Marwadi University	**Marwadi University** **Faculty of Technology** **Department of Information and Communication Technology**	
Subject: Introduction to R and R Studio (01CT0106)	**Aim: Obtain the best fit line using linear regression in R**	
Experiment: 10	**Date: 13/04/2023**	**Enrollment No: 92200133030**

Aim: Obtain the best fit line using linear regression in R

IDE: R Studio

Theory:

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variables is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.

The general mathematical equation for a linear regression is −

y = ax + b

Following is the description of the parameters used −

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

Linear regression is a method for determining the best linear relationship between two variables *X* and *Y*. If variables *X* and *Y* are uncorrelated, it is pointless embarking upon linear regression. However, if a reasonable degree of correlation exists between *X* and *Y* then linear regression may be a useful means to describe the relationship between the two variables. The usual approach is to use the *least-squares* method, which minimizes the squared difference between the actual data points and a straight line. Let [*xi,yi*], *i* = 1,2,3,….,*N* be the *N* pairs of data values of the variables *X* and *Y*. The straight-line relating *X* and *Y* is *y* = *mx* + *c*, where *m* and *c* are the gradient and constant values (to be determined) defining the straight line. Thus, *y(xi)* - *yi* is the difference between the line and data point *i* (see Fig. 1). Taking all the data points, we seek values of *m* and *c* that minimize the squared difference *SD.*

$$\sum_{1}^{N}\left[y(x_i)-y_i\right]^2$$

This is achieved by calculating the partial derivatives of *SD* with respect to *m* and *c* and finding the pair [*m*,*c*] such that *SD* is at a minimum.

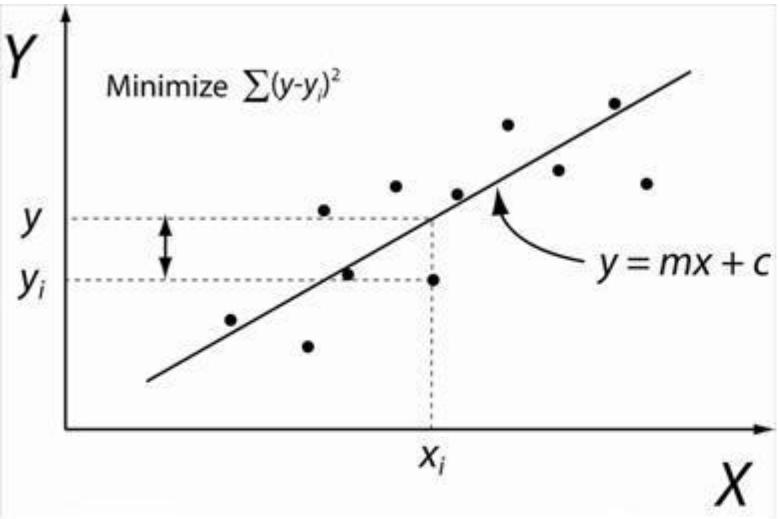


Figure 1: Illustration of Linear Regression. Linear least squares regression, the idea is to find the line y = mx + c that minimizes the mean squared difference between the line and the data points

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for lm() function in linear regression is −

```
lm(formula,data)
```

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between x and y.
- **data** is the vector on which the formula will be applied.

predict() Function

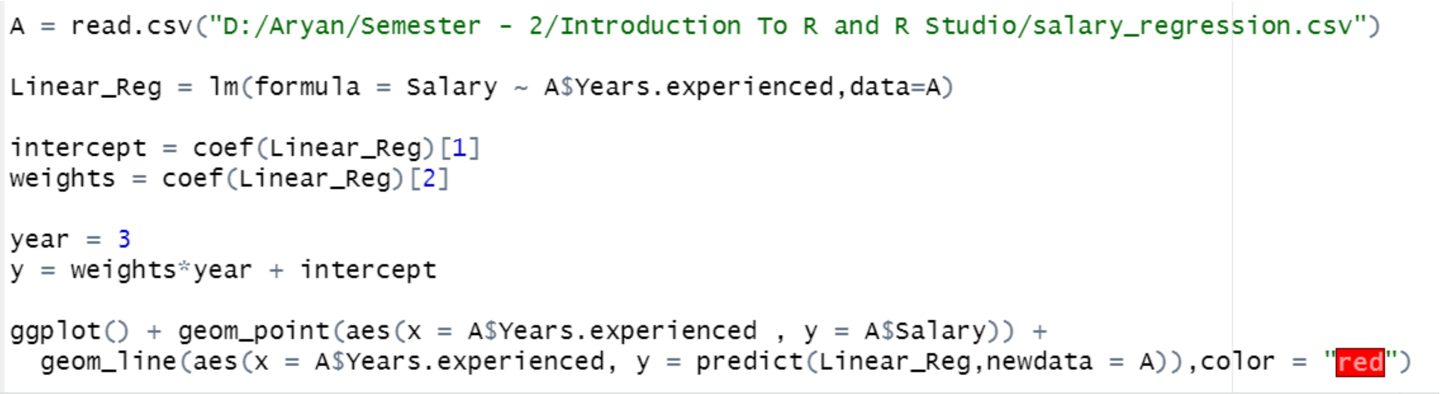
Now, we will predict the weight of new persons with the help of the predict() function. There is the following syntax of predict function:

```
predict(object, newdata)
```

Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Introduction to R and R Studio (01CT0106)	**Aim: Obtain the best fit line using linear regression in R**	
Experiment: 10	**Date: 13/04/2023**	**Enrollment No: 92200133030**

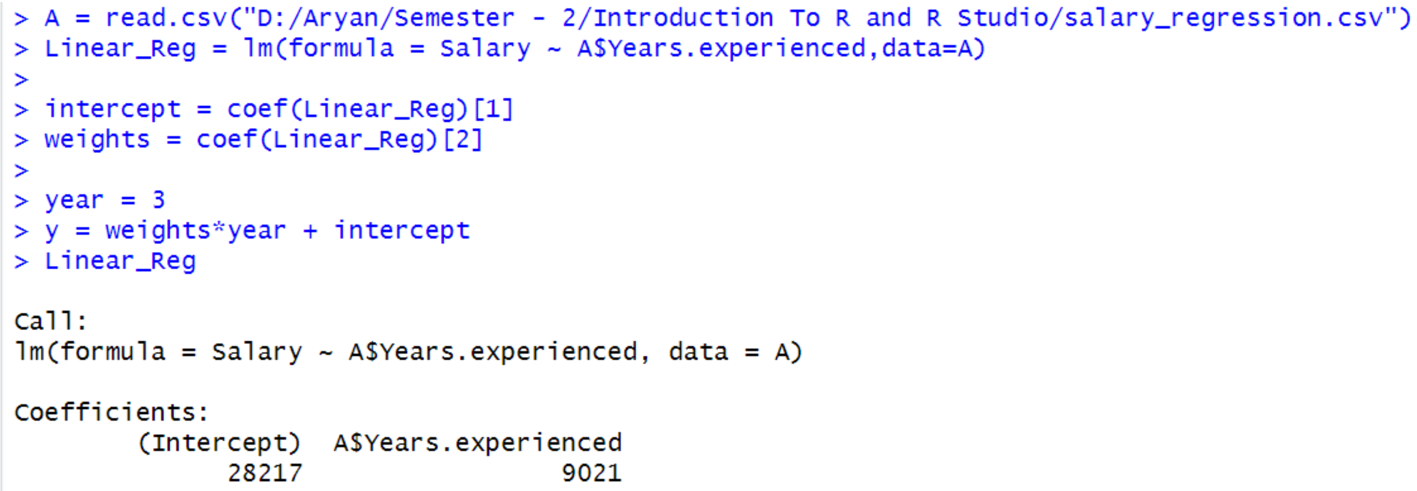
Program:

Write a R script to obtain the best fit line using linear regression

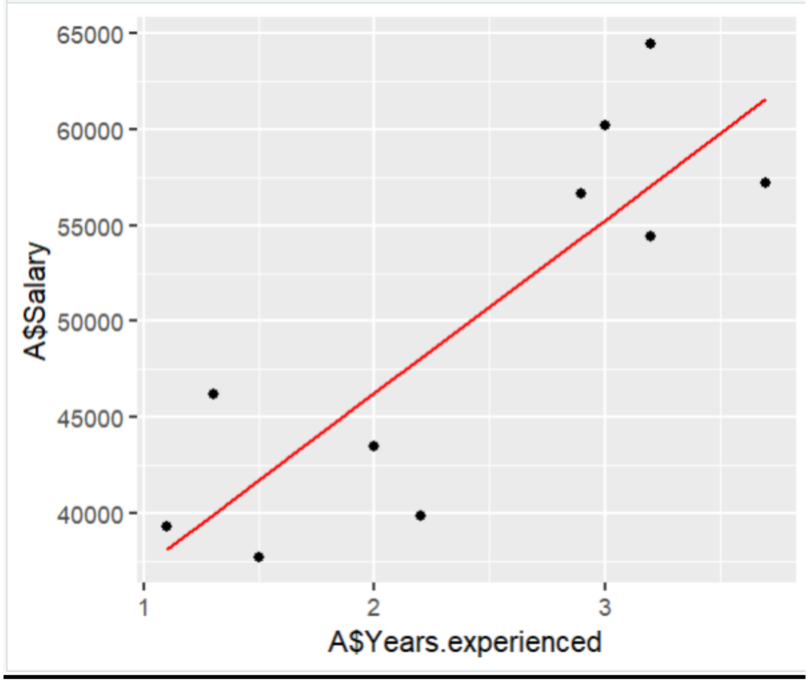


```
A = read.csv("D:/Aryan/Semester - 2/Introduction To R and R Studio/salary_regression.csv")

Linear_Reg = lm(formula = Salary ~ A$Years.experienced,data=A)

intercept = coef(Linear_Reg)[1]
weights = coef(Linear_Reg)[2]

year = 3
y = weights*year + intercept

ggplot() + geom_point(aes(x = A$Years.experienced , y = A$Salary)) +
  geom_line(aes(x = A$Years.experienced, y = predict(Linear_Reg,newdata = A)),color = "red")
```

Output:



```
> A = read.csv("D:/Aryan/Semester - 2/Introduction To R and R Studio/salary_regression.csv")
> Linear_Reg = lm(formula = Salary ~ A$Years.experienced,data=A)
> 
> intercept = coef(Linear_Reg)[1]
> weights = coef(Linear_Reg)[2]
> 
> year = 3
> y = weights*year + intercept
> Linear_Reg

Call:
lm(formula = Salary ~ A$Years.experienced, data = A)

Coefficients:
        (Intercept)  A$Years.experienced  
              28217                 9021  
```



Observation and Learnings: