M.Sc. Applications of Mathematics

Project : Application of Least Squares Method in Regression Analysis

Presented by : Tamal Kanti Panja, Shouvik Sardar, Suvadip Roy

■ INTRODUCTION :

• What is Regression?

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a single or a series of other changing variables (known as independent variables).

• What is Regression Analysis?

In statistics, Regression Analysis is a statistical technique for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us to understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

• What is the objective?

In Regression Analysis, our objective is to get the exact form of the approximate relationship between a dependent variable and one or more independent variables, which is widely used for prediction and forecasting purposes.

The idea behind Regression Analysis is to verify that a function $Y = f(X)$ fits a given data set $\{(x_1, y_1), (x_2, y_2), \dots\dots, (x_n, y_n)\}$ after obtaining the parameters that identify the function $f(X)$. The value X represents one or more independent variables.

The function $f(X)$ can be, for example, a linear function, i.e.

$$Y = mX + b \text{ or } Y = b_0 + b_1X_1 + b_2X_2 + \cdots\cdots + b_mX_m \ ,$$

a polynomial function, i.e.

$$Y = b_0 + b_1X + b_2X^2 + \cdots\cdots + b_pX^p \ ,$$

or other non-linear functions such as :

(i) Exponential function, i.e. $Y = ab^X$,

(ii) Logistic function, i.e.

$$Y = \frac{e^{(b_0+b_1X)}}{e^{(b_0+b_1X)}+1} = \frac{1}{e^{-(b_0+b_0X)}+1} \ ,$$

(iii) Trigonometric function, i.e. $Y = g(sin(bX)) + E$, where g(.) is a known polynomial function and E stands for error.

Note: We use the above non-linear functions in some particular phenomena such as :

(i) Exponential function is used in census and for population prediction.

(ii) Logistic function is used for predicting the outcome of a categorical data. Categorical data refers to a data which has more than one source of variation.

(iii) Trigonometric function is used in case of share market purpose.

Though there are various types of regression model, in general purposes we emphasize on linear and polynomial equations only.

- **Why linear and polynomial equations only?**

We use linear and polynomial equations because

(i) Linear and polynomial equations are linear in parameters. So these are easier to deal with than any other types of equations.

(ii) Analysis of linear and polynomial equations are less time, money and labour consuming.

(iii) In most of the cases, the regression equation can be well approximated by linear or polynomial equations with less error.

- **What is the procedure of obtaining a regression equation?**

The procedure of obtaining a regression equation consists in postulating a form of the function to be fitted, $Y = f(X)$, which will depend, in general, on a number of parameters, say $\{b_0, b_1, \ldots, b_k\}$. Then we choose a suitable method to determine the values of those parameters. The most commonly used method for this purpose is the **Least Squares Method**. In this context, we will describe how least squares method is useful in regression analysis. However, there are many other useful ways of regression analysis but here we are interested only on least squares method since this is the most useful, suitable and commonly used method for fitting a linear or polynomial equation to a given set of observations on a dependent and one or more independent variable(s).

■ METHOD OF LEAST SQUARES :

Suppose $(x_i, y_i), i = 1, 2, \ldots\ldots, n$ be n paired observations on two variables X and Y where Y is the study variable and X is the explanatory variable. Let $Y = f(a_1, a_2, \ldots, a_k; X)$ be the empirical formula to be fitted on the data where the form of the empirical formula is obtained either by graphical display or through any other method. Here, $a_1, a_2, \ldots, a_k$ are the unknown parameters to be estimated on the basis of the given data.

Here, we assume that the empirical formula to be fitted is linear in parameters $a_1, a_2, \ldots, a_k$. We also assume that only Y values are subjected to error while the X values are free from errors. Naturally, we measure the residuals (error part) along the Y-axis only.

Here, e_i = Residual for i^{th} paired observation $= y_i - f(a_1, a_2, \ldots, a_k; x_i), i = 1, 2, \ldots, n$

We define $S^2 = \sum_{i=1}^{n} e_i^2$ as the Residual Sum of Squares (RSS)

In Least Squares Method, we assume that the best fitted curve is that for which RSS is minimum. In ideal situation, best fitted curve will be provided by

$$\begin{aligned} & RSS = 0 \\ \Leftrightarrow\ & \sum_{i=1}^{n} e_i^2 = 0 \\ \Leftrightarrow\ & e_i = 0 \quad \forall \quad i = 1, 2, \ldots, n \end{aligned}$$

But such a situation will be realised very rarely in practice. For other situations, we shall try to minimize the RSS with respect to the unknown parameters $a_1, a_2, \ldots, a_k$. For this purpose, we equate the partial derivatives of S^2 with respect to $a_1, a_2, \ldots, a_k$ separately to 0 to generate as many equations as the no. of parameters. These equations are known as normal equations.

Mathematically,

$$\frac{\partial S^2}{\partial a_j} = 0 \quad \forall \quad j = 1, 2, \ldots, k \qquad \text{are k normal equations in k unknowns}$$

Finally, these normal equations are solved simultaneously to get the estimates of $a_1, a_2, \ldots, a_k$. Let the estimates be $\hat{a_1}, \hat{a_2}, \ldots, \hat{a_k}$. Then the fitted equation is

$$Y = f(\hat{a_1}, \hat{a_2}, \ldots, \hat{a_k}; X)$$

- **Fitting of equation when it is linear in parameters :**

Polynomial and multiple linear equations are linear in parameters. So, in this context we will fit polynomial and multiple linear equations on the basis of observed data on dependent and independent variables.

1. **To fit $Y = a_0+a_1X+a_2X^2+\cdots+a_pX^p$ ($\mathbf{p^{th}}$ degree polynomial) on n paired observations $(x_i, y_i), i = 1, 2, \ldots, n$.**

$$\text{Here, } S^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p)^2$$

Normal equations are :

$$\left.\begin{aligned}
\frac{\partial S^2}{\partial a_0} &= 0 \Rightarrow \sum_{i=1}^{n} (y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p) = 0\\
\frac{\partial S^2}{\partial a_1} &= 0 \Rightarrow \sum_{i=1}^{n} x_i(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p) = 0\\
\frac{\partial S^2}{\partial a_2} &= 0 \Rightarrow \sum_{i=1}^{n} x_i^2(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p) = 0\\
&\vdots\\
&\vdots\\
\frac{\partial S^2}{\partial a_p} &= 0 \Rightarrow \sum_{i=1}^{n} x_i^p(y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_px_i^p) = 0
\end{aligned}\right\} \cdots\cdots(*)$$

$(*)$: (p+1) normal equations for fitting a p^{th} degree polynomial.

Matrix Representation of Normal Equations :
The (p+1) normal equations can be expressed in matrix notation as

$$\begin{pmatrix}
n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^p\\
\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{p+1}\\
\vdots & \vdots & \ddots & \vdots\\
\sum_{i=1}^{n} x_i^p & \sum_{i=1}^{n} x_i^{p+1} & \cdots & \sum_{i=1}^{n} x_i^{2p}
\end{pmatrix}
\begin{pmatrix} a_0\\ a_1\\ a_2\\ \vdots\\ \vdots\\ \vdots\\ a_p \end{pmatrix}
=
\begin{pmatrix} \sum_{i=1}^{n} y_i\\ \sum_{i=1}^{n} x_iy_i\\ \vdots\\ \sum_{i=1}^{n} x_i^py_i \end{pmatrix}
\Leftrightarrow Ma = c, \text{ say}$$

Given n paired observations $(x_i, y_i), i = 1, 2, \ldots, n$; we can obtain the elements of M and c whereas $a = (a_0, a_1, a_2, \ldots, a_p)^T$ is unknown. So, we have a linear system of equations with (p+1) unknowns. So, the unknown a can be obtained by solving the linear system of equations $Ma = c$ and we get $\hat{a} = (\hat{a_0}, \hat{a_1}, \hat{a_2}, \ldots, \hat{a_p})^T$ as the Least Square Estimate of a and the fitted equation is

$$Y = \hat{a_0} + \hat{a_1}X + \hat{a_2}X^2 + \cdots + \hat{a_p}X^p$$

Here, if we take p = 1, we get the linear equation as $Y = \hat{a_0} + \hat{a_1}X$
For p = 2, we get the quadratic equation as $Y = \hat{a_0} + \hat{a_1}X + \hat{a_2}X^2$,
for p = 3, the cubic equation as $Y = \hat{a_0} + \hat{a_1}X + \hat{a_2}X^2 + \hat{a_3}X^3$ and so on.

Ways to solve the linear system of equations $Ma = c$:
(i) We can solve the linear system of equations $Ma = c$ by Gauss Elimination method.
(ii) Since the matrix M is a square matrix, we can apply LU factorisation method also.
(iii) Again, we can see that the matrix M is symmetric and it can be shown that it is a positive definite matrix (since the determinant value of all the principal order sub-matrices are positive). So, we can use Cholesky factorisation here to solve the linear system.
(iv) Moreover, depending on the data, the matrix M may have other properties that would lead us to solve the system by some other methods but those properties are completely uncertain.
(v) If the matrix M be non-singular, we can use the crude method to solve the system and get the estimate of a as $\hat{a} = M^{-1}c$.
(vi) If the matrix M be non-singular, i.e. M has full column rank, we can apply QR factorisation also.

2. **To fit $Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m$ (multiple linear equation) on n sets of observations $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i), i = 1, 2, \ldots, n$.**

$$\text{Here, } S^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1x_{1i} - b_2x_{2i} - \cdots - b_mx_{mi})^2$$

Normal equations are :

$$\left.\begin{aligned}
\frac{\partial S^2}{\partial b_0} = 0 &\Rightarrow \sum_{i=1}^{n} (y_i - b_0 - b_1x_{1i} - b_2x_{2i} - \cdots - b_mx_{mi}) = 0 \\
\frac{\partial S^2}{\partial b_1} = 0 &\Rightarrow \sum_{i=1}^{n} x_{1i}(y_i - b_0 - b_1x_{1i} - b_2x_{2i} - \cdots - b_mx_{mi}) = 0 \\
\frac{\partial S^2}{\partial b_2} = 0 &\Rightarrow \sum_{i=1}^{n} x_{2i}(y_i - b_0 - b_1x_{1i} - b_2x_{2i} - \cdots - b_mx_{mi}) = 0 \\
&\vdots \\
&\vdots \\
\frac{\partial S^2}{\partial b_m} = 0 &\Rightarrow \sum_{i=1}^{n} x_{mi}(y_i - b_0 - b_1x_{1i} - b_2x_{2i} - \cdots - b_mx_{mi}) = 0
\end{aligned}\right\} \cdots\cdots (**)$$

$(**)$: (m+1) normal equations for fitting a multiple linear equation on m independent variables.

Matrix Representation of Normal Equations :
The (m+1) normal equations can be expressed in matrix notation as

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{1i} & \cdots & \sum_{i=1}^{n} x_{mi} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \cdots & \sum_{i=1}^{n} x_{1i}x_{mi} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{mi} & \sum_{i=1}^{n} x_{mi}x_{1i} & \cdots & \sum_{i=1}^{n} x_{mi}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{mi}y_i \end{pmatrix} \Leftrightarrow Nb = d, \text{ say}$$

Given n sets of observations $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i), i = 1, 2, \ldots, n$; we can obtain the elements of N and d whereas $b = (b_0, b_1, b_2, \ldots, b_m)^T$ is unknown. So, we have a linear system of equations with (m+1) unknowns. So, the unknown b can be obtained by solving the linear system of equations $Nb = d$ and we get $\hat{b} = (\hat{b_0}, \hat{b_1}, \hat{b_2}, \ldots, \hat{b_m})^T$ as the Least Square Estimate of b and the fitted equation is

$$Y = \hat{b_0} + \hat{b_1}X_1 + \hat{b_2}X_2 + \cdots + \hat{b_m}X_m$$

Here, if we take p = 1, we get the simple linear equation as $Y = \hat{b_0} + \hat{b_1}X$. (Since here we have only one independent variable X_1, we can take it as X.)

Ways to solve the linear system of equations $Nb = d$:
(i) We can solve the linear system of equations $Nb = d$ by Gauss Elimination method.
(ii) Since the matrix N is a square matrix, we can apply LU factorisation method also.
(iii) Again, we can see that the matrix N is symmetric and it can be shown that it is a positive definite matrix (since the determinant value of all the principal order sub-matrices are positive). So, we can use Cholesky factorisation here to solve the linear system.
(iv) Moreover, depending on the data, the matrix N may have other properties that would lead us to solve the system by some other methods but those properties are completely uncertain.
(v) If the matrix N be non-singular, we can use the crude method to solve the system and get the estimate of a as $\hat{b} = N^{-1}d$.
(vi) If the matrix N be non-singular, i.e. N has full column rank, we can apply QR factorisation also.

- **A measure of goodness of fit :**

A measure of goodness of fit for fitting a p^{th} degree polynomial $Y = \hat{a_0} + \hat{a_1}X + \hat{a_2}X^2 + \cdots + \hat{a_p}X^p$, where $\hat{a_0}, \hat{a_1}, \hat{a_2}, \ldots, \hat{a_p}$ are least square estimates of $a_0, a_1, a_2, \ldots, a_p$ obtained by solving (p+1) normal equations, is given by

$$\begin{aligned}
\text{RSS} = \sum_{i=1}^{n} e_i^2 = & \sum_{i=1}^{n} (y_i - \hat{y_i})^2, \quad \text{where} \quad \hat{y_i} = \hat{a_0} + \hat{a_1}x_i + \hat{a_2}x_i^2 + \cdots + \hat{a_p}x_i^p \\
= & \sum_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p)(y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p) \\
= & \sum_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p)y_i - \hat{a_0}\sum_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p) \\
& -\hat{a_1}\sum_{i=1}^{n} x_i(y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p) - \hat{a_2}\sum_{i=1}^{n} x_i^2(y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p) \\
& - \cdots\cdots - \hat{a_p}\sum_{i=1}^{n} x_i^p(y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p) \\
= & \sum_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1}x_i - \cdots - \hat{a_p}x_i^p)y_i \quad \text{(by normal equations)} \\
= & \sum_{i=1}^{n} y_i^2 - \hat{a_0}\sum_{i=1}^{n} y_i - \hat{a_1}\sum_{i=1}^{n} x_iy_i - \hat{a_2}\sum_{i=1}^{n} x_i^2y_i - \cdots\cdots - \hat{a_p}\sum_{i=1}^{n} x_i^py_i
\end{aligned}$$

Note that, in addition to the calculations made earlier for solving normal equations, only $\sum_{i=1}^{n} y_i^2$ is required for the determination of RSS.

Similarly, for fitting a multilinear equation with m independent variables $Y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots + \hat{b}_m X_m$, where $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_m$ are least square estimates of $b_0, b_1, b_2, \ldots, b_m$ obtained by solving (m+1) normal equations, a measure of goodness of fit can be given by

$$\begin{aligned}
\text{RSS} = \sum_{i=1}^{n} e_i^2 = &\sum_{i=1}^{n} (y_i - \hat{y_i})^2, \quad \text{where} \quad \hat{y_i} = \hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_m x_{mi} \\
= &\sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi})(y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) \\
= &\sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) y_i - \hat{b}_0 \sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) \\
&- \hat{b}_1 \sum_{i=1}^{n} x_{1i}(y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) - \hat{b}_2 \sum_{i=1}^{n} x_{2i}(y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) \\
&- \cdots\cdots - \hat{b}_m \sum_{i=1}^{n} x_{mi}(y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) \\
= &\sum_{i=1}^{n} (y_i - \hat{b}_0 - \hat{b}_1 x_{1i} - \cdots - \hat{b}_m x_{mi}) y_i \quad \text{(by normal equations)} \\
= &\sum_{i=1}^{n} y_i^2 - \hat{b}_0 \sum_{i=1}^{n} y_i - \hat{b}_1 \sum_{i=1}^{n} x_{1i} y_i - \hat{b}_2 \sum_{i=1}^{n} x_{2i} y_i - \cdots\cdots - \hat{b}_m \sum_{i=1}^{n} x_{mi} y_i
\end{aligned}$$

Here also, we can see that, in addition to the calculations made earlier for solving normal equations, only $\sum_{i=1}^{n} y_i^2$ is required for the determination of RSS.

We see that RSS is actually the measure of sum of square errors (SSE). A small value of RSS denotes that an equation fitted to the given data set is good. In case of polynomial fitting, it can be shown that RSS is a non-increasing function of the degree of the polynomial to be fitted. So, RSS reduces with the increase of the degree of polynomial. Again, the RSS will decrease with the increase of the number of explanatory variables in case of multilinear fitting. By doing this we can minimize the error upto a certain threshold.

- **Fitting of equation when it is not linear in parameters :**

We can use least square method for fitting transformed equation using the transformed data if the empirical relation can be made linear in parameters with suitable transformations.

1. **To fit $Y = ab^X$ (exponential equation) on n paired observations $(x_i, y_i), i = 1, 2, \ldots, n$.** Here, we take logarithm to both sides of the equation and get

$$\begin{aligned}
&\log Y = \log a + X \log b \\
\Rightarrow \quad &Z = A + BX, \qquad\qquad \text{where} \quad A = \log a, \quad B = \log b \quad \text{and} \quad Z = \log Y
\end{aligned}$$

We fit $Z = A + BX$ to the transformed data $(x_i, z_i), i = 1, 2, \ldots, n$; where $z_i = \log y_i$, $i = 1, 2, \ldots, n$. Suppose, $\hat{A}$ and $\hat{B}$ are the least squares estimates of A and B. Then we get the estimates of a and b as $\hat{a} = e^{\hat{A}}$ and $\hat{b} = e^{\hat{B}}$ respectively and get the fitted equation as

$$Y = \hat{a}\hat{b}^X$$

■ APPLICATION OF LEAST SQUARES METHOD IN REGRESSION :

Our main objective is to show how the least squares method is applied in the regression analysis to get the regression equation rather than doing a whole regression analysis. In this context, we will discuss only two types of regression model as follows:

(i) Simple Linear Regression
(ii) Multiple Linear Regression

We will not discuss the polynomial regression, since the problem of polynomial regression can be extended from the problem of simple linear regression. Actually, simple linear regression is a particular form of polynomial regression. Again, the problem of exponential regression can be reduced to a simple linear regression problem. So,we will discuss these two regression models and will compare them with an example. Before starting that discussion, we first need to know what is simple linear regression and what is multiple linear regression.

1. **Simple Linear Regression :**
   Definition : Simple Linear Regression is a statistical technique that uses only one explanatory variable to predict the outcome of a response variable. The goal of simple linear regression (SLR) is to model the relationship between an explanatory and a response variable.

   For example, we can think of the height of a human body as the explanatory variable and weight as the response variable. Then we could try to predict the weight on the basis of height. Some more such examples are:

   (i) Production and sale of a large business house,
   (ii) Age and blood pressure of a human body,
   (iii) Weight of a new born baby and age of the mother, etc.

   A simple linear regression equation can be taken as $Y = a_0 + a_1X$, where Y is the dependent or response variable and X is the independent or explanatory variable. Here, we try to estimate the value of Y on the basis of X. Given a set of paired observations on the variables X and Y, we can use the least square method to get the estimates of a_0 and a_1 as $\hat{a_0}$ and $\hat{a_1}$ respectively and obtain the simple linear regression equation of Y on X as $Y = \hat{a_0} + \hat{a_1}X$. (When we take the regression equation in such a way that Y is a function of X, i.e. we take Y as the response variable and X as the explanatory variable, then we call the regression equation as Y on X. In the reverse, we call it as the regression equation of X on Y.)

2. **Multiple Linear Regression :**
   Definition : Multiple Linear Regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the relationship between the explanatory and response variables.

   It is often the case that a response variable may depend on more than one explanatory variable. For example, human weight could reasonably be expected to depend on both the height and the age of the person. Furthermore, possible explanatory variables often co-vary with one another (e.g. sea surface temperatures and sea-level pressures). This makes it impossible to subtract out the effects of the factors separately by performing successive linear regressions for each individual factor. It is necessary in such cases to perform multiple regression defined by an extended linear model.

   A multiple linear regression equation can be taken as $Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m$, where Y is the dependent or response variable and $X_1, X_2, \ldots, X_m$ are the independent or explanatory variables. Here, we try to estimate the value of Y on the basis of $X_1, X_2, \ldots, X_m$. Given a set of tuples of observations on the variables $X_1, X_2, \ldots, X_m$ and Y, we can use the least square method to get the estimates of $b_0, b_1, \ldots, b_m$ as $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$ respectively and obtain the multiple linear regression equation as $Y = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \cdots + \hat{b}_mX_m$.

- **Elaboration with an Example :**

Here, we have a dataset (Table 1) on systolic blood pressure of 11 persons with their ages (in years) and weight (in pounds). In this context, we will deal with this dataset and will try to obtain a simple and a multiple linear regression equation to obtain an estimate of the systolic blood pressure of a person with respect to age (in case of simple linear regression) and with respect to both age and weight (in case of multiple linear regression). Before that, we need to clear our concept about systolic blood pressure.

What is Systolic Blood Pressure?

Systolic blood pressure is the amount of pressure that blood exerts on vessels while the heart is beating. In a blood pressure reading (such as 120/80, this is the normal value), it is the number on the top. If the top and bottom blood pressures are both too high, a person is said to have high blood pressure. If only the top number is higher than 140, the person has a condition called isolated systolic hypertension.

There are many physical factors, such as age, weight, diet, exercise, disease, drugs or alcohol etc. that influence systolic blood pressure. Here, we consider the first two physical factors only.

i	y_i	x_{1i}	x_{2i}
1	132	52	173
2	143	59	184
3	153	67	194
4	162	73	211
5	154	64	196
6	168	74	220
7	137	54	188
8	149	61	188
9	159	65	207
10	128	46	167
11	166	72	217

Table 1 : Showing the data on systolic blood pressure (y_i) with age in years (x_{1i}) and weight in pounds (x_{2i}) corresponding to i^{th} person of total 11 persons

At first, we fit a simple linear regression equation to the data to estimate systolic blood pressure (BP) with respect to age only. For this, we take our regression equation to be fitted as $Y = a_0 + a_1 X_1$, where Y is the response variable denoting the systolic BP and X_1 is the explanatory variable denoting the age (in years).

Now, by Least Squares Method, for fitting $Y = a_0 + a_1 X_1$, a simple linear equation, we have the normal equations in matrix form as:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{1i} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i} y_i \end{pmatrix} \Leftrightarrow Ma = c, \text{ say}$$

Here, $n = 11$
Using scilab, we get

$$\sum_{i=1}^{n} x_{1i} = 687, \quad \sum_{i=1}^{n} x_{1i}^2 = 43737, \quad \sum_{i=1}^{n} y_i = 1651, \quad \sum_{i=1}^{n} x_{1i} y_i = 104328$$

$$\therefore\ M = \begin{pmatrix} 11 & 687 \\ 687 & 43737 \end{pmatrix} \quad \text{and} \quad c = \begin{pmatrix} 1651 \\ 104328 \end{pmatrix}$$

We get the estimate of a as

$$\hat{a} = \begin{pmatrix} \hat{a_0} \\ \hat{a_1} \end{pmatrix} = M^{-1}c = \begin{pmatrix} 58.705515 \\ 1.4632305 \end{pmatrix}$$

Hence, the fitted equation is $\quad Y = 58.705515 + 1.4632305\, X_1$

The estimated values of Y obtained by this equation are given below (in Table 2):

i	y_i	$\hat{y}_{i\,simple}$	$e_i = y_i - \hat{y}_{i\,simple}$
1	132	134.7935	-2.7934997
2	143	145.03611	-2.0361129
3	153	156.74196	-3.7419567
4	162	165.52134	-3.5213395
5	154	152.35227	1.6477347
6	168	166.98457	1.0154301
7	137	137.71996	-0.7199606
8	149	147.96257	1.0374261
9	159	153.8155	5.1845043
10	128	126.01412	1.9858831
11	166	164.05811	1.941891

Table 2 : Showing the observed (y_i) and expected ($\hat{y}_{i\,simple}$) value on systolic BP with the error ($e_i = y_i - \hat{y}_{i\,simple}$) due to this fitting corresponding to i^{th} person of total 11 persons

From Table 2, considering the errors, we can say that the fitted simple linear regression equation to the observed data is moderately good. We compute RSS (Residual Sum of Squares) for comparison with other models as:

$$\text{RSS}_{simple} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{i\,simple})^2 = 78.285949$$

Now, we fit a multiple linear regression equation to the data to estimate systolic BP with respect to both age and weight. For this, we take our regression equation to be fitted as $Y = b_0 + b_1X_1 + b_2X_2$, where Y is the response variable denoting the systolic BP, X_1 and X_2 are the explanatory variables denoting the age (in years) and the weight (in pounds) respectively.

Now, by Least Squares Method, for fitting $Y = b_0 + b_1X_1 + b_2X_2$, a multiple linear equation, we have the normal equations in matrix form as:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{2i}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \sum_{i=1}^{n} x_{2i}y_i \end{pmatrix} \Leftrightarrow Nb = d, \text{ say}$$

Using scilab, with the previous calculations of $\sum_{i=1}^{n} x_{1i}$, $\sum_{i=1}^{n} x_{1i}^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_{1i}y_i$, we get

$$\sum_{i=1}^{n} x_{2i} = 2146, \quad \sum_{i=1}^{n} x_{1i}x_{2i} = \sum_{i=1}^{n} x_{2i}x_{1i} = 135530, \quad \sum_{i=1}^{n} x_{2i}^2 = 421708, \quad \sum_{i=1}^{n} x_{2i}y_i = 324401$$

$$\therefore\; N = \begin{pmatrix} 11 & 687 & 2146 \\ 687 & 43737 & 135530 \\ 2146 & 135530 & 421708 \end{pmatrix} \quad \text{and} \quad d = \begin{pmatrix} 1651 \\ 104328 \\ 324401 \end{pmatrix}$$

We get the estimate of b as

$$\hat{b} = \begin{pmatrix} \hat{b_0} \\ \hat{b_1} \\ \hat{b_2} \end{pmatrix} = N^{-1}d = \begin{pmatrix} 31.60052 \\ 0.8663002 \\ 0.3300308 \end{pmatrix}$$

Hence, the fitted equation is $\quad Y = 31.60052 + 0.8663002\, X_1 + 0.3300308\, X_2$
The estimated values of Y obtained by this equation are given below (in Table 3):

i	y_i	$\hat{y}_{i_{multiple}}$	$e_i = y_i - \hat{y}_{i_{multiple}}$
1	132	133.74345	-1.7434543
2	143	143.43789	-0.4378943
3	153	153.6686	-0.6686038
4	162	164.47693	-2.4769282
5	154	151.72976	2.2702353
6	168	168.31351	-0.3135054
7	137	140.42652	-3.4265163
8	149	146.49062	2.5093821
9	159	156.2264	2.7735967
10	128	126.56547	1.4345317
11	166	165.92084	0.0791566

Table 3 : Showing the observed (y_i) and expected $(\hat{y}_{i_{multiple}})$ value on systolic BP with the error $(e_i = y_i - \hat{y}_{i_{multiple}})$ due to this fitting corresponding to i^{th} person of total 11 persons

From Table 3, considering the errors, we can say that the fitted multiple linear regression equation to the observed data is very good. We compute RSS (Residual Sum of Squares) for comparison with other models as:

$$\text{RSS}_{multiple} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{i_{multiple}})^2 = 42.860841$$

Now, comparing RSS$_{simple}$ and RSS$_{multiple}$, we can clearly say that for the observed data, multiple linear regression equation fitting is better than simple linear regression equation fitting. This implies that we can estimate the systolic BP better taking into consideration the weight alongwith the age than taking the age only.

From the above example, we can conclude that for estimating the value of response variable taking into consideration several explanatory variables is better than taking only one explanatory variable. The RSS will decrease with the increase of the number of explanatory variables but that does not mean that we can increase the number of explanatory variables as much as we want because the increasing of the number of explanatory variables is more time, money and labour consuming.

■ <u>CONCLUSION</u> :

Regression analysis is itself a huge subject of discussion and subject to various considerations also. It is important to state few points about this, such as:

✓ While we are doing some linear regression problem, we may deal with simple linear regression or multiple linear regressions. In a model, different factors have different extent of effects, so if we consider larger number of variables in the model we would be able to explain larger amount of information for the dependent variable. So it is clear that error is inversely related with number of factors taken into account of the model.

✓ Here, in this project, we have deliberately considered that a mathematical model can be fitted on some data set and then we have shown that there exist various types of mathematical model and of course those models have different extent of fitting with the original data, whereas in regression analysis, there is another notion called inferential problem where we test whether the effect of a particular factor is present in the problem or not. Then if some factors have no effect on a particular problem, we drop those effects when we fix the model. And then we go on with the important factors but there is some technical problems related to Statistics that should be taken in to account. While we test the presence of a factor's effect it may happen that our test gives the conclusion that there is no significant effect of a factor in the model but we should practically think about it and then make a conclusion. For an example, if we are to fit a regression model for predicting the cut off marks for admission in an engineering college, taking marks of the students of that college in different subjects, it may happen that someone is interested to test whether a particular subject has significant effect in the model or not. For that a test is conducted for cross validation. In practice, testing the effect of a particular subject both of English and Mathematics, the test may give the result that there is no significant effect of that subject in each case but we should accept the result for English and reject the result for Mathematics to make the model appropriate.

✓ Another problem related to regression theory is the problem of predicting a value of a dependent variable while the independent variable is a very unlikely value for the set, on which the model is fitted. In this case, though the model is well fitted with observed data so far, it may give high error for the unlikely value of the independent variable. For an example, if a regression model is fitted on heights and weights for some adults, where weight is to be predicted for some given height, and the data set for height lies between 5 feet to 5.11 feet. Then if we want to predict the weight while the height is above 7 feet, it may results in high error.

✓ Sometimes it may happen that a regression model can be fitted on some data set, but in reality there is no meaning of fitting such a model. For an example, we may construct some linear relationship between sizes of shoes with their I.Q level, and may be the model is well fitted with the original data set, but actually there is no relationship between these two.

So, for regression analysis, we should take various factor in to our account which are related to the problem, while we are to fit an appropriate model, otherwise it will be worthless, and the decision should be taken with practical ideas about the field of interest. So, it is convenient to state a quotation at the end: *"I am in full favor of keeping dangerous weapons from the fools. Let's start with STATISTICS."*

- **References :**

  - Fundamentals of Statistics (Volume I), Eighth Edition (2008) [The World Press Private Limited, Kolkata] – A. M. Gun, M. K. Gupta, B. Dasgupta.
  - Linear Algebra and Its Applications, Fourth Edition – Gilbert Strang.
  - Statistical Inference, Second Edition – George Casella, Roger L. Berger.
  - Least squares - Wikipedia, the free encyclopedia
    weblink: `en.wikipedia.org/wiki/Least_squares`
  - Linear regression - Wikipedia, the free encyclopedia
    weblink: `en.wikipedia.org/wiki/Linear_regression`
  - Regression analysis - Wikipedia, the free encyclopedia
    weblink: `en.wikipedia.org/wiki/Regression_analysis`
  - Simple linear regression - Wikipedia, the free encyclopedia
    weblink: `en.wikipedia.org/wiki/Simple_linear_regression`
  - Multiple Linear Regression
    weblink: `www.stat.yale.edu/Courses/1997-98/101/linmult.htm`
  - Multiple Linear Regression (MLR) Definition - Investopedia
    weblink: `www.investopedia.com/terms/m/mlr.asp`
  - Data source for the elaborate example:
    weblink: `http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr02.html`