Sanjivani Rural Education Society's

College of Engineering, Kopargaon-423603

DEPARTMENT OF COMPUTER ENGINEERING

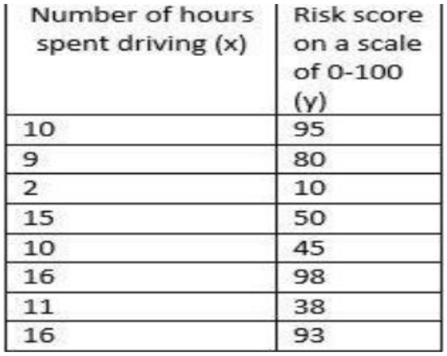
Instruction No. 01 and 02 ML Lab / Sr. No.01 and 02 Rev 00 Date: 27/12/17	**Title: Assignment on Simple Linear Regression**

Aim:

Implement Simple Linear Regression on given Problem

Problem Definition/Objective:

The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data.



Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

Input:CSV Dataset

Answer-

y = 4.59x + 12.58

Hints: For each x calculate the value of y using the given equations. Then calculate error for each equation. Equation with lowest error is the desired answer. For error calculation

Given (x1,y1),(x2,y2),…,(xn,yn), best fitting data to y = f(x) by least squares requires minimization of

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2$$

Outcomes:

After completion of this assignment students are able to understand the How to find the correlation between to Two variable, How to Calculate Accuracy of the Linear Model and how to plot graph using **matplotlib.**

Theory:

Linear Regression

Regression analysis is used in stats to find trends in data. For example, you might guess that there's a connection between how much you eat and how much you weight; regression analysis can help you quantify that.

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

Prerequisites for Regression:

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable *Y* has a linear relationship to the independent variable *X*. To check this, make sure that the XY scatterplot is linear and that the residual plotshows a random pattern. For each value of X, the probability distribution of Y has the same standard deviation σ.
- When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.
- For any given value of X,
    - The Y values are independent, as indicated by a random pattern on the residual plot.
    - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogramor a dotplot will show the shape of the distribution.

The Least Squares Regression Line:

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

$$\mathbf{Y = B_0 + B_1X}$$

Where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\mathbf{\hat{y} = b_0 + b_1x}$$

Where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and $\hat{y}$ is the *predicted* value of the dependent variable.

How to Define a Regression Line:

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator- to find b_0 and b_1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter. In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$B_1 = \Sigma\ [\ (x_{i\,–}\ x)(y_i - y)\ ]\ /\ \Sigma\ [\ (x_i - x)^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = y – b_1 * x$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the *X* value of observation *i*, y_i is the *Y* value of observation *i*, x is the mean of *X*, y is the mean of *Y*, s_x is the standard deviation of *X*, and s_y is the standard deviation of *Y*.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{\ (\ 1\ /\ N\ )\ *\ \Sigma\ [\ (x_i - x)\ *\ (y_i - y)\ ]\ /\ (\sigma_x * \sigma_y)\ \}^2$$

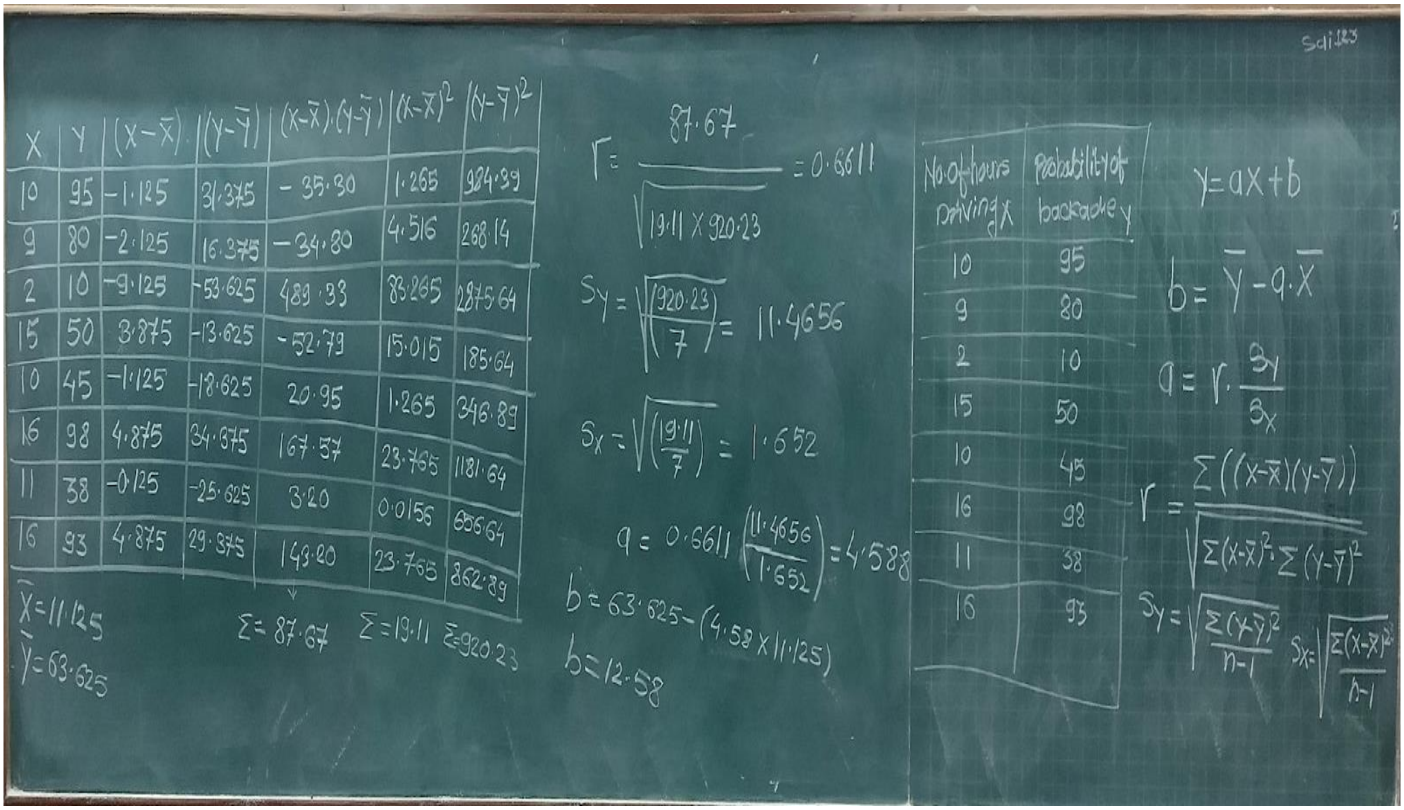
where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$.

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Manual Solution:



X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X}).(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
10	95	-1.125	31.375	-35.30	1.265	984.39
9	80	-2.125	16.375	-34.80	4.516	268.14
2	10	-9.125	-53.625	489.33	83.265	2875.64
15	50	3.875	-13.625	-52.79	15.015	185.64
10	45	-1.125	-18.625	20.95	1.265	346.89
16	98	4.875	34.375	167.57	23.765	1181.64
11	38	-0.125	-25.625	3.20	0.0156	656.64
16	93	4.875	29.375	143.20	23.765	862.89

$\Sigma = 87.67$ $\quad \Sigma = 19.11$ $\quad \Sigma = 920.23$

$\bar{X} = 11.125$

$\bar{Y} = 63.625$

$$r = \frac{87.67}{\sqrt{19.11 \times 920.23}} = 0.6611$$

$$S_y = \sqrt{\left(\frac{920.23}{7}\right)} = 11.4656$$

$$S_x = \sqrt{\left(\frac{19.11}{7}\right)} = 1.652$$

$$a = 0.6611\left(\frac{11.4656}{1.652}\right) = 4.588$$

$$b = 63.625 - (4.58 \times 11.125)$$

$$b = 12.58$$

Sai123

No.of hours driving x	Probability of backache y
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

$$y = ax + b$$

$$b = \bar{y} - a.\bar{x}$$

$$a = r.\frac{S_y}{S_x}$$

$$r = \frac{\Sigma((x-\bar{x})(y-\bar{y}))}{\sqrt{\Sigma(x-\bar{x})^2 . \Sigma(y-\bar{y})^2}}$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} \quad S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Example:

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

How to Find the Regression Equation?

In the table below ,the xi column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

And finally, for each student, we need to compute the product of the deviation scores.

Student	x_i	y_i	$(x_i-x)(y_i-y)$
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96
5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y – b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$

How to Use the Regression Equation:

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (*x*), perform the computation, and you have an estimated value (ŷ) for the dependent variable. In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade (ŷ) would be:

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

Algorithm:

1.Import the Required Packages

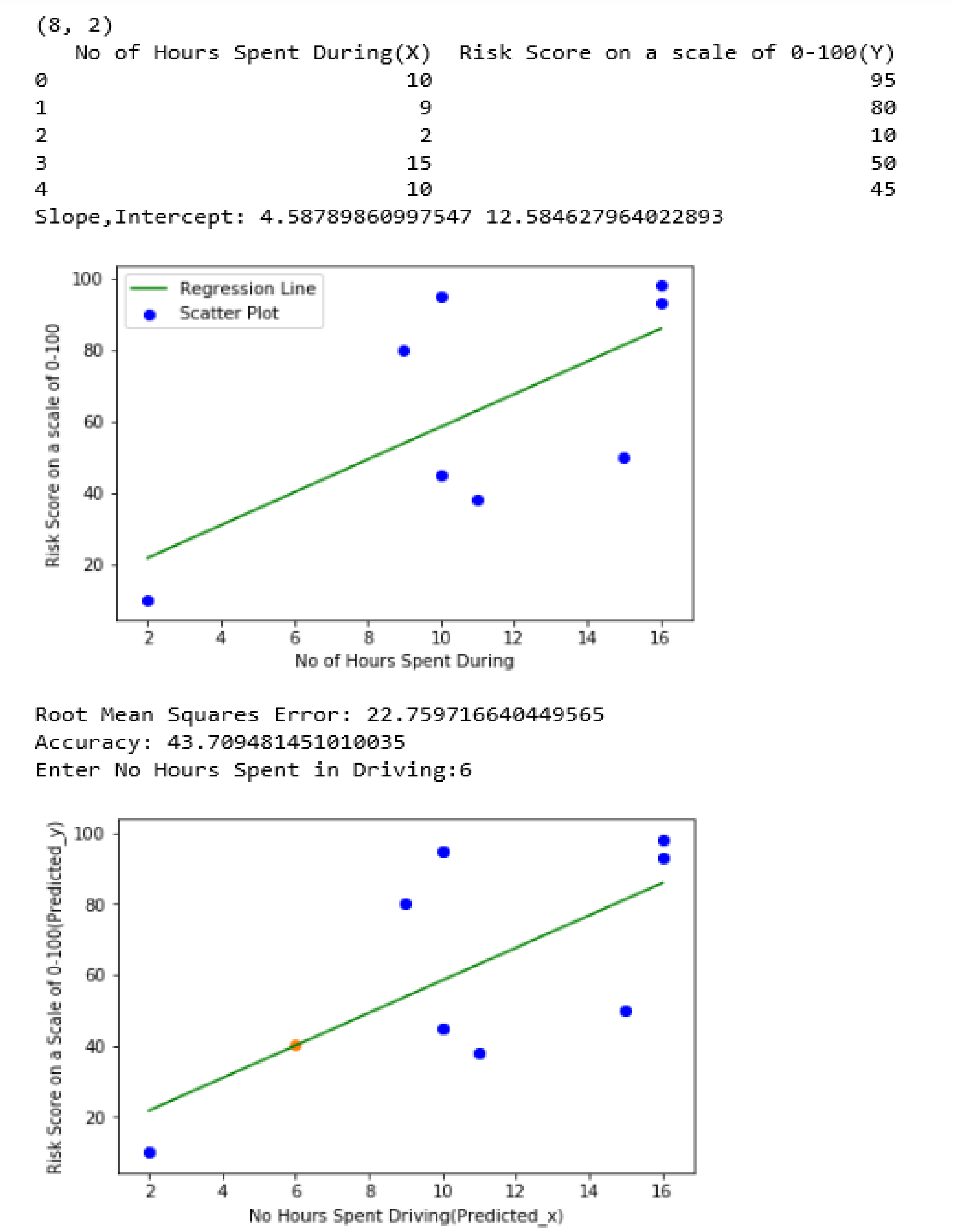
2.Read Given Dataset

3.Import the Linear Regression and Create object of it

4.Find the Accuracy of Model using Score Function

5.Predict the value using Regressor Object
6.Take input from User
7.Calculate the value of y
8.Draw Scatter PLot

Output:



```
(8, 2)
   No of Hours Spent During(X)  Risk Score on a scale of 0-100(Y)
0                           10                                 95
1                            9                                 80
2                            2                                 10
3                           15                                 50
4                           10                                 45
Slope,Intercept: 4.58789860997547 12.584627964022893
```

Regression Line
Scatter Plot
Risk Score on a scale of 0-100
100
80
60
40
20
2
4
6
8
10
12
14
16
No of Hours Spent During

```
Root Mean Squares Error: 22.759716640449565
Accuracy: 43.709481451010035
Enter No Hours Spent in Driving:6
```

Risk Score on a Scale of 0-100(Predicted_y)
100
80
60
40
20
2
4
6
8
10
12
14
16
No Hours Spent Driving(Predicted_x)

Linear Regression Applications:

1. **Trend lines:**A trend line represents the variation in some quantitative data with passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values. However, this method suffers from a lack of scientific validity in cases where other potential changes can affect the data.
2. **Economics:**Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country’s exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.
3. **Finance:**Capital price asset model uses linear regression to analyze and quantify the systematic risks of an investment.
4. **Biology**:Linear regression is used to model causal relationships between parameters in biological systems.

Conclusion:

Thus student can learn that to how to find the trend of data using X as Independent Variable and Y is and Dependent Variable by using Linear Regression.

Prepared by:	**Approved by:**
Dr.T.Bhaskar	Dr. D.B. Kshirsagar
Subject Teacher	HoD-Computer