REGRESSION LINE (FITTING OF CURVE)

Concept of Regression:

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as ‘Regression Analysis’.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which are extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Regression analysis was explained by M. M. Blair as follows:

“Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.”

Some of the examples of dependent and independent variables

(i) Hours spent studying Vs Marks scored by students

(ii) Amount of rainfall Vs Agricultural yield

(iii) Electricity usage Vs Electricity bill

(iv) Suicide rates Vs Number of stressful people

(v) Years of experience Vs Salary

(vi) Demand Vs Product price

(vii) Age Vs Beauty

(viii) Age Vs Health issues

(ix) Number of Degrees Vs Salary

(x) Number of Degrees Vs Education expenditure

Review/summary of objectives of regression:

[1] To determine whether a relationship exists between two variables

[2] To describe the nature of the relationship, should one exist, in the form of a mathematical equation

[3] To assess the degree of accuracy of description or prediction achieved by the regression equation, and

Assumptions of Linear Regression:

[1] Relationship is approximately linear (approximates a straight line in scatter plot of Y, X)

[2] For each value of X there is a probability distribution of independent values of Y, and from each of these Y distributions one or more values are sampled at random.

[3] The means of the Y distributions fall on the regression line.

Lines of Regression and Equation:

Simple regression:

It is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known.

The regression line (known as the least squares line):

It is a plot of the expected value of the dependent variable for all values of the independent variable. Technically, it is the line that "minimizes the squared residuals". The regression line is the one that **best fits the data** on a scatterplot.

Using the **regression equation**, the dependent variable may be predicted from the independent variable. The slope of the regression line (b) is defined as the rise divided by the run. The y intercept (a) is the point on the y axis where the regression line would intercept the y axis.

The slope and y intercept are incorporated into the regression equation. The intercept is usually called the constant, and the slope is referred to as the coefficient. Since the regression model is usually not a perfect predictor, there is also an error term in the equation.

Here is a way to mathematically describe a linear regression model:

$y = a + bx + e$

If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable.

If the slope is zero. It has no prediction ability because for every value of the independent variable, the prediction for the dependent variable would be the same. Knowing the value of the independent variable would not improve our ability to predict the dependent variable. Thus, if the slope is not significantly different than zero, don't use the model to make predictions.

The standard error of the estimate for regression measures the amount of variability in the points around the regression line. It is the standard deviation of the data points as they are distributed around the regression line. The standard error of the estimate can be used to develop confidence intervals around a prediction.

A line minimizes the sum of squares of differences value given by straight line, is chosen. This principle is called as least square principle. The equation so obtained is called as least square regression line.

Regression as Prediction Model:-

Suppose we have a sample of size ‘n’ and it has two sets of measures, denoted by x and y. We can predict the values of ‘y’ given the values of ‘x’ by using the equation, called the Regression Equation.

$Y = a + bX$

where,

Y is the dependent variable, measured in units of the dependent variable.

X is the independent variable, measured in units of the independent variable. ‘a’ is the Y-intercept is the value of Y when X = 0. ‘b’ is the slope of the line and is known as the **regression coefficient** and is the change in Y associated with a one-unit change in X.

The greater the slope or regression coefficient, the more influence the independent variable has on the dependent variable, and the more change in Y associated with a change in X.

The regression coefficient is typically more important than the intercept from a policy researcher perspective as we are usually interested in the effect of one variable on another.

Coming back to the equation, we also have a term to capture the error in our estimating equation, denoted by ε or *e*. e_i is the difference between observed and estimated value and is the error or residue. It reflects the unexplained variation in Y, and its magnitude reflects the goodness of fit of the regression line. The smaller the error, the closer the points are to our line. So our general equation describing a line is:

$Y = a + bX + e$

Note: While deriving the regression model following things are important.

[i] Paired values of X, Y

[ii] Regression equation

[iii] Apply Principal of least squares to the regression equation.

[iv] Take the partial derivatives w.r.t. a and b.

[v] We get normal equations.

[vi] Determine constant a from first normal equation.

[vii] Determine constant b from second normal equation.

[viii] Substitute values of the constants a and b in regression equation as mentioned in point number [ii].

Derivation of Linear Regression Model of Y on X:

Suppose (x_i, y_i); i= 1,2,…,n., are n pairs of observations on variables X, Y. We assume that Y as dependent variable, which can be expressed in terms of X. The simplest form is the linear relation. Suppose $Y = bX + a$ However when we observe the numerical values of x and y, the relation may not be observed perfectly.

We assume the model $Y = bX + a + e$ ... (1)

We assume E(e) = 0 and Var (e) = o.

The equation (1) contains three unknown quantities and our main aim is to estimate these quantities by using least square principle, whereas e is a random variable, we estimate its parameters E (e) and Var (e).

We estimate a and b so that the error is minimum

e = Y-a-bx

By using principal of least square Symbolically we write $S = \sum_{i=1}^{n} e_i^2$ as sum of squares of errors. We find the points minima using calculus methods. The solution of equation $\frac{\partial s}{\partial a} = 0$ and $\frac{\partial s}{\partial b} = 0$ gives extreme points.

$$S=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i\text{-a-b}x_i)^2$$

$$\frac{\partial s}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n} (y_i\text{-a-b}x_i)^2$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial a}(y_i\text{-a-b}x_i)^2$$

$$0 = -2\sum \left(Y_i - \text{a-b}x_i\right)$$

$$\sum \left(y_i - a - bx_i\right) = 0$$

$$\sum y_i - na - b \sum x_i = 0$$

$$\sum y_i = na + b \sum x_i \qquad \text{………..} \ (2)$$

Similarly,

$$\frac{\partial s}{\partial b} = \frac{\partial}{\partial b} \sum \left(\mathrm{y_i} - \mathrm{a} - \mathrm{bx_i} \right)^2 = 0 \text{ gives}$$

$$2 \sum \left(\mathrm{y_i} - \mathrm{a} - \mathrm{bx_i} \right) \left(-\mathrm{x_i} \right) = 0$$

$$\sum \mathrm{x_i y_i} - \mathrm{a} \sum \mathrm{x_i} - \mathrm{b} \sum \mathrm{x_i}^2 = 0$$

$$\sum \mathrm{x_i y_i} = \mathrm{a} \sum \mathrm{x_i} + \mathrm{b} \sum x_i^2 \qquad \ldots\ldots (3)$$

The equation (2) & (3) are referred to as normal equations.

Solving equations (2) & (3) simultaneously, we get a & b .

$$\sum \mathrm{y_i} = \mathrm{na} + \mathrm{b} \sum x_i$$

$$\mathrm{na} = \sum \mathrm{y_i} - \mathrm{b} \sum x_i$$

$$\mathrm{a} = \frac{\sum \mathrm{y_i}}{\mathrm{n}} - \mathrm{b} \frac{\sum x_i}{n}$$

$$\mathrm{a} = \bar{\mathrm{y}} - \mathrm{b}\bar{\mathrm{x}} \qquad \ldots (4)$$

Substituting,

$a = \bar{y} - b\bar{x}$ in eq$^{\mathrm{n}}$ (3) we get ,

$$\sum \mathrm{x_i y_i} = \left(\bar{y} - \mathrm{b}\bar{x} \right) \sum \mathrm{x_i} + \mathrm{b} \sum \mathrm{x_i}^2$$

$$\sum x_i y_i = n\bar{x}\,\bar{y} - nb(\bar{x})^2 + b\sum x_i^2 \quad \because \left(\bar{x} = \frac{\sum x}{n} \Rightarrow \sum x_i = n\bar{x} \right)$$

$$\sum x_i y_i - n\bar{x}\,\bar{y} = b\left(\sum x_i^2 - n\left(\bar{x}\right)^2 \right)$$

Dividing by n we get

$$\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} = b\left[\frac{\sum x_i^2}{n} - \left(\bar{x}\right)^2 \right]$$

But, $\mathrm{Cov(x,y)} = \dfrac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$ and $\mathrm{Var}\left(x\right) = \sigma_{\mathrm{x}}^2 = \dfrac{\sum x_i^2}{n} - \left(\bar{x}\right)^2$

$$\mathrm{Cov(x,y)} = b \ \mathrm{Var}\left(x\right)$$

$$b=\frac{Cov(x,y)}{\sigma_x^2}$$

$$\Rightarrow b_{yx}=\frac{Cov(x,y)}{\sigma_x^2} \qquad \cdots\cdots(5)$$

Substituting equation (4) and (5) in the regression equation, $y=a+bx$, we get

$$y=\bar{y}-b_{yx}\bar{x}+b_{yx}x$$

$$(y-\bar{y})=b_{yx}(x-\bar{x})$$

$$(y-\bar{y})=\frac{Cov(x,y)}{\sigma_x^2}(x-\bar{x})$$

Therefore, $(y-\bar{y})=b_{yx}(x-\bar{x})$ represents a least square regression equation of Y on X

Derivation of Linear Regression Model of X on Y:

Suppose (x_i, y_i); i= 1,2,…,n., are n pairs of observations on variables X, Y. We assume that X as dependent variable, which can be expressed in terms of Y. The simplest form is the linear relation. Suppose $X = bY + a$; However when we observe the numerical values of x and y, the relation may not be observed perfectly.

We assume the model $X = bY + a + e$ ... (1)

We assume E(e) = 0 and Var (e) = o.

The equation (1) contains three unknown quantities and our main aim is to estimate these quantities by using least square principle, whereas e is a random variable, we estimate its parameters E (e) and Var (e).

We estimate a and b so that the error is minimum

$e = x-by-a$

By using principal of least square

Symbolically we write $S=\sum_{i=1}^{n} e_i^2$ as sum of squares of errors. We find the points minima using calculus methods. The solution of equation $\frac{\partial s}{\partial a}=0$ and $\frac{\partial s}{\partial b}=0$, gives extreme points.

$$S=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}(x_i\text{-a-by}_i)^2$$
$$\frac{\partial s}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}(x_i\text{-a-by}_i)^2$$
$$\frac{\partial s}{\partial a} =\sum_{i=1}^{n}\frac{\partial}{\partial a}(x_i\text{-a-by}_i)^2$$
$$0 = -2\sum x_i\text{-a-by}_i$$
$$\sum (x_i\text{-a-by}_i) = 0$$
$$\sum x_i - na - b \sum y_i = 0$$
$$\sum x_i = na + b \sum y_i \qquad \ldots\ldots\ldots.. (2)$$
$$\sum x_i - na - b \sum y_i = 0$$
$$\sum x_i = na + b \sum y_i \qquad \ldots\ldots\ldots.. (3)$$

Similarly, $\frac{\partial s}{\partial b} = \frac{\partial}{\partial b}\sum(x_i - a - by_i)^2 = 0$ gives

$$2 \sum \left(x_i - a - by_i\right) \left(-y_i\right) = 0$$
$$\sum x_i y_i - a \sum y_i - b\sum y_i^2 = 0$$
$$\sum x_i y_i = a \sum y_i + b \sum y_i^2 \qquad \ldots\ldots (4)$$

Equations (2) & (3) are referred to as normal equations.

Solving equations (2) & (3) simultaneously, we get a & b .

$$\sum x_i = na + b \sum y_i$$
$$na = \sum x_i - b \sum y_i$$
$$a = \frac{\sum x_i}{n} - b \frac{\sum y_i}{n}$$
$$a = \overline{x} - b\overline{y} \qquad \ldots(4)$$

Substituting, $a = \bar{x} - b\bar{y}$ in equation (3), we get

$$\sum x_i y_i = \left(\bar{x} - b\bar{y}\right)\sum y_i + b \sum y_i^2$$

$$\sum x_i y_i = \left(\bar{x} - b\bar{y}\right)n\bar{y} + b \sum y_i^2 \qquad \because \left(\bar{y} = \frac{\sum y}{n} \Rightarrow \sum y_i = n\bar{y}\right)$$

$$\sum x_i y_i = n\bar{y}\,\bar{x} - nb\left(\bar{y}\right)^2 + b \sum y_i^2$$

$$\sum x_i y_i - n\bar{y}\,\bar{x} = b \sum y_i^2 - nb\left(\bar{y}\right)^2$$

$$\sum x_i y_i - n\bar{y}\,\bar{x} = b \left[\sum y_i^2 - n\left(\bar{y}\right)^2\right]$$

Dividing by n we get

$$\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} = b\left[\frac{\sum y_i^2}{n} - \left(\bar{y}\right)^2\right]$$

But, Cov(x,y) = $\dfrac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$ and $\mathrm{Var}\left(y\right) = \sigma_y^2 = \dfrac{\sum y_i^2}{n} - \left(\bar{y}\right)^2$

$$\mathrm{Cov(x,y)} = \mathrm{b}\mathrm{Var}\left(y\right)$$

$$\mathrm{b} = \frac{\mathrm{Cov(x,y)}}{\sigma_y^2}$$

$$\Rightarrow \mathrm{b}_{xy} = \frac{\mathrm{Cov(x,y)}}{\sigma_y^2} \qquad \cdots\cdots(5)$$

Substituting equation (4) and (5) in the regression equation, y=a + bx , we get

$$\mathrm{x} = \bar{x} - \mathrm{b}_{\mathrm{xy}}\bar{y} + \mathrm{b}_{\mathrm{xy}}y$$

$$\left(\mathrm{x} - \bar{x}\right) = \mathrm{b}_{\mathrm{xy}}\left(y - \bar{y}\right)$$

$$\left(\mathrm{x} - \bar{x}\right) = \frac{\mathrm{Cov(x,y)}}{\sigma_y^2}\left(y - \bar{y}\right)$$

Therefore, $\left(\mathrm{x} - \bar{x}\right) = \mathrm{b}_{\mathrm{xy}}\left(y - \bar{y}\right)$ represents a least square regression equation of X on Y

Slope of the line:-

Slope is the ratio of rise over the run. It is given by

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}}$$

$$\text{Slope} = \frac{y_2 - y_1}{x_2 - x_1}$$

Rise means how much does it go up or down and Run means how much does it go side to side.

Sign of the slope is depend on rise and run i.e.

If the line is upward then slope is positive.

If the line is downward then slope is negative.

If the line is parallel to X-axis then slope of the line is zero.

$$\text{Slope} = \frac{0}{\text{run}} = 0$$

If the line is parallel to Y-axis then slope of the line is undefined.

$$\text{Slope} = \frac{\text{Rise}}{0} = \infty$$

Interpretation of Regression Coefficient:-

Definition: The **Regression Coefficient** is the constant 'b' in the regression equation that tells about the change in the value of dependent variable corresponding to the unit change in the independent variable. If there are two regression equations, then there will be two regression coefficients:

Regression Coefficient of X on Y: The regression coefficient of X on Y is represented by the symbol b_{xy} that measures the change in X for the unit change in Y. Symbolically, it can be represented as:

$$b_{xy} = \frac{\mathrm{Cov}(x,y)}{\sigma_y^2} \qquad \text{but,} \quad r = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} \Rightarrow b_{xy} = \frac{r\sigma_x\sigma_y}{\sigma_y^2} \qquad \Rightarrow b_{xy} = \frac{r\sigma_x}{\sigma_y}$$

Regression Coefficient of Y on X: The symbol b_{yx} is used that measures the change in Y corresponding to the unit change in X. Symbolically, it can be represented as:

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} \qquad \text{but,} \quad r = \frac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow b_{yx} = \frac{r\sigma_x \sigma_y}{\sigma_x^2} \qquad \Rightarrow b_{yx} = \frac{r\sigma_y}{\sigma_x}$$

The Regression Coefficient is also called as a **slope coefficient** because it determines the slope of the line i.e. the change in the dependent variable for the unit change in the independent variable.

Interpretation of Regression coefficients:

Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values. Suppose you have the following regression equation: y = 3X + 5. In this equation, +3 is the coefficient, X is the predictor, and +5 is the constant.

The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable.

A positive sign indicates that as the predictor variable increases, the response variable also increases.

A negative sign indicates that as the predictor variable increases, the response variable decreases.

The coefficient value represents the mean change in the response given a one unit change in the predictor. For example, if a coefficient is +3, the mean response value increases by 3 for every one unit change in the predictor.

Properties of Regression Coefficients:

[1] Correlation coefficient and regression coefficients have same algebraic sign

Proof: $b_{yx} = \frac{Cov(x,y)}{\sigma_x^2}$; $b_{xy} = \frac{Cov(x,y)}{\sigma_y^2}$ and $r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

Clearly, numerator of each coefficient is same and denominator of each coefficient is positive. Hence, numerator decides algebraic sign. Thus all coefficients have same algebraic sign. Hence, If $r > 0$, then $b_{yx} > 0$ and $b_{xy} > 0$.

If $r = 0$, then $b_{yx} = 0 = b_{xy}$.

If $r < 0$, then $b_{yx} < 0$ and $b_{xy} < 0$.

[2] Correlation coefficient is a square root of product of regression coefficients. $\left(\text{i.e. } r = \sqrt{b_{yx} \times b_{xy}}\right)$ **or correlation coefficient is geometric mean of regression coefficients.**

Proof:

$$b_{yx} \times b_{xy} = \frac{Cov(x,y)}{\sigma_x^2} \times \frac{Cov(x,y)}{\sigma_y^2}$$

$$b_{yx} \times b_{xy} = \left(\frac{Cov(x,y)}{\sigma_x \times \sigma_y}\right)^2$$

$$b_{yx} \times b_{xy} = (r)^2$$

$$\therefore r = \sqrt{b_{yx} \times b_{xy}}$$

Note: Choose positive square root if regression coefficients are positive, otherwise, negative.

[3] Both regression coefficients cannot exceed unity simultaneously.

Proof: If possible, let us assume $b_{yx} > 1$ and $b_{xy} > 1$.

Hence, $b_{yx} \times b_{xy} > 1$

$\therefore r^2 > 1$

Hence, which is impossible $\because r < 1$. Thus our assumption is incorrect.

[4] Regression coefficient can be expressed in terms of correlation coefficient.

i.e. $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$

Proof:

We have

$$b_{yx}=\frac{Cov(x,y)}{\sigma_x^2} \qquad \text{but} \qquad r=\frac{Cov(x,y)}{\sigma_x\,\sigma_y} \Rightarrow Cov(x,y)=r\sigma_x\,\sigma_y$$

$$\therefore b_{yx}=\frac{r\sigma_x\,\sigma_y}{\sigma_x^2} \qquad \Rightarrow b_{yx}=\frac{r\sigma_y}{\sigma_x}$$

$$b_{xy}=\frac{Cov(x,y)}{\sigma_y^2} \qquad \text{but} \qquad r=\frac{Cov(x,y)}{\sigma_x\,\sigma_y} \Rightarrow Cov(x,y)=r\sigma_x\,\sigma_y$$

$$\therefore b_{xy}=\frac{r\sigma_x\,\sigma_y}{\sigma_y^2} \qquad \Rightarrow b_{xy}=\frac{r\sigma_x}{\sigma_y}$$

[6] Regression coefficients are invariant to the change of origin.

Note that Cov (x ,y), σ_x and σ_y are invariant to the change of origin, hence the regression coefficients are invariant to the change of origin. This property makes the computations of regression coefficients simple. We can subtract a constant from each observation for computations

[7] Regression coefficients are invariant to the change of origin but not of scale.

Proof:

Let

$$u=\frac{x\text{-}a}{h} \quad \text{and} \quad v=\frac{y\text{-}b}{k}$$

$$Cov\left(\frac{x\text{-}a}{h},\frac{y\text{-}b}{k}\right)=Cov(u,v)=\frac{1}{hk}Cov(x,y)$$

$$\sigma^2_{\left(\frac{x\text{-}a}{h}\right)}=\sigma_u^2=\frac{1}{h^2}\sigma_x^2 \quad \text{and} \quad \sigma^2_{\left(\frac{y\text{-}b}{k}\right)}=\sigma_v^2=\frac{1}{k^2}\sigma_y^2$$

$$b_{uv}=\frac{Cov(u,v)}{\sigma_v^2}=\frac{Cov(x,y)/hk}{\sigma_y^2/k^2}=\frac{k}{h}\frac{Cov(x,y)}{\sigma_y^2}$$

$$b_{vu}=\frac{Cov(u,v)}{\sigma_u^2}=\frac{Cov(x,y)/hk}{\sigma_x^2/h^2}=\frac{h}{k}\frac{Cov(x,y)}{\sigma_x^2}$$

[8] If $r=\pm 1$, then regression coefficients are reciprocals of each other.

Proof: We have,

$$b_{yx} \times b_{xy} = r^2$$

$$b_{yx} \times b_{xy} = 1$$

$$b_{xy} = \frac{1}{b_{yx}} \quad \text{or} \quad b_{yx} = \frac{1}{b_{xy}}$$

[9] If $\sigma_x = \sigma_y$ then prove that regression coefficients are equal.

Proof:

We have

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = \frac{Cov(x,y)}{\sigma_y^2} = \frac{r\sigma_x}{\sigma_y}$$

but $\sigma_x = \sigma_y \quad \therefore b_{yx} = r = b_{xy}$

[10] Product of regression coefficients is less than unity.

Proof:

$b_{yx} \times b_{xy} = r^2 \quad$ but $r^2 < 1$

Hence, $b_{yx} \times b_{xy} < 1$

[11] The acute angle (θ) between the regression lines is

Note :

(i) We see from the above expression that larger the r^2, smaller is the angle between the lines.

(ii) The point of intersection of two regression lines is $(\bar{x}, \bar{y})$

(iii) When, $r = \pm 1$, then $\tan\theta = 0$, therefore, $\theta = 0$.

When the angle $\theta = 0$, there are two possibilities. First, the lines will be **coincident** and the second, the lines will be **parallel.** However, the regression lines **intersect** at $(\bar{x}, \bar{y})$. Hence the second possibility is ruled out. Therefore, for r = ± 1, the regression lines are coincident. In other words, if there is perfect correlation, then the regression lines coincide.

iv. If r = 0, then $\tan\theta = \infty$, therefore, $\theta = \frac{\pi}{2}$. Hence, the lines are perpendicular to each other. The points on scatter diagram will show maximum spread. In other words, if the variables are uncorrelated, then the regression lines are perpendicular to each other.

Different type of Variation in Regression Y on X:-

Residual: The difference between the observed value of Y and its predicted or estimated value of Y is called residual or error in prediction. It is denoted by e and is given by

$\therefore e = (y_i - \hat{y}_i)$

It is measured in two ways: residual plot and residual sum of squares.

Residual plot:

It is obtained by plotting residuals $(y_i - \hat{y}_i)$ on y-axis against the x_i values.

Total Variation (Total sum of squares):-

Variation of the regression of dependent variable y is caused by the variation in x is called total variation. It is denoted by SST or $\sum(y_i - \bar{y})^2$ and is given by

$$\text{Average Total Sum of squares (SST)} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

SST=SSR + SSE

$$(y_i - \bar{y}) = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

Unexplained Variation (Residual sum of squares):-

Variation not explained by the regression of y on x is called unexplained variation. It is denoted by SSE or $\sum(y_i - \hat{y}_i)^2$ and is given by

$$\text{Average Sum of squares due to error (SSE)} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

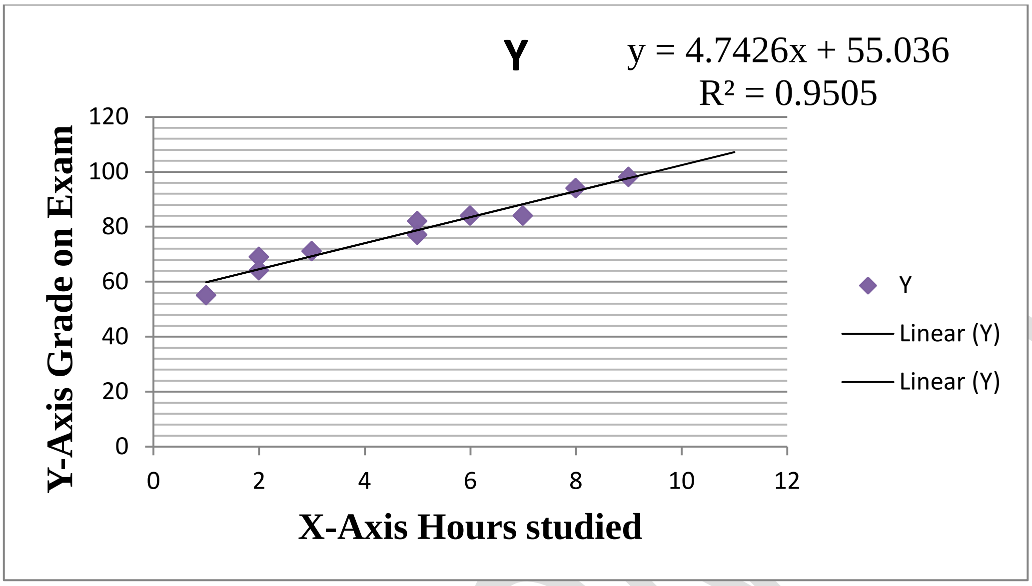
Explained Variation (sum of squares due to regression):-

Variation explained by the regression of y on x is called explained variation. It is denoted by SSR or $\sum(\hat{y}_i - \bar{y})^2$ and is given by

$$\text{Average Sum of squares due to regression (SSR)} = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$$

Geometrical Interpretation:-

X	2	9	5	5	3	7	1	8	6	2
Y	69	98	82	77	71	84	55	94	84	64



Coefficient of determination:-

Definition: The **Coefficient of determination** is the square of the coefficient of correlation r^2 which is calculated to interpret the value of the correlation. It is useful because it explains the level of variance in the dependent variable caused or explained by its relationship with the independent variable.

The coefficient of determination explains the proportion of the explained variation or the relative reduction in variance corresponding to the regression equation rather than about the mean of the dependent variable. For example, if the value of $r = 0.8$, then r^2 will be 0.64, which means that 64% of the variation in the dependent variable is explained by the independent variable while 36% remains unexplained.

Thus, the coefficient of determination is the ratio of explained variance to the total variance that tells about the strength of linear association between the variables, say X and Y. The value of $\mathbf{r^2}$ lies between **0 and 1** and observes the following relationship with **'r'**. With the decrease in the value of 'r' from its maximum value of 1, the 'r^2' also decreases much more rapidly.The value of 'r' will always be greater than 'r^2' unless the r^2 =0 or 1.The coefficient of determination also explains that how well the regression line fits the statistical data. The closer the regression line to the points plotted on a scatter diagram, the more likely it explains all the variation and the farther the line from the points the lesser is the ability to explain the variance.

Understanding the P Value

The P value is another statistic displayed on a spectrum of 0 to 1 that you'll see after a regression analysis. Unlike R-squared, the P value tells you **how likely it is that there is no correlation whatsoever**. A high P value tells you that it's likely there is zero correlation, whereas a low P value indicates that the two variables are correlated.

If the outcome of the dependent variable truly does depend on the independent variable, the P value will be low. If you're way off base and comparing apples to oranges, the P value will be high.

Regression Formulae

[1] $\bar{x} = \frac{1}{n}\sum x_i$ [2] $\bar{y} = \frac{1}{n}\sum y_i$

[3] $\sigma_x^2 = \frac{1}{n}\sum x_i^{\,2} - (\bar{x})^2$ [4] $\sigma_y^2 = \frac{1}{n}\sum y_i^{\,2} - (\bar{y})^2$

[5] $\mathrm{Cov}(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}$ [6] $\mathrm{Corr}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$

[7] Equation of a line of X on Y is

X= a + bY

$\therefore$ regression line of equation of X on Y is

$(x-\bar{x}) = b_{xy}(y-\bar{y}) \quad \Rightarrow (x-\bar{x}) = \frac{r\sigma_x}{\sigma_y}(y-\bar{y}) \; (\because a = \bar{x} - b\bar{y})$

[8] Equation of a line of X on Y is

y = a + bx

Regression line of equation of Y on X is

$$(y-\overline{y})=b_{yx}(x-\overline{x}) \qquad \Rightarrow (y-\overline{y})=\frac{r\sigma_y}{\sigma_x}(x-\overline{x})\left(\because a=\overline{y}-b\overline{x}\right)$$

[9]Regression coefficient of X on Y

$$b_{xy}=\frac{\text{cov}(x,y)}{\sigma_y{}^2} \qquad \Rightarrow b_{xy}=\frac{r\sigma_x}{\sigma_y}$$

[10]Regression coefficient of Y on X

$$b_{yx}=\frac{\text{cov}(x,y)}{\sigma_x{}^2} \qquad \Rightarrow b_{yx}=\frac{r\sigma_y}{\sigma_x}$$

[11]Sum of squares due to regression (SSR)=$\sum\left(\hat{y}_i-\overline{y}\right)^2$

[12]Sum of squares due to error (SSE)=$\sum\left(y_i-\hat{y}_i\right)^2$

[13]Total Sum of squares (SST)=$\sum\left(y_i-\overline{y}\right)^2$

[14] SST=SSR + SSE

[15]Mean Sum of squares due to error (MSSE)=$\dfrac{\sum\left(y_i-\hat{y}_i\right)^2}{n-2}$

[16] Coefficient of determination = $r^2=\dfrac{SSR}{SST}=\dfrac{SST-SSE}{SST}=1-\dfrac{SSE}{SST}$

Coefficient of determination =$r^2=\dfrac{\sum\left(\hat{y}_i-\overline{y}\right)^2}{\sum\left(y_i-\overline{y}\right)^2}=\dfrac{\text{Explained variation}}{\text{Total variation}}$

[17] Adjusted $r^2=1-\dfrac{MSSE}{MSST}=\dfrac{SSE/n-2}{SST/n-1}$

Numerical Examples:

[1] A panel of examiners A and B based seven candidates independently and awarded the following marks,

Candidate	1	2	3	4	5	6	7
Marks A	40	34	28	30	44	38	31
Marks B	32	39	26	30	38	34	28

Eight candidates was awarded 36 marks by examiner A using regression
line estimate the marks awarded by the examiner B

Candidate	Marks by A (X)	Marks by B (Y)	X^2	Y^2	XY
1	40	32	1600	1024	1280
2	34	39	1156	1521	1326
3	28	26	784	676	728
4	30	30	900	900	900
5	44	38	1936	1089	1452
6	38	34	1444	1156	1292
7	31	28	961	784	868
Total	245	227	8781	7150	7846

$\bar{x} = \frac{1}{n} \times \sum xi \qquad \Rightarrow \bar{x} = \frac{1}{7} \times 245 \qquad \Rightarrow \bar{x} = 35$

$\bar{y} = \frac{1}{n} \times \sum yi \qquad \Rightarrow \bar{y} = \frac{1}{7} \times 227 \qquad \Rightarrow \bar{y} = 32.42$

$\sigma_x^2 = \sum x_i^2 - n(\bar{x})^2 \qquad \Rightarrow \sigma_x^2 = 8781 - 7(35)^2$

$\Rightarrow \sigma_x^2 = 8781 - 8775 \qquad \Rightarrow \sigma_x^2 = 6$

$\sigma_y^2 = \sum y_i^2 - n(\bar{y})^2 \qquad \Rightarrow \sigma_y^2 = 7150 - 7(32.42)^2$

$\sigma_y^2 = 7150 - 7357.39 \qquad \Rightarrow \sigma_y^2 = -207.39$

$cov(x,y) = \sum x_i y_i - n(\bar{x})(\bar{y})$

$cov(x,y) = 7846 - 7(35)(32.42)$

$cov(x,y) = 7846 - 7942.9 \qquad \Rightarrow cov(x,y) = -96.9$

$b_{xy} = \frac{cov(x,y)}{\sigma_y^2}$ $\Rightarrow b_{xy} = \frac{-96.9}{7357.39}$

b_{xy}=-0.4672 $\Rightarrow b_{yx} = \frac{cov(x,y)}{\sigma_x^2}$

$b_{yx} = \frac{-96.9}{6}$ $\Rightarrow b_{yx}$=-16.15

$\therefore X-\bar{x} = b_{xy}(Y-\bar{y})$

$\therefore$ X-35= -0.4672(Y-32.42)

$\therefore$ X-35 = -0.4672Y+50.14

$\therefore$ X= -16.81+50.14

$\therefore$ X= 33.33

$\therefore$ X=33

The marks given by examiner B is 33

[2] The following data related to age of husband & wife in years at the time of marriage

Age of husband	23	24	25	26	27
Age of wife	19	19	20	21	22

Estimate the age of husband if age of wife is 20 year

Solution:-

Age of husband(x)	Age of wife (y)	X^2	Y^2	XY
23	19	529	361	437
24	19	576	361	456
25	20	625	400	500
26	21	676	441	546
27	22	729	484	594
		3135	2047	2533

$\bar{x}=\frac{1}{n}\sum x_i \quad \Rightarrow \bar{x}=\frac{1}{5}\times 125 \quad \Rightarrow \bar{x}=25$

$\bar{y}=\frac{1}{n}\sum y_i \quad \Rightarrow \bar{y}=\frac{1}{5}\times 101 \quad \Rightarrow \bar{y}=2.2$

$\sigma_x^2=\sum x_i^2\text{-}n(\bar{x})^2 \quad \Rightarrow \sigma_x^2= 3135\text{-}20(25)^2$

$\sigma_x^2=\text{-}9365$

$\sigma_y^2=\sum y_i^2\text{-}n(\bar{y})^2 \quad \Rightarrow \sigma_y^2= 2047\text{-}20(2.2)$

$\sigma_y^2=1950.2$

$\text{Cov(x,y)} =\sum x_i y_i\text{-}n(\bar{x})(\bar{y}) \quad \Rightarrow \text{Cov(x,y)}=2533\text{-}20(25)(2.2)$

$\text{Cov(x,y)}=1433$

$b_{xy}=\frac{\text{Cov(x,y)}}{\sigma_x^2} \quad \Rightarrow b_{xy}=\frac{1433}{\text{-}9362}$

$b_{xy}=\text{-}0.153$

X be the age of husband and y be the age of wife

$\therefore \; x\text{-}\bar{x} = b_{xy}(y\text{-}\bar{y})$

$\therefore x-25=-0.153(y-2.2)$
$\therefore x-25=-0.153y+0.3366$
$\therefore x=-0.153y+25.33$
$\therefore x=3.06+25.33$
$\therefore x=28.39$
$\therefore x=28$

The age of husband is 28 years when the age of wife is 20 years

[3] Given the following information:

Mean height $(\bar{x})$ =120.5cm, mean age $(\bar{y})$=10.37year ,S.D of x =12.7cm,

S.D of y =2.39 year correlation coefficient between x and y = 0.93

(i) Fit the regression line (ii) Estimate the height of boy of 12 years

Solution:

Given,

$\bar{x} = 120.5\text{cm} \quad \Rightarrow \bar{y} = 10.37\text{year}$

$\sigma_x = 12.7\text{cm} \quad \Rightarrow \sigma_y = 2.39\text{year}$

r = 0.93

$\sigma_x^2 = (12.7)^2 \quad \Rightarrow \therefore \sigma_x^2 = 161.29$

$\sigma_y^2 = (2.39)^2 \quad \Rightarrow \therefore \sigma_y^2 = 5.712$

$\therefore$ Now, we have to find regression line x on y

$$\therefore x\text{-}\bar{x} = r\frac{\sigma_x}{\sigma_y}(y\text{-}\bar{y})$$

$$\therefore x - 120.5 = 0.93\left(\frac{12.7}{2.39}\right)(y - 10.37)$$

$$\therefore x - 120.5 = 4.9418(y - 10.37)$$

$$\therefore x - 120.5 = 4.9418\,y - 51.246$$

$$\therefore x = 4.9418\,y + 69.254 \qquad \text{......(1)}$$

Now regression line y on x

$$\therefore y\text{-}\bar{y} = r.\frac{\sigma_y}{\sigma_x}(x\text{-}\bar{x})$$

$$\therefore y - 10.37 = 0.93\left(\frac{2.39}{12.7}\right)(x - 120.5)$$

$$\therefore y - 10.37 = 0.1750\,x - 21.089$$

$$\therefore y = 0.1750\,x - 10.719 \qquad \text{....(2)}$$

(ii) Now, we have to find x i.e. height of boy if y i.e age is 12 years

From equation (1)

X=4.89418(12)+69.254

X=128.55

[4] Following is the information about the bivariate frequency distribute

$\sum x = 80, \sum y = 40 \;\; \sum x^2 = 1680, \; \sum y^2 = 320, \; \sum xy = 480, \; n = 20,$

(i) Obtain the regression line (ii) Estimate y for x=3 & estimate x of y=3

Solution:

$\sum x = 80, \sum y = 40 \;\; \sum x^2 = 1680, \; \sum y^2 = 320, \; \sum xy = 480, \; n = 20,$

$\bar{x} = \frac{1}{n} \times \sum x_i \qquad \Rightarrow \bar{x} = \frac{1}{20} \times 80$

$\bar{x} = 4$

$\bar{y} = \frac{1}{n} \times \sum y_i \qquad \Rightarrow \bar{y} = \frac{1}{20} \times 40$

$\bar{y} = 2$

$\sigma_x^2 = \frac{1}{n} \times \sum x_i^2 - (\bar{x})^2 \qquad \Rightarrow \sigma_x^2 = \frac{1}{20} \times (1680) - (4)^2$

$\sigma_x^2 = 68 \qquad \Rightarrow \sigma_x^2 = 8.246$

$\sigma_y^2 = \frac{1}{n} \times \sum y_i^2 - (\bar{y})^2 \qquad \Rightarrow \sigma_y^2 = \frac{1}{20} \times 320 - 4$

$\sigma_y^2 = 12 \qquad \Rightarrow \sigma_y^2 = 3.464$

$\text{Cov(x,y)} = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y}) \qquad \Rightarrow \text{Cov}(x, y) = \frac{1}{20}(480) - (4)(2)$

$\text{Cov(x,y)} = 24 - 8 \qquad \Rightarrow \text{Cov(x,y)} = 16$

Regression line y on x is given by

$b_{yx} = \frac{\text{Cov(x,y)}}{\sigma_x^2} \qquad \Rightarrow b_{yx} = \frac{16}{68} = .02352$

$b_{xy} = \frac{\text{Cov(x,y)}}{\sigma_y^2} \qquad \Rightarrow b_{xy} = \frac{16}{12} = 1.33$

$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad \Rightarrow (y-2) = 0.2352(x-4)$

$(y-2) = 0.2352x - 0.9408 \qquad \Rightarrow y = .02$

Regression line of x on y is

$x - \bar{x} = b_{xy}(y - \bar{y}) \qquad \Rightarrow x-4 = b_{xy}(y-2)$

$x-4 = 1.333y - 2.666 \qquad \Rightarrow x = 1.333y + 1.334$

(ii)

Y = 0.2352X+1.0592

Y = 0.2352(3) +1.0592

Y = 0.7056+1.0592

Y = 1.7647

X = 1.333Y+1.334

X = 1.333(3) +1.134

X = 5.3333

[5] Following is the information about the bivariate frequency distribution Result of capital employed and profit earn by a firm in ten successive year of calculated.

	Mean	S.D.
Capital employed (000'Rs.)	55	28.7
Profit earned (000'Rs.)	13	85

Coefficient of correlation = 0.96. Estimate the amount of capital to be employed to even profit of Rs.20000

Solution:- Given, r = 0.96

Consider, capital employed =X (000'Rs.) and Profit employed =Y(000'Rs.)

$\therefore \quad \bar{x} = 55$, $\sigma_x = 28.7$, $\bar{y} = 13$, $\sigma_y = 85$, n =10

Regression line of x on y is

$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \Rightarrow x - 55 = 0.96\left(\frac{28.7}{85}\right)(y-13)$

$x - 55 = 0.3241(y-13) \qquad \Rightarrow x - 55 = 0.3241y - 4.2138$

$x = 0.3241y + 50.786$

Given that the amount of capital to be employed to even profit of RS 20,000

Y=20,000

X=0.3241(20,000) + 50.786

X= 6532.78

[6] Determine the two regression lines from the following data

X	7	6	10	14	13
Y	22	18	20	26	24

Solution:-

Xi	Yi	Xi^2	Yi^2	XiYi
7	22	49	484	154
6	18	36	324	108
10	20	100	400	200
14	26	196	676	364
13	24	169	576	312
50	110	550	2460	1138

n = no. of pairs of observation is 5

$$\bar{x} = \frac{1}{n}\times\sum x_i \qquad \Rightarrow \bar{x} = \frac{1}{5}\times 50$$

$$\bar{x} = 10$$

$$\bar{y} = \frac{1}{n}\times\sum y_i \qquad \Rightarrow \bar{y} = \frac{1}{5}\times 110$$

$$\bar{y} = 22$$

$$\sigma_x^2 = \frac{1}{n}\times\sum x_i^2 - (\bar{x})^2 \qquad \Rightarrow \sigma_x^2 = \frac{1}{5}\times(550) - (10)^2$$

$$\sigma_x^2 = 110 - 100 \qquad \Rightarrow \sigma_x^2 = 10$$

$$\sigma_y^2 = \frac{1}{n}\times\sum y_i^2 - (\bar{y})^2 \qquad \Rightarrow \sigma_y^2 = \frac{1}{5}\times(2460) - (22)^2$$

$$\sigma_y^2 = 492 - 484 \qquad \Rightarrow \sigma_y^2 = 8$$

$$\mathrm{Cov}(x,y) = \frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y}) \qquad \Rightarrow \mathrm{Cov}(x,y) = \frac{1}{5}(1138) - (10)(22)$$

$$\mathrm{Cov}(x,y) = 227.6 - 220 \qquad \Rightarrow \mathrm{Cov}(x,y) = 7.6$$

Regression coefficent X on Y is given by

$$b_{xy} = \frac{Cov(x,y)}{\sigma_y^2} \qquad \Rightarrow b_{xy} = \frac{7.6}{8} = 0.95$$

Regression coefficent Y on X is given by

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} \qquad \Rightarrow b_{yx} = \frac{7.6}{10} = 0.76$$

Regression line of Y on X is

$$y-\bar{y} = b_{yx}(x-\bar{x}) \qquad \Rightarrow y-22 = 0.76(x-10)$$

$$y = 0.76x-7.6+22 \qquad \Rightarrow y = 0.76x+14.4$$

Regression line x on y is

$$x-\bar{x} = b_{xy}(y-\bar{y}) \qquad \Rightarrow x-10 = 0.95(y-22)$$

$$x = 0.95y-20.9+10 \qquad \Rightarrow x = 0.95y-10.9$$

[7] Following data are related to marks in Mathematics (X) and Marks in Statistics (Y) of 10 candidates.

$$\overline{U} = \frac{1}{n}\sum U_i = \frac{1}{10}(10) = 1 \qquad \& \qquad \overline{V} = \frac{1}{n}\sum V_i = \frac{1}{10}(-2) = -0.2$$

$$\bar{x} = a + \overline{U} \qquad \Rightarrow \bar{x} = 66 + 1 \qquad \Rightarrow \bar{x} = 67$$

$$\bar{y} = b + \overline{V} \qquad \Rightarrow \bar{y} = 68 + (-0.2) \qquad \Rightarrow \bar{y} = 67.8$$

X	Y	U=X-66	V=Y-68	U^2	V^2	U*V
66	68	0	0	0	0	0
65	67	-1	-1	1	1	1
68	67	2	-1	4	1	-2
68	70	2	2	4	4	-4
67	65	1	-3	1	9	-3
66	68	0	0	0	0	0
70	70	4	2	16	4	8
64	66	-2	-2	4	4	4
69	68	3	0	9	0	0
67	69	1	1	1	1	1
		10	-2	40	24	13

$$\mathrm{Cov}(U,V)=\frac{1}{n}\sum U_iV_i-\overline{U}\times\overline{V} \qquad \Rightarrow \mathrm{Cov}(U,V)=\frac{1}{10}(13)-(1)(-0.2)$$

$$\mathrm{Cov}(U,V)=1.3+0.2 \qquad \Rightarrow \mathrm{Cov}(U,V)=1.5$$

$$\mathrm{Cov}(U,V)=\mathrm{Cov}(X,Y) \qquad (\because \text{Covariance is independent of change of origin})$$

$$\therefore \mathrm{Cov}(X,Y)=1.5$$

$$\sigma_U^2=\frac{1}{n}\sum U_i^{\,2}-(\overline{U})^2 \qquad \Rightarrow \sigma_U^2=\frac{1}{10}(40)-(1)^2$$

$$\sigma_U^2=3 \qquad \Rightarrow \sigma_U^2=\sigma_X^2 \qquad (\because \text{variance is independent of change of origin})$$

$$\therefore \sigma_X^2=3$$

$$\sigma_V^2=\frac{1}{n}\sum V_i^2-(\overline{V})^2 \qquad \Rightarrow \sigma_V^2=\frac{1}{10}(24)-(0.2)^2$$

$$\sigma_V^2=2.36 \qquad \Rightarrow \sigma_V^2=\sigma_Y^2 \qquad (\because \text{variance is independent of change of origin})$$

$$\therefore \sigma_Y^2=2.36$$

i] Regression coefficient X on Y is,

$$b_{xy}=\frac{\mathrm{cov}(X,Y)}{\sigma_Y^2} \qquad \Rightarrow b_{xy}=\frac{1.5}{2.36}=0.63$$

Rrgression coefficient Y on X is,

$$b_{YX}=\frac{\mathrm{cov}(X.Y)}{\sigma_X^2} \qquad \Rightarrow b_{YX}=\frac{1.5}{3}=0.5$$

$$r^2=b_{XY}b_{YX} \qquad \Rightarrow r^2=(0.63)(0.5)$$

$$r^2=0.315 \qquad \Rightarrow r=0.56$$

ii]Given that X= Mathamatics Y=Statistics

X=76 then Y is...

Regression line Y on X is

$$Y-\overline{Y}=b_{yx}(X-\overline{X}) \qquad \Rightarrow Y-61.8=0.5(X-67)$$

$$Y=0.5\times X-33.5+61.8 \qquad \Rightarrow X=76$$

$$Y=0.5\times(76)+34.3 \qquad \Rightarrow Y=72.30$$

then marks in statiscs is72.30

iii]marks obtained in Statistics

Y=60

Regression line of X on Y is

$X-\bar{x} = b_{xy}(Y-\bar{y}) \quad \Rightarrow X-67 = 0.63(y-67.8)$

$X-67 = 0.63Y-42.714 \quad \Rightarrow X = 0.63y + 24028$

Y= 60 then X is

$X=0.63(60)+24028 \quad \Rightarrow X=62.08$

Marks in Mathematics is 62

[8] Following are data of retail food price index(x) &whole sale food price index (y) for 10 years. Find the regression lines hence find correlation coefficient.

X	89	86	74	65	65	63	66	67	72	79
Y	92	91.5	84	75	73.5	72	70.5	75	77.5	84

Solution:- Given that,

x = retail food price index and Y= whole sale food price index.

X	Y	U=X-65	V=Y-73.5	U^2	V^2	$U\times V$
89	92	24	18.5	576	342.25	444
86	91.5	21	18	441	324	378
74	84	9	10.5	81	110.25	94.5
65	75	0	1.5	0	0	0
65	73.5	0	0	0	2.25	0
63	72	-2	-1.5	4	9	3
66	70.5	1	-3	1	2.25	-3
67	75	2	1.5	4	16	3
72	77.5	7	4	49	2.25	28
79	84	14	10.5	196	110.25	147
		76	60	1352	918.50	1094.5

$\overline{U}=\frac{1}{n}\sum U_i \qquad \Rightarrow \overline{U}=\frac{1}{10}\times 76 \qquad \Rightarrow \overline{U}=7.6$

$\overline{V}=\frac{1}{n}\sum V_i \qquad \Rightarrow \overline{V}=\frac{1}{10}\times 60 \qquad \Rightarrow \overline{V}=6$

$\overline{X} = a + \overline{U} \qquad \Rightarrow \overline{X}= 65+7.6 \qquad \Rightarrow \overline{X}=72.6$

$\overline{Y} = b + \overline{V} \qquad \Rightarrow \overline{Y}=73.5+6 \qquad \Rightarrow \overline{Y}=79.5$

$\text{Cov}(U,V)=\frac{1}{n}\sum U_iV_i-\overline{U}\times\overline{V} \qquad \Rightarrow \text{Cov}(U,V)=\frac{1}{10}(1094.5)-(7.6)(6)$

$\text{Cov}(U,V)=109.45+45.6 \qquad \Rightarrow \text{Cov}(U,V)=63.85$

$\text{Cov}(U,V)=\text{Cov}(X,Y) \qquad (\because \text{Covariance is independent of change of origin})$

$\therefore \text{Cov}(X,Y)=63.85$

$\sigma_U^2=\frac{1}{n}\sum U_i{}^2-(\overline{U})^2 \qquad \Rightarrow \sigma_U^2=\frac{1}{10}(1352)-(7.6)^2$

$\sigma_U^2=77.44$

$\sigma_V^2=\frac{1}{n}\sum V_i^2-(\overline{V})^2 \qquad \Rightarrow \sigma_V^2=\frac{1}{10}(918.50)-(6)^2$

$\sigma_V^2=55.85$

Variance is independent of change of origin

$\sigma_U^2=\sigma_X^2 \;\; \& \;\; \sigma_V^2=\sigma_Y^2 \; ; \; \sigma_X^2=77.44 \;\; \& \;\; \sigma_Y^2=55085$

Regression coefficient of x on y & y on x

$$b_{XY}=\frac{\text{Cov}(x,y)}{\sigma_Y^2}=\frac{63.85}{55.85}=1.1432$$

$$b_{YX}=\frac{\text{Cov}(x,y)}{\sigma_X^2}=\frac{63.85}{77.44}=0.8242$$

Regression line of x on y

$x-\bar{x} = b_{XY}(y-\bar{y}) \qquad \Rightarrow x-72.6=1.1432(y-79.6)$

$x=1.1432y-90.88+7206 \qquad \Rightarrow x=1.1432y-18.28$

Regression line of y on x is

$y-\bar{y} = b_{YX}(x-\bar{x})$ $\Rightarrow x-79.6=0.8242(x-72.6)$

$y = 0.8242x-59.859+79.6$ $\Rightarrow y = 0.8242x+19.641$

Correlation coefficient is

$r^2 = b_{XY} \times b_{YX}$ $\Rightarrow r^2 = (1.1432)(0.8242)$

$r^2 = 0.9425$

[9] Following are the results of B.com examination in a certain for the last 10 years

Year	No of candidate appeared	No of successful candidate
1981	120	100
1982	150	137
1983	200	164
1984	350	302
1985	371	356
1986	385	379
1987	400	375
1988	386	381
1989	362	331
1990	350	350

Using regression line estimate no of successful candidate for the year 1996 if 400 candidate appears for examination

Solution:-Let X =no of candidate of appeared and Y= no of successful candidate

X	Y	U= X-200	V=Y-164	U^2	V^2	UV
120	100	-80	-64	6400	4096	5120
150	137	-50	-27	2500	729	1350
200	164	0	0	0	0	0
350	302	150	138	22500	19044	20700
371	356	171	192	29241	36864	32832
385	379	185	215	34225	46245	39775
400	375	200	211	40000	44521	42200
386	381	186	217	34596	47089	40362
365	331	162	167	26244	27889	27054

350	350	150	186	22500	34596	27900
		1024	866	218206	261053	237293

$\overline{U}=\frac{1}{n}\sum U_i \qquad \Rightarrow \overline{U}=\frac{1}{10}\times 1076 \qquad \Rightarrow \overline{U}=107.4$

$\overline{V}=\frac{1}{n}\sum V_i \qquad \Rightarrow \overline{V}=\frac{1}{10}\times 866 \qquad \Rightarrow \overline{V}=86.6$

$\overline{X} = a + \overline{U} \qquad \Rightarrow \overline{X}= 200+107.4 \quad \Rightarrow \overline{X}=307.4$

$\overline{Y} = b + \overline{V} \qquad \Rightarrow \overline{Y}=164+86.6 \qquad \Rightarrow \overline{Y}=250.6$

$\mathrm{Cov}(U,V)=\frac{1}{n}\sum U_iV_i-\overline{U}\times\overline{V} \qquad \Rightarrow \mathrm{Cov}(U,V)=\frac{1}{10}(237293)-(107.4)(86.6)$

$\mathrm{Cov}(U,V)=23729.3-9300.84 \qquad \Rightarrow \mathrm{Cov}(U,V)=14428.46$

$\mathrm{Cov}(U,V)=\mathrm{Cov}(X,Y) \qquad (\because \text{Covariance is independent of change of origin})$

$\therefore \mathrm{Cov}(X,Y)=14428.46$

$\sigma_U^2=\frac{1}{n}\sum U_i^{\,2}-(\overline{U})^2 \qquad \Rightarrow \sigma_U^2=\frac{1}{10}(218206)-(107.4)^2$

$\sigma_U^2=\frac{1}{10}(21820.6)-(11534.76) \qquad \Rightarrow \sigma_U^2=10285.84$

$\Rightarrow \sigma_U^2=\sigma_X^2 \qquad (\because \text{variance is independent of change of origin})$

$\therefore \sigma_X^2=10285.84$

Regression coefficient of Y on X is

$b_{YX}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} \qquad \Rightarrow b_{YX}=\frac{14428.46}{10285.84}=1.4027498$

Regression line on y and x is

$Y-\bar{y}=b_{YX}(X-\bar{x}) \qquad \Rightarrow Y-250.6=1.4027498(X-307.4)$

$Y-250.6=1.4027498X-431.2052 \qquad \Rightarrow Y=1.4027498-180.6052 \qquad \ldots.(1)$

Estimate no. of successful candidate for the year 1996 if 400 candidates appear examination i.e. X = 400 using equation (1)

Y = 1.4027498(400)-180.6052

Y = 380.51

Y = 381

[10] No. of successful candidates for year 1996 is 381 the following data gives the sales and expense of 10 firs

Firm no	1	2	3	4	5	6	7	8	9	10
Sales (in 000)	45	70	65	30	90	40	50	75	85	60
expenses	35	90	70	40	95	40	60	80	80	50

Obtain the least square regression line of expenses on sales estimate expenses if n sales x75000 also draw residual plot find the residual sum of square .

X	Y	X^2	Y^2	XY	Regression estimate of $\hat{y}$	Residual $y-\hat{y}$	$(y-\hat{y})^2$
45	35	2025	1225	1575	47.79	-12.79	163.584
70	90	4900	8100	6300	73.110	16.890	285.272
65	70	4225	4900	4550	68.046	1.954	3.8181
30	40	900	1600	1200	32.598	7.402	54.7896
90	95	8100	9025	8550	93.366	1.634	2.6699
40	40	1600	1600	1600	42.726	-2.726	7.43107
50	60	2500	3600	3000	52.854	7.146	51.0653
75	80	5625	6400	6000	78.174	1.826	3.3342
85	80	7225	6400	6800	88.302	-8.302	68.9232
60	50	3600	2500	3000	62.982	-12.982	168.532
610	640	40700	45350	42575	639.948	0.0520	809.4196

First of all we fit regression line of $(Y-\bar{y})=b_{YX}(X-\bar{x})$

$$\bar{x}=\frac{1}{n}\sum x_i \qquad \Rightarrow \bar{x}=\frac{1}{10}\times 610 = 61$$

$$\bar{y}=\frac{1}{n}\sum y_i \qquad \Rightarrow \bar{y}=\frac{1}{10}\times 640 = 64$$

$$\sigma_x^2 = \frac{1}{n}\times\sum x_i^2 - (\bar{x})^2 \qquad \Rightarrow \sigma_x^2 = \frac{1}{10}\times(40700)-(61)^2$$

$$\sigma_x^2 = 4070-3721 \qquad \Rightarrow \sigma_x^2 = 349 \qquad \Rightarrow \sigma_x = 18.68$$

$$\sigma_y^2 = \frac{1}{n}\times\sum y_i^2 - (\bar{y})^2 \qquad \Rightarrow \sigma_y^2 = \frac{1}{10}\times(45350)-(64)^2$$

$$\sigma_y^2 = 4535-4096 \qquad \Rightarrow \sigma_y^2 = 439 \qquad \Rightarrow \sigma_Y = 20.95$$

$$\text{Cov(x,y)}=\frac{\sum x_i y_i}{n}-(\bar{x})(\bar{y}) \qquad \Rightarrow \text{Cov}(x,y)=\frac{1}{10}(42575)-(61)(64)$$

$$\text{Cov(x,y)} = 4257.5-3904 \qquad \Rightarrow \text{Cov(x,y)} = 353.5$$

Regression coefficient of y on x is

$$b_{yx} = \frac{\text{Cov(x,y)}}{\sigma_x^2} \qquad \Rightarrow b_{yx} = \frac{353.5}{349} = 1.0128$$

Regression line of y on x is

$$y-\bar{y} = b_{yx}(x-\bar{x}) \qquad \Rightarrow y-64 = 1.0128(x-61.7808)$$

$$y = 1.0128x-61.7808+64 \qquad \Rightarrow y = 1.0128x+2.2192 \qquad \cdots(1)$$

Given that estimate Y if sales are Rs. 75000 i.e. X=75000, From equation (1)

Y=1.0128(75000) + 2.2192

Y=75960 + 2.2192

Y=75962.22

Substitute X in equation (1), we compute Y, Hence $y-\hat{y}$ and $\left(y-\hat{y}\right)^2$, thus we complete the column (6) (7) (8) in the above table

Residual sum of squares

$$\text{SSE}=\sum\left(y-\hat{y}\right)^2 \qquad \Rightarrow \text{SSE}=809.4196$$

Total sum of square

$$\text{SST}=\sum(y-\bar{y})^2 \qquad \Rightarrow \text{SST}=\sum y^2-n\left(\bar{y}\right)^2$$

$$\text{SST}=45350-10\times\left(64\right)^2 \qquad \Rightarrow \text{SST}=45350-40960$$

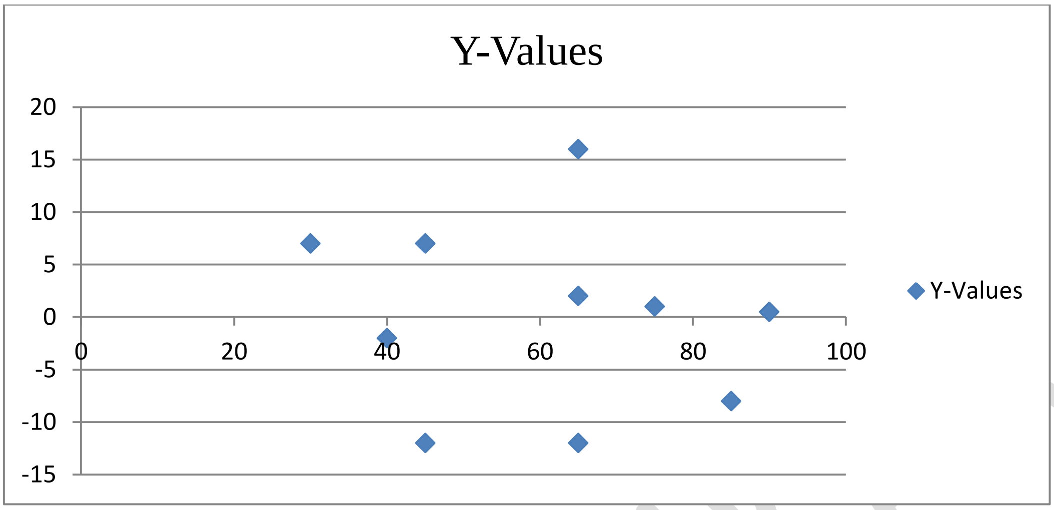
$$\text{SST}=4390$$

$$r^2=1-\frac{\text{SSE}}{\text{SST}} \qquad \Rightarrow r^2=1-\frac{809.4196}{4390}$$

$$r^2 = \frac{4390-809.4196}{4390} \qquad \Rightarrow r^2 = \frac{3580.58}{4390}$$

$$r^2 = 0.81562$$

Residual Plot:-



Interpretation:-There is no pattern seen on residual plot.

[11] The two lines of regression of x+2y-5=0 & 2x+3y-8=0

(i) Compute the correlation between X & Y

(ii) Estimate X when y =2.5

Solution:- Given

x+2y-5=0…………(1)

2x+3y-8=0..……….(2)

(i) To compute the correction between x &y assume the regression equation (1) be X on Y

$x+2y-5 = 0$

$x+2y = 5$

$x = 5-2y$

$x = a+by$

$b_{xy} = -2.$

Now the regression equation (2) be Y on X

2x+3y-8=0

3y=8-2x

$$y=\frac{8}{3}-\frac{2}{3}x$$

Compare y=a + b x

$$b_{yx}=\frac{-2}{3}$$

As we know

$r^2=b_{XY}\times b_{YX}$ $\Rightarrow r^2=-2\times\frac{-2}{3}$

$r^2=\frac{4}{3}$ $\Rightarrow r^2=1.33$

$r=1.15$ but $r\le 1$

[12] For bivariate data the regretion equation are 4x-5y+33=0 & 20x-9y=107 find means of x & y find correction coefficient between x&y also estimate y when x=10

Solution:- Given equations are

4x-5y+33=0...........[i]

20x-9y=107..........[ii]

i] To find means of x & y since two regresion lines intersected at $\bar{x}$ $\bar{y}$ therefore equations [i] and [ii]

$4\bar{x}-5\bar{y}=-33$...............[iii]

$20\bar{x}-9\bar{y}=107$............[iv]

Multiplying a constant 5 by equation [iii], we get

$20\bar{x}-25\bar{y}=-165$.........[v]

Substract equation [iv] from [v], we get

$$(20\bar{x}-25\bar{y})-(20\bar{x}-9\bar{y})=-165-107$$

$$-16\bar{y}=-272$$

$$\bar{y}=\frac{-272}{16}$$

$$\bar{y}=17$$

Put value of $\bar{y}$ in eqn ..[i]

$$4\bar{x} - 5(17) + 33 = 0 \quad \Rightarrow \quad 4\bar{x} - 85 + 33 = 0$$

$$4\bar{x} = 52 \quad \Rightarrow \quad \bar{x} = 13$$

ii] To find correlation coefficient between x&y

Let, regression equation [ii] is y on x

$4x - 5y + 33 = 0 \qquad \Rightarrow 4x - 5y = -33$

$-5y = -33 - 4x$

hence our supposition is wrong

Let regression equation [i] be y on x

$x + 2y - 5 = 0 \qquad \Rightarrow 2y = 5 - x$

$y = \frac{5}{2} - \frac{1}{2}x$

Hence by $b_{YX} = -\frac{1}{2}$

Now regression equation [ii] be x on y

$2x + 3y - 8 = 0 \qquad \Rightarrow 2x = 8 - 3y$

$x = 4 - \frac{3}{2}y$

Hence by $b_{XY} = -\frac{3}{2}$

We know,

$r^2 = b_{xy} \times b_{yx} \qquad \Rightarrow r^2 = \frac{-3}{2} \times \frac{1}{2}$

$r^2 = \frac{-3}{4} \qquad \Rightarrow r^2 = 0.75$

$r = 0.866$

2] Estimate x when $y = 2.5$

$2x + 3y - 8 = 0 \qquad \Rightarrow 2x + 3(2.5) - 8 = 0$

$2x - 0.5 = 0 \qquad \Rightarrow 2x = 0.5$

$x = \frac{0.5}{2} \qquad \Rightarrow x = 0.25$

$y = \frac{33}{5} + \frac{4}{5}x \qquad \Rightarrow y = a + b_{YX}x$

Hence, $b_{YX} = \frac{4}{5}$

Then regression equation [ii] is x on y

$20x - 9y = 107 \qquad \Rightarrow 20x = 107 + 9y$

$x = \frac{107}{20} + \frac{9}{20}y \qquad \Rightarrow \therefore b_{XY} = \frac{9}{20}$

As we know

$r^2 = b_{XY}.b_{YX} \qquad \Rightarrow r^2 = \frac{9}{20} \times \frac{4}{5}$

$r^2 = \frac{36}{100} \qquad \Rightarrow r = 0.6$

iii] For estimating y when x=10

$y = \frac{33}{5} + \frac{4}{5}x \qquad \Rightarrow y = \frac{33}{5} + \frac{4}{5} \times 10$

$y = 14.6$

[13] For a certain bivariate dada the list square lines of regression are 4y-x=19 and 9x-y=39 obtain

(i) Regression coefficient of x on y

(ii) Regression coefficient y on x

(iii) Correlation coefficient between x and y

Answer:- Given equations are,

4y-x=90………… [i]

9x-y=39………….[ii]

Let equation [i] become a regression line x on y

$4y - x = 19 \qquad \Rightarrow -x = 19 - 4y$

$x = -19 + 4y \qquad \Rightarrow x = a + b_{XY}y$

$b_{XY} = 4$

Let equation [ii] become regression line y on x

$9x - y = 39 \qquad \Rightarrow -y = 39 - 9x$

$y = -39 + 9x \qquad \Rightarrow y = a + b_{YX}x$

$\therefore b_{YX} = 9$

As we know

$r^2 = b_{XY}.b_{YX}$ $\Rightarrow r^2 = 4 \times 9$

$r^2 = 36$ $\Rightarrow r = 6$

But $r \le 1$, Hence our assumption was wrong and therefore alternate the equations

$9x - y = 39$ $\Rightarrow 9x = 39 + y$

$x = \frac{39}{9} + \frac{1}{9}y$ $\Rightarrow x = a + b_{XY}y$

$\therefore b_{XY} = \frac{1}{9}$

Regression equation [i] is y on x

$4y - x = 19$ $\Rightarrow 4y = 19 + x$

$y = \frac{19}{4} + \frac{1}{4}x$ $\Rightarrow y = a + b_{YX}x$

$b_{YX} = \frac{1}{4}$

Correlation coefficient

$r^2 = b_{XY}.b_{YX}$ $\Rightarrow r^2 = \frac{1}{4} \times \frac{1}{9}$

$r^2 = \frac{1}{36}$ $\Rightarrow r = 0.16$

[14] The equation of the two regression lines are 2x+3y-6=0 & 5x+7y-12=0 obtain

(a) correlation coefficient between x & y (b) $\frac{\sigma_x}{\sigma_y}$.

Solution :- Let 2x+3y-6=0…….[i]

5x+7y-12=0……[ii]

(a) Now, assume regression equation (i) is y on x

$2x + 3y = 6$ $\Rightarrow 3y = 6 - 2x$

$y = 2 - \frac{2}{3}x$ $\Rightarrow y = a + b_{YX}x$

$\therefore b_{YX} = -\frac{2}{3}$

Let regression equation (ii) is x on y

$5x+7y=12 \qquad \Rightarrow 5x=12-7y$

$x=\frac{12}{5}-\frac{7}{5}y \qquad \Rightarrow$ x = a + b_{XY}y

$b_{XY}=-\frac{7}{5}$

As we know

$r^2=b_{XY}.b_{YX} \qquad \Rightarrow r^2=\left(\frac{-7}{5}\right)\times\left(\frac{-2}{3}\right)$

$r^2=-\left(\frac{14}{15}\right) \qquad \Rightarrow r^2=-0.93$

r = -0.96

(b) To find $\frac{\sigma_x}{\sigma_y}$ as we know,

$b_{yx}=\frac{Cov(x,y)}{\sigma_x^2}$ …….(1)

$b_{xy}=\frac{Cov(x,y)}{\sigma_y^2}$ ……(2)

Divide equation (1) by (2), we get

$\frac{b_{yx}}{b_{xy}}=\frac{\frac{Cov(x,y)}{\sigma_x^2}}{\frac{Cov(x,y)}{\sigma_y^2}} \qquad \Rightarrow \frac{b_{yx}}{b_{xy}}=\frac{Cov(x,y)}{\sigma_x^2}\times\frac{\sigma_y^2}{Cov(x,y)}$

$\frac{b_{yx}}{b_{xy}}=\frac{\sigma_y^2}{\sigma_x^2} \qquad \Rightarrow \frac{\sigma_x^2}{\sigma_y^2}=\frac{b_{xy}}{b_{yx}}$

$\frac{\sigma_x^2}{\sigma_y^2}=\frac{-7}{5}\times\frac{-3}{2} \qquad \Rightarrow \frac{\sigma_x^2}{\sigma_y^2}=\frac{21}{10}$

$\frac{\sigma_x}{\sigma_y}=3.16$

[15] For a bivariate data on x and y the regression equation of two lines of regression are 3x-2y+1=0 & 3x-8y+13=0 predict the value of y for x = 4 and value of x for y =3

Solution:- Given equations are

– 3x-2y +1 = 0 ……(1)

3x-8y +13=0…….(2)

Assume regression equation (1) is x on y

$3x-2y=-1 \qquad \Rightarrow 3x=-1+2y$

$x=\frac{-1}{3}+\frac{2}{3}y \qquad \Rightarrow \therefore b_{XY}=\frac{2}{3}$

Let regression equation (2) is y on x

$3x-8y+13=0 \qquad \Rightarrow 3x-8y=-13$

$-8y=-13-3x \qquad \Rightarrow y=\frac{13}{8}+\frac{3}{8}x$

$y=a+b_{YX}x \qquad \Rightarrow b_{YX}=\frac{3}{8}$

$y=a+bx \qquad \Rightarrow b_{yx}=\frac{3}{8}$

We know,

$r^2=b_{yx}\times b_{xy} \qquad \Rightarrow r^2=\frac{3}{8}\times\frac{2}{3}$

$r^2=\frac{2}{8} \qquad \Rightarrow r=0.5$

Now equation [i] is x on y, Put the value of y for getting x in equation [iii]

$x=\frac{2}{3}y-\frac{1}{3} \qquad \Rightarrow x=\frac{2}{3}\times3-\frac{1}{3}$

$x=\frac{5}{3} \qquad \Rightarrow x=1.66$

Put value in equation [iv]

$y=\frac{13}{8}+\frac{3}{8} \qquad \Rightarrow y=\frac{25}{8}$

$y=3.125$

[16] You are given the following information about two variables x & y, n=10,

$\bar{x}=5.5, \bar{y}=4, \sum x^2=385, \sum y^2=192, \sum xy=185$.Find

(i) Regression line of y on x (ii) Regression line x on y

Solution:- Regression line of x on y is

$$(x-\bar{x}) = b_{XY}(y-\bar{y}) \qquad \Rightarrow (x-\bar{x}) = r \times \frac{\sigma_X}{\sigma_Y}(y-\bar{y})\ldots\ldots\ldots(i)$$

Regression line y on x is

$$(y-\bar{y}) = b_{YX}(x-\bar{x}) \qquad \Rightarrow (y-\bar{y}) = r \times \frac{\sigma_Y}{\sigma_X}(x-\bar{x})\ldots\ldots.(i)$$

$$\bar{x} = 5.5 \qquad \Rightarrow \bar{y} = 4$$

$$\sigma_X^2 = \frac{1}{n}\sum x_i^{\ 2} \text{-} \bar{x}^2 \qquad \Rightarrow \sigma_X^2 = \frac{1}{10} \times 185\text{-}5.5^2$$

$$\sigma_X^2 = 38.5\text{-}30.25 \qquad \Rightarrow \sigma_X^2 = 8.25$$

$$\sigma_Y^2 = \frac{1}{n}\sum y_i^{\ 2} \text{-} \bar{y}^2 \qquad \Rightarrow \sigma_Y^2 = \frac{1}{10} \times 192\text{-}4^2$$

$$\sigma_Y^2 = 19.2\text{-}16 \qquad \Rightarrow \sigma_Y^2 = 3.2$$

$$\text{Cov}(x,y) = \frac{1}{n}\sum x_i y_i \text{-} \bar{x}\,\bar{y} \qquad \Rightarrow \text{Cov}(x,y) = \frac{1}{10} \times 185\text{-}(5.5)(4)$$

$$\text{Cov}(x,y) = 18.5\text{-}22 \qquad \Rightarrow \text{Cov}(x,y) = 3.2$$

$$b_{XY} = \frac{\text{cov}(x,y)}{\sigma_Y^2} \qquad \Rightarrow b_{XY} = \text{-}\frac{3.5}{3.2}$$

$$b_{XY} = \text{-}1.0937$$

$$b_{YX} = \frac{\text{Cov}(x,y)}{\sigma_X^2} \qquad \Rightarrow b_{YX} = \frac{\text{-}3.5}{8.25}$$

$$b_{YX} = \text{-}0.4242$$

(i) Regression line X on Y is

$$(x-\bar{x}) = b_{XY}(y-\bar{y}) \qquad \Rightarrow (x-5.5) = -1.0937(y-4)$$

$$x = -1.0937y + 4.3748 \qquad \Rightarrow x = -1.0937y + 9.8748$$

$$x = 9.8748 - 1.0937y$$

(ii) Regression line of Y on X

$$(y-\bar{y}) = b_{YX}(x-\bar{x}) \qquad \Rightarrow y - 4 = -0.4242x + 2.3331$$

$$y = -0.4242x + 6.3331 \qquad \Rightarrow y = 6.3331 - 0.4242x$$

[17] Compute regression coefficient from the following data

n = 8, $\sum(x\text{-}45) = \text{-}40$, $\sum(x\text{-}45)^2 = 4400$, $\sum(y\text{-}150) = 280$, $\sum(y\text{-}150)^2 = 167432$, $\sum(x\text{-}45)(y\text{-}150) = 21680$

Answer:

$\bar{x} = \frac{1}{n}\sum x_i \qquad \Rightarrow \bar{x} = \frac{1}{8}\times -40$

$\bar{x} = -5$

$\bar{y} = \frac{1}{n}\sum y_i \qquad \Rightarrow \bar{y} = \frac{1}{8}\times 280$

$\bar{y} = 35$

$\sigma_X^2 = \frac{1}{n}\sum x_i^{\ 2} - \bar{x}^2 \qquad \Rightarrow \sigma_X^2 = \frac{1}{8}\times 4400 - (-5)^2$

$\sigma_X^2 = 550 - 25 \qquad \Rightarrow \sigma_X^2 = 525$

$\sigma_Y^2 = \frac{1}{n}\sum y_i^{\ 2} - \bar{y}^2 \qquad \Rightarrow \sigma_Y^2 = \frac{1}{8}\times 167432 - (35)^2$

$\sigma_Y^2 = 20929 - 1225 \qquad \Rightarrow \sigma_Y^2 = 19704$

$\mathrm{Cov}(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y} \qquad \Rightarrow \mathrm{Cov}(x,y) = \frac{1}{8}\times 21680 - (-5)(35)$

$\mathrm{Cov}(x,y) = 2710 + 175 \qquad \Rightarrow \mathrm{Cov}(x,y) = 2885$

$b_{XY} = \frac{\mathrm{cov}(x,y)}{\sigma_Y^2} \qquad \Rightarrow b_{XY} = \frac{2885}{19704}$

$b_{XY} = 0.1464$

$b_{YX} = \frac{\mathrm{Cov}(x,y)}{\sigma_X^2} \qquad \Rightarrow b_{YX} = \frac{2885}{525}$

$b_{YX} = 5.4952$

[18] From the following data obtain the yield when the rainfall is 29 inches

	Rainfall(inches)	Yield(per acre)
A.M.	27	40 quintal
S.D.	3	6 quintal

Correlation coefficient between rainfall and yield is 0.8

Answer:-

Let $\bar{x} = 27$, $\bar{y} = 40$, $\sigma_x = 3$, $\sigma_y = 6$, $r = 0.8$

Regression equation x on y

$x-\bar{x}= r\times\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \qquad \Rightarrow x-27=0.8\times\frac{3}{6}(y-40)$

x=11+0.44.............(i)

Regression equation y on x

$y-\bar{y} =r\times\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \qquad \Rightarrow y-40 =0.8\times\frac{6}{3}(x-27)$

y =1.6x-43.2+40

y =1.6x-3.2..................(ii)

Put x=29 in equation (ii) we get,

Y=43.2

The yield is 43.2 per acre when the rainfall is 29 inches

[19] For a bivariate data we have

$\bar{x}=53$, $\bar{y}=28$, $b_{YX}=-1.5$, $b_{XY}=-0.2$ find

i] Correlation coefficient between x & y

ii] Estimate of y for x=60

iii] Estimate x for y=30

Solution:-

[i] $r^2 = b_{XY}\times b_{YX} \qquad \Rightarrow r^2 = -0.2\times(-1.5)$

$r^2 = -0.3 \qquad \Rightarrow r = -0.5477$

[ii]Regression line x on y is

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

(x-53)=-0.2(y-28)

x=-0.2y+58.6.............(i)

Put y =30 in equation (i)

x=58.6-0.2(30)

x=52.6

[iii]Regression line y on x is

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

(y-28)=-1.5(x-53)

y=-1.5x+107.5

$y = 107.5 - 1.5x$................(ii)

$x=60$ put this value in equation (ii),we get

$y=107.5-1.5(60)$

$y=17.5$

[20] Obtain the coefficient of correlation & the regression lines from the following data.

	X	Y
No. of observation	15	15
Sum of squares of deviation from mean	136	138
Sum of product of deviation from mean	122	

Solution:-

Here, $\sum(x-\bar{x})^2=136, \sum(y-\bar{y})^2=138$, $n=15, \sum(x_i-\bar{x})(y_i-\bar{y})=122$

Now,

$Cov(x,y)=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$ $\Rightarrow Cov(x,y)=\frac{1}{15}\times 122$

$Cov(x,y)=8.1333$

$\sigma_X^2=\frac{1}{n}\sum(x_i-\bar{x})^2$ $\Rightarrow \sigma_X^2=\frac{1}{15}\times 136$

$\sigma_X^2=9.0666$ $\Rightarrow \sigma_X=3.01107$

$\sigma_Y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \sigma_Y^2 = \frac{1}{15} \times 138 \sigma_Y^2 = 9.2 \sigma_Y = 3.0331$

Regression coefficient of x on y & y on x

$b_{XY} = \frac{\text{cov}(x,y)}{\sigma_Y^2}$ $\Rightarrow b_{XY} = \frac{8.1333}{9.2}$

$b_{XY} = 0.8840$

$b_{YX} = \frac{\text{Cov}(x,y)}{\sigma_X^2}$ $\Rightarrow b_{YX} = \frac{8.1333}{9.066}$

$b_{YX} = 0.8971$

$r^2 = b_{XY} \times b_{YX}$ $\Rightarrow r^2 = 0.8840 \times 0.8971$

$r^2 = 0.7930$ $\Rightarrow r = 0.8905$

Regression line x on y is

$(x-\bar{x}) = r.\frac{\sigma_x}{\sigma_y}(y-\bar{y})$ $\Rightarrow (x-\bar{x}) = 0.8905 \times \frac{3.01109}{3.0331}(y-\bar{y})$

$(x-\bar{x}) = 0.8970(y-\bar{y})$

Regression line y on x is

$(y-\bar{y}) = r.\frac{\sigma y}{\sigma x}(x-\bar{x})$ $\Rightarrow (y-\bar{y}) = \frac{0.8905}{1.0073}(x-\bar{x})$

$(y-\bar{y}) = 0.8840(x-\bar{x})$

QUESTION BANK ON REGRESSION

[1] ________ gives the mathematical relations of the variables

(a) correlation (b) regression

(c) both (d) none

Answer: (b) regression

[2] Under Algebraic Method we get ________ linear equations.

(a) one (b) two

(c) three (d) none

Answer: (c) three

[3] In linear equations Y= a+bX and X=a+bY ‘a’ is the

(a) intercept of the line (b) slope

(c) both (d) none

Answer: (b) slope

[4] In linear equation Y=a+bX and X=a+bY ‘b’ is the

(a) intercept of the line (b) slope of the line

(c) both (d) none

Answer: (b) slope of the line

[5] The regression equtions is Y=a+bX and X a+bY are based on the Method of the

(a) greatest squares (b) least squares

(c) both (d) none

Answer: (a) greatest squares

[6] The line Y= a + bX represents the regressions equations of

(a) Y on X (b) X on Y

(c) both (d) none

Answer: (a) Y on X

[7] The lines X= a+bY represents the regression equation of

(a) Y on X (b) X on Y

(c) both (d) none

Answer: (b) X on Y

[8] Two regression lines always intersect at the means

(a) true (b) false

(c) none (d) both

Answer: (a) true

[9] r, b_{xy}, b_{yx} all have _______ sign

(a) different (b) same

(c) both (d) done

Answer: (b) same

[10] The regression coefficients are zero if r is equal to

(a) 2 (b) -1

(c) 1 (d) 0

Answer: (d) 0

[11] The regression lines are identical if r is equal to

(a) +1 (b) -1

(c) ±1 (d) 0

Answer: (b) -1

[12] The regression lines are perpendicular to each other if r is equal to

(a) 0 (b) +1

(c) -1 (d) ±1

Answer: (d) ±1

[13] Feature of least square regression lines are ________ The sum of the deviations at the Y' s or the X's from their regressions lines are zero

(a) true (b) false

(c) both (d) none

Answer: (c) both

[14] The coefficient of determination is defined by the formula

(a) $r^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}}$

(b) $r^2 = \frac{\text{unexplained variance}}{\text{total variance}}$

(c) both

(d) none

Answer: (a) $r^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}}$

[15] If the line $Y = \frac{13}{2} - \frac{3X}{2}$ is the regression equation of y on the x then b_{yx} is

(a) 2/3 (b)-2/3

(c) 3/2 (d)-3/2

Answer: (a) 2/3

[16] The line, $X=19-\frac{5}{2}Y$ is the regression equation x on y then b_{xy} is

(a) 19/2 (b) 5/2

(c) -5/2 (d) -2/5

Answer: (c) -5/2

[17] The line $X=\frac{31}{6}-\frac{1}{6}Y$ is the regression equation of

(a) Y on X (b) X on Y

(c) both (d) we can not say

Answer: (d) we can not say

[18] In the regression equation x on y, $X=\frac{35}{8}-\frac{2}{5}Y$, b_{xy} is equal to

(a)-2/5 (b) 35/8 (c) 2/5 (d) 5/2

Answer: (a)-2/5

[19] The correlation coefficient being +1 if the slope of the straight line in a scatter dingram is

(a) positive (b) negative (c) zero (d) none

Answer: (a) positive

[20] The correlation coefficient being -1 if the slope of the straight line in a scatter diagram is

(a) positive (b) negative (c) zero (d) none

Answer: (b) negative

[21] The more scattered the points are around a straight line in a scattered diagram the……………. is the correlation coefficient.

(a) zero (b) more (c) less (d) none

Answer: (c) less

[22] If the values of y are not affected by changes in the values of x, the variables are said to be

(a) correlated (b) uncorrelated

(c) both (d) zero

Answer: (b) uncorrelated

[23] If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then correlatiom in said to be

(a) non -linear (b) linear (c) both (d) none

Answer: (b) linear

[24] Two regression lines councide when r is equal to

(a) 0 (b) 2

(c) ± 1 (d) none

Answer: (c) ± 1

[25] Neither y nor x can be estimated by a linear function of the other variable when r is equal to

(a) +1 (b) -1

(c) 0 (d) none

Answer: (c) 0

[26] When r = 0 then cov (x, y) is equal to

(a) +1 (b) -1

(c) 0 (d) none

Answer: (c) 0

[27] When the variables are not independent, the correlation coefficient may be zero.

(a) true (b) false

(c) both (d) none

Answer: (a) true

[28] b_{xy} is called regression coefficient of

(a) x on y (b)y on x

(c) both (d)none

Answer: (a) x on y

[29] b_{yx} is called regression coefficient of

(a) x on y (b) y on x

(c) both (d)none

Answer: (b) y on x

[30] The slopes of the regression line of y on x is denoted by

(a) b_{yx} (b) b_{xy} (c) b_{xx} (d) b_{yy}

Answer: (a) b_{yx}

[31] The slopes of the regression line of x on y is denoted by

(a) b_{yx} (b) b_{xy} (c) b_{xx} (d) b_{yy}

Answer: (b) b_{xy}

[32] The angle between the regression lines depends on

(a) correlation coefficient (b) regression coefficient

(c) both (d) none

Answer: (a) correlation coefficient

[33] If x and y satisfy the relationship y= -5+7x, the value of r is

(a) 0 (b) -1

(c) +1 (d) none

Answer: (c) +1

[34] If b_{yx} and b_{xy} are negative the r is

(a) positive (b) negative

(c) zero (d)none

Answer: (b) negative

[35] Correlation coefficient r lie between the regression coefficients b_{yx} and b_{xy}

(a) true (b)false (c) both (d)none

Answer: (a) true

[36] Since the correlation coefficient r cannot be greater than 1 numerically, the product of the regression coefficient must

(a) not exceed 1 (b) exceed 1

(c) be zero (d) none

Answer: (a) not exceed 1

[37] The correlation coefficient r is the -------of the two regression coefficient b_{yx} and b_{xy}

(a) A.M. (b) G.M.

(c) H.M. (b) none

Answer: (b) G.M.

[38] Which is true?

(a) $b_{yx} = r\frac{\sigma_x}{\sigma_y}$ (b) $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

(c) $b_{yx} = r\frac{\sigma_x}{\sigma_y^2}$ (d) none

Answer: (b) $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

[39] Maximum value of rank correlation coefficient is

(a) -1 (b) +1 (c) 0 (d) none

Answer: (b) +1

[40] The partial correlation coefficient lies between

(a) -1 and +1inclusive of these two value (b) 0 and +1

(c) -1 and (d) none

Answer: (a) -1 and +1inclusive of these two value

[41] Regression analysis is concerned with

[a] Establishing a mathematical relationship between two variables

[b] Measuring the extent of association between two variables

[c] Predicting the value of the dependent variable for a given value of the independent variable.

[d] Both (a) and (c).

Answer: [d]

[42] If case the correlation coefficient between two variables is 1, the relationship between the two variables would be

(a) $y = a + bx$ (b) $y = a + bx, \quad b > 0$

(c) $y = a + bx, \quad b < 0$ (d) $y = a + bx$, both a and b being positive

Answer:[b]

[43] If the relationship between two variables x and y is giving by 2x+3x+4=0, then the value of the correlation between x and y is

(a) 0 (b) 1

(c) -1 (d) negative.

Answer:[c]

[44] If there are two variables x and y, the number of regression equation could be

(a) 1 (b) 2 (c) Any other (d) 3

Answer:[b]

[45] Since Blood Pressure of a person depends on age, we need consider

(a) The regression equation of Blood Pressure on age

(b) The regression equation of age on Blood Pressure

(c) Both (A) and (b)

(d) Either (a) or (b)

Answer: [a]

[46] The method applied for deriving the regression equations is knows as

(a) Least square (b) Concurrent deviation

(c) Product moment (d) Normal equation

Answer:[a]

[47] The different between the observed value and the estimated value in regression analysis is knows as

(a) Error (b) Residue

(c) Deviation (d) (a) or (b)

Answer:[d]

[48] The error in case of regression equations are

(a) Positive (b) Negative

(c) Zero (d) All the above

Answer:[d]

[49] The regression line of y on x is derived by

(a) The minimization of vertical distances in the scatter diagram

(b) The minimization of horizontal distance in the scatter diagram

(c) Both (a) and (b)

(d) (a) or (B)

Answer:[a]

[50] The two lines of regression become identical when

(a) r = 1 (b) r = -1

(c) r = 0 (d) (a) or (b)

Answer:[d]

[51] What are the limits of the two regression coefficients?

(a) No limit

(b) Must be positive

(c) one positive and the other negative

(d)Product of the regression coefficient must be numerically less than unit.

Answer: [d]

[52] The regression coefficients remain unchanged due to a

(a) Shift of origin (b) Shift of scale

(c) Both (a) and (b) (d) (a) or (b).

Answer: (a) Shift of origin

[53] If the coefficient of correlation between two variables is -0 9, then the coefficient of determination is

(a) 0.9 (b) 0.81 (c) 0.1 (d) 0.19

Answer: (b) 0.81

[54] If the coefficient of correlation between two variables is 0.7 then the percentage of variation unaccounted for is

(a) 70% (b) 30% (c) 51% (d) 49%

Answer: (c) 51%

[55] If y = a + bx, then coefficient of correlation between x and y?

(a) 1 (b) -1

(c) 1 or -1 according as b > 0 or b < 0 (d) none of these.

Answer: (c)

[56] If u + 5x = 6 and 3y – 7v =20 and the correlation coefficient between x and y is 0.58 then what would be the correlation coefficient between u and v?

(a) 0.58 (b) -0.58

(c) -0.84 (d) 0.84

Answer: (b) -0.58

[57] If the relation between x and u is 3x + 4u +7 = 0 and the correlation coefficient between x and y is -0.6, then what is the correlation coefficient between u and y?

(a) -0.6 (b) 0.8

(c) 0.6 (d) -0.8

Answer: (c) 0.6

[58] Following are the two normal equation obtained for deriving the regression line of y and x: 5a+10b=40 and 10a+25b=95. The regression line of y on x is given by

(a) 2x+3y=5 (b) 2y+3x=5

(c) y =2+3x (d) y=3+5x

Answer: (c) y =2+3x

[59] If the regression line of y on x and of x on y is given by 2x+3y=-1 and 5x+6y=-1 then the arithmetic means of x and y are given by

(a) (1, -1) (b) (-1, 1) (c) (-1, -1) (d) (2, 3)

Answer: (a) (1, -1)

[60] Given the regression equations as 3x+y=13 and 2x+5y=20, which one is the regression equation of y on x?

(a) 3x+y=13 (b) 2x+5y=20

(c) both (a) and (b) (d) none of these

Answer: (b) 2x+5y=20

[61] Given the following equation: 2x-3y=10 and 3x+4y=15, which one is the regression equation of x on y ?

(a) 2x-3y=10 (b) 3x+4y=15

(c) both the equation (d) none of these

Answer: (d) none of these

[62] If u=2x+5 and v=-3y-6 and regression coefficient of y on x is 2.4, what is the regression coefficient of v on u?

(a) 3.6 (b) -3.6 (c) 2.4 (d) -2.4

Answer: (b) -3.6

[63] If 4y-5x=15 is the regression line of y on x and coefficient of correlation between x and y is 0.75, what is the value of the regression coefficient of x on y ?

(a) 0.75 (b) 0.9375

(c) 0.6 (d) none of these

Answer: (a) 0.75

[64] If the regression line of y on x and that of x on y are given by y=-2x+3 and 8x=-y+3 respectively, what is the coefficient of correlation between x and y?

(a) 0.5 (b) $-1/\sqrt{2}$

(c) -0.5 (d) none of these

Answer: (c) -0.5

[65] If the regression coefficient of y on x, the coefficient of correlation between x and y and variance of y are -3/4, $\sqrt{\frac{3}{2}}$ and 4 respectively, what is the variance of x?

(a) $\frac{2}{\sqrt{\frac{3}{2}}}$ (b) 16/3 (c) 4/3 (d) 4

Answer: (b) 16/3

[66] If y=3x+4 is the regression line of y on x and the arithmetic mean of x is -1, what is the arithmetic mean of y ?

(a) 1 (b) -1

(c) 7 (d) none of these

Answer: (a) 1

[67] The regressions equation of y on x for the following dada

X	4 1	8 2	6 2	3 7	5 8	9 6	12 7	7 4	12 3	10 0
Y	2 8	5 6	3 5	1 7	4 2	8 5	10 5	6 1	98	73

(a) Y=1.2x-15 (b) Y =1.2x+15

(c) Y=o.93x-14.64 (d) Y =1.5x-10.89

Answer: (c)

[68] The following data relate to the heights of 10 pairs of fathers and sons;
(175,173) , (172, 172), (167, 171), (168,171), (172,173), (171,170), (174,173), (176,175), (169,170) (170,173)

The regression equation of height of son on that of father is given by

(a) y=100+5x (b) y=99.708+0.405x

(c) y=89.653 +0.582 x (d) y=88.758+0.562x

Answer: (b)

[69] The two regression coefficients for the following data ;

X	38	23	43	33	28
Y	28	23	43	38	8

(a) 1.2 and 0.4 (b) 1.6and0.8

(c) 1.7 and 0.8 (d) 1.8 and 0.3

Answer: (a)

[70] For y =25, what is the estimated value of x , from the following data:

X	11	12	15	16	18	19	21
Y	21	15	13	12	11	10	9

(a) 15 (b) 13.926

(c) 13,588 (d) 14.986

Answer: (c)

[71] Given the following data

Variable	X	Y
Mean	80	98
Variance	4	9

Coefficient of correlation=0.6. What is the most likely value of y when x = 90?

(a) 90 (b) 103 (c) 104 (d) 107

Answer: (d) 107

[72] The two lines of regression are 8x+10y=25 and 16x+5y=12 respectively; If the variance of x is 25, what is the standard deviation of y ?

(a) 16 (b) 8 (c) 64 (d) 4

Answer: (b) 8

[73] Given below the information about the capital employed and profit earned by a company over the last twenty five years;

	mean	S.D.
Capital employed (0000' Rs.)	62	5
Profit earned (000Rs.)	25	6

Coefficient of correlation between capital employed and profit = 0.92. The sum of the regression coefficients for the above data would be;

(a) 1.871 (b) 2.358 (c) 1.968 (d) 2.346

Answer: (a) 1.871

[74] The coefficient of correlation between cost of advertisement and sales of a product on the basis of the following data;

Ad cost (000 Rs.)	75	81	85	105	93	113	121	125
Sales (000 Rs.)	35	45	59	75	43	79	87	95

(a) 0.85 (b) 0.89 (c) 0.95 (d) o.98

Answer: (c) 0.95

[A] THEORY QUESTIONS:

[1] Define the term 'regression' in details.

[2] State utility of regression lines.

[3] Define regression coefficients and state its properties.

[4] How would you interpret regression coefficients?

[5] State the situations where regression analysis is used

[6] Derive the expression for regression lines of Y on X.

[7] Derive the expression for regression lines of X on Y.

[8] Derive standard error of regression estimate

[9] Explain the following terms:

[i] Explained variation of dependent variable

[ii] Unexplained variation of dependent variable

[iii] Coefficient of determination

[10] Show that regression lines intersect at $(\bar{x}, \bar{y})$.

[11] Show that r, b_{YX}, b_{XY} have same algebraic sign.

[B] Numerical Problems:

[1] Determine the two regression lines from the following data:

X	1	2	3	4	5
Y	5	4	3	2	1

[2] Determine the two regression lines from the following data:

X	2	4	5	8	10
Y	4	16	25	64	100

[3] Following are the data of marks in Statistics and Mathematics of 5 students

Statistics	78	82	88	90	95
Mathematics	71	76	80	88	100

(i) Calculate Correlation coefficients.

(ii) Calculate regression coefficients.

(iii) Estimate marks in Mathematics when he has scored 93 marks in Statistics.

(iv) Estimate marks in Statistics when he has scored 85 marks in Mathematics.

[4] From the following data, correlation coefficient between rainfall and yield is 0.8. Obtain the yield when the rainfall is 30 inches.

	Rainfall (inches)	Yield (per acre)
Arithmetic mean	28	40
Standard deviation	4	6

[5] For a bivariate data:

Arithmetic means	$\bar{X} = 53$	$\bar{Y} = 28$
Regression coefficient	$b_{YX} = -1.5$	$b_{XY} = -0.2$

Find

(i) Correlation coefficient between X and Y.

(ii) Estimate of Y when X = 60

(iii) Estimate of X when Y = 30

[6] The two regression equations of variables X and Y are 3X-Y-5 = 0 and 4X-3Y = 0. Find (i) Arithmetic mean of X and Y. (ii) Coefficient of variations of X and Y, if $\sigma_x = 2$. (iii) Correlation coefficient between X and Y.

[7] The two regression equations of variables X and Y are 8X-10Y = -66 and 40X-18Y = 214. Find (i) Arithmetic mean of X and Y. (ii) Correlation coefficient between X and Y.

[8] Find the regression line of Y on X from the following data:

$$n = 10, \sum x_i^2 = 385, \sum y_i^2 = 192, \bar{x} = 5.5, \bar{y} = 4, \sum (x_i - \bar{x})(y_i - \bar{y}) = 185$$

[9] Find the regression line of Y on X from the following data. Also, estimate Y when X = 0

$$n = 100, \sum x_i = 25, \sum y_i = 68, \sum x_i^2 = 167, \sum y_i^2 = 162, \sum (x_i - \bar{x})(y_i - \bar{y}) = 130$$

[10] Find the regression line of X on Y from the following data:

$$n = 20, \sum x_i^2 = 285, \sum y_i^2 = 172, \bar{x} = 4.5, \bar{y} = 3, \sum (x_i - \bar{x})(y_i - \bar{y}) = -40$$