DA-2 Foundations of Data Analytics(CSE3505)

Reg.No : 20BCE1025
Name : Abhishek N N
Slot : F2
Faculty : Dr. Trilok Nath Pandey

Suppose you have the following dataset

Registration Number	Study Time in Hrs	Attendance in %	CGPA
1	4	100	6.2
2	6	50	5.3
3	16	95	9.9
4	12	85	9.0
5	18	100	10.0
6	2	50	4.0
7	5	70	6.9
8	9	80	7.7
9	15	80	8.9
10	3	75	6.5
11	7	7.5	8.0

The above table shows the marks obtained by students based on their study hours and attendance in the class.

a) Derived the multiple regression equation to predict CGPA based on study Time and Attendance.
b) Apply multiple regression to predict the CGPA of a student if he has 78% attendance and 8hr Study time.
c) Finally write an R script to perform the multiple regression and predict the CGPA of the student as per the condition given in bit (b).
d) Interpret the results and various statistics measures obtained after executing the script and attach the outputs.

Foundation of Data Analytics

Digital Assignment 2

Abhishek N N 20BCE1025

let study time = X_1, Attendance = X, CGPA = Y

Reg No	X_1	X_2	Y	$X_1-\bar{X_1}$	$X_2-\bar{X_2}$	$(X_1-X_1)^2$	$(X_2-X_2)^2$	$X_1 Y$	$X_2 \cdot Y$	$X_1 \cdot X_2$
1	4	100	6.2	-4.818	27.955	23.215	781.457	6.22	-36.087	-134.69
2	16	50	5.3	-2.818	-22.045	7.942	486.002	6.174	48.3	62.12
3	16	95	9.9	7.182	22.955	51.579	526.111	17.302	55.3	164.85
4	12	85	9.0	8.182	12.955	10.124	167.82	4.802	19.55	43.219
5	18	100	10.0	9.182	27.955	84.306	781.457	23.032	70.14	256.67
6	2	50	4.0	-6.818	-22.075	46.488	486.002	23.032	76.957	150.314
7	5	70	6.9	-3.818	-2.045	14.579	4.184	2.056	1.209	7.81
8	9	80	7.7	0.182	7.955	0.033	63.275	0.038	1.663	1.446
9	15	80	8.9	6.182	7.955	38.215	63.275	8.711	11.209	49.174
10	3	75	6.5	-5.818	9.955	33.851	8.729	5.765	-2.918	17.19
11	7	75	8.0	-1.818	-64.544	3.306	4165.927	-0.926	-32.816	117.355

$\varepsilon X_1 = 97$, $\varepsilon X_2 = 792.5$, $\varepsilon y = 82.4$, $\bar{X_1} = 8.8182$

$\bar{X_2} = 72.0455$, $\bar{Y} = 7.4909$

$\varepsilon X_1^2 = 313.636$, $\varepsilon X_2^2 = 7535.227$

$\varepsilon X_1 Y = 97.182$, $\varepsilon X_2 Y = 212.455$

$\varepsilon X_1 \cdot X_2 = 699.091$

$$b_1 = \frac{(\varepsilon X_1 Y)(\varepsilon X_2^2) - (\varepsilon X_1 X_2)(\varepsilon X_2 \cdot Y)}{(\varepsilon X_1^2)(\varepsilon X_2^2) - (\varepsilon X_1 X_2)^2}$$

$$= \frac{97.182 \times 7535.227 - 699.091 \times 212.455}{313.636 \times 7535.227 - (699.091)^2}$$

$b_1 = 0.31141$ //

$$b_2 = \frac{(\varepsilon X_1^2)(\varepsilon X_2 \cdot Y) - (\varepsilon X_1 X_2)(\varepsilon X_2 Y)}{(\varepsilon X_1^2)(\varepsilon X_2^2) - (\varepsilon X_1 X_2)^2}$$

$$b_2 = \frac{313.636 \times 212.455 - 699.091 \times 212.455}{313.636 \times 7535.227 - (699.091)^2}$$

$b_2 = -0.007$ //

$a = \overline{Y} - b_1\overline{X_1} - b_2\overline{X_2}$

$= 7.49 - (0.31 \times 8.82) - (72.05 * (-0.0007))$

$= 4.79503$

∴ Regression equation

$\hat{Y} = 0.31141 * X_1 - 0.0007 * X_2 + 4.79503$

(b) Apply multiple regression to predict the CGPA of a student if he has 78.7 attendance and 8 Hrs of study time

Sol Given $X_1 = 8$, $X_2 = 78$, $Y = ?$

∴ from equation found in (a)

$\hat{Y} = 0.3141 * X_1 - 0.0007 X_2 + 4.79503$

$= 0.3141 * 8 - 0.0007 * 78.7 + 4.79503$

$= 7.23171$

So predicted CGPA is 7.23

c) Finally write an R script to perform the multiple regression and predict the CGPA of the student as per the condition given in bit (b).

```
> df <- data.frame(
+     studyHr = c(4,6,16,12,18,2,5,9,15,3,7),
+     attendance = c(100,50,95,85,100,50,70,80,80,75,7.5),
+     cgpa = c(6.2,5.3,9.9,9.0,10.0,4.0,6.9,7.7,8.9,6.5,8.0)
+ )
> df
   studyHr attendance cgpa
1        4      100.0  6.2
2        6       50.0  5.3
3       16       95.0  9.9
4       12       85.0  9.0
5       18      100.0 10.0
6        2       50.0  4.0
7        5       70.0  6.9
8        9       80.0  7.7
9       15       80.0  8.9
10       3       75.0  6.5
11       7        7.5  8.0
```

```
> model <- lm(cgpa~studyHr+attendance,data=df)
> model

Call:
lm(formula = cgpa ~ studyHr + attendance, data = df)

Coefficients:
(Intercept)      studyHr   attendance
  4.7950347    0.3114073   -0.0006964
```

```
> predictionDf <- data.frame(
+     studyHr = c(8),
+     attendance = c(78)
+ )
> predict(model, newdata = predictionDf)
       1
7.231975
```

d) Interpret the results and various statistics measures obtained after executing the script and attach the outputs.

```
> summary(model)

Call:
lm(formula = cgpa ~ studyHr + attendance, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3830 -0.4206  0.1886  0.5620  1.0303 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.7950347  0.8073421   5.939 0.000346 ***
studyHr      0.3114073  0.0572976   5.435 0.000620 ***
attendance  -0.0006964  0.0116896  -0.060 0.953957    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9037 on 8 degrees of freedom
Multiple R-squared:  0.8217,	Adjusted R-squared:  0.7771 
F-statistic: 18.44 on 2 and 8 DF,  p-value: 0.00101

> summary(df)
    studyHr         attendance          cgpa       
 Min.   : 2.000   Min.   :  7.50   Min.   : 4.000  
 1st Qu.: 4.500   1st Qu.: 60.00   1st Qu.: 6.350  
 Median : 7.000   Median : 80.00   Median : 7.700  
 Mean   : 8.818   Mean   : 72.05   Mean   : 7.491  
 3rd Qu.:13.500   3rd Qu.: 90.00   3rd Qu.: 8.950  
 Max.   :18.000   Max.   :100.00   Max.   :10.000  
```

From the linear model we got the predicted value of 7.231975 which is matching with calculation result got at question (b)

While checking the sumarry of the linear model created, residuals and coefficients are obtained.

Residuals are essentially the difference between actual observed response values and response ~~e~~ value that model predicted.

We can observe that difference of minimum and maximum as difference greater than 1 and ~~Q~~ median and 3Q are in between 0 to 1 only. Hence we can that dispersion measures are consistent and hence, the model ~~ean~~-cover most of the data point and accuracy is greater.

The Fstatistic ÷ 18.44 and p-value ÷ 0.00101 so result ~~is~~ p is less than 0.05 then the result is significant.