8

Regression Analysis

INTRODUCTION

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable given the value of another. For example, if we know that advertising and sales are correlated, we may find out the expected amount of sales for a given advertising expenditure or the required amount of expenditure for achieving a fixed sales target. *The statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression**. With the help of regression analysis, we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.

The dictionary meaning of the term 'regression' is the act of returning or going back. The term 'regression' was first used in 1877 by Francis Galton while studying the relationship between the height of fathers and sons. His study of height of about one thousand fathers and sons revealed a very interesting relationship, *i.e.*, tall fathers tend to have tall sons and short fathers, short sons; but the average height of the sons of a group of tall fathers is less than that of the tall fathers and the average height of the sons of a group of short fathers is greater than that of the short fathers. The line describing this tendency to regress or going back was called by Galton a 'Regression Line'. The term is still used to describe that line drawn for a group of points to represent the trend present, but it no longer necessarily carries the original implication that Galton intended. These days there is a growing tendency of the modern writers to use the term *estimating line or predicting line* instead of *regression line.*

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy. The regression analysis helps in three important ways :

1. It provides estimates of values of the dependent variables from values of independent variables. The device used to accomplish the estimation procedure is the regression line which describes the average relationship existing between X and Y variables.

2. The second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimations. For this purpose, the standard error of estimate is calculated. If the line fits the data closely, that is, if there is relatively little scatter of the observations around the regression line, good estimate can be made of Y variable. On the other hand, if there is a great deal of scatter of the observations around the fitted regression line, the line will not produce accurate estimates of the dependent variable.

3. With the help of regression analysis, we can obtain a measure of the degree of association or correlation that exists between the two variables. The coefficient of determination calculated for this purpose measures the strength of the relationship that exists between the variables. It assesses the proportion of variance that has been accounted for by the regression equation.

The tool of regression analysis can be extended to three or more variables. But in this text we shall confine ourselves to the problems of two variables only, *i.e.*, simple regression.

Difference between Correlation and Regression Analysis

There are two important points of difference between correlation and regression analysis :

1. Whereas correlation coefficient is a measure of degree of relationship between X and Y, the objective of regression analysis is to study the '*nature of relationship*' between the variables.

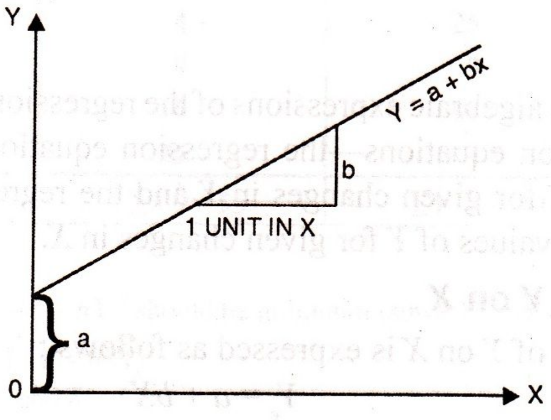
2. The cause and effect relation is clearly indicated through regression analysis than by correlation. Correlation is merely a tool of ascertaining the degree of relationship between two variables and, therefore, we cannot say that one variable is the cause and the other the effect.

THE LINEAR BIVARIATE REGRESSION MODEL

In regression analysis, as in other types of statistical studies, we usually proceed by observing the sample data and using the results obtained as estimates of the corresponding population relationship. To make valid inferences, we must assume some population model. For a bivariate population, there are many possible models that can be constructed to describe the mutual variations of the two variables. The particular one in which we are interested is called the *simple linear regression model.* This model is constructed under the following set of assumptions :

(1) The value of the dependent variable, Y, is dependent in some degree upon the value of the independent variable, X. The dependent variable is assumed to be a random variable, but the values of X are assumed to be fixed quantities that are selected and controlled by the experimenter. The requirement that the independent variable assumes fixed values, however, is not a critical one. Useful results can still be obtained by regression analysis in the case where both X and Y are random variables.

(2) The average relationship between X and Y can be adequately described by a linear equation $Y = a + bX$ whose geometrical presentation is a straight line as in the diagram that follows :



As is clear from the above diagram, the height of the line tells the average value of Y at a fixed value of X. When $X = 0$, the average value of Y is equal to a. The value of a is called the Y intercept, since it is the point at which the straight line crosses the Y-axis. The slope of the line is measured by b, which gives the average amount of change of Y per unit change in the value of X. The sign of b also indicates the type of relationship between Y and X.

(3) Associated with each value of X there is a sub-population of Y. The distribution of the sub-population may be assumed to be normal or non-specified in the sense that it is unknown. In any event, the distribution of each population Y is conditional to the value of X.

(4) The mean of each sub-population Y is called the expected value of Y for a given X: $E(Y/X) = \mu_{yx}$. Furthermore, under the assumption of a linear relationship between X and Y, all values of $E(Y/X)$ or μ_{yx} must fall on a straight line. That is

$$E(Y/X) = \mu_{yx} = a + bX$$

which is the population regression equation for our bivariate linear model. In this equation a and b are called the population regression coefficients.

(5) An individual value in each sub-population Y, may be expressed as :

$$Y = E(Y/X) + e$$

where e is the deviation of a particular value of Y from μ_{yx} and is called the *error term* or *the stochastic disturbance term*. The errors are assumed to be independent random variables because Y's are random variables and independent. The expectations of these errors are zero; $E(e) = 0$. Moreover, if Y's are normal variables, the error can also be assumed to be normal.

(6) It is assumed that the variances of all sub-populations, called variances of the regression, are identical.

Regression Lines

If we take the case of two variables X and Y, we shall have two regression lines as the regression line of X on Y and the regression line of Y on X. 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. The farther the two regression lines are from each other, the lesser is the degree of correlation and the nearer the two regression lines to each other, the higher is the degree of correlation. If the variables are independent, r is zero and the lines of regression are at right angles, *i.e.*, parallel to X-axis and Y-axis.

It should be noted that the regression lines cut each other at the point of average of X and Y, *i.e.*, if from the point where both the regression lines cut each other, a perpendicular is drawn on the X-axis, we will get the mean value of X and if from the point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.

Regression Equations

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations—the regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation of *Y* on *X*

The regression equation of Y on X is expressed as follows :

$$Y_e = a + bX$$

Where Y_e is the dependent variable to be estimated and X is the independent variable.

In this equation a and b are two unknown constants (fixed numerical values) which determine the position of the line completely. The constants are called the parameters of the line. If the value of either or both of them is changed, another line is determined. The parameter 'a' determines the *level* of the fitted line (*i.e.*, the distance of the line directly above or below the origin). The parameter 'b' determines the *slope* of the line, *i.e.*, the change in Y for unit change in X.

If the values of the constants 'a' and 'b' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the *method of least squares* which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the vertical deviations of the actual Y values from the estimated Y values is the least, or, in other words, in order to obtain a line which fits the points best, $(Y - Y_c)^2$ should be minimum*. Such a line is known as the line of best fit.

With a little algebra and differential calculus, it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters a and b such that the least squares requirement is fulfilled.

$$\Sigma Y = Na + b\Sigma X \quad \text{... (i)}$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{... (ii)}$$

These equations are usually called the *normal equations*. In the equations $\Sigma X, \Sigma Y, \Sigma XY, \Sigma X^2$ indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the total number of observed pairs of values.

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows :

$$X = a + bY$$

To determine the values of a and b the following two normal equations are to be solved simultaneously :

$$\Sigma X = Na + b\Sigma Y \quad \text{...(i)}$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \quad \text{...(ii)}$$

Illustration 1. Calculate the regression equations of X on Y and Y on X from the following data :

X	: 1	2	3	4	5
Y	: 2	5	3	8	7

Solution :

CALCULATION OF REGRESSION EQUATIONS

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	3	9	9	9
4	8	16	64	32
5	7	25	49	35
$\Sigma X = 15$	$\Sigma Y = 25$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 151$	$\Sigma XY = 88$

*$\Sigma(Y - Y_c)^2$ should be minimum or $\Sigma(Y - a - bX)^2$ should be minimum (since $Y_c = a + bX$).

Let $$S = \Sigma(Y - a - bX)^2$$

Differentiating partially with respect to a and b

$$\frac{\partial S}{\partial a} = \Sigma(Y - a - bX)(-1) = 0$$

$$\frac{\partial S}{\partial b} = \Sigma(Y - a - bX)(-X) = 0$$

or
$$\Sigma(Y - a - bX) = 0$$
$$\Sigma(Y - a - bX)X = 0$$

or
$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Regression equation of X on Y is given by

$$X = a + bY$$

The normal equations are :

$$\Sigma X = Na + b\Sigma Y$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get

$$15 = 5a + 25b \quad \text{...(i)}$$
$$88 = 25a + 151b \quad \text{...(ii)}$$

Solving (*i*) and (*ii*), we get

$$a = 0.5 \text{ and } b = 0.5$$

Hence the required regression equation of X on Y is given by

$$X = 0.5 + 0.5Y$$

Regression equations of Y on X is : $Y = a + bX$

The normal equations are:

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values, we get

$$25 = 5a + 15b \quad \text{...(iii)}$$
$$88 = 15a + 55b \quad \text{...(iv)}$$

Solving (*iii*) and (*iv*), we get

$$a = 1.10 \text{ and } b = 1.3$$

Hence the required regression equation of Y on X is given by

$$Y = 1.10 + 1.30X$$

Illustration 2. After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are given figures for the sales of automobiles in the five cities for the year 2003 and the number of families residing in those cities.

City	*No. of families in lakhs (X)*	*Sale of Automobiles in 000's (Y)*
A	70	25.2
B	75	28.6
C	80	30.2
D	60	22.3
E	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2006 for city *A* which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution. CALCULATION OF REGRESSION EQUATION

City	*X*	*Y*	X^2	*XY*
A	70	25.2	4,900	1,764
B	75	28.6	5,625	2,145
C	80	30.2	6,400	2,416
D	60	22.3	3,600	1,338
E	90	35.4	8,100	3,186
	$\Sigma X = 375$	$\Sigma Y = 141.7$	$\Sigma X^2 = 28,625$	$\Sigma XY = 10,849$

Regression equation of Y on X is $Y = a + bX$.

To determine the values of a and b, we shall solve the normal equations

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values from the table, the normal equations become

$$141.7 = 5a + 375b \quad \text{...(i)}$$
$$10,849 = 375a + 28,625b \quad \text{...(ii)}$$

Multiplying Eqn. (*i*) by 75 and subtracting from Eqn. (*ii*), we get

$$221.5 = 500b \text{ or } b = 0.443$$

Substituting the value of b in Eqn. (*i*), we have

$$-24.425 = 5a \text{ or } a = -4.885$$

Therefore, the regression equation of Y on X is

$$Y = -4.885 + 0.443X$$

Estimated sales for the year 2006 for city A

$$Y = -4.885 + 0.443\,(100)$$
$$= -4.885 + 44.3 = 39.415$$

Hence it is expected that about 39,415 autos would be sold in city A having a population of 100 lakh families.

Deviations taken from Arithmetic Means of X and Y

The calculations by the direct method discussed above are quite cumbersome when the values of X and Y are large. The work can be simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means. In such a case, the equation $Y = a + bX$ is changed to :

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

The value of b_{yx} can be easily obtained as follows :

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

where $\quad x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

The two normal equations which we had written earlier when changed in terms of x and y become

$$\Sigma y = Na + b\Sigma x \qquad \ldots (i)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \ldots (ii)$$

Since $\quad \Sigma x = \Sigma y = 0$ [deviations being taken from means]

Equation (*i*) reduces to

$$Na = 0, \quad \therefore a = 0$$

Equation (*ii*) reduces to

$$\Sigma xy = b\Sigma x^2 \quad \therefore b \text{ or } b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

After obtaining the value of b_{yx} the regression equation can easily be written in terms of X and Y by substituting for y, $(Y - \bar{Y})$ and for x, $(X - \bar{X})$.

Similarly, the regression equation $X = a + bY$ is reduced to $(X - \bar{X}) = b_{xy}\,(Y - \bar{Y})$ and the value of b_{xy} can be similarly obtained as

$$b_{xy} = \frac{\Sigma xy}{\Sigma x^2}$$

Illustration 3. In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales :

Salesmen :	1	2	3	4	5	6	7	8	9	10
Test score :	40	70	50	60	80	50	90	40	60	60
Sales ('000 Rs.) :	2.5	6.0	4.0	5.0	4.0	2.5	5.5	3.0	4.5	3.0

Calculate the regression equation of sales on test scores and estimate the probable weekly sales volume if a salesman makes a score of 100.

Solution. Let sales be denoted by Y and test scores by X. We have to fit a regression equation of Y on X, *i.e.*, $Y - \bar{Y} = b_{yx}\,(X - \bar{X})$.

CALCULATION OF REGRESSION EQUATION

Salesmen	Test Score X	$(X-\bar{X})$ x	x^2	Sales Y	$(Y-\bar{Y})$ y	y^2	xy
1	40	−20	400	2.5	−1.5	2.25	+30
2	70	+10	100	6.0	+2.0	4.00	+20
3	50	−10	100	4.0	0	0	0
4	60	0	0	5.0	1.0	1.00	0
5	80	+20	400	4.0	0	0	0
6	50	−10	100	2.5	−1.5	2.25	+15
7	90	+30	900	5.5	+1.5	2.25	+45
8	40	−20	400	3.0	−1.0	1.00	+20
9	60	0	0	4.5	+0.5	0.25	0
10	60	0	0	3.0	−1.0	1.00	0
$N = 10$	$\Sigma X = 600$	$\Sigma x = 0$	$\Sigma x^2 = 2{,}400$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 14$	$\Sigma xy = 130$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{600}{10} = 60;\quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{10} = 4$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{130}{2{,}400} = 0.054$$

The regression equation of sales and test scores is given as :

$$Y - 4 = 0.054\,(X - 60)$$

$$Y = 0.76 + 0.054\,X$$

When X is 100, Y would be

$$Y = 0.76 + 0.054\,(100) = 6.16.$$

Thus the most probable weekly sales volume if salesman makes a score of 100 is 6.16 thousand rupees.

Deviations taken from Assumed Means

When actual means of X and Y variables are in fractions, the calculations can be simplified by taking the deviations from the assumed mean. The value of b, *i.e.*, the regression coefficient, will be calculated as follows :

Regression equation of X on Y : $(X - \bar{X}) = b_{xy}\,(Y - \bar{Y})$

where $$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^{\,2} - (\Sigma d_y)^2}$$

Regression equation of Y on *X* : $(Y - \bar{Y}) = b_{yx}\,(X - \bar{X})$

where $$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^{\,2} - (\Sigma d_x)^2}$$

Once the values of b_{xy} and b_{yx} are determined in the above manner, the regression equations can be obtained very easily.

Illustration. 4. A company wants to assess the impact of R & D expenditure on its annual profit. The following table presents the information for the last eight years :

Years	:	2010	2009	2008	2007	2006	2005	2004	2003
R & D expenditure (Rs. '000)	:	9	7	5	10	4	5	3	2
Annual Profit (Rs. '000)	:	45	42	41	60	30	34	25	20

Estimate the regression equation and predict the annual profit for 2009 for an allocated sum of Rs. 100,000 as R & D expenditure.

Solution. Let R & D expenditure be denoted by X and annual profit by Y.

CALCULATION OF REGRESSION EQUATION

Year	X	(X − 6) d_x	d_x^2	Y	(Y − 37) d_y	d_y^2	d_xd_y
2003	2	−4	16	20	−17	289	+68
2004	3	−3	9	25	−12	144	+36
2005	5	−1	1	34	−3	9	+3
2006	4	−2	4	30	−7	49	+14
2007	10	+4	16	60	+23	529	+92
2008	5	−1	1	41	+4	16	−4
2009	7	+1	1	42	+5	25	+5
2010	9	+3	9	45	+8	64	+24
	$\Sigma X = 45$	$\Sigma d_x = -3$	$\Sigma d_x^2 = 57$	$\Sigma Y = 297$	$\Sigma d_y = 1$	$\Sigma d_y^2 = 1125$	$\Sigma d_xd_y = 238$

Fitting regression equation of *Y* on *X*, we get

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

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{297}{8} = 37.125; \quad \bar{X} = \frac{\Sigma X}{N} = \frac{45}{8} = 5.625$$

$$b_{yx} = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 238 - (-3)(1)}{8 \times 57 - (-3)^2} = \frac{1904 + 3}{456 - 9} = \frac{1907}{447} = 4.266$$

$$Y - 37.125 = 4.266\,(X - 5.625) \text{ Or } Y - 37.125 = 4.266\,X - 23.996$$

$$Y = 13.129 + 4.266X; \text{ When } X \text{ is 10, } Y \text{ shall be}$$

$$Y = 13.129 + 4.266(100) = 439.729$$

Thus the likely expenditure on Research and Development for an allocation of Rs. 100,000 is Rs. 439.729.

Regression Coefficients

The Quantity *b* in the regression equations is called the "regression coefficient" or "slope coefficient". Since there are two regression equations, therefore, there are two regression coefficients—regression coefficient of *X* on *Y* and regression coefficient of *Y* on *X*.

Regression Coefficient of X on Y

The regression coefficient of *X* on *Y* is represented by the symbol b_{xy} or b_1. It measures the amount of change in *X* corresponding to a unit change in *Y*. The regression coefficient of *X* on *Y* is given by

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

When deviations are taken from the means of *X* and *Y*, the regression coefficient is obtained by

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

When deviations are taken from assumed means, the value of b_{xy} is obtained as follows :

$$b_{xy} = \frac{N\Sigma d_xd_y - \Sigma d_x\Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression Coefficient of Y on X

The regression coefficient of *Y* on *X* is represented by b_{yx} or b_2. It measures the amount of change in *Y* corresponding to a unit change in *X*. The value of b_{yx} is given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

When deviations are taken from actual means of *X* and *Y*,

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

When deviations are taken from assumed means,

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^{\,2} - (\Sigma d_x)^2}$$

Properties of the Regression coefficients

(1) The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically :

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

Proof. $$b_{xy} = r\frac{\sigma_x}{\sigma_y}\ ;\ b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\therefore \quad b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2.$$

(2) If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity. For example, if $b_{xy} = 1.2$ and $b_{yx} = 1.4$, r would be $\sqrt{1.2 \times 1.4} = 1.29$ which is not possible.

(3) Both the regression coefficients will have the same sign, *i.e.*, they will be either positive or negative. In other words, it is not possible that one of the regression coefficients is having minus sign and the other plus sign.

(4) The coefficient of correlation will have the same sign as that of regression coefficients, *i.e.*, if regression coefficients have a negative sign, r will also have negative sign and if the regression coefficients have a positive sign, r would also be positive. For example,

$$\text{if } b_{xy} = -0.2 \text{ and } b_{yx} = -0.8$$

$$r = -\sqrt{0.2 \times 0.8} = -0.4$$

(5) The average value of the two regression coefficients would be greater than the value of coefficient of correlation. In symbols $(b_{xy} + b_{yx})/2 > r$. For example, if $b_{xy} = 0.8$ and $b_{yx} = 0.4$, the average of the two values would be $(0.8 + 0.4)/2 = 0.6$ and the value of r would be $\sqrt{0.8 \times 0.4} = 0.566$ which is less than 0.6.

(6) Regression coefficients are independent of change of origin but not scale.*

---

*Proof

$$b_{yx} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \quad \text{or} \quad b_{yx} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2}$$

Let $$u = \frac{X-a}{h} \text{ and } v = \frac{Y-b}{k}$$

Then $$X = a + hu, \text{ and } Y = b + kv$$

and $$\overline{X} = a + h\overline{u}\ ;\quad \overline{Y} = b + k\overline{v}$$

Subtracting, we get

$$(X - \overline{X}) = h(u - \overline{u})\ ;\ (Y - \overline{Y}) = k(v - \overline{v})$$

Substituting these values in the above formula, we get

$$b_{yx} = \frac{\Sigma hk\,(u-\overline{u})(v-\overline{v})}{\Sigma h^2\,(u-\overline{u})^2}$$

$$= \frac{k}{h}\,\frac{\Sigma(u-\overline{u})(v-\overline{v})}{\Sigma(u-\overline{u})^2} = \frac{k}{h}\,b_{vu}$$

Similarly, we have $$b_{xy} = \frac{h}{k}\,b_{uv}$$

Hence the result.

Illustration 5. On the basis of figures recorded below for 'Supply' and 'Price' for nine years, calculate the regression coefficients and the value of r :

Year	:	2002	2003	2004	2005	2006	2007	2008	2009	2010
Supply	:	80	82	86	91	83	85	89	96	93
Price	:	145	140	130	124	133	127	120	110	116

Solution. Let the price be denoted by Y and supply by X.

CALCULATION OF REGRESSION COEFFICIENTS

Year	*Supply* X	$(X-90)$ d_x	d_x^2	*Price* Y	$(Y-127)$ d_y	d_y^2	$d_x d_y$
2002	80	−10	100	145	+18	324	−180
2003	82	− 8	64	140	+13	169	− 104
2004	86	− 4	16	130	+3	9	− 12
2005	91	+1	1	124	−3	9	−3
2006	83	−7	49	133	+ 6	36	−42
2007	85	−5	25	127	0	0	0
2008	89	−1	1	120	−7	49	+ 7
2009	96	+ 6	36	110	− 17	289	−102
2010	93	+ 3	9	116	− 11	121	−33
$N = 9$	$\Sigma X = 785$	$\Sigma d_x = -25$	$\Sigma d_x^2 = 301$	$\Sigma Y = 1,145$	$\Sigma d_y = +2$	$\Sigma d_y^2 = 1,006$	$\Sigma d_x d_y = -469$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{9 \times -469 - (-25)(2)}{9 \times 301 - (-25)^2} = \frac{-4221 + 50}{2709 - 625} = -\frac{4171}{2084} = -2.001$$

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{9 \times -469 - (-25)(2)}{9 \times 1006 - (-2)^2} = \frac{-4221 + 50}{9054 - 4} = -\frac{4171}{9050} = -0.461$$

$$r = \sqrt{b_{yx} \times b_{xy}} = -\sqrt{2.001 \times .461} = -0.96.$$

It is a case of very high degree of negative correlation.

Illustration 6. The following data relate to advertising expenditure (in lakhs of rupees) and their corresponding sales (in crores of rupees) :

Advertising Expenditure	:	10	12	15	23	20
Sales	:	14	17	23	25	21

Estimate (*i*) the sales corresponding to advertising expenditure of Rs. 30 lakhs and (*ii*) the advertising expenditure for a sales target of Rs. 35 crores.

Solution. Let advertising expenditure be denoted by X and sales by Y.

CALCULATION OF REGRESSION EQUATIONS

X	$(X-16)$ x	x^2	Y	$(Y-20)$ y	y^2	xy
10	−6	36	14	−6	36	+36
12	−4	16	17	−3	9	+12
15	−1	1	23	+3	9	−3
23	+7	49	25	+5	25	+35
20	+4	16	21	+1	1	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 118$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 80$	$\Sigma xy = +84$

(*i*) *Regression equation of Y on X* : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{100}{5} = 20; \quad \overline{X} = \frac{\Sigma X}{N} = \frac{80}{5} = 16$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{84}{118} = 0.712$$

$$Y - 20 = .712\,(X - 16)$$

$$Y - 20 = .712X - 11.392 \;\text{ or }\; Y = 8.608 + 0.712X$$

$$Y_{30} = 8.608 + 0.712\,(30) = 8.608 + 21.36 = 29.968$$

Thus the likely sales corresponding to advertising expenditure of Rs. 30 lakhs is Rs. 29.968 crores.

(*ii*) *Regression equation of X on Y* : $X - \overline{X} = b_{xy}\,(Y - \overline{Y})$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{84}{80} = 1.05$$

$$X - 16 = 1.05\,(Y - 20)$$

$$X = -5 + 1.05Y$$

$$X_{35} = -5 + 1.05(35) = -5 + 36.75 = 31.75.$$

Thus, the advertising expenditure for a sales target of Rs. 35 crores is Rs. 31.75 lakhs.

Regression Equations in Bivariate Grouped Frequency Distributions

While calculating regression equations for bivariate grouped frequency distributions, first of all we will have to prepare a correlation tables as was discussed in the chapter on Correlation. Then we will find out the value of $\overline{X}$, $\overline{Y}$ and the two regression coefficients and proceed in the usual manner. However, special care must be exercised while calculating the value of regression coefficient because regression coefficients are independent of the change of origin but not of scale. The values of b_{xy} and b_{yx} shall be obtained as follows :

$$b_{xy} = \frac{N\Sigma fd_x d_y - \Sigma fd_x\,\Sigma fd_y}{N\Sigma fd_x^{\,2} - (\Sigma fd_x)^2} \times \frac{h}{k}$$

where h = width of the class-interval of the X variable

and k = width of class-interval of the Y variable.

$$b_{yx} = \frac{N\Sigma fd_x d_y - \Sigma fd_x\,\Sigma fd_y}{N\Sigma fd_x^{\,2} - (\Sigma fd_x)^2} \times \frac{h}{k}$$

(For proof see section on properties of Regression Coefficients.)

Illustration 7. Obtain the two regression equations from the following bivariate frequency distribution :

Sales Revenue (in Rs. lakhs)	*Advertising Expenditure (in Rs. thousand)* 5 – 15	15 – 25	25 – 35	35 – 45
75 – 125	3	4	4	8
125 – 175	8	6	5	7
175 – 225	2	2	3	4
225 – 275	3	3	2	2

Estimate (*i*) the sales corresponding to advertising expenditure of Rs. 50 thousand (*ii*) the advertising expenditure for a sales revenue of Rs. 300 lakhs (*iii*) the coefficient of correlation and interpret its value. (*MBA, Delhi Univ., 2004, 2007*)

Solution. Let sales revenue be denoted by X and advertising expenditure by Y.

CALCULATION OF REGRESSION LINES

X m.p.	X	dx \ dy	Y m.p. 10 5 – 15 -2	20 15 – 25 -1	30 25 – 35 0	40 35 – 45 $+1$	f	fd_x	fd_x^2	fd_xd_y
100	75–125	-1	3 (6)	4 (4)	4 (0)	8 (-8)	19	-19	19	2
150	125–175	0	8 (0)	6 (0)	5 (0)	7 (0)	26	0	0	0
200	175–225	$+1$	2 (-4)	2 (-2)	3 (0)	4 (4)	11	11	11	-2
250	225–275	$+2$	3 (-12)	3 (-6)	2 (0)	2 (4)	10	20	40	-14
f			16	15	14	21	$N = 66$	$\Sigma fd_x = 12$	$\Sigma fd_x^2 = 70$	$\Sigma fd_xd_y = -14$
fd_y			-32	-15	0	21	$\Sigma fd_y = -26$			
fd_y^2			64	15	0	21	$\Sigma fd_y^2 = 100$			
fd_xd_y			-10	-4	0	0	$\Sigma fd_xd_y = -14$			

Regression equation of X on Y : $X - \overline{X} = b_{xy}(Y - \overline{Y})$

$$\overline{X} = A + \frac{\Sigma fd_x}{N} \times h = 150 + \frac{12}{66} \times 50 = 150 + 9.09 = 159.09$$

$$\overline{Y} = B + \frac{\Sigma fd_y}{N} \times k = 30 + \frac{-26}{66} \times 10 = 30 - 3.94 = 26.06$$

$$b_{xy} = \frac{N\Sigma fd_xd_y - \Sigma fd_x\Sigma fd_y}{N\Sigma fd_y^2 - (\Sigma fd_y)^2} \times \frac{h}{k} = \frac{66(-14) - 12\,(-26)}{66\,(100) - (-26)^2} \times \frac{50}{10}$$

$$= \frac{-924 + 312}{6600 - 676} \times \frac{50}{10} = -\frac{3060}{5924} = -0.5165.$$

Therefore, the regression equation of X on Y is : $X - 159.09 = -0.5165\,(Y - 26.06)$

$$X - 159.09 = -0.5165Y + 13.46 \text{ or } X = 172.55 - 0.5165Y$$

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{N\Sigma fd_xd_y - \Sigma fd_x\Sigma fd_y}{N\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{k}{h} = \frac{66(-14) - 12(-26)}{66(70) - (12)^2} \times \frac{10}{50} = -0.0273$$

Therefore, the regression equation of Y on X is : $Y - 26.06 = -0.0273\,(X - 159.09)$

$$Y = 26.06 - 0.0273X + 4.343 = 30.40 - 0.0273X.$$

The sales revenue corresponding to advertising expenditure of Rs. 50 thousand

(i)
$$X_{50} = 172.55 - 0.5165\,(50)$$
$$= 172.55 - 25.825 = 146.725$$

(*ii*)

$$Y_{300} = 30.40 - 0.0273\,(300)$$
$$= 30.40 - 8.19 = 22.21$$

Hence, to attain sales revenue of Rs. 300 lakhs, the advertising expenditure required is Rs. 22.21 lakhs.

(*iii*)

$$r = \sqrt{b_{xy} \times b_{yx}}$$
$$= \sqrt{.5165 \times .0273} = -0.119.$$

Standard Error of Estimate

As we find it necessary to supplement an average with a measure of dispersion or variation, so in order to see how good or representative the regression line is, we look for a measure of variation about it. If we have a wide scatter or variation of the dots about the regression line, then it would have to be considered a poor representative of the relationship. The more closely the dots cluster around the line, the more representative it is and the better the estimate based on the equation for this line. And if the dots should all lie on the regression line a (hypothetical situation), then there is no variation about the line and the correlation is perfect.

The variation about the line of average relationship can be measured in the manner similar to the measuring of the variation of the items about an average. Thus, we use here a measuring of variation similar to the standard deviation—**the standard error of estimate.**

The measure of variation of the observations around the computed regression line is referred to as the standard error of estimate. Just as the standard deviation is a measure of the scatter of observations in a frequency distribution around the mean of that distribution, the standard error of estimate is a measure of the scatter of the observed values of Y around the corresponding computed values of Y on the regression line. It is computed as a standard deviation, being also a square root of the mean of the squared deviation. But the deviations here are not the deviations of the items from the arithmetic mean ; they are rather the vertical distances of every dot from the line of average relationship.

The deviation of each dot from the regression line is symbolised by $Y - Y_c$. Thus the square root of mean of the squared deviation is :

$$\sqrt{\frac{\Sigma\,(Y - Y_c)^2}{N - 2}}$$

This formula is not convenient from the computational point of view because it requires the computation of Y_c, *i.e.,* estimated values of Y. A more convenient formula is given below :

$$S_{y.x} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y + b\Sigma YX}{N - 2}}$$

where $S_{y.x}$ denote the S.E. of estimate of regression equation of y on x.

Similarly, we can calculate $S_{x.y}$.

$$S_{x.y} = \sqrt{\frac{\Sigma(X - X_c)^2}{N - 2}}$$

or

$$S_{x.y} = \sqrt{\frac{\Sigma X^2 - a\Sigma X + b\Sigma XY}{N - 2}}$$

The standard error of estimate can very easily be calculated with the help of the following formula :

$$S_{x.y} = S_y\sqrt{1-r^2}\ ; \qquad S_{y.x} = S_x\sqrt{1-r^2}$$

The standard error of estimate measures the accuracy of the estimated figures. The smaller the value of standard error of estimate, the closer will be dots to the regression line and the better the estimates based on the equation for this line. If standard error of estimate is zero, then there is no variation about the line and the correlation will be perfect. Thus with the help of standard error of estimate it is possible for us to ascertain how good and representative the regression line is as a description of the average relationship between two series.

Coefficient of Determination

The ratio of the unexplained variation to the total variation represents the proportion of variation in Y that is not explained by regression on X. Subtraction of this proportion from 1.0 gives the proportion of variation in Y that is explained by regression on X. The statistic used to express this proportion is called the coefficient of determination and is denoted by R^2. It may be written as follows :

$$R^2 = 1 - \frac{\text{Variation in } Y \text{ remaining after regression on } X}{\text{Total variation in } Y}$$

$$R^2 = 1 - \frac{\text{Error sum of squares}}{\text{Total sum of squares}}$$

The value of R^2 is the proportion of the variation in the dependent variable Y explained by regression on the independent variable X.

MISCELLANEOUS ILLUSTRATIONS

Illustration 8. Given the following bivariate data :

X :	–1	5	3	2	1	1	7	3
Y :	–6	1	0	0	1	2	1	5

(*a*) Fit a regression line of Y on X and predict Y if $X = 10$.

(*b*) Fit a regression line of X on Y and predict X if $Y = 2.5$.

(MBA, Osmania Univ., 2001)

Solution.

FITTING REGRESSION EQUATIONS

X	$(X-3)$ d_x	d_x^2	Y	$(Y-2)$ d_y	d_y^2	d_xd_y
–1	–4	16	–6	–8	64	+32
5	+2	4	+1	–1	1	–2
3	0	0	0	–2	4	0
2	–1	1	0	–2	4	+2
1	–2	4	+1	–1	1	+2
1	–2	4	+2	0	0	–0
7	+4	16	+1	–1	1	–4
3	0	0	+5	+3	9	0
$\Sigma X = 21$	$\Sigma d_x = -3$	$\Sigma d_x^2 = 45$	$\Sigma Y = 4$	$\Sigma d_y = -12$	$\Sigma d_y^2 = 84$	$\Sigma d_xd_y = 30$

(*a*) $Y-\overline{Y} = b_{yx}(X-\overline{X})$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{(8)(30) - (-3)(-12)}{(8)(45) - (-1)^2}$$

$$= \frac{240-36}{360-1} = \frac{204}{359} = 0.568$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{4}{8} = 0.5; \quad \overline{X} = \frac{\Sigma X}{N} = \frac{21}{8} = 2.625$$

$$Y - 0.5 = .568\,(X - 2.625)$$

$$Y = .568X - 0.991$$

(*b*) $X-\overline{X} = b_{xy}(Y-\overline{Y})$

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{(8)(30) - (-3)(-12)}{(8)(84) - (-12)^2} = \frac{204}{528} = 0.386$$

$$X - 2.625 = .386\,(Y - .5)$$

$$X = .386Y + 2.432$$

If $Y = 2.5$, X shall be

$$X = .386\,(2.5) + 2.432 = 3.397.$$

Illustration 9. From the following data obtain the two regression equations :

Sales :	91	97	108	121	67	124	51	73	111	57
Purchase :	71	75	69	97	70	91	39	61	80	47

Solution : CALCULATION OF REGRESSION EQUATIONS

Sales X	$(X-\overline{X})$ $\overline{X}=90$ x	x^2	*Purchase* Y	$(Y-\overline{Y})$ $\overline{Y}=70$ y	y^2	xy
91	+1	1	71	+1	1	+1
97	+7	49	75	+5	25	+35
108	+18	324	69	−1	1	−18
121	+31	961	97	+27	729	+837
67	−23	529	70	0	0	0
124	+34	1156	91	+21	441	+714
51	−39	1521	39	−31	961	+1209
73	−17	289	61	−9	81	+153
111	+21	441	80	+10	100	+210
57	−33	1089	47	−23	529	+759
$\Sigma X = 900$	$\Sigma x = 0$	$\Sigma x^2 = 6360$	$\Sigma Y = 700$	$\Sigma y = 0$	$\Sigma y^2 = 2868$	$\Sigma xy = 3900$

Regression equation of X on Y : $X-\overline{X} = b_{xy}(Y-\overline{Y})$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{900}{10} = 90; \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{700}{10} = 70$$

$$b_{xy} = \frac{\Sigma XY}{\Sigma y^2} = \frac{3900}{2868} = 1.36$$

$$X - 90 = 1.36\,(Y - 70)$$
$$X - 90 = 1.36Y - 95.2 \text{ or } X = -5.2 + 1.36Y$$

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{3900}{6360} = 0.613$$
$$Y - 70 = 0.613\,(X - 90)$$
$$Y - 70 = 0.613X - 55.17 \quad \text{or} \quad Y = 14.83 + 0.613\,X.$$

Illustration 10. The personnel manager of an electronic manufacturing company devises a manual dexterity test for job applicants to predict their production rating in the assembly department. In order to do this he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. Results are as follows :

Worker	:	A	B	C	D	E	F	G	H	I	J
Test Score	:	53	36	88	84	86	64	45	48	39	69
Production Rating	:	45	43	89	79	84	66	49	48	43	76

Fit a linear least square regression equation of production rating on test score. *(MBA, Delhi Univ., 2002)*

Solution. Let test score be denoted by X and production rating by Y. We have to fit a regression equation of Y on X.

FITTING REGRESSION EQUATION OF Y ON X

Worker	X	$(X-61)$ d_x	d_x^2	Y	$(Y-62)$ d_y	d_xd_y
A	53	−8	64	45	−17	+136
B	36	−25	625	43	−19	+475
C	88	+27	729	89	+27	+729
D	84	+23	529	79	+17	+391
E	86	+25	625	84	+22	+550
F	64	+3	9	66	+4	+12
G	45	−16	256	49	−13	+208
H	48	−13	169	48	−14	+182
I	39	−22	484	43	−19	+418
J	69	+8	64	76	+14	+112
	$\Sigma X = 612$	$\Sigma d_x = 2$	$\Sigma d_x^2 = 3554$	$\Sigma Y = 622$	$\Sigma d_y = 2$	$\Sigma d_xd_y = 3213$

Regression Equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10 \times 3213 - 2 \times 2}{10 \times 3554 - (2)^2} = \frac{32130 - 4}{35536} = \frac{32126}{35536} = +\,0.904$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{622}{10} = 62.2; \quad \overline{X} = \frac{\Sigma X}{N} = \frac{612}{10} = 61.2$$

Hence $\quad Y - 62.2 = 0.904(X - 61.2)$

$$Y = .904X - 55.325 + 62.2$$

$Y = 6.875 + 0.904X$ is the required regression equation to predict production rating on test score.

Illustration 11. The following data give the ages and blood pressure of 10 women :

Age (X)	:	56	42	36	47	49	42	60	72	63	55
Blood Pressure (Y)	:	147	125	118	128	145	140	155	160	149	150

(*i*) Find the correlation coefficient between X and Y.

(*ii*) Determine the least square regression equation of Y on X.

(*iii*) Estimate the blood pressure of a woman whose age is 45 years.

Solution. CALCULATION OF CORRELATION COEFFICIENT

Age X	$(X-49)$ d_x	d_x^2	Blood pressure Y	$(Y-145)$ d_y	d_y^2	d_xd_y
56	+7	49	147	+2	4	+14
42	−7	49	125	−20	400	+140
36	−13	169	118	−27	729	+351
47	−2	4	128	−17	289	+34
49	0	0	145	0	0	0
42	−7	49	140	−5	25	+35
60	+11	121	155	+10	100	+110
72	+23	529	160	+15	225	+345
63	+14	196	149	+4	16	+56
55	+6	36	150	+5	25	+30
$\Sigma X = 522$	$\Sigma d_x = 32$	$\Sigma d_x^2 = 1202$	$\Sigma Y = 1417$	$\Sigma d_y = -33$	$\Sigma d_y^2 = 1813$	$\Sigma d_xd_y = 1115$

(*i*) Coefficient of correlation is given by

$$r = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

$$= \frac{11150 + 1056}{\sqrt{12020 - 1024}\sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892$$

There is a high degree of positive correlation between age and blood pressure.

(*ii*) The least square regression equation of Y on X is given by

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

$$\overline{X} = \frac{\Sigma X}{N} = \frac{522}{10} = 52.2; \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{1417}{10} = 141.7$$

and $$b_{yx} = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$$

Substituting these values in the above equation, we have

$$Y - 141.7 = 1.11(X - 52.2)$$

or $$Y = 1.11X + 141.7 - 57.942 = 83.758 + 1.11X$$

This is the required least square regression equation of Y on X.

(*iii*) When $X = 45$, then

$$Y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the most likely blood pressure of a woman of 45 years is 134.

Illustration 12. For the following data determining to production and capacity utilisation :

	Average	*Standard deviation*
Production (in lakh units)	35.6	10.5
Capacity utilisation (in percentage)	84.8	8.5

$r = 0.62$

(*i*) Estimate the production when the capacity utilisation is 70 per cent.

(*ii*) The capacity utilisation to achieve the production of 50 lakh units.

(*MBA, Pune Univ.; MBA, Delhi Univ., 2005*)

Solution. Let production be denoted by the variable X and capacity utilisation by Y. Then the regression equation showing the regression equation of capacity utilisation of production will be given by the following formula :

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

where $\quad b_{yx} = r\dfrac{\sigma_y}{\sigma_x} = 0.62 \times \dfrac{8.5}{10.5} = \dfrac{5.27}{10.5} = 0.5019$

and $\quad \overline{X} = 35.6;\ \overline{Y} = 84.8$

Substituting all these values in the above equation, we get

$$Y - 84.8 = 0.5019\,(X - 35.6)$$

or $\quad Y = 84.8 + 0.5019\,X - 17.8676$ or $Y = 66.9324 + 0.5019X$

which is the required regression of capacity utilisation on production.

To estimate the production, we shall have to find the regression equation of X on Y, *i.e.*,

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

where $\quad b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = 0.62 \times \dfrac{10.5}{8.5} \dfrac{6.51}{8.50} = 0.7659$

Substituting the values, we have

$$(X - 35.6) = 0.7659\,(Y - 84.8)$$

or $\quad X = 35.6 + 0.7659\,Y - 64.9483$

$\quad = -29.3483 + 0.7659\,Y$

when $\quad Y = 70,\ X = -29.3483 + 0.7659\,(70)$

$\quad = -29.3483 + 53.613 = 24.2647$

Hence the estimated production is 2,42,647 units when the capacity utilisation is 70 per cent.

Illustration 13. There are two series of index numbers, P for price index and S for stock of a commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off values of S for various values of P?

Solution. We have to fit an equation $P = a + bS$

$$(P - \overline{P}) = r\frac{\sigma_P}{\sigma_s}(S - \overline{S})$$

$$\overline{P} = 100,\ \overline{S} = 103,\ \sigma_P = 8,\ \sigma_s = 4,\ r = 0.4$$

$$\therefore \quad P - 100 = .4\ \frac{8}{4}(S - 103) \quad \text{or} \quad P = 17.6 + 0.8\,S.$$

The same equation cannot be used to read off values of S for various values of P. For that we have to fit an equation $S = a + bP$.

$$(S - \overline{S}) = r\frac{\sigma_s}{\sigma_P}(P - \overline{P})$$

$$S - 103 = 0.4\frac{4}{8}(P - 100) \text{ or } S = 83 + 0.2\,P.$$

Illustration 14. The following data show the experience of machine operators and their performance ratings as given by the number of good parts turned out per 100 pieces :

Operator	:	1	2	3	4	5	6	7	8
Experience (X)	:	16	12	18	4	3	10	5	12
Performance Rating (Y)	:	87	88	89	68	78	80	75	83

Calculate the regression line of performance ratings on experience and estimate the probable performance if an operator has 10 years' experience. *(MBA, Kumaun Univ., 1999)*

Solution. Let performance rating be denoted by Y and experience by X. We have to calculate the regression line of Y on X.

CALCULATION OF REGRESSION EQUATIONS

Experience X	$(X-10)$ x	x^2	Performance Y	$(Y-81)$ y	y^2	xy
16	+6	36	87	+6	36	+36
12	+2	4	88	+7	49	+14
18	+8	64	89	+8	64	+64
4	−6	36	68	−13	169	+78
3	−7	49	78	−3	9	+21
10	0	0	80	−1	1	0
5	−5	25	75	−6	36	+30
12	+2	4	83	+2	4	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 218$	$\Sigma Y = 648$	$\Sigma y = 0$	$\Sigma y^2 = 368$	$\Sigma xy = 247$

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{247}{218} = 1.133;\quad \overline{Y} = \frac{648}{8} = 81,\ \overline{X} = \frac{80}{8} = 10$$

$\therefore$ $$Y - 81 = 1.133\,(X - 10) = 1.133\,X - 11.33$$

$$Y = 69.67 + 1.133X$$

When $X = 10$, Y will be

$$Y = 69.67 + 1.133\,(10) = 69.67 + 11.33 = 81$$

Thus the probable performance of an operator who has 10 years' experience is 81 good parts out of 100.

Illustration 15. Find the most likely production corresponding to a rainfall of 40" from the following data :

	Rainfall	*Production*
Average	30"	50 quintals
S.D.	5"	10 quintals
Coefficient of correlation	0.8	

Solution. Let rainfall be denoted by X and production by Y. The expected yield corresponding to a rainfall 40" will be obtained by the regression equation of Y on X.

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$\overline{Y} = 50,\ \sigma_y = 10,\ \overline{X} = 30,\ \sigma_x = 5,\ r = 0.8$$

$$Y - 50 = 0.8\frac{10}{5}(X - 30) \text{ or } Y - 50 = 1.6\,(X - 30)$$

$$Y - 50 = 1.6X - 48 \text{ or } Y = 2 + 1.6X$$

When rainfall (X) is 40", the expected production, *i.e.*, Y would be

$$Y = 2 + 1.6\,(40) = 66 \text{ quintals.}$$

Illustration 16. Obtain the regression equations from the data given below :

X:	1	2	3	4	5	6	7	8	9
Y:	9	8	10	12	11	13	14	16	15

Plot the regression equation on a graph paper and determine $\overline{X}$ and $\overline{Y}$. Also calculate the value of correlation coefficient.

(BBA, BHU, 2000; MBA, Hyderabad Univ., 2005)

Solution. CALCULATION OF REGRESSION EQUATIONS

X	$(X-\overline{X})$ x	x^2	Y	$(Y-\overline{Y})$ y	y^2	xy
1	−4	16	9	−3	9	+12
2	−3	9	8	−4	16	+12
3	−2	4	10	−2	4	+ 4
4	−1	1	12	0	0	0
5	0	0	11	−1	1	0
6	+1	1	13	+1	1	+1
7	+2	4	14	+2	4	+4
8	+3	9	16	+4	16	+12
9	+4	16	15	+3	9	+12
$\Sigma X = 45$	$\Sigma x = 0$	$\Sigma x^2 = 60$	$\Sigma Y = 108$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 57$

Regression Equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{45}{9} = 5;\ \overline{Y} = \frac{\Sigma Y}{N} = \frac{108}{9} = 12$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{57}{60} = 0.95$$

$$Y - 12 = 0.95\,(X - 5) = 0.95\,X - 4.75 \text{ or } Y = 7.25 + 0.95\,X$$

Regression Equation of X on Y : $X - \overline{X} = b_{xy}(Y - \overline{Y})$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{57}{60} = 0.95$$

$$X - 5 = 0.95\,(Y - 12) = 0.95Y - 11.4 \text{ or } X = -6.4 + 0.95Y$$

Graphing regression equations

From the regression equation of Y on X, we can estimate the most probable value of Y for various values of X and from the regression equation of X on Y, we can estimate the most probable values of X for various values of Y.

(*Estimated value of Y*)

$Y = 7.25 + 0.95X$

when $X = 1$, $Y = 7.25 + .95(1) = 8.20$

when $X = 2$, $Y = 7.25 + .95(2) = 9.15$

when $X = 3$, $Y = 7.25 + .95(3) = 10.10$

when $X = 4$, $Y = 7.25 + .95(4) = 11.05$

when $X = 5$, $Y = 7.25 + .95(5) = 12.00$

when $X = 6$, $Y = 7.25 + .95(6) = 12.95$

when $X = 7$, $Y = 7.25 + .95(7) = 13.90$

when $X = 8$, $Y = 7.25 + .95(8) = 14.85$

when $X = 9$, $Y = 7.25 + .95(9) = 15.80$

To plot the regression line of Y on X, we will take the actual values of X and estimated values of Y.

(*Estimated values of X*)

$X = -6.4 + 0.95Y$

when $Y = 9$, $X = .95(9) - 6.4 = 2.15$

when $Y = 8$, $X = .95(8) - 6.4 = 1.20$

when $Y = 10$, $X = .95(10) - 6.4 = 3.10$

when $Y = 11$, $X = .95(11) - 6.4 = 4.05$

when $Y = 12$, $X = .95(12) - 6.4 = 5.00$

when $Y = 13$, $X = .95(13) - 6.4 = 5.95$

when $Y = 14$, $X = .95(14) - 6.4 = 6.90$

when $Y = 15$, $X = .95(15) - 6.4 = 7.85$

when $Y = 16$, $X = .95(16) - 6.4 = 8.80$

Coefficient of Correlation :

We are given $b_{xy} = 0.95$ and $b_{yx} = 0.95$

$r = \sqrt{b_{yx} \times b_{xy}}$ or $r = \sqrt{.95 \times .95} = 0.95$

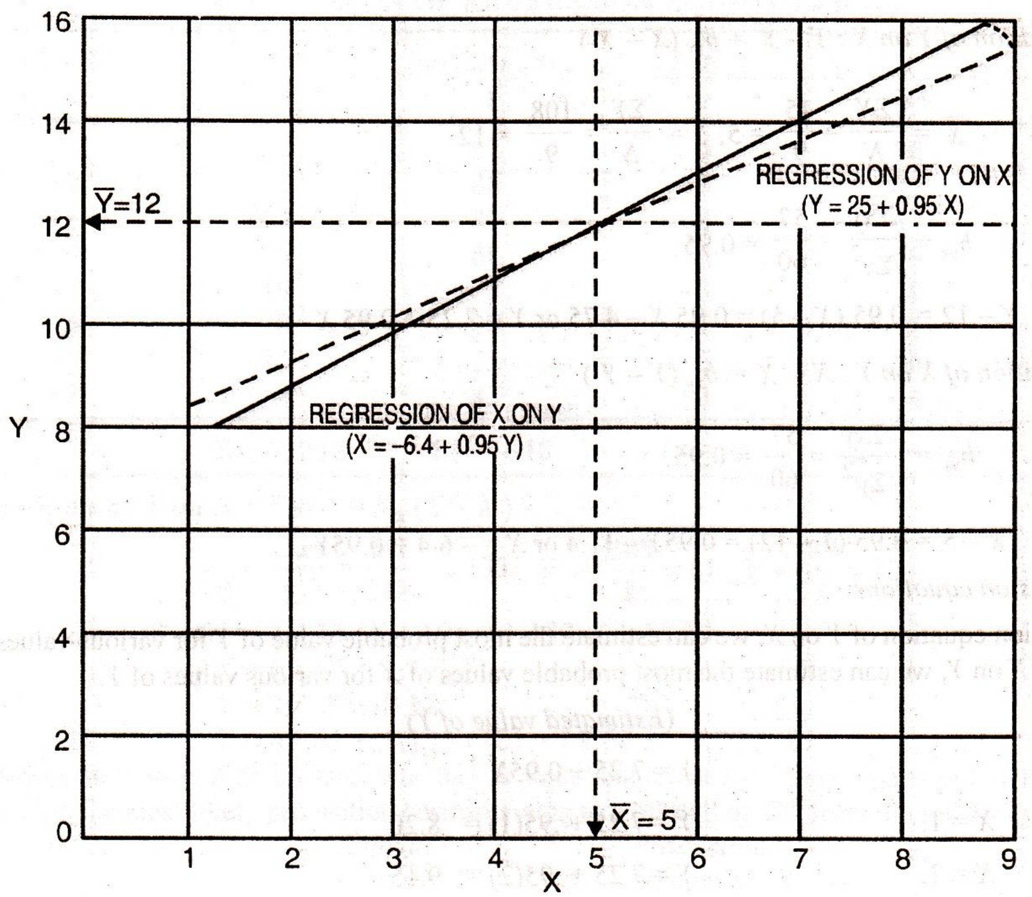


Illustration 17. The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of ready-made men's wears—is toying with the idea of increasing his sales to Rs. 80,000. On checking the records of sales during the last 10 years, it was found that the annual sale proceeds and advertisement expenditure were highly correlated to the extent of 0.8. It was further noted that the annual average sale has been Rs. 45,000 and annual average advertisement expenditure Rs. 30,000, with a variance of Rs. 1,600 and Rs. 626 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement you would suggest the General Sales Manager of the enterprise to incur to meet his target of sales. ***(MBA, Kurukshetra Univ., 2004)***

Solution. Let advertisement expenditure be denoted by X and sales by Y. We are required to find out the regression equation of Y on X given by the equation,

$$(Y - \overline{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}).$$

$$r = 0.8, \quad \sigma_x = 400, \quad \sigma_y = 25, \quad \overline{X} = 45{,}000, \quad \overline{Y} = 30{,}000$$

Substituting the values, we get

$$(Y - 30{,}000) = 0.8\ \frac{25}{400}\ (X - 45{,}000) = .05\ (X - 45{,}000)$$

$$Y = 30{,}000 + .05X - 2{,}250 = 27{,}750 + .05X$$

When $\quad X = 80{,}000$

$$Y = 27{,}750 + .05 \times 80{,}000 = 27{,}750 + 4{,}000 = 31{,}750$$

Hence the General Sales Manager should spend Rs. 31,750 to have the target sales of Rs. 80,000.

Illustration 18. Suppose that you are interested in using past expenditure on research and development by a firm to predict current expenditures on R & D. You got the following data by taking a random sample of firms, where X is the amount on R & D (in lakhs of rupees) 5 years ago and Y is the amount spent on R & D (in lakhs of rupees) in the current year :

X	:	30	50	20	80	10	20	20	40
Y	:	50	80	30	110	20	20	40	50

(*i*) Find the regression equation of Y on X.

(*ii*) If a firm is chosen randomly and $X = 10$, can you use the regression to predict the value of Y? Discuss.

(MBA, Madurai-Kamaraj Univ., 2000)

Solution. CALCULATION OF REGRESSION EQUATION

X	$(X-33)$ d_x	d_x^2	Y	$(Y-50)$ d_y	d_y^2	d_xd_y
30	–3	9	50	0	0	0
50	+17	289	80	+30	900	+510
20	–13	169	30	–20	400	+260
80	+47	2209	110	+60	3600	+2820
10	–23	529	20	–30	900	+690
20	–13	169	20	–30	900	+390
20	–13	169	40	–10	100	+130
40	+7	49	50	0	0	0
$\Sigma X = 270$	$\Sigma d_x = +6$	$\Sigma d_x^2 = 3592$	$\Sigma Y = 400$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 6800$	$\Sigma d_xd_y = 4800$

(*i*) *Regression equation of Y on X* : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{270}{8} = 33.75; \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{400}{8} = 50$$

$$b_{yx} = \frac{N\Sigma d_xd_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 4800 - 6 \times 0}{8 \times 3592 - (6)^2} = \frac{38400}{28700} = 1.338$$

$$Y - 50 = 1.338\,(X - 33.75) \text{ or } Y = 4.84 + 1.338X$$

(*ii*) When X is 10 : $\quad Y = 4.84 + 1.338\,(10) = 18.22$

For $\quad X = 10$, Y is 18.22.

Illustration 19. You are given the following information about advertising expenditure and sales :

	Adv. Exp. (*X*) (*Rs. lakhs*)	*Sales* (*Y*) (*Rs. lakhs*)
$\overline{X}$	10	90
σ	3	12

Correlation coefficient = 0.8

(*i*) Obtain the two regression equations.

(*ii*) Find the likely sales when advertisement budget is Rs. 15 lakhs.

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

(*MBA, Kumaun Univ., 2000; MBA, DU, 2002, MBA (HCA), DU, 2003*)

Solution. (*i*) *Regression equation of X on Y* : $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$\overline{X} = 10, r = 0.8, \sigma_x = 3, \sigma_y = 12, \overline{Y} = 90$$

$$X - 10 = 0.8\frac{3}{12}(Y - 90)$$

$$X - 10 = 0.2\,(Y - 90)$$

$$X - 10 = 0.2Y - 18 \quad \text{or} \quad X = -8 + 0.2Y$$

Regression equation of Y on X : $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

$$Y - 90 = .8\frac{12}{3}(X - 10)$$

$$Y - 90 = 3.2\,(X - 10) \text{ or } Y = 58 + 3.2X$$

(*ii*) By putting 15 in regression equation of Y on X, we can find out the likely sales,

$$Y = 58 + 3.2\,(15) = 58 + 48 = 106$$

Thus the likely sales for advertisement budget of Rs. 15 lakhs is Rs. 106 lakhs.

(*iii*) By putting 120 in regression equation of X on Y, we can find what should be the advertisement budget.

$$X = -8 + 0.2\ (120) = 16$$

Thus for attaining sales target of Rs. 120 lakhs, the advertisement budget should be Rs. 16 lakhs.

Illustration 20. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random :

Aptitude scores :	60	62	65	70	72	48	53	73	65	82
Productivity index :	68	60	62	80	85	40	52	62	60	81

Estimate (*i*) the productivity index of a worker whose test score is 92, (*ii*) the test score of a worker whose productivity index is 75.

(MBA, Delhi Univ., 2001; MBA, Hyderabad, Univ., 2004)

Solution. Since productivity depends on aptitude scores, let Y denote the productivity and X the aptitude score.

CALCULATION OF REGRESSION EQUATIONS

Aptitude Score X	*(X–65)* $\bar{X} = 65$ x	x^2	*Productivity Index* Y	*(Y–65)* $\bar{Y} = 65$ y	y^2	xy
60	–5	25	68	+3	9	–15
62	–3	9	60	–5	25	+15
65	0	0	62	–3	9	0
70	+5	25	80	+15	225	+75
72	+7	49	85	+20	400	+140
48	–17	289	40	–25	625	+425
53	–12	144	52	–13	169	+156
73	+8	64	62	–3	9	–24
65	0	0	60	–5	25	0
82	+17	289	81	+16	256	+272
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 894$	$\Sigma Y = 650$	$\Sigma y = 0$	$\Sigma y^2 = 1752$	$\Sigma xy = 1044$

For answering part (*i*) of the question we have to fit a regression equation of Y on X.

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

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{650}{10} = 65$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1044}{894} = 1.168$$

$$Y - 65 = 1.168\ (X - 65)$$

$$Y - 65 = 1.168\ X - 75.92 \quad \text{or} \quad Y = 1.168\ X - 10.92$$

$$Y_{92} = 1.168\ (92) - 10.92 = 107.456 - 10.92 = 96.536$$

For answering part (*ii*) of the question we have to fit a regression equation of X on Y.

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

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{1044}{1752} = 0.596$$

$$X - 65 = 0.596\ (Y - 65) \quad \text{or} \quad X - 65 = 0.596Y - 38.74$$

$$X = 0.596Y + 26.26$$

$$Y_{75} = .596\ (75) + 26.26 = 44.7 + 26.26 = 70.96.$$

Illustration 21. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :

Variance of $X = 9$

Regression equation $8X - 10Y + 66 = 0$

$40X - 18Y = 214$

Find on the basis of the above information :

(*i*) The mean values of X and Y,

(*ii*) Coefficient of correlation between X and Y, and

(*iii*) Standard deviation of Y. (*MBA, Pune Univ., 2002; MBA, Anna Univ., 2003*)

Solution. (*i*) *Calculating mean values of X and Y*

$$8X - 10Y = -66 \quad \ldots(i)$$

$$40X - 18Y = 214 \quad \ldots(ii)$$

Multiplying eq. (*i*) by 5

$$40X - 50Y = -330$$

$$40X - 18Y = 214$$

$$- \quad + \quad -$$

$$-32Y = -544$$

$$Y = 17 \quad \text{or } \overline{Y} = 17$$

Putting the value of Y in eq. (*i*)

$$8X - 10\,(17) = -66$$

$$8X = -66 + 170$$

$$8X = 104 \text{ or } X = 13 \text{ or } \overline{X} = 13$$

(*ii*) *Coefficient of correlation between X and Y*

For finding the value of r, we have to determine the value of regression coefficients. Since we don't know which equation is regression of X on Y and which is of Y on X, we have to make an assumption. Assuming eq. (*i*) as the regression of X on Y.

$$8X = 10Y - 66$$

$$X = -\frac{66}{8} + \frac{10}{8}Y \quad \text{or } b_{xy} = \frac{10}{8}$$

From eq. (*ii*) $-18Y = 214 - 40X$

$$Y = -\frac{214}{18} + \frac{40}{18}X \text{ or } b_{yx} = \frac{40}{18}$$

Since both the regression coefficients are greater than 1, our assumption is wrong. Hence eq. (*i*) is regression eq. of Y on X.

$$-10\,Y = -66 - 8X$$

$$Y = \frac{66}{10} + \frac{8}{10}X \quad or \quad b_{yx} = \frac{8}{10}$$

From eq. (*ii*) $40X = 214 + 18Y$

$$X = \frac{214}{40} + \frac{18}{40}Y \quad or \quad b_{xy} = \frac{18}{40}$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{18}{40} \times \frac{8}{10}} = \sqrt{0.36} = 0.6$$

(*iii*) The value of standard deviation of Y can be determined from any regression coefficient.

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

$$b_{xy} = \frac{18}{40}, \; r = .6, \; \sigma_x = \sqrt{9} = 3$$

Substituting the values

$$\frac{18}{40} = .6\frac{3}{\sigma_y} \quad \text{or} \quad 18\sigma_y = 72 \text{ or } \sigma_y = 4.$$

Illustration 22. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8, the average of husbands age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations :

(*a*) the expected age of husband when wife's age is 16 years, and

(*b*) the expected age of wife when husband's age is 33 years. (*MBA, Osmania Univ., 2000*)

Solution. Let age of wife be denoted by Y and age of husband by X. We are given

$$\overline{X} = 25, \quad \overline{Y} = 22, \quad \sigma_x = 4, \quad \sigma_y = 5, \quad r = 0.8$$

For answering part (*a*) we have to fit a regression equation X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$$X - 25 = .8\frac{4}{5}(Y - 22) \text{ or } X - 25 = .64\,(Y - 22)$$

$$X - 25 = .64Y - 14.08 \text{ or } X = 10.92 + 0.64Y$$

When $\quad Y = 16,\ X = 10.92 + 0.64\,(16) = 10.92 + 10.24 = 21.16$

Thus, the expected age of husband when wife's age is 16 years shall be 21.16 years.

For answering part (*b*) we have to fit a regression equation of Y on X.

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$Y - 22 = .8\,\frac{5}{4}\,(X - 25)$$

$$Y - 22 = (X - 25) \text{ or } Y = -3 + X; \text{ when } X = 33,$$

$$Y = -3 + 33 = 30$$

Thus, the expected age of wife when husband's age is 33 is 30 years.

Illustration 23. The following data relate to marks obtained by 250 students in Accountancy and Statistics in an examination of a university :

Subject	*Arithmetic Mean*	*Standard Deviation*
Accountancy	48	4
Statistics	55	5

Coefficient of correlation between marks in accountancy and statistics is +0.8. Find the two regression equations and estimate the marks obtained by a student in Statistics who secured 50 marks in Accountancy. (*M.Com., Sukhadia Univ., 2001*)

Solution. Let marks in accountancy be denoted by X and in statistics by Y.

Regression equation of X on Y

$$X - \overline{X} = r\,\frac{\sigma_x}{\sigma_y}\,(Y - \overline{Y})$$

$$\overline{X} = 48,\ \overline{Y} = 55,\ \sigma_x = 4,\ \sigma_y = 5,\ r = 0.8$$

$$X - 48 = .8\,\frac{4}{5}\,(Y - 55)$$

$$X - 48 = .64\,(Y - 55)$$

$$X = .64Y + 12.8$$

Regression equation of Y on X

$$Y - \overline{Y} = r\,\frac{\sigma_y}{\sigma_x}\,(X - \overline{X})$$

$$Y - 55 = .8\,\frac{5}{4}\,(X - 48)$$

$$Y - 55 = (X - 48) \text{ or } Y = 7 + X$$

If marks in accountancy, *i.e.*, X is 50; the marks in statistics shall be 57.

Illustration 24. The following figures relate to length of service and income of the employees of an organisation :

Length of Service (Years)	:	11	7	2	5	8	6	10
Income (Rs. hundred)	:	7	5	3	2	6	4	8

Compute the coefficient of correlation for the above data. Find the two regression equations and examine the relationship.

Solution. Let length of service be denoted by X and income by Y.

CALCULATION OF REGRESSION EQUATIONS AND CORRELATION COEFFICIENT

X	$(X-7)$ x	x^2	Y	$(Y-5)$ y	y^2	xy
11	+4	16	7	+2	4	+8
7	0	0	5	0	0	0
2	−5	25	3	−2	4	+10
5	−2	4	2	−3	9	+6
8	+1	1	6	+1	1	+1
6	−1	1	4	−1	1	+1
10	+3	9	8	+3	9	+9
$\Sigma X = 49$	$\Sigma x = 0$	$\Sigma x^2 = 56$	$\Sigma Y = 35$	$\Sigma y = 0$	$\Sigma y^2 = 28$	$\Sigma xy = 35$

Regression equation of X on Y : $X - \overline{X} = b_{xy}(Y - \overline{Y})$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{49}{7} = 7; \quad \overline{Y} = \frac{35}{7} = 5$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{35}{28} = 1.25$$

$$X - 7 = 1.25\,(Y - 5)$$

$$X - 7 = 1.25Y - 6.25 \text{ or } X = 0.75 + 1.25Y$$

Regression equation of Y on X : $Y - \overline{Y} = b_{yx}(X - \overline{X})$

$$b_{yx} = \frac{\Sigma xy}{\Sigma y^2} = \frac{35}{56} = 0.625$$

$$X - 5 = .625\,(X - 7)$$

$$X - 5 = .625X - 4.375 \text{ or } X = 0.625 + 0.625\,X$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{1.25 \times .625} = 0.884$$

Thus, there is a high degree of positive correlation between length of service and experience.

Illustration 25. In a correlation study the following values are obtained :

	X	Y
Mean	65	67
S.D.	2.5	3.5
Coefficient of Correlation	$r = 0.8.$	

Find the two regression equations. *(M.Com., Madurai-Kamaraj Univ., 2007)*

Solution : *Regression equation of X on Y* :

$$X - \overline{X} = r\,\frac{\sigma_x}{\sigma_y}\,(Y - \overline{Y})$$

$$\overline{X} = 65,\ \sigma_x = 2.5,\ \sigma_y = 3.5,\ r = 0.8,\ \overline{Y} = 67$$

$$X - 65 = 0.8\,\frac{2.5}{3.5}\,(Y - 67)$$

$$X - 65 = 0.571\,(Y - 67)$$

$$X - 65 = 0.571\,Y - 38.26$$

$$X = 0.571Y + 26.74$$

Regression equation of Y on X :

$$Y - \overline{Y} = r\,\frac{\sigma_y}{\sigma_x}\,(X - \overline{X})$$

$$Y - 67 = 0.8\,\frac{3.5}{2.5}\,(X - 65)$$

$$Y - 67 = 1.12\,(X - 65)$$

$$Y - 67 = 1.12\,X - 72.8$$

$$Y = 1.12\,X - 5.8$$

Illustration 26. In trying to evaluate the effectiveness in its advertising campaign, a firm compiled the following information :

Year	2003	2004	2005	2006	2007	2008	2009	2010
Adv. Expenditure ('000 Rs.)	12	15	15	23	24	38	42	48
Sales (Rs. lakh)	5.0	5.6	5.8	7.0	7.2	8.8	9.2	9.5

Calculate the regression equation of sales on advertising expenditure. Estimate the probable sales when advertisement expenditure is Rs. 60 thousand.

Solution : CALCULATION OF REGRESSION EQUATION

X	$(X-24)$ d_x	d_x^2	Y	$(Y-7.0)$ d_y	d_y^2	d_xd_y
12	−12	144	5.0	−2.0	4.00	24.0
15	−9	81	5.6	−1.4	1.96	12.6
15	−9	81	5.8	−1.2	1.44	10.8
23	−1	1	7.0	0	0	0
24	0	0	7.2	+ 0.2	.04	0
38	+14	196	8.8	+1.8	3.24	25.2
42	+18	324	9.2	+ 2.2	4.84	39.6
48	+24	576	9.5	+2.5	6.25	60.0
$\Sigma X = 217$	$\Sigma d_x = 25$	$\Sigma d_x^2 = 1403$	$\Sigma Y = 58.1$	$\Sigma d_y = 2.1$	$\Sigma d_y^2 = 21.77$	$\Sigma d_xd_y = 172.2$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{217}{8} = 27.125; \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{58.1}{8} = 7.26$$

Regression equation of sales on advertisement expenditure is given by :

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

where
$$b_{yx} = \frac{N\Sigma d_xd_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{8(172.2) - (25)(2.1)}{8(1403) - (25)^2} = \frac{1377.6 - 52.5}{11224 - 625} = \frac{1325.1}{10599}$$
$$= 0.125$$

Substituting the values, we have

$$Y - 7.2625 = 0.125(X - 27.125)$$
$$Y - 7.2625 = 0.125X - 3.3906$$
$$Y = 3.8719 + 0.1250X$$

When $X = 60$, the estimated value of Y shall be :

$$Y = 3.8719 + 0.1250(60) = 3.8719 + 7.5 \approx 11.37.$$

Illustration 27. A resarch company summarized advertising expenditure and sales results as follows :

	Ad. Expenditure (Rs. crore)	*Sales (Rs. crore)*
Mean	20	200
S.D.	18	170
r	= 0.6.	

Derive two regression equations. *(MBA, GGDIP Univ., 2009)*

Solution : Since sales depend on advertisement expenditure, we take sales as Y and advertisement expenditure as X.

Regression equation of X on Y :

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$
$$\overline{X} = 20, \sigma_x = 18, \sigma_y = 170, r = 0.6, \overline{Y} = 200$$

$$X - 20 = 0.6\frac{18}{170}(Y - 200)$$
$$X - 20 = 0.64(Y - 200)$$
$$X - 20 = 0.64Y - 12.8$$
$$X = 0.64Y + 7.2$$

Regression equation of Y on X :

$$Y - \overline{Y} = r\,\frac{\sigma_y}{\sigma_x}\,(X - \overline{X})$$

$$\overline{Y} = 200,\ \sigma_y = 170,\ \sigma_x = 18,\ r = 0.6,\ \overline{X} = 20$$

$$Y - 200 = 0.6\,\frac{170}{18}\,(X - 20)$$

$$Y - 200 = 5.667\,(X - 20)$$

$$Y - 200 = 5.667\,X - 113.34$$

$$Y = 5.667\,X - 86.66$$

PROBLEMS

Answer the following questions, each question carries **one** mark:

(*i*) What is regression ?
(*ii*) What is the use of studying regression ?
(*iii*) When will regression coefficients become coefficient of correlation ? (*MBA, Madurai-Kamaraj Univ., 2003*)
(*iv*) Write down the two regression equations.
(*v*) Write down the formula for regression coefficient of x and y ?
(*vi*) What do you understand by the term 'regression line' ? (*M.Com., M.K.Univ., 2003*)
(*vii*) What are regression coefficients ?
(*viii*) Can both the regression coefficients exceed one ?
(*ix*) Are regression coefficients independent of change of scale and origin or only origin ?
(*x*) In the regression equation of y on x how do you interpret the values of 'a' and 'b'?
(*xi*) Who had coined the term 'regression' ?

Answer the following questions, each question carries **four** marks:

(*i*) Distinguish between 'correlation' and 'regression analysis'. Why there are two regression lines? (*MBA, UP Tech. Univ., 2007*)
(*ii*) What are regression coefficients ? How do you interpret them ?
(*iii*) What are the important characteristics of regression coefficients ?
(*iv*) If two regression coefficients are –1.2 and – 0.8, what would be the value of r ?
(*v*) What are the important uses of regression analysis ?

(*a*) Explain the concept of regression and point out its usefulness in dealing with business problems.
(*b*) Distinguish between correlation and regression. Also point out the properties of regression coefficients.

(*a*) Compare and contrast the role of correlation and regression in studying the interdependence of two variates.
(*b*) Explain the concept of regression and point out its importance in business forecasting.

Under what conditions can there be one regression line? Explain.

"The regression line gives only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in this estimate by calculating a quantity known as the standard error of estimate". Elucidate.

Do you agree with the view that regression equations are irreversible, *i.e.*, we cannot find out the regression of X on Y from that of Y on X?

(*a*) Point out the usefulness of regression analysis in business and industry.
(*b*) What is linear regression? When is it used? (*MBA, Madurai-Kamaraj Univ., 2003*)
(*c*) Discuss the role of correlation and regression analysis in business. Illustrate.

What are regression lines ? With the help of an example, illustrate how they help in business decision-making. (*MBA, Delhi Univ., 2004*)

What do you understand by the term "regression analysis"? Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients? (*MBA, Osmania Univ.; MBA, Delhi Univ., 2006*)

(*a*) Write any two differences between correlation and regression. (*M.Com., Madras univ., 2009*)
(*b*) What are regression coefficients? State some of the important properties of regression coefficients.
(*c*) Write down the mathematical properties of Correlation Coefficient and Regression Coefficient. (*MBA, Hyderabad Univ., 2005*)
(*d*) State the utility of regression in economic analysis.

The following data give the hardness (X) and tensile strength (Y) of 7 samples of metal in certain units. Find the linear regression equation of Y on X.

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

[$Y = 29.45 + 0.31X$]

12. The average daily wage for working class in Nagpur is Rs. 12 and for that in Delhi Rs. 18, their respective standard deviations are Rs. 2 and Rs. 3 and the coefficient of correlation is 0.67. Find the most likely wage in Delhi corresponding to the wage of Rs. 20 in Nagpur.
[$Y_{20} = 26.04$]

13. There are two series of index numbers D for disposable personal income and S for a salary of the company. The mean and standard deviations of the D series are 120 and 15 respectively and of the S series 115 and 10. The coefficient of correlation between the two series is 0.75. From the given information obtain a linear equation for estimating the values of S for different values of D. How will you interpret the values of S corresponding to different values of D obtained from the equation? Can the same equation be used for estimating values of D for different values of S?
[$S = 0.5$; $D = 55$; No]

14. The following calculations have been made for closing prices of 12 stocks (X) on the Bombay Stock Exchange on a certain day along with the volume of sales in thousand of shares (Y). From these calculations find the regression equations.
$\Sigma X = 580$, $\Sigma Y = 370$, $\Sigma XY = 11{,}494$
$\Sigma X^2 = 41{,}658$, $\Sigma Y^2 = 17{,}206$
[$Y = 53.55 - 0.47X$, $X = 79.16 - 1.1Y$]

15. Given the following data, what will be the possible yield when the rainfall is 29"?

	Rainfall	*Production*
Mean	29"	40 units per acre
S.D.	3"	6 units per acre

Coefficient of correlation between rainfall and production = 0.8.
[40 units]

16. In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales:

Salesmen :	1	2	3	4	5	6	7	8	9	10
Test Scores :	45	75	50	60	80	90	85	40	80	55
Sales ('000) :	2.0	6.5	3.5	5.0	4.5	6.0	6.5	2.5	5.5	4.5

Calculate the regression line of sales on test score and estimate the most probable weekly sales volume if a salesman makes a score of 70.
[$Y = -0.541 + 0.078X$, 4919]

17. The following marks have been obtained by a group of students in Statistics (out of 100):

Paper I :	80	45	55	56	58	60	65	68	70	75	85
Paper II :	82	56	50	48	60	62	64	65	70	74	90

Compute the coefficient of correlation for the above data. Find the lines of regression and examine the relationship.
[$r = 0.75$, $Y = -1 + 0.75\,X$, $X = 4.25 + 0.75\,Y$]

18. The following table gives marks out of 50 awarded in a French and a German test to the same group of boys. Assume there is a linear relation between the sets of marks, calculate the equations of the lines of regression.

French :	10	10	18	25	28	33	34	39	42	43
German :	11	22	22	19	35	27	33	40	42	47

[$Y = 6.25 + 0.13\,X$, $X = -0.34 + 0.96\,Y$]

19. You are given the following result of the height (X) and weight (Y) of 1,000 managers:

Mean (X)	= 68.00"
Mean (Y)	= 150 lbs
Standard deviation (X)	= 2.50"
Standard deviation (Y)	= 20 lbs

Coefficient of correlation between X and $Y = 0.6$. Estimate from the above data the height of a manager whose weight is 200 lbs. *(MBA, Kurukshetra Univ., 2002)*

20. The following table shows the mean and standard deviation of the prices of two shares on a stock exchange :

Shares	*Mean* (*in Rs.*)	*Standard deviation* (*in Rs.*)
A Ltd.	39.5	10.8
B Ltd.	47.5	16.8

If the coefficient of correlation between the prices of two shares is 0.42, find the most likely price of share A corresponding to a price of Rs. 55 observed in the case of share B.

21. Catalogues listing textbooks were examined to discover the relationship between the cost of a book and number of pages it contains. The perusal gives the following data for ten books:

Pages :	700	540	210	625	380	910	610	420	750	400
Price (Rs.) :	12	11	5	10	7	15	9	8	12	9

(*a*) Obtain the line of regression for estimating the price of a book.
(*b*) What is your estimate for the price of a book containing 500 pages ?
(*c*) What increase would you expect for a book if it is decided to increase the number of pages of the book by 100 ?
(*d*) Calculate the standard error of the estimate.

22. From the data given below find the two regression equations.

Age of wife	*Age of Husband* 20–25	25–30	30–35	Total
16–20	4	9	—	13
20–24	1	4	1	6
24–28	4	4	3	11
Total	9	17	4	30

(*M.Phil, Kurukshetra Univ., 2003*)

23. The data given below relate to the scores obtained by 9 salesmen in an intelligence test and their weekly sales, in lakh of rupees :

Salesman :	1	2	3	4	5	6	7	8	9
Test Score :	50	60	50	60	80	50	80	40	70
Sales (Rs. lakh) :	3	6	4	5	6	3	7	5	6

Obtain regression equation of sales on the intelligence test scores. If a salesman has obtained a score of 65, what would be his expected weekly sales ?
[$Y = 0.075X + 0.5$, Rs. 5.375 lakh]

24. The following figures relate to advertisement expenditure and sales :

Adv. Exp. (in lakh of Rs.) :	60	62	65	70	73	75	71
Sales (in crore of Rs.) :	10	11	13	15	16	19	14

Estimate (*i*) the sales for advertisement expenditure of Rs. 80 lakh and (*ii*) the advertisement for a sales target of Rs. 25 crore.
[20.1; 87.75]

25. You are given the following data about the sales and advertisement expenditure of a firm :

	Sales (Rs. crore)	*Advertisement Expenditure (Rs. crore)*
Arithmetic Mean	50	10
Standard Deviation	10	2
Coefficient of Correlation	+0.9	

(*a*) Calculate the two regression equations.
(*b*) Estimate the likely sales for a proposed advertisement expenditure of Rs. 13.5 crore.
(*c*) What should be the advertisement budget if the company wants to achieve a sales target of Rs. 70 crore ?

(*MBA, Delhi Univ., 2005*)

[(*a*) $Y = 4.5X + 5$, $X = .18Y + 1$. (*b*) 65.75 crore. (*c*) 13.6 crore]

26. The following bivariate frequency distribution relates to sales turnover (in lakh Rs.) and money spent on advertising budget (in thousand Rs.). Obtain the two regression equations.

Sales Turnover (in lakh Rs.)	*Advertising budget (in thousand Rs.)* 50–60	60–70	70–80	80–90
25–50	2	1	2	5
50–75	3	4	7	6
75–100	1	5	8	6
100–125	2	7	9	2

Estimate (*i*) the sales turnover corresponding to advertising budget of Rs. 150 thousand, (*ii*) the advertising budget to achieve a sales turnover of Rs. 200 lakh, and (*iii*) compute the coefficient of correlation.

(*MBA, Delhi Univ., 2008*)

27. The following data give the test scores and sales made by nine salesmen during the last one year :

Test Scores :	14	19	24	21	26	22	15	20	19
Sales ('000 Rs.) :	31	36	48	37	50	45	33	41	39

Obtain (*i*) the regression equation of test scores on sales, (*ii*) the regression equation of sales on test scores, and (*iii*) coefficient of correlation.
[$X = -2.312 + 0.5578\,Y$, (*ii*) $Y = 7.834 + 1.6083\,X$, (*iii*) $r = 0.947$]

28. A study of share prices of Textile group and Fertiliser group of companies yielded the following results :

	Textiles	*Fertilisers*
Mean	12.8	985.0
Standard Deviation	1.6	70.1
Coefficient of Correlation	+0.52	

The financial expert has estimated the likely price of textiles shares at the close of the next accounting year as 92. What would be your estimate of the likely price of fertiliser shares at the corresponding time ?

29. Following are the data on business turnover and staff of a company for eight years from 2003 to 2010 :

	2003	2004	2005	2006	2007	2008	2009	2010
Business Turnover (Rs. crore) :	45	50	60	75	80	110	150	170
Staff :	2,600	3,000	3,100	3,530	3,850	4,300	5,870	7,150

Fit a proper regression equation to estimate manpower in terms of business turnover. Estimate the staff requirement when the business turnover reaches Rs. 200 crore.

[$Y = 33.24X + 1100.3$; 7748.3]

30. The data on sales and promotion expenditure on a product for 10 years are given below :

Sales (Rs. lakh) :	8	10	9	12	10	11	12	13	14	15
Promotion Exp. (Rs. thousand) :	2	2	3	4	5	5	5	6	7	8

Use two-variable regression model to estimate the effect of promotion on sales. Forecast the sales for next year when the company hopes to spend Rs. 10 thousand on promotion.

[$X = 0.815\ Y - 4.591$, $Y = 1.003X + 6.686$, $Y_{10} = 16.716$]

31. Table below shows the power and top speeds of different brands of sports cars :

Brand :	*A*	*B*	*C*	*D*	*E*	*F*
Power X [kW] :	70	63	72	60	66	70
Speed Y [km/h] :	155	150	180	135	156	168
Brand :	*G*	*H*	*I*	*J*	*K*	*L*
Power X [kW] :	74	65	62	67	65	68
Speed Y [km/h] :	178	160	132	145	139	152

(*i*) Find the best linear relationship that fits the given data.
(*ii*) Estimate the speed of a car that has a power of 63 kW and find a 95% confidence interval for this estimate.
(*iii*) Determine how much of the variability in speed may be explained by the regression hypothesis.

32. Calculate the coefficient of correlation from the following data :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Also obtain the regression equations and find an estimate of Y which should correspond on an average to $X = 6.2$.

(*MBA, Madurai-Kamaraj Univ., 2006*)

[$Y = 0.95X + 7.25$; $Y_{6.2} = 13.14$]

33. Family income and its percentage spent on food gave the following bivariate frequency table :

Food Expenditure (in%)	*Monthly Family Income (in hundred Rs.)* 25–35	35–45	45–55	55–65	65–70
15–20	8	9	12	13	8
20–25	6	3	6	11	14
25–30	—	7	9	—	4
30–35	5	8	10	14	13

(*i*) Estimate the family income for a food expenditure of 40%.
(*ii*) What amount should be spent on food expenditure for a monthly family income of Rs. 10,000.
(*iii*) Compute coefficient of correlation.

34. You are given below the following information about advertisement and sales.

	Adv. Exp. (X) *(Rs. crore)*	*Sales (Y)* *(Rs. crore)*
Mean	20	120
S.D.	5	25
Correlation coefficient	+0.8	

(*i*) Calculate the two regression equations.
(*ii*) Find the likely sales when advertisement expenditure is Rs. 25 crore.
(*iii*) What should be the advertisement budget if the company wants to attain sales target of Rs. 150 crore ?

[$Y = 4X + 40$; $X = 0.16Y + 0.8$; $Y_{25} = 140$; $X_{150} = 24.8$]

35. From the following data obtain the regression equation. Also find the correlation coefficient with the help of regression coefficient :

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

[$Y = 11.9 - 0.65\,X$; $X = 16.4 - 1.3Y$, $r = -0.919$]

36. The monthly expenditure on advertisement and sales of a firm are given for 2010. It is generally found that expenditure on advertisement has its impact after two months. Allowing for time lag:
(a) calculate the correlation between expenditure on advertisement and sales.
(b) estimate the sales of the firm in February 2015.

Months/Year (2010)	*Expenditure on Advertisement (Rs.)*	*Sales (Rs.)*
January	50	1200
February	60	1500
March	70	1600
April	90	2000
May	120	2200
June	150	2500
July	140	2400
August	160	2600
September	170	2800
October	190	2900
November	200	3100
December	250	3900

37. The following figures relate to advertisement expenditure and sales :

Advertisement (in Rs. lakh) :	60	62	65	70	73	75	71
Sales (in Rs. crore) :	10	11	13	15	16	19	14

Estimate (*i*) the sales for advertisement expenditure of Rs. 80 lakh; and (*ii*) the advertisement expenditure for a sales target of Rs. 25 crore.

38. Given the regression equation of Y on X and X on Y are respectively $Y = 2X$ and $6X - Y = 4$ and the second moment of X about the origin is 3. Find (*i*) the correlation coefficient, and (*ii*) standard deviation of Y.

39. Find the regression coefficient of Y on X from the following regression equations :

$$5X = 22 + Y$$
$$64X = 24 + 45Y$$

Is it possible to calculate the standard deviation of Y from the given information ? Answer with reason.

40. A financial analyst has gathered the following data about the relationship between income and investment in securities in respect of 8 randomly selected families :

Income (Rs. '000) :	8	12	9	24	143	37	19	16
Per cent invested in securities :	36	25	33	15	28	19	20	22

(a) Develop an estimating equation that best describes these data.
(b) Find the coefficient of determination and interpret it.
(c) Calculate the standard error of estimate for this relationship.
(d) Find an approximate 90 per cent confidence interval for the percentage of income invested in securities by a family earning Rs. 25,000 annually.

41. From the data given below find :
(*i*) The two regression equations.
(*ii*) The coefficient of correlation between marks in Economics and Statistics.
(*iii*) The most likely marks in Statistics when the marks in Economics are 30.

Marks in Economics (X) :	25	28	35	32	31	36	29	38	34	32
Marks in Statistics (Y) :	43	46	49	41	36	32	31	30	33	39

42. A financial analyst obtained the following information relating to return on security A and that of market portfolio M for the past 8 years :

Year :	1	2	3	4	5	6	7	8
Return on A :	10	15	18	14	16	16	18	4
Return on M :	12	14	13	10	9	13	14	7

(*i*) Develop an estimating equation that best describes these data.
(*ii*) Find the coefficient of determination and interpret it.
(*iii*) Determine the percentage of total variation in security return being explained by the return on the market portfolio.

(*MFC, Delhi Univ., 2005*)

43. Given the bivariate data :

X	:	1	5	3	2	1	1	7	3
Y	:	6	1	0	0	1	2	1	5

(*i*) Fit a regression equation of Y on X.

(*ii*) If a person has scored 8 on X variable, what would be his score on Y variable ?

44. Personnel Manager of a large industrial unit is interested to find a measure that can be used to fix the wages (yearly) of skilled workers. On experimental basis, the data on the length of service and their yearly wages (in Rs. '000) from a group of 10 randomly selected skilled workers are given below :

Length of service (X)	:	11	7	9	5	8	6	10	12	3	4
Yearly wages (Y)	:	14	11	10	9	13	10	14	16	6	7

(*a*) Develop the regression equation of wage (Y) on the length of service X.

(*b*) On the basis of (*a*) what initial pay the personnel manager should give to a skilled worker who has put in thirteen years of service on a similar basis, in another industry.

[$Y = 3.455 + 1.006\,X$; $Y = 16.533$] *(DIM, IGNOU, 2000)*

45. In a laboratory experiment on correlation research study, the equation to the two regression lines were to be $2X - Y + 1 = 0$ and $3X - 2Y + 7 = 0$. Find (*i*) the means of X and Y. Also work out the values of the regression coefficients and the coefficient of correlation between the two variables X and Y.

[$\overline{X} = 5$, $\overline{Y} = 11$; $bxy = 0.5$, $byx = 1.5$; $r = 0.866$]

46. An industrial engineer collected the following data on experience & performance rating of 8 operators :

Operators	:	1	2	3	4	5	6	7	8
Experience (years)	:	16	12	18	4	3	10	5	12
Performance Rating	:	87	88	89	68	58	80	70	85

(*a*) Does the data give evidence that experience improves performance ?

(*b*) Estimate the performance rating of an operator having (*a*) 9 years and (*b*) 15 years of experience.

[$Y = 69.67 + 1.133\,X$] *(MBA, Kumaun Univ., 2002)*

47. The following table gives the age of cars of certain make and the annual maintenance costs. Find (*i*) the coefficient of correlation between the variables and (*ii*) Regression equation for costs related to age.

Age of Cars (in years)	:	2	4	6	8
Maintenance costs (in hundred Rs.)	:	0	20	25	30

(MBA, HPU, 2002)

48. A firm administers a test to sales trainees before they go into the field. The management of the firm is interested in determining the relationship between the test scores and the sales made by the trainees at the end of one year in the field. The following data were collected for ten sales personnel who have been in the field for one year :

Sales Person Number	*Test Score*	*Number of Units Sold*
1	2.6	95
2	3.7	140
3	2.4	85
4	4.5	180
5	2.6	100
6	5.0	195
7	2.8	115
8	3.0	136
9	4.0	175
10	3.4	150

(*i*) Find the regression line which would be used to predict sales from trainees test scores.

(*ii*) Predict the number of units which would be sold by trainee who received an average test score. *(MBA, DU, 2001)*

49. For the data given below :

	Average	*S.D.*
Production (in units)	35	10
Capacity utilisation (%)	85	8
Coefficient of correlation	0.6	

Obtain the two regression equations.

Estimate the production when the capacity utilisation is 70 per cent. *(MBA, D.U., 2003)*

50. Explain why there are two regression lines ? What happens if the two lines are identical ? For the data given below, find the relevant line of regression to estimate the price, if supply is 25 million tonnes.

Supply (m.t.)	:	5	10	12	15	18
Price (Rs./kg.)	:	16	15	12	12	10

(M. Com., A.M.U., 2001)

51. The following table shows the ages (X) and blood pressure (Y) of 8 persons :

X	:	52	63	45	36	72	65	47	25
Y	:	62	53	51	25	79	43	60	33

Obtain the regression equation of Y on X and find out the expected blood pressure of a person who is 49 years old.

(M.Com., Madurai-Kamaraj Univ., 2008)

52. Determine the equation of the straight line which best fits the following data :

X	Y
10	19
12	22
13	24
16	27
17	29
20	33
25	37

(MBA., IGNOU, 2005)

53. Regression calculations were carried out as follows :

$\sum X = 32, \sum Y = 24, \sum XY = 218$

$\sum X^2 = 296, \sum Y^2 = 162.5, n = 4$

Find the lines of regression and coefficient of correlation and comment. *(MBA, M.D. Univ., 2000)*

54. From the following data obtain the two regression equations :

Sales	:	91	97	103	121	67	124	52	73	111	57
Purchases	:	97	75	69	97	70	91	39	61	83	47

(MBA, Madurai Kamaraj Univ., Nov. 2001)

55. Obtain the regression of Y on X and X on Y from the following data and estimate the blood pressure when the age is 50.

Age	*Blood Pressure*	*Age*	*Blood Pressure*
50	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	150
47	128	60	155

(MBA, Bharathidasan Univ., 2001)

56. From the data given below, find the two regression equations and the most likely marks in statistics when marks in Economics are 30.

Marks in Economics	:	25	28	35	32	31	36	24	38	34	32
Marks in Statistics	:	43	46	49	41	36	32	31	30	33	39

(MBA, M.K. Univ., 2003)

57. Cost accountants often estimate overheads based on the level of production. At BFL company, the data collected are as follows. Find the best fit equation between production and overhead costs. Predict overheads when 50 units are produced.

Overhead	:	191	170	272	155	280	173	234	116	153	178
Production units	:	40	42	53	35	56	39	48	30	37	40

(MBA, Bharathidasan Univ., 2007)

*****