Experiment-3

Applying correlation and simple linear regression model to real data set; computing and interpreting the coefficient of determination

Aim: To understand the simple correlation and linear regression with computation and interpretation

Introduction

The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation.

Correlation:

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Regression:

Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

Simple Linear Regression:

Simple linear regression model we have the following two regression lines:

1. Regression line of Y on X: This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable). Regression line of Y on X : $Y - \dot{Y} = byx\ (X - \dot{X})$ OR : $Y = a + bX$

2. Regression line of X on Y: This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable). Regression line of X on Y : $X - \dot{X} = bxy\ (Y - \dot{Y})$ OR : $X = a + bY$

In the above two regression lines or regression equations, there are two regression parameters, which are "a" and "b". Here "a" is unknown constant and "b" which is also denoted as "byx" or "bxy", is also another unknown constant popularly called as regression coefficient. Hence, these "a" and "b" are two unknown constants (fixed numerical values) which determine the position of the line completely.

Procedure:

- Input/Import the data set
- Determine the correlation and regression line using R functions
- Visualize the regression line using R functions

Code and Result:

```
# Problem-1
# Import the inbuilt data set "cars"
data=cars
data
```

```
##    speed dist
## 1      4    2
## 2      4   10
```

```
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
## 8     10   26
## 9     10   34
## 10    11   17
## 11    11   28
## 12    12   14
## 13    12   20
## 14    12   24
## 15    12   28
## 16    13   26
## 17    13   34
## 18    13   34
## 19    13   46
## 20    14   26
## 21    14   36
## 22    14   60
## 23    14   80
## 24    15   20
## 25    15   26
## 26    15   54
## 27    16   32
## 28    16   40
## 29    17   32
## 30    17   40
## 31    17   50
## 32    18   42
## 33    18   56
## 34    18   76
## 35    18   84
## 36    19   36
## 37    19   46
## 38    19   68
## 39    20   32
## 40    20   48
## 41    20   52
## 42    20   56
## 43    20   64
## 44    22   66
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

```r
# Summary of the data set
summary(data)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```r
# Variance of "speed"
v1=var(data$speed)
v1
```

```
## [1] 27.95918
```

```r
# Variance of "dist"
v2=var(data$dist)
v2
```

```
## [1] 664.0608
```

```r
# Covariance between "speed" and "dist"
covariance=cov(data$speed,data$dist)
covariance
```

```
## [1] 109.9469
```

```r
#or
covariance=var(data$speed,data$dist)
covariance
```

```
## [1] 109.9469
```

```r
# correlation coefficient using Pearson's formula
corr=covariance/(sd(data$speed)*sd(data$dist))
corr
```

```
## [1] 0.8068949
```

```r
# or
corr=cor(data$speed,data$dist)
corr
```

```
## [1] 0.8068949
```

```r
# Test for association between paired samples
cor.test(data$speed,data$dist)
```

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  data$speed and data$dist
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
```

```
cor.test(data$speed,data$dist,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$speed and data$dist
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
```

```
cor.test(data$speed,data$dist,method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$speed and data$dist
## S = 3532.8, p-value = 8.825e-14
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8303568
```

```
# Visualize the samples
plot(data$speed,data$dist)
# Linear Regression model of "speed" with respect to "dist"
regression1=lm(data$speed~data$dist)
regression1
```

```
##
## Call:
## lm(formula = data$speed ~ data$dist)
##
## Coefficients:
## (Intercept)    data$dist
##      8.2839       0.1656
```

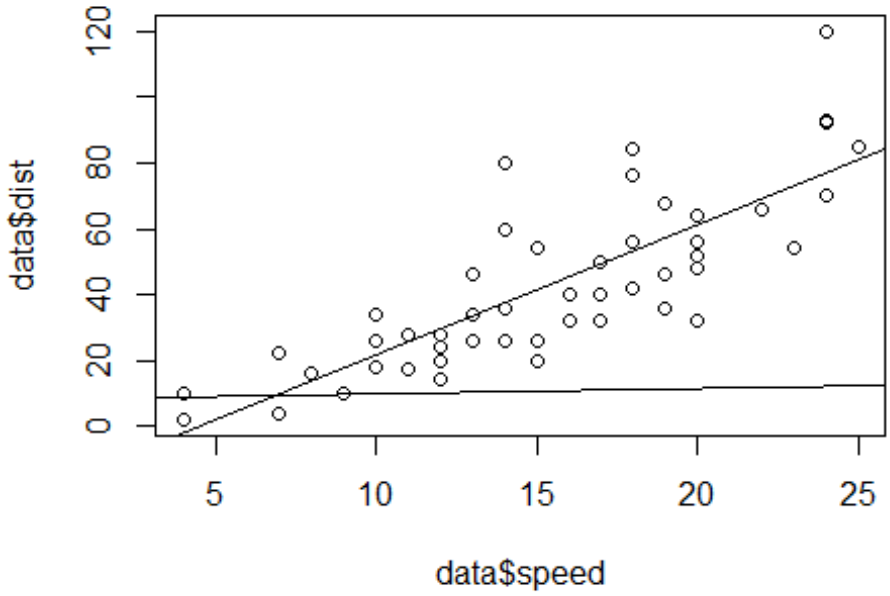
```r
# Visualize linear regression line
abline(regression1)
summary(regression1)
```

```
##
## Call:
## lm(formula = data$speed ~ data$dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5293 -2.1550  0.3615  2.4377  6.4179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.28391    0.87438   9.474 1.44e-12 ***
## data$dist    0.16557    0.01749   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.156 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
# Linear Regression model of "dist" with respect to "speed"
regression2=lm(data$dist~data$speed)
regression2
```

```
##
## Call:
## lm(formula = data$dist ~ data$speed)
##
## Coefficients:
## (Intercept)   data$speed
##     -17.579        3.932
```

```r
abline(regression2)
```




```
summary(regression2)

##
## Call:
## lm(formula = data$dist ~ data$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## data$speed    3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Problem:-

The body weight and the BMI of 12 school going children are given in the following table

Weight	***15***	***26***	***27***	***25***	***25.5***	***27***	***32***	***18***	***22***	***20***	***26***	***24***
BMI	***13.35***	***16.12***	***16.74***	***16.00***	***13.59***	***15.73***	***15.65***	***13.85***	***16.07***	***12.8***	***13.65***	***14.42***

Let us fit a simple regression model BMI on weight and examine the results.

```
#Problem-2
weight=c(15,26,27,25,25.5,27,32,18,22,20,26,24)
weight
```

```
##  [1] 15.0 26.0 27.0 25.0 25.5 27.0 32.0 18.0 22.0 20.0 26.0 24.0
```

```
bmi=c(13.35,16.12,16.74,16,13.59,15.73,15.65,13.85,16.07,12.8,13.65,14.42)
bmi
```

```
##  [1] 13.35 16.12 16.74 16.00 13.59 15.73 15.65 13.85 16.07 12.80 13.65
14.42
```

```
cor(weight,bmi)
```

```
## [1] 0.5790235
```

```
mode1<-lm(bmi~weight)
summary.lm(mode1)
```

```
##
## Call:
## lm(formula = bmi ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52988 -0.75527  0.04426  0.95286  1.57397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.73487    1.85405   5.790 0.000175 ***
## weight       0.17096    0.07612   2.246 0.048524 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.155 on 10 degrees of freedom
## Multiple R-squared:  0.3353, Adjusted R-squared:  0.2688
## F-statistic: 5.044 on 1 and 10 DF,  p-value: 0.04852
```

Interpretation :

Correlation r=0.5790,which is the correlation coefficient between the body 'weight' and BMI.There is a positive correlation between these two variables.The Value of R^2 is 0.3353,which means that about 33.53% variation in BMI can be explained by 'weight'through this linear model.This is apparently low because more than 67% of variation remains unexplained.There could be several reasons for this and one of them is that there might be some other influencing variables that have not been included in the present model.

The F value shown in the above output gives the statistics for the variance ratio test of the regression model.The significance of F, which is given as 0.0485,is the p value of the F-test carried out in ANOVA. If this value is less than 0.05 we say that the regression is statistical significant at 5% level of significance .Here regression is significant which means that the relationship is not an occurrence by chance

In the above output we find b_0 is the intercept which value of 10.73487 and b_1 is the regression coefficient due to weight with a value of 0.1710.The regression coefficient is positive ,which shows that the BMI is positively related to weight,

The regression output can be written as mathematical equation

$$BMI = 10.7349 + 0.1710 * weight$$

Suppose body weight of one student is known as 25 kg.Using the above equation,the estimated BMI is 15.01.since this is only an estimate we have to intercept it as the average BMI corresponding to the given weight assuming that other parameters are unchanged.

Conclusion: The simple correlation and linear regression equation have been computed and interpreted.