Data Science and Business Analytics

Task of Spark Foundation

Topic : Simple Linear Regression Analysis

Student Name : Shruti Rajendra Divekar

> students=read.csv("C:/Users/atula/OneDrive/Desktop/spark.csv")

> students

Hours Scores

1 2.5 21

2 5.1 47

3 3.2 27

4 8.5 75

5 3.5 30

6 1.5 20

7 9.2 88

8 5.5 60

9 8.3 81

10 2.7 25

11 7.7 85

12 5.9 62

13 4.5 41

14 3.3 42

15 1.1 17

16 8.9 95

17 2.5 30

18 1.9 24

19 6.1 67

20 7.4 69

21 2.7 30

22 4.8 54

23 3.8 35

24 6.9 76

25 7.8 86

> x=students$Hours

> x

[1] 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

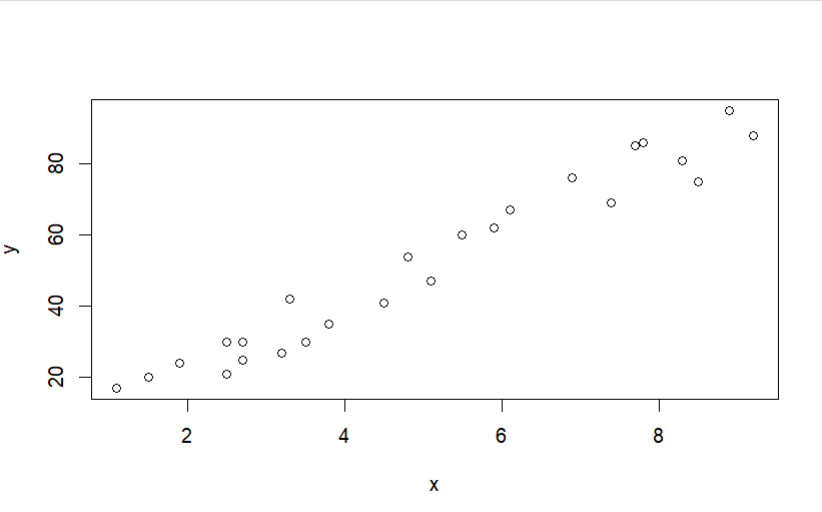
[24] 6.9 7.8

> y=students$Scores

> y

[1] 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

> plot (x, y)



# There is linear relationship between scores and hours

> model=lm(y~x)

> model

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

2.484 9.776

# we find R-squared adjusted R-squared

> 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

> #R-squared=0.9529, adjusted R-squared=0.9509

> #simple linear regression model is given by Y=2.484+9.776X

> #suppose we take the value X=9.25

> X=9.25

> X

[1] 9.25

> Y=2.484+9.776\*X

> Y

[1] 92.912

**# Therefore the predicted score of a student if they studied for 9.25 hour per day is 92.912**