NOTES ON REGRESSION ANALYSIS

This notes contain the following topics

1. Meaning of Regression
2. Regression Equations
   - Regression equation of Y on X
   - Regression equation of X on Y
3. Regression coefficients
   - Regression coefficient of Y on X
   - Regression coefficient of X on Y
   - Properties of regression coefficients
4. Properties of regression lines
5. Identification of regression lines
6. Comparison between correlation and regression

REGRESSION ANALYSIS

1. Meaning of Regression

Regression analysis is a mathematical measure of the **nature of relationship** between two or more variables. It is the functional form of the nature of relation. Regression analysis is the study of the **cause and effect relationship**.

Regression analysis was introduced by **Sir Francis Galton**.

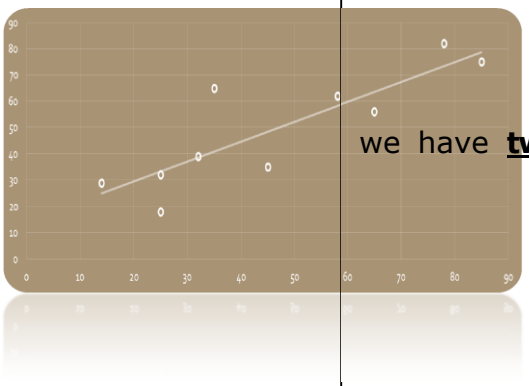
There are different types of regression. We have to study only the **linear regression** between two variables. In this type of regression only linear nature is considered. That is the two variables are assumed to be linearly dependent.

The following are some **uses** of regression analysis.

- It is used to determine the relationship between independent and dependent variables.
- The most important use is to predict and forecast.

2. Regression Equations

Here we discuss only the linear regression between two variables as mentioned earlier. Suppose we have a bivariate data. We can plot this data in scatter diagram. If the points in the scatter diagram concentrate around a straight line, that line is called the **regression line** or the **line of best fit**. The equation to that line is called the **regression equation**. This equation is a first degree equation in the variables X and Y (say). The variables X and Y are not reversible in regression analysis. So we have **two** regression lines. The regression lines are fitted by the assumption that one variable under analysis is an independent variable and the other is a dependent variable. If X is the independent variable and Y is the dependent variable we get a regression line called **the regression line of Y on X** and if Y is the independent variable and X is the dependent variable we get another regression line called the **regression line of X on Y**. This two lines are not reversible. Hence we have **two regression lines.**



Regression equation of Y on X

The regression equation of Y on X is constructed by the assumption that X is the independent variable and Y is the dependent variable.

The regression equation of Y on X is of the form,

$$y - \bar{y} = b_{YX}\left(x - \bar{x}\right)$$

Where b_{YX} is called the regression coefficient of Y on X, defined as $b_{YX} = \frac{Cov(X,Y)}{V(X)}$. (more on regression coefficients is described in next session).

Regression equation of X on Y

The regression equation of X on Y is constructed by the assumption that Y is the independent variable and X is the dependent variable.

The regression equation of X on Y is of the form,

$$x - \bar{x} = b_{XY}\left(y - \bar{y}\right)$$

Where b_{XY} is called the regression coefficient of X on Y, defined as $b_{XY} = \frac{Cov(X,Y)}{V(Y)}$.

3. The regression coefficients

As mentioned in the above, there are two types of regression coefficients.

Regression coefficient of Y on X

It is defined as $b_{YX} = \frac{Cov(X,Y)}{V(X)} = \frac{\frac{1}{n}\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\frac{1}{n}\sum\left(x - \bar{x}\right)^2}$.

The simplified formula used for calculation is:

$$b_{YX} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

Regression coefficient of X on Y

It is defined as $b_{XY} = \frac{Cov(X,Y)}{V(Y)} = \frac{\frac{1}{n}\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\frac{1}{n}\sum\left(y - \bar{y}\right)^2}$.

The simplified formula used for calculation is:

$$b_{XY} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$$

Properties of regression coefficients

- In the regression line of Y on X, b_{YX} is the regression coefficient of Y on X.
- In the regression line of X on Y, b_{XY} is the regression coefficient of X on Y.
- If the regression line of Y on X is expressed in the form $y = ax + b$, then $b_{YX} = a$.
- If the regression line of X on Y is expressed in the form $x = cy + d$, then $b_{XY} = c$.
- The sign of both regression coefficients will be the same. Either both are +ve or both are –ve.
- The product of both the coefficients is less than or equal to 1.
- If one of the regression coefficients is greater than 1, then the other must be less than 1. Both of them can be less than 1 simultaneously.
- Correlation coefficient is the geometric mean of the regression coefficients.

  ie, $r = \sqrt{b_{YX} \cdot b_{XY}}$

  Explanation

  We have $b_{YX} = \frac{Cov(X,Y)}{V(X)}$ *and* $b_{XY} = \frac{Cov(X,Y)}{V(Y)}$

  $$b_{YX} \cdot b_{XY} = \frac{Cov(X,Y)}{V(X)} \cdot \frac{Cov(X,Y)}{V(Y)} = \frac{\left(Cov(X,Y)\right)^2}{V(X) \cdot V(Y)}$$

  ie, $$b_{YX} \cdot b_{XY} = \frac{\left(Cov(X,Y)\right)^2}{\sigma_X^2 \cdot \sigma_Y^2} = \left(\frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}\right)^2 = r^2$$

  $$ie, r^2 = b_{YX} \cdot b_{XY} \Rightarrow r = \sqrt{b_{YX} \cdot b_{XY}}$$

- $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y}$

  Explanation

  We have the correlation coefficient, $r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$

  $$\text{m}\, Cov(X,Y) = r \cdot \sigma_X \cdot \sigma_Y$$

  Now, $$b_{YX} = \frac{Cov(X,Y)}{V(X)} = \frac{Cov(X,Y)}{\sigma_X^2} = \frac{r \cdot \sigma_X \cdot \sigma_Y}{\sigma_X^2} = r \cdot \frac{\sigma_Y}{\sigma_X}$$

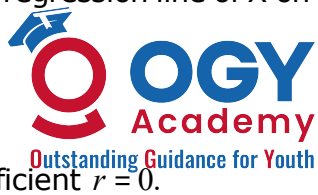
Similarly. $b_{XY} = \frac{Cov(X,Y)}{V(Y)} = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}$

- Regression coefficients are not symmetric. ie, in general, b_{YX} s b_{XY} .

4. Properties of regression lines

We have two regression lines, the regression line Y on X and the regression line of X on Y.

The following are some properties of regression lines.



- The two lines intersect at the point $(\bar{x}, \bar{y})$.
- The regression lines will be perpendicular if the correlation coefficient $r = 0$.
- The regression lines will coincide if the correlation coefficient is perfect ($r = \pm 1$).

5. Identification of regression lines

Suppose we are given two regression lines. Sometimes we have to identify them. That is we have to identify which is the regression line of Y on X and which the regression line of X on Y.

At the first step we have to assume one of them as the regression line of Y on X and the other as the regression line of X on Y.

Express the regression line of Y on X in the form $y = ax + b$. Then $b_{YX} = a$

Express the regression line of X on Y in the form $x = cy + d$. Then $b_{XY} = c$

Find $b_{YX} \cdot b_{XY}$. If it is less than or equal to 1, our assumption is right.

6. Comparison between correlation and regression

Regression	**Correlation**
Regression deals with the study of the nature of relationship between the variables	Correlation deals with the study of the degree of relationship between the variables.
Regression is not symmetric	Correlation is symmetric
Regression is the cause and effect relationship between the variables	Correlation is the association between the variables.
Regression is used for prediction	Correlation is not used for prediction
Regression is not reversible	Correlation is reversible