CHAPTER # 07

REGRESSION & MULTIPLE REGRESSION

- **Regression**

The dependence of one variable upon the other variable is called regression. For example, weights depend upon the heights.

OR

Regression is a mathematical relationship b/w one dependent and one independent variable.
For example, Demand depends upon price. When price is independent variable and demand is dependent variable.

- **Linear Regression**

When the dependence of the variable is represented by a straight line, then it is called the linear regression otherwise it said to be non-linear or curvilinear regression. For example, If X is independent variable and Y is dependent variable, then the relation **Y=a+bX** is called linear regression.

- **Properties of Least Square Regression line or Regression line**

1. The least square regression line always passes through the mean values i.e. ($\bar{X}, \bar{Y}$).
2. Regression Coefficient i.e. b , d always have the same size.
3. The sum of deviation from observed as estimated values is always equal to zero. i.e. $\sum(Y-\hat{Y})=0$, $\sum(X-\hat{X})=0$
4. The sum of square deviation b/w observed estimated values always minimum. i.e. $\sum(Y-\hat{Y})^2$ = minimum , $\sum(X-\hat{X})^2$ = minimum
5. Sum of trend values always equal to sum of observed values. i.e. $\sum Y=\sum \hat{Y}$, $\sum X=\sum \hat{X}$

- **Types of Linear Regression / Regression Equations**

Regression equations are the algebraic expressions of the regression lines. There are two regression equations, because there are two regression lines. These are:

1. Regression Equations of Y and X
2. Regression Equation of X and Y

1. **Regression Equations of Y and X or Y on X**

b is the regression coefficient of regression line Y on X.

Liner Regression / Regression line / Least square regression line

$$Y=a+bX$$

Or

$$Y - \bar{Y} = b(X - \bar{X})$$

Or

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

General Method

Normal Equations

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

We get the value of "a" and "b" solving the above equations simultaneously.

Alternative Methods

Direct formula of obtaining the value of "a" and "b"

Direct formula of "a"

(1). $a = \bar{Y} - b\bar{X}$

(2). $a = a_{YX} = \dfrac{\left(\sum X^2\right)\left(\sum Y\right) - \left(\sum X\right)\left(\sum XY\right)}{n\sum X^2 - \left(\sum X\right)^2}$

(3). $a = a_{YX} = \dfrac{\begin{vmatrix} \sum X^2 & \sum XY \\ \sum X & \sum Y \end{vmatrix}}{\begin{vmatrix} \sum X^2 & \sum X \\ \sum X & n \end{vmatrix}}$

Direct formula of "b"

(1). $b = b_{YX} = \dfrac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2}$

(2). $b = b_{YX} = \dfrac{\begin{vmatrix} \sum XY & \sum X \\ \sum Y & n \end{vmatrix}}{\begin{vmatrix} \sum X^2 & \sum X \\ \sum X & n \end{vmatrix}}$

(3). $b = b_{YX} = \dfrac{\sum XY - \dfrac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}}$

(4). $b = b_{YX} = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$

(5). $b = b_{YX} = \dfrac{\sum XY - n\bar{X}\bar{Y}}{nS_X^2}$ When $S_X^2 = \left[\dfrac{\sum X^2}{n} - \left(\dfrac{\sum X}{n}\right)^2\right]$

(6). $b = b_{YX} = \dfrac{\sum (X-\bar{X})\sum (Y-\bar{Y})}{\sum (X-\bar{X})^2}$

(7). $b = b_{YX} = \dfrac{\sum D_X D_Y - \dfrac{\left(\sum D_X\right)\left(\sum D_Y\right)}{n}}{\sum D_X{}^2 - \dfrac{\left(\sum D_X\right)^2}{n}}$

Where $D_X = X - A$ A= Constant

$D_Y = Y - B$ B= Constant

(8). $b = b_{YX} = r\dfrac{S_Y}{S_X}$

(9). $b = b_{YX} = \dfrac{S_{XY}}{S_X^2}$

Where

$$S_{XY} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{n}$$

$$S_X = \sqrt{\frac{\sum (X-\bar{X})^2}{n}} \quad , \quad S_X = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

$$S_Y = \sqrt{\frac{\sum (Y-\bar{Y})^2}{n}} \quad , \quad S_Y = \sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}$$

2. **<u>Regression Equation of X and Y or X on Y</u>**

d is the regression coefficient of regression line X on Y.

Liner Regression / Regression line / Least square regression line

$$X = c + dY$$

Or

$$X - \bar{X} = d(Y - \bar{Y})$$

Or

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

General Method

Normal Equations

$$\sum X = nc + d\sum Y$$
$$\sum XY = c\sum Y + d\sum Y^2$$

We get the value of "c" and "d" solving the above equations simultaneously.

Alternative Methods
Direct formula of obtaining the value of "c" and "d"

Direct formula of "c"

(1). $c = \bar{X} - d\bar{Y}$

(2). $c = a_{XY} = \dfrac{\left(\sum X\right)\left(\sum Y^2\right) - \left(\sum Y\right)\left(\sum XY\right)}{n\sum Y^2 - \left(\sum Y\right)^2}$

(3). $c = a_{XY} = \dfrac{\begin{vmatrix} \sum X & \sum XY \\ \sum Y & \sum Y^2 \end{vmatrix}}{\begin{vmatrix} \sum Y^2 & \sum Y \\ \sum Y & n \end{vmatrix}}$

Direct formula of "d"

(1). $d = b_{XY} = \dfrac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(\sum Y\right)^2}$

(2). $$d = b_{XY} = \frac{\begin{vmatrix} \sum XY & \sum X \\ \sum Y & n \end{vmatrix}}{\begin{vmatrix} \sum Y^2 & \sum Y \\ \sum Y & n \end{vmatrix}}$$

(3). $$d = b_{XY} = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}$$

(4). $$d = b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

(5). $$d = b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{nS_Y^2}$$ When $S_Y^2 = \left[\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2\right]$

(6). $$d = b_{XY} = \frac{\sum (X-\bar{X})\sum (Y-\bar{Y})}{\sum (Y-\bar{Y})^2}$$

(7). $$d = b_{XY} = \frac{\sum D_X D_Y - \frac{\left(\sum D_X\right)\left(\sum D_Y\right)}{n}}{\sum {D_Y}^2 - \frac{\left(\sum D_Y\right)^2}{n}}$$

Where $D_X = X - A$ A= Constant

$D_Y = Y - B$ B= Constant

(8). $$d = b_{XY} = r\frac{S_X}{S_Y}$$

(9). $$d = b_{XY} = \frac{S_{XY}}{S_Y^2}$$

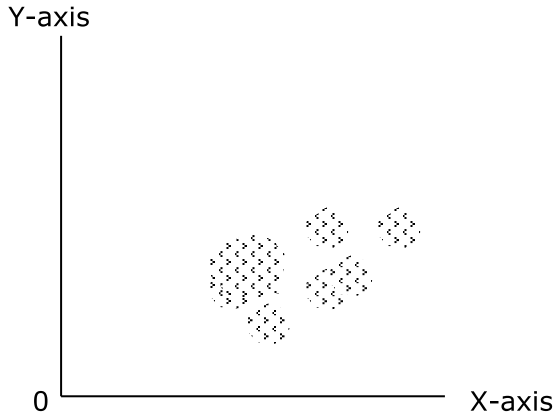
Where

$$S_{XY} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{n}$$

$$S_X = \sqrt{\frac{\sum (X-\bar{X})^2}{n}} \quad , \quad S_X = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

$$S_Y = \sqrt{\frac{\sum (Y-\bar{Y})^2}{n}} \quad , \quad S_Y = \sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}$$

- **Scatter Diagram**
  If we plot the paired observation $(X_1, Y_1)(X_2, Y_2)(X_3, Y_3)$ on a graph, the resulting set of points is called a scatter diagram.



- **Standard Deviation of Regression or Standard Error of Estimate**
  To observed values of (X, Y) do not all fall on the regression line but they scatter away from it. The degree of scatter (or dispersion) of the observed values about the regression line is measured by what is called the standard deviation of regression or the standard error of estimate of Y on X and X on Y.

1. **Y on X (** $Y = a + bX$ **)**

For Ungrouped Data

$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

Or

$$s_{y.x} = \sqrt{\frac{\sum \left(Y - \hat{Y}\right)^2}{n-2}}$$

Where $\hat{Y}$ is Trend values

For Grouped Data

$$s_{y.x} = k\sqrt{\frac{\sum fv^2 - a\sum fv - b\sum fuv}{\sum f - 2}}$$

Where k is constant

2. **X on Y (** $X = c + dY$ **)**

For Ungrouped Data

$$s_{x.y} = \sqrt{\frac{\sum X^2 - c\sum X - d\sum XY}{n-2}}$$

Or

$$s_{x.y} = \sqrt{\frac{\sum\left(X - \hat{X}\right)^2}{n-2}}$$

Where $\hat{X}$ is Trend values

For Grouped Data

$$s_{x.y} = h\sqrt{\frac{\sum fu^2 - c\sum fu - d\sum fuv}{\sum f - 2}}$$

Where h is constant

- ## Multiple Regression

  A regression which involves two or more independent variable is called a multiple regression. For example; the yield of a crop depends upon fertility of the land, fertilizer applied, rain fall, quality of seeds etc. likewise, the systolic blood pressure of a person depends upon one's weight, age, etc

  $$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} .... + \beta_k X_{ki} + \varepsilon_i$$

- ## Multiple Liner Regression With Two Independent Variables

  Multiple Regression Line

  $$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

  The estimated multiple liner regression based on sample data is

  $$Y = a + b_1 X_1 + b_2 X_2$$

  Normal Equations are

  $$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$
  $$\sum X_1 Y = a\sum X_1 + b_1 \sum X_1^{\ 2} + b_2 \sum X_1 X_2$$
  $$\sum X_2 Y = a\sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^{\ 2}$$

  We get the value of "a", "b_1" and "b_2" solving the above equations simultaneously.

- ## Type of Multiple Liner Regression / Multiple Regression Equations With Two Independent Variables

  Multiple Regression equations are the algebraic expressions of the regression lines. There are two regression equations, because there are two regression lines. These are:

  1. Multiple Regression Equations of X_1 on X_2 and X_3
  2. Multiple Regression Equation of X_2 on X_1 and X_3

3. Multiple Regression Equation of X_3 on X_1 and X_2

1. **Multiple Regression Equations of X_1 on X_2 and X_3**

$b_{12.3}$ and $b_{13.2}$ is the regression coefficient of Multiple regression line X_1 on X_2 and X_3.

Multiple Liner Regression / Multiple Regression line / Least Square Multiple Regression line

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$

Or

$$(X_1 - \bar{X}_1) = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

General Method

Normal Equations

$$\sum X_1 = na + b_{12.3}\sum X_2 + b_{13.2}\sum X_3$$
$$\sum X_1X_2 = a\sum X_2 + b_{12.3}\sum X_2^{\,2} + b_{13.2}\sum X_2X_3$$
$$\sum X_1X_3 = a\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^{\,2}$$

We get the value of "a", "$b_{12.3}$" and "$b_{13.2}$"solving the above equations simultaneously.

Alternative Methods

Direct formula of obtaining the value of "a", "$b_{12.3}$" and "$b_{13.2}$"

Direct formula of "a"

$$a = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3$$

Direct formula of "$b_{12.3}$"

(1). $b_{12.3} = -\frac{S_1}{S_2}.\frac{\Delta_{12}}{\Delta_{11}}$

(2). $b_{12.3} = -\frac{S_1}{S_2}.\frac{r_{13}r_{23} - r_{12}}{1 - r_{23}^2}$ Where r=correlation

(3). $b_{12.3} = \frac{S_1}{S_2}.\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$

Direct formula of "$b_{13.2}$"

(1). $b_{13.2} = -\frac{S_1}{S_3}.\frac{\Delta_{13}}{\Delta_{11}}$

(2). $b_{13.2} = -\frac{S_1}{S_3} . \frac{r_{12}r_{32} - r_{13}}{1 - r_{23}^2}$ Where r=correlation

(3). $b_{13.2} = \frac{S_1}{S_3} . \frac{r_{13} - r_{12}r_{32}}{1 - r_{23}^2}$

Where $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$\Delta_{11} = 1 - r_{23}^2$

Direct Method to Solve Multiple Regression equation X_1 on X_2 and X_3

$$\frac{X_1 - \bar{X}_1}{S_1}\Delta_{11} + \frac{X_2 - \bar{X}_2}{S_2}\Delta_{12} + \frac{X_3 - \bar{X}_3}{S_3}\Delta_{13} = 0$$

Or

$$X_1 = \left(\frac{S_1}{S_2}\right)\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)X_2 + \left(\frac{S_1}{S_3}\right)\left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)X_3$$

2. **Multiple Regression Equations of X_2 on X_1 and X_3**

$b_{21.3}$ and $b_{23.1}$ is the regression coefficient of Multiple regression line X_2 on X_1 and X_3.

Multiple Liner Regression / Multiple Regression line / Least Square Multiple Regression line

$$X_2 = a + b_{21.3}X_1 + b_{23.1}X_3$$

Or

$$(X_2 - \bar{X}_2) = b_{21.3}(X_1 - \bar{X}_1) + b_{23.1}(X_3 - \bar{X}_3)$$

General Method

Normal Equations

$$\sum X_2 = na + b_{21.3}\sum X_1 + b_{23.1}\sum X_3$$
$$\sum X_2X_1 = a\sum X_1 + b_{21.3}\sum X_1^2 + b_{23.1}\sum X_1X_3$$
$$\sum X_2X_3 = a\sum X_3 + b_{21.3}\sum X_1X_3 + b_{23.1}\sum X_3^2$$

We get the value of "a", "$b_{21.3}$" and "$b_{23.1}$"solving the above equations simultaneously.

Alternative Methods

Direct formula of obtaining the value of "a", "$b_{21.3}$" and "$b_{23.1}$"

Direct formula of "a"

$$a = \bar{X}_2 - b_{21.3}\bar{X}_1 - b_{23.1}\bar{X}_3$$

Direct formula of "$b_{21.3}$"

(1). $b_{21.3} = -\frac{S_2}{S_1}.\frac{\Delta_{21}}{\Delta_{22}}$

(2). $b_{21.3} = -\frac{S_2}{S_1}.\frac{r_{23}r_{13} - r_{21}}{1 - r_{13}^2}$ Where r=correlation

(3). $b_{21.3} = \frac{S_2}{S_1}.\frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2}$

Direct formula of "$b_{23.1}$"

(1). $b_{23.1} = -\frac{S_2}{S_3}.\frac{\Delta_{23}}{\Delta_{22}}$

(2). $b_{23.1} = -\frac{S_2}{S_3}.\frac{r_{21}r_{31} - r_{23}}{1 - r_{13}^2}$ Where r=correlation

(3). $b_{23.1} = \frac{S_2}{S_3}.\frac{r_{23} - r_{21}r_{31}}{1 - r_{13}^2}$

Where $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$\Delta_{11} = 1 - r_{23}^2$

Direct Method to Solve Multiple Regression equation X_2 on X_1 and X_3

$$\frac{X_2 - \bar{X}_2}{S_2}\Delta_{22} + \frac{X_1 - \bar{X}_1}{S_1}\Delta_{21} + \frac{X_3 - \bar{X}_3}{S_3}\Delta_{23} = 0$$

Or

$$X_2 = \left(\frac{S_2}{S_1}\right)\left(\frac{r_{21} - r_{23}r_{13}}{1 - r_{13}^2}\right)X_1 + \left(\frac{S_2}{S_3}\right)\left(\frac{r_{23} - r_{21}r_{31}}{1 - r_{13}^2}\right)X_3$$

3. **Multiple Regression Equations of X_3 on X_2 and X_1**

$b_{31.2}$ and $b_{32.1}$ is the regression coefficient of Multiple regression line X_3 on X_2 and X_1.

Multiple Liner Regression / Multiple Regression line / Least Square Multiple Regression line

$$X_3 = a + b_{31.2}X_1 + b_{32.1}X_2$$

Or

$$(X_3 - \bar{X}_3) = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

General Method

Normal Equations

$$\sum X_3 = na + b_{31.2}\sum X_1 + b_{32.1}\sum X_2$$
$$\sum X_3X_1 = a\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2$$
$$\sum X_3X_2 = a\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2$$

We get the value of "a", "$b_{31.2}$" and "$b_{32.1}$" solving the above equations simultaneously.

Alternative Methods

Direct formula of obtaining the value of "a", "$b_{31.2}$" and "$b_{32.1}$"

Direct formula of "a"

$$a = \bar{X}_3 - b_{31.2}\bar{X}_1 - b_{32.1}\bar{X}_2$$

Direct formula of "$b_{31.2}$"

(1). $b_{31.2} = -\frac{S_3}{S_1}.\frac{\Delta_{31}}{\Delta_{33}}$

(2). $b_{31.2} = -\frac{S_3}{S_1}.\frac{r_{32}r_{12} - r_{31}}{1 - r_{12}^2}$ Where r=correlation

(3). $b_{31.2} = \frac{S_3}{S_1}.\frac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2}$

Direct formula of "$b_{32.1}$"

$$(1).\ b_{32.1} = -\frac{S_3}{S_2}.\frac{\Delta_{32}}{\Delta_{33}}$$

$$(2).\ b_{32.1} = -\frac{S_3}{S_2}.\frac{r_{31}r_{21} - r_{32}}{1 - r_{12}^2}$$

Where r=correlation

$$(3).\ b_{32.1} = \frac{S_3}{S_2}.\frac{r_{32} - r_{31}r_{21}}{1 - r_{12}^2}$$

Where $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\Delta_{11} = 1 - r_{23}^2$$

Direct Method to Solve Multiple Regression equation X_3 on X_2 and X_1

$$\frac{X_3 - \bar{X}_3}{S_3}\Delta_{33} + \frac{X_1 - \bar{X}_1}{S_1}\Delta_{31} + \frac{X_2 - \bar{X}_2}{S_2}\Delta_{32} = 0$$

Or

$$X_3 = \left(\frac{S_3}{S_1}\right)\left(\frac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2}\right)X_1 + \left(\frac{S_3}{S_2}\right)\left(\frac{r_{32} - r_{31}r_{21}}{1 - r_{12}^2}\right)X_2$$