3 Linear Regression

Let's Study

- Meaning and Types of Regression
- Fitting Simple Linear Regression
  - Least Square Method
  - Regression of Y on X
  - Regression of X on Y
- Properties of Regression Coefficients

Let's Recall

- Concept of Correlation
- Coefficient of Correlation
- Interpretation of Correlation

Introduction

We have already learns that correlation is used to measure the strength and direction of association between two variables. In statistics, correlation denotes association between two quantitative variables. It is assumed that this association is linear. That is, one variable increase or decreases by a fixed amount for every unit of increase or decrease in the other variable. Consider the relationship between the two variables in each of the following examples.

1. Advertising and sales of a product. (Positive correlation)
2. Height and weight of a primary school student. (Positive correlation)
3. The amount of fertilizer and the amount of crop yield. (Positive correlation)
4. Duration of exercise and weight loss. (Positive correlation)
5. Demand and price of a commodity. (Positive correlation)
6. Income and consumption. (Positive correlation)
7. Supply and price of a commodity. (Negative correlation)
8. Number of days of absence (in school) and performance in examination. (Negative correlation)
9. The more vitamins one consumes, the less likely one is to have a deficiency.(Negative correlation)

Correlation coefficient measures association between two variables but cannot determine the value of one variable when the value of the other variable is known or given. The technique used for predicting the value of one variable for a given value of the other variable is called regression. Regression is a statistical tool for investigating the relationship between variables. It is frequently used to predict the value and to identify factors that cause an outcome. Karl Pearson defined the coefficient of correlation known as *Pearson's Product Moment correlation coefficient.* Carl Friedrich Gauss developed the method known as the *Least Squares Method* for finding the linear equation that best describes the relationship between two or more variables. R.A. Fisher combined the work of Gauss and Pearson to develop the complete theory of least squares estimation in linear regression. Due to Fisher's work, linear regression is used for prediction and understanding correlations.

Note: Some statistical methods attempt to determine the value of an unknown quantity, which may be a parameter or a random variable. The method used for this purpose is called estimation if the unknown quantity is a parameter, and prediction if the unknown quantity is a variable.

Let's :Learn

3.1 Meaning and Types of Regression

Meaning of Regression

Linear regression is a method of predicting the value of one variable when the values of all other variables are known or specified. The variable being predicted is called the response or dependent variable. The variables used for predicting the response or dependent variable are called predictors or independent variables. Linear regression proposes that the relationship between two or more variables is described by a linear equation. The linear equation used for this purpose is called a linear regression model. A linear regression model consists of a linear equation with unknown coefficients. The unknown coefficients in the linear regression model are called parameters of the linear regression model. Observed values of the variables are used to estimate the unknown parameters of the model. The process of developing a linear equation to represent the relationship between two or more variables using the available sample data is known as fitting the linear regression model to observed data. Correlation analysis is used for measuring the strength or degree of the relationship between the predictors or independent variables and the response or dependent variable. The sign of correlation coefficient indicates the direction (positive or negative) of the relationship between the variables, while the absolute value(that is, magnitude) of correlation coefficient is used as a measure of the strength of the relationship. Correlation analysis, however,does not go beyond measuring the direction and strength of the relationship between predictor or independent variables and the response or dependent variable. The linear regression model goes beyond correlation analysis and develops a formula for predicting the value of the response or dependent variable when the values of the predictor or independent variables are known. Correlation analysis is therefore a part of regression analysis and is performed before performing regression analysis. The purpose of correlation analysis is to find whether there is a strong correlation between two variables. Linear regression will be useful for prediction only if there is strong correlation between the two variables.

Types of Linear Regression

The primary objective of a linear regression is to develop a linear equation to express or represent the relationship between two or more variables. Regression equation is the mathematical equation that provides prediction of values of the dependent variable based on the known or given values of the independent variables.

When the linear regression model represents the relationship between the dependent variable (Y) and only one independent variable (X), then the corresponding regression model is called a simple linear regression model. When the linear regression model represents the relationship between the dependent variable and two or more independent variables, then the corresponding regression model is called a multiple linear regression model.

Following examples illustrate situations for simple linear regression.

1. A firm may be interested in knowing the relationship between advertising (X) and sales of its product (Y), so that it can predict the amount of sales for the allocated advertising budget.
2. A botanist wants to find the relationship between the ages (X) and heights (Y) of seedling in his experiment.
3. A physician wants to find the relationship between the time since a drug is administered (X) and the concentration of the drug in the blood-stream (Y).

Following examples illustrate situations for multiple linear regression

1. The amount of sales of a product (dependent variable) is associated with

several independent variables such as price of the product, amount of expenditure on its advertisement, quality of the product, and the number of competitors.

2. Annual savings of a family (dependent variable) are associated with several independent variables such as the annual income, family size, health conditions of family members, and number of children in school or college.

3. The blood pressure of a person (dependent variable) is associated with several independent variables such as his or her age, weight, the level of blood cholesterol, and the level of blood sugar.

The linear regression model assumes that the value of the dependent variable changes in direct proportion to changes in the value of an independent variable, regardless of values of other independent variables. Linear regression is the simplest form of regression and there are more general and complicated regression models. We shall restrict our attention only to linear regression model in this chapter.

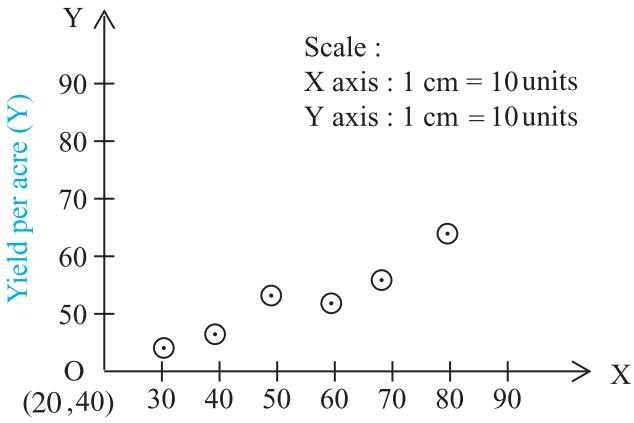
3.2 Fitting Simple Linear Regression

Consider an example where we wish to predict the amount of crop yield (in kg. per acre) as a linear function of the amount of fertilizer applied (in kg. per acre). In this example, the crop yield is to be predicted. Therefore, It is dependent variable and is denoted by Y. The amount of fertilizer applied is the variable used for the purpose of making the prediction. Therefore, it is the independent variable and is denoted by X.

Amount of fertilizer (X) (Kgs. In per acre)	Yield (Y) (in '00 kg)
30	43
40	45
50	54
60	53
70	56
80	63

Table : 3.1

Table 3.1 shows the amount of fertilizer and the crop yield for six cases. These pairs of observations are used to obtain the scatter diagram as shown in Fig. 2.1



Fertilizer (X)

Fig. 3.1: Scatter digram of the yield of grain and amount of fertilizer used.

We want to draw a straight line that is closest to the points in the scatter diagram (Fig. 3.1). If all the points were collinear (that is, on a straight line), there would have been no problem in drawing such a line. There is a problem because all the points are not on a straight line.

Since the points in the scatter diagram do not form a straight line, we want to draw a straight line that is closest to these points. Theoretically, the number of possible line is unlimited. It is therefore necessary to specify some condition in order to ensure that we draw the straight line that is closest to all the data points in the scatter diagram. The method of least squares provide the line of best fit because it is closest to the data points in the scatter diagram according to the least squares principle.

3.2.1 Method of Least Squares

The principle used in obtaining *the line of best fit* is called the **method of least squares**. The method of least squares was developed by Adrien-Maire Lagendre and Carl Friedrich Gauss independently of each other. Let us understand the central idea behind the principle of least squares.

Suppose the data consists of n pairs of values (x_1, y_1), (x_2, y_2), …, (x_n, y_n) and suppose that the line that fits best to the given data is written as follows. $\boxed{\hat{Y} = a + bX}$ (Here, $\hat{Y}$ is to be read as *Y cap*.)This equation is called the prediction equation. That is using the same values of constants *a* and *b*, the predicted value of *Y* are given by $\hat{Y}_i = a + bx_i$, where x_i is the value of the independent variable and $\hat{Y}_i$ is the corresponding predicted value of *Y*. Note that the observed value y_i of the independent variable *Y* is different from the predicted value $\hat{Y}_i$.

The observed valued (y_i) and predicted values $\left(\hat{Y}_i\right)$ of Y do not match perfectly because the observations do not fall on a straight line. The Difference between the observed values and the predicted values are called errors or residuals. Mathematically speaking the quantities

$y_1 - \hat{Y}_1$, $y_2 - \hat{Y}_2$ -------- $y_n - \hat{Y}_n$ or equivalently, the quantities $y_1 - (a + bx_1)$, $y_2 - (a + bx_2)$, ..., $y_n - (a + bx_n)$ are deviations of observed values of *Y* from the corresponding predicted values and are therefore called errors or residuals. We write $e_i = y_i - \hat{Y}_1 = y_i - (a + bx_i)$, for $i = 1, 2, ..., n$.

Geometrically, the residual e_i, which is given by $y_i - (a + bx_i)$, denotes the vertical distance (which may be positive or negative) between the observed value (y_i) and the predicted value $\left(\hat{Y}_i\right)$.

The principle of the method of least squares can be stated as follows.

Among all the possible straight lines that can be drawn on the scatter diagram, the line of best fit is defined as the line that minimizes the sum of squares of residuals, that is, the sum of squares of deviations of the predicted *y*-values from the observed *y*-values. In other words, the line of the best fit is obtained by determining the constant *a* and *b* so that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

is minimum.

The straight line obtained using this principle is called *the least regression line*.

Symbolically, we write

$$S^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

as the sum of squared errors. It can also be written as

$$S^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

We want to determine the constants *a* and *b* in such a way that S^2 is minimum.

Note that S^2 is a continuous and differentiable function of both *a* and *b*. We differentiate S^2 with respect to *a* (assuming *b* to be constant) and with respect to *b* (assuming *a* constant) and with respect to b (assuming a to be constant). We then equate both these derivatives to zero in order to minimize S^2. As the result, we get the following two linear equations in two unknowns *a* and *b*.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

When we solve these two linear equations, the values of *a* and *b* that minimize S^2 are given by

$$a = \bar{y} - b\bar{x} \text{ , } b = \frac{\text{cov}(X, Y)}{\sigma_x^2} \text{ ,}$$

where

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \text{ ,}$$

and

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Substituting the values of a and b obtained as indicated above in the regression equation

$Y = a + bX$,

we get the equation

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

Note: The constant b is called the regression coefficient (or the slope of the regression line) and the constant a is called the Y-intercept (that is, the Y value when $X = 0$). Recall that the equation $Y = a + bX$ defining a straight line is called the slope intercept formula of the straight line.

When observations on two variables, X and Y, are available, it is possible to fit a linear regression of Y on X as well as a linear regression of X on Y. Therefore, we consider both the models in order to understand the difference between the two and also the relationship between the two.

3.2.2 Regression of *Y* on *X*.

We now introduce notation b_{YX} for b when Y is the dependent variable and X is the independent variable.

Linear regression of Y on X assumes that the variable X is the independent variable and the variable Y is the dependent variable. In order to make this explicit, we express the linear regression model as follows.

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

or $$Y = b_{YX}X.$$

Here, note that b is replaced by b_{YX}.

$$b_{YX} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

$$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2}$$

2.2.3 Regression of *X* on *Y*

The notation b_{XY} stands for b when X is the dependent variable and Y is the independent variable.

Linear regression of X on Y assumes that the variable Y is the independent variable and the variable X is the dependent variable. In order to make it clear that this model is different from the linear regression of Y on X, we express the linear regression model as follows.

$$X = a' + b'Y.$$

The method of least squares, when applied to this model leads to the following expressions for the constant a' and b'.

$$a' = \bar{x} - b'\bar{y},$$

and

$$b' = \frac{\text{cov}(X,Y)}{\text{var}(Y)}$$

Substituting the values of a' and b' in the linear regression model, we get

$$X = a' + b'Y$$

$$X = \bar{x} - b'\bar{y} + b'Y$$

$$\text{i.e. } (X - \bar{x}) = b'(Y - \bar{y}).$$

Note: The constant b′ in the above equation is called the regression coefficient of X on Y. In order to make this explicit, it will henceforth be written as b_{XY} instead of b′. The least squares regression of X on Y will therefore be written as

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

Note that the linear regression of X on Y is expressed as

$$X = a' + b_{XY}Y.$$

Here note that b is replaced by b_{XY}. This can be written as

$$Y = \frac{1}{b_{XY}}(X - a')$$

Showing that the constant $\left(\frac{1}{b_{XY}}\right)$ is the slope of the line of regression of X on Y.

Further, note that the regression coefficient b_{XY} involved in the linear regression of X on Y is given by

$$b_{XY} = \frac{\text{cov}(X,Y)}{\text{var}(Y)}$$

$$= \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum(y_i - \bar{y})^2}$$

$$= \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum y_i^2 - n\bar{y}^2}$$

Also,

$$b_{XY} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n(\bar{x})^2}$$

Observed that the point $(\bar{x}, \bar{y})$ satisfies equation of both the lines of regression. Therefore, the point $(\bar{x}, \bar{y})$ is the point of intersection of the two lines regression.

SOLVED EXAMPLES

Ex. 1 : For the data on fertilizer application and yield of grain is given in the table 3.2

i) Obtain the line of regression of yield of grain on the amount of fertilizer used.

ii) Draw the least squares regression line.

iii) Estimate the yield of grain when 90kgs. of fertilizer is applied.

Solution :

Amount of fertilizer $X = x_i$	Yield $Y = y_i$	x_i	$x_i y_i$
30	43	900	1290
40	45	1600	1800
50	54	2500	2700
60	53	3600	3180
70	56	4900	3920
80	63	6400	5040
330	314	19900	17930

Table : 3.2

i) Since $n = 6$, $\sum_{i=1}^{6} x_i = 330$, $\sum_{i=1}^{6} y_i = 314$.

$\sum_{i=1}^{6} x_i^2 = 19900$ and $\sum_{i=1}^{6} x_i y_i = 17930$

$\bar{x} = \frac{\sum x_i}{n} = 55$, $\bar{y} = \frac{\sum y_i}{n} = 52.3$

$$b_{XY} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$= \frac{17930 - 6 \times 55 \times 52.3}{19900 - 6 \times 3025}$$

$$= \frac{671}{1750} = 0.38$$

and $a = \bar{y} - b_{XY} \cdot \bar{x} = 31.4$.

Finally, the line of regression of Y on X is given by

$$Y = 31.4 + 0.38\,X$$

Predicted yield; Fixed yield; change in the yield per one kg. increase in fertilizer; kgs. of fertilizer used

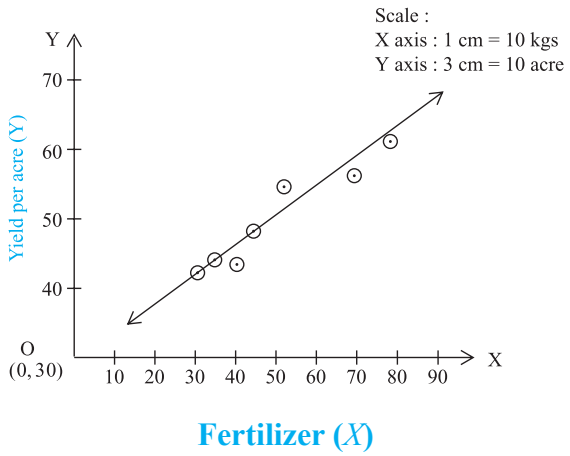


Fig. 3.2 : Scatter diagram and the line of regression of yield on amount of fertilizer

ii) To draw the least squares regression line, we pick any two convenient values of X and find the corresponding values of Y.

For $x = 35$. $y = 44.7$

$x = 45$, $y = 48.5$

Joining the two points (35,44.8) and (45,48.6), we get the line in Fig 3.2

iii) Putting $x = 90$ in the regression equation

$$\hat{Y} = 31.4 + 0.38 \times 90 = 65.6$$

Ex. 2: A departmental store gives in service training to the salesmen followed by a test. It is experienced that the performance regarding sales of any salesmen is linearly related to the scores secured by him. The following data give test scores and sales made by nine salesmen during fixed period.

Test scores (X)	16	22	28	24	29	25	16	23	24
Sales ('00Rs.) (Y)	35	42	57	40	54	51	34	47	45

i) Obtain the line of regression of Y on X.

ii) Estimate Y when $X = 17$.

Solution : To show the calculations clearly, it is better to prepare the following table

$X=x_i$	$Y=y_i$	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
16	35	−7	−10	49	70
22	42	−1	−3	1	3
28	57	5	12	25	60
24	40	1	−5	1	−5
29	54	6	9	36	54
25	51	2	6	4	12
16	34	−7	−11	49	77
23	47	0	2	00	00
24	45	1	0	1	00
207	405	00	00	166	271

i) $n = 9$ and $\sum_{i=1}^{9} x_i = 207$, $\sum_{i=1}^{9} y_i = 405$.

$$\bar{x} = \frac{\sum x_i}{n} = 23, \quad \bar{y} = \frac{\sum y_i}{n} = 45$$

Since the means of X and Y are whole numbers, it is preferable to use the formula

$$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$ for the calculation of b_{YX}.

Line of regression of Y on X is

$$Y = a + b_{XY}X$$

where $$b_{XY} = \frac{\text{cov}(X,Y)}{\sigma_Y^2}$$

$$= \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}}{\frac{\sum(x_i-\bar{x})^2}{n}}$$

$$= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$= \frac{271}{166}$$

$$= 1.6325$$

and $$a = \bar{y} - b_{XY}\bar{x}$$

$$= 7.4525$$

Here, the line of regression of Y on X is

$$Y = 7.4525 + 1.6325X$$

ii) Estimate of Y when $X = 17$ is

$$Y = 7.4525 + (1.6325)(17)$$

$$= 35.205$$

Ex 3: The management of a large furniture store would like to determine sales (in thousands of Rs.) (X) on a given day on the basis of number of people (Y) that visited the store on that day. The necessary records were kept, and a random sample of ten days was selected for the study. The summary results were as follows:

$\sum x_i = 370,\ \sum y_i = 580,\ \sum x_i^2 = 17200,$

$\sum y_i^2 = 41640,\ \sum x_i y_i = 11500,\ n = 10$

Obtain the line of regression of X on Y.

Solution:

Line of regression of X on Y is

$$X = a' + b_{XY} Y$$

where $$b_{XY} = \frac{\text{cov}(X,Y)}{\sigma_Y^2}$$

$$= \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum y_i^2}{n} - (\bar{y})^2}$$

$$= \frac{\left(\frac{11500}{10}\right) - \left(\frac{370}{10}\right)\left(\frac{580}{10}\right)}{\left(\frac{41640}{10}\right) - \left(\frac{580}{10}\right)^2}$$

$$= \frac{1150 - 37 \times 58}{4164 - (58)^2}$$

$$= -\frac{996}{800}$$

$$= -1.245$$

and $$a' = \bar{x} - b_{XY}\,\bar{y}$$

$$= 37 - (-1.245)\,(58)$$

$$= 109.21$$

$\therefore$ Line of regression of X on Y is

$$X = 109.21 - 1.245Y$$

EXERCISE 3.1

1. The HRD manager of a company wants to find a measure which he can use to fix the monthly income of persons applying for the job in the production department. As an experimental project, he collected data of 7 persons from that department referring to years of service and their monthly incomes.

Years of service (X)	11	7	9	5	8	6	10
Monthly Income (Rs.1000's) (Y)	10	8	6	5	9	7	11

   (i) Find the regression equation of income on years of service.

   (ii) What initial start would you recommend for a person applying for the job after having served in similar capacity in another company for 13 years?

2. Calculate the regression equations of X on Y and Y on X from the following data:

X	10	12	13	17	18
Y	5	6	7	9	13

3. For a certain bivariate data on 5 pairs of observations given

   $\sum x = 20,\ \sum y = 20,\ \sum x^2 = 90,$

   $\sum y^2 = 90,\ \sum xy = 76$

   Calculate (i) cov(x,y) (ii) b_{yx} and b_{xy}, (iii) r

4. From the following data estimate y when $x = 125$

X	120	115	120	125	126	123
Y	13	15	14	13	12	14

5. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random.

Aptitude score (X)	60	62	65	70	72	48	53	73	65	82
Productivity Index (Y)	68	60	62	80	85	40	52	62	60	81

Obtain the two regression equations and estimate:

(i) The productivity index of a worker whose test score is 95.

(ii) The test score when productivity index is 75.

6. Compute the appropriate regression equation for the following data:

X [Independent Veriable]	2	4	5	6	8	11
Y [Dependent Veriable]	18	12	10	8	7	5

7. The following are the marks obtained by the students in Economics (X) and Mathematics (Y)

X	59	60	61	62	63
Y	78	82	82	79	81

Find the regression equation of Y on X.

8. For the following bivariate data obtain the equations of two regression lines:

X	1	2	3	4	5
Y	5	7	9	11	13

9. From the following data obtain the equation of two regression lines:

X	6	2	10	4	8
Y	9	11	5	8	7

10. For the following data, find the regression line of Y on X

X	1	2	3
Y	2	1	6

Hence find the most likely value of y when $x = 4$.

11. From the following data, find the regression equation of Y on X and estimate Y when $X = 10$.

X	1	2	3	4	5	6
Y	2	4	7	6	5	6

12. The following sample gives the number of hours of study (X) per day for an examination and marks (Y) obtained by 12 students.

X	3	3	3	4	4	5	5	5	6	6	7	8
Y	45	60	55	60	75	70	80	75	90	80	75	85

Obtain the line of regression of marks on hours of study.

3.3 Properties of Regression Coefficients

The line of regression of Y on X is given by $Y = a + b_{YX}X$ and the line of regression of X on Y is given by $X = a' + b_{XY}Y$.

Here, b_{YX} $\dfrac{\text{cov}(X,Y)}{\text{var}(X)}$, the slope of the line of regression of Y on X is called the regression coefficient of Y on X. Simillary,

$b_{XY} = \dfrac{\text{cov}(X,Y)}{\text{var}(Y)}$, the slope of the line of regression of X on Y is called the regression coefficient of X on Y. These two regression coefficients have the following property.

(a) $b_{XY} \cdot b_{YX} = r^2$

where r is the correlation coefficient between X and Y,

b_{XY} is the regression coefficient of X on Y. and b_{YX} is the regression coefficient of Y on X.

Proof: Note that

$$b_{XY} \cdot b_{YX} = \frac{\text{cov}(X,Y)}{\text{var}(Y)} \cdot \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

$$= \left[\frac{\text{cov}(X,Y)}{\sigma_X . \sigma_Y}\right]^2$$

$$= r^2.$$

Can it be said that the correlation coefficient is the square root of the product of the two regression coefficients?

(b) If $b_{YX} > 1$, then $b_{XY} < 1$.

Proof: Let, if possible, $b_{XY} > 1$ and $b_{YX} > 1$.

Then, using the above result, $b_{XY} \cdot b_{YX} > 1$, implies that $r^2 > 1$, which is impossible. (Can you provide the reason?)

This shows that our assumption must be invalid. That is, both the regression coefficients cannot simultaneously exceed unity.

We already know that the two variances σ_x^2, σ_y^2, and the correlation coefficient r satisfy the relation.

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

$$\therefore \quad r = \frac{\text{cov}(X,Y)}{\sigma_X . \sigma_Y}$$

The regression coefficients can also be written as follows.

$$b_{YX} = \frac{\text{cov}(X.Y)}{\sigma_X^2} = \frac{r.\sigma_X.\sigma_Y}{\sigma_x^2} = r \cdot \frac{\sigma_Y}{\sigma_X}$$

and

$$b_{XY} = \frac{\text{cov}(X.Y)}{\sigma_y^2} = \frac{r.\sigma_X.\sigma_Y}{\sigma_y^2} = r \cdot \frac{\sigma_X}{\sigma_Y}$$

(c) $\left|\frac{b_{YX} + b_{XY}}{2}\right| \geq |r|$

Proof: We have already seen that

$$b_{YX} = r \frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{XY} = r \frac{\sigma_X}{\sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.
Therefore,

$$b_{YX} + b_{XY} = r \frac{\sigma_Y}{\sigma_X} + r \frac{\sigma_X}{\sigma_Y}$$

$$= \left[\frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y}\right]$$

$$= r \left[\frac{\sigma_Y + \sigma_X}{\sigma_X . \sigma_Y}\right] \qquad (1)$$

But $(\sigma_X - \sigma_Y)^2 > 0$ and therefore

$$\sigma^2_X - \sigma^2_Y - 2\sigma_X\sigma_Y \geq 0$$

$$\sigma^2_X + \sigma^2_Y \geq 2\sigma_X\sigma_Y$$

$$\frac{\sigma_y^2 + \sigma_x^2}{\sigma_X . \sigma_Y} \geq 2$$

$$\therefore \quad r \frac{\sigma_y^2 + \sigma_x^2}{\sigma_X . \sigma_Y} \geq 2r. \qquad (2)$$

From (1) and (2), we have

$$b_{YX} + b_{XY} \geq 2r.$$

$$\therefore \quad \frac{b_{YX} + b_{XY}}{2} \geq r.$$

this result shows that the arithmetic mean of the two regression coefficients, namely b_{YX} and b_{XY} is greater than or equal to r. This result, however, holds only when b_{YX}, b_{XY} and r are positive. (Can you find the reason?)

Consider the case where $b_{YX} = -0.8$ and $b_{XY} = -0.45$. In this case, we have $r = -0.6$. (Can you find the reason?)

Note that $b_{YX} + b_{XY} = -1.25$, and $2r = -1.2$. This shows that $b_{YX} + b_{XY} \leq 2r$.

It may be interesting to note that

$$b_{YX} = \frac{\text{cov}(X.Y)}{\sigma_X^2}$$

$$b_{XY} = \frac{\text{cov}(X.Y)}{\sigma_y^2}$$

$$r = \frac{\text{cov}(X,Y)}{\sigma_X . \sigma_Y}$$

It is evident from the above three equations that all the coefficients have the same numerator and this numerator determines their sign. As the result, all these coefficients have the same sign. In other words, if $r > 0$, then $b_{YX} > 0$, and $b_{XY} > 0$. Similary. if $r < 0$. then $b_{YX} < 0$, and $b_{XY} < 0$. Finally. if $r = 0$, then $b_{YX} = b_{XY} = 0$.

(d) b_{YX} and b_{XY} are not affected by change of origin, but are affected by change of scale. This property is known as *invariance* property.

The invariance property states that b_{YX} and b_{XY} are invariant under change of origin, but are not invariant under change of scale.

Proof: Let $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$,

where a, b, h and k are constants with the condition that $h, k \neq 0$

We have already proved that $\sigma_X^2 = h^2\sigma_U^2$, $\sigma_Y^2 = k^2\sigma_V^2$, and $\text{cov}(X, Y) = hk\text{cov}(U, V)$.

Therefore,

$$b_{YX} = \frac{\text{cov}(X.Y)}{\sigma_X^2}$$

$$= \frac{hk\,\text{cov}(U,V)}{h^2\sigma_U^2}$$

$$= \frac{k}{h}\,\frac{\text{cov}(U,V)}{\sigma_U^2}$$

that is,

$$\boxed{b_{YX} = \frac{k}{h} b_{VU}}$$

Similatly,

$$\boxed{b_{XY} = \frac{k}{h} b_{UV}}$$

These two results show that regression coefficients are invariant under change of origin, but are not invariant under change of scale.

SOLVED EXAMPLES

Ex. 1 : The table below gives the heights of fathers (X) and heights of their sons (Y) respectively.

Heights of fathers (inches) (X)	64	62	66	63	67	61	69	65	67	66
Heights of sons (inches) (Y)	67	65	67	64	68	65	67	64	70	66

(i) Find the regression line of Y on X.

(ii) Find the regression line of X on Y.

(iii) Predict son's height if father's height is 68 inches.

(iv) Predict father's height if son's height is 59 inches.

Solution :

Let us use the change of origin for computations of regression coefficients.

Let $u_i = x_i - 65$ and $v_i = y_i - 67$

x_i	y_i	u_i	v_i	u_i^2	v_i^2	u_iv_i
64	67	−1	0	1	0	0
62	65	−3	−2	9	4	6
66	67	1	0	1	0	0
63	64	−2	−3	4	9	6
67	68	2	1	4	1	2
61	65	−4	−2	16	4	8
69	67	4	0	16	0	0
65	64	0	−3	0	9	0
67	70	2	3	4	9	6
66	66	1	−1	1	1	−1
Total		0	−7	56	37	27

Here, $n = 10$, $\sum u_i = 0$, $\sum v_i = -7$,

$\sum u_i^2 = 56$, $\sum v_i^2 = 37$, and $\sum u_iv_i = 27$

$$\bar{u} = \frac{\sum u_i}{n} = \frac{0}{10} = 0,$$

$$\sigma_u^2 = \frac{\sum u_i^2}{n} - \left(\bar{u}\right)^2$$

$$= \frac{56}{10} - 0^2$$

$$\sigma_u^2 = 5.6$$

$$\bar{v} = \frac{\sum v_i}{n}, \qquad \sigma_v^2 = \frac{\sum v_i^2}{n} - \left(\bar{v}\right)^2$$

$$= \frac{-7}{10} \qquad = \frac{37}{10} - (-0.7)^2$$

$$= -0.7 \qquad = 3.21$$

$$\text{cov}\,(u, v) = \frac{\sum u_i v_i}{n} - \bar{u}\ \bar{v}$$

$$= \frac{27}{10} - 0 \times -0.7$$

$$= 2.7$$

Now, you know that, regression coefficients are independent of change of origin.

$$\therefore \quad b_{XY} = b_{UV} = \frac{\text{cov}(u,v)}{\sigma_V} = \frac{2.7}{3.21} = 0.84$$

$$\text{and } b_{YX} = b_{VU} = \frac{\text{cov}(u,v)}{\sigma_U} = \frac{2.7}{5.6} = 0.48$$

You are also aware that mean is affected by change of origin.

$$\therefore \quad \bar{x} = \bar{u} + 65 = 0 + 65 = 65$$

$$\text{and} \quad \bar{y} = \bar{v} + 67 = -0.7 + 67 = 66.3$$

(i) Line of regression of Y on X is

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

$$\therefore \ (Y - 66.3) = 0.48\ (X - 65)$$

$$\therefore \qquad Y = 0.48.\ X + 35.1$$

ii) Regression line of X on Y is

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

$$\therefore \ (X - 65) = 0.84\ (Y - 66.3)$$

$$\therefore \qquad X = 0.84.\ Y + 9.31$$

iii) Estimate of sons height Y for $X = 68$

$$Y = 0.48 \times 68 + 35.1$$

$$= 67.74 \text{ inches}$$

iv) Estimate of fathers height X for $Y = 59$

$$X = 0.84 \times 59 + 9.31$$

$$= 58.87 \text{ inches}$$

v) Correlation coefficient

$$r = \sqrt{b_{YX}.b_{XY}}$$

$$= \sqrt{0.84 \times 0.48}$$

$$= 0.635$$

We choose positive square root ! (why?)

Ex. 2: Compute regression coefficient from the following data on the variable weight (X) and height (Y) of 8 individuals :

$$n = 8, \ \sum (x_i - 45) = 48$$

$$\sum (x_i - 45)^2 = 4400,$$

$$\sum (y_i - 150) = 280,$$

$$\sum (y_i - 150)^2 = 167432,$$

$$\sum (x_i - 45).(y_i - 150) = 21680$$

Solution: Let $u_i = x_i - 45$ and $v_i = y_i - 150$

So $\sum u_i = 48, \quad \sum u_i^2 = 4400,$

$\sum v_i = 280, \ \sum v_i^2 = 167432,$

$\sum u_i v_i = 21680$

$$\therefore \quad \bar{u} = \frac{\sum u_i}{n} = \frac{48}{8} = 6$$

$$\bar{v} = \frac{\sum v_i}{n} = \frac{280}{8} = 35$$

$$\sigma_u^2 = \frac{\sum u_i^2}{n} - \left(\bar{u}\right)^2$$

$$= \frac{4400}{8} - (6)^2 = 514$$

$$\sigma_v^2 = \frac{\sum v_i^2}{n} - \left(\bar{v}\right)^2$$

$$= \frac{167432}{8} - (35)^2 = 19704$$

$$\text{cov}(u, v) = \frac{\sum u_i v_i}{n} - \bar{u}\ \bar{v}$$

$$= \frac{21680}{8} - (6)(35)$$

$$= 2500$$

From the properties of regression coefficients, you know they are independent of change of origin.

$$\therefore b_{YX} = b_{VU} = \frac{\text{cov}(u,v)}{\sigma_U} = \frac{2500}{514} = 4.86$$

$$\text{and } b_{XY} = b_{UV} = \frac{\text{cov}(u,v)}{\sigma_V} = \frac{2500}{19704} = 0.12$$

(Have you noticed $b_{YX} > 1$ and $b_{XY} < 1$?)

Ex. 3: The following results were obtained from records of age (X) and systolic blood pressure (Y) of a group of 10 women.

	X	Y
Mean	53	142
Variance	130	165

$$\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = 1170$$

Find the appropriate regression equation and use it to estimate the blood pressure of a woman with age 47 years.

Solution:

Here, we need to find line of regression of Y on X, which is given as

$$Y = a + b_{YX} X$$

where, $b_{YX} = \frac{\text{cov}(X,Y)}{\sigma_x^2}$

$$= \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n} \Big/ \sigma_x^2$$

$$= \frac{\left(\frac{1170}{10}\right)}{130}$$

$$= 0.9$$

and $a = \bar{y} - b_{YX}\ \bar{x}$

$$= 142 - (0.9) \times (53)$$

$$= 94.3$$

Therefore, regression equation of Y on X is

$$Y = 94.3 + 0.9\ X$$

Now, the estimate of blood pressure of women with age 47 years is

$$Y = 94.3 + 0.9 \times 47$$

$$= 136.6$$

Ex. 4: Given the following data, obtain the linear regression & estimate of X for $Y = 10$

$\bar{x} = 7.6$, $\bar{y} = 14.8$, $\sigma_X = 3.5$, $\sigma_Y = 28$ and $r = 0.8$

Solution:

Here, we need to obtain line of regression of X on Y which can be expressed as

$$X = a' - b_{YX} Y$$

where $b_{XY} = \frac{\text{cov}(X,Y)}{\sigma_Y^2}$

$$= r\,\frac{\sigma_X}{\sigma_Y}$$

$$= 0.8\,\frac{(3.5)}{(28)}$$

$$= 0.1$$

and $a' = \bar{x} - b_{YX}\ \bar{y}$

$= 7.6 - (0.1)\ (14.8)$

$= 6.12$

$\therefore$ Line of regression of X on Y is

$X = 6.12 + 0.1\ Y$

Estimate of X for $Y = 10$ is

$X = 6.12 + 0.1 \times 10$

$X = 7.12$

EXERCISE 3.2

1. For a bivariate data.

$\bar{x} = 53,\ \bar{y} = 28,\ b_{YX} = -1.2,\ b_{XY} = -0.3$

Find

i) Correlation coefficient between X and Y.

ii) Estimate of Y for $X = 50$

iii) Estimate of X for $Y = 25$

2. From the data of 20 pairs of observation on X and Y, following results are obtained.

$\bar{x} = 199,\ \bar{y} = 94,$

$\sum(x_i - \bar{x})^2 = 1200 \quad \sum(y_i - \bar{y})^2 = 300$

$\sum(x_i - \bar{x})(y_i - \bar{y}) = -250$

Find

i) The line of regression of Y on X.

ii) The line of regression of X on Y.

iii) Correlation coefficient between X and Y

3. From the data of 7 pairs of observations on X and Y, following results are obtained.

$\sum(x_i - 70) = -35 \quad \sum(y_i - 60) = -7$

$\sum(x_i - 70)^2 = 2989 \quad \sum(y_i - 60)^2 = 476$

$\sum(x_i - 70)(y_i - 60) = 1064$

[*Given* $\sqrt{0.7884} = 0.8879$]

Obtain

i) The line of regression of Y on X.

ii) The line of regression of X on Y

iii) The correlation coefficient between X and Y.

4. You are given the following information about advertising expenditure and sales.

	Advertisment expenditure (Rs.in lakh) (X)	Sales (Rs.in lakh) (Y)
Arithmetic mean	10	90
Standard deviation	3	12

Correlation coefficient between X and Y is 0.8

(i) Obtain the two regression equations.

(ii) What is the likely sales when the advertising budget is Rs 15 lakh?

(iii) What should be the advertising budget if the company wants to attain sales target of Rs.120 lakh?

5. Bring out inconsistency if any, in the following :

(i) $b_{YX} + b_{XY} = 1.30$ and $r = 0.75$

(ii) $b_{YX} = b_{XY} = 1.50$ and $r = -0.9$

(iii) $b_{YX} = 1.9$ and $b_{XY} = -0.25$

(iv) $b_{YX} = 2.6$ and $b_{XY} = \frac{1}{2.6}$

6. Two samples from bivariate populations have 15 observations each. The sample means of X and Y are 25 and 18 respectively. The corresponding sum of squares of deviations from respective means are 136 and 150. The sum of product of deviations from respective means is 123. Obtain the equation of line of regression of X on Y.

7. For a certain bivariate data

	X	Y
Mean	25	20
S.D.	4	3

And $r = 0.5$. estimate y when $x = 10$ and estimate x when $y = 16$

8. Given the following information about the production and demand of a commodity obtain the two regression lines :

	Production (X)	Demand (Y)
Mean	85	90
S.D.	5	6

Coefficient of correlation between X and Y is 0.6. Also estimate the production when demand is 100.

9. Given the following data, obtain linear regression estimate of X for $Y = 10$

$\bar{x} = 7.6$, $\bar{y} = 14.8$, $\sigma_X = 3.2$, $\sigma_Y = 16$ and $r = 0.7$

10. An inquiry of 50 families to study the relationship between expenditure on accommodation (Rs. x) and expenditure on food and entertainment (Rs. y) gave the following results. :

$\sum x = 8500$, $\sum y = 9600$, $\sigma_X = 60$, $\sigma_Y = 20$, $r = 0.6$

Estimate the expenditure on food and entertainment when expenditure on accommodation is Rs 200.

11. The following data about the sales and advertisement expenditure of a firms is given below (in Rs. Crores)

	Sales	Adv. Exp.
Mean	40	6
S.D.	10	1.5

i) Estimate the likely sales for a proposed advertisement expenditure of Rs.10 crores.

ii) What should be the advertisement expenditure if the firm proposes a sales target Rs.60 crores.

12. For a certain bivariate data the following information are available.

	X	Y
A.M.	13	17
S.D.	3	2

Correlation coefficient between x and y is 0.6. estimate x when $y = 15$ and estimate y when $x = 10$.

SOLVED EXAMPLES

Ex.1: The equations of the two lines of regression are $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$

(i) Find the means of X and Y.

(ii) Obtain correlation coefficient between X and Y.

(iii) Estimate Y for $X = 2$.

Solution:

(i) We know that the co-ordinates of the point of intersection of the two lines are $\bar{x}$ and $\bar{y}$, the means of X and Y.

The regression equations are

$$3x + 2y - 26 = 0$$

and $6x + y - 31 = 0$

Solving these equations simultaneously, we get

$$\begin{aligned} 6x + 4y - 52 &= 0 \\ 6x + y - 31 &= 0 \\ (-)\ \ (-)\ \ (+)& \end{aligned}$$

$$\begin{aligned} 3y - 21 &= 0 \\ \therefore\ 3y &= 21 \\ \text{i.e. } y &= 7 \\ \text{and } x &= 4 \end{aligned}$$

Hence the means of X and Y are $\bar{x} = 4$ and $\bar{y} = 7$

(ii) Now, to find correlation coefficient, we have to find the regression coefficients b_{YX} and b_{XY}

For this, we have to choose one of the lines as that of line of regression of Y on X and other the line of regression of X on Y

Let $3x + 2y - 26 = 0$ be the line of regression on Y on X this gives

$$Y = -\frac{3}{2} X + 13$$

The coefficient of X in this equation is $b_{YX} = -\frac{3}{2}$

Then the other equation is that of line of regression of X on Y which can be written as

$$X = -\frac{1}{6} Y + \frac{31}{6}$$

Here, the regression coefficient $b_{XY} = -\frac{1}{6}$.

Now, you know that

$$r^2 = b_{XY} \cdot b_{YX}$$
$$= 0.25$$
$$\therefore \quad r = \pm 0.5$$

The correlation coefficient has the sign as that of b_{YX} and b_{XY}

$$\therefore \quad r = -0.5$$

Note: We choose arbitrarily the lines as that of regression of Y on X or X on Y. if the product $b_{YX} \cdot b_{XY}$ is less than unity, our choice is correct. Fortunately, there are only two choices.

Ex 2.: The regression equation of Y on X is

$y = \frac{4}{3} x$ and the regression equation X on Y is

$$x = \frac{y}{3} + \frac{5}{3}$$

Find

(i) Correlation coefficient between X and Y.

(ii) σ_Y^2 if $\sigma_X^2 = 4$

Solution:

Here, the regression lines are specified.

So $b_{YX} = \frac{4}{3}$ and $b_{XY} = \frac{1}{3}$

(i) $\therefore r^2 = b_{YX} \cdot b_{XY}$

$$= \frac{4}{3} \cdot \frac{1}{3}$$
$$= \frac{4}{9}$$
$$\therefore \quad r = +\frac{2}{3} \quad (\text{why} + \frac{2}{3} \text{ only?})$$

(ii) You know that

$$b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X}$$
$$\therefore \quad \frac{4}{3} = \frac{2}{3} \cdot \frac{\sigma_Y}{2}$$
$$\therefore \quad \sigma_Y = 4$$
$$\therefore \quad \sigma_Y^2 = 16$$

EXERCISE 3.3

1. From the two regression equations find r, $\bar{x}$ and $\bar{y}$.

   $4y = 9x + 15$ and $25x = 4y + 17$

2. In a partially destroyed laboratory record of an analysis of regression data, the following data are legible:

   Variance of $X = 9$

   Regression equations:

   $8x - 10y + 66 = 0$

   and $40x - 18y = 214$.

Find on the basis of above information

(i) The mean values of X and Y.

(ii) Correlation coefficient between X and Y.

(iii) Standard deviation of Y.

3. For 50 students of a class, the regression equation of marks in statistics (X) on the marks in Accountancy (Y) is $3y - 5x + 180 = 0$. The mean marks in accountancy is 44 and the variance of marks in statistics is $\left(\frac{9}{16}\right)^{th}$ of the variance of marks in accountancy. Find the mean marks in statistics and the correlation coefficient between marks in two subjects.

4. For a bivariate data, the regression coefficient of Y on X is 0.4 and the regression coefficient of X on Y is 0.9. Find the value of variance of Y if variance of X is 9.

5. The equations of two regression lines are

$$2x + 3y - 6 = 0$$

and $\quad 2x + 2y - 12 = 0$

Find (i) Correlation coefficient

(ii) $\frac{\sigma_X}{\sigma_Y}$

6. For a bivariate data: $\bar{x} = 53$, $\bar{y} = 28$, $b_{YX} = -1.5$ and $b_{XY} = -0.2$. Estimate Y when $X = 50$.

7. The equations of two regression lines are $x - 4y = 5$ and $16y - x = 64$. Find means of X and Y. Also, find correlation coefficient between X and Y.

8. In a partially destroyed record, the following data are available variance of $X = 25$. Regression equation of Y on X is $5y - x = 22$ and Regression equation of X on Y is $64x - 45y = 22$ Find

(i) Mean values of X and Y

(ii) Standard deviation of Y

(iii) Coefficient of correlation between X and Y.

9. If the two regression lines for a bivariate data are $2x = y + 15$ (x on y) and $4y = 3x + 25$ (y on x), find

(i) $\bar{x}$, (ii) $\bar{y}$, (iii) b_{YX},

(iv) b_{XY}, (v) r $\left[Given \sqrt{0.375} = 0.61\right]$

10. The two regression equations are

$5x - 6y + 90 = 0$ and $15x - 8y - 130 = 0$.

Find $\bar{x}$, $\bar{y}$, r.

11. Two lines of regression are $10x + 3y - 62 = 0$ and $6x + 5y - 50 = 0$ Identify the regression of x on y. Hence find $\bar{x}$, $\bar{y}$ and r.

12. For certain X and Y series, which are correlated the two lines of regression are $10y = 3x + 170$ and $5x + 70 = 6y$. Find the correlation coefficient between them. Find the mean values of X and Y.

13. Regression equations of two series are

$2x - y - 15 = 0$ and $3x - 4y + 25 = 0$

Find $\bar{x}$, $\bar{y}$ and regression coefficients. Also find coefficients of correlation.

$\left[Given \sqrt{0.375} = 0.61\right]$

14. The two regression lines between height (X) in inches and weight (Y) in kgs of girls are,

$$4y - 15x + 500 = 0$$

and $\quad 20x - 3y - 900 = 0$

Find mean height and weight of the group. Also, estimate weight of a girl whose height is 70 inches.

Let's Remember

- Line of regression of Y on X is

$$Y = a + bX$$

where $b = b_{YX}$ = regression coefficient of Y on X

$$b_{YX} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

$$= \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}}{\frac{\sum(x_i-\bar{x})^2}{n}}$$

$$= \frac{\frac{\sum x_i y_i}{n} - \bar{x}.\bar{y}}{\frac{\sum x_i^2}{n} - (\bar{x})^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}.\bar{y}}{\sum x_i^2 - n(\bar{x})^2}$$

and $a = \bar{y} - b\,\bar{x}$

- Line of regression of X on Y is

$$X = a' + b'y$$

where $b' = b_{XY}$ = regression coefficient of Y on X

$$b_{XY} = \frac{\text{cov}(X,Y)}{\text{var}(Y)}$$

$$= \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}}{\frac{\sum(y_i-\bar{y})^2}{n}}$$

$$= \frac{\frac{\sum x_i y_i}{n} - \bar{x}.\bar{y}}{\frac{\sum y_i^2}{n} - (\bar{y})^2}$$

$$= \frac{\sum x_i y_i - n\bar{x}.\bar{y}}{\sum y_i^2 - n(\bar{y})^2}$$

and $a' = \bar{x} - b'\,\bar{y}$

- Line of regression of Y on X is also given as

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

- Line of regression of X on Y is also given as

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

- $r^2 = b_{YX} \cdot b_{XY} = b \cdot b'$
- If $b_{YX} > 1$ then $b_{XY} < 1$
- $\left|\frac{b_{yx} + b_{xy}}{2}\right| \geq |r|$
- Regression coefficients are independent of change of origin but not of scale.
- Lines of regression have a point of intersection$(\bar{x}, \bar{y})$

MISCELLANEOUS EXERCISE - 3

I) Choose the correct alternative.

1. Regression analysis is the theory of
   a) Estimation b) Prediction
   c) Both a and b d) Calculation

2. We can estimate the value of one variable with the help of other known variable only if they are
   a) Correlated
   b) Positively correlated
   c) Negatively correlated
   d) Uncorrelated

3. There are _____ types of regression equations.
   a) 4 b) 2 c) 3 d) 1

4. In the regression equation of Y on X
   a) X is independent and Y is dependent.
   b) Y is independent and X is dependent.
   c) Both X and Y are independent.
   d) Both X and Y are dependent.

5. In the regression equation of X on Y
   a) X is independent and Y is dependent.
   b) Y is independent and X is dependent.
   c) Both X and Y are independent.
   d) Both X and Y are dependent.

6. b_{XY} is _____
   a) Regression coefficient of Y on X
   b) Regression coefficient of X on Y
   c) Correlation coefficient between X and Y
   d) Covariance between X and Y

7. b_{YX} is _____
   a) Regression coefficient of Y on X
   b) Regression coefficient of X on Y
   c) Correlation coefficient between X and Y
   d) Covariance between X and Y

8. 'r' is _____
   a) Regression coefficient of Y on X
   b) Regression coefficient of X on Y
   c) Correlation coefficient between X and Y
   d) Covariance between X and Y

9. $b_{XY} . b_{YX}$ _____
   a) $v(x)$ (b) σ_x (c) r^2 (d) $(\sigma_y)^2$

10. If $b_{yx} > 1$ then b_{xy} is _____
    a) >1 (b) < 1 (c) > 0 (d) < 0

11. $| b_{xy} + b_{yx}| \geq$ _____
    a) $|r|$ (b) $2|r|$ (c) r (d) $2r$

12. b_{xy} and b_{yx} are _____
    a) Independent of change of origin and scale
    b) Independent of change of origin but not of scale
    c) Independent of change of scale but not of origin
    d. Affected by change of origin and scale

13. If $u = \frac{x-a}{c}$ and $v = \frac{y-b}{d}$ then b_{yx} =_____
    a) $\frac{d}{c} b_{vu}$ b) $\frac{c}{d} b_{vu}$
    c) $\frac{a}{b} b_{vu}$ d) $\frac{b}{a} b_{vu}$

14. If $u = \frac{x-a}{c}$ and $v = \frac{y-b}{d}$ then b_{xy} =_____
    a) $\frac{d}{c} b_{uv}$ b) $\frac{c}{d} b_{uv}$
    c) $\frac{a}{b} b_{uv}$ d) $\frac{b}{a} b_{uv}$

15. Corr (x,x) = _____
    a) 0 (b) 1 (c) -1 (d) can't be found

16. Corr (x,y) = _____
    a) corr(x,x) (b) corr(y,y)
    c) corr(y,x) (d) cov(y,x)

17. Corr $\left(\frac{x-a}{c}, \frac{y-b}{d}\right) = -\, corr(x,y)$ if,
    a) c and d are opposite in sign
    b) c and d are same in sign
    c) a and b are opposite in sign
    d) a and b are same in sign

18. Regression equation of X on Y is ____
    a) $y - \bar{y} = b_{yx}(x - \bar{x})$
    b) $x - \bar{x} = b_{xy}(y - \bar{y})$
    c) $y - \bar{y} = b_{xy}(x - \bar{x})$
    d) $x - \bar{x} = b_{yx}(y - \bar{y})$

19. Regression equation of Y on X is _____

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

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

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

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

20. b_{yx} = _____

a) $r\frac{\sigma_x}{\sigma_y}$ b) $r\frac{\sigma_y}{\sigma_x}$

c) $\frac{1}{r}\frac{\sigma_y}{\sigma_x}$ d) $\frac{1}{r}\frac{\sigma_x}{\sigma_y}$

21. b_{xy} = _____

a) $r\frac{\sigma_x}{\sigma_y}$ b) $r\frac{\sigma_y}{\sigma_x}$

c) $\frac{1}{r}\frac{\sigma_y}{\sigma_x}$ d) $\frac{1}{r}\frac{\sigma_x}{\sigma_y}$

22. Cov (x,y) = _____

a) $\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$

b) $\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n}$

c) $\frac{\sum xy}{n} - \bar{x}\bar{y}$

d) b and c both

23. If $b_{xy} < 0$ and $b_{yx} < 0$ then 'r' is _____

a) > 0 (b) < 0 (c) > 1 (d) not found

24. If equations of regression lines are $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$ then means of x and y are _____

a) (7,4) b) (4,7) c) (2,9) d) (−4,7)

II) Fill in the blanks :

1. If $b_{xy} < 0$ and $b_{yx} < 0$ then 'r' is _____
2. Regression equation of Y on X is_____
3. Regression equation of X on Y is _____
4. There are _____ types of regression equations.
5. Corr $(x, -x)$ = _____
6. If $u = \frac{x-a}{c}$ and $v = \frac{y-b}{d}$ then b_{xy} =_____
7. If $u = \frac{x-a}{c}$ and $v = \frac{y-b}{d}$ then b_{yx} =_____
8. $|b_{xy} + b_{yx}| \geq$ _____
9. If $b_{yx} > 1$ then b_{xy} is _____
10. $b_{xy}.b_{yx}$ = _____

III) State whether each of the following is True or False.

1. Corr (x,x) = 1
2. Regression equation of X on Y is $y - \bar{y} = b_{yx}(x - \bar{x})$
3. Regression equation of Y on X is $y - \bar{y} = b_{yx}(x - \bar{x})$
4. Corr (x,y) = Corr (y,x)
5. b_{xy} and b_{yx} are independent of change of origin and scale.
6. 'r' is regression coefficient of Y on X
7. b_{yx} is correlation coefficient between X and Y
8. If $u = x - a$ and $v = y - b$ then $b_{xy} = b_{uv}$
9. If $u = x - a$ and $v = y - b$ then $r_{xy} = r_{uv}$
10. In the regression equation of Y on X, b_{yx} represents slope of the line.

IV) Solve the following problems.

1. The data obtained on X, the length of time in weeks that a promotional project has been in progress at a small business, and Y, the percentage increase in weekly sales over the period just prior to the beginning of the campaign.

X	1	2	3	4	1	3	1	2	3	4	2	4
Y	10	10	18	20	11	15	12	15	17	19	13	16

Find the equation of regression line to predict the percentage increase in sales if the campaign has been in progress for 1.5 weeks.

2. The regression equation of y on x is given by $3x + 2y - 26 = 0$ Find b_{yx}.

3. If for a bivariate data $\bar{x} = 10$, $\bar{y} = 12$, v(x) = 9, $\sigma_y = 4$ and $r = 0.6$. Estimate y when $x = 5$.

4. The equation of the line of regression of y on x is $y = \frac{2}{9}x$ and x on y is $x = \frac{y}{2} + \frac{7}{6}$. Find

   (i) r (ii) σ_y^2 if $\sigma_x^2 = 4$.

5. Identify the regression equations of x on y and y on x from the following equations,
   $2x + 3y = 6$ and $5x + 7y - 12 = 0$

6. (i) If for a bivariate data $b_{yx} = -1.2$ and $b_{xy} = -0.3$ then find r.

   (ii) From the two regression equations $y = 4x - 5$ and $3x = 2y + 5$, find $\bar{x}$ and $\bar{y}$.

7. The equations of the two lines of regression are $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$ Find

   (i) Means of X and Y

   (ii) Correlation coefficient between X and Y

   (iii) Estimate of Y for X = 2

   (iv) var (X) if var (Y) = 36

8. Find the line of regression of X on Y for the following data:

   $n = 8$, $\sum(x_i - \bar{x})^2 = 36$, $\sum(y_i - \bar{y})^2 = 44$

   $\sum(x_i - \bar{x})(y_i - \bar{y}) = 24$

9. Find the equation of line of regression of Y on X for the following data:

   $n = 8$, $\sum(x_i - \bar{x})(y_i - \bar{y}) = 120$,

   $\bar{x} = 20$, $\bar{y} = 36$, $\sigma_x = 2$, $\sigma_y = 3$.

10. The following results were obtained from records of age (X) and systolic blood pressure (Y) of a group of 10 men.

	X	*Y*
Mean	50	140
Variance	150	165

and $\sum(x_i - \bar{x})(y_i - \bar{y}) = 1120$

Find the prediction of blood pressure of a man of age 40 years.

11. The equations of two regression lines are $10x - 4y = 80$ and $10y - 9x = -40$

    Find:

    (i) $\bar{x}$ and $\bar{y}$

    (ii) b_{YX} and b_{XY}

    (iii) If var (Y) = 36, obtain var (X)

    (iv) r

12. If $b_{YX} = -0.6$ and $b_{XY} = -0.216$ then find correlation coefficient between X and Y. Comment on it.

Activities

1) Consider a group of 70 students of your class to take their heights in cm (x) and weights kg (y). Hence find both the regression equations.

2) The age in years of 7 young couples is given below:

Husband (x)	21	25	26	24	22	30	20
Wife (y)	19	20	24	20	22	24	18

i) Find the equation of regression line of age of husband on age of wife.

ii) Draw the regression line of y on x

iii) Predict the age of wife whose husband's age is 27 years.

3) The equations of two regression lines are

$10x - 4y = 80$ ..... (1)

$10y - 9x = -40$ ..... (2)

$\therefore$ $(\bar{x}, \bar{y})$ is the point of intersection of both the regression lines.

$\therefore$ Solve equations (i) and (ii), we get

$\bar{x} =$ ☐ and $\bar{y} =$ ☐

Now, consider $10x - 4y = 80$

$\therefore a =, \quad b =$

$\therefore$ slope$(m_1) = -\frac{a}{b} =$ ☐/☐

Consider, $10y - 9x = -40$

$\therefore a =, \quad b =$

$\therefore$ slope$(m_2) = -\frac{a}{b} =$ ☐/☐

$\therefore |m_1| > |m_2|$

$\therefore b_{yx} =$ ☐ and $b_{xy} =$ $\frac{1}{\square}$

$\therefore 10x - 4y = 80$ is the regression equations of ☐ on ☐ and

$\therefore 10y - 9x = -40$ is the regression equations of ☐ on ☐.

Now, $r = \pm\sqrt{b_{yx}.b_{xy}}$

$\therefore r = \pm\sqrt{\square/\square} \times \frac{4}{10}$

$r =$ ☐

If V(y) = 36 then $\sigma_y = \sqrt{\square} =$ ☐

$\therefore b_{xy} = r\frac{\sigma_y}{\sigma_x}$

$\therefore$ ☐/☐ = ☐ × ☐/σ_x

$\therefore \sigma_x = (\ )$

$\therefore V(x) = \sigma_x^2 =$ ☐

4) Given $n = 8$, $\sum(x_i - \bar{x})^2 = 36$,

$\sum(y_i - \bar{y})^2 = 40$,

$\sum(x_i - \bar{x})(y_i - \bar{y}) = 24$

$\therefore b_{yx} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$=$ ☐/☐ $=$ ☐/☐

$\therefore b_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}$

$=$ ☐/☐ $=$ ☐/☐

$\therefore$ Regression equation of Y on X :

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

$y -$ ☐ $=$ ☐/☐ $(x -$ ☐$)$

$\therefore$ Regression equation of X on Y :

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

$x -$ ☐ $=$ ☐/☐ $(y -$ ☐$)$

5) Consider, given

$n = 8 \ \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = 120,$

$\bar{y} \ = 36, \sigma_x = 2, \sigma y = 3$

$\therefore \ \text{cov}(x, y) = \dfrac{\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{n}$

$= \dfrac{\square}{\square} = \square$

$\therefore \ b_{yx} = \dfrac{\square}{\sigma_x^4} = \dfrac{150}{\square}$

$b_{xy} = \dfrac{\text{cov}(x, y)}{\square}$

$= \dfrac{\square}{9} = \dfrac{\square}{\square}$

$\therefore$ Regression equation of Y on X :

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

$y - \square = \square(x - \square)$

$\therefore$ Regression equation of Y on X :

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

$x - \square = \square(y - \square)$