Linear Regression with R

In this section we will learn to use R library in Machine learning for prediction using Supervised ML. Here, I consider simple linear regression problem with 2 variable i.e, dependent and independent variable.

Simple Linear Regression

The simple linear regression is used to predict dependent variable (y) on the basis of single independent/predictor variable (x). It is a mathematical model that defines y as a function of x variable.
In this task we will predict the expected percentage of marks gain by student based on number of hours they studied. In this task , there is only two variable (Hours is independent/predictor variable and Scores in dependent variable). Hence, it is a simple linear regression problem.

R Code

Import require package for this problem.

```
{library(readxl)\\
library(ggplot2)\\
library(tidyverse)}\\
```

Import the data

```
url<-"http://bit.ly/w-data"
data<- read_excel("TSF.xlsx",sheet = "Sheet1")
head(data)
```

```
# A tibble: 6 x 2
   Hours Scores
   <dbl>  <dbl>
1   2.5     21
2   5.1     47
3   3.2     27
```

```
4   8.5     75
5   3.5     30
6   1.5     20
```

Successfully imported data

Then, we will read the nature and structure of data i.e., summary of imported data

```
summary(data)
```

```
Hours            Scores
Min.   :1.100   Min.   :17.00
1st Qu.:2.700   1st Qu.:30.00
Median :4.800   Median :47.00
Mean   :5.012   Mean   :51.48
3rd Qu.:7.400   3rd Qu.:75.00
Max.   :9.200   Max.   :95.00
```

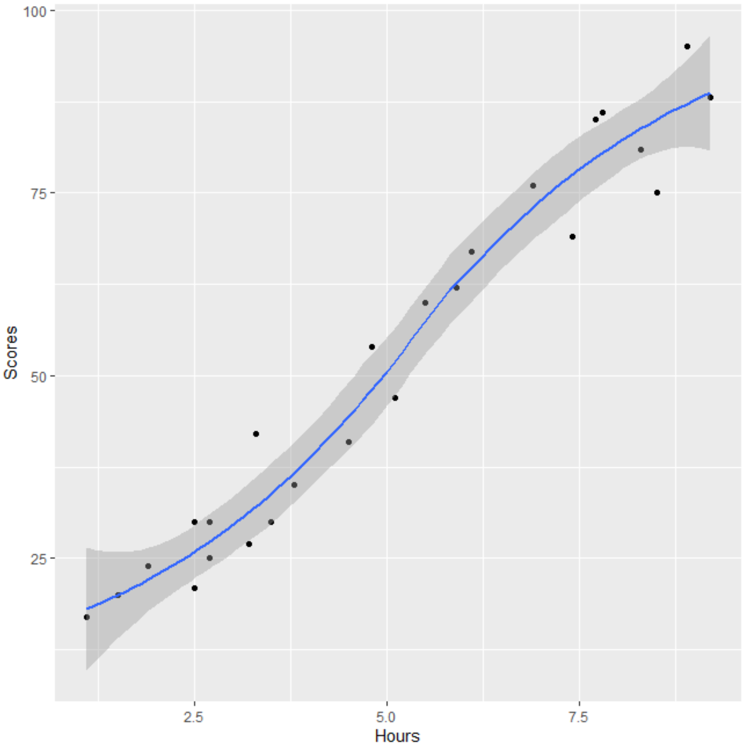
```
str(data) # structure of data
```

```
tibble [25 x 2] (S3: tbl_df/tbl/data.frame)
$ Hours : num [1:25] 2.5 5.1 3.2 8.5 3.5 1.5 9.2 5.5 8.3 2.7 ...
$ Scores: num [1:25] 21 47 27 75 30 20 88 60 81 25 ...
```

Both variable have numeric structure.

Now, we plot 2-d scatter plot of provided data and plot a smooth curve for it.

```
ggplot(data,aes(x=Hours,y=Scores))+geom_point()+geom_smooth()
```



Above plot represent high correlation between two variable and the correlation is

```
cor(data$Scores,data$Hours)
```

```
 0.9761907
```

Simple linear regression tries to find the best predicted line on the basis no of hours student studied daily.
The linear model equation can be written as

$$Scores = b_0 + b_1 Hours$$

using lm() we will determine the beta coefficient of model

```
 model<-lm(data$Scores~data$Hours,data = data);model
```

```
Call:
lm(formula = data$Scores ~ data$Hours, data = data)

Coefficients:
(Intercept)   data$Hours
2.484        9.776
```

Interpretation

```
summary(model)


Call:
lm(formula = data$Scores ~ data$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
data$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
```

Calculate predicted value for given model

```
 predicted<-predict(model)
 Data<-data.frame(data,predicted);Data


    Hours Scores predicted
1     2.5     21  26.92318
2     5.1     47  52.34027
3     3.2     27  33.76624
4     8.5     75  85.57800
5     3.5     30  36.69899
6     1.5     20  17.14738
7     9.2     88  92.42106
8     5.5     60  56.25059
9     8.3     81  83.62284
10    2.7     25  28.87834
11    7.7     85  77.75736
12    5.9     62  60.16091
```

```
13   4.5      41   46.47479
14   3.3      42   34.74382
15   1.1      17   13.23706
16   8.9      95   89.48832
17   2.5      30   26.92318
18   1.9      24   21.05770
19   6.1      67   62.11607
20   7.4      69   74.82462
21   2.7      30   28.87834
22   4.8      54   49.40753
23   3.8      35   39.63173
24   6.9      76   69.93672
25   7.8      86   78.73494
```

Now, predict score if a student studies 9.25 hours/day.

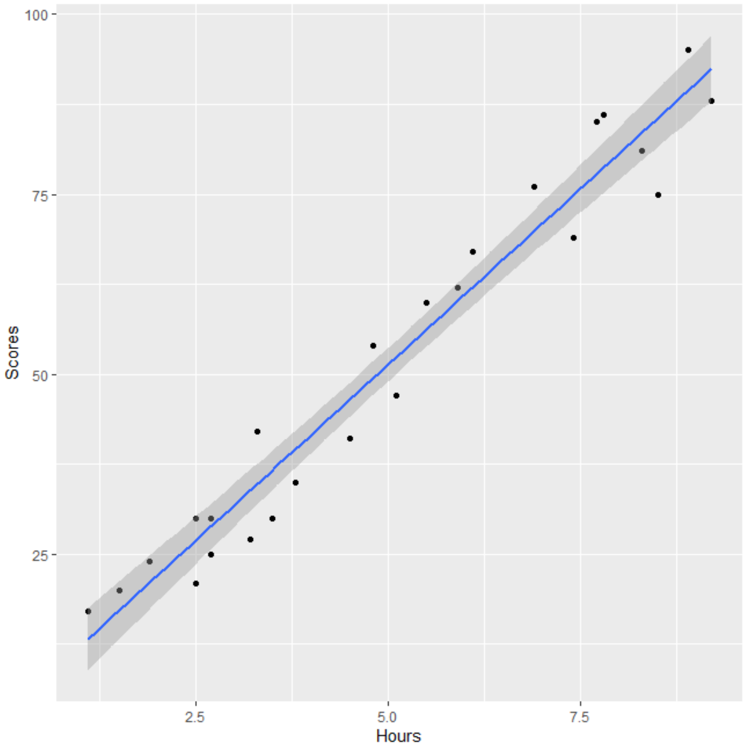
```
 hours<-9.25
 est_scores<-model$coef[1]+model$coef[2]*hours;est_scores
```

```
(Intercept)
92.90985
```

```
ggplot(data,aes(x=Hours,y=Scores))+geom_point()+stat_smooth(method = lm)
```

so, 92.90985 will be the predicted score if a student studies for 9.25 hours daily. This is predicted using provided data.

Plot best fit line to the scatter plot.



Using plot also we can observe that a student can score approximately 92 if they study 9.25 hours per day.