**DATA SCIENCE &**

**BUSINESS ANALYTICS**

**INTERN**

**@THESPARKFOUNDATION**

**TASK1- SIMPLE LINEAR REGRESSION ANALYSIS**

**By Namrata Bhosale**

**Prediction using supervised ML**

Analysis of data :

* Predict the percentage of an students base on the number of study hours
* This is a simple linear regression task as it involves just two variables.

> Y=c(21,47,27,75,30,20,88,60,81,25,85,62,41,42,17,95,30,24,67,69,30,54,35,76,86)

> X=c(2.5,5.1,3.2,8.5,3.5,1.5,9.2,5.5,8.3,2.7,7.7,5.9,4.5,3.3,1.1,8.9,2.5,1.9,6.1,7.4,2.7,4.8,3.8,6.9,7.8)

> model=lm(Y~X)

> model

Call:

lm(formula = Y ~ X)

Coefficients:

(Intercept) X

2.484 9.776

> Y=2.484+9.776\*X

> mean(X)

[1] 5.012

> mean(Y)

[1] 51.48131

* **if students study at least 5 hours per day then they will get minimum 51 score in exam**

> summary(model)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-10.578 -5.340 1.839 4.593 7.265

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.4837 2.5317 0.981 0.337

X 9.7758 0.4529 21.583 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.603 on 23 degrees of freedom

Multiple R-squared: 0.9529, Adjusted R-squared: 0.9509

F-statistic: 465.8 on 1 and 23 DF, p-value: < 2.2e-16

* the final three lines are model diagnostics - the most important think to note is the p value (here it is 2.2e-16,or almost 0),which will indicate whether the model fits the data well. from this result, we can say that there is a significant positive relationship between score and hours .

> anova(model)

Analysis of Variance Table

Response: Y

Df Sum Sq Mean Sq F value Pr(>F)

X 1 14624.2 14624.2 465.82 < 2.2e-16 \*\*\*

Residuals 23 722.1 31.4

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* R-squared value, it represents the proportion of variance in the dependent variable that can be explained by independent variable. Higher R-Squared value indicates better fit of model to the data. From our data R-Squared value is 0.95 i.e close to 1 indicates stronger relationship between the scores and hours.

Data visualization:

> plot(X,Y,xlab="Hours(X)",ylab="scores(Y)",main="scatter plot")



* the distribution of data point could be described with a straight line, hence the relationship between the hours and scores variable must be linear .

Prediction:

> yi=predict(model)

> yi

1 2 3 4 5 6 7 8

26.92318 52.34027 33.76624 85.57800 36.69899 17.14738 92.42106 56.25059

9 10 11 12 13 14 15 16

83.62284 28.87834 77.75736 60.16091 46.47479 34.74382 13.23706 89.48832

17 18 19 20 21 22 23 24

26.92318 21.05770 62.11607 74.82462 28.87834 49.40753 39.63173 69.93672

25

78.73494

> data=data.frame(Y,yi)

> data

Y yi

1 21 26.92318

2 47 52.34027

3 27 33.76624

4 75 85.57800

5 30 36.69899

6 20 17.14738

7 88 92.42106

8 60 56.25059

9 81 83.62284

10 25 28.87834

11 85 77.75736

12 62 60.16091

13 41 46.47479

14 42 34.74382

15 17 13.23706

16 95 89.48832

17 30 26.92318

18 24 21.05770

19 67 62.11607

20 69 74.82462

21 30 28.87834

22 54 49.40753

23 35 39.63173

24 76 69.93672

25 86 78.73494

>plot(Y,yi, ,xlab="Actual value",ylab="predicted value",main="Actual vs predicted plot")



> x=9.25

> Y=2.484+9.776\*x

> Y

[1] 92.912

* **if student study maximum 9.25 hours per day then there score will be 92.912.**

**From this analysis, we can pridict that the student get high score in exam if they study more hours.**