CORRELATION AND REGRESSION

Experiment-2

JUNE 18, 2021

BIMAL PARAJULI
20BDS0405

Correlation Definition:-

Correlation refers to the relationship between two or more variables. Simple correlation studies the relationship between two variables. Correlation analysis attempts to determine the degree of relationship between variables.

Measures of Correlation:

- Scatter Diagram
- Karl Pearson's Coefficient of Correlation
  It is defined as the ratio of covariance between x and y say Cov (X, Y) to the product of the standard deviations of X and Y, say σ (X) and σ (Y)

$$i.e \qquad r_{XY} = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

- SPEARMAN'S RANK CORRELATION COEFFICIENT
  Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient ρ is given by

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2 - 1)}\right]$$

- KENDALL'S COEFFICIENT OF CONCURRENT DEVIATIONS
  The Kendall's coefficient of concurrent deviations is denoted by rc and defined

$$r_c = \pm\sqrt{\pm\left[\frac{2C - n}{n}\right]}$$

  as

  Where, C = Number of concurrent deviations or position signs of (DX, DY); n = Number of pairs of deviations

Regression:

DEFINITION

Regression analysis is a statistical method of determining the mathematical functional relationship connecting independent variable(s) and a dependent variable.

Its types are:

- Simple linear Regression

  In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of relationship is linear. This relationship can be expressed using a straight line equation (linear regression) that best approximates all the individual data points.
  The general form of the simple linear regression equation is Y = a + bX + e, where 'X' is independent variable, 'Y' is dependent variable, a' is intercept, 'b' is slope of the line and 'e' is error term.

- Multiple linear Regression

  Multiple linear regression uses two or more independent variables to estimate the value(s) of the response variable (Y). The general form of the multiple linear regression equation is Y = a + b1X1 + b2X2 + b3X3 + … + btXt + e

- Non Linear Regression

Problem 1:

1. Using R obtain Correlation coefficient between X and Y and regression line of X and Y and regression line of Y on X for the following data

X	62	58	68	48	72	44	52	56
Y	68	64	75	50	64	80	40	55

R- Code:

```
Console ~/R/
> #Using R obtain Correlation coefficient between X and Y and regression line
> #of X and Y and regression line of Y on X for the following data
> # X  =  62 58 68 48 72 44 52 56
> # Y  =  68 64 75 50 64 80 40 55
>
>
> x=c(62, 58, 68, 48, 72, 44, 52, 56)
> x
[1] 62 58 68 48 72 44 52 56
> y=c(68, 84, 75, 50, 64, 80, 40, 55)
> y
[1] 68 84 75 50 64 80 40 55
>
> r=cor(x,y)
> r
[1] 0.1998941
>
> regxony=lm(x~y)
> summary.lm(regxony)                    #regression analysis of x on Y

Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-15.442  -3.743  -1.127   5.342  14.563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.4179    16.5678   2.983   0.0245 *
y             0.1253     0.2507   0.500   0.6351
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.17 on 6 degrees of freedom
Multiple R-squared:  0.03996,   Adjusted R-squared:  -0.12
F-statistic: 0.2497 on 1 and 6 DF,  p-value: 0.6351
```

```
>
> regyonx=lm(y~x)
> summary.lm(regyonx)            #regression analysis of Y on X

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-22.746  -9.634  -1.529  10.199  19.805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.1641    37.1374   1.243    0.260
x             0.3189     0.6381   0.500    0.635

Residual standard error: 16.22 on 6 degrees of freedom
Multiple R-squared:  0.03996,   Adjusted R-squared:  -0.12
F-statistic: 0.2497 on 1 and 6 DF,  p-value: 0.6351

> |
```

* Using R, obtain Correlation Coefficient between X and Y and regression line of X and Y and regression line of Y and x for following data:-

X	62	58	68	48	72	44	52	56
Y	68	64	75	50	64	80	40	55

R-codes:-

```
> X=c (62,58,68,48,72,44,52,56)
> Y=c (68,64,75,50,64,80,40,55)
> r=Cor (x,y)                # Correlation Coefficient
>r
  [1] 0.1998941
>regyonx = lm (y~x)
>summary.lm(regyonx)         # summary of regression of y on x.
>reg.xony = lm (x~y)
>summary.lm(regx on y).      # Summary of regression of x on y.
```

Problem 2:

Calculate the Coefficient of correlation of x and y from the given data:

X	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

R- Code:

```
Console ~/R/
> x=c(23,27,28,28,29,30,31,33,35,36)                  # Given X data
> x
 [1] 23 27 28 28 29 30 31 33 35 36
> y=c(18,20,22,27,21,29,27,29,28,29)                  # Given Y data
> y
 [1] 18 20 22 27 21 29 27 29 28 29
>
> var(x)                                              #variance of X
[1] 15.33333
> var(y)                                              #variance of Y
[1] 18.22222
> var(x,y)                                            #Co-variance of X and Y
[1] 13.66667
>
> var(x,y)/sqrt(var(x)*var(y))                        #Coefficient of Correlation
[1] 0.8176052
> cor.test(x,y,method="pearson")

	Pearson's product-moment correlation

data:  x and y
t = 4.0164, df = 8, p-value = 0.003861
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3874142 0.9554034
sample estimates:
      cor
0.8176052

> |
```

Calculate the coefficient of correlation of x and y from given data.

x	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

R-code:-

```
>x = c(23,27,28,28,29,30,31,33,35,36).
>y = c(18,20,22,27,21,29,27,29,28,29).
>var(x)
  [1] 15.33333
>var(y)
  [1] 18.22222
>var(x,y)
  [1] 13.6667
>r = var(x,y)./sqrt(var(x)*var(y))
>r
 [1] 0.8176052
>cor.test(x,y, method = "pearson").
```

output:-

```
            Pearson's product-moment correlation.
  data: x and y.
  t=4.0164, df=8, p-value= 0.003861
  alternative hypothesis: true correlation is not equal to 0.
  95 percent confidence interval:
    0.387142      0.9554034
  Sample estimates:
     cor
  0.8176052
```

Problem 3:

Twelve recruits were subjected to selection test to ascertain their sustainability 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	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency test Score	48	55	45	60	43	80	58	50	77	46	47	65

R- Code:

```
Console ~/R/
> selection=c(44, 49, 52, 54, 47, 76, 65, 60, 63, 58, 50, 67)        #Selection Test Score
> selection
 [1] 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)     # Proficiency Test Score
> proficiency
 [1] 48 55 45 60 43 80 58 50 77 46 47 65
>
>
> 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 

>
```

Problem :- Twelve recruits were subjected to selection test to ascertain their sustainability 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	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test score	48	55	45	60	43	80	58	50	77	46	47	65

Calculate the rank correlation coefficient and comment on result

Solution:-

```
>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.test (selection, proficiency, method= "spearman").
```

output:

```
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.
```

Problem 4:

The body weight and BMI of 12 school going children are given in the following table. Fit a simple regression model of BMI on weight and examine the results.

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

R- Code:

```
Console ~/R/
> weight=c(15, 26, 27, 25, 25.5, 27, 32, 18, 22, 20, 26, 24)          #Given weight data
> 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, 16.00, 13.59, 15.73, 15.65, 13.85, 16.07, 12.8, 13.65, 14.42)   #Given BMI data
> 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)                 #Correlation between weight and BMI
[1] 0.5790235
>
> model<- lm(bmi~weight)
> summary.lm(model)

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

>
```

Problem:-

The body weight and BMI of 12 school going children are given in the following table:- Fit a simple regression model BMI on weight and examine the results.

Weight	15	26	27	25	25.5	27	32	18	22	20	26	2
BMI	13.35	16.12	16.74	16.00	13.59	15.73	15.65	13.85	16.07	12.8	13.65	14

R-code:

```
>weight = c(15,26,27,25,25.5, 27,32,18,22,20,26,24).
>bmi = c(13.35,16.12,16.74,16.00,13.59,15.73,15.65,13.85,16.07,12.8,
13.65,14.42).
>cor (weight,bmi).
>model <- lm(bmi~weight).
>summary.lm(model).
```

Output:

```
Call:
lm(formula= bmi~weight).

Residuals:
  Min      1Q      Median   3Q     Max
-1.52988 -0.75527 0.04426 0.958 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-Statistics: 5.044 on 1 and 10 DF , p-value: 0.04852.
```