9142875 1.0000000 0.9724102
Happyness  0.9741866 0.9598953 0.9037547  0.8679587 0.9724102 1.0000000
```

To fix the digit after decimal we can write code

```
round(cor(data3),3)
```

It gives us the correlation matrix like

```
           Age Income Education Experience Saving Happyness
Age        1.000  0.987     0.913      0.889  0.976     0.974
Income     0.987  1.000     0.918      0.865  0.963     0.960
Education  0.913  0.918     1.000      0.810  0.923     0.904
Experience 0.889  0.865     0.810      1.000  0.914     0.868
Saving     0.976  0.963     0.923      0.914  1.000     0.972
Happyness  0.974  0.960     0.904      0.868  0.972     1.000
```

Now to test the significance of correlation
We use code cor.test
In our example we can write code

```
cor.test(Education,Happyness)
```

```
Pearson's product-moment correlation
data:  Education and Happyness
t = 14.628, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 0.8356119 0.9445047
sample estimates:
      cor
0.9037547
```

Here correlation coefficient is significant

While finding correlation test in a data file

cor(data3)

Unfortunately, the function cor() returns only the correlation coefficients between variables. We should use Hmisc R package to calculate the correlation p-values. The function rcorr() [in Hmisc package] can be used to compute the significance levels for pearson and spearman correlations. It returns both the correlation coefficients and the p-value of the correlation for all possible pairs of columns in the data table.

Install package Hmisc and call library(Hmisc)

```
library(Hmisc)
```

convert data in the matrix form

```
data4=as.matrix(data3)
```

Then find out correlation matrix along with test

```
rcorr(data4)
```

The result may be like

```
             Age Income Education Experience Saving Happyness
Age         1.00   0.99      0.91       0.89   0.98      0.97
Income      0.99   1.00      0.92       0.86   0.96      0.96
Education   0.91   0.92      1.00       0.81   0.92      0.90
Experience  0.89   0.86      0.81       1.00   0.91      0.87
Saving      0.98   0.96      0.92       0.91   1.00      0.97
Happyness   0.97   0.96      0.90       0.87   0.97      1.00
n= 50
P
            Age Income Education Experience Saving Happyness
Age               0      0          0           0      0
Income       0            0          0           0      0
Education    0    0                  0           0      0
Experience   0    0      0                       0      0
Saving       0    0      0          0                   0
Happyness    0    0      0          0           0
```

Package Corrplot

The function corrplot(), in the package of the same name, creates a graphical display of a correlation matrix, highlighting the most correlated variables in a data table.

In this plot, correlation coefficients are colored according to the value. Correlation matrix can be also reordered according to the degree of association between variables.

Correlogram can be created using the following R code:

```
install.packages("corrplot")

library(corrplot)

corrplot(cor(data3))

corrplot(cor(data3),type="upper")
```

Simple Linear Regression

Between two sets of vectors

Example IThe weight and blood pressure of 10 persons are given below-

```
x = c(65, 87, 79, 52, 89, 68, 93, 112, 76, 55)        ## weight

y = c(110, 135, 145, 120, 135, 115, 145, 125, 100, 125)    ##  blood pressure

lm(y ~ x)

reg = lm(y ~ x)

reg

### some more example man hour and lot size (check the scatter plot)

Manhr=c(30,20,60,80,40,50,60,30,70,60)

Lotsize=c(73,50,128,170,87,108,135,69,148,132)

plot(Manhr,Lotsize)
```

```
abline(lm(Lotsize~Manhr))

lm(Lotsize~Manhr)

### Next example for time to study and grade obtained

hrs=c(2,9,5,5,3,7,1,8,6,2)

grd=c(69,98,82,77,71,84,55,94,84,64)

lm(grd~hrs)

# predict the grd when hrs of study is 4

hrs1=data.frame(hrs=4)

predict(lm(grd~hrs),hrs1)
```

To display more information on linear regression carried above

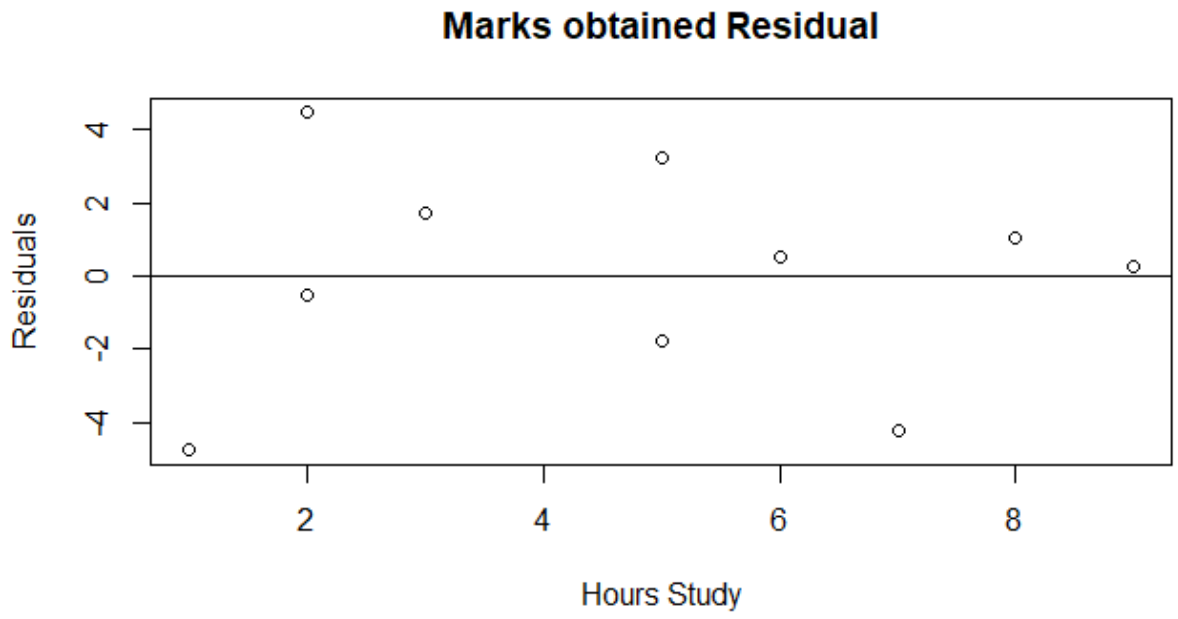
```
summary(lm(grd~hrs))
```

Regression from out dataset

```
reg1=lm(mpg~weight,data=data6)

reg1[1]        ### to see the coefficients

summary.aov(reg1)      ## ANOVA in regression

summary(reg1)       ## ## total summary of regression

## To predict when the weight is 120

newdata = data.frame(Weight = 120)

predict(reg1, newdata)
```

<u>To display residual plot</u>

```
cc = resid(reg2)

plot(hrs, cc, main="Marks obtained Residual", xlab = "Hours Study", ylab="Residuals")

abline(0,0)                    ## to display horizontal line
```



To display normal probability plot of residuals

The normal probability plot is a graphical tool for comparing a data set with the normal distribution. We can use it with the standardized residual of the linear regression model and see if the error term ε is actually normally distributed. For this purpose 'qqnorm()' function is used.

```
dd = rstandard(reg2)

qqnorm(dd, ylab="Standardized Residuals", xlab="Normal Scores", main="Marks obtained")

qqline(dd)
```

Multiple Regression

The dataset 'stackloss' is shown below –

```
attach(stackloss)

head(stackloss)
```

Apply the multiple linear regression model for the data set stackloss, and predict the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

```
aa=lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)

aa

##  To predict specified value

newdt = data.frame(Air.Flow=72, Water.Temp=20, Acid.Conc.=85)

newdt = data.frame(Air.Flow=72, Water.Temp=20, Acid.Conc.=85, data=stackloss)

predict(aa, newdt)

## To display entire information of multiple regression analysis

summary(aa)
```

Result

```
Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

To display coefficient of determination

```
summary(aa) $ r.squared
```

To display adjusted coefficient of determination

```
summary(aa) $ adj.r.squared
```

Test of Significance of Multiple Linear Regression

Use summary() function and observe the p-value.

Prediction Interval and Confidence Interval for MLR

```
attach(stackloss)
aa = lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)
newdt = data.frame(Air.Flow=72, Water.Temp=20, Acid.Conc.=85)
predict(aa, newdt, interval = "confidence")                    ## for confidence interval
predict(aa, newdt, interval = "predict")              ## for prediction interval
```

Multicollearity in Regression

We need car package to diagnose the Multicollinearity present in the model

```
library(car)
vif(summary(reg3))
summary(reg2)
head(data6,2)
attach(data6)
reg3=lm(mpg~dis+hp+weight)
vif(reg3)
reg4=lm(mpg~hp+weight)
summary(reg4)
vif(reg4)
```