# REGRESSION ANALYSIS

## What is regression?

Ans: The probable movement of one variable in terms of the other variables is called regression.

In other words the statistical technique by which we can estimate the unknown value of one variable (dependent) from the known value of another variable is called regression.

The term "regression" was used by a famous Biometrician Sir. F. Galton (1822-1911) in 1877.

Example: The productions of paddy of amount y is dependent on rainfall of amount x. Here x is independent variable and y is dependent variable.

## Regression analysis.

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

## Regression coefficient.

Ans: The mathematical measures of regression are called the coefficient of regression.

Let, $(x_1,y_1)$, $(x_2,y_2)$……….. $(x_n,y_n)$ be the pairs of n observations. Then the regression coefficient of y on x is denoted by $\mathbf{b_{yx}}$ and defined by

$$b_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Again, the regression coefficient of x on y is denoted by $\mathbf{b_{xy}}$ and defined by

$$b_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

## Regression lines:

If we consider two variables X and Y, we shall have two regression lines as the regression line of Y on X and the regression line of X on Y. The regression line of Y on X gives the most probable values of Y for given values of X and The regression line of X on Y gives the most probable values of X for given values of Y. Thus we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression lines will coincide i.e, we will have one line.

## Regression equation:

The regression equation of y on x is expressed as follows:

y = a + bx , where y is the dependent variable to be estimated and x is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

Here, a = $\overline{y}$ - $b\overline{x}$ $= \frac{\sum y}{n} - b\frac{\sum x}{n}$ and b= $\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Similarly, The regression equation of x on y is expressed as follows:

x = a + by, where x is the dependent variable to be estimated and y is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

Here, a = $\overline{x}$ - $b\overline{y}$

And b = $\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

## Properties of regression coefficient.

**Ans:** 1. Regression coefficient is independent of change of origin but not of scale.

2. Regression coefficient lies between $-\propto$ to $+\propto$ . i.e, $-<\propto \quad b_{yx} < \propto$ .

3. Regression coefficient is not symmetric. i.e, $b_{xy} \neq b_{yx}$

4. The geometric mean of regression coefficients is equal to correlation coefficient

   i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

5. The arithmetic mean of two regression coefficient is greater than correlation

   Coefficient. i.e, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

6. If one of regression coefficient is greater than unity the other must be less than

   unity. i.e, $b_{yx} \geq 1$ and $b_{xy} < 1$

7. Regression coefficient is not pure number.

## COEFFICIENT OF DETERMINATION

The coefficient of correlation between two variable series is a measure of linear relationship between them and indicates the amount of variation of one variable which is associated with or is accounted for by another variable. A more useful and readily comprehensible measure for this purpose is the *coefficient of determination* which gives the percentage variation in the dependent variable that is accounted for by the independent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient of determination is given by the square of the correlation coefficient, *i.e.*, $r^2$. Thus,

$$\text{Coefficient of determination} = r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

The coefficient of determination is a much useful and better measure for interpreting the value of *r*.

For example, if the value of $r = 0{\cdot}8$, we cannot conclude that 80% of the variation in the relative series (dependent variable) is due to the variation in the subject series (independent variable). But the coefficient of determination in this case is $r^2 = 0{\cdot}64$ which implies that only 64% of the variation in the relative series has been explained by the subject series and the remaining 36% of the variation is due to other factors. By the same argument while comparing two correlation coefficients, one of which is 0·4 and the other is 0·8 it is misleading to conclude that the correlation in the second case is twice as high as correlation in the first case. The coefficient of determination clearly explains this viewpoint, since in the case $r = 0{\cdot}4$, the coefficient of determination is 0·16 and in the case $r = 0{\cdot}8$, the coefficient of determination is 0·64, from which we conclude that correlation in the second case is four times as high as correlation in the first case.

The coefficient of determination, $r^2$, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The coefficient of determination is the ratio of the explained variation to the total variation. The coefficient of determination is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y.

The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained. The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variations. The further the line is away from the points, the less it is able to explain.

### Show that correlation coefficient is the geometric mean of regression coefficients. i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

Proof: Let, $(x_1,y_1)$, $(x_2,y_2)$……….$(x_n,y_n)$ be the pairs of n observations. Then the correlation coefficient between x and y is denoted by $r_{xy}$ and defined as,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Again, the regression coefficient of y on x is, $b_{yx} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Again, the regression coefficient of x on y is, $b_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$b_{yx} \times b_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \times \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\sqrt{b_{yx} \times b_{xy}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$= r_{xy}$ (proved)

**The arithmetic mean of two regression coefficient is greater than correlation coefficient. i.e, $\left(\frac{b_{yx}+b_{xy}}{2}\right) \geq r_{xy}$**

Proof: Let, $(x_1,y_1)$, $(x_2,y_2)$………. $(x_n,y_n)$ be the pairs of n observations. Then the regression coefficient of y on x is denoted by $b_{yx}$ and the regression coefficient of x on y is denoted by $b_{xy}$.

The arithmetic mean of $b_{yx}$ and $b_{xy}$ is A.M=$\left(\frac{b_{yx}+b_{xy}}{2}\right)$ and the geometric mean is

G.M= $\sqrt{b_{yx} \times b_{xy}}$

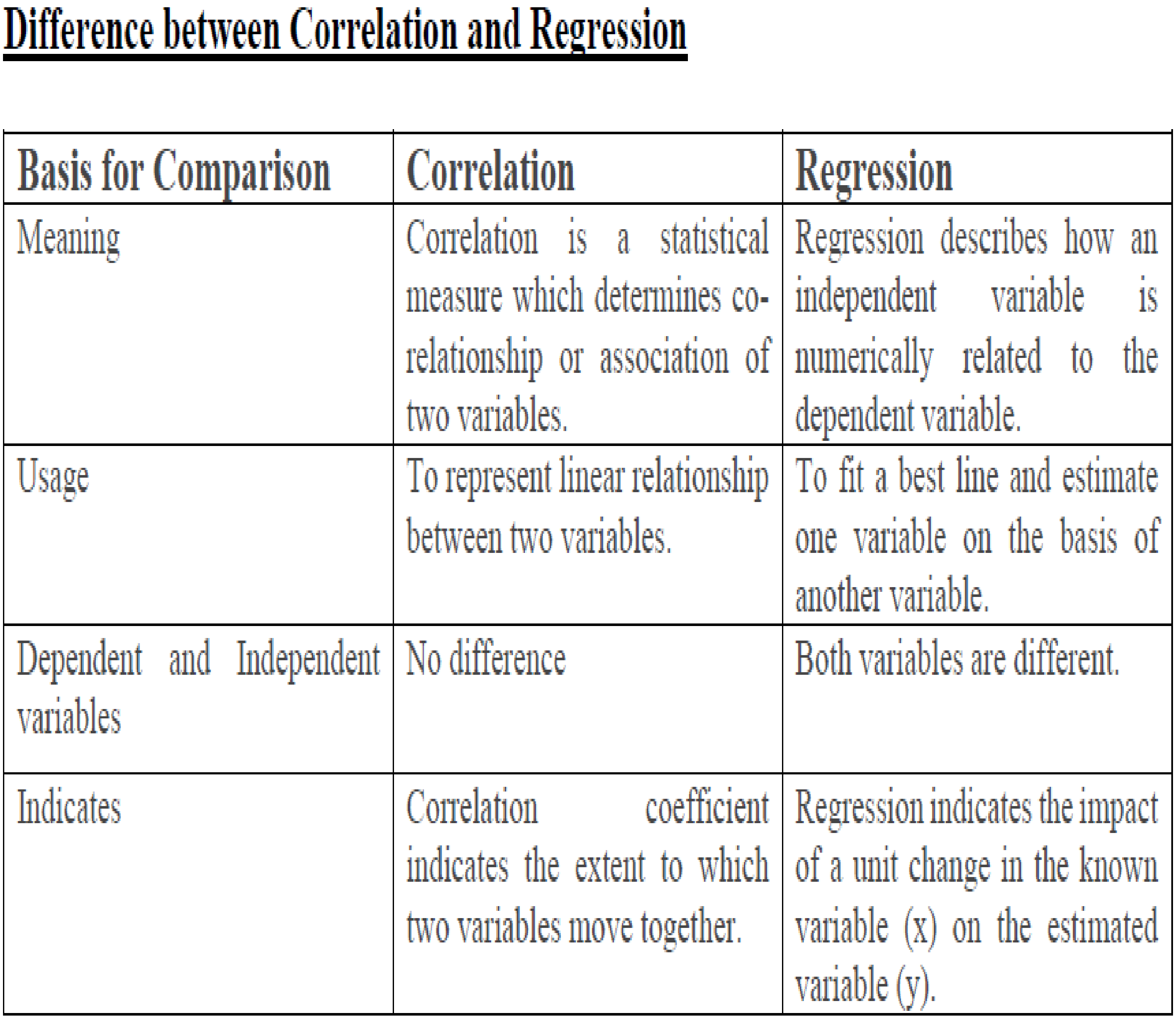
We know, Correlation coefficient is the geometric mean of regression coefficients.

i.e, $r_{xy}=\sqrt{b_{yx} \times b_{xy}}$

Since, A.M $\geq$ G.M

or, $\left(\frac{b_{yx}+b_{xy}}{2}\right) \geq \sqrt{b_{yx} \times b_{xy}}$

or, $\left(\frac{b_{yx}+b_{xy}}{2}\right) \geq r$ (proved



## Difference between Correlation and Regression

| Basis for Comparison | Correlation | Regression |
|---|---|---|
| Meaning | Correlation is a statistical measure which determines co-relationship or association of two variables. | Regression describes how an independent variable is numerically related to the dependent variable. |
| Usage | To represent linear relationship between two variables. | To fit a best line and estimate one variable on the basis of another variable. |
| Dependent and Independent variables | No difference | Both variables are different. |
| Indicates | Correlation coefficient indicates the extent to which two variables move together. | Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y). |

| | | |
|---|---|---|
| Objective | To find a numerical value expressing the relationship between variables. | To estimate values of random variable on the basis of the values of fixed variable. |

## General Difference between Correlation and Regression

| Sl.No. | Correlation | Regression |
|---|---|---|
| 1 | Correlation is the relationship between variables. It is expressed numerically | Regression means going back. The average relationship between variables is given as an equation |
| 2 | Between two variables, none is identified as independent or dependent variable. | One of the variable is independent and other is dependent variable in any particular context |
| 3 | Correlation does not mean causation. One variable need not be the cause and the other effect | Independent variable may be 'the cause and depndent variable,' the effect'. |
| 4 | There is spurious or non sense correlation | Regression is considered. Regression is considered only when the variables are related. |
| 5 | Correlation co efficient is independent change of origin and scale. | Correlation co efficient is independent change of origin but are affected by change of scale. |
| 6 | Correlation coefficient is a number between -1 to +1 | The two regression coefficients have the same sign, + or-. One of them can be greater than unity. But they can not be greater than one numerically simultaneously |
| 7 | Correlation coefficient is not in any unit of measurement | The regression coefficients is in the unit of measurement in the dependent variable |
| 8 | Correlation coefficient indicates the direction of co variation and the closeness of the linear relation between two variables | Regression equations give the value of dependent variable corresponding to any value of the independent variable. |

**Application problem-1:** A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Age of Husband(in years):x | 39 | 25 | 29 | 35 | 32 | 27 | 37 |
| Age of wife(in years):y | 37 | 18 | 20 | 25 | 25 | 20 | 30 |

(a) Compute the regression line of y on x.

(b) Predict the age of wife whose husband's age in 45 years.

(c) Find the regression line of x on y and estimate the age of husband if the age of his wife is 28 years.

(d) Compute the value of correlation coefficient with the help of regression coefficients.

**Solution**: The equation of the best –fitted regression line of y on x is $\hat{y}$ = a + bx

Where,
$$b = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad \text{and } a = \bar{y} - b\bar{x}$$

Computation table

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 39 | 37 | 1521 | 1369 | 1443 |
| 25 | 18 | 625 | 324 | 450 |
| 29 | 20 | 841 | 400 | 580 |
| 35 | 25 | 1225 | 625 | 875 |
| 32 | 25 | 1024 | 625 | 800 |
| 27 | 20 | 729 | 400 | 540 |
| 37 | 30 | 1369 | 900 | 1110 |
| $\sum x = 224$ | $\sum y = 175$ | $\sum x^2 = 7334$ | $\sum y^2 = 4643$ | $\sum xy = 5798$ |

(a) Here,
$$b = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{5798 - \frac{(224)(175)}{7}}{7334 - \frac{(224)^2}{7}} = 1.193$$

And $a = \bar{y} - b\bar{x}$

$$= \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$=\frac{175}{7}-(1.193)\ \frac{(224)}{7}\ = 25\text{-}38.176 = \text{-}13.176$$

Hence the fitted regression line is $\hat{y}$ = a + bx = -13.176 + 1.193x

(b) Hence, if the age of husband is 45, the probable age of wife would be $\hat{y}$ = -13.176 + 1.193x = -13.176 + 1.193 × 45 = 40.51 years.

(c) The equation of the best –fitted regression line of y on x is $\hat{x}$ = a + by

$$\text{Where, b } = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

$$= \frac{5798 - \frac{(224)(175)}{7}}{4643 - \frac{(175)^2}{7}} = 0.739$$

And $a = \bar{x}\ - b\bar{y}$

$$= \frac{\sum x}{n} - b\frac{\sum y}{n}$$

$$= \frac{224}{7} - 0.739\frac{175}{7}\ = 13.525$$

Hence the fitted regression line is $\hat{x}$ = a + by = 13.525+ 0.739y

Hence, if the age of wife is 28 years, the estimate age of husband is

$\hat{x}$ = a + by

= 13.525+ (0.739)(28) = 34.22 years.

**Application problem-2:** A research physician recorded the pulse rates and the temperatures of water submerging the faces of ten small children in cold water to control the abnormally rapid heartbeats. The results are presented in the following table. Calculate the correlation coefficient and regression coefficients between temperature of water and reduction in pulse rate.

| Temperature of water | 68 | 65 | 70 | 62 | 60 | 55 | 58 | 65 | 69 | 63 |
|---|---|---|---|---|---|---|---|---|---|---|
| Reduction in pulse rate. | 2 | 5 | 1 | 10 | 9 | 13 | 10 | 3 | 4 | 6 |

Also show that (i) $\left(\dfrac{b_{yx}+b_{xy}}{2}\right) \geq r_{xy}$

Solution: Calculating table of correlation coefficient and regression coefficients.

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_iy_i$ |
|---|---|---|---|---|
| 68 | 2 | 4624 | 4 | 136 |
| 65 | 5 | 4225 | 25 | 325 |
| 70 | 1 | 4900 | 1 | 70 |
| 62 | 10 | 3844 | 100 | 620 |
| 60 | 9 | 3600 | 81 | 540 |
| 55 | 13 | 3025 | 169 | 715 |
| 58 | 10 | 3364 | 100 | 580 |
| 65 | 3 | 4225 | 9 | 195 |
| 69 | 4 | 4761 | 16 | 276 |
| 63 | 6 | 3969 | 36 | 378 |
| $\sum x_i = 635$ | $\sum y_i = 63$ | $\sum x_i^2 = 40537$ | $\sum y_i^2 = 541$ | $\sum x_iy_i = 3835$ |

We know, $r_{xy} = \dfrac{\sum x_iy_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$

$$= \frac{3835 - \frac{635 \times 63}{10}}{\sqrt{\left\{40537 - \frac{(635)^2}{10}\right\}\left\{541 - \frac{(63)^2}{10}\right\}}}$$

$= -0.94$

We know, the regression coefficient of y on x is, $b_{yx} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{3835 - \frac{635 \times 63}{10}}{40537 - \frac{(635)^2}{10}} = \ = \frac{-1655}{2145} = \underline{-0.77}$$

Again, the regression coefficient of x on y is, $b_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}} = \frac{3835 - \frac{635 \times 63}{10}}{541 - \frac{(63)^2}{10}} = \frac{-1655}{1441} = \underline{-1.1}$$

(i) $\left(\dfrac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

Here, $\left(\dfrac{b_{yx} + b_{xy}}{2}\right) = \dfrac{(-0.77) + (-1.1)}{2} = \ -0.94 = r_{xy}$

**Assignment Problem-1**: The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Test Scores: y | 14 | 19 | 24 | 21 | 26 | 22 | 15 | 20 | 19 |
| Sales(in lakh Taka) | 31 | 36 | 48 | 37 | 50 | 45 | 33 | 41 | 39 |

(a) Find the regression equation of test scores on sales.

Ans: $\hat{y}$ = -2.4 + 0.56x

(b) Find the test scores when the sale is Tk. 40 lakh.

Ans: 20 lakh

(c) Find the regression equation of sales on test scores.

Ans: $\hat{x}$ = 7.8 + 1.61y

(d) Predict the value of sale if the test score is 30

Ans: 56.1 lakh

(e) Compute the value of correlation coefficient with the help of regression coefficients.

**Assignment Problem-2:** The following table gives the ages and blood pressure of 10 women:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age in years x | 56 | 42 | 36 | 47 | 49 | 42 | 72 | 63 | 55 | 60 |
| Blood pressure y | 147 | 125 | 118 | 128 | 125 | 140 | 155 | 160 | 149 | 150 |

(i) Obtain the regression line of y on x. Ans: $\hat{y}$ = 83.76+ 1.11x

(ii) Estimate the blood pressure of a women whose age is 50 years. Ans: 139.26

(iii) Obtain the regression line of x on y.

(iv) Find correlation coefficient between x and y and comment.

**Assignment Problem-3**: Consider the following data set on two variables x and y:

x : 1 2 3 4 5 6

y : 6 4 3 5 4 2

(a) Find the equation of the regression line y on x. Ans: $\hat{y}$ = 5.799-0.541x

(b) Graph the line on a scatter diagram.

(c) Estimate the value of y when x = 4.5 Ans: $\hat{y}$ = 3.486

(d) Predict the value of y when x = 8. Ans: $\hat{y}$ =1.687

**Assignment Problem-4:** Cost accountants often estimate overhead based on production. At the standard knitting company, they have collected information on overhead expenses and units produced at different plants and what to estimate a regression equation to predict future overhead.

| Units | 56 | 40 | 48 | 30 | 41 | 42 | 55 | 35 |
|---|---|---|---|---|---|---|---|---|
| Overhead | 282 | 173 | 233 | 116 | 191 | 171 | 274 | 152 |

(i)Draw a scatter diagram and comment

(ii)Fit a regression equation.

(iii)Estimate overhead when 65 units are produced.

**Assignment Problem-5:** The following data refer to information about annual sales

( Tk.'000) and year of experience of a super store of 8 salesmen:

| Salesmen | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Annual sales (Tk.'000) | 90 | 75 | 78 | 86 | 95 | 110 | 130 | 145 |
| Year of experience | 7 | 4 | 5 | 6 | 11 | 12 | 13 | 17 |

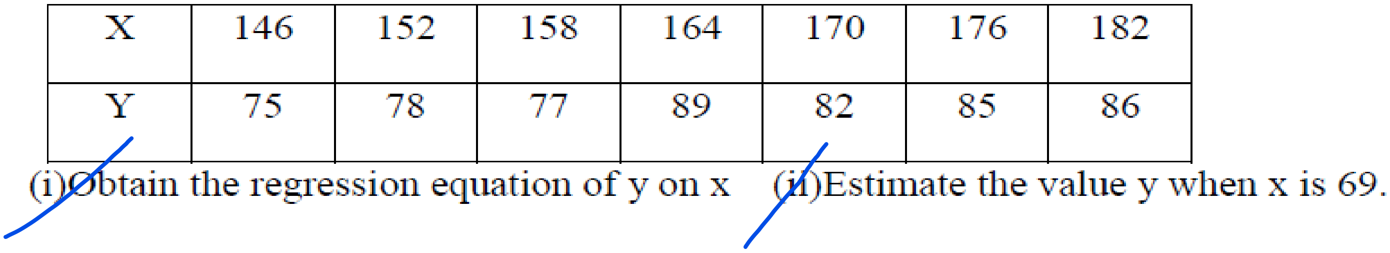
(i)Fit two regression lines.

(ii)Estimate sales for year of experience is 10

(iii)Estimate year of experience for sales 100000

**Assignment Problem (1-5): Same as solution-1**

**Assignment problem: 06**



| X | 146 | 152 | 158 | 164 | 170 | 176 | 182 |
|---|---|---|---|---|---|---|---|
| Y | 75 | 78 | 77 | 89 | 82 | 85 | 86 |

(i)Obtain the regression equation of y on x (ii)Estimate the value y when x is 69.