Regression

- The following table that shows recently conducted study on the correlation of number of hours spent driving with the risk of developing acute backache. Find the equation of best fit line.

No 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

- **Theory:**

  I. Regression:
  It is basically a statistical approach to find the relationship between variables. In machine learning, this is used to predict the outcome of an event based on the relationship between variables obtained from the data-set. It is the process where an algorithm is used to predict a result based on the previously entered values and the results generated from them.

Types of Regression:

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. ElasticNet Regression

There are two key variables in every experiment: The independent variable and the dependent variable.

a. Independent variable: What the scientist changes or what changes on its own. The independent variable is the variable whose change isn't affected by any other variable in the experiment

b. Dependent variable: The dependent variable is what is being studied and measured in the experiment. It is what changes as a result of the changes to the independent variable.

II. Linear Regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable. In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

With simple linear regression we want to model our data as follows:

1. Equation of line is: y = m * x + b

2. Calculate slope of line (m) or coefficient, which is as follows: $m = \sum_{i=0}^{n} \frac{(x-X)(y-Y)}{(x-X)2}$

3. Calculate constant value i.e intercept for the given data point: b = Y – m * X

4. Final step is to calculate the value of y, by substituting the above values of m and b with respective value x.

Mathematical implementation of Linear Regression:

X	y	x - X	y - Y	(x - X)(y - Y)	$(x - X)^2$
10	95	-1.125	31.375	-35.296	1.265
9	80	-2.125	16.375	-34.796	4.515
2	10	-9.125	-53.625	489.32	83.265
15	50	3.875	-13.625	-52.796	15.015
10	45	-1.125	-18.625	20.95	1.265
16	98	4.875	34.375	167.578	23.765
11	38	-0.125	-25.625	3.203	0.015
16	93	4.875	29.375	143.203	23.765
∑ = 89	509			701.366	152.87

Where, Y = Mean of y values

X = Mean of x values.

∴ Y = 509 / 8 = 63.625 and X = 89 / 8 = 11.125

Now, calculate value of m = 701.366 / 152.87

m = 4.587

Calculating the value for b = 63.625 – 4.587 * 11.125 = 12.59

Hence, the equation of **best fit line** will be:

y = 4.587 * x + 12.59

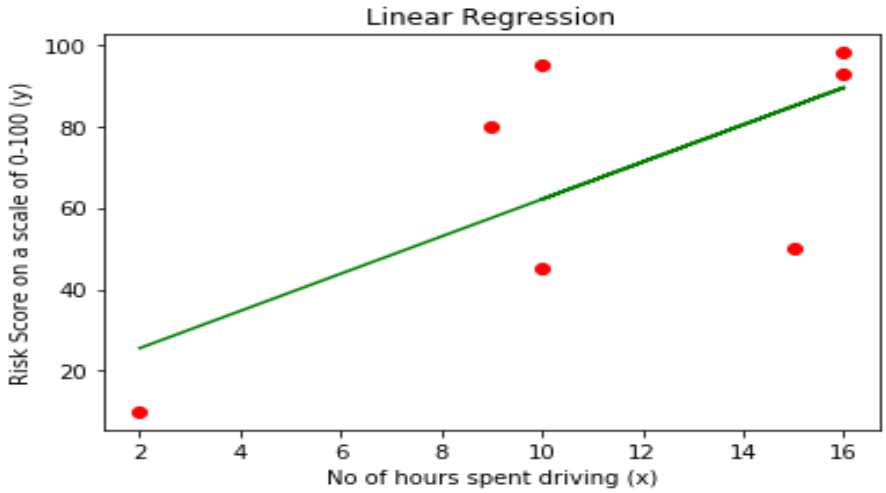


Fig: Best fit line using Linear Regression (Mean Squared method)

III. Multiple Linear Regression:

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.

When selecting the model for the multiple linear regression analysis, another important consideration is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable (typically expressed as R^2). Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model.

Multiple Linear Regression using two independent variables id as follows:

a. The equation of line for multiple linear regression using two variables is –

$$y = b1x1 + b2x2 + a$$

b. Calculating the values for slopes of x1 and x2 as follows -

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

c. Calculating the value of intercept i.e $a = Y - b1X1 - b2X2$

d. Calculate the required values to calculate the slopes respectively –

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$
$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$
$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

e. Substitute the above calculated values of slopes and intercept to find the best fit line using multiple linear regression.

IV. Polynomial Regression:

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an *nth* degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y|x).

The basic goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable x. In simple regression, we used following equation: **y = m*x + b**

Here, y is dependent variable, b is y intercept, m is the slope.

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$Y = m_1x_1 + m_2x_1^2 + m_2x_1^3 + \ldots\ldots + m_nx_1^n + c$$

Mathematical implementation of Polynomial Regression:

x	y	x^2	x^3	x^4	x*y	$x^{2*}y$
10	95	100	1000	10000	950	9500
9	80	81	729	6561	720	6480
2	10	4	8	16	20	40
15	50	225	3375	50625	750	11250
10	45	100	1000	10000	450	4500
16	98	256	4096	65536	1568	25088
11	38	121	1331	14641	418	4598
16	93	256	4096	65536	1488	23808
$\sum$ = 89	509	1143	15635	222915	6364	85264

1. Calculate the best fit curve for degree = 2.

2. Calculating the values of intercept and coefficients of the polynomial equation:

$$y = m_1*x + m_2*x^2 + c \ldots\ldots. (\because \text{degree} = 2)$$

3. Hence there are three constant values to be calculated, they are m_1, m_2 and c. Hence the representation of equations is as follows:

$$\begin{pmatrix} n & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{pmatrix} \begin{pmatrix} c \\ m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum x*y \\ \sum x^{2*}y \end{pmatrix}$$

n = Total no of data points.

4. Substituting the values and forming the required matrices and vector.

$$\begin{pmatrix} 8 & 89 & 1143 \\ 89 & 1143 & 15635 \\ 1143 & 15635 & 222915 \end{pmatrix} \begin{pmatrix} c \\ m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} 509 \\ 6364 \\ 85264 \end{pmatrix}$$

5. Similarly, forming the other matrices. Form matrix M_0 by replacing the first column of original matrix by vector, similarly for matrix M_1 replace the second column and for M_2 replace the third column. Matrices M_0, M_1, M_2 are as follows-

$$M = \begin{pmatrix} 8 & 89 & 1143 \\ 89 & 1143 & 15635 \\ 1143 & 15635 & 222915 \end{pmatrix}$$

$$M_0 = \begin{pmatrix} 509 & 89 & 1143 \\ 6364 & 1143 & 15635 \\ 85264 & 15635 & 222915 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 8 & 509 & 1143 \\ 89 & 6364 & 15635 \\ 1143 & 85264 & 222915 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 8 & 89 & 509 \\ 89 & 1143 & 6364 \\ 1143 & 15635 & 85264 \end{pmatrix}$$

6. Calculate the Determinant values for all the four matrices.

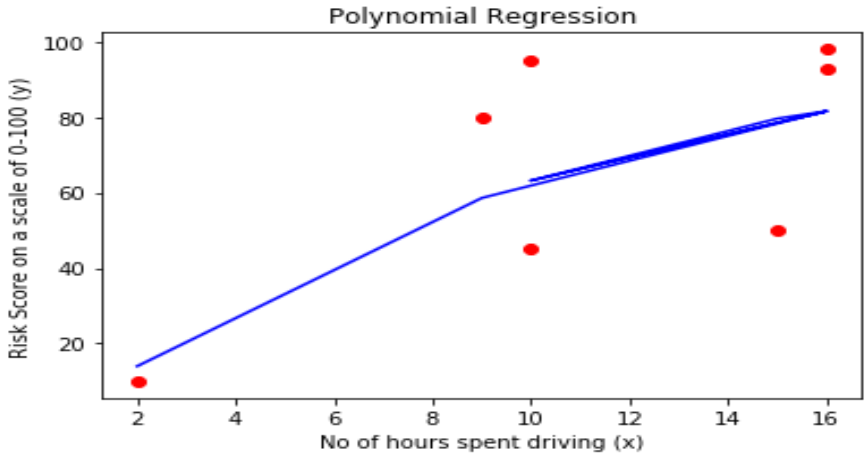
7. Now, substitute the calculated determinant values in the below formulae in order to calculate the values of m_1, m_2 and c.

$$\boldsymbol{m1} = \frac{|M1|}{|M|} = \frac{\mathbf{41604182}}{\mathbf{4731328}} = 8.7933 \qquad \boldsymbol{c} = \frac{|M0|}{|M|} = \frac{\mathbf{-12936316}}{\mathbf{4731328}} = -2.7341$$

$$\boldsymbol{m2} = \frac{|M2|}{|M|} = \frac{\mathbf{-1042026}}{\mathbf{4731328}} = -0.2202$$

8. Hence the equation of curve for degree = 2 is –

$y = 8.7933*x - 0.2202*x^2 - 2.7341$



V. R^2 (R-squared in Regression Analysis):

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

R-squared is always between 0 and 100%:

- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 100% represents a model that explains all of the variation in the response variable around its mean.

VI. Adjusted $R^{2:}$

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared. Adjusted r-squared value always be less than or equal to r-squared value.