**>data=read.csv("https://raw.githubusercontent.com/AdiPersonalWorks/Random/master/student\_scores%20-%20student\_scores.csv")**

**>data**

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

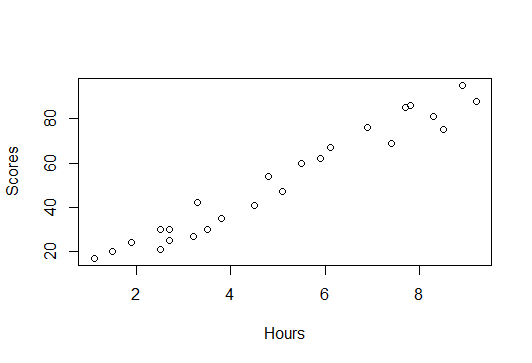
23 3.8 35

24 6.9 76

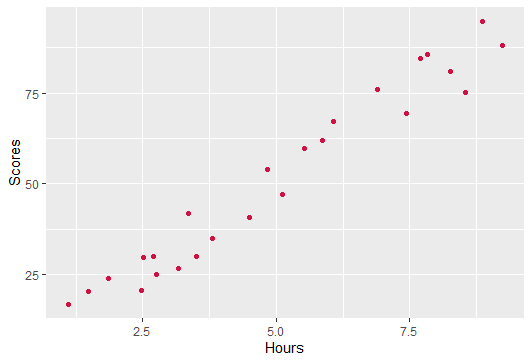
25 7.8 86

**> plot(data)**

>

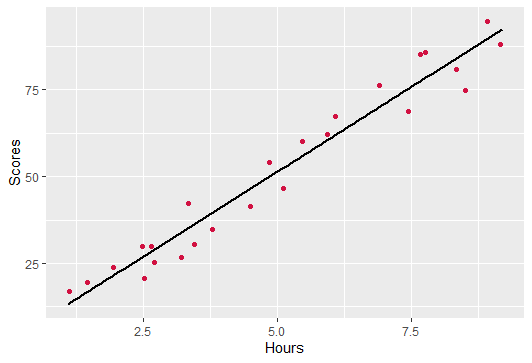


**> sjPlot::plot\_scatter(data,Hours,Scores)**



**> sjPlot::plot\_scatter(data,Hours,Scores,fit.line = "lm")**

`geom\_smooth()` using formula 'y ~ x'



**> cor(data)**

Hours Scores

Hours 1.0000000 0.9761907

Scores 0.9761907 1.0000000

**From the above value, correlation between these two variables is positive and they strongly correlated. That is Any change in Hours impacts the Scores of the student.**

**Accuracy check:**

**T-test: (p-value)**

**H0= Any change in Hours doesnot impact the Scores of the student.**

**H1= Any change in Hours impacts the Scores of the student.**

**F-test: (p-value)**

**H0=Model is not good.**

**H1= Model is good.**

**> ##Build a linear regression model - simple**

**> ##here score is a dependent variable y and Hours is an inependent variable x**

**> model=lm(Scores~Hours,data=data)**

**> summary(model)**

Call:

lm(formula = Scores ~ Hours, data = data)

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

Hours 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

**> jtools::summ(model)**

MODEL INFO:

Observations: 25

Dependent Variable: Scores

Type: OLS linear regression

MODEL FIT:

**F(1,23) = 465.82, p = 0.00**

**R² = 0.95**

**Adj. R² = 0.95**

Standard errors: OLS

-----------------------------------------------

Est. S.E. t val. p

----------------- ------ ------ -------- ------

(Intercept) **2.48** 2.53 0.98 0.34

Hours **9.78**  0.45 21.58 **0.00**

Eqn – y=a+b(x)

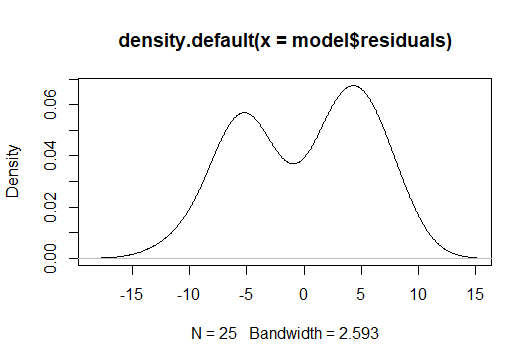
Scores=a+b(Hours)

Scores=2.48+9.78(Hours)

**> ##here, R2 value = 0.95 an p value is less than 0.05 so our assumption that student study hours influences his score is right.**

**Assumptions check:**

**> plot(density(model$residuals))**



**> shapiro.test(model$residuals)**

Shapiro-Wilk normality test

data: model$residuals

W = 0.90401, **p-value = 0.02246**

**Here, H0: residuals are normally distributed.**

**H1: residuals are not normally distributed**

**From the p-value, which is less than 0.05, H1 is accepted.**

**residuals are not normally distributed**

**> performance::check\_autocorrelation(model)**

OK: Residuals appear to be independent and not autocorrelated (p = 0.168).

**> performance::check\_heteroscedasticity(model)**

OK: Error variance appears to be homoscedastic (p = 0.283).

**> ## now, to predict the student score when he studies 9.25hrs/day is**

**> predict(model,newdata=data.frame(Hours=9.25))**

1

**Output: -92.90985**

> data=log(data)

> data

Hours Scores

1 0.91629073 3.044522

2 1.62924054 3.850148

3 1.16315081 3.295837

4 2.14006616 4.317488

5 1.25276297 3.401197

6 0.40546511 2.995732

7 2.21920348 4.477337

8 1.70474809 4.094345

9 2.11625551 4.394449

10 0.99325177 3.218876

11 2.04122033 4.442651

12 1.77495235 4.127134

13 1.50407740 3.713572

14 1.19392247 3.737670

15 0.09531018 2.833213

16 2.18605128 4.553877

17 0.91629073 3.401197

18 0.64185389 3.178054

19 1.80828877 4.204693

20 2.00148000 4.234107

21 0.99325177 3.401197

22 1.56861592 3.988984

23 1.33500107 3.555348

24 1.93152141 4.330733

25 2.05412373 4.454347

> model=lm(Scores~Hours,data=data)

> summary(data)

Hours Scores

Min. :0.09531 Min. :2.833

1st Qu.:0.99325 1st Qu.:3.401

Median :1.56862 Median :3.850

Mean :1.46346 Mean :3.810

3rd Qu.:2.00148 3rd Qu.:4.317

Max. :2.21920 Max. :4.554

> jtools::summ(data)

Error in UseMethod("summ") :

no applicable method for 'summ' applied to an object of class "data.frame"

> model

Call:

lm(formula = Scores ~ Hours, data = data)

Coefficients:

(Intercept) Hours

2.5171 0.8834

> shapiro.test(model$residuals)

Shapiro-Wilk normality test

data: model$residuals

W = 0.9336, p-value = 0.1052

> model2=lm(Scores~Hours,data=data)

> model2

Call:

lm(formula = Scores ~ Hours, data = data)

Coefficients:

(Intercept) Hours

2.5171 0.8834

> shapiro.test(model$residuals)

Shapiro-Wilk normality test

data: model$residuals

W = 0.9336, p-value = 0.1052

> From the p-value, which is greater than 0.05, H0 is accepted.

residuals are normally distributed

> predict(model,newdata=data.frame(Hours=9.25))

1

10.6882

> data=exp(data)

> data

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