In this chapter, the following methods of finding the correlation coefficient between two variables x and y are discussed:

1. Scatter Diagram method
2. Karl Pearson's Coefficient of Correlation method
3. Spearman's Rank Correlation method
4. Method of Least-squares

Figure 13.1 shows how the strength of the association between two variables is represented by the coefficient of correlation.

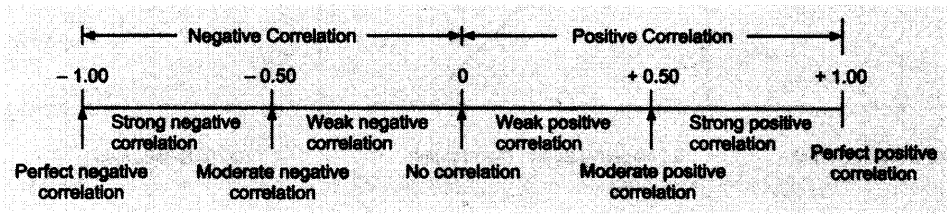


Figure 13.1
Interpretation of Correlation Coefficient

13.5.1 Scatter Diagram Method

The **scatter diagram** method is a quick at-a-glance method of determining of an apparent relationship between two variables, if any. A scatter diagram (or a graph) can be obtained on a graph paper by plotting observed (or known) pairs of values of variables x and y, taking the independent variable values on the x-axis and the dependent variable values on the y-axis.

Scatter diagram: A graph of pairs of values of two variables that is plotted to indicate a visual display of the pattern of their relationship.

It is common to try to draw a straight line through data points so that an equal number of points lie on either side of the line. The relationship between two variables x and y described by the data points is defined by this straight line.

In a scatter diagram the horizontal and vertical axes are scaled in units corresponding to the variables x and y, respectively. Figure 13.2 shows examples of different types of relationships based on pairs of values of x and y in a sample data. The pattern of data points in the diagram indicates that the variables are related. If the variables are related, then the dotted line appearing in each diagram describes relationship between the two variables.

The patterns depicted in Fig. 13.2(a) and (b) represent linear relationships since the patterns are described by straight lines. The pattern in Fig. 13.2(a) shows a *positive* relationship since the value of y tends to increase as the value of x increases, whereas pattern in Fig. 13.2(b) shows a *negative* relationship since the value of y tends to decrease as the value of x increases.

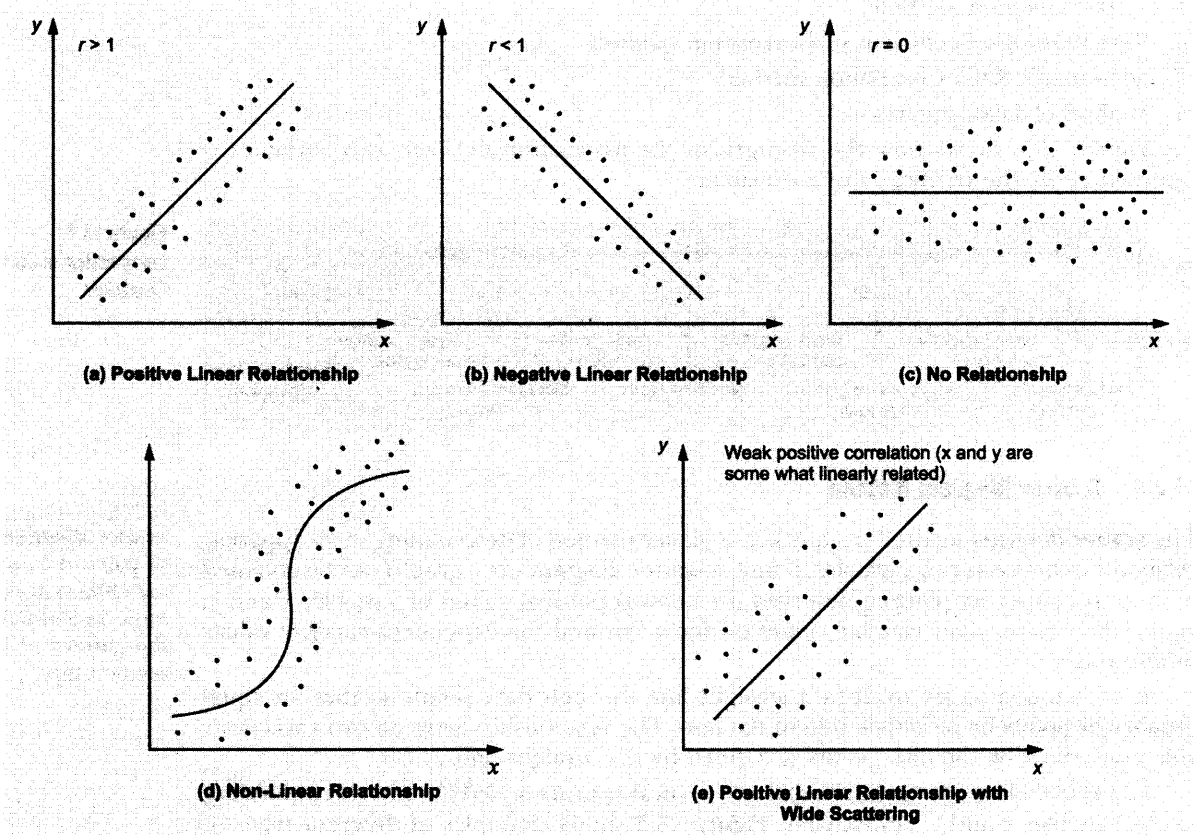
The pattern depicted in Fig. 13.2(c) illustrates very low or no relationship between the values of x and y, whereas Fig. 13.2(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. Figure 13.2(e) illustrates a positive linear relationship with a widely scattered pattern of points. The wider scattering indicates that there is a lower degree of association between the two variables x and y than there is in Fig. 13.2(a).

Interpretation of Correlation Coefficients While interpretating correlation coefficient r, the following points should be taken into account:

(i) A low value of r does not indicate that the variables are unrelated but indicates that the relationship is poorly described by a straight line. A non-linear relatinship may also exist.

(ii) A correlation does not imply a *cause-and-effect* relationship, it is merely an observed association.

Figure 13.2
Typical Examples of Correlation Coefficient



Types of Correlation Coefficients Table 13.1 shows several types of correlation coefficients used in statistics along with the conditions of their use. All of them are appropriate for quantifying linear relationship between two variables x and y.

Table 13.1: Types of Correlation Coefficients

Coefficient	Conditions Applied for Use
• ϕ (phi)	Both x and y variables are measured on a nominal scale
• ρ (rho)	Both x and y variables are measured on, or changed to, ordinal scales (rank data)
• r	Both x and y variables are measured on an interval or ratio scale scales (numeric data)

The correlation coefficient, denoted by η (eta) is used for quantifying nonlinear relationships (It is beyond the scope of this text). In this chapter we will calculate only the commonly used Pearson's r and Spearman's ρ correlation coefficients.

Specific Features of the Correlation Coefficient Regardless of the type of correlation coefficient we use, the following are the common among all of them.

(i) The value of r depends on the slope of the line passing through the data points and the scattering of the pair of values of variables x and y about this line (for detail see Chapter 14).

(ii) The sign of the correlation coefficient indicates the direction of the relationship. A positive correlation denoted by + (positive sign) indicates that the two variables tend to increase (or decrease) together (a positive association) and a negative correlation by – (minus sign) indicates that when one variable increases the other is likely to decrease (a negative association).

(iii) The values of the correlation coefficient range from + 1 to – 1 regardless of the units of measurements of x and y.

(iv) The value of $r = +1$ or -1 indicates that there is a perfect linear relationship between two variables, x and y. A perfect correlation implies that every observed pair of values of x and y falls on the straight line.

(v) The magnitude of the correlation indicates the strength of the relationship, which is the overall closeness of the points to a straight line. The sign of the correlation does indicates about the strength of the linear relationship.

(vi) Correlation coefficient is independent of the change of origin and scale of reference. In other words, its value remains unchanged when we subtract some constant from every given value of variables x and y (change of origin) and when we divide or multiply by some constant every given value of x and y (change of scale).

(vii) Correlation coefficient is a pure number independent of the unit of measurement.

(viii) The value of $r = 0$ indicates that the straight line through the data is exactly horizontal, so that the value of variable x does not change the predicated value of variable y.

(ix) The square of r, i.e., r^2 is referred to as *coefficient of determination.*

Further, from Fig. 13.2(a) to (e) we conclude that the closer the value of r is to either + 1 or – 1, the stronger is the association between x and y. Also, closer the value of r to 0, the weaker the association between x and y appears to be.

Example 13.1: Given the following data:

Student	: 1	2	3	4	5	6	7	8	9	10
Management aptitude score	: 400	675	475	350	425	600	550	325	675	450
Grade point average	: 1.8	3.8	2.8	1.7	2.8	3.1	2.6	1.9	3.2	2.3

(a) Draw this data on a graph paper.

(b) Is there any correlation between per capita national income and per capita consumer expenditure? If yes, what is your opinion.

Solution: By taking an appropriate scale on the x and y axes, the pair of observations are plotted on a graph paper as shown in Fig. 13.3. The scatter diagram in Fig. 13.3 with straight line represents the relationship between x and y 'fitted' through it.

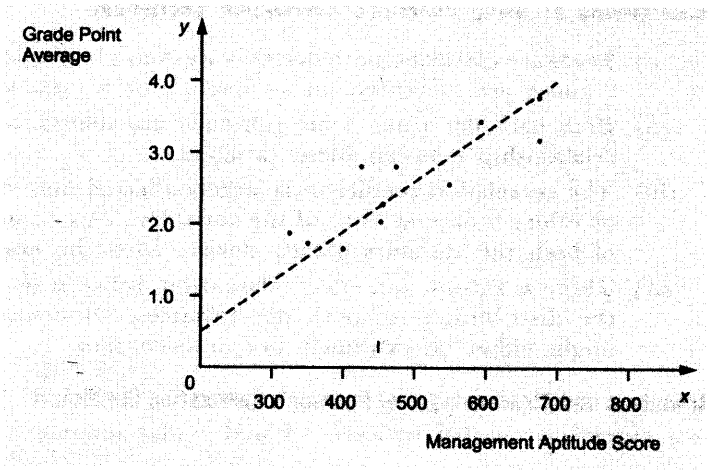


Figure 13.3
Scatter Diagram

Interpretation: From the scatter diagram shown in Fig. 13.3, it appears that there is a high degree of association between two variable values. It is because the data points are very close to a straight line passing through the points. This pattern of dotted points also indicates a high degree of linear positive correlation.

13.5.2 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures quantitatively the extent to which two variables x and y are correlated. For a set of n pairs of values of x and y, Pearson's correlation coefficient r is given by

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

where $$\text{Cov}(x, y) = \frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})$$

$$\sigma_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} \leftarrow \text{standard deviation of sample data on variable } x$$

$$\sigma_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}} \leftarrow \text{standard deviation of sample data on variable } y$$

Substituting mathematical formula for Cov (x, y) and σ_x and σ_y, we have

$$r = \frac{\frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}\sqrt{\frac{\Sigma(y - \bar{y})^2}{n}}} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (13\text{-}1)$$

Step Deviation Method for Ungrouped Data When actual mean values $\bar{x}$ and $\bar{y}$ are in fraction, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of x and y values from their assumed means A and B, respectively. That is, $d_x = x - A$ and $d_y = y - B$, where A and B are assumed means of x and y values. The formula (13-1) becomes

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}} \quad (13\text{-}2)$$

Step Deviation Method for Grouped Data When data on x and y values are classified or grouped into a frequency distribution, the formula (13-2) is modified as:

$$r = \frac{n\Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{n\Sigma f d_x^2 - (\Sigma f d_x)^2}\sqrt{n\Sigma f d_y^2 - (\Sigma f d_y)^2}} \quad (13\text{-}3)$$

Assumptions of Using Pearson's Correlation Coefficient

(i) Pearson's correlation coefficient is appropriate to calculate when both variables x and y are measured on an interval or a ratio scale.

(ii) Both variables x and y are normally distributed, and that there is a linear relationship between these variables.

(iii) The correlation coefficient is largely affected due to truncation of the range of values in one or both of the variables. This occurs when the distributions of both the variables greatly deviate from the normal shape.

(iv) There is a cause and effect relationship between two variables that influences the distributions of both the variables. Otherwise correlation coefficient might either be extremely low or even zero.

Advantage and Disadvantages of Pearson's Correlation Coefficient The correlation coefficient is a numerical number between – 1 and 1 that summarizes the magnitude as well

as direction (positive or negative) of association between two variables. The chief limitations of Pearson's method are:

(i) The correlation coefficient always assumes a linear relationship between two variables, whether it is true or not.

(ii) Great care must be exercised in interpreting the value of this coefficient as very often its value is misinterpreted.

(iii) The value of the coefficient is unduly affected by the extreme values of two variable values.

(iv) As compared with other methods the computational time required to calculate the value of r using Pearson's method is lengthy.

13.5.3 Probable Error and Standard Error of Coefficient of Correlation

The probable error (PE) of coefficient of correlation indicates extent to which its value depends on the condition of random sampling. If r is the calculated value of correlation coefficient in a sample of n pairs of observations, then the standard error SE_r of the correlation coefficient r is given by

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

The probable error of the coefficient of correlation is calculated by the expression:

$$PE_r = 0.6745\, SE_r = 0.6745\, \frac{1-r^2}{\sqrt{n}}$$

Thus with the help of PE_r we can determine the range within which population coefficient of correlation is expected to fall using following formula:

$$\rho = r \pm PE_r$$

where ρ (rho) represents population coefficient of correlation.

Remarks

1. If $r < PE_r$ then the value of r is not significant, that is, there is no relationship between two variables of interest.
2. If $r > 6PE_r$ then value of r is significant, that is, there exists a relationship between two variables.

Illustration: If $r = 0.8$ and $n = 25$, then PE_r is

$$PE_r = 0.6745\, \frac{1-(0.8)^2}{\sqrt{25}} = 0.6745\, \frac{0.36}{5} = 0.048$$

Thus the limits within which population correlation coefficient (ρ_r) should fall are

$$r \pm PE_r = 0.8 \pm 0.048 \quad \text{or} \quad 0.752 \leq \rho_r \leq 0.848$$

13.5.4 The Coefficient of Determination

The squared value of the correlation coefficint r is called **coefficient of determination**, denoted as r^2, It always has a value between 0 and 1. By squaring the correlation coefficient we retain information about the strength of the relationship but we lose information about the direction. *This measure represents the proportion (or percentage) of the total variability of the dependent variable, y that is accounted for or explained by the independent variable, x.* The proportion (or percentage) of variation in y that x can explain determines more precisely the extent or strength of association between two variables x and y (see Chapter 14 for details).

Coefficient of determination: A statistical measure of the proportion of the variation in the dependent variable that is explained by independent variable.

- The coefficient of correlation r has been grossly overrated and is used entirely too much. Its square, coefficient of determination r^2, is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between two correlated variables. —Tuttle

Interpretation of Coefficient of Determination: Coefficient of determination is preferred for interpreting the strength of association between two variables becuase it is easier to interpret a percentage. Figure 13.4 illustrates the meaning of the coefficient of determination:

- If $r^2 = 0$, then *no variation* in y can be explain by the variable x. It is shown in Fig 13.2(c) where x is of no value in predicting the value of y. There is *no association* between x and y.
- If $r^2 = 1$, then values of y are *completely explained* by x. There is *perfect association* between x and y.
- If $0 \leq r^2 \leq 1$, the degree of explained variation in y as a result of *variation in values of* x depends on the value of r^2. Value of r^2 closer to 0 shows low proportion of variation in y explained by x. On the other hand value of r^2 closer to 1 show that variable x can predict the actual value of the variable y.

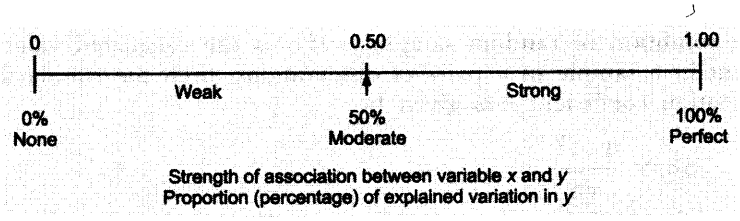


Figure 13.4
Interpretation of Coefficient of Determination

Mathematically, the coefficient of determination is given by

$$r^2 = 1 - \frac{\text{Explained variability in } y}{\text{Total variability in } y}$$

$$= 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{n\Sigma y^2 - a\Sigma y - b\Sigma xy}{n\Sigma y^2 - (\bar{y})^2}$$

where $\hat{y} = a + bx$ is the estimated value of y for given values of x. One minus the ratio between these two variations is referred as the *coefficient of determination.*

For example, let correlation between variable x (height) and variable y (weight) be $r = 0.70$. Now the coefficient of determination $r^2 = 0.49$ or 49 per cent, implies that only 49 per cent of the variation in variable y (weight) can be accounted for in terms of variable x (height). The remaining 51 per cent of the variability may be due to other factors, say for instance, tendency to eat fatty foods.

It may be noted that even a relatively high correlation coefficient $r = 0.70$ accounts for less than 50 per cent of the variability. In this context, it is important to know that 'variability' refers to how values of variable y are scattered around its own mean value. That is, as in the above example, some people will be heavy, some average, some light. So we can account for 49 per cent of the total variability of weight (y) in terms of height (x) if $r = 0.70$. The greater the correlation coefficient, the greater the coefficient of determination, and the variability in dependent variable can be accounted for in terms of independent variable.

Example 13.2: The following table gives indices of industrial production and number of registered unemployed people (in lakh). Calculate the value of the correlation coefficient.

Year	:	1991	1992	1993	1994	1995	1996	1997	1998
Index of Production	:	100	102	104	107	105	112	103	99
Number Unemployed	:	15	12	13	11	12	12	19	26

Solution: Calculations of Karl Pearson's correlation coefficient are shown in the table below:

Year	*Production* x	$dx = (x - \bar{x})$	d_x^2	*Unemployed* y	$d_y^2 = (y - \bar{y})$	d_y^2	$d_x d_y$
1991	100	- 4	16	15	0	0	0
1992	102	- 2	4	12	- 3	9	+ 6
1993	104	0	0	13	- 2	4	0
1994	107	+ 3	9	11	- 4	16	- 12
1995	105	+ 1	1	12	- 3	9	- 3
1996	112	+ 8	64	12	- 3	9	- 24
1997	103	- 1	1	19	+ 4	16	- 4
1998	99	- 5	25	26	+ 11	121	- 55
Total	832	0	120	120	0	184	- 92

$$\bar{x} = \frac{\Sigma x}{n} = \frac{832}{8} = 104; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{120}{8} = 15$$

Applying the formula, $r = \dfrac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$

$$= \frac{8 \times -92}{\sqrt{8 \times 120}\sqrt{8 \times 184}} = \frac{-92}{10.954 \times 13.564}$$

$$= \frac{-92}{148.580} = -0.619$$

Interpretation: Since coefficient of correlation $r = -0.619$ is moderately negative, it indicates that there is a moderately large inverse correlation between the two variables. Hence we conclude that as the production index increases, the number of unemployed decreases and vice-versa.

Example 13.3: The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size-group	:	15–16	16–17	17–18	18–19	19–20	20–21
No. of items	:	200	270	340	360	400	300
No. of defective items	:	150	162	170	180	180	114

[*Delhi Univ., BCom, 1999*]

Solution: Let group size be denoted by variable x and number of defective items by variable y. Calculations for Karl Pearson's correlation coefficient are shown below:

Size-Group	*Mid-value* m	$d_x = m - 17.5$	d_x^2	*Percent of Defective Items*	$d_y = y - 50$	d_y^2	$d_x d_y$
15–16	15.5	- 2	4	75	+ 25	625	- 50
16–17	16.5	- 1	1	60	+ 10	100	- 10
17–18	17.5	0	0	50	0	0	0
18–19	18.5	+ 1	1	50	0	0	0
19–20	19.5	+ 2	4	45	- 5	25	- 10
20–21	20.5	+ 3	9	38	- 12	144	- 36
		3	19		18	894	- 106

Substituting values in the formula of Karl Pearson's correlation coefficient r, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{6\times -106 - 3\times 18}{\sqrt{6\times 19-(3)^2}\sqrt{6\times 894-(18)^2}} = \frac{-636-54}{\sqrt{105}\sqrt{5040}}$$

$$= -\frac{690}{727.46} = -0.949$$

Interpretation: Since value of r is negative, and is moderately close to –1, statistical association between x (size group) and y (percent of defective items) is moderate and negative, we conclude that when size of group increases, the number of defective items decreases and vice-versa.

Example 13.4: The following data relate to age of employees and the number of days they reported sick in a month.

Employees :	1	2	3	4	5	6	7	8	9	10
Age :	30	32	35	40	48	50	52	55	57	61
Sick days :	1	0	2	5	2	4	6	5	7	8

Calculate Karl Pearson's coefficient of correlation and interpret it.

[*Kashmir Univ., BCom, 1997*]

Solution: Let age and sick days be represented by variables x and y, respectively. Calculations for value of correlation coefficient are shown below:

Age			*Sick days*			
x	$dx = x - \bar{x}$	d_x^2	y	$d_y = y - \bar{y}$	d_y^2	$d_x d_y$
30	– 16	256	1	– 3	9	48
32	– 14	196	0	– 4	16	56
35	– 11	121	2	– 2	4	22
40	– 6	36	5	1	1	– 6
48	2	4	2	– 2	4	– 4
50	4	16	4	0	0	0
52	6	36	6	2	4	12
55	9	81	5	1	1	9
57	11	121	7	3	9	33
61	15	225	8	4	16	60
460	0	1092	40	0	64	230

$$\bar{x} = \frac{\Sigma x}{n} = \frac{460}{10} = 46 \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{40}{10} = 4$$

Substituting values in the formula of Karl Pearson's correlation coefficient r, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10\times 230}{\sqrt{10\times 1092}\sqrt{10\times 64}}$$

$$= \frac{230}{264.363} = 0.870$$

Interpretation: Since value of r is positive, therefore age of employees and number of sick days are positively correlated to a high degree. Hence we conclude that as the age of an employee increases, he is likely to go on sick leave more often than others.

Example 13.5: The following table gives the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

Test Marks	Age in years 18	19	20	21	Total
200–250	4	4	2	1	11
250–300	3	5	4	2	14
300–350	2	6	8	5	21
350–400	1	4	6	10	21
Total	10	19	20	18	67

[*Allahabad Univ., BCom, 1999*]

Solution: Let age of students and marks obtained by them be represented by variables x and y, respectively. Calculations for correlation coefficient for this bivariate data is shown below:

y \ x	d_y \ d_x	18 / −1	19 / 0	20 / 1	21 / 2	Total, f	fd_y	fd_y^2	fd_xd_y
200–250	−1	(4) 4	(0) 4	(−2) 2	(−2) 1	11	−11	11	0
250–300	0	(0) 3	(0) 5	(0) 4	(0) 2	14	0	0	0
300–350	1	(−2) 2	(0) 6	(8) 8	(10) 5	21	21	21	16
350–400	2	(−2) 1	(0) 4	(12) 6	(40) 10	21	42	84	50
Total, f		10	19	20	18	$n = 67$	$\Sigma fd_y = 52$	$\Sigma fd_y^2 = 116$	$\Sigma fd_xd_y = 66$
fd_x		−10	0	20	36	$\Sigma fd_x = 46$			
fd_x^2		10	0	20	72	$\Sigma fd_x^2 = 102$			
fd_xd_y		0	0	18	48	$\Sigma fd_xd_y = 66$			

where $d_x = x - 19$, $d_y = (m - 275)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{67\times66 - 46\times52}{\sqrt{67\times102 - (46)^2}\sqrt{67\times116 - (52)^2}}$$

$$= \frac{4422 - 2392}{\sqrt{6834 - 2116}\sqrt{7772 - 2704}} = \frac{2030}{\sqrt{4718}\sqrt{5068}}$$

$$= \frac{2030}{68.688\times71.19} = 0.415$$

Interpretation: Since the value of r is positive, therefore age of students and marks obtained in an intelligence test are positively correlated to the extent of 0.415. Hence, we conclude that as the age of students increases, score of marks in intelligence test also increases.

Example 13.6: Calculate the coefficient of correlation from the following bivariate frequency distribution:

Sales Revenue (Rs in lakh)	*Advertising Expenditure (Rs in '000)*				*Total*
	5–10	*10–15*	*15–20*	*20–25*	
75–125	4	1	—	—	5
125–175	7	6	2	1	16
175–225	1	3	4	2	10
225–275	1	1	3	4	9
Total	13	11	9	7	40

[*Delhi Univ., MBA, 1997*]

Solution: Let advertising expenditure and sales revenue be represented by variables x and y, respectively. The calculations for correlation coefficient are shown below:

Revenue y	*Mid-value* (m)	$x \rightarrow$ *Mid-value* (m) d_x / d_y	*Advertising Expenditure* *5–10* 7.5 −1	*10–15* 12.5 0	*15–20* 17.5 1	*20–25* 22.5 2	*Total*, f	fd_y	fd_y^2	fd_xd_y
75–125	100	−2	(8) 4	(0) 1	(0) —	(0) —	5	−10	20	8
125–175	150	−1	(7) 7	(0) 6	(−2) 2	(−2) 1	16	−16	16	3
175–225	200	0	(0) 1	(0) 3	(0) 4	(0) 2	10	0	0	0
225–275	250	1	(−1) 1	(0) 1	(3) 3	(8) 4	9	9	9	10
		Total, f	13	11	9	7	$n = 40$	$\Sigma d_y = -17$	$\Sigma d_y^2 = 45$	$\Sigma fd_xd_y = 21$
		fd_x	−13	0	9	14	$\Sigma fd_x = 10$			
		fd_x^2	13	0	9	28	$\Sigma fd_x^2 = 50$			
		fd_xd_y	14	0	1	6	$\Sigma fd_xd_y = 21$			

where, $d_x = (m - 12.5)/5$ and $d_y = (m - 200)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{40\times 21 - 10\times -17}{\sqrt{40\times 50 - (10)^2}\sqrt{40\times 45 - (-17)^2}}$$

$$= \frac{840 + 170}{\sqrt{1900}\sqrt{1511}} = \frac{1010}{1694.373} = 0.596$$

Interpretation: Since the value of r is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence we conclude that as expenditure on advertising increases, the sales revenue also increases.

Example 13.7: A computer, while calculating the correlation coefficient between two variables x and y from 25 pairs of observations, obtained the following results:

$n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$ and $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$

It was, however, discovered at the time of checking that he had copied down two pairs of observations as:

x	y
6	14
8	6

instead of

x	y
8	12
6	8

Obtain the correct value of correlation coefficient between x and y.

[MD Univ., MCom, 1998; Kumaon Univ., MBA, 2000]

Solution: The corrected values for termed needed in the formula of Pearson's correlation coefficient are determined as follows:

$$\text{Correct } \Sigma x = 125 - (6 + 8 - 8 - 6) = 125$$
$$\text{Correct } \Sigma y = 100 - (14 + 6 - 12 - 8) = 100$$
$$\text{Correct } \Sigma x^2 = 650 - \{(6)^2 + (8)^2 - (8)^2 - (6)^2\}$$
$$= 650 - \{36 + 64 - 64 - 36\} = 650$$
$$\text{Correct } \Sigma y^2 = 460 - \{(14)^2 + (6)^2 - (12)^2 - (8)^2\}$$
$$= 460 - \{196 + 36 - 144 - 64\} = 436$$
$$\text{Correct } \Sigma xy = 508 - \{(6\times14) + (8\times6) - (8\times12) - (6 \times 8)\}$$
$$= 508 - \{84 - 48 - 96 - 48\} = 520$$

Applying the formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{25\times520 - 125\times100}{\sqrt{25\times650 - (125)^2}\sqrt{25\times436 - (100)^2}}$$

$$= \frac{13{,}000 - 12{,}500}{\sqrt{625}\sqrt{900}} = \frac{500}{25\times30} = 0.667$$

Thus, the correct value of correlation coefficient between x and y is 0.667.

Self-Practice Problems 13A

13.1 Making use of the data summarized below, calculate the coefficient of correlation.

Case	x_1	x_2	*Case*	x_1	x_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

13.2 Find the correlation coefficient by Karl Pearson's method between x and y and interpret its value.

x :	57	42	40	33	42	45	42	44	40	56	44	43
y :	10	60	30	41	29	27	27	19	18	19	31	29

13.3 Calculate the coefficient of correlation from the following data:

x :	100	200	300	400	500	600	700
y :	30	50	60	80	100	110	130

13.4 Calculate the coefficient of correlation between x and y from the following data and calculate the probable errors. Assume 69 and 112 as the mean value for x and y respectively.

x :	78	89	99	60	50	79	68	61
y :	125	137	156	112	107	136	123	108

13.5 Find the coefficient of correlation from the following data:

Cost :	39	65	62	90	82	75	25	98	36	78
Sales :	47	53	58	86	62	68	60	91	51	84

[Madras Univ., BCom, 1997]

13.6 Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Also calculate the probable error and comment on the value:

Age :	20	21	22	23	24	25
No. of students :	500	400	300	240	200	160
Regular players :	400	300	180	96	60	24

[HP Univ., MBA, 1997]

13.7 Find the coefficient of correlation between age and the sum assured (in 1000 Rs) from the following table:

Age Group (years)	*Sum Assured (in Rs)*				
	10	*20*	*30*	*40*	*50*
20–30	4	6	3	7	1
30–40	2	8	15	7	1
40–50	3	9	12	6	2
50–60	8	4	2	—	—

[Delhi Univ., MBA, 1999]

13.8 Family income and its percentage spent on food in the case of one hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and interpret its value.

Food Expenditure (in percent)	*Monthly Family Income (Rs)*				
	2000–3000	*3000–4000*	*4000–5000*	*5000–6000*	*6000–7000*
10–15	—	—	—	3	7
15–20	—	4	9	4	3
20–25	7	6	12	5	—
25–30	3	10	19	8	—

[*Delhi Univ., MBA, 2000*]

13.9 With the following data in 6 cities, calculate Pearson's coefficient of correlation between the density of population and death rate:

City	*Area in Kilometres*	*Population (in '000)*	*No. of Deaths*
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

[*Sukhadia Univ., BCom, 1998*]

13.10 The coefficient of correlation between two variables x and y is 0.3. The covariance is 9. The variance of x is 16. Find the standard deviation of y series.

Hints and Answers

13.1 $\bar{x}_1 = 80/8 = 10$, $\bar{x}_2 = 64/8 = 8$;

$$r = \frac{43}{\sqrt{32}\sqrt{72}} = 0.896$$

13.2 $r = -0.554$ **13.3** $r = 0.997$

13.4 $r = 0.014$ **13.5** $r = 0.780$

13.6 $r = 0.005$ **13.7** $r = -0.256$

13.8 $r = -0.438$ **13.9** $r = 0.988$

13.10 Given $\sigma_x = \sqrt{16} = 4$; $r = \dfrac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$

or $0.3 = \dfrac{9}{4\sigma_y}$ or $\sigma_y = 7.5$.

13.5.5 Spearman's Rank Correlation Coefficient

This method of finding the correlation coefficient between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This method is applied to measure the association between two variables when only *ordinal (or rank) data* are available. In other words, this method is applied in a situation in which quantitative measure of certain qualitative factors such as judgement, brands personalities, TV programmes, leadership, colour, taste, cannot be fixed, but individual observations can be arranged in a definite order (also called rank). The ranking is decided by using a set of ordinal rank numbers, with 1 for the individual observation ranked first either in terms of quantity or quality; and n for the individual observation ranked last in a group of n pairs of observations. Mathematically, Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad (13\text{-}4)$$

where R = rank correlation coefficient

R_1 = rank of observations with respect to first variable

R_2 = rank of observations with respect to second variable

$d = R_1 - R_2$, difference in a pair of ranks

n = number of paired observations or individuals being ranked

The number '6' is placed in the formula as a scaling device, it ensures that the possible range of R is from – 1 to 1. While using this method we may come across three types of cases.

Advantages and Disadvantages of Spearman's Correlation Coefficient Method

Advantages

(i) This method is easy to understand and its application is simpler than Pearson's method.

(ii) This method is useful for correlation analysis when variables are expressed in qualitative terms like beauty, intelligence, honesty, efficiency, and so on.

(iii) This method is appropriate to measure the association between two variables if the data type is at least ordinal scaled (ranked)

(iv) The sample data of values of two variables is converted into ranks either in ascending order or decending order for calculating degree of correlation between two variables.

Disadvantages

(i) Values of both variables are assumed to be normally distributed and describing a linear relationship rather than non-linear relationship.

(ii) A large computational time is required when number of pairs of values of two variables exceed 30.

(iii) This method cannot be applied to measure the association between two variable grouped data.

Case I: When Ranks are Given

When observations in a data set are already arranged in a particular order (rank), take the differences in pairs of observations to determine d. Square these differences and obtain the total Σd^2. Apply, formula (13-4) to calculate correlation coefficient.

Example 13.8: The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of the squares of the differences in ranks is given to be 48, find the values of n.

Solution: The formula for Spearman's correlation coefficient is as follows:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Given, R = 0.143, $\Sigma d^2 = 48$ and $n = 7$. Substituting values in the formula, we get

$$0.143 = 1 - \frac{6 \times 48}{n(n^2-1)} = 1 - \frac{288}{n^3-n}$$

$$0.143\ (n^3 - n) = (n^3 - n) - 288$$

$$n^3 - n - 336 = 0 \quad \text{or} \quad (n - 7)\ (n^2 + 7n + 48) = 0$$

This implies that either $n - 7 = 0$, that is, $n = 7$ or $n^2 + 7n + 48 = 0$. But $n^2 + 7n + 48 = 0$ on simplification gives undesirable value of n because its discriminant $b^2 - 4ac$ is negative. Hence $n = 7$.

Example 13.9: The ranks of 15 students in two subjects A and B, are given below. The two numbers within brackets denote the ranks of a student in A and B subjects respectively.

(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1),
(9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13)

Find Spearman's rank correlation coefficient. *[Sukhadia Univ., MBA, 1998]*

Solution: Since ranks of students with respect to their performance in two subjects are given, calculations for rank correlation coefficient are shown below:

Rank in A R_1	*Rank in B* R_2	*Difference* $d = R_1 - R_2$	d^2
1	10	−9	81
2	7	−5	25
3	2	1	1
4	6	−2	4
5	4	1	1
6	8	−2	4
7	3	4	16
8	1	7	49
9	11	−2	4
10	15	−5	25
11	9	2	4
12	5	7	49
13	14	−1	1
14	12	2	4
15	13	2	4
			$\Sigma d^2 = 272$

Apply the formula, $R = 1 - \Sigma\frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 272}{15\{(15)^2 - 1\}}$

$$= 1 - \frac{1632}{3360} = 1 - 0.4857 = 0.5143$$

The result shows a moderate degree positive correlation between performance of students in two subjects.

Example 13.10: An office has 12 clerks. The long-serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service.

Ranking according to length of service :	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency :	2	3	5	1	9	10	11	12	8	7	6	4

Do the data support the clerks' claim for seniority increment?

[*Sukhadia Univ., MBA, 1991*]

Solution: Since ranks are already given, calculations for rank correlation coefficient are shown below:

Rank According to Length of Service R_1	*Rank According to Efficiency* R_2	*Difference* $d = R_1 - R_2$	d^2
1	2	−1	1
2	3	−1	1
3	5	−2	4
4	1	3	9
5	9	−4	16
6	10	−4	16
7	11	−4	16
8	12	−4	16
9	8	1	1
10	7	3	9
11	6	5	25
12	4	8	64
			$\Sigma d^2 = 178$

Applying the formula, $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$$= 1 - \frac{6\times178}{12(144-1)} = 1 - \frac{1068}{1716} = 0.378$$

The result shows a low degree positive correlation between length of service and efficiency, the claim of the clerks for a seniority increment based on length of service is not justified.

Example 13.11: Ten competitors in a beauty contest are ranked by three judges in the following order:

Judge 1:	1	6	5	10	3	2	4	9	7	8
Judge 2:	3	5	8	4	7	10	2	1	6	9
Judge 3:	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty. *[MD Univ., MBA, 1996]*

Solution: The pair of judges who have the nearest approach to common taste in beauty can be obtained in ${}^3C_2 = 3$ ways as follows:

(i) Judge 1 and judge 2.

(ii) Judge 2 and judge 3.

(iii) Judge 3 and judge 1.

Calculations for comparing their ranking are shown below:

Judge 1 R_1	*Judge 2* R_2	*Judge 3* R_3	$d^2 = (R_1 - R_2)^2$	$d^2 = (R_2 - R_3)^2$	$d^2 = (R_3 - R_1)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			$\Sigma d^2 = 200$	$\Sigma d^2 = 214$	$\Sigma d^2 = 60$

Applying the formula

$$R_{12} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times200}{10(100-1)} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times214}{10(100-1)} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times60}{10(100-1)} = 1 - \frac{360}{990} = 0.636$$

Since the correlation coefficient $R_{13} = 0.636$ is largest, the judges 1 and 3 have nearest approach to common tastes in beauty.

Case 2: When Ranks are not Given

When pairs of observations in the data set are not ranked as in Case 1, the ranks are assigned by taking either the highest value or the lowest value as 1 for both the variable's values.

Example 13.12: Quotations of index numbers of security prices of a certain joint stock company are given below:

Year	*Debenture Price*	*Share Price*
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Using the rank correlation method, determine the relationship between debenture prices and share prices. [*Calicut Univ., BCom, 1997*]

Solution: Let us start ranking from the lowest value for both the variables, as shown below:

Debenture Price (x)	*Rank*	*Share Price* (y)	*Rank*	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
97.8	3	73.2	1	2	4
99.2	7	85.8	6	1	1
98.8	6	78.9	4	2	4
98.3	4	75.8	2	2	4
98.4	5	77.2	3	2	4
96.7	1	87.2	7	−6	36
97.1	2	83.8	5	−3	9
					$\Sigma d^2 = 62$

Applying the formula $R = 1 - \frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 62}{(7)^3 - 7}$

$$= 1 - \frac{372}{336} = 1 - 0.107 = -0.107$$

The result shows a low degree of negative correlation between the debenture prices and share prices of a certain joint stock company.

Example 13.13 An economist wanted to find out if there was any relationship between the unemployment rate in a country and its inflation rate. Data gathered from 7 countries for the year 2004 are given below:

Country	*Unemployment Rate (Percent)*	*Inflation Rate (Per cent)*
A	4.0	3.2
B	8.5	8.2
C	5.5	9.4
D	0.8	5.1
E	7.3	10.1
F	5.8	7.8
G	2.1	4.7

Find the degree of linear association between a country's unemployment rate and its level of inflation.

Solution: Let us start ranking from the lowest value for both the variables as shown below:

Unemployment Rate (x)	*Rank* R_1	*Inflation Rate (y)*	*Rank* R_2	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4.0	3	3.2	1	2	4
8.5	7	8.2	5	2	4
5.5	4	9.4	6	- 2	4
0.8	1	5.1	3	- 2	4
7.3	6	10.1	7	- 1	1
5.8	5	7.8	4	1	1
2.1	2	4.7	2	0	0
					$\Sigma d^2 = 18$

Applying the formula,

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 18}{(7)^3 - (7)} = 1 - \frac{108}{336} = 0.678$$

The result shows a moderately high degree of positive correlation between unemployment rate and inflation rate of seven countries.

Case 3: When Ranks are Equal

While ranking observations in the data set by taking either the highest value or lowest value as rank 1, we may come across a situation of more than one observations being of equal size. In such a case the rank to be assigned to individual observations is an average of the ranks which these individual observations would have got had they differed from each other. For example, if two observations are ranked equal at third place, then the average rank of (3 + 4)/2 = 3.5 is assigned to these two observations. Similarly, if three observations are ranked equal at third place, then the average rank of (3 + 4 + 5)/3 = 4 is assigned to these three observations.

While equal ranks are assigned to a few observations in the data set, an adjustment is made in the Spearman rank correlation coefficient formula as given below:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\right\}}{n(n^2 - 1)}$$

where m_i (i = 1, 2, 3, . . .) stands for the number of times an observation is repeated in the data set for both variables.

Example 13.14: A financial analyst wanted to find out whether inventory turnover influences any company's earnings per share (in per cent). A random sample of 7 companies listed in a stock exchange were selected and the following data was recorded for each.

Company	*Inventory Turnover (Number of Times)*	*Earnings per Share (Per cent)*
A	4	11
B	5	9
C	7	13
D	8	7
E	6	13
F	3	8
G	5	8

Find the strength of association between inventory turnover and earnings per share. Interpret this finding.

Solution: Let us start ranking from lowest value forboth the variables. Since there are tied ranks, the sum of the tied ranks is averaged and assigned to each of the tied observations as shown below.

Inventory Turnover (x)	*Rank* R_1	*Earnings Per Share (y)*	*Rank* R_2	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
4	2	11	5	- 3.0	9.00
5	3.5	9	4	- 0.5	0.25
7	6	13	6.5	0.5	0.25
8	7	7	1	6.0	36.00
6	5	13	6.5	- 1.5	2.25
3	1	8	2.5	- 1.5	2.25
5	3.5	8	2.5	1.0	1.00
					$\Sigma d^2 = 51$

If may be noted that a value 5 of variable x is repeated twice ($m_1 = 2$) and values 8 and 13 of variable y is also reperated twice, so $m_2 = 2$ and $m_3 = 2$. Applying the formula:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{51 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right\}}{7(49-1)}$$

$$= 1 - \frac{6\{51 + 0.5 + 0.5 + 0.5\}}{336} = 1 - 0.9375 = 0.0625$$

The result shows a very week positive association between inventory turnover and earningings per share.

Example 13.15: Obtain the rank correlation coefficient between the variables x and y from the following pairs of observed values.

x :	50	55	65	50	55	60	50	65	70	75
y :	110	110	115	125	140	115	130	120	115	160

[*Mangalore Univ., BCom, 1997*]

Solution: Let us start ranking from lowest value for both the variables. Moreover, certain observations in both sets of data are repeated, the ranking is done in accordance with suitable average value as shown below.

Variable x	*Rank* R_1	*Variable* y	*Rank* R_2	*Difference* $d = R_1 - R_2$	$d^2 = (R_1 - R_2)^2$
50	2	110	1.5	0.5	0.25
55	4.5	110	1.5	3.0	9.00
65	7.5	115	4	3.5	12.25
50	2	125	7	- 5.0	25.00
55	4.5	140	9	- 4.5	20.25
60	6	115	4	2.0	4.00
50	2	130	8	- 6.0	36.00
65	7.5	120	6	1.5	2.25
70	9	115	4	5.0	25.00
75	10	160	10	0.0	00.00
					$\Sigma d^2 = 134.00$

It may be noted that for variable x, 50 is repeated thrice ($m_1 = 3$), 55 is repeated twice ($m_2 = 2$), and 65 is repeated twice ($m_3 = 2$). Also for variable y, 110 is repeated twice ($m_4 = 2$) and 115 thrice ($m_5 = 3$). Applying the formula:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \frac{1}{12}(m_4^3 - m_4) + \frac{1}{12}(m_5^3 - m_5)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10(100-1)}$$

$$= 1 - \frac{6[134 + 2 + 0.5 + 0.5 + 0.5 + 2]}{990} = 1 - \frac{6 \times 139.5}{990} = 1 - \frac{837}{990}$$

$$= 1 - 0.845 = 0.155$$

The result shows a weak positive association between variables x and y.

13.5.6 Method of Least-Squares

The method of least-squares to calculate the correlation coefficient requires the values of regression coefficients b_{xy} and b_{yx}, so that

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

In other words, correlation coefficient is the geometric mean of two regression coefficients (see Chapter 14 for details).

13.5.7 Auto-Correlation Coefficient

The auto correlation coefficient describes the association or mutual dependence between values of the same variable but at different time periods. The auto correlation coefficient provides important information on how a variable relates to itself for a specific time lag. The difference in the period before a cause-and-effect relationship is established is called 'lag'. While computing the correlation, the time gap must be considered, otherwise misleading (deceptive) conclusions may be arrived at. For example, the decrease or increase in supply of a commodity may not immediately reflect on its price, it may take some lead time or time lag.

The formula for auto-correlation coefficient at time lag k is stated as:

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where k = length of time lag
n = number of observations
$\bar{x}$ = mean of all observations

Example 13.16: The monthly sales of a product, in thousands of units, in the last 6 months are given below:

Month :	1	2	3	4	5	6
Sales :	1.8	2.5	3.1	3.0	4.2	3.4

Compute the auto-correlation coefficient upto lag 2. What conclusion can be derived from these values regarding the presence of a trend in the data?

Solution: The calculations for auto-correlation coefficient are shown below:

Time	*Sales* (x)	x_1 = *One Time Lag Variable Constructed From x*	x_2 = *Two Time Lags Variable Constructed From x*
1	1.8	2.5	3.1
2	2.5	3.1	3.0
3	3.1	3.0	4.2
4	3.0	4.2	3.4
5	4.2	3.4	—
6	3.4	—	—

For $k = 1$, $\bar{x} = \frac{1}{6}(1.8 + 2.5 + \ldots + 3.4) = 3$

$$r_1 = \frac{\{(1.8-3)(2.5-3)+(2.5-3)(3.1-3)+(3.1-3)(3-3)+(3-3)(4.2-3)+(4.2-3)(3.4-3)\}}{(1.8-3)^2+(2.5-3)^2+(3.1-3)^2+(3-3)^2+(4.2-3)^2+(3.4-3)^2}$$

$$= \frac{(-1.2)(-0.5)+(-0.5)(0.1)+(0.1)(0)+(0)(1.2)+(1.2)(0.4)}{1.44+0.25+0.01+0+1.44+0.16}$$

$$= \frac{(0.6-0.5+0.48)}{3.3} = 0.312$$

For $k = 2$

$$r_2 = \frac{(1.8-3)(3.1-3)+(2.5-3)(3-3)+(3.1-3)(4.2-3)+(3-3)(3.4-3)}{(1.8-3)^2+(2.5-3)^2+\ldots+(3.4-3)^2}$$

$$= \frac{(-1.2\times0.1)+(-0.5\times0)+(0.1\times1.2)+(0\times0.4)}{3.3} = \frac{-0.12+0.12}{3.3} = 0$$

Since the value of r_1 is positive, it implies that there is a seasonal pattern of 6 months duration and $r_2 = 0$ implies that there is no significant change in sales.

Self-Practice Problems 13B

13.11 The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.2. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct coefficient of rank correlation.

[Delhi Univ., BCom, 1996]

13.12 The ranking of 10 students in accordance with their performance in two subjects A and B are as follows:

A:	6	5	3	10	2	4	9	7	8	1
B:	3	8	4	9	1	6	10	7	5	2

Calculate the rank correlation coefficient and comment on its value.

13.13 Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges x and y in a certain competitive test as shown below:

Student	*Marks by Judge x*	*Marks by Judge y*
1	52	65
2	53	68
3	42	43
4	60	38
5	45	77
6	41	48
7	37	35
8	38	30
9	25	25
10	27	50

13.14 An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the accountancy and statistics papers, compute the rank correlation coefficient.

Applicant :	A	B	C	D	E	F	G	H
Marks in accountancy:	15	20	28	12	40	60	20	80
Marks in statistics :	40	30	50	30	20	10	30	60

13.15 Seven methods of imparting business education were ranked by the MBA students of two universities as follows:

Method of Teaching :	1	2	3	4	5	6	7
Rank by students of Univ. A :	2	1	5	3	4	7	6
Rank by students of Univ. B :	1	3	2	4	7	5	6

Calculate the rank correlation coefficient and comment on its value.

13.16 An investigator collected the following data with respect to the socio-economic status and severity of respiratory illness.

Patient :	1	2	3	4	5	6	7	8
Socio-economic status (rank) :	6	7	2	3	5	4	1	8
Severity of illness rank) :	5	8	4	3	7	1	2	6

Calculate the rank correlation coefficient and comment on its value.

13.17 You are given the following data of marks obtained by 11 students in statistics in two tests, one before and other after special coaching:

First Test (Before coaching)	*Second Test (After coaching)*
23	24
20	19
19	22
21	18
18	20
20	22
18	20
20	22
18	20
17	20
23	23
16	20
19	17

Do the marks indicate that the special coaching has benefited the students? [*Delhi Univ., MCom, 1989*]

13.18 Two departmental managers ranked a few trainees according to their perceived abilities. The ranking are given below:

Trainee :	A	B	C	D	E	F	G	H	I	J
Manager A :	1	9	6	2	5	8	7	3	10	4
Manager B :	3	10	8	1	7	5	6	2	9	4

Calculate an appropriate correlation coefficient to measure the consistency in the ranking.

13.19 In an office some keyboard operators, who were already ranked on their speed, were also ranked on accuracy by their supervisor. The results were as follows:

Operator :	A	B	C	D	E	F	G	H	I	J
Speed :	1	2	3	4	5	6	7	8	9	10
Accuracy :	7	9	3	4	1	6	8	2	10	5

Calculate the appropriate correlation coefficient between speed and accuracy.

13.20 The personnel department is interested in comparing the ratings of job applicants when measured by a variety of standard tests. The ratings of 9 applicants on interviews and standard psychological test are shown below:

Applicant :	A	B	C	D	E	F	G	H	I
Interview :	5	2	9	4	3	6	1	8	7
Standard test :	8	1	7	5	3	4	2	9	6

Calculate Spearman's rank correlation coefficient and comment on its value.

Hints and Answers

13.11 Given R = 0.2, n = 10; $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$ or

$$0.2 = 1 - \frac{6\Sigma d^2}{10(100-1)} \text{ or } \Sigma d^2 = 100$$

$$\text{Correct value of } R = 1 - \frac{6\times 100}{10\times 99} = 0.394$$

13.12 $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 36}{10(100-1)} = 0.782$

13.13 $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 76}{10(100-1)} = 0.539$

13.14 $$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{81.5 + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\right\}}{8(64-1)} = 0$$

13.15 R = 0.50 **13.16** R = 0.477

13.17 R = 0.71 **13.18** R = 0.842

13.19 R = 0.006 **13.20** R = 0.817

13.6 HYPOTHESIS TESTING FOR CORRELATION COEFFICIENT

We often use the sample correlation coefficient r as an estimator to test whether the possible strength of association between two random variables in the population exist. In other words, we use r as an estimator in testing null and alternative hypotheses about true *population correlation coefficient* ρ(Greek letter rho). When such hypotheses are tested, the assumptions of normal distribution of two random variables, say x and y is required.

13.6.1 Hypothesis Testing about Population Correlation Coefficient (Small Sample)

The test of hypothesis for the existence of a linear relationship between two variables x and y involves the determination of sample correlation coefficient r. This test of linear relationship between x and y is the same as determining whether there is any significant correlation between them. For determining the correlation, we start by hypothesizing the population correlation coefficient ρ equal to zero. The population correlation coefficient ρ measures the degree of association between two variables in a population of interest. The null and alternative hypotheses are expressed as:

- **Two-tailed Test**

  $H_0 : \rho = 0$ (No correlation between variables x and y)

  $H_1 : \rho \neq 0$ (Correlation exists between variables x and y)

- **One-tailed Test**

  $H_0 : \rho = 0$ and $H_1 : \rho > 0$ (or $\rho < 0$)

The t-test statistic for testing the null hypothesis is given by:

$$t = \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{r \times \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

where r = sample correlation coefficient

s_r = standard error of correlation coefficient

n = sample size

The t-test statistic follows t-distribution with $n - 2$ degrees of freedom. If the sample size is large, then the standard error of correlation coefficient is given by $s_r = (1 - r^2)/\sqrt{n}$.

Decision Rule: The calculated value of t-test statistic is compared with its critical (or table) value at $n - 2$ degrees of freedom and level of significance α to arrive at a decision as follows:

One-tailed Test	*Two-tailed Test*
• Reject H_0 if $t_{cal} > t_{\alpha, n-2}$ or $t_{cal} < -t_{\alpha}$.	• Reject H_0 if $t_{cal} > t_{\alpha/2, n-2}$
• Otherwise accept H_0	• Otherwise accept H_0

Example 13.17: A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated? [*Delhi Univ., MCom. 1997*]

Solution: Let us take the null hypothesis that there is no significant difference in the sample and population correlation coefficients, that is,

$$H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \neq 0$$

Given $n = 27$, $df = n - 2 = 25$, $r = 0.42$. Applying t-test statistic to test the null hypothesis, H_0:

$$t = \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.42}{\sqrt{\{1 - (0.42)^2\}/(27 - 2)}}$$

$$= \frac{0.42}{0.908/5} = 2.312$$

Since the calculated value of $t = 2.312$ is more than the table value $t = 1.708$ at $\alpha = 0.05$ and $df = 25$, the null hypothesis is rejected. Hence it is likely that the variables in the population are not correlated.

Example 13.18: How many pairs of observations must be included in a sample so that an observed correlation coefficient of value 0.42 shall have a calculated value of t greater than 2.72?

Solution: Given, $r = 0.42$, $t = 2.72$. Applying t-test statistic, we get

$$\frac{r}{\sqrt{(1-r^2)/(n-2)}} = t \quad \text{or} \quad r^2 \times \frac{n-2}{1-r^2} = t^2$$

$$(0.42)^2 \times \frac{(n-2)}{1-(0.42)^2} = (2.72)^2$$

$$n - 2 = \frac{(2.72)^2\,[1-(0.42)^2]}{(0.42)^2} = \frac{7.3984(0.8236)}{0.1764}$$

$$= \frac{6.0933}{0.1764} = 34.542$$

$$n = 2 + 34.542 = 36.542 \cong 37$$

Hence, the sample size should be of 37 pairs of observations.

Example 13.19: To study the correlation between the stature of father and son, a sample of 1600 is taken from the universe of fathers and sons. The sample study gives the correlation between the two to be 0.80. Within what limits does it hold true for the universe?

Solution: Since the sample size is large, the standard error of the correlation coefficient is given by

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

Given, correlation coefficient, $r = 0.8$ and $n = 1600$. Thus

$$\text{Standard error } SE_r = \frac{1-(0.8)^2}{\sqrt{1600}} = \frac{1-0.64}{40} = \frac{0.36}{40} = 0.009$$

The limits within which the correlation coefficient should hold true is given by

$$r \pm 3SE_r = 0.80 \pm 3\,(0.009) \quad \text{or} \quad 0.773 \le r \le 0.827$$

13.6.2 Hypothesis Testing about Population Correlation Coefficient (Large Sample)

Since the distribution of sample correlation coefficient r is not normal and its probability curve is skewed in the neighbourhood of population correlation coefficient $\rho = \pm 1$, even for large sample size n, therefore we use Fisher's z-transformation for transforming r into z, using the formula:

$$z = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

The value of z can be seen for different values of r from the standard tables given in the Appendix.

Changing common logarithm to the base e to natural logarithm to the base 10 by multiplying with the constant 2.3026, that is,

$$\log_e x = 2.3026 \log_{10} x$$

where x is a positive integer. Thus the transformation formula becomes

$$z = \frac{1}{2}(2.3026)\log_{10}\frac{1+r}{1-r} = 1.1513 \log_{10}\frac{1+r}{1-r}$$

Fisher showed that the distribution of z is approximately normal with

$$\text{Mean } z_\rho = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} = 1.1513 \log_{10}\frac{1+\rho}{1-\rho}$$

and Standard deviation $\sigma_z = \dfrac{1}{\sqrt{n-3}}$

This approximation is useful for large sample sizes, say $n > 50$. However it can also be used for small sample sizes but at least $n \ge 10$.

The Z-test statistic to test the null hypothesis $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$ is given by

$$Z = \frac{z - z_\rho}{\sigma_z} = \frac{z - z_\rho}{1/\sqrt{n-3}}$$

where σ is the standard error of Z.

Decision rule

- Accept null hypothesis H_0 if $|Z_{cal}| <$ Table value of $Z_{\alpha/2}$
- Otherwise reject H_0

13.6.3 Hypothesis Testing about the Difference between Two Independent Correlation Coefficients

The formula for Z-test statistic given above can be generalized to test the hypothesis of two correlation coefficients r_1 and r_2 derived from two independent samples as follows:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where $z_1 = \frac{1}{2}\log_e \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+r_1}{1-r_1}$, and

$$z_2 = \frac{1}{2}\log_e \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+r_2}{1-r_2}$$

are approximately normally distributed with zero mean and unit standard deviation.

The null hypothesis H_0 is accepted if the absolute value $|Z_{cal}|$ is less than the table value $Z_{\alpha/2}$. Otherwise reject H_0.

Example 13.20: What is the probability that a correlation coefficient of 0.75 or less arises in a sample of 30 pairs of observations from a normal population in which the true correlation is 0.9?

Solution: Given, $r = 0.75$, $n = 30$, and $\rho = 0.9$. Applying Fisher's z-transformation, we get

$$z = 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1.75}{0.25}$$

$$= 1.1513 [\log_{10} 1.75 - \log_{10} 0.25]$$

$$= 1.1513 (0.24304 - \bar{1}.39794) = 0.973$$

The distribution of z is normal around the true population correlation value $\rho = 0.9$. Therefore

$$\text{Mean } z_\rho = 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.90}{1-0.90} = 1.1513 \log_{10} \frac{1.90}{0.10}$$

$$= 1.1513 [\log_{10} 1.90 - \log_{10} 0.10] = 1.1513 (0.27875 + 1) = 1.47$$

Thus, the Z-test statistic is given by

$$Z = \frac{|z - z_\rho|}{\sigma_z} = \frac{|z - z_\rho|}{1/\sqrt{n-3}} = \frac{|0.973 - 1.47|}{1/\sqrt{30-3}} = 0.498 \times 5.196 = 2.59$$

Hence $\rho (r \leq 0.75) = P[Z \leq 2.59] = 1 - 0.9952 = 0.0048.$

Example 13.21: Test the significance of the correlation $r = 0.5$ from a sample of size 18 against hypothesized population correlation $\rho = 0.70$.

Solution: Let us take the null hypothesis that the difference is not significant, that is,

$$H_0 : \rho = 0.70 \quad \text{and} \quad H_1 : \rho \neq 0.70$$

Given $n = 18$, $r = 0.5$. Applying z-transformation, we have

$$z = 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$$= 1.1513 \log_{10} \frac{1.50}{0.5} = 1.1513 \log_{10} 3$$

$$= 1.1513 \ (0.4771) = 0.5492$$

and $$\text{Mean } z_\rho = 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.70}{1-0.70}$$

$$= 1.1513 \log_{10} \frac{1.70}{0.30} = 1.1513 \log_{10} 5.67$$

$$= 1.1513 \ (0.7536) = 0.8676$$

Applying Z-test statistic, we get

$$Z = \frac{|z - z_\rho|}{\sigma_z} = \frac{|z - z_\rho|}{1/\sqrt{n-3}} = |z - z_\rho| \sqrt{n-3}$$

$$= |\ 0.5492 - 0.8676\ |\ \sqrt{15} = 0.3184\ (3.872) = 1.233$$

Since calculated value of $Z_{cal} = 1.233$ is less than its table value $Z_{\alpha./2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence we conclude that the difference (if any) is due to sampling error.

Example 13.22: Two independent samples of size 23 and 21 pairs of observations were analysed and their coefficient of correlation was found as 0.5 and 0.8, respectively. Do these value differ significantly?

Solution: Let us take the null hypothesis that two values do not differ significantly, that is, the samples are drawn from the same population.

Given $n_1 = 23$, $r_1 = 0.5$; $n_2 = 28$, $r_2 = 0.8$. Applying Z-test statistic as follows:

$$Z = \frac{|z_1 - z_2|}{\sigma_{z_1 - z_2}}; \qquad z_1 = 1.1513 \log_{10} \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$$= \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{|0.55 - 1.10|}{\sqrt{\frac{1}{20} + \frac{1}{25}}} = 1.1513 \log_{10} 3 = 0.55$$

$$= \frac{0.55}{0.30} = 1.833 \qquad z_2 = 1.1513 \log_{10} \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+0.8}{1-0.8}$$

$$= 1.1513 \log_{10} 9 = 1.10$$

Since the calculated value of $Z_{cal} = 1.833$ is less than its table value $Z_{\alpha/2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence the difference in correlation values is not significant.

Conceptual Questions 13A

1. What is the meaning of the coefficient of correlation?
2. Explain the meaning and significance of the term correlation. [*Delhi Univ., MBA, 1995*]
3. What is meant by 'correlation'? Distinguish between positive, negative, and zero correlation. [*Ranchi Univ., MBA, 1996*]
4. What are the numerical limits of r^2 and r? What does it mean when r equals one? zero? minus one?
5. What is correlation? Clearly explain its role with suitable illustration from simple business problems. [*Delhi Univ., MBA, 1997*]

6. What is the relationship between the coefficient of determination and the coefficient of correlation? How is the coefficient of determination interpreted?
7. Does correlation always signify a cause-and-effect relationship between the variables?
[*Osmania Univ., MBA, 1990*]
8. What information is provided by the coefficient of correlation of a sample? Why is it necessary to perform a test of a hypothesis for correlation?
9. When the result of a test of correlation is significant, what conclusion is drawn if r is positive? If r is negative?
10. What is the t-statistic that is used in a test for correlation? What is meant by the number of degrees of freedom in a test for correlation and how is it used?
11. What is coefficient of rank correlation? Bring out its usefulness. How does this coefficient differ from the coefficient of correlation? [*Delhi Univ., MBA, 2000*]
12. What is Spearman's rank correlation coefficient? How does it differ from Karl Pearson's coefficient of correlation?
13. (a) What is a scatter diagram? How do you interpret a scatter diagram?
(b) What is a scatter diagram? How does it help in studying the correlation between two variables, in respect of both its direction and degree?
[*Delhi Univ., MBA, 1999*]
14. Define correlation coefficient 'r' and give its limitations. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is (–) 0.60 if only drivers with at least one accident are considered?

Self-Practice Problems 13C

13.21 The correlation between the price of two commodities x and y in a sample of 60 is 0.68. Could the observed value have arisen
(a) from an uncorrelated population?
(b) from a population in which true correlation was 0.8?

13.22 The following data give sample sizes and correlation coefficients. Test the significance of the difference between two values using Fisher's z-transformation.

Sample Size	*Value of r*
5	0.870
12	0.560

13.23 A company wants to study the relationship between R&D expenditure (in Rs 1000's) and annual profit (in Rs 1000's). The following table presents the information for the last 8 years.

Year :	1988	87	86	85	84	83	82	81
R&D expenses :	9	7	5	10	4	5	3	2
Annual profit :	45	42	41	60	30	34	25	20

(a) Estimate the sample correlation coefficient.
(b) Test the significance of correlation coefficient at a α = 5 per cent level of significance.

13.24 Find the least value of r in a sample of 27 pairs from a bivariate normal population at α = 0.05 level of significance, where $t_{\alpha = 0.05}$ = 2.06 at df = 25.

13.25 A small retail business has determined that the correlation coefficient between monthly expenses and profits for the past year, measured at the end of each month, is r = 0.56. Assuming that both expenses and profits are approximately normal, test at α = 0.05 level of significance the null hypothesis that there is no correlation between them.

13.26 The manager of a small shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000's and analyzed the results. Has the rise been significant?

Week :	1	2	3	4	5	6
Sales :	2.69	2.62	2.80	2.70	2.75	2.81

Find the correlation coefficient between sales and week and test it for significance at α = 0.05.

Hints and Answers

13.21 (a) $z = 1.1513 \log_{10} \dfrac{1+r}{1-r} = 1.1513 \log_{10} \dfrac{1+0.68}{1-0.68}$

$= 1.1513 \log_{10} \dfrac{1.68}{0.32} = 0.829$

Standard error, $\sigma_z = \dfrac{1}{\sqrt{n-3}} = \dfrac{1}{\sqrt{57}} = 0.13$

Test statistic $Z = \dfrac{z - z_\rho}{\sigma_z} = \dfrac{0.829 - 0}{0.13} = 6.38$

Since deviation of z from z_ρ is 6 times more than σ_z, the hypothesis is not correct, that is, population is correlated.

$$\text{Mean } z_\rho = 1.1513 \log_{10} \frac{1+\rho}{1-\rho}$$

$$= 1.1513 \log_{10} \frac{1.8}{1.2} = 1.099$$

$$\therefore \quad Z = \frac{|z - z_\rho|}{\sigma_z} = \frac{|0.829 - 1.099|}{0.13} = 2.08 > 2$$

times standard error, ρ is likely to be less than 0.8.

13.22 Let H_0 : samples are drawn from the same population.

$$z_1 = 1.1513 \log_{10} \frac{1+r_1}{1-r_1}$$

$$= 1.1513 \log \frac{1+0.87}{1-0.87} = 1.333$$

$$z_2 = 1.1513 \log_{10} \frac{1+r_2}{1-r_2}$$

$$= 1.1513 \log_{10} \frac{1+0.56}{1-0.56} = 0.633$$

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} = \sqrt{\frac{1}{5-3} + \frac{1}{12-3}}$$

$$= 0.782$$

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{0.7}{0.782} = 0.895$$

Since the calculated value Z = 0.895 is less than its table value $Z_\alpha = 2.58$ at $\alpha = 0.01$ level of significance, H_0 is accepted.

13.23 (a) $r = 0.95$ (*b*) Let $H_0 : r = 0$ and $H_1 : r \neq 0$

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

$$= \frac{0.95}{\sqrt{\{1-(0.95)^2\}/(8-2)}} = 7.512$$

Since $t_{cal} = 7.512 > t_{\alpha/2} = 2.447$ for $df = 6$, the H_0 is rejected.

13.24 $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{r\sqrt{27-2}}{\sqrt{1-r}} = \dfrac{5r}{\sqrt{1-r}} > 2.06$

or $\quad |r| = 0.381$

13.25 $r = 0.560$ and $t_{cal} = 0.576$, H_0 is rejected.

13.26 $r = 0.656$ and $t_{cal} = 0.729$, H_0 is rejected.

Formulae Used

1. Karl Pearson's correlation coefficient

$$r = \frac{\text{Covariance between } x \text{ and } y}{\sigma_x \sigma_y}$$

- Deviation from actual mean

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

- Deviation from assumed mean

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$d_x = x - A$, $d_y = y - B$

A, B = constants

- Bivariate frequency distribution

$$r = \frac{n\Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{n\Sigma f d_x^2 - (\Sigma f d_x)^2}\sqrt{n\Sigma f d_y^2 - (\Sigma f d_y)^2}}$$

- Using actual values of x and y

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

2. Standard error of correlation coefficient, r

$$SE_r = \frac{1-r^2}{\sqrt{n}}$$

- Probable error of correlation coefficient, r

$$PE_r = 0.6745 \frac{1-r^2}{\sqrt{n}}$$

3. Coefficient of determination

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$$

4. Spearman's rank correlation coefficient

- Ranks are not equal

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

- Ranks are equal

$$R = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m_i^3 - m_i)\right]}{n(n^2-1)}$$

$t = 1, 2, \ldots$

5. Hypothesis testing

- Population correlation coefficient r for a small sample

$$t = \frac{r-\rho}{SE_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

- Population correlation coefficient for a large sample

$$Z = \frac{z - z_\rho}{\sigma_z} = \frac{z - z_\rho}{1/\sqrt{n-3}}$$

Chapter Concepts Quiz

True or False

1. There are several types of correlation coefficients, the selection of which is determined by the level of scaling of the two variables. (T/F)
2. When both variables use measured on an interval or ratio scale, Pearson's correlation coefficient is most appropriate. (T/F)
3. To use Pearson's correlation coefficient, it is assumed that both variables are continuous and normally distributed. (T/F)
4. When there is no linear association between two variables, the value of r will be close to zero. (T/F)
5. A correlation coefficient $r = -1$ represents a very low linear correlation. (T/F)
6. The coefficient of determination is the square of the correlation coefficient. (T/F)
7. As the correlation coefficient approaches zero, the possible error in linear prediction increases. (T/F)
8. The closer the correlation coefficient is to zero, the greater the predictive validity of a test. (T/F)
9. If a correlation coefficient for reliability of a test is close to 1, then the test is unreliable. (T/F)
10. Even a high correlation is not necessarily indicative of a casual relationship between two variables. (T/F)
11. As the value of r increases, the proportion of variability of one variable y that can be accounted for another variable x decreases. (T/F)
12. If the relationship between two variables is nonlinear, the value of the correlation coefficient must be negative.
13. Spearman's correlation coefficient is used where one or both variables are at least of interval scaling. (T/F)
14. A scatter diagram is used to help to decide if the relationship between two variables is linear or curvilinear. (T/F)
15. When calculating Spearman's correlation coefficient, Σd^2 is the sum of the square of the difference between the means (T/F)

Multiple Choice

16. A scatter diagram
    (a) is a statistical test (b) must be linear
    (c) must be curvilinear (d) is a graph of x and y values
17. If the relationship between variables x and y is linear, then the points on the scatter diagram
    (a) will fall exactly on a straight
    (b) will fall on a curve
    (c) must represent population parameters
    (d) are best represented by a straight line
18. If the relationship between x and y is positive, as variable y decreases, variable x
    (a) increases (b) decreases
    (c) remains same (d) changes linearly
19. In a 'negative' relationship
    (a) as x increases, y increases
    (b) as x decreases, y decreases
    (c) as x increases, y decreases
    (d) both (a) and (b)
20. The lowest strength of association is reflected by which of the following correlation coefficients?
    (a) 0.95 (b) − 0.60
    (c) − 0.35 (d) 0.29
21. The highest strength of association is reflected by which of the following correlation coefficients?
    (a) − 1.0 (b) − 0.95
    (c) 0.1 (d) 0.85
22. There is a high inverse association between measures 'overweight' and 'life expectancy'. A correlation coefficient consistent with the above statement is:
    (a) $r = 0.80$ (b) $r = 0.20$
    (c) $r = -0.20$ (d) $r = -0.80$
23. Of the following measurement levels, which is the required level for the valid calculation of the Pearson correlation coefficient
    (a) nominal (b) ordinal
    (c) internal (d) ratio
24. Of the following measurement levels, which is required for the valid calculation of the Spearman correlation coefficient?
    (a) nominal (b) ordinal
    (c) internal (d) ratio
25. There is a high direct association between measures of 'cigarette smoking' and 'lung damage'. The correlation coefficient consistent with the above statement is:
    (a) 0.30 (b) 0.80
    (c) − 0.80 (d) − 0.30
26. The correlation coefficient appropriate for establishing the degree of correlation between the two variables (assuming a linear relationship)
    (a) is determined by the sample size
    (b) is Spearman's R
    (c) is Pearson's r
    (d) both (b) and (c)
27. When deciding which measure of correlation to employ with a specific set of data, you should consider
    (a) whether the relationship is linear or nonlinear (b) the type of scale of measurement for each variable
    (c) both (a) and (b)
    (d) neither (a) nor (b)
28. The proportion of variance accounted for by the level of correlation between two variables is calculated by
    (a) $\bar{x}$ (b) r^2
    (c) Σx (d) not possible
29. The value of correlation coefficient
    (a) depends on the origin
    (b) depends on the unit of scale
    (c) depends on both origin and unit of scale
    (d) is independent with respect to origin and unit of scale

30. Which of the following statements is false?
(a) In a perfect positive correlation, each individual obtains the same z value on each variable
(b) Spearman's correlation coefficient is used when one or both variables are at least of interval scaling
(c) The range of the correlation coefficient is from -1 to $+1$
(d) A correlation of $r = 0.85$ implies a stronger association than $r = -0.70$

31. The strength of a linear relationship between two variables x and y is measured by
(a) r (b) r^2
(c) R^2 (d) b_{xy} or b_{yx}

32. If value of $r^2 = 0.64$, then what is the coefficient of correlation
(a) 0.40 (b) 0.04
(c) 0.80 (d) 0.08

33. If both dependent and independent variables increase in an estimating equation, then coefficient of correlation falls in the range.
(a) $-1 \le r \le 1$ (b) $0 \le r \le 1$
(c) $-3 \le r \le 3$ (d) none of these

34. If unexplained variation between variables x and y is 0.25, then r^2 is
(a) 0.25 (b) 0.50
(c) 0.75 (d) none of these

35. What type of relationhship between the two variables is indicated by the sign of r
(a) direct relation (b) indirect relation
(c) both (a) and (b) (d) none of these

Concepts Quiz Answers

1. T	**2.** T	**3.** T	**4.** T	**5.** F	**6.** T	**7.** T	**8.** F	**9.** F
10. T	**11.** F	**12.** F	**13.** F	**14.** T	**15.** F	**16.** (d)	**17.** (d)	**18.** (b)
19. (c)	**20.** (d)	**21.** (a)	**22.** (d)	**23.** (c)	**24.** (c)	**25.** (b)	**26.** (c)	**27.** (c)
28. (b)	**29.** (d)	**30.** (b)	**31.** (a)	**32.** (c)	**33.** (d)	**34.** (c)	**35.** (d)	

Review-Self Practice Problems

13.27 The following are the monthly figures of the advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after 2 months. Allowing for this time lag, calculate the coefficient of correlation.

Months	*Advertising Expenditure*	*Sales*	*Months*	*Advertising Expenditure*	*Sales*
Jan.	50	1200	July	140	2400
Feb.	60	1500	Aug.	160	2600
March	70	1600	Sep.	170	2800
April	90	2000	Oct.	190	2900
May	120	2200	Nov.	200	3100
June	150	2500	Dec.	250	3900

13.28 The coefficient of correlation between two variables x and y is 0.64. Their covariance is 16. The variance of x is 19. Find the standard deviation of y series.

13.29 Given $r = 0.8$, $\Sigma xy = 60$, $\sigma_y = 2.5$ and $\Sigma x^2 = 90$, find the number of observations, items. x and y are deviations from arithmetic mean.

[Delhi Univ., BCom, 1998]

13.30 Calculate the Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Comment on the value

Age	:	20	21	22	23	24	25
No. of students	:	500	400	300	240	200	160
Regular players	:	400	300	180	96	60	24

[Osmania Univ., MBA, 1998]

13.31 A survey regarding income and savings provided the following data:

Income (Rs)	*Saving (Rs)*			
	500	*1000*	*1500*	*2000*
40,000	8	4	—	—
6000	—	12	24	6
8000	—	9	7	2
10,000	—	—	10	5
12,000	—	—	9	4

Compute Karl Pearson's coefficient of correlation and interpret its value.

[Kurukshetra Univ., MBA, 1997]

13.32 A company gives on-the-job training to its salesmen, followed by a test. It is considering whether it should terminate the services of any salesman who does not do well in the test. Following data give the test scores and sales (in 1000 Rs) made by nine salesmen during the last one year

Test scores : 14 19 24 21 26 22 15 20 19
Sales : 31 36 48 37 50 45 33 41 39

Compute the coefficient of correlation between test scores and sales. Does it indicate that termination of the services of salesman with low test scores is justified?

[Madurai Univ., MBA, 1999]

13.33 Calculate the coefficient of correlation and its probable error from the following:

S.No.	Subject	Per cent Marks in Final Year Exams	Percent Marks in Sessionals
1	Hindi	75	62
2	English	81	68
3	Physics	70	65
4	Chemistry	76	60
5	Maths	77	69
6	Statistics	81	72
7	Botany	84	76
8	Zoology	75	72

13.34 Following figures give the rainfall in inches for the year and the production (in 100's kg) for the Rabi crop and Kharif crops. Calculate Karl Pearson's coefficient of correlation, between rainfall and total production

Rainfall	:	20	22	24	26	28	30	32
Rabi production	:	15	18	20	32	40	39	40
Kharif production	:	15	17	20	18	20	21	15

[Pune Univ., MBA, 1996]

13.35 President of a consulting firm is interested in the relationship between environmental work factors and the employees turnover rate. He defines environmental factors as those aspects of a job other than salary and benefits. He visited to similar plants and gave each plant a rating 1 to 25 on its environmental factors. He then obtained each plant's turnover rate (Annual in percentage) examined the relationship.

Environmental rating	:	11	19	7	12	13	10	16	22	14	12
Turnover rate	:	6	4	8	3	7	8	3	2	5	6

Compute the correlation coefficient between turnover rate and environmental rating and test it.

[IGNOU, 1996]

13.36 Sixteen companies in a state have been ranked according to profit earned during a particular financial year, and the working capital for that year. Calculate the rank correlation coefficient

Company	Rank(Profit)	Rank(Working capital)
A	1	13
B	2	16
C	3	14
D	4	15
E	5	10
F	6	12
G	7	4
H	8	11
I	9	5
J	10	9
K	11	8
L	12	3
M	13	1
N	14	6
O	15	7
P	16	2

13.37 Following are the percentage figures of expenditure incurred on clothing (in Rs 100's) and entertainment (in Rs 100's) by an average working class family in a period of 10 years

Year	:	1989	90	91	92	93	94	95	96	97	98
Expenditure on clothing	:	24	27	31	32	20	25	33	30	28	22
Expenditure on entertainment	:	11	8	5	3	13	10	2	7	9	2

Compute Spearman's rank correlation coefficient and comment on the result.

Hints and Answers

13.27 $r = 0.918$

13.28 $r = \dfrac{\Sigma xy}{n\sigma_x\sigma_y}$; $\sigma_x = \sqrt{9} = 3$;

$$0.64 = 16\,\frac{1}{3\sigma_y} \quad \text{or} \quad \sigma_y = 8.33$$

13.29 $r = \dfrac{\Sigma xy}{n\sigma_x\sigma_y}$ or $r^2 = \dfrac{(\Sigma xy)^2}{n^2\sigma_x^2\sigma_y^2}$;

$$(0.8)^2 = \frac{(60)^2}{n^2(90/n)\times 6.25} = \frac{3600}{90n\times 6.25};$$

$$n = 10$$

13.30 $r = -0.991$ **13.31** $r = 0.0522$

13.22 $r = 0.947$ **13.33** $r = 0.623$, $PE_r = 0.146$

13.34 $r = 0.917$ **13.35** $r = -0.801$

13.36 R = −0.8176 **13.37** R = −0.60

Chapter 14

The cause is hidden, but the result is known.

—Ovid

I never think of the future, it comes soon enough.

—Albert Einstein

Regression Analysis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- use simple linear regression for building models to business data.
- understand how the method of least squares is used to predict values of a dependent (or response) variable based on the values of an independent (or explanatory) variable.
- measure the variability (residual) of the dependent variable about a straight line (also called regression line) and examine whether regression model fits to the data.

14.1 INTRODUCTION

In Chapter 13 we introduced the concept of statistical relationship between two variables such as: level of sales and amount of advertising; yield of a crop and the amount of fertilizer used; price of a product and its supply, and so on. The relationship between such variables indicate the degree and direction of their association, but fail to answer following question:

- Is there any functional (or algebraic) relationship between two variables? If yes, can it be used to estimate the most likely value of one variable, given the value of other variable?

The statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called *regression analysis*. The variable whose value is estimated using the algebraic equation is called *dependent* (or *response*) *variable* and the variable whose value is used to estimate this value is called *independent* (*regressor* or *predictor*) *variable*. The linear algebraic equation used for expressing a dependent variable in terms of independent variable is called *linear regression equation*.

The term regression was used in 1877 by Sir Francis Galton while studying the relationship between the height of father and sons. He found that though 'tall father has tall sons', the average height of sons of tall father is x above the general height, the average height of sons is $2x/3$ above the general height. Such a fall in the average height was described by Galton as 'regression to mediocrity'. However, the theory of Galton is not universally applicable and the term regression is applied to other types of variables in business and economics. The term regression in the literary sense is also referred as 'moving backward'.

The basic differences between correlation and regression analysis are summarized as follows:

1. Developing an algebraic equation between two variables from sample data and predicting the value of one variable, given the value of the other variable is referred to as regression analysis, while measuring the strength (or degree) of the relationship between two variables is referred as correlation analysis. The sign of correlation coefficient indicates the nature (direct or inverse) of relationship between two variables, while the absolute value of correlation coefficient indicates the extent of relationship.
2. Correlation analysis determines an association between two variables x and y but not that they have a cause-and-effect relationship. Regression analysis, in contrast to correlation, determines the cause-and-effect relationship between x and y, that is, a change in the value of independent variable x *causes* a corresponding change (*effect*) in the value of dependent variable y if all other factors that affect y remain unchanged.
3. In linear regression analysis one variable is considered as dependent variable and other as independent variable, while in correlation analysis both variables are considered to be independent.
4. *The coefficient of determination r^2 indicates the proportion of total variance in the dependent variable that is explained or accounted for by the variation in the independent variable.* Since value of r^2 is determined from a sample, its value is subject to sampling error. Even if the value of r^2 is high, the assumption of a linear regression may be incorrect because it may represent a portion of the relationship that actually is in the form of a curve.

14.2 ADVANTAGES OF REGRESSION ANALYSIS

The following are some important advantages of regression analysis:

1. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable with respect to the regression line. Smaller the variance and error of estimate, the closer the pair of values (x, y) fall about the regression line and better the line fits the data, that is, a good estimate can be made of the value of variable y. When all the points fall on the line, the standard error of estimate equals zero.
3. When the sample size is large ($df \geq 29$), the interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either x or y. The magnitude of r^2 remains the same regardless of the values of the two variables.

14.3 TYPES OF REGRESSION MODELS

The primary objective of regression analysis is the development of a *regression model* to explain the association between two or more variables in the given population. A regression model is the mathematical equation that provides prediction of value of dependent variable based on the known values of one or more independent variables.

The particular form of regression model depends upon the nature of the problem under study and the type of data available. However, each type of association or relationship can be described by an equation relating a dependent variable to one or more independent variables.

14.3.1 Simple and Multiple Regression Models

If a regression model characterizes the relationship between a dependent y and only one independent variable x, then such a regression model is called a *simple regression model.* But if more than one independent variables are associated with a dependent variable,

then such a regression model is called a *multiple regression model.* For example, sales turnover of a product (a dependent variable) is associated with multiple independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors, and so on. Now if we want to estimate possible sales turnover with respect to only one of these independent variables, then it is an example of a simple regression model, otherwise multiple regression model is applicable.

14.3.2 Linear and Nonlinear Regression Models

If the value of a dependent (response) variable y in a regression model tends to increase in direct proportion to an increase in the values of independent (predictor) variable x, then such a regression model is called a *linear model.* Thus, it can be assumed that the mean value of the variable y for a given value of x is related by a straight-line relationship. Such a *relationship is called simple linear regression model* expressed with respect to the population parameters β_0 and β_1 as:

$$E(y|x) = \beta_0 + \beta_1 x \qquad (14\text{-}1)$$

where β_0 = y-intercept that represents mean (or average) value of the dependent variable y when $x = 0$

β_1 = slope of the regression line that represents the expected change in the value of y (either positive or negative) for a unit change in the value of x.

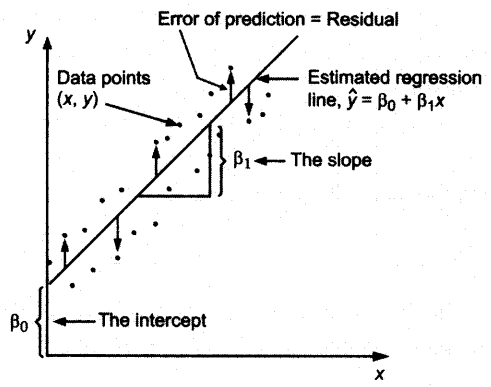


Figure 14.1
Straight Line Relationship

The intercept β_0 and the slope β_1 are *unknown regression coefficients.* The equation (14-1) requires to compute the values of β_0 and β_1 to predict average values of y for a given value of x. However Fig. 14.1 presents a scatter diagram where each pair of values (x_i, y_i) represents a point in a two-dimensional coordinate system. Although the mean or average value of y is a linear function of x, but not all values of y fall exactly on the straight line rather fall around the line.

Since few points do not fall on the regression line, therefore values of y are not exactly equal to the values yielded by the equation: $E(y|x) = \beta_0 + \beta_1 x$, also called *line of mean deviations of observed y value from the regression line.* This situation is responsible for *random error* (also called *residual variation or residual error*) in the prediction of y values for given values of x. In such a situation, it is likely that the variable x does not explain all the variability of the variable y. For instance, sales volume is related to advertising, but if other factors related to sales are ignored, then a regression equation to predict the sales volume (y) by using annual budget of advertising (x) as a predictor will probably involve some error. Thus for a fixed value of x, the actual value of y is determined by the *mean value function plus a random error term as follows:*

$$\begin{aligned} y &= \text{Mean value function} + \text{Deviation} \\ &= \beta_0 + \beta_1 x + e = E(y) + e \end{aligned} \qquad (14\text{-}2)$$

where e is the *observed random error.* This equation is also called *simple probabilitic linear regression model.*

The error component e allows each individual value of y to deviate from the line of means by a small amount. The random errors corresponding to different observations (x_i, y_i) for $i=1, 2, ..., n$ are assumed to follow a normal distribution with mean zero and (unknown) constant standard deviation.

The term e in the expression (14-2) is called the *random error* because its value, associated with each value of variable y, is assumed to vary unpredictably. The extent of this error for a given value of x is measured by the error variance σ_e^2. Lower the value of σ_e^2, better is the fit of linear regression model to a sample data.

If the line passing through the pair of values of variables x and y is curvilinear, then the relationship is called *nonlinear*. A nonlinear relationship implies a varying absolute change in the dependent variable with respect to changes in the value of the independent variable. A nonlinear relationship is not very useful for predictions.

In this chapter, we shall discuss methods of simple linear regression analysis involving single independent variable, whereas those involving two or more independent variables will be discussed in Chapter 15.

14.4 ESTIMATION : THE METHOD OF LEAST SQUARES

To estimate the values of regression coefficients β_0 and β_1, suppose a sample of n pairs of observations (x_1, y_1), (x_2, y_2), . . ., (x_n, y_n) is drawn from the population under study. A method that provides the best linear unbiased estimates of β_0 and β_1 is called the *method of least squares*. The estimates of β_0 and β_1 should result in a straight line that is *'best fit'* to the data points. The straight line so drawn is referred to as *'best fitted' (least squares or estimated) regression line because the sum of the squares of the vertical deviations (difference between the acutal values of y and the estimated values $\hat{y}$ predicted from the fitted line) is as small as possible.*

Using equation (14-2), we may express given n observations in the sample data as:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{or} \quad e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for all } i$$

Mathematically, we intend to minimize

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let b_0 and b_1 be the least-squares estimators of β_0 and β_1 respectively. The least-squares estimators b_0 and b_1 must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{b_0 b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{b_0 b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$

After simplifying these two equations, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \qquad (14\text{-}3)$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Equations (14-3) are called the *least-squares normal equations*. The values of least squares estimators b_0 and b_1 can be obtained by solving equations (14-3). Hence the *fitted* or *estimated regression line* is given by:

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ (called y hat) is the value of y lying on the fitted regression line for a given x value and $e_i = y_i - \hat{y}_i$ is called the *residual* that describes the error in fitting of the regression line to the observation y_i. The fitted value $\hat{y}$ is also called the *predicted value of* y because if actual value of y is not known, then it would be predicted for a given value of x using the estimated regression line.

Remark: The sum of the residuals is zero for any least-squares regression line. Since $\Sigma y_i = \Sigma \hat{y}_i$, therefore so $\Sigma e_i = 0$.

14.5 ASSUMPTIONS FOR A SIMPLE LINEAR REGRESSION MODEL

To make valid statistical inference using regression analysis, we make certain assumptions about the bivariate population from which a sample of paired observations is drawn and the manner in which observations are generated. These assumptions form the basis for application of simple linear regression models. Figure 14.2 illustrates these assumptions.

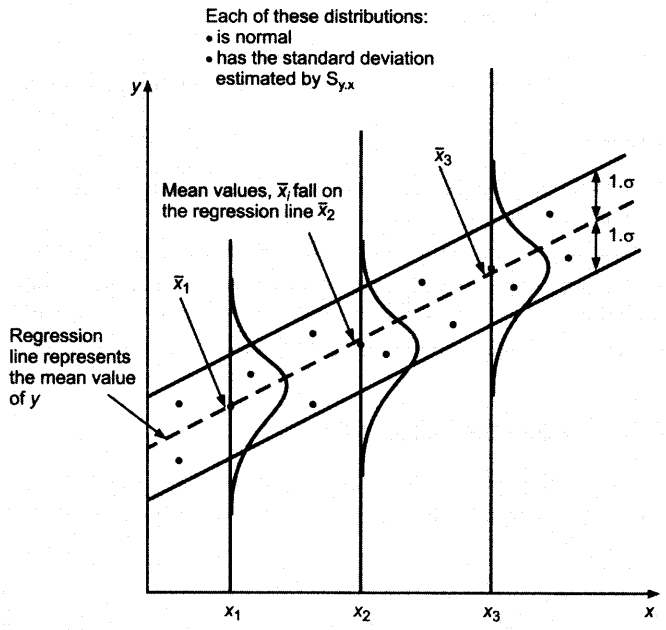


Figure 14.2
Graphical Illustration of Assumptions in Regression Analysis

Assumptions

1. The relationship between the dependent variable y and independent variable x exists and is linear. The average relationship between x and y can be described by a simple linear regression equation $y = a + bx + e$, where e is the deviation of a particular value of y from its expected value for a given value of independent variable x.
2. For every value of the independent variable x, there is an expected (or mean) value of the dependent variable y and these values are normally distributed. The mean of these normally distributed values fall on the line of regression.
3. The dependent variable y is a continuous random variable, whereas values of the independent variable x are fixed values and are not random.
4. The sampling error associated with the expected value of the dependent variable y is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The errors are not related with each other in successive observations.
5. The standard deviation and variance of expected values of the dependent variable y about the regression line are constant for all values of the independent variable x within the range of the sample data.
6. The value of the dependent variable cannot be estimated for a value of an independent variable lying outside the range of values in the sample data.

14.6 PARAMETERS OF SIMPLE LINEAR REGRESSION MODEL

The fundamental aim of regression analysis is to determine a regression equation (line) that makes sense and fits the representative data such that the error of variance is as small as possible. This implies that the regression equation should adequately be used for prediction. J. R. Stockton stated that

- *The device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression.*

The two variables x and y which are correlated can be expressed in terms of each other in the form of straight line equations called *regression equations*. Such lines should be able to provide the best fit of sample data to the population data. The algebraic expression of regression lines is written as:

- The regression equation of y on x

$$y = a + bx$$

  is used for estimating the value of y for given values of x.

- Regression equation of x on y

$$x = c + dy$$

  is used for estimating the value of x for given values of y.

Remarks

1. When variables x and y are correlated perfectly (either positive or negative) these lines coincide, that is, we have only one line.
2. Higher the degree of correlation, nearer the two regression lines are to each other.
3. Lesser the degree of correlation, more the two regression lines are away from each other. That is, when $r = 0$, the two lines are at right angle to each other.
4. Two linear regression lines intersect each other at the point of the average value of variables x and y.

14.6.1 Regression Coefficients

To estimate values of population parameter β_0 and β_1, under certain assumptions, the fitted or estimated regression equation representing the straight line regression model is written as:

$$\hat{y} = a + bx$$

where $\hat{y}$ = estimated average (mean) value of dependent variable y for a given value of independent variable x.

a or b_0 = y-intercept that represents average value of $\hat{y}$

b = slope of regression line that represents the expected change in the value of y for unit change in the value of x

To determine the value of $\hat{y}$ for a given value of x, this equation requires the determination of two unknown constants a (intercept) and b (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable y for a given value of independent variable x.

The particular values of a and b define a specific linear relationship between x and y based on sample data. The coefficient 'a' represents the *level of fitted line* (i.e., the distance of the line above or below the origin) when x equals zero, whereas coefficient 'b' represents the *slope of the line* (a measure of the change in the estimated value of y for a one-unit change in x).

The regression coefficient 'b' is also denoted as:

- b_{yx} (*regression coefficient of y on x*) in the regression line, $y = a + bx$
- b_{xy} (*regression coefficient of x on y*) in the regression line, $x = c + dy$

Properties of regression coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, that is, $r = \sqrt{b_{yx} \times b_{xy}}$.
2. If one regression coefficient is greater than one, then other regression coefficient must be less than one, because the value of correlation coefficient r cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the case of opposite sign of two regression coefficients.

4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \times 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients b_{xy} and b_{yx} is more than or equal to the correlation coefficient r, that is, $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then the arithmetic mean of these two values is $(-0.664 - 0.234)/2 = -0.449$, and this value is more than the value of $r = -0.394$.
6. Regression coefficients are independent of origin but not of scale.

14.7 METHODS TO DETERMINE REGRESSION COEFFICIENTS

Following are the methods to determine the parameters of a fitted regression equation.

14.7.1 Least Squares Normal Equations

Let $\hat{y} = a + bx$ be the least squares line of y on x, where $\hat{y}$ is the estimated average value of dependent variable y. The line that minimizes the sum of squares of the deviations of the observed values of y from those predicted is the best fitting line. Thus the sum of residuals for any least-square line is minimum, where

$$L = \Sigma(y - \hat{y})^2 = \Sigma\{y - (a + bx)\}^2; \quad a, b = \text{constants}$$

Differentiating L with respect to a and b and equating to zero, we have

$$\frac{\partial L}{\partial a} = -2\Sigma\{y - (a + bx)\} = 0$$

$$\frac{\partial S}{\partial b} = -2\Sigma\{y - (a + bx)\}x = 0$$

Solving these two equations, we get the same set of equations as equations (14-3)

$$\Sigma y = na + b\Sigma x \qquad (14\text{-}4)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

where n is the total number of pairs of values of x and y in a sample data. The equations (14-4) are called *normal equations* with respect to the regression line of y on x. After solving these equations for a and b, the values of a and b are substituted in the regression equation, $y = a + bx$.

Similarly if we have a least squares line $\hat{x} = c + dy$ of x on y, where $\hat{x}$ is the estimated mean value of dependent variable x, then the normal equations will be

$$\Sigma x = nc + d\Sigma y$$
$$\Sigma xy = n\Sigma y + d\Sigma y^2$$

These equations are solved in the same manner as described above for constants c and d. The values of these constants are substituted to the regression equation $x = c + dy$.

Alternative method to calculate value of constants

Instead of using the algebraic method to calculate values of a and b, we may directly use the results of the solutions of these normal equation.

The gradient 'b' (regression coefficient of y on x) and 'd' (regression coefficient of x on y) are calculated as:

$$b = \frac{S_{xy}}{S_{xx}}, \quad \text{where} \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$\text{and} \quad d = \frac{S_{yx}}{S_{yy}}, \quad \text{where} \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y\right)^2$$

Since the regression line passes through the point $(\bar{x}, \bar{y})$, the mean values of x and y and the regression equations can be used to find the value of constants a and c as follows:

$a = \bar{y} - b\bar{x}$ for regression equation of y on x

$c = \bar{x} - d\bar{y}$ for regression equation of x on y

The calculated values of a, b and c, d are substituted in the regression line $y = a + bx$ and $x = c + dy$ respectively to determine the exact relationship.

Example 14.1: Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 per cent in advertising expenditure.

Firm	*Annual Percentage Increase in Advertising Expenditure*	*Annual Percentage Increase in Sales Revenue*
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

Solution: Assume sales revenue (y) is dependent on advertising expenditure (x). Calculations for regression line using following normal equations are shown in Table 14.1

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Table 14.1: Calculation for Normal Equations

Sales Revenue y	*Advertising Expenditure,* x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
40	56	524	373

Approach 1 (Normal Equations):

$\Sigma y = na + b\Sigma x$ or $40 = 8a + 56b$

$\Sigma xy = a\Sigma x + b\Sigma x^2$ or $373 = 56a + 524b$

Solving these equations, we get

$a = 0.072$ and $b = 0.704$

Substituting these values in the regression equation

$$y = a + bx = 0.072 + 0.704x$$

For $x = 7.5\%$ or 0.075 increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704\ (0.075) = 0.1248 \text{ or } 12.48\%$$

Approach 2 (Short-cut method):

$$b = \frac{S_{xy}}{S_{xx}} = \frac{93}{132} = 0.704,$$

where $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 373 - \frac{40 \times 56}{8} = 93$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 524 - \frac{(56)^2}{8} = 132$$

The intercept 'a' on the y-axis is calculated as:

$$a = \bar{y} - b\bar{x} = \frac{40}{8} - 0.704 \times \frac{56}{8} = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of $a = 0.072$ and $b = 0.704$ in the regression equation, we get

$$y = a + bx = 0.072 + 0.704\,x$$

For $x = 0.075$, we have $y = 0.072 + 0.704\ (0.075) = 0.1248$ or 12.48%.

Example 14.2: The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs 1000's and analysed the results

Week :	1	2	3	4	5	6
Sales :	2.69	2.62	2.80	2.70	2.75	2.81

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

Solution: Assume sales (y) is dependent on weeks (x). Then the normal equations for regression equation: $y = a + bx$ are written as:

$$\Sigma y = na + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 14.2.

Table 14.2: Calculations of Normal Equations

Week (x)	*Sales (y)*	x^2	xy
1	2.69	1	2.69
2	2.62	4	5.24
3	2.80	9	8.40
4	2.70	16	10.80
5	2.75	25	13.75
6	2.81	36	16.86
21	16.37	91	57.74

The gradient 'b' is calculated as:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{0.445}{17.5} = 0.025; \quad S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 57.74 - \frac{21 \times 16.37}{6} = 0.445$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 91 - \frac{(21)^2}{6} = 17.5$$

The intercept 'a' on the y-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 \times \frac{21}{6}$$

$$= 2.728 - 0.025 \times 3.5 = 2.64$$

Substituting the values $a = 2.64$ and $b = 0.025$ in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For $x = 7$, we have $y = 2.64 + 0.025\ (7) = 2.815$

Hence the expected sales during the 7th week is likely to be Rs 2.815 (in Rs 1000's).

14.7.2 Deviations Method

Calculations to least squares normal equations become lengthy and tedious when values of x and y are large. Thus the following two methods may be used to reduce the computational time.

(a) **Deviations Taken from Actual Mean Values of x and y** If deviations of actual values of variables x and y are taken from their mean values $\bar{x}$ and $\bar{y}$, then the regression equations can be written as:

- Regression equation of y on x

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

where b_{yx} = regression coefficient of y on x.

The value of b_{yx} can be calculated using the using the formula

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

- Regression equation of x on y

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

where b_{xy} = regression coefficient of x on y.

The value of b_{xy} can be calculated formula

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

(b) **Deviations Taken from Assumed Mean Values for x and y** If mean value of either x or y or both are in fractions, then we must prefer to take deviations of actual values of variables x and y from their assumed means.

- Regression equation of y on x

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

where $b_{yx} = \dfrac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2}$

n = number of observations

$d_x = x - A$; A is assumed mean of x

$d_y = y - B$; B is assumed mean of y

- Regression equation of x on y

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

where $b_{xy} = \dfrac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2}$

n = number of observations

$d_x = x - A$; A is assumed mean of x

$d_y = y - B$; B is assumed mean of y

(c) **Regression Coefficients in Terms of Correlation Coefficient** If deviations are taken from actual mean values, then the values of regression coefficients can be alternatively calculated as follows:

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\text{Covariance}(x, y)}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2} = \frac{\text{Covariance}(x, y)}{\sigma_y^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

Example 14.3: The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (in Rs 1000's)

Salesmen	:	A	B	C	D	E	F	G	H	I
Test scores	:	50	60	50	60	80	50	80	40	70
Weekly sales	:	30	60	40	50	60	30	70	50	60

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman in 65, what would be his expected weekly sales. *[HP Univ., MCom, 1996]*

Solution: Assume weekly sales (y) as dependent variable and test scores (x) as independent variable. Calculations for the following regression equation are shown in Table 14.3.

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

Table 14.3: Calculation for Regression Equation

Weekly Sales, x	$dx = x - 60$	d_x^2	*Test Score, y*	$dy = y - 50$	d_y^2	$d_x d_y$
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	20	400	60	10	100	200
50	-10	100	30	-20	400	200
80	20	400	70	20	400	400
40	-20	400	50	0	0	0
70	10	100	60	10	100	100
540	0	1600	450	0	1600	1200

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{540}{9} = 60; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{450}{9} = 50$

$$b_{yx} = \frac{\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1200}{1600} = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75\,(x - 60) \text{ or } y = 5 + 0.75x$$

For test score $x = 65$ of salesman, we have

$$y = 5 + 0.75\,(65) = 53.75$$

Hence we conclude that the weekly sales is expected to be Rs 53.75 (in Rs 1000's) for a test score of 65.

Example 14.4: A company is introducing a job evaluation scheme in which all jobs are graded by points for skill, responsibility, and so on. Monthly pay scales (Rs in 1000's) are then drawn up according to the number of points allocated and other factors such as experience and local conditions. To date the company has applied this scheme to 9 jobs:

Job	:	A	B	C	D	E	F	G	H	I
Points	:	5	25	7	19	10	12	15	28	16
Pay (Rs)	:	3.0	5.0	3.25	6.5	5.5	5.6	6.0	7.2	6.1

(a) Find the least squares regression line for linking pay scales to points.

(b) Estimate the monthly pay for a job graded by 20 points.

Solution: Assume monthly pay (y) as the dependent variable and job grade points (x) as the independent variable. Calculations for the following regression equation are shown in Table 14.4.

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

Table 14.4: Calculations for Regression Equation

Grade Points, x	$d_x = x - 15$	d_x^2	*Pay Scale, y*	$d_y = y - 5$	d_y^2	$d_x d_y$
5	-10	100	3.0	-2.0	4	20
25	10	100	(5.0) ← B	0	0	0
7	-8	64	3.25	-1.75	3.06	14
19	4	16	6.5	1.50	2.25	6
10	-5	25	5.5	0.50	0.25	-2.5
12	-3	9	5.6	0.60	0.36	-1.8
(15) ← A	0	0	6.0	1.00	1.00	0
28	13	169	7.2	2.2	4.84	28.6
16	1	1	6.1	1.1	1.21	1.1
137	2	484	48.15	3.15	16.97	65.40

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{137}{9} = 15.22; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{48.15}{9} = 5.35$

Since mean values $\bar{x}$ and $\bar{y}$ are non-integer value, therefore deviations are taken from assumed mean as shown in Table 14.4.

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9\times 65.40 - 2\times 3.15}{9\times 484 - (2)^2} = \frac{582.3}{4352} = 0.133$$

Substituting values in the regression equation, we have

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \text{or} \quad y - 5.35 = 0.133\,(x - 15.22) = 3.326 + 0.133x$$

(b) For job grade point x=20, the estimated average pay scale is given by

$$y = 3.326 + 0.133x = 3.326 + 0.133\,(20) = 5.986$$

Hence, likely monthly pay for a job with grade points 20 is Rs 5986.

Example 14.5: The following data give the ages and blood pressure of 10 women.

Age	:	56	42	36	47	49	42	60	72	63	55
Blood pressure	:	147	125	118	128	145	140	155	160	149	150

(a) Find the correlation coefficient between age and blood pressure.

(b) Determine the least squares regression equation of blood pressure on age.

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

[Ranchi Univ. MBA; South Gujarat Univ., MBA, 1997]

Solution: Assume blood pressure (y) as the dependent variable and age (x) as the independent variable. Calculations for regression equation of blood pressure on age are shown in Table 14.5.

Table 14.5: Calculations for Regression Equation

Age, x	$d_x = x - 49$	d_x^2	Blood, y	$d_y = y - 145$	d_y^2	$d_x d_y$
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) ← A	0	0	(145) ← B	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
522	32	1202	1417	−33	1813	1115

(a) Coefficient of correlation between age and blood pressure is given by

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

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

We may conclude that there is a high degree of positive correlation between age and blood pressure.

(b) The regression equation of blood pressure on age is given by

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

$$\bar{x} = \frac{\Sigma x}{n} = \frac{522}{10} = 52.2; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{1417}{10} = 141.7$$

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

Substituting these values in the above equation, we have

$$y - 141.7 = 1.11\ (x - 52.2) \text{ or } y = 83.758 + 1.11x$$

This is the required regression equation of y on x.

(*c*) For a women whose age is 45, the estimated average blood pressure will be

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

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

Example 14.6: The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of readymade men's wear—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 1600 and Rs 625 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement would you suggest the General Sales Manager of the enterprise to incur to meet his target of sales?

[*Kurukshetra Univ., MBA, 1998*]

Solution: Assume advertisement expenditure (y) as the dependent variable and sales (x) as the independent variable. Then the regression equation advertisement expenditure on sales is given by

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

Given $r = 0.8$, $\sigma_x = 40$, $\sigma_y = 25$, $\bar{x} = 45{,}000$, $\bar{y} = 30{,}000$. Substituting these value in the above equation, we have

$$(y - 30{,}000) = 0.8\ \frac{25}{40}\ (x - 45{,}000) = 0.5\ (x - 45{,}000)$$

$$y = 30{,}000 + 0.5x - 22{,}500 = 7500 + 0.5x$$

When a sales target is fixed at $x = 80{,}000$, the estimated amount likely to the spent on advertisement would be

$$y = 7500 + 0.5 \times 80{,}000 = 7500 + 40{,}000 = \text{Rs } 47{,}500$$

Example 14.7: You are given the following information about advertising expenditure and sales:

	Advertisement (x) (*Rs in lakh*)	*Sales* (y) (*Rs in lakh*)
Arithmetic mean, $\bar{x}$	10	90
Standard deviation, σ	3	12

Correlation coefficient = 0.8

(a) Obtain the two regression equations.

(b) Find the likely sales when advertisement budget is Rs 15 lakh.

(c) What should be the advertisement budget if the company wants to attain sales target of Rs 120 lakh? [*Kumaon Univ., MBA, 2000, MBA, Delhi Univ., 2002*]

Solution: (a) Regression equation of x on y is given by

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

Given $\bar{x} = 10$, $r = 0.8$, $\sigma_x = 3$, $\sigma_y = 12$, $\bar{y} = 90$. Substituting these values in the above regression equation, we have

$$x - 10 = 0.8 \frac{3}{12} (y - 90) \quad \text{or} \quad x = -8 + 0.2y$$

Regression equation of y on x is given by

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

$$y - 90 = 0.8 \frac{12}{3} (x - 10) \quad \text{or} \quad y = 58 + 3.2x$$

(b) Substituting $x = 15$ in regression equation of y on x. The likely average sales volume would be

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

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

(c) Substituting $y = 120$ in the regression equation of x on y. The likely advertisement budget to attain desired sales target of Rs 120 lakh would be

$$x = -8 + 0.2y = -8 + 0.2\,(120) = 16$$

Hence, the likely advertisement budget of Rs 16 lakh should be sufficient to attain the sales target of Rs 120 lakh.

Example 14.8: In a partially destroyed laboratory record of an analysis of regression data, the following results only are legible:

Variance of $x = 9$

Regression equations : $8x - 10y + 66 = 0$ and $40x - 18y = 214$

Find on the basis of the above information:

(a) The mean values of x and y,

(b) Coefficient of correlation between x and y, and

(c) Standard deviation of y. [*Pune Univ., MBA, 1996; CA May 1999*]

Solution: (a) Since two regression lines always intersect at a point $(\bar{x}, \bar{y})$ representing mean values of the variables involved, solving given regression equations to get the mean values $\bar{x}$ and $\bar{y}$ as shown below:

$$8x - 10y = -66$$
$$40x - 18y = 214$$

Multiplying the first equation by 5 and subtracting from the second, we have

$$32y = 544 \quad \text{or} \quad y = 17, \text{ i.e. } \bar{y} = 17$$

Substituting the value of y in the first equation, we get

$$8x - 10(17) = -66 \quad \text{or} \quad x = 13, \text{ that is, } \bar{x} = 13$$

(b) To find correlation coefficient r between x and y, we need to determine the regression coefficients b_{xy} and b_{yx}.

Rewriting the given regression equations in such a way that the coefficient of dependent variable is less than one at least in one equation.

$$8x - 10y = -66 \text{ or } 10y = 66 + 8x \text{ or } y = \frac{66}{10} + \frac{8}{10} x$$

That is, $b_{yx} = 8/10 = 0.80$

$$40x - 18y = 214 \text{ or } 40x = 214 + 18y \text{ or } x = \frac{214}{40} + \frac{18}{40} y$$

That is, $b_{xy} = 18/40 = 0.45$

Hence coefficient of correlation r between x and y is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.80} = 0.60$$

(c) To determine the standard deviation of y, consider the formula:

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad \text{or} \quad \sigma_y = \frac{b_{yx}\, \sigma_x}{r} = \frac{0.80 \times 3}{0.6} = 4$$

Example 14.9: 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: The regression equation to read off values of P for various values S is given by

$$P = a + bS \quad \text{or} \quad (P - \bar{P}) = r\frac{\sigma_p}{\sigma_s}(S - \bar{S})$$

Given $\bar{P} = 100$, $\bar{S} = 103$, $\sigma_p = 8$, $\sigma_s = 4$, $r = 0.4$. Substituting these values in the above equation, we have

$$P - 100 = 0.4\frac{8}{4}(S - 103) \text{ or } P = 17.6 + 0.8\,S$$

This equation cannot be used to read off values of S for various values of P. Thus to read off values of S for various values of P we use another regression equation of the form:

$$S = c + dP \quad \text{or} \quad S - \bar{S} = \frac{\sigma_s}{\sigma_p}(P - \bar{P})$$

Substituting given values in this equation, we have

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

Example 14.10: The two regression lines obtained in a correlation analysis of 60 observations are:

$$5x = 6x + 24 \quad \text{and} \quad 1000y = 768x - 3708$$

What is the correlation coefficient and what is its probable error? Show that the ratio of the coefficient of variability of x to that of y is 5/24. What is the ratio of variances of x and y?

Solution: Rewriting the regression equations

$$5x = 6y + 24 \quad \text{or} \quad x = \frac{6}{5}y + \frac{24}{5}$$

That is, $b_{xy} = 6/5$

$$1000y = 768x - 3708 \quad \text{or} \quad y = \frac{768}{1000}x - \frac{3708}{1000}$$

That is, $b_{yx} = 768/1000$

We know that $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{768}{1000}$, therefore

$$b_{xy}b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} = 0.9216$$

Hence $r = \sqrt{0.9216} = 0.96$.

Since both b_{xy} and b_{yx} are positive, the correlation coefficient is positive and hence $r = 0.96$.

$$\text{Probable error of } r = 0.6745\,\frac{1 - r^2}{\sqrt{n}} = 0.6745\,\frac{1 - (0.96)^2}{\sqrt{60}}$$

$$= \frac{0.0528}{7.7459} = 0.0068$$

Solving the given regression equations for x and y, we get $\bar{x} = 6$ and $\bar{y} = 1$ because regression lines passed through the point $(\bar{x}, \bar{y})$.

$$\text{Since } r\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } 0.96\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } \frac{\sigma_x}{\sigma_y} = \frac{6}{5 \times 0.96} = \frac{5}{4}$$

$$\text{Also the ratio of the coefficient of variability} = \frac{\sigma_x/\bar{x}}{\sigma_y/\bar{y}} = \frac{\bar{y}}{\bar{x}} \cdot \frac{\sigma_x}{\sigma_y} = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}.$$

14.7.3 Regression Coefficients for Grouped Sample Data

The method of finding the regression coefficients b_{xy} and b_{yx} would be little different than the method discussed earlier for the case when data set is grouped or classified into frequency distribution of either variable x or y or both. The values of b_{xy} and b_{yx} shall be calculated using the formulae:

$$b_{xy} = \frac{n\Sigma d_x d_y - \Sigma f d_x \Sigma f d_y}{n\Sigma f d_y^2 - (\Sigma f d_y)^2} \times \frac{h}{k}$$

$$b_{yx} = \frac{n\Sigma f d_x d_y - \Sigma f d_x \Sigma f d_y}{n\Sigma f d_x^2 - (\Sigma f d_x)^2} \times \frac{k}{h}$$

where h = width of the class interval of sample data on x variable

k = width of the class interval of sample data on y variable

Example 14.11: The following bivariate frequency distribution relates to sales turnover (Rs in lakh) and money spent on advertising (Rs in 1000's). Obtain the two regression equations

Sales Turnover	*Advertising Budget (Rs in 1000's)*			
(Rs in lakh)	*50–60*	*60–70*	*70–80*	*80–90*
20– 50	2	1	2	5
50– 80	3	4	7	6
80–110	1	5	8	6
110–140	2	7	9	2

Estimate (a) the sales turnover corresponding to advertising budget of Rs 1,50,000, and (b) the advertising budget to achieve a sales turnover of Rs 200 lakh.

Solution: Let x and y represent sales turnover and advertising budget respectively. Then the regression equation for estimating the sales turnover (x) on advertising budget (y) is expressed as:

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

where $b_{xy} = \dfrac{n\Sigma f d_x d_y - \Sigma f d_x \Sigma f d_y}{n\Sigma f d_y^2 - (\Sigma f d_y)^2}$

Table 14.6: Calculations for Regression Coefficients

Sales x	*m.v.*	y *m.v* d_y / d_x	*Advertising Budget* 50–60 (55, −2)	60–70 (65, −1)	70–80 (75, 0)	80–90 (85, 1)	f	fd_x	fd_x^2	$fd_x d_y$
20–50	35	1	2 (4)	1 (1)	2 —	5 (−5)	10	−10	10	0
50–80	65	0	3 —	4 —	7 —	6 —	20	0	0	0
80–110	95	1	1 (−2)	5 (−5)	8 —	6 (6)	20	20	20	−1
110–140	125	2	2 (−8)	7 (−14)	9 —	2 (4)	20	40	80	−18
		f	8	17	26	19	$n = 70$	50 = $\Sigma f d_x$	110 = $\Sigma f d_x^2$	−19 $\Sigma f d_x d_y$
		fd_y	−16	−17	0	19	−14 = $\Sigma f d_y$			
		fd_y^2	32	17	0	19	68 = $\Sigma f d_y^2$			
		$fd_x d_y$	−6	−18	0	5	−19 = $\Sigma f d_x d_y$			

Similarly, the regression equation for estimating the advertising budget (y) on sales turnover of Rs 200 lakh is written as:

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

where $b_{yx} = \dfrac{n\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_x^2 - (\Sigma fd_x)^2}$

The calculations for regression coefficients b_{xy} and b_{yx} are shown in Table 14.6.

$$\bar{x} = A + \frac{\Sigma fd_x}{n} \times h = 65 + \frac{50}{70} \times 30 = 65 + 21.428 = 86.428$$

$$\bar{y} = B + \frac{\Sigma fd_y}{n} \times k = 75 - \frac{14}{70} \times 10 = 75 - 2 = 73$$

$$b_{xy} = \frac{n\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_y^2 - (\Sigma fd_y)^2} \times \frac{h}{k} = \frac{70 \times -19 - (50)(-14)}{70 \times 68 - (-14)^2} \times \frac{30}{10}$$

$$= \frac{-1330 + 700}{4760 - 196} \times \frac{30}{10} = \frac{-18{,}900}{45{,}640} = -0.414$$

$$b_{yx} = \frac{n\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{k}{h} = \frac{70 \times -19 - (50)(-14)}{70 \times 110 - (50)^2} \times \frac{10}{30}$$

$$= \frac{-1330 + 700}{7700 - 2500} \times \frac{10}{30} = \frac{-6300}{1{,}56{,}000} = -0.040$$

Substituting these values in the two regression equations, we get

(a) Regression equation of sales turnover (x) to advertising budget (y) is:

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

$x - 86.428 = -0.414\,(y - 73)$, or $x = 116.65 - 0.414y$

For $y = 150$, we have $x = 116.65 - 0.414 \times 150 =$ Rs 54.55 lakh

(*b*) Regression equation of advertising budget (y) on sales turnover (x) is:

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

$y - 73 = -0.040\,(x - 86.428)$ or $y = 76.457 - 0.04x$

For $x = 200$, we have $y = 76.457 - 0.04\,(200) =$ Rs 68.457 thousand.

Self-Practice Problems 14A

14.1 The following calculations have been made for prices of twelve stocks (x) at the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares (y). From these calculations find the regression equation of price of stocks on the volume of sales of shares.

$\Sigma x = 580$, $\Sigma y = 370$, $\Sigma xy = 11494$,
$\Sigma x^2 = 41658$, $\Sigma y^2 = 17206$.

[*Rajasthan Univ., MCom, 1995*]

14.2 A survey was conducted to study the relationship between expenditure (in Rs) on accommodation (x) and expenditure on food and entertainment (y) and the following results were obtained:

	Mean	*Standard Deviation*
• Expenditure on accommodation	173	63.15
• Expenditure on food and entertainment	47.8	22.98

Coefficient of correlation $r = 0.57$

Write down the regression equation and estimate the expenditure on food and entertainment if the expenditure on accommodation is Rs 200.

[*Bangalore Univ., BCom, 1998*]

14.3 The following data give the experience of machine operators and their performance ratings 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 ratings (y)	:	87	88	89	68	78	80	75	83

Calculate the regression lines of performance ratings on experience and estimate the probable performance if an operator has 7 years experience.

[*Jammu Univ., MCom; Lucknow Univ., MBA, 1996*]

14.4 A study of prices of a certain commodity at Delhi and Mumbai yield the following data:

	Delhi	*Mumbai*
• Average price per kilo (Rs)	2.463	2.797
• Standard deviation	0.326	0.207
• Correlation coefficient between prices at Delhi and Mumbai $r = 0.774$		

Estimate from the above data the most likely price (a) at Delhi corresponding to the price of Rs 2.334 per kilo at Mumbai (b) at Mumbai corresponding to the price of 3.052 per kilo at Delhi.

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

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

Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92, (b) the test score of a worker whose productivity index is 75. *[Delhi Univ., MBA, 2001]*

14.6 A company wants to assess the impact of R&D expenditure (Rs in 1000s) on its annual profit; (Rs in 1000's). The following table presents the information for the last eight years:

Year	*R & D expenditure*	*Annual profit*
1991	9	45
1992	7	42
1993	5	41
1994	10	60
1995	4	30
1996	5	34
1997	3	25
1998	2	20

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of Rs 1,00,000 as R&D expenditure.

[Jodhpur Univ., MBA, 1998]

14.7 Obtain the two regression equations from the following bivariate frequency distribution:

Sales Revenue	*Advertising Expenditure (Rs in thousand)*			
(Rs in lakh)	*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 (a) the sales corresponding to advertising expenditure of Rs 50,000, (b) the advertising expenditure for a sales revenue of Rs 300 lakh, (c) the coefficient of correlation. *[Delhi Univ., MBA, 2002]*

14.8 The personnel manager of an electronic manufacturing company devises a manual 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. The 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 squares regression equation of production rating on test score. *[Delhi Univ., MBA, 200]*

14.9 Find the regression equation showing the capacity utilization on production from the following data:

	Average	*Standard Deviation*
• Production (in lakh units) :	35.6	10.5
• Capacity utilization (in percentage) :	84.8	8.5
• Correlation coefficient $r = 0.62$		

Estimate the production when the capacity utilization is 70 per cent.

[Delhi Univ., MBA, 1997; Pune Univ., MBA, 1998]

14.10 Suppose that you are interested in using past expenditure on R&D 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 spent on R&D (in lakh of rupees) 5 years ago and y is the amount spent on R&D (in lakh of rupees) in the current year:

x :	30	50	20	80	10	20	20	40
y :	50	80	30	110	20	20	40	50

(a) Find the regression equation of y on x.

(b) If a firm is chosen randomly and $x = 10$, can you use the regression to predict the value of y? Discuss.

[Madurai-Kamraj Univ., MBA, 2000]

14.11 The following data relates to the scores obtained by a salesmen of a company in an intelligence test and their weekly sales (in Rs. 1000's):

Salesman intelligence :	A	B	C	D	E	F	G	H	I
Test score :	50	60	50	60	80	50	80	40	70
Weekly sales :	30	60	40	50	60	30	70	50	60

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

[HP Univ., M.com., 1996]

14.12 Two random variables have the regression equations:

$$3x + 2y - 26 = 0 \quad \text{and} \quad 6x + y - 31 = 0$$

(a) Find the mean values of x and y and coefficient of correlation between x and y.

(b) If the varaince of x is 25, then find the standard deviation of y from the data.

[MD Univ., M.Com., 1997; Kumaun Univ., MBA, 2001]

14.13 For a given set of bivariate data, the fiollowing results were obtained

$$\bar{x} = 53.2,\ \bar{y} = 27.9,$$

Regression coefficient of y on $x = -1.5$, and Regression coefficient of x and $y = -0.2$.

Find the most probable value of y when $x = 60$.

14.14 In trying to evaluate the effectiveness in its advertising compaign, a firm compiled the following information:

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

Year	*Adv. expenditure (Rs. 1000's)*	*Sales (in lakhs Rs)*
1996	12	5.0
1997	15	5.6
1998	17	5.8
1999	23	7.0
2000	24	7.2
2001	38	8.8
2002	42	9.2
2003	48	9.5

[*Bharathidasan Univ., MBA, 2003*]

Hints and Answers

14.1 $\bar{x} = \Sigma x/n = 580/12 = 48.33;$

$\bar{y} = \Sigma y/n = 370/12 = 30.83$

$$b_{xy} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n(\bar{y})^2} = \frac{11494 - 12 \times 48.33 \times 30.83}{17206 - 12(30.83)^2}$$

$$= -1.102$$

Regression equation of x on y:

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

$$x - 48.33 = -1.102\,(y - 30.83)$$

or $x = 82.304 - 1.102y$

14.2 Given $\bar{x} = 172$, $\bar{y} = 47.8$, $\sigma_x = 63.15$, $\sigma_y = 22.98$, and $r = 0.57$

Regression equation of food and entertainment (y) on accomodation (x) is given by

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

$$y - 47.8 = 0.57\,\frac{22.98}{63.15}(x - 173)$$

or $y = 11.917 + 0.207x$

For $x = 200$, we have $y = 11.917 + 0.207(200) = 53.317$

14.3 Let the experience and performance rating be represented by x and y respectively.

$\bar{x} = \Sigma x/n = 80/8 = 10;\quad \bar{y} = \Sigma y/n = 648/8 = 81$

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{247}{218} = 1.133;$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$

Regression equation of y on x

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

or $y - 81 = 1.133\,(x - 10)$

or $y = 69.67 + 1.133x$

When $x = 7$, $y = 69.67 + 1.133\,(7) = 77.60 \cong 78$

14.4 Let price at Mumbai and Delhi be represented by x and y, respectively

(a) Regression equation of y on x

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

$$y - 2.463 = 0.774\,\frac{0.326}{0.207}(x - 2.797)$$

For $x =$ Rs 2.334, the price at Delhi would be $y =$ Rs 1.899.

(b) Regression on equation of x on y

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

or $x - 2.791 = 0.774\,\frac{0.207}{0.326}\,(y - 2.463)$

For $y =$ Rs 3.052, the price at Mumbai would be $x =$ Rs 3.086.

14.5 Let aptitude score and productivity index be represented by x and y respectively.

$\bar{x} = \Sigma x/n = 650/10 = 65;\ \bar{y} = \Sigma y/n = 650/10 = 65$

$$b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - \Sigma(d_y)^2} = \frac{1044}{1752} = 0.596;$$

where $d_x = x - \bar{x};\ dy = y - \bar{y}$

(a) Regression equation of x on y

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

or $x - 65 = 0.596\,(y - 65)$

or $x = 26.26 + 0.596y$

When $y = 75$, $x = 26.26 + 0.596\,(75) = 70.96 \cong 71$

(b) $$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1044}{894} = 1.168$$

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

or $y - 65 = 1.168\,(x - 65)$

or $y = -10.92 + 1.168x$

When $x = 92$, $y = -10.92 + 1.168\,(92) = 96.536 \cong 97$

14.6 Let R&D expenditure and annual profit be denoted by x and y respectively

$\bar{x} = \Sigma x/n = 40/8 = 5.625$; $\bar{y} = \Sigma y/n = 297/8 = 37.125$

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8\times 238 - (-3)(1)}{8\times 57 - (-3)^2}$$

$$= 4.266;$$

where $d_x = x - 6$, $d_y = y - 37$

Regression equation of annual profit on R&D expenditure

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

$$y - 37.125 = 4.26\,(x - 5.625)$$

or $y = 13.163 + 4.266x$

For x = Rs 1,00,000 as R&D expenditure, we have from above equation y = Rs 439.763 as annual profit.

14.7 Let sales revenue and advertising expenditure be denoted by x and y respectively

$$\bar{x} = A + \frac{\Sigma fd_x}{n}\times h = 150 + \frac{12}{66}\times 50 = 159.09$$

$$\bar{y} = B + \frac{\Sigma fd_y}{n}\times k = 30 - \frac{26}{66}\times 10 = 26.06$$

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

$$= \frac{66(-14) - 12(-26)}{66(100) - (-26)^2}\times\frac{50}{10} = -0.516$$

(a) Regression equation of x on y

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

$$x - 159.09 = -0.516\,(y - 26.06)$$

or $x = 172.536 - 0.516y$

For $y = 50$, $x = 147.036$

(b) Regression equation of y on x

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

$$= \frac{66(-14) - 12(-26)}{66(70) - (12)^2}\times\frac{10}{50} = -0.027.$$

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

$$y - 26.06 = -0.027\,(x - 159.09)$$

$$y = 30.355 - 0.027x$$

For $x = 300$, $y = 22.255$

(c) $r = \sqrt{b_{xy}\times b_{yx}} = -\sqrt{0.516\times 0.027} = -0.1180$

14.8 Let test score and production rating be denoted by x and y respectively.

$\bar{x} = \Sigma x/n = 612/10 = 61.2$;

$\bar{y} = \Sigma y/n = 622/10 = 62.2$

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10\times 3213 - 2\times 2}{10\times 3554 - (2)^2} = 0.904$$

Regression equation of production rating (y) on test score (x) is given by

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

$$y - 62.2 = 0.904\,(x - 61.2)$$

$$y = 6.876 + 0.904x$$

14.9 Let production and capacity utilization be denoted by x and y, respectively.

(a) Regression equation of capacity utilization (y) on production (x)

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

$$y - 84.8 = 0.62\,\frac{8.5}{10.5}\,(x - 35.6)$$

$$y = 66.9324 + 0.5019x$$

(b) Regression equation of production (x) on capacity utilization (y)

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

$$x - 35.6 = 0.62\,\frac{10.5}{8.5}\,(y - 84.8)$$

$$x = -29.3483 + 0.7659y$$

When $y = 70$, $x = -29.3483 + 0.7659(70) = 24.2647$

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

14.10 $\bar{x} = \Sigma x/n = 270/8 = 33.75$; $\bar{y} = \Sigma y/n = 400/8 = 50$

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8\times 4800 - 6\times 0}{8\times 3592 - (6)^2}$$

$$= 1.338;$$

where $d_x = x - 33$ and $d_y = y - 50$

Regression equation of y on x

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

$$y - 50 = 1.338\,(x - 33.75)$$

$$y = 4.84 + 1.338x$$

For $x = 10$, $y = 18.22$

14.11 Let intelligence test score be denoted by x and weekly sales by y

$\bar{x} = 540/9 = 60$; $\bar{y} = 450/9 = 50$,

$$b_{yx} = \frac{n\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9\times 1200}{9\times 1600} = 0.75$$

Regression equation of y on x :

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

$$y - 50 = 0.75\,(x - 60)$$

$$y = 5 + 0.75x$$

For $x = 65$, $y = 5 + 0.75\,(65) = 53.75$

14.12 (a) Solving two regression lines:

$$3x + 2y = 6 \quad \text{and} \quad 6x + y = 31$$

we get mean values as $\bar{x} = 4$ and $\bar{y} = 7$

(b) Rewritting regression lines as follows:

$$3x + 2y = 26 \quad \text{or} \quad y = 13 - (3/2)x,$$

So $b_{yx} = -3/2$

$$6x + y = 31 \quad \text{or} \quad x = 31/6 - (1/6)y,$$

So $b_{xy} = -1/6$

Correlation coefficient,

$$r = \sqrt{b_{xy}\times b_{yx}} = -\sqrt{(3/2)(1/6)} = -0.5$$