Given, Var(x) = 25, so $\sigma_x = 5$. Calculate σ_y using the formula:

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

or $$-\frac{3}{2} = 0.5\frac{\sigma_y}{5} \quad \text{or} \quad \sigma_y = 15$$

14.13 The regression equation of y on x is stated as:

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

Given, $\bar{x} = 53.20$; $\bar{y} = 27.90$, $b_{yx} = -1.5$; $b_{xy} = -0.2$

Thus $y - 27.90 = -1.5(x - 53.20)$

or $y = 107.70 - 1.5x$

For $x = 60$, we have $y = 107.70 - 1.5(60) = 17.7$

Also $$r = \sqrt{b_{yx} \times b_{xy}} = -\sqrt{1.5 \times 0.2} = -0.5477$$

14.14 Let advertising expenditure and sales be denoted by x and y respectively.

$\bar{x} = \Sigma x/n = 217/8 = 27.125$; $\bar{y} = \Sigma y/n = 58.2/8 = 7.26$

$$b_{yx} = \frac{n\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{n\Sigma d_x^2 - (\Sigma dx)^2}$$

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

Thus regression equation of y on x is:

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

or $y - 7.26 = 0.125(x - 27.125)$

$y = 3.86 + 0.125x$

When $x = 60$, the estimated value of $y = 3.869 + 0.125(60) = 11.369$

14.8 STANDARD ERROR OF ESTIMATE AND PREDICTION INTERVALS

The pattern of dot points on a scatter diagram is an indicator of the relationship between two variables x and y. Wide scatter or variation of the dot points about the regression line represents a poor relationship. But a very close scatter of dot points about the regression line represents a close relationship between two variables. The variability in observed values of dependent variable y about the regression line is measured in terms of *residuals*. A residual is defined as the difference between an observed value of dependent variable y and its estimated (or fitted) value $\hat{y}$ determined by regression equation for a given value of the independent variable x. The residual about the regression line is given by

$$\text{Residual } e_i = y_i - \hat{y}_i$$

The residual values e_i are plotted on a diagram with respect to the least squares regression line $\hat{y} = a + bx$. These residual values represent error of estimation for individual values of dependent variable and are used to estimate, the variance σ^2 of the error term. In other words, residuals are used to estimate the amount of variation in the dependent variable with respect to least squares regression line. Here it should be noted that the variations are not the variations (deviations) of observations from the mean value in the sample data set, rather these variations are the vertical distances of every observation (dot point) from the least squares line as shown in Fig. 14.3.

Since sum of the residuals is zero, therefore it is not possible to determine the total amount of error by summing the residuals. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing them. That is

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftarrow \text{Error or Residual sum of squares}$$

This quantity is called the *sum of squares of errors* (*SSE*).

The estimate of variance of the error term σ_e^2 or $S_{y.x}^2$ is obtained as follows:

$$S_{yx}^2 \text{ or } \hat{\sigma}_e^2 = \frac{\Sigma e_i^2}{n-2} = \frac{\Sigma (y_i - y_i)^2}{n-2} = \frac{SSE}{n-2}$$

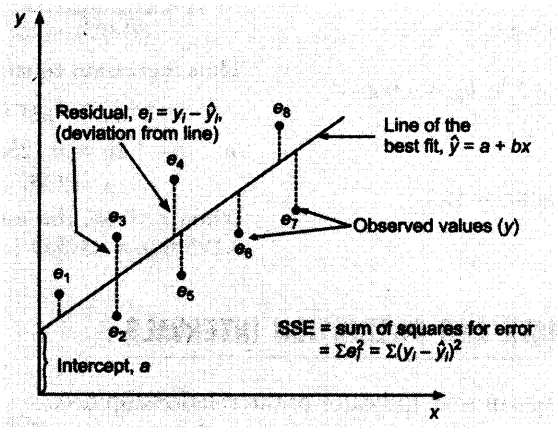
The denominator, $n - 2$ represents the *error* or *residual degrees of freedom* and is determined by subtracting from sample size n the number of parameters β_0 and β_1 that are estimated by the sample parameters a and b in the least squares equation. The subscript 'yx' indicates that the standard deviation is of dependent variable y, given (or conditional) upon independent variable x.

The *standard error of estimate* S_{yx} also called *standard deviation of the error term t* measures the variability of the observed values around the regression line, i.e. the amount

by which the $\hat{y}$ values are away from the sample y values (dot points). In other words, S_{yx} is based on the deviations of the sample observations of y-values from the least squares line or the estimated regression line of $\hat{y}$ values. The standard deviation of error about the least squares line is defined as:

$$S_{yx} \text{ or } \sigma_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} \tag{14-4}$$

Figure 14.3
Residuals



To simplify the calculations of S_{yx}, generally the following formula is used

$$S_{yx} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

The variance S_{yx}^2 measures how the least squares line 'best fits' the sample y-values. A large variance and standard error of estimate indicates a large amount of scatter or dispersion of dot points around the line. Smaller the value of S_{yx}, the closer the dot points (y-values) fall around the regression line and better the line fits the data and describes the better average relationship between the two variables. When all dot points fall on the line, the value of S_{yx} is zero, and the relationship between the two variables is perfect.

A smaller variance about the regression line is considered useful in predicting the value of a dependent variable y. In actual practice, some variability is always left over about the regression line. It is important to measure such variability due to the following reasons:

(i) This value provides a way to determine the usefulness of the regression line in predicting the value of the dependent variable.

(ii) This value can be used to construct interval estimates of the dependent variable.

(iii) Statistical inferences can be made about other components of the problem.

Figure 14.4 displays the distribution of conditional average values of y about a least squares regression line for given values of independent variable x. Suppose the amount of deviation in the values of y given any particular value of x follow normal distribution. Since average value of y changes with the value of x, we have different normal distributions of y-values for every value of x, each having same standard deviation. When a relationship between two variables x and y exists, the standard deviation (also called *standard error of estimate*) is less than the standard deviation of all the x-values in the population computed about their mean.

Based on the assumptions of regression analysis, we can describe sampling properties of the sample estimates such as a, b, and S_{yx}, as these vary from sample to sample. Such knowledge is useful in making statistical inferences about the relationship between the two variables x and y.

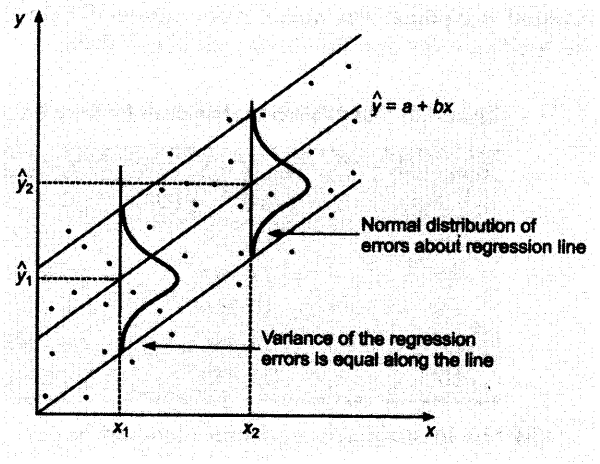


Figure 14.4
Regression Line Showing the Error Variance

The standard error of estimate can also be used to determine an approximate interval estimate based on sample data ($n < 30$) for the value of the dependent variable y for a given value of the independent variable x as follows:

$$\text{Approximate interval estimate} = \hat{y} \pm t_{df}\, S_{yx}$$

where value of t is obtained using t-distribution table based upon a chosen probability level. The interval estimate is also called a *prediction interval.*

Example 14.12: The following data relate to advertising expenditure (Rs in lakh) and their corresponding sales (Rs in crore)

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

(a) Find the equation of the least squares line fitting the data.

(b) Estimate the value of sales corresponding to advertising expenditure of Rs 30 lakh.

(c) Calculate the standard error of estimate of sales on advertising expenditure.

Solution: Let the advertising expenditure be denoted by x and sales by y.

(a) The calculations for the least squares line are shown in Table 14.7

Table 14.7: Calculations for Least Squares Line

Adv. Expenditure, x	$d_x = x - 16$	d_x^2	Sales y	$d_y = y - 20$	d_y^2	$d_x d_y$
10	−6	36	14	−6	36	36
12	−4	16	17	−3	9	12
15	−1	1	23	3	9	−3
23	7	49	25	5	25	35
20	4	16	21	1	1	4
80	0	118	100	0	80	84

$$\bar{x} = \Sigma x/n = 80/5 = 16; \quad \bar{y} = \Sigma y/n = 100/5 = 20$$

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{5\times 84}{5\times 118} = 0.712$$

(a) Regression equation of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 20 = 0.712\,(x - 16)$$
$$y = 8.608 + 0.712\,x$$

where parameter $a = 8.608$ and $b = 0.712$.

Table 14.8 gives the fitted values and the residuals for the data in Table 14.7. The fitted values are obtained by substituting the value of x into the regression equation (equation for the least squares line). For example, 8.608 + 0.712(10) = 15.728. The

residual is equal to the actual value minus fitted value. The residuals indicate how well the least squares line fits the actual data values.

Table 14.8: Fitted Values and Residuals for Sample Data

Value, x	Fitted Value $y = 8.608 + 0.712x$	Residuals
10	15.728	– 5.728
12	17.152	– 5.152
15	19.288	– 4.288
23	24.984	– 1.984
20	22.848	– 2.848

(b) The least squares equation obtained in part (a) may be used to estimate the sales turnover corresponding to the advertising expenditure of Rs 30 lakh as:

$$\hat{y} = 8.608 + 0.712x = 8.608 + 0.712\ (30) = \text{Rs } 29.968 \text{ crore}$$

(c) Calculations for standard error of estimate $S_{y \cdot x}$ of sales (y) on advertising expenditure (x) are shown in Table 14.9.

Table 14.9: Calculations for Standard Error of Estimate

x	y	y^2	xy
10	14	196	140
12	17	289	204
15	23	529	345
23	25	625	575
20	21	441	420
80	100	2080	1684

$$S_{y \cdot x} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} = \sqrt{\frac{2080 - 8.608 \times 100 - 0.712 \times 1684}{5-2}}$$

$$= \sqrt{\frac{2080 - 860.8 - 1199}{3}} = 2.594$$

14.8.1 Coefficient of Determination: Partitioning of Total Variation

The objective of regression analysis is to develop a regression model that best fits the sample data, so that the residual variance $S^2_{y.x}$ is as small as possible. But the value of $S^2_{y.x}$ depends on the scale with which the sample y-values are measured. This drawback with the calculation of $S^2_{y.x}$ restricts its interpretation unless we consider the units in which the y-values are measured. Thus, we need another measure of fit called *coefficient of determination* that is not affected by the scale with which the sample y-values are measured. *It is the proportion of variability of the dependent variable, y accounted for or explained by the independent variable, x,* i.e. it measures how well (i.e. strength) the regression line fits the data. The coefficient of determination is denoted by r^2 and its value ranges from 0 to 1. A particular r^2 value should be interpreted as high or low depending upon the use and context in which the regression model was developed. The coefficient of determination is given by

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\text{Residual variation in response variable } y\text{-values from least-squares line}}{\text{Total variance of } y\text{-values}}$$

where SST = total sum of square deviations (or total variance) of sampled response variable y-values from the mean value of y.

$$= S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

SSE = sum of squares of error or *unexplained variation* in response variable y-values from the least squares line due to sampling errors, i.e. it measures the residual variation in the data that is not explained by predictor variable x

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

SSR = sum of squares of regression *or explained variation* is the sample values of response variable y accounted for or explained by variation among x-values

= SST – SSE

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = a \sum_{i=1}^{n} y_i + b \sum_{i=1}^{n} x_i y_i - n(\bar{y})^2$$

The three variations associated with the regression analysis of a data set are shown in Fig 14.5. Thus

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{S_{yx}^2}{S_y^2}; \quad S_{y \cdot x} = S_y \sqrt{1 - r^2}$$

where $\frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$ = fraction of the total variation that is explained or accounted for

$S_{y \cdot x} = \frac{\Sigma(y - \hat{y})^2}{n - 2}$, variance of response variable y-values from the least squares line

$S_y^2 = \frac{1}{n - 2} \Sigma(y - \bar{y})^2$, total variance of response variable y-values

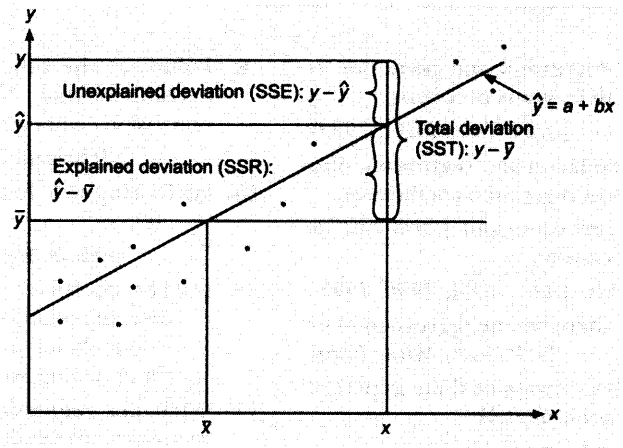


Figure 14.5
Relationship Between Three Types of Variations

Since the formula of r^2 is not convenient to use therefore an easy formula for the sample coefficient of determination is given by

$$r^2 = \frac{a\Sigma y + b\Sigma xy - n(\bar{y})^2}{\Sigma y^2 - n(\bar{y})^2} \quad \leftarrow \text{Short-cut method}$$

For example, the coefficient of determination that indicates the extent of relationship between sales revenue (y) and advertising expenditure (x) is calculated as follows from Example 14.1:

$$r^2 = \frac{a\Sigma y + b\Sigma xy - n(\bar{y})^2}{\Sigma y^2 - n(\bar{y})^2} = \frac{0.072 \times 40 + 0.704 \times 373 - 8(5)^2}{270 - 8(5)^2}$$

$$= \frac{2.88 + 262.592 - 200}{270 - 200} = \frac{65.47}{70} = 0.9352$$

The value $r^2 = 0.9352$ indicates that 93.52% of the variance in sales revenue is accounted for or statistically explained by advertising expenditure.

A comparison between bivariate correlation and regression summarized in Table 14-10 could provide further insight about the relationship between two variables x and y in the data set.

Table 14.10: Comparison between Linear Correlation and Regression

	Correlation	*Regression*
• Measurement level	Interval or ratio scale	Interval or ratio scale
• Nature of variables	Both continuous, and linearly related	Both continuous, and linearly related
• $x-y$ relationship	x and y are symmetric	y is dependent, x is independent; regression of x on y differs from y on x
• Correlation	$b_{xy} = b_{yx}$	Correlation between x and y is the same as the correlation between y and x
• Coefficient of determination	Explains common variance of x and y	Proportion of variability of x explained by its least-squares regression on y

Conceptual Questions 14A

1. (a) Explain the concept of regression and point out its usefulness in dealing with business problems. *[Delhi Univ., MBA, 1993]*
   (b) Distinguish between correlation and regression. Also point out the properties of regression coefficients.
2. Explain the concept of regression and point out its importance in business forecasting. *[Delhi Univ., MBA, 1990, 1998]*
3. Under what conditions can there be one regression line? Explain. *[HP Univ., MBA, 1996]*
4. Why should a residual analysis always be done as part of the development of a regression model?
5. What are the assumptions of simple linear regression analysis and how can they be evaluated?
6. What is the meaning of the standard error of estimate?
7. What is the interpretation of y-intercept and the slope in a regression model?
8. What are regression lines? With the help of an example illustrate how they help in business decision-making. *[Delhi Univ., MBA, 1998]*
9. Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients? *[Osmania Univ., MBA; Delhi Univ., MBA, 1999]*
10. (a) Distinguish between correlation and regression analysis. *[Dipl in Mgt., AIMA, Osmania Univ., MBA, 1998]*
    (b) The coefficient of correlation and coefficient of determination are available as measures of association in correlation analysis. Describe the different uses of these two measures of association.
11. What are regression coefficients? State some of the important properties of regression coefficients. *[Dipl in Mgt., AIMA, Osmania Univ., MBA, 1989]*
12. What is regression? How is this concept useful to business forecasting? *[Jodhpur Univ., MBA, 1999]*
13. What is the difference between a prediction interval and a confidence interval in regression analysis?
14. Explain what is required to establish evidence of a cause-and-effect relationship between y and x with regression analysis.

15. What technique is used initially to identify the kind of regression model that may be appropriate.
16. (a) What are regression lines? Why is it necessary to consider two lines of regression?
    (b) In case the two regression lines are identical, prove that the correlation coefficient is either + 1 or − 1. If two variables are independent, show that the two regression lines cut at right angles.
17. What are the purpose and meaning of the error terms in regression?
18. Give examples of business situations where you believe a straight line relationship exists between two variables. What would be the uses of a regression model in each of these situations.
19. 'The regression lines give only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in the estimate by calculating a quantity known as the standard error of estimate' Elucidute.
20. Explain the advantages of the least-squares procedure for fitting lines to data. Explain how the procedure works.

Formulae Used

1. Simple linear regression model
$$y = \beta_0 + \beta_1 x + e$$
2. Simple linear regression equation based on sample data
$$y = a + bx$$
3. Regression coefficient in sample regression equation
$$b = \hat{y}$$
$$a = \bar{y} - b\bar{x}$$
4. Residual representing the difference between an observed value of dependent variable y and its fitted value
$$e = y - \hat{y}$$
5. Standard error of estimate based on sample data
   - Deviations formula
$$S_{y.x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$$
   - Computational formula
$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$
6. Coefficient of determination based on sample data
   - Sums of squares formula
$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$
   - Computational formula
$$r^2 = \frac{a\Sigma y + b\Sigma xy - n(\bar{y})^2}{\Sigma y^2 - n(\bar{y})^2}$$
7. Regression sum of squares
$$S_{y.x} = S_y\sqrt{1 - r^2}$$
8. Interval estimate based on sample data: $\hat{y} \pm t_{df} S_{yx}$

Chapter Concepts Quiz

True or False

1. A statistical relationship between two variables does not indicate a perfect relationship. (T/F)
2. A dependent variable in a regression equation is a continuous random variable. (T/F)
3. The residual value is required to estimate the amount of variation in the dependent variable with respect to the fitted regression line. (T/F)
4. Standard error of estimate is the conditional standard deviation of the dependent variable. (T/F)
5. Standard error of estimate is a measure of scatter of the observations about the regression line. (T/F)
6. If one of the regression coefficients is greater than one the other must also be greater than one. (T/F)
7. The signs of the regression coefficients are always same. (T/F)
8. Correlation coefficient is the geometric mean of regression coefficients. (T/F)
9. If the sign of two regression coefficients is negative, then sign of the correlation coefficient is positive. (T/F)
10. Correlation coefficient and regression coefficient are independent. (T/F)
11. The point of intersection of two regression lines represents average value of two variables. (T/F)
12. The two regression lines are at right angle when the correlation coefficient is zero. (T/F)
13. When value of correlation coefficient is one, the two regression lines coincide. (T/F)
14. The product of regression coefficients is always more than one. (T/F)
15. The regression coefficients are independent of the change of origin but not of scale. (T/F)

Multiple Choice

16. The line of 'best fit' to measure the variation of observed values of dependent variable in the sample data is
(a) regression line (b) correlation coefficient
(c) standard error (d) none of these

17. Two regression lines are perpendicular to each other when
(a) $r = 0$ (b) $r = 1/3$
(c) $r = -1/2$ (d) $r = \pm 1$

18. The change in the dependent variable y corresponding to a unit change in the independent variable x is measured by
(a) b_{xy} (b) b_{yx}
(c) r (d) none of these

19. The regression lines are coincident provided
(a) $r = 0$ (b) $r = 1/3$
(c) $r = -1/2$ (d) $r = \pm 1$

20. If b_{yx} is greater than one, then b_{xy} is
(a) less than one (b) more than one
(c) equal to one (d) none of these

21. If b_{xy} is negative, then b_{yx} is
(a) negative (b) positive
(c) zero (d) none of these

22. If two regression lines are: $y = a + bx$ and $x = c + dy$, then the correlation coefficient between x and y is
(a) $\sqrt{bc}$ (b) $\sqrt{ac}$ (c) $\sqrt{ad}$ (d) $\sqrt{bd}$

23. If two regression lines are: $y = a + bx$ and $x = c + dy$, then the ratio of standard deviations of x and y are
(a) $\sqrt{c/b}$ (b) $\sqrt{c/a}$ (c) $\sqrt{d/a}$ (d) $\sqrt{d/b}$

24. If two regression lines are: $y = a + bx$ and $x = c + dy$, then the ratio of a/c is equal to
(a) b/d (b) $\frac{1-b}{1-d}$ (c) $\frac{1+b}{1+d}$ (d) $\frac{b-1}{d-1}$

25. If two coefficients of regression are 0.8 and 0.2, then the value of coefficient of correlation is
(a) 0.16 (b) −0.16 (c) 0.40 (d) −0.40

26. If two regression lines are: $y = 4 + kx$ and $x = 5 + 4y$, then the range of k is
(a) $k \le 0$ (b) $k \ge 0$
(c) $0 \le k \le 1$ (d) $0 \le 4k \le 1$

27. If two regression lines are: $x + 3y + 7 = 0$ and $2x + 5y = 12$, then $\bar{x}$ and $\bar{y}$ are respectively
(a) 2, 1 (b) 1, 2
(c) 2, 3 (d) 2, 4

28. The residual sum of square is
(a) minimized (b) increased
(c) maximized (d) decreased

29. The standard error of estimate $S_{y \cdot x}$ is the measure of
(a) closeness (b) variability
(c) linearity (d) none of these

30. The standard error of estimate is equal to
(a) $\sigma_y\sqrt{1-r^2}$ (b) $\sigma_y\sqrt{1+r^2}$
(c) $\sigma_x\sqrt{1-r^2}$ (d) $\sigma_x\sqrt{1+r^2}$

Concepts Quiz Answers

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

Review Self-Practice Problems

14.15 Given the following bivariate data:

x:	−1	5	3	2	1	1	7	3
y:	−6	1	0	0	1	2	1	5

(a) Fit a regression line of y on x and predict y if $x = 10$.
(b) Fit a regression line of x on y and predict x if $y = 2.5$.
[*Osmania Univ., MBA, 1996*]

14.16 Find the most likely production corresponding to a rainfall of 40 inches from the following data:

	Rainfall (in inches)	*Production (in quintals)*
Average	30	50
Standard deviation	5	10

Coefficient of correlation $r = 0.8$.
[*Bharthidarsan Univ., MCom, 1996*]

14.17 The coefficient of correlation between the ages of husbands and wives in a community was found to be + 0.8, the average of husbands age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations:
(a) the expected age of husband when wife's age is 16 years, and
(b) the expected age of wife when husband's age is 33 years. [*Osmania Univ., MBA, 2000*]

14.18 You are given below the following information about advertisement expenditure and sales:

	Adv. Exp. (*x*) (*Rs in crore*)	*Sales* (*y*) (*Rs in crore*)
Mean	20	120
Standard deviation	5	25

Correlation coefficient 0.8

(a) Calculate the two regression equations.

(b) Find the likely sales when advertisement expenditure is Rs 25 crore.

(c) What should be the advertisement budget if the company wants to attain sales target of Rs 150 crore?

[*Jammu Univ., MCom, 1997; Delhi Univ., MBA, 1999*]

14.19 For 50 students of a class the regression equation of marks in Statistics (x) on the marks in Accountancy (y) is $3y - 5x + 180 = 0$. The mean marks in Accountancy is 44 and the variance of marks in Statistics is 9/16th of the variance of marks in Accountancy. Find the mean marks in Statistics and the coefficient of correlation between marks in the two subjects.

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

Years of service :	11	7	9	5	8	6	10
Income (Rs in 1000's) :	10	8	6	5	9	7	11

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

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

(c) Do you think other factors are to be considered (in addition to the years of service) in fixing the income with reference to the above problems? Explain.

14.21 The following table gives the age of cars of a certain make and their annual maintenance costs. Obtain the regression equation for costs related to age.

Age of cars (in years) :	2	4	6	8
Maintenance costs (Rs in 100's) :	10	20	25	30

[*HP Univ., MBA, 1994*]

14.22 An analyst in a certain company was studying the relationship between travel expenses in rupees (y) for 102 sales trips and the duration in days (x) of these trips. He has found that the relationship between y and x is linear. A summary of the data is given below:

$\Sigma x = 510$; $\Sigma y = 7140$; $\Sigma x^2 = 4150$; $\Sigma xy = 54{,}900$, and $\Sigma y^2 = 7{,}40{,}200$

(a) Estimate the two regression equations from the above data.

(b) A given trip takes seven days. How much money should a salesman be allowed so that he will not run short of money?

14.23 The quantity of a raw material purchased by ABC Ltd. at specified prices during the post 12 months is given below.

Month	*Price per kg (in Rs)*	*Quantity (in kg)*	*Month*	*Price per kg (in Rs)*	*Quantity (in kg)*
Jan	96	250	July	112	220
Feb	110	200	Aug	112	220
March	100	250	Sept	108	200
April	90	280	Oct	116	210
May	86	300	Nov	86	300
June	92	300	Dec	92	250

(a) Find the regression equations based on the above data.

(b) Can you estimate the approximate quantity likely to be purchased if the price shoots up to Rs 124 per kg?

(c) Hence or otherwise obtain the coefficient of correlation between the price prevailing and the quantity demanded.

14.24 With ten observations on price (x) and supply (y), the following data were obtained (in appropriate units): $\Sigma x = 130$, $\Sigma y = 220$, $\Sigma x^2 = 2288$, $\Sigma y^2 = 5506$, $\Sigma xy = 3467$. Obtain the line of regression of y on x and estimate the supply when the price is 16 units. Also find out the standard error of the estimate.

14.25 Data on the annual sales of a company in lakhs of rupees over the past 11 years is shown below. Determine a suitable straight line regression model $y = \beta_0 + \beta_1 x + \epsilon$ for the data. Also calculate the standard error of regression of y for values of x.

Year :	1978	79	80	81	82	83	84	85	86	87	88
sales :	1	5	4	7	10	8	9	13	14	13	18

From the regression line of y on x, predict the values of annual sales for the year 1989.

14.26 Find the equation of the least squares line fitting the following data:

x:	1	2	3	4	5
y:	2	6	5	3	4

Calculate the standard error of estimate of y on x.

14.27 The following data relating to the number of weeks of experience in a job involving the wiring of an electric motor and the number of motors rejected during the past week for 12 randomly selected workers.

Workers	*Experience (weeks)*	*No. of Rejects*
1	2	26
2	9	20
3	6	28
4	14	16
5	8	23
6	12	18
7	10	24
8	4	26
9	2	38
10	11	22
11	1	32
12	8	25

(a) Determine the linear regression equation for estimating the number of components rejected given the number of weeks of experience. Comment on the relationship between the two variables as indicated by the regression equation.

(b) Use the regression equation to estimate the number of motors rejected for an employee with 3 weeks of experience in the job.

(c) Determine the 95 per cent approximate prediction interval for estimating the number of motors rejected for an employee with 3 weeks of experience in the job, using only the standard error of estimate.

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

Income (Rs in 1000's)	:	8	12	9	24	43	37	19	16
Per cent invested in securities	:	36	25	33	15	28	19	20	22

(a) Develop an estimating equation that best describes these data.

(b) Find the coefficient of determination and interpret it.

(c) Calculate the standard error of estimate for this relationship.

(d) Find an approximate 90 per cent confidence interval for the percentage of income invested in securities by a family earning Rs 25,000 annually.

[Delhi Univ., MFC, 1997]

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

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

(a) Develop an estimating equation that best describes these data.

(b) Find the coefficient of determination and interpret it.

(c) Determine the percentage of total variation in security return being explained by the return on the market portfolio.

14.30 The equation of a regression line is

$$\hat{y} = 50.506 - 1.646x$$

and the data are as follows:

x:	5	7	11	12	19	25
y:	47	38	32	24	22	10

Solve for residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?

14.31 Graph the followign residuals and indicate which of the assumptions underlying regresion appear to be in jeopardy on the basis of the graph:

x	:	13	16	27	29	37	47	63
$y - \hat{y}$	:	−11	−5	−2	−1	6	10	12

Hints and Answers

14.15 $\bar{x} = \Sigma x = 21/8 = 2.625$; $\bar{y} = \Sigma y/n = 4/8 = 0.50$

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

$$= 0.568;$$

$d_x = x - 3$; $d_y = y - 3$.

Regression equation:

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

or $y - 0.5 = 0.568\,(x - 2.625)$

$$y = -0.991 + 0.568x$$

(b) $$b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2} = \frac{8\times 30 - (-3)(-12)}{8\times 84 - (-12)^2}$$

$$= 0.386$$

Regression equation:

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

or $x - 2.625 = 0.386\,(y - 5)$

$$x = 0.695 + 0.386y$$

14.16 Let x = rainfall y = production by y. The expected yield corresponding to a rainfall of 40 inches is given by regression equation of y on x.

Given $\bar{y} = 50, \sigma_y = 10, \bar{x} = 30, \sigma_x = 5, r = 0.8$

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

$$y - 50 = 0.8\,\frac{10}{5}\,(x - 30)$$

$$y = 2 + 1.6x$$

For $x = 40, y = 2 + 1.6\,(40) = 66$ quintals.

14.17 Let x = age of wife y = age of husband.

Given $\bar{x} = 25$, $\bar{y} = 22, \sigma_x = 4, \sigma_y = 5, r = 0.8$

(a) Regression equation of x on y

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

$$x - 25 = 0.8\,\frac{4}{5}\,(y - 22)$$

$$x = 10.92 + 0.64y$$

When age of wife is $y = 16$; $x = 10.92 + 0.64\,(16) = 22$ approx.(husband's age)

(b) Left as an exercise

14.18 (a) Regression equation of x on y

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

$$x - 20 = 0.8\,\frac{5}{25}\,(y - 120)$$

$$x = 0.8 + 0.16y$$

Regression equation of y on x

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

$$y - 120 = 0.8\frac{25}{5}\,(x - 20)$$

$$y = 40 + 4x$$

(b) When advertisement expenditure is of Rs 25 crore, likely sales is

$$y = 40 + 4x = 40 + 4\,(25) = 140 \text{ crore.}$$

(c) For $y = 150$, $x = 0.8 + 0.16y = 0.8 + 0.16(150)$
$= 24.8$

14.19 Let x = marks in Statistics and y = marks in Accountancy,

Given: $3y - 5x + 180 = 0$

or $x = (3/5)y + (180/5)$

For $y = 44$, $x = (3/5) \times 44 + (180/5) = 62.4$

Regression coefficient of x on y, $b_{xy} = 3/5$

Coefficient of regression

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sqrt{9}}{\sqrt{16}} \text{ (given)}$$

or $$\frac{3}{5} = r\frac{\sqrt{9}}{\sqrt{16}} \quad \text{or} \quad \frac{3}{5} = \frac{3r}{4}$$

Hence $3r = 2.4$ or $r = 0.8$

14.20 Let x = years of service and y = income.

(a) Regression equation of y on x

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7\times469 - 56\times56}{7\times476 - (56)^2} = 0.75$$

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

$$y - 8 = 0.75\,(x - 8)$$

$$y = 2 + 0.75x$$

(b) When $x = 13$ years, the average income would be

$$y = 2 + 0.75x = 2 + 0.75(13) = \text{Rs } 11{,}750$$

14.21 Let x = age of cars and y = maintenance costs.

The regression equation of y on x

$\bar{x} = \Sigma x/n = 20/4 = 5$; $\bar{y} = \Sigma y/n = 85/4 = 21.25$

and $$b_{yx} = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7\times490 - 20\times85}{7\times120 - (20)^2} = 3.25$$

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

$$y - 21.25 = 3.25\,(x - 5)$$

$$y = 5 + 3.25x$$

14.22 $\bar{x} = \Sigma x/n = 510/102 = 5$; $\bar{y} = \Sigma y/n = 7140/102 = 70$

Regression coefficients:

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

$$= \frac{102\times54900 - 510\times7140}{102\times740200 - (7140)^2} = 0.08$$

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

$$= \frac{102\times54900 - 510\times7140}{102\times4150 - (510)^2} = 12$$

Regression lines:

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

$$x - 5 = 0.08\,(y - 70) \quad \text{or} \quad x = 0.08y - 0.6$$

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

$$y - 70 = 12\,(x - 5) \quad \text{or} \quad y = 12x + 10$$

When $x = 7$, $\bar{y} = 12\times7 + 10 = 94$

14.23 Let price be denoted by x and quantity by y

$\bar{x} = \Sigma x/n = 1200/12 = 100$;

$\bar{y} = \Sigma y/n = 2980/12 = 248.33$

(a) Regression coefficients:

$$b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma dx)(\Sigma dy)}{n\Sigma d_y^2 - (\Sigma d_y)^2} = -0.26$$

$$b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = -3.244$$

Regression lines:

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

$$x - 100 = -0.26\,(y - 248.33)$$

or $$x = -0.26y + 164.56$$

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

$$y - 248.33 = -3.244\,(x - 100)$$

$$y = -3.244x + 572.73$$

(b) For $x = 124$,

$$y = -3.244 \times 124 + 572.73 = 170.474$$

14.24 (a) Regression line of y on x is given by

Given $\bar{y} = \Sigma y/n = 220/10 = 22$;

$\bar{x} = \Sigma x/n = 130/10 = 13$

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

$$y - 22 = 1.015\,(x - 13)$$

$$y = 8.805 + 1.015x$$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10\times3467 - 130\times220}{10\times2288 - (130)^2}$$

$$= \frac{34670 - 28600}{22880 - 16900} = \frac{6070}{5980} = 1.015$$

(b) When $x = 16$,

$$y = 8.805 + 1.015\,(16) = 25.045$$

(c) $S_{yx} = S_y\sqrt{1 - r^2} = 8.16\sqrt{1 - (0.9618)^2} = 2.004$

14.25 Take years as $x = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$, where 1983 = 0. The regression equation is

$$\hat{y} = 9.27 + 1.44x$$

For $x = 1989$, $\hat{y} = 9.27 + 1.44\,(6) = 17.91$

$$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{21.2379}{10}} = 1.4573$$

14.26 $\bar{x} = \Sigma x/n = 15/5 = 3$, $\bar{y} = \Sigma x/n = 20/5 = 4$

The regression equation is:

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

$$y - 4 = 0.7\,(x - 3) \text{ or } \hat{y} = 1.9 + 0.7x$$

Standard error of estimate,

$$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})}{n - 2}} = \sqrt{\frac{5.1}{3}} = 1.303$$

14.27 (a) $b = \dfrac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \dfrac{2048 - 12(7.67)\,(24.83)}{876 - 12(7.67)^2} = -1.40$

$$a = \bar{y} - b\bar{x} = 24.83 - (-1.40)(7.67) = 35.57$$

Thus $\hat{y} = a + bx = 35.57 - 1.40x$

Since $b = -1.40$, it indicate an inverse (negative) relationship between weeks of experience (x) and the number of rejects (y) in the sample week

(b) For $x = 3$, we have $\hat{y} = 35.57 - 1.40(3) \cong 31$

(c) $$S_{yx} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\,\Sigma xy}{n - 2}}$$

$$= \sqrt{\frac{7{,}798 - (35.57)\,(298) - 1.40\,(2048)}{12 - 2}}$$

$$= 2.56$$

95 per cent approximate prediction interval

$$\hat{y} \pm t_{df} S_{y \cdot x} = 31.37 \pm 2.228\,(2.56)$$

$$= 25.67 \text{ to } 37.07 \text{ or } 26 \text{ to } 37 \text{ rejects.}$$

14.28 4.724; – 0.983; – 0.399, – 6.753, 2.768, 0.644

14.29 Error term non-independent.

Case Studies

Case 14.1: Made in China

The phrase 'made in China' has become an issue of concern in the last few years, as Indian Companies try to protect their products from overseas competition. In these years a major trade imbalance in India has been caused by a flood of imported goods that enter the country and are sold at lower price than comparable Indian made goods. One prime concern is the electronic goods in which total imported items have steadily increased during the year 1990s to 2004s. The Indian companies have been worried on complaints about product quality, worker layoffs, and high prices and has spent millions in advertising to produce electronic goods that will satisfy consumer demands. Have these companies been successful in stopping the flood these imported goods purchased by Indian consumers? The given data represent the volume of imported goods sold in India for the years 1999-2004. To simplify the analysis, we have coded the year using the coded variable x = Year 1989.

Year	*x = Year 1989*	*Volume of Import (in Rs billion)*
1989	0	1.1
1990	1	1.3
1991	2	1.6
1992	3	1.6
1993	4	1.8
1994	5	1.4
1995	6	1.6
1996	7	1.5
1997	8	2.1
1998	9	2.0
1999	10	2.3
2000	11	2.4
2001	12	2.3
2002	13	2.2
2003	14	2.4
2004	15	2.4

Questions for Discussion

1. Find the least-squares line for predicting the volume of import as a function of year for the years 1989-2000.
2. Is there a significant linear relationship between the volume of import and the year?
3. Predict the volume of import of goods using 95% prediction intervals for each of the years 2002, 2003 and 2004.
4. Do the predictions obtained in Step 4 provide accurate estimates of the actual values observed in these years? Explain.
5. Add the data for 1989-2004 to your database, and recalculate the regression line. What effect have the new data points had on the slope> What is the effect of SSE?
6. Given the form of the scattered diagram for the years 1989-2004, does it appear that a straight line provides an accurate model for the data? What other type of model might be more appropriate?

Chapter 15

Life can only be understood backward, but it must be lived forward.

—Niels Bohr

Other than food and sex, nothing is quite as universally interesting as the size of our pay cheques.

—N. L. Preston and E. R. Fiedler

Partial and Multiple Correlation and Regression Analysis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- describe the relationship between two variables when influence of one or more other variables is held constant or involved.
- establish a regression equation for estimating value of a dependent variable given the values of two or more independent variables.
- determine the extent of random error involved in the estimation of dependent variable value.
- measure the coefficient of determination to understand the proportion of variation in the dependent variable which is explained by the independent variables.

15.1 INTRODUCTION

The word 'multiple' describes the number of independent variables involved in estimating the value of a dependent variable. Thus the concepts of simple linear correlation and regression presented in Chapters 13 and 14 can be extended by adding more than one independent variables (predictors) in a regression equation. Such an extension may serve as a basis to explain with more accuracy the variability in the dependent variable and hence estimating its value. This equation can also be used to obtain better explanation about the influence of independent variables on the predicted value of the dependent variable.

A linear regression equation with more than one independent variables is called a *multiple linear regression model.* Although multiple linear regression analysis in many ways is an extension of simple linear regression analysis, but the calculations become quite complicated when estimating value of the dependent variable. Consequently, problems involving only 2 or 3 independent variables are discussed in this Chapter in order to provide some awareness of multiple regression analysis.

Through *multiple correlation analysis* we can attempt to measure the degree of association between a dependent (response) variable y and two or more independent variables

(predictors) x_1, x_2, ... taken together as a group. In multiple correlation analysis, it is also possible to measure the degree of association between a dependent variable and any one of the independent variables included in the analysis, while the effect of the other independent variables included in the analysis is held constant. This measure of association is called *partial correlation coefficient.* It differs from a simple correlation coefficient in the manner that in simple correlation analysis the effect of all other variables in ignored rather than being statistically controlled as in partial correlation analysis.

The main advantage of multiple regression analysis is that it allows us to include the values of more independent variables known to us to estimate the value of the dependent variable. The data on the values of independent variables enable us to determine the statistical error associated with this estimated value of the dependent variable, and hence the relationship between a dependent variable and two or more independent variables can be described with more accuracy.

Illustrations: The following examples illustrate multiple regression model (or equation)

1. Suppose a farmer who wishes to relate the yield (y) of wheat crop, a dependent variable, to the amount of fertilizer used, an independent variable, he can certainly find a simple regression equation that relates these two variables. However, true prediction of yield of crop is possible by including in the regression equation a few more variables such as: quality of seed, amount of water given to the crop, and so on.
2. Suppose, if the management of an organization is wishes to understand the expected variance in the performance (y), a dependent variable, to be explained by four independent variables, say, pay, task difficulty, supervisory support, and organizational culture. These four independent variables are of course correlated to the dependent variable in varying degrees, but they might also be correlated among themselves, e.g. task difficulty is likely to be related to supervisory support, pay is likely to be related to task difficulty. These three variables together are likely to influence the organizational culture.

15.2 ASSUMPTIONS IN MULTIPLE LINEAR REGRESSION

A multiple regression equation or model is used in the same way as that a simple linear regression equation is used for understanding the relationship among variables of interest and for estimating the average value of the dependent variable y, given a set of values of independent variables in the data set. Assumptions for the multiple linear regression model are same as for the simple linear regression model mentioned in Chapter 14. However, the summary of assumptions for multiple linear regression is as follows:

1. The population linear regression equation involving a dependent variable y and a set of k independent variables x_1, x_2,..., x_k is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e \qquad (15\text{-}1)$$

where
y = value of dependent variable to be estimated
β_0 = a constant which represents the value of y when value of all independent variables is zero.
$\beta_1, \ldots, \beta_k$ = parameters or *regression coefficients* associated with each of the x_k independent variables.
x_k = value of kth independent variable.
e = random error associated with the sampling process. It represents the unpredictable variation in y values from the population regression model

The term *linear* is used because equation (15-1) is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$,

2. The dependent variable is a continuous random variable, whereas independent variables are not random because their values are controlled by assigning different values.
3. The variance and standard deviation of the dependent variable are equal for each combination of values of independent variables.

4. The random error e associated with the dependent variable y for various combinations of values of independent variables are statistically independent of each other and normally distributed.
5. The error of variance, σ_e^2 is same for all values of independent variables. Thus, the range of deviations of the y-values from the regression line is same regardless of the values of the independent variables.

   The magnitude of error of variance, σ_e^2 measures the closeness of observed values of the dependent variable to the regression line (line of 'best fit'). The smaller the value of σ_e^2 (also called residual variance), the better the predicted value of the dependent variable.
6. The random error e is a random variable with zero expected (or mean) value. Consequently the expected or mean value of the dependent variable is denoted by

$$\hat{y} \text{ or } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

The parameter β_j ($j = 0, 1, 2, \ldots, k$) is also called *partial regression coefficient* because it measures the expected change in response variable y per unit change in x_j when all remaining independent variables x_i ($i \neq j$) are held constant.

15.3 ESTIMATING PARAMETERS OF MULTIPLE REGRESSION MODEL

Multiple regression analysis requires that we obtain sample data and calculate values of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ by fitting the model to the data. To fit the general linear multiple regression model using method of least-squares, we choose the estimated regression model

$$E(y) \text{ or } \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad (15\text{-}2)$$

that minimize the sum of squares errors (SSE) = $\Sigma(y_i - \hat{y}_i)^2$, where y_i and $\hat{y}_i$ ($i = 1, 2\ldots, k$) represent the observed and estimated (or predicted) value of the dependent variable for the ith observation. The terms b_j ($j = 0, 1, 2, \ldots, k$) are the least-squares estimates of population regression parameter β_j.

15.3.1 Estimation : The Method of Least Squares

The method of least squares discussed in Chapter 14 to compute the values of regression coefficients (or parameters) a and b and estimating the value of the dependent variable y in a linear regression model can also be extended for estimating the unknown parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ based on sample data. The least squares estimators of these parameters are denoted by $b_0, b_1, b_2, \ldots, b_k$ respectively. Given these values, the least squares *multiple regression equation* can be written as:

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad (15\text{-}3)$$

where $\hat{y}$ = estimated value of dependent variable y

a = y-intercept

x_1, x_2 = independent variables

b_j = slope associated with variable x_j ($j = 0, 1, 2, \ldots, k$)

To visualize a multiple regression model, consider the following regression equation involving two independent variables x_1 and x_2 and a dependent variable y:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

This equation describes a plane in a 3-dimensional space of y, x_1 and x_2 as shown in Fig. 15.1.

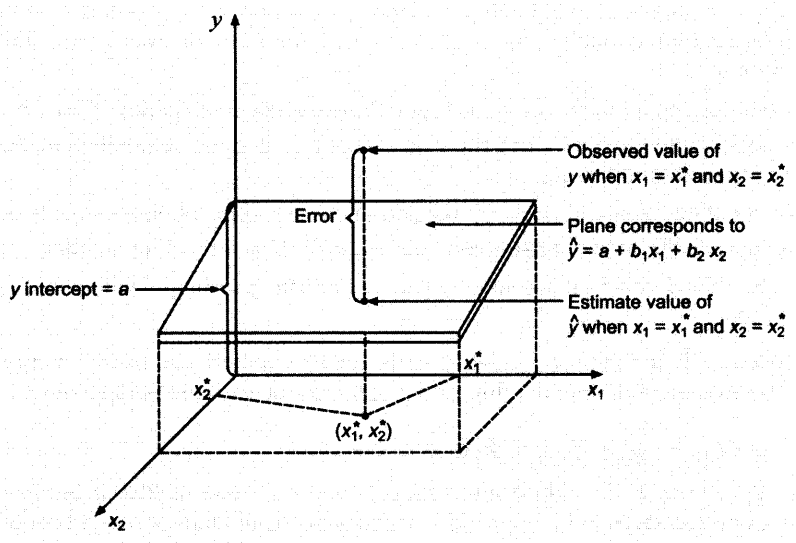


Figure 15.1
Graph of Multiple Regression Equation with Two Independent Variables

15.3.2 Partial Regression Coefficients

The partial regression coefficient β_j ($j = 1, 2, ..., k$) represents the expected change in the dependent (response) variable y with respect to per unit change in independent variable x_j when all other independent variables x_i ($i \neq j$) are held constant. The partial regression coefficients occur because more than one independent variable is present in the regression model. The partial regression coefficients are analogous to β_1, the slope of the simple regression model.

The partial regression coefficients and regression constant of a multiple regression model are population values and are unknown. In practice, these values are estimated by using sample information. The least-squares multiple regression equation is same as equation (15-3).

Estimated multiple regression equation: The estimate of the multiple regression equation based on sample data and the least-squares method.

Least-Squares Normal Equations To demonstrate the calculation of partial regression coefficients we consider the particular form of regression equation (15-3) involving two independent variables x_2 and x_3 and a dependent variable x_1:

$$\hat{x}_1 = a + b_1x_2 + b_2x_3 = a + b_{12.3}x_2 + b_{13.2}x_3 \qquad (15\text{-}4)$$

where $\hat{x}_1$ = estimated value of dependent variable

a = a regression constant representing intercept on y-axis; its value is zero when the regression equation passes through the origin.

$b_{12.3}, b_{13.2}$ = partial regression coefficients; $b_{12.3}$ corresponds to change in x_1 for each unit change in x_2 while x_3 is held constant; $b_{13.2}$ represents the change in x_1 for each unit change in x_3 while x_2 is held constant.

To obtain the value of $\hat{x}_1$ it is necessary to determine the values of the constants a (or $b_{1.23}$), $b_{12.3}$, and $b_{13.2}$ in accordance with the least-squares criterion, i.e. minimizing the sum of the squares of the residuals. Thus

$$\text{Min } z = \Sigma(x_1 - \hat{x}_1)^2 = \Sigma(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3)^2$$

To minimize this sum, we use the concept of maxima and minima, where derivatives of z with respect to these constants are equated to zero. Consequently we get the following three normal equations:

$$\begin{aligned} \Sigma x_1 &= na + b_{12.3}\Sigma x_1 + b_{13.2}\Sigma x_3 \\ \Sigma x_1x_2 &= a\Sigma x_2 + b_{12.3}\Sigma x_2^2 + b_{13.2}\Sigma x_3x_2 \\ \Sigma x_1x_3 &= a\Sigma x_3 + b_{12.3}\Sigma x_2x_3 + b_{13.2}\Sigma x_3^2 \end{aligned} \qquad (15\text{-}5)$$

Regression equation of x_2 on x_1 and x_3: $x_2 = b_{2.13} + b_{21.3}x_1 + b_{23.1}x_3$

The normal equations for fitting this regression equation will be

$$\Sigma x_2 = n\, b_{2.13} + b_{21.3}\Sigma x_1 + b_{23.1}\Sigma x_3$$
$$\Sigma x_2 x_1 = b_{2.13}\Sigma x_1 + b_{21.3}\Sigma x_1^2 + b_{23.1}\Sigma x_1 x_3 \quad (15\text{-}6)$$
$$\Sigma x_2 x_3 = b_{2.13}\Sigma x_3 + b_{21.3}\Sigma x_1 x_3 + b_{23.1}\Sigma x_3^2$$

Regression equation of x_3 on x_1 and x_2: $x_3 = b_{3.12} + b_{31.2}\, x_1 + b_{32.1} x_2$

The normal equations for fitting this regression equation will be

$$\Sigma x_3 = n\, b_{3.12} + b_{31.2}\Sigma x_1 + b_{32.1}\Sigma x_2$$
$$\Sigma x_3 x_1 = b_{3.12}\Sigma x_1 + b_{31.2}\Sigma x_1^2 + b_{32.1}\Sigma x_2 x_1 \quad (15\text{-}7)$$
$$\Sigma x_3 x_2 = b_{3.12}\Sigma x_2 + b_{31.2}\Sigma x_1 x_2 + b_{32.1}\Sigma x_2^2$$

The values of constants can be calculated by solving the system of simultaneous equations (15-5), (15-6) or (15-7) as per the nature of the regression equation.

Short-cut Method : For solving above stated normal equations for constants, e.g. $a = b_{1.23}$, $b_{12.3}$ and $b_{13.2}$, take deviations of the values of the variables from their actual mean values. Let $X_1 = x_1 - \bar{x}_1$, $X_2 = x_2 - \bar{x}_2$ and $X_3 = x_3 - \bar{x}_3$ be the deviations from the actual mean values of variables x_1, x_2, and x_3, respectively. Since the sum of deviations of the values of a variable from its actual mean is zero, therefore

$$\Sigma X_1 = \Sigma(x_1 - \bar{x}_1) = 0,\ \Sigma X_2 = (x_2 - \bar{x}_2) = 0, \text{ and } \Sigma X_3 = (x_3 - \bar{x}_3) = 0$$

Summing the variables and dividing by n in the regression equation of x_1 on x_2 and x_3, we have

$$\bar{x}_1 = b_{12.3} + b_{12.3}\,\bar{x}_2 + b_{13.2}\,\bar{x}_3 \quad (15\text{-}8)$$

Subtracting equation (15-8) from equation (15-4), we have

$$x_1 - \bar{x}_1 = b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3)$$
$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

The second and third equations of (15-5) can be rewritten as:

$$\Sigma X_1 X_2 = b_{12.3}\Sigma X_2^2 + b_{13.2}\Sigma X_2 X_3 \quad (15\text{-}9)$$
$$\Sigma X_1 X_3 = b_{12.3}\Sigma X_2 X_3 + b_{13.2}\Sigma X_3^2$$

Regression equation of x_2 on x_1 and x_3: $X_2 = b_{21.3} X_1 + b_{23.1} X_3$

The normal equations for getting the values of $b_{21.3}$ and $b_{23.1}$ are written as:

$$\Sigma X_2 X_1 = b_{21.3}\Sigma X_1^2 + b_{23.1}\Sigma X_1 X_3 \quad (15\text{-}10)$$
$$\Sigma X_2 X_3 = b_{21.3}\Sigma X_1 X_3 + b_{23.1}\Sigma X_3^2$$

Regression equation of x_3 on x_1 and x_2: $X_3 = b_{31.2} X_1 + b_{32.1} X_2$

The normal equations for getting the values of $b_{31.2}$ and $b_{32.1}$ are written as:

$$\Sigma X_3 X_1 = b_{31.2}\Sigma X_1^2 + b_{32.1}\Sigma X_1 X_2 \quad (15\text{-}11)$$
$$\Sigma X_3 X_2 = b_{31.2}\Sigma X_1 X_2 + b_{32.1}\Sigma X_2^2$$

When simultaneous equations (15-9) are solved for $b_{12.3}$ and $b_{13.2}$, we have

$$b_{12.3} = \frac{(\Sigma X_1 X_2)(\Sigma X_3^2) - (\Sigma X_1 X_3)(\Sigma X_2 X_3)}{\Sigma X_2^2 \times \Sigma X_3^2 - (\Sigma X_2 X_3)^2} \quad (15\text{-}12)$$

and

$$b_{13.2} = \frac{(\Sigma X_1 X_3)(\Sigma X_2^2) - (\Sigma X_1 X_2)(\Sigma X_2 X_3)}{(\Sigma X_2^2)(\Sigma X_3^2) - (\Sigma X_1 X_3)^2}$$

15.3.3 Relationship Between Partial Regression Coefficients and Correlation Coefficients

Let r_{12} = correlation coefficient between variable x_1 and x_2

x_{13} = correlation coefficient between variable x_1 and x_3

These correlation coefficients are known as *zero order correlation coefficients*. The variance of sample data on variables x_1, x_2, and x_3 is given by

$$s_1^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2}{n} = \frac{\Sigma X_1^2}{n}$$

$$s_2^2 = \frac{\Sigma (x_2 - \bar{x}_2)^2}{n} = \frac{\Sigma X_2^2}{n}$$

$$s_3^2 = \frac{\Sigma (x_3 - \bar{x}_3)^2}{n} = \frac{\Sigma X_3^2}{n}$$

We know that

$$r_{12} = \frac{\text{Cov.}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}} = \frac{\Sigma (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\sqrt{\Sigma (x_1 - \bar{x}_1)^2} \sqrt{\Sigma (x_2 - \bar{x}_2)^2}} = \frac{\Sigma X_1 X_2}{\sqrt{\Sigma X_1^2 \, \Sigma X_2^2}}$$

$$= \frac{\Sigma X_1 X_2}{\sqrt{n s_1^2} \sqrt{n s_2^2}} = \frac{\Sigma X_1 X_2}{n s_1 s_2}$$

or $\Sigma X_1 X_2 = n r_{12} s_1 s_2$

Similarly $\Sigma X_1 X_3 = n r_{13} s_1 s_3$ and $\Sigma X_2 X_3 = n r_{23} s_2 s_3$

Substituting these values in equation (15-9) and on further simplification, we get

$$\begin{aligned} b_{12.3} s_2 + b_{13.2} s_3 r_{23} &= s_1 r_{12} \\ b_{13.2} s_3 + b_{12.3} s_2 r_{23} &= s_1 r_{13} \end{aligned} \qquad (15\text{-}13)$$

Solving equations (15-13) for $b_{12.3}$ and $b_{13.2}$, we have

$$b_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \left(\frac{s_1}{s_2} \right)$$

and

$$b_{13.2} = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \left(\frac{s_1}{s_3} \right) \qquad (15\text{-}14)$$

Thus the regression equation of x_1 on x_2 and x_3 can be written as:

$$X_1 = b_{12.3} X_2 + b_{13.2} X_2$$

$$x_1 - \bar{x}_1 = \left(\frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_2} \right) (x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \right) \left(\frac{s_1}{s_3} \right) (x_3 - \bar{x}_3)$$

Regression equation of x_2 on x_1 and x_3:

$$x_2 - \bar{x}_2 = \left(\frac{r_{12} - r_{23} r_{31}}{1 - r_{31}^2} \right) \left(\frac{s_2}{s_1} \right) (x_1 - \bar{x}_1) + \left(\frac{r_{23} - r_{12} r_{13}}{1 - r_{13}^2} \right) \left(\frac{s_2}{s_3} \right) (x_3 - \bar{x}_3)$$

Regression equation of x_3 on x_1 and x_2:

$$x_3 - \bar{x}_3 = \left(\frac{r_{13} - r_{23} r_{12}}{1 - r_{12}^2} \right) \left(\frac{s_3}{s_1} \right) (x_1 - \bar{x}_1) + \left(\frac{r_{23} - r_{12} r_{13}}{1 - r_{12}^2} \right) \left(\frac{s_3}{s_2} \right) (x_2 - \bar{x}_2)$$

Example 15.1: A sample survey of 5 families was taken and figures were obtained with respect to their annual savings x_1 (Rs in 100's), annual income x_2 (Rs in 1000's), and family size x_3. The data is summarized in the table below:

Family	*Annual Savings* (x_1)	*Annual Income* (x_2)	*Family Size* (x_3)
1	10	16	3
2	5	13	6
3	10	21	4
4	4	10	5
5	8	13	3

(a) Find the least-squares regression equations of x_1 on x_2 and x_3.

(b) Estimate the annual savings of a family whose size is 4 and annual income is Rs 16,000.

Solution: The calculations needed in the three variables regression problem are shown in Table 15.1.

Table 15.1: Calculations for Regression Equation

Family	*Savings* x_1	*Income* x_2	*Size* x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
1	10	16	3	100	256	9	160	30	48
2	5	13	6	25	169	36	65	30	78
3	10	21	4	100	441	16	210	40	84
4	4	10	5	16	100	25	40	20	50
5	8	13	3	64	169	9	104	24	39
$n = 5$	37	73	21	305	1135	95	579	144	299

Substituting values from Table 15.1 in the normal equations for the regression equation of x_1 on x_2 and x_3, we get

(i) $\Sigma x_1 = na + b_1\Sigma x_2 + b_2\Sigma x_3$ or $37 = 5a + 73b_1 + 21b_2$

(ii) $\Sigma x_1x_2 = a\Sigma x_2 + b_1\Sigma x_2^2 + b_2\Sigma x_2x_3$ or $579 = 73a + 1135b_1 + 299b_2$

(iii) $\Sigma x_1x_3 = a\Sigma x_3 + b_1\Sigma x_2x_3 + b_2\Sigma x_3^2$ or $144 = 21a + 299b_1 + 95b_2$

Multiply Eqn. (i) by 73 and Eqn. (ii) by 5 to eliminate a, we get

$$\begin{array}{rrrrr} 2701 = & 365a & + 5329b_1 & + 1533b_2 \\ + 2895 = & + 365a & + 5675b_1 & + 1495b_2 \\ - & - & - & - \\ \hline \text{(iv)} \quad 194 = & & 346b_1 - & 38b_2 \end{array}$$

Multiply Eqn. (i) by 21 and Eqn. (iii) by 5 to eliminate a, we get

$$\begin{array}{rrrrr} 777 = & 105a & + 1533b_1 & + 441b_2 \\ + 720 = & + 105a & + 1495b_1 & + 475b_2 \\ - & - & - & - \\ \hline \text{(v)} \quad 57 = & & 38b_1 - & 34b_2 \end{array}$$

Multiplying Eqn. (iv) by 34 and Eqn. (v) by 38 to eliminate b_2, we get

$$\begin{array}{rrrr} 6596 = & 11764b_1 & - 1292b_2 \\ + 1406 = & + 1444b_1 & - 1292b_2 \\ - & - & + \\ \hline 5190 = & 10320b_1 & \text{or } b_1 = 0.502 \end{array}$$

Substituting the value of b_1 in Eqn. (i) and (ii), we get

(vi) $5a + 21b_2 = 0.354$

(vii) $21a + 95b_2 = -6.098$

Multiplying Eqn. (vi) by 21 and Eqn. (vii) by 5 to eliminate a, we get

$$\begin{array}{rrrr} 105a & + 441b_2 & = & 7.434 \\ + 105a & + 475b_2 & = & - 30.49 \\ - & - & + & \\ \hline - 34b_2 & & = 37.924 & \text{or } b_2 = -1.115 \end{array}$$

Substituting the value of b_1 and b_2 in Eqn. (i), we get $a = 4.753$.

(a) The least-squares regression equation of x_1 on x_2 and x_3 is given by

$$\begin{aligned} x_1 &= a + b_1x_2 + b_2x_3 \\ &= 4.753 + 0.502x_2 - 1.115x_3 \end{aligned}$$

(b) The estimated value of annual savings (x_1) is obtained by substituting annual income x_2 = Rs. 1600 and family size $x_3 = 4$, as:

$$\hat{x}_1 = 4.753 + 0.502(1600) - 1.115(4) = \text{Rs } 803.493$$

Example 15.2: An instructor of mathematics wishes to determine the relationship of grades on the final examination to grades on two quizes given during the semester. Let x_1, x_2, and x_3 be the grades of a student on the first quiz, second quiz, and final examination respectively. The instructor made the following computations for a total of 120 students:

$$\bar{x}_1 = 6.80 \qquad \bar{x}_2 = 0.70 \qquad \bar{x}_3 = 74.00$$
$$s_1 = 1.00 \qquad s_2 = 0.80 \qquad s_3 = 9.00$$
$$r_{12} = 0.60 \qquad r_{13} = 0.70 \qquad r_{23} = 0.65$$

(a) Find the least-squares regression equation of x_3 on x_1 and x_2.

(b) Estimate the final grades of two students who scored respectively 9 and 7 and 4 and 8 marks in the two quizes. *[HP Univ., MBA, 1998]*

Solution: (a) The regression equation of x_3 on x_2 and x_1 can be written as:

$$(x_3 - \bar{x}_3) = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(x_1 - \bar{x}_1)$$

Substituting the given values, we have

$$(x_3 - 74) = \left(\frac{0.65 - 0.7 \times 0.6}{1 - (0.6)2}\right)\left(\frac{9}{0.8}\right)(x_2 - 7) + \left(\frac{0.7 - 0.65 \times 0.6}{1 - (0.6)^2}\right)\left(\frac{9}{1}\right)$$
$$(x_1 - 6.8)$$

$$(x_3 - 74) = \left(\frac{0.65 - 0.42}{0.64}\right)\left(\frac{9}{0.8}\right)(x_2 - 7) + \left(\frac{0.7 - 0.39}{0.64}\right)(9)(x_1 - 6.8)$$

$$(x_3 - 74) = 4.04\,(x_2 - 7) + 4.36\,(x_1 - 6.8)$$

$$x_3 = 16.07 + 4.36\,x_1 + 4.04\,x_2$$

(b) The final grade of student who scored 9 and 7 marks is obtained by substituting $x_1 = 9$ and $x_2 = 7$ in the regression equation:

$$x_3 = 16.07 + 4.36\,(9) + 4.04\,(7)$$
$$= 16.07 + 39.24 + 28.28 = 83.59 \text{ or } 84$$

Similarly, the final grade of student who scored 4 and 8 marks can also be obtained by substituting $x_1 = 4$ and $x_2 = 8$ in the regression equation:

$$x_3 = 16.07 + 4.36\,(4) + 4.04\,(8)$$
$$= 16.07 + 17.44 + 32.32 = 65.83 \text{ or } 66$$

Example 15.3: The following data show the corresponding values of three variables x_1, x_2, and x_3. Find the least-square regression equation of x_3 on x_1 and x_2. Estimate x_3 when $x_1 = 10$ and $x_2 = 6$.

$\bar{x}_1$:	3	5	6	8	12	14
$\bar{x}_2$:	16	10	7	4	3	2
$\bar{x}_3$:	90	72	54	42	30	12

Solution: The regression equation of x_3 on x_2 and x_1 can be written as follows:

$$x_3 - \bar{x}_3 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(x_1 - \bar{x}_1)$$

Calculations for regression equations are shown in the table below:

x_1	$(x_1 - \bar{x}_1) = X_1$	X_1^2	x_2	$(x_2 - \bar{x}_2) = X_2$	X_2^2	x_3	$(x_3 - \bar{x}_3) = X_3$	X_3^2	X_1X_2	X_1X_3	X_2X_3
3	- 5	25	16	9	81	90	40	1600	- 45	- 200	360
5	- 3	9	10	3	9	72	22	484	- 9	- 66	66
6	- 2	4	7	0	0	54	4	16	0	- 8	0
8	0	0	4	- 3	9	42	- 8	64	0	0	24
12	4	16	3	- 4	16	30	- 20	400	- 16	- 80	80
14	6	36	2	- 5	25	12	- 38	1444	- 30	- 228	190
48	0	90	42	0	140	300	0	4008	- 100	- 582	720

$$\bar{x}_1 = \frac{\Sigma x_1}{n} = \frac{48}{6} = 8;\ \bar{x}_2 = \frac{\Sigma x_2}{n} = \frac{42}{6} = 7;\ \bar{x}_3 = \frac{\Sigma x_3}{n} = \frac{300}{6} = 50$$

$$s_1 = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2}{n}} = \sqrt{\frac{90}{8}} = \sqrt{15} = 3.87$$

$$s_2 = \sqrt{\frac{\Sigma (x_2 - \bar{x}_2)^2}{n}} = \sqrt{\frac{140}{6}} = \sqrt{23.33} = 4.83$$

$$s_3 = \sqrt{\frac{\Sigma (x_3 - \bar{x}_3)^2}{n}} = \sqrt{\frac{4008}{6}} = \sqrt{668} = 25.85$$

$$r_{12} = \frac{\Sigma X_1 X_2}{\sqrt{\Sigma X_1^2}\sqrt{\Sigma X_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = -0.891$$

$$r_{13} = \frac{\Sigma X_1 X_3}{\sqrt{\Sigma X_1^2}\sqrt{\Sigma X_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = -0.969$$

$$r_{23} = \frac{\Sigma X_2 X_3}{\sqrt{\Sigma X_2^2}\sqrt{\Sigma X_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = 0.961$$

Substituting values in the regression equation, we have

$$(x_3 - \bar{x}_3) = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(x_1 - \bar{x}_1)$$

$$x_3 - 50 = \left[\frac{0.961 - (-0.969 \times -0.891)}{1 - (-0.891)^2}\right]\left(\frac{25.85}{4.83}\right)(x_2 - 7) + \left[\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.891)^2}\right]\left(\frac{25.85}{3.87}\right)(x_1 - 8)$$

$$x_3 - 50 = 2.546\,(x_2 - 7) - 3.664\,(x_1 - 8)$$

$$x_3 = 2.546x_2 - 3.664x_1 + 61.49$$

When $x_1 = 10$ and $x_2 = 6$, we have

$$x_3 = 2.546(6) - 3.664(10) + 61.49 = 15.276 - 36.64 + 61.49 = 40 \text{ approx.}$$

Self-Practice Problems 15A

15.1 The estimated regression equation for a model involving two independent variables and 10 observations is as follows:

$$\hat{y} = 25.1724 + 0.3960\,x_1 + 0.5980\,x_2$$

(a) Interpret b_1 and b_2 in this estimated regression equation

(b) Estimate y when $x_1 = 150$ and $x_2 = 190$

15.2 Consider the linear regression model

$$y = a + b_1x_1 + b_2x_2$$

where y = demand for a product (in units)

x_1 = annual average price of the product (in Rs/unit)

x_2 = $1/a$; a is the advertising expenditure (Rs in lakh)

The data for regression analysis is as follows:

Year	1	2	3	4	5	6	7
y	53.52	51.34	49.31	45.93	51.65	38.26	44.29
x_1	1.294	1.344	1.332	1.274	1.056	1.102	0.930
a	1.837	1.053	0.905	0.462	0.576	0.260	0.363

Write the least-squares prediction equation and interpret the b_1 and b_2 estimates.

15.3 In a trivariate distribution we have: $s_1 = 3$, $s_2 = 4$, $s_3 = 5$, $r_{23} = 0.4$, $r_{31} = 0.6$, $r_{12} = 0.7$

Determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from their means.

15.4 The following constants are obtained from measurement on length in mm(x_1), volume in cc(x_2), and weight in gm(x_3) of 300 eggs:

$\bar{x}_1 = 55.95$, $s_1 = 2.26$, $r_{12} = 0.578$

$\bar{x}_2 = 51.48$, $s_2 = 4.39$, $r_{13} = 0.581$

$\bar{x}_3 = 56.03$, $s_3 = 4.41$, $r_{23} = 0.974$

Obtain the linear regression equation of egg weight on egg length and egg volume. Hence estimate the weight of an egg whose length is 58 mm and volume is 52.5 cc.

15.5 Given the following data:

x_1:	20	25	15	20	26	24
x_2:	3.2	6.5	2.0	0.5	4.5	1.5
x_3:	4.0	5.2	7.5	2.5	3.4	1.5

(a) Obtain the least-squares equation to predict x_1 values from those of x_2 and x_3.

(b) Predict x_1 when $x_2 = 3.2$ and $x_3 = 3.0$.

15.6 The incharge of an adult education centre in a town wants to know as to how happy and satisfied the adult education centre students are. The following four factors were studied to measure the degree of satisfaction:

x_1 = age at the time of completing education

x_2 = number of living children

x_3 = annual income

x_4 = average number of social activities per week

The multiple regression equation was determined to be

$$y = -20 + 0.04x_1 + 30x_2 + 0.04x_3 + 36.3x_4$$

(a) Calculate the satisfaction level for a person who passed out at the age of 45, has two living children, has an annual income of Rs 12,000, and has only one social activity in a week.

(b) Interpret the value of $a = -20$.

(c) Would a person be more satisfied with an additional income of Rs 2000?

15.7 Find the multiple regression equation of x_1, x_2, x_3 from the data relating to three variable given below:

x_1 :	4	6	7	9	13	15
x_2 :	15	12	8	6	4	3
x_3 :	30	24	20	14	10	4

15.8 Data Given the following data:

$\bar{x}_1 = 6$, $\bar{x}_2 = 7$, $\bar{x}_3 = 8$,

$s_1 = 1$, $s_2 = 1$, $s_3 = 3$,

$r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.8$,

(a) Find the regression equation of x_3 on x_1 and x_2.

(b) Estimate the value of x_3 when $x_1 = 4$ and $x_2 = 5$

Hints and Answers

15.1 (a) $b_1 = 0.3960$; every unit increase in the value of x_1 accounted for an increase of 0.3960 in the value of y when the influence of x_2 is held constant.

$b_2 = 0.5980$; every unit increase in the value of x_2 accounted for an increase of 0.5980 in the value of y when influence of x_1 is held constant.

(b) $y = 25.1724 + 0.3960\,(150) + 0.5980\,(190) = 198.1924$

15.2 $\hat{y} = 69.753 - 10.091x_1 - 5.334x_2$

15.3 $$b_{12.3} = \frac{r_{12} - r_{23}r_{13}}{1 - r_{23}^2}\left(\frac{s_1}{s_2}\right) = \frac{0.7 - 0.4 \times 0.6}{1 - (0.4)^2}\left(\frac{3}{4}\right) = 0.411$$

$$b_{13.2} = \frac{r_{13} - r_{12}r_{32}}{1 - r_{23}^2}\left(\frac{s_1}{s_3}\right) = \frac{0.6 - 0.7 \times 0.4}{1 - (0.4)^2}\left(\frac{3}{5}\right) = 0.229$$

Required regression equation is:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 = 0.411x_2 + 0.229x_3$$

15.4 The regression equation of x_3 on x_1 and x_2 can be written as:

$$x_3 - \bar{x}_3 = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\left(\frac{s_3}{s_2}\right)(x_2 - \bar{x}_2) + \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2}\left(\frac{s_3}{s_1}\right)(x_1 - \bar{x}_1)$$

$$x_3 - 56.03 = \left(\frac{0.974 - (0.581) \times 0.578)}{1 - (0.578)^2}\right)\left(\frac{4.41}{4.39}\right)(x_2 - 51.48)$$

$$+ \left(\frac{0.581 - (0.974 \times 0.578)}{1 - (0.581)^2}\right)\left(\frac{4.41}{2.26}\right)(x_2 - 55.95)$$

$$x_3 - 56.03 = \left(\frac{0.974 - 0.336}{1 - 0.334}\right)\left(\frac{4.41}{4.39}\right)(x_2 - 51.48)$$

$$+ \left(\frac{.581 - .563}{1 - 0.334}\right)\left(\frac{4.41}{2.26}\right)(x_2 - 55.95)$$

$$x_3 - 56.03 = 0.962\,(x_2 - 51.48) + 0.053\,(x_1 - 55.95)$$

$$x_3 = 3.54 + 0.053x_1 + 0.962x_2$$

When length, $x_1 = 58$ and volume, $x_2 = 52.5$, weight of the egg would be:

$$x_3 = 3.54 + 0.053(58) + 0.962(52.5)$$
$$= 3.54 + 3.074 + 50.50 = 57.119 \text{ gm}$$

15.5 (a) The least-squares equation of x_1 on x_2 and x_3 is:

$$\hat{x}_1 = a + b_{12.3}x_2 + b_{13.2}x_3 = 21.925 - 0.481x_2 + 0.299x_3$$

(b) For $x_2 = 3.2$ and $x_3 = 3.0$, we have

$$\hat{x}_1 = 21.925 - 0.481\,(3.2) + 0.299\,(3.0) = 21.283$$

15.6 (a) $y = -20 + 0.04x_1 + 30x_2 + 0.04x_3 + 36.3x_4$

$$= -20 + 0.04\,(45) + 30\,(2) + 0.04(12000) + 36.3\,(1)$$
$$= -20 + 1.8 + 60 + 480 + 36.3 = 558.10$$

(b) $a = -20$ implies that in the absence of all variables the person will be negatively satisfied (dissatisfied).

(c) For $x_3 = 14{,}000$ (=12,000 + 2000), we have

$$y = -20 + 0.04(45) + 30\,(2) + 0.04(14000) + 6.48.10 = 638.10$$

15.7 Normal equations are

$$6a + 48b_{12.3} + 102b_{13.2} = 54$$
$$48a + 449b_{12.3} + 1034b_{13.2} = 339$$
$$102a + 1034b_{12.3} + 2188b_{13.2} = 720$$

The required regression equation is

$$x_1 = 16.479 + 0.389x_2 - 0.623x_3.$$

15.8 (a) The regression equation of x_3 on x_1 and x_2 is

$$x_3 - \bar{x}_3 = \frac{s_3}{s_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](x_2 - \bar{x}_2)$$

$$+ \frac{s_3}{s_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](x_1 - \bar{x}_1)$$

$$= \frac{3}{2}\left[\frac{8-(0.6\times0.7)}{1-(0.6)^2}\right](x_2-7) + \frac{3}{1}\left[\frac{0.7-(0.8\times0.6)}{1-(0.6)^2}\right](x_1-6)$$

$$x_3 - 8 = 1.5\left[\frac{0.8-0.42}{0.64}\right](x_2-7) + 3\left[\frac{0.7-0.48}{0.64}\right](x_1-6)$$

$$x_3 = -4.41 + 0.89x_2 + 1.03x_1$$

(b) The value of x_3 when $x_1 = 4$ and $x_2 = 5$ would be

$$x_3 = -4.41 + 0.89\times5 + 1.03\times4 = 4.16$$

15.4 STANDARD ERROR OF ESTIMATE FOR MULTIPLE REGRESSION

When we measure variation in the dependent variable y in multiple linear regression analysis we calculate three types of variations exactly in the same way as in simple linear regression. These variations are

- Total variation or total sum of squares deviation $\quad SST = \sum_{i=1}^{n}(y-\bar{y})^2$
- Explained variation resulting from regression relationship between x and y $\quad SSR = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2$
- Unexplained variation resulting from sampling error $\quad SSE = \sum_{i=1}^{n}(y-\hat{y})^2$

The calculation of each of these sum of squares is associated with a certain number of degrees of freedom. SST has $n-1$ degrees of freedom (n sample observations minus one due to fixed sample mean), SSR has k degrees of freedom (k independent variables involved in estimating value of y), SSE has $n-(k+1)$ degrees of freedom (n sample observations minus $k+1$ constants a, b_1, b_2 . . ., b_k), because in multiple regression we estimate k slope parameters $b_1, b_2, \ldots, b_k$ and an intercept a from a data set containing n observations

These three types of variations are related by the following equation:

$$\sum_{i=1}^{n}(y-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2 + \sum_{i=1}^{n}(y-\hat{y})^2$$

$$SST = SSR + SSE$$

This implies that the total variation in y can be divided into two parts: explained part and unexplained part as shown in Fig. 15.2.

If the multiple regression equation fits the entire data perfectly (i.e., all observations in the data set fall on the regression line), then the estimation of the value of the dependent variable is accurate and there is no error. But if it does not happen, then to know the degree of accuracy in the predication of the value of dependent variable y, we use a measure called *standard error of estimate*.

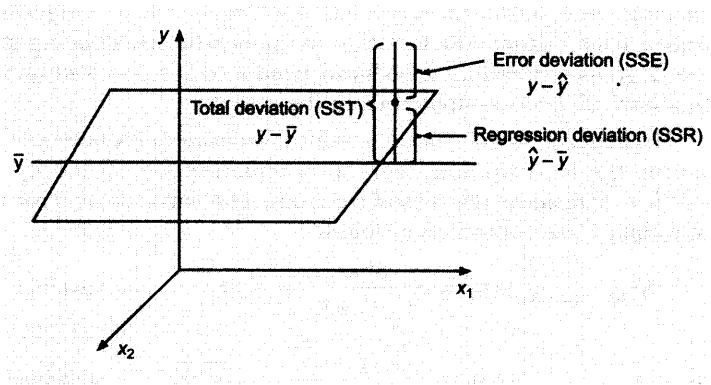


Figure 15.2
Decomposition of Total Variation

In multiple regression analysis, the sample standard error of the estimate represents the extent of variation (dispersion) of observations in the data set with respect to a plane (two independent variables) or a hyperplane (three or more independent variables).

15.4.1 Significance Test of Regression Model

To understand whether all independent variables x_i taken together significantly explain the variability observed in the dependent variable y.The F-test used to test the singnificance of the regression model is based on the fact that at least one of the regression parameters in the regression equation must be zero. To apply F-test, the null hypothesis that all of the true regression parameters are zero, is defined as:

$H_0 : b_1 = b_2 = \ldots = b_k = 0$ ← null hypothesis that y does not depend on x_is'

H_1 : at least one $b_i \neq 0$ ← alternative hypothesis that y depends on at least one of the xs_i'

The significance of regression effect is tested by computing the F-test statistic as shown in Table 15.2

Table 15.2: ANOVA Table for Multiple Regression

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F-Ratio*
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Residual (or Error)	SSE	$n-(k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
Total	SST	$n-1$		

Decision Rule: If the calculated value of F_{cal} is more than its table value at a given level of significance and degrees of freedom k for numerator and $n - k - 1$ for denominator, then H_0 is rejected.

Hence, we conclude that the regression model has no significant prediction for the dependent variable. In other words, rejection of H_0 implies that at least one of the independent variables is adding significant prediction for y.

If two variables are involved in the least-squares regression equation to predict the value of the dependent variable, then the *standard error of estimate* denoted by $S_{y.12}$ is given by

$$S_{y.12} = \sqrt{\frac{SSE}{n-3}} = \sqrt{\frac{\Sigma (y-\hat{y})^2}{n-3}} = \sqrt{\frac{\Sigma y^2 - b_1 \Sigma x_1 y - b_2 \Sigma x_2 y}{n-3}} \qquad (15\text{-}15)$$

The subscript $(y.12)$ lists the dependent variable y for which the prediction is being made with respect to the values of two independent variables coded as 1 and 2. The denominator of this equation indicates that in a multiple regression with 2 independent variables the standard error has $n - (2 + 1) = n - 3$ degrees of freedom (number of unrestricted chances for variation in the measurement being made). This occurs because the degrees of freedom is reduced from n to $2 + 1 = 3$ numerical constants a, b_1 and b_2 that have all been estimated from the sample.

In general, to determine $\hat{y}$ values, we have to estimate $k + 1$ parameters $(a, b_1, b_2, \ldots, b_k)$ for the least-squares regression equation $y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$ with $n - (k + 1)$ residual degrees of freedom. The variance of error term e and standard error of estimate are computed as follows:

$$S^2_{y.12\ldots(k+1)} \text{ or } \sigma_e^2 = \frac{SSE}{n-(k+1)} = MSE \quad \leftarrow \text{variance of error term } e \qquad (15\text{-}16)$$

$$\text{or} \quad S_{y.12\ldots(k+1)} \text{ or } \sigma_e = \sqrt{\frac{\Sigma (y-\hat{y})^2}{n-(k+1)}} = \sqrt{MSE} \quad \leftarrow \text{standard error of estimate}$$

where y = sample values of dependent variable

$\hat{y}$ = estimated values of dependent variable from the regression equation

n = number of observations in the sample

k = number of independent variables

Using the result from Chapter 14, the standard error of estimate (15.15) of y (say x_1) on x_2 and x_3 is defined as:

$$S_{1.23} = \sqrt{\frac{(x_1 - \hat{x}_1)^2}{n-3}}$$

An alternative method of computing $S_{1.23}$ in terms of correlation coefficients r_{12}, r_{13}, and r_{23} is

$$S_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

By symmetry, we may write

$$S_{2.13} = s_2 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}$$

$$S_{3.12} = s_3 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

15.5 COEFFICIENT OF MULTIPLE DETERMINATION

The coefficient of determination in multiple regression denoted $R^2_{y.12}$ or $R^2_{1.23}$ is similar to the coefficient of determination r^2 in the linear regression described in Chapter 14. It represents the *proportion (fraction) of the total variation in the multiple values of dependent variable y, accounted for or explained by the independent variables in the multiple regression model.* The value of R^2 varies from zero to one .

The difference SST – SSE = SSR measures the variation in the sample values of the dependent variable y due to the changes (variations) among sample values of several independent variables in the least-squares equation, Thus $R^2_{y.12}$ (for two independent variables coded as 1 and 2) which measures the ratio of the sum of squares due to regression to the total sum of squared deviations is given by

$$R^2_{y.12} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{S^2_{y.12}}{S^2_y}$$

Adjusted R^2: The value of SST($= S^2_y$) remains same even if additional independent variables are added to the regression model because it represents the sum of squares of the dependent variable. However, additional independent variables are likely to increase value of SSR, so value of R^2 is also likely to increase for additional independent variables.

Sometimes, additional variables added to regression analysis do not provide any significant information, yet R^2 increases. A higher R^2 value can be achieved by adding few additional independent variables, some of which may contribute very little (may be insignificant) in explaining the variation in y-values. A value of R^2 aims to establish the existence of some relationship between the y-values and the independent variables in the regression model. An *adjusted* R^2 value takes into consideration: (i) the additional information which each additional independent variable brings to the regression analysis, and (ii) the changed degress of freedom of SSE and SST.

In particular, the coefficient of multiple determination as a measure of the proportion of total variation in the dependent variable x_1 which is explained by the combined influence of the variations in the independent variables x_2 and x_3 can be defined as:

$$R^2_{1.23} = 1 - \frac{S^2_{1.23}}{S^2_1}$$

where S^2_1 is the variance of the dependent variable x_1.

By symmetry we may write

$$R^2_{2.13} = 1 - \frac{S^2_{2.13}}{S^2_2} \text{ and } R^2_{3.13} = 1 - \frac{S^2_{3.12}}{S^2_3}$$

An adjusted (or corrected) coefficient of determination is defined as:

$$\text{Adjusted } R^2_a = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-(k+1)}$$

Since MSE = SSE/$(n - k - 1)$, therefore adjust R^2_a may be considered as the mixture of the two measures of the performance of a regression model: R^2 and MSE. *The decision to add an additional independent variable in a regression model should weigh the increase in R^2 against the loss of one digree of freedom for error resulting from the addition of the variable.*

The three measures of performance of a regression model: (i) coefficient of determination R^2, (ii) mean square error, MSE, and (iii) adjusted coefficient of determination R^2_a, can be obtained from ANOVA table as shown in Table 15.3.

Table 15.3: Measures of Performance of Regression Analysis

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F-ratio*
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Residual error	SSE	$n-(k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
Total	SST	$n-1$		

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ = Multiple coefficient of determination

$\bar{R}^2_a = 1 - \frac{MSE}{SST/(n-1)}$ = Adjusted multiple coefficient of determination

MSE = Unbiased estimator of the variance of the errors in the multiple regression analysis

$F = \frac{R^2}{(1-R^2)}\left[\frac{n-(k+1)}{k}\right]$ F-ratio to test the existence of a regression relationship between independent variable y and any of the dependent variables.

15.6 MULTIPLE CORRELATION ANALYSIS

Multiple regression model: A mathematical equation that describes how the dependent variable, say y, is related to the independent variables $x_1, x_2, \ldots, x_k$ and a random error term e.

The **multiple correlation coefficient** is denoted by $R_{1.23} = \sqrt{R^2_{1.23}}$ for a dependent variable x_1 and two independent variables x_2 and x_3 involved in the regression equation. In general, the coefficient of multiple correlation measures the extent of the association between a dependent variable and several independent variables taken together.

The multiple correlation coefficient $R_{y.12\ldots}$ is always measured as an absolute number without any arithmetic sign. This is because a few independent variables may have a negative (inverse) relationship with the dependent variable while the remaining may have a positive relationship.

The coefficient of multiple correlation can be expressed in terms of simple linear correlation coefficients r_{12}, r_{13}, and r_{23} as

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

By symmetry we may also write

$$R_{2.13} = \sqrt{\frac{r^2_{21} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{13}}} \text{ and } R_{3.12} = \sqrt{\frac{r^2_{31} + r^2_{32} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{12}}}$$

The value of the multiple correlation coefficient always lies between 0 and 1. The closer it is to 1, better is the linear relationship between the variables. When $R_{y.12} = 0$, it implies no linear relationship between the variables. Further, when $R_{y.12} = 1$, the correlation is called *perfect*.

15.7 PARTIAL CORRELATION ANALYSIS

Partial coefficient of correlation: It describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically.

The **partial correlation coefficients** measure the correlation between the dependent variable and one of the independent variables, holding other independent variables constant rather than ignored in the analysis. If a dependent variable x_1 and two independent variables x_2 and x_3 are included in the partial correlation analysis, then the partial correlation between x_1 and x_2 holding x_3 constant is denoted by $r_{12.3}$. Similarly the partial correlation between x_1 and x_3 holding x_2 constant is denoted by $r_{13.2}$. Depending upon the number of independent variables held constant, we often call partial correlation coefficients as zero-order, first-order, second-order correlation coefficients.

The partial correlation coefficient between x_1 and x_2 keeping x_3 constant is determined as:

$$r^2_{12.3} = b_{12.3} \times b_{21.3} = \frac{r_{12} - r_{13}\, r_{23}}{1 - r^2_{23}}\left(\frac{s_1}{s_2}\right) \times \frac{r_{12} - r_{13}\, r_{23}}{1 - r^2_{13}}\left(\frac{s_2}{s_1}\right)$$

$$= \frac{(r_{12} - r_{13}\, r_{23})^2}{\left(1 - r^2_{23}\right)\left(1 - r^2_{13}\right)}$$

or $r_{12.3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}}$; $r_{13.2} = \dfrac{r_{13} - r_{12}\, r_{32}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{32}}}$ and $r_{23.1} = \dfrac{r_{23} - r_{21}\, r_{31}}{\sqrt{1 - r^2_{21}}\sqrt{1 - r^2_{31}}}$

These results can also be obtained by applying the following formulae:

(a) $r_{12.3} = \sqrt{R^2_{12.3}} = \sqrt{1 - \dfrac{S^2_{1.23}}{S^2_{1.3}}}$ and $r_{13.2} = \sqrt{R^2_{13.2}} = \sqrt{1 - \dfrac{S^2_{1.23}}{S^2_{1.2}}}$

(b) $r_{12.3} = \sqrt{b_{12.3} \times b_{21.3}}$; $r_{13.2} = \sqrt{b_{13.2} \times b_{31.2}}$, and $r_{23.1} = \sqrt{b_{23.1} \times b_{32.1}}$

Limitations of Partial Correlation Analysis

1. While calculating partial correlation coefficient it is assumed that the simple correlation or zero order, correlation from which partial correlation is studied have linear relationship between the variables. In actual practice, particularly in social sciences this assumption is not desirable because linear relationship does not generally exist in such situations.
2. The effects of the independent variables are studied additively not jointly. It means it is presumed that the various independent variables are independent of each other. In actual practice this may not be true and there may be a relationship among the variables.
3. The reliability of the partial correlation coefficient decreases as their order goes up. This means that the second order partial coefficients are not dependable compared to first order ones. Therefore it is necessary that the size of the items in the gross correlation should be large.
4. A lot of computational work is involved and its analysis is not easy.

Standard Error of Partial Correlations The reliability of the partial correlation coefficients is studied though the standard error of $z = 1/\sqrt{n-3}$. In case of partial correlation like $r_{12.3}$, the standard error would be $z = 1/\sqrt{n-(3+1)} = 1/\sqrt{n-4}$, i.e. one more degree of freedom is lost when one independent variable is made constant. Thus if there are four variables under study and two of them have been made constant, then two more degrees of freedom are lost. Thus for partial correlation like, $r_{12.34}$ the degrees of freedom would be $n - 3 - 2$ and the standard error would be $z = 1/\sqrt{n-(3+2)} = 1/\sqrt{n-5}$.

Example 15.4: Varify whether following partial correlation coefficients are significant.

(a) $r_{12.3} = 0.50$; $n = 29$

(b) $r_{12.34} = 0.60$; $n = 54$

Also set up the 95 per cent confidence limits of the correlation.

Solution: (a) Converting r into z, we get

$$z = \frac{1}{2}\log_e \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.50}{1-0.50}$$

$$= 1.1513 \log 3 = 1.1513 \times 0.4771 = 0.549$$

$$\text{The standard error of } z = \frac{1}{\sqrt{n-4}} = \frac{1}{\sqrt{29-4}} = \frac{1}{5} = 0.20$$

The calculated value of z is more than 1.96 times of the standard error at $\alpha = 0.05$ level of sifnificance. Hence the correlation is significant

The 95% conficence limits would be: $0.549 \pm (1.96 \times 0.2) = 0.157$ and 0.941

(b) Converting of r into z, we get

$$z = \frac{1}{2}\log_e \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.60}{1-0.60} = 1.1513 \log 4$$

$$= 1.1513 \times 0.6021 = 0.693$$

$$\text{The standard error of } z = \frac{1}{\sqrt{n-5}} = \frac{1}{\sqrt{54-5}} = \frac{1}{\sqrt{49}} = 0.143$$

The calculated value of z is more than 1.96 times the standard error at $\alpha = 0.05$ level of significance. Hence the correlation is significant

The 95% conficence limits would be: $0.693 \pm (1.96 \times 0.143) = 0.413$ and 0.973

15.7.1 Relationship Between Multiple and Partial Correlation Coefficients

The results connecting multiple and partial correlation coefficients are as follows:

- $1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$

  $1 - R^2_{2.13} = (1 - r^2_{21})(1 - r^2_{23.1})$,

  $1 - R^2_{3.12} = (1 - r^2_{31})(1 - r^2_{32.1})$
- $1 - R_{1.234} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})$

Example 15.5: The correlation between a general intelligence test and school achievement in a group of children aged 6 to 15 years is 0.80. The correlation between the general intelligence test and age in the same group is 0.70 and the correlation between school achievement and age is 0.60. What is the correlation between general intelligence and school achievement in children of the same age? Comment on the result.

Solution: Let x_1 = general intelligence test; x_2 = school achievement; x_3 = age of children.

Given $r_{12} = 0.8$, $r_{13} = 0.7$, and $r_{23} = 0.6$. Then the correlation between general intelligence test and school achievement, keeping the influence of age as constant, we have

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}} = \frac{0.8 - 0.7 \times 0.6}{\sqrt{1 - (0.7)^2}\sqrt{1 - (0.6)^2}}$$

$$= \frac{0.8 - 0.42}{\sqrt{0.51}\sqrt{0.64}} = \frac{0.38}{0.57} = 0.667$$

Hence, we conclude that intelligence and school achievement are associated to each other to the extent of $r_{12.3} = 0.667$ while the influence of the children's age is held constant.

Example 15.6: In a trivariate distribution it is found that $r_{12} = 0.70$, $r_{13} = 0.61$, and $r_{23} = 0.40$. Find the values of $r_{23.1}$, $r_{13.2}$, and $r_{12.3}$. *[Delhi Univ., MCom, 1998]*

Solution: The partial correlation between variables 2 and 3 keeping the influence of variable 1 constant is given by

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{13}}}$$

Substituting the given values, we get

$$r_{23.1} = \frac{0.40 - 0.70 \times 0.61}{\sqrt{1-(0.70)^2}\sqrt{1-(0.61)^2}} = \frac{0.40 - 0.427}{\sqrt{0.51}\sqrt{0.6279}}$$

$$= \frac{0.027}{0.714 \times 0.7924} = \frac{0.027}{0.5657} = 0.0477$$

Similarly, we get $r_{13.2} = \dfrac{r_{13} - r_{12} \bullet r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.61 - 0.70 \times 0.40}{\sqrt{1-(0.70)^2}\sqrt{1-(0.40)^2}} = 0.504$

$$r_{12.3} = \frac{r_{12} - r_{13} \bullet r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.70 - 0.61 \times 0.40}{\sqrt{1-(0.61)^2}\sqrt{1-(0.40)^2}} = 0.633$$

Example 15.7: Based on the following data, calculate $R_{1.23}$, $R_{3.12}$, and $R_{2.13}$.

$\bar{x}_1 = 6.8$	$\bar{x}_2 = 7.0$	$\bar{x}_3 = 74$
$s_1 = 1.0$	$s_2 = 0.8$	$s_3 = 9$
$r_{12} = 0.6$	$r_{13} = 0.7$	$r_{23} = 0.65$

[HP Univ., MBA, 1997]

Solution: The coefficient of multiple determination of two independent variables coded as 2 and 3 is given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

Substituting the given values, we get

$$R_{1.23} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2 \times 0.6 \times 0.7 \times 0.65}{1-(0.65)^2}} = \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}}$$

$$= \sqrt{0.526} = 0.725$$

Similarly

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}} = \sqrt{\frac{(0.7)^2 + (0.65)^2 - 2 \times 0.6 \times 0.7 \times 0.65}{1-(0.6)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.573} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}} = \sqrt{\frac{(0.6)^2 + (0.65)^2 - 2 \times 0.6 \times 0.7 \times 0.65}{1-(0.7)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{0.51}} = \sqrt{0.464} = 0.681$$

Example 15.8: Suppose, a computer has found, for a given set of values of variables x_1, x_2, and x_3 the correlation coefficients are: $r_{12} = 0.91$, $r_{13} = 0.33$, and $r_{23} = 0.81$. Explain whether these computations may be said to be free from errors.

[Madurai Kamaraj Univ., BCom, 1989]

Solution: For determining whether the given computations are correct or not, we calculate the value of the partial correlation coefficient $r_{12.3}$ for variables 1 and 2 keeping the influence of variable 3 constant. If the value of $r_{12.3}$ is less than one, then the computations may be said to be free from errors.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.91 - (0.33)(0.81)}{\sqrt{1-(0.33)^2}\sqrt{1-(0.81)^2}}$$

$$= \frac{0.91 - 0.2673}{\sqrt{1-0.1089}\sqrt{1-0.6561}} = \frac{0.6427}{\sqrt{0.8911 \times 0.3439}} = 1.161$$

Since the calculated values of $r_{12.3}$ is more than one, the computations given in the question are not free from errors.

Example 15.9: The simple correlation coefficients between temperature (x_1), yield of corn (x_2), and rainfall (x_3) are $r_{12} = 0.59$, $r_{13} = 0.46$, and $r_{23} = 0.77$. Calculate the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

[Delhi Univ., MCom; HP Univ., MBA, 1996]

Solution: Partial correlation coefficient $r_{12.3}$ is defined as:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the given values, we get

$$r_{12.3} = \frac{0.59 - 0.46 \times 0.77}{\sqrt{1-(0.46)^2}\sqrt{1-(0.77)^2}} = \frac{0.59 - 0.3542}{\sqrt{1-0.2116}\sqrt{1-0.5529}}$$

$$= \frac{0.2358}{0.5665} = 0.416$$

Multiple correlation coefficient is defined as:

$$R_{1.32} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.59)^2 + (0.46)^2 - 2\,(0.59 \times 0.46 \times 0.77)}{1-(0.77)^2}}$$

$$= \sqrt{\frac{0.3481 + 0.2116 - 0.418}{0.4071}} = \sqrt{\frac{0.5597 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.1417}{0.4071}} = 0.589$$

Example 15.10: A random sample of 5 years on the yield of a crop when observed for seed (x_1), rainfall (x_2) and temperature (x_3) revealed the following information:

$r_{12} = 0.80$	$r_{13} = -0.40$	$r_{23} = -0.56$
$s_1 = 4.42$	$s_2 = 1.10$	$s_3 = 8.50$

Calculate the following:

(a) Partial regression coefficient $b_{12.3}$ and $b_{13.2}$

(b) Standard error of estimate $S_{1.23}$

(c) Coefficient of multiple correlation $R_{1.23}$

(d) Coefficient of partial correlation $r_{12.3}$ between x_1 and x_2 holding x_3 constant

Solution: (a) Substituting the given values in the formulae for partial regression coefficients $b_{12.3}$ and $b_{13.2}$, we get

$$b_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\left(\frac{s_1}{s_2}\right) = \frac{0.80 - (-0.40)(-0.56)}{1-(-0.56)^2}\left(\frac{4.42}{1.10}\right) = 3.370$$

$$b_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\left(\frac{s_1}{s_3}\right) = \frac{-0.40 - 0.80\,(-0.56)}{1-(-0.56)^2}\left(\frac{4.42}{8.50}\right) = 0.036$$

(b) Standard error of estimate

$$S_{1.23} = s_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= 4.42\sqrt{\frac{1-(0.8)^2 - (-0.4)^2 - (-0.56)^2 + 2\,(0.8) - (-0.4)\,(-0.56)}{1-(-0.56)^2}}$$

$$= 4.42\sqrt{\frac{1 - 0.64 - 0.16 - 0.313 + 0.358}{0.6864}} = 2.642$$

(c) Coefficient of multiple correlation

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.4)^2 - 2\,(0.8)\,(-0.4)\,(-0.56)}{1 - (-0.56)^2}} = 0.6433$$

(d) Coefficient of partial correlation $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.8 - (-0.40)\,(-0.56)}{\sqrt{1 - (-0.4)^2}\sqrt{1 - (-0.56)^2}}$$

$$= \frac{0.576}{0.759} = 0.758.$$

Example 15.11: On the basis of observations made on 35 cotton plants the total correlations of yield of cotton (x_1) and number of balls i.e. (the seed vessels) (x_2), and height (x_3) are found to be $r_{12} = 0.863$, $r_{13} = 0.648$, and $r_{23} = 0.709$. Determine the multiple correlation coefficient and partial correlation coefficient $r_{12.3}$ and $r_{13.2}$ and interpret your results.

Solution: The multiple correlation coefficient $R_{1.23}$ is given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.863)^2 + (0.648)^2 - 2\,(0.863)\,(0.648)\,(0.709)}{1 - (0.709)^2}}$$

$$= \sqrt{\frac{0.744 + 0.419 - 0.792}{0.498}} = \sqrt{\frac{0.371}{0.498}} = 0.863$$

The partial correlation coefficients $r_{12.3}$ and $r_{13.2}$ are calculated as:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.863 - 0.648 \times 0.709}{\sqrt{1 - (0.648)^2}\sqrt{1 - (0.709)^2}} = \frac{0.409}{0.762 \times 0.705} = 0.761$$

$$r_{13.2} = \frac{r_{13} - r_{23}\, r_{12}}{\sqrt{1 - r_{23}^2}\sqrt{1 - r_{12}^2}} = \frac{0.648 - 0.709 \times 0.863}{\sqrt{1 - (0.709)^2}\sqrt{1 - (0.863)^2}} = \frac{0.037}{0.705 \times 0.505} = 0.103$$

Example 15.12: The following values have been obtained from the measurement of three variables x_1, x_2, and x_3.

$\bar{x}_1 = 6.80$	$\bar{x}_2 = 7.00$	$\bar{x}_3 = 7.40$
$s_1 = 1.00$	$s_2 = 0.80$	$s_3 = 0.90$
$r_{12} = 0.60$	$r_{13} = 0.70$	$r_{23} = 0.65$

(a) Obtain the regression equation of x_1 on x_2 and x_3

(b) Estimate the value of x_1 for $x_2 = 10$ and $x_2 = 9$

(c) Find the coefficient of multiple determination $R_{1.23}^2$ from r_{12} and $r_{13.2}$

Solution: (a) The regression equation of x_1 on x_2 and x_3 is given by

$$(x_1 - \bar{x}_1) = \left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)(x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)(x_3 - \bar{x}_3)$$

$$x_1 - 6.8 = \left(\frac{0.60 - 0.70 \times 0.65}{1 - (0.65)^2}\right)\left(\frac{1}{0.80}\right)(x_2 - 7)$$

$$+ \left(\frac{0.70 - 0.60 \times 0.65}{1 - 10.657^2}\left(\frac{1}{0.90}\right)(x_3 - 7.4)\right)$$

$$x_1 - 6.8 = \left(\frac{0.60 - 0.455}{0.578}\right)(1.25)\,(x_2 - 7) + \left(\frac{0.70 - 0.39}{0.578}\right)(1.111)\,(x_3 - 7.4)$$

$$= 0.181\,(x_2 - 7) + 0.595\,(x_3 - 7.4)$$

or $$x_1 = 5.670 + 0.181x_2 + 0.595x_3$$

(b) Substituting $x_2 = 10$ and $x_2 = 9$ in the regression equation obtained in part (a), we have

$$x_1 = 5.670 + 0.181(10) + 0.595\ (9) = 12.835$$

(c) Multiple and partial correlation coefficients are related as

$$R^2_{1.23} = 1 - (1 - r^2_{12})\ (1 - r^2_{13.2})$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{23}}} = \frac{0.70 - 0.60 \times 0.65}{\sqrt{1 - (0.60)^2}\sqrt{1 - (0.65)^2}}$$

$$= \frac{0.70 - 0.39}{0.8 \times 0.760} = 0.509$$

or $r^2_{13.2} = 0.259$

Substituting values for r^2_{12} and $r^2_{13.2}$ for $R^2_{1.23}$, we have

$$R^2_{1.23} = 1 - (1 - 0.36)\ (1 - 0.259) = 0.526.$$

Conceptual Questions 15A

1. What is the relationship between a residual and the standard error of estimate? If all residuals about a regression line are zero, what is the value of the standard error of estimate?
2. What information is provided by the variance computed about the regression line by the standard error of estimate?
3. What are 'normal equations' and how are they used in multiple regression analysis?
4. Define the following terms
   (a) Standard error of estimate,
   (b) Coefficient of multiple determination, and
   (c) Coefficient of partial and multiple correlation.
5. Explain the objectives of performing a multiple regression and correlation analysis?
6. Why may it be of interest to perform tests of hypotheses in multiple regression problems?
7. Define partial and multiple correlation. With the help of an example distinguish between partial and multiple correlation.
8. What is multiple linear regression? Explain the difference between simple linear and multiple linear regression.
9. Explain the concept of multiple regression and try to find out an example in the practical field where multiple regression analysis is likely to be helpful.
10. Distinguish between partial and multiple correlation and point out their usefulness in statistical analysis.
11. (a) In the multiple regression equation of x_1 on x_2 and x_3, what are the two progression coefficients and how do your interpret them?
    (b) Explain the concepts of simple, partial, and multiple correlation.
    (c) When is multiple regression needed? Explain with the help of an example.
12. Under what conditions is it important to use the adjusted multiple coefficient of determination?
13. Can you judge how well a regression model fits the data by considering the mean square error only. Explain.
14. Explain why the multiple coefficient of determination never decreases as variables are added to the multiple regression equation.

Self-Practice Problems 15B

15.9 In a trivariate distribution, it is found that $r_{12} = 0.7$, $r_{13} = 0.61$, and $r_{23} = 0.4$. Find the values of the partial correlation coefficients $r_{12.3}$, $r_{23.1}$, and $r_{13.2}$.

15.10 The following data present the values of the dependent variable y and the two independent variables x_1 and x_2:

y :	6	8	9	11	12	14
x_1 :	14	16	17	18	20	23
x_2 :	21	22	27	29	31	32

Compute the following:

(a) Multiple regression coefficients $b_{12.3}$, $b_{13.2}$, $b_{23.1}$
(b) Multiple correlation coefficient $R_{1.23}$
(c) Partial correlation coefficients $r_{12.3}$, $r_{13.2}$, and $r_{23.1}$
(d) Standard error of estimate $S_{1.23}$

15.11 For a given set of values of x_1, x_2, and x_3, the computer has found that $r_{12} = 0.96$, $r_{13} = 0.36$, and $r_{23} = 0.78$. Examine whether these computations may be said to be free from errors.

15.12 The simple correlation coefficients between variable x_1, x_2, and x_3 are, respectively, $r_{12} = 0.41$, $r_{13} = 0.71$, and $r_{23} = 0.50$. Calculate the partial correlation coefficients $r_{12.3}$, $r_{13.2}$, and $r_{23.1}$.

15.13 In a trivariate distribution, it is found that $r_{12} = 0.8$, $r_{13} = 0.4$, and $r_{23} = 0.56$. Find the value of $r_{23.2}$, $r_{13.2}$, and $r_{23.1}$. *[Madras Univ., MCom, 1998]*

15.14 The simple correlation coefficients between profits (x_1), sales (x_2), and advertising expenditure (x_3) of a factory are $r_{12} = 0.69$, $r_{13} = 0.45$, and $r_{23} = 0.58$. Find the partial correlation coefficients $r_{12.3}$, and $r_{13.3}$ and interpret them.

15.15 (a) The simple coefficients of correlation between two variables out of three are as follows.

$r_{12} = 0.8$; $r_{13} = 0.7$, and $r_{23} = 0.6$

Find the partial coefficents of correlation, $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

(b) If $r_{12} = 0.86$, $r_{13} = 0.65$ and $r_{23} = 0.72$, then prove that $r_{12.3} = 0.743$

15.16 On the basis of observations made on 30 cotton plants, the total correlation of yield of cotton (x_1), the number of balls, i.e. seed vessels (x_2) and height (x_3) are found to be;

$r_{12} = 0.80$; $r_{13} = 0.65$, and $r_{23} = 0.70$

Compute the partial correlation between yield of cotton and the number of balls, eliminating the effect of height.

15.17 The following simple correlation coefficients are given:

$r_{12} = 0.98$, $r_{13} = 0.44$, and $r_{23} = 0.54$

Calculate the partial coefficient of correlation between first and third variables keeping the effect second variable constant.

15.18 (a) Do you find the following data consistent:

$r_{12} = 0.07$, $r_{13} = -0.6$, and $r_{23} = 0.90$

(b) The simple correlation coeffiecint between temperature (x_1), yield (x_2) and rainfall (x_3) are $r_{12} = 0.6$, $r_{13} = 0.5$ and $r_{23} = 0.8$. Determine the multiple correlation $R_{1.23}$.

15.19 (a) The following zero order correlation coefficient are given:

$r_{12} = 0.98$, $r_{13} = 0.44$, and $r_{23} = 0.54$

Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independents.

(b) The following zero order correlation coefficients are given: $r_{12} = 0.5$, $r_{13} = 0.6$, and $r_{23} = .7$

Calculate the multiple correlation coefficients: $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Hints and Answers

15.9 $r_{12.3} = \dfrac{r_{12} - r_{13}\,r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$

$= \dfrac{0.70 - 0.60 \times 0.40}{\sqrt{1-(0.60)^2}\sqrt{1-(0.40)^2}} = 0.633$

Similarly $r_{23.1} = 0.049$ and $r_{13.2} = 0.504$

15.10 (a) $b_{12.3} = 0.057$, $b_{13.2} = 0.230$; $b_{23.1} = 0.093$

(b) $R_{1.23} = 0.99$

(c) $r_{12.3} = 0.88$; $r_{13.2} = 0.78$, $r_{23.1} = 0.28$

(d) $S_{1.23} = 0.34$

15.11 Since $r_{12.3} = 1.163$ (> 1), the given computations are not free from error.

15.12 $r_{12.3} = 0.09$, $r_{23.1} = 0.325$ and $r_{13.2} = 0.639$

15.13 $r_{12.3} = 0.759$; $r_{23.1} = 0.436$ and $r_{13.2} = -0.097$

15.14 $r_{12.3} = 0.589$; $r_{13.2} = 0.085$

15.15 (a) $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.8-(0.7\times0.6)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.6)^2}}$

$= \dfrac{0.8-0.42}{0.714\times0.8} = \dfrac{0.38}{0.5712} = 0.665$

$r_{13.2} = \dfrac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.7-(0.8\times0.6)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.6)^2}}$

$= \dfrac{0.7-0.48}{0.6\times0.8} = \dfrac{0.22}{0.48} = 0.458$

$r_{23.1} = \dfrac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} = \dfrac{0.6-(0.8\times0.7)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.7)^2}}$

$= \dfrac{0.6-0.56}{0.6\times0.49} = \dfrac{.04}{0.4284} = 0.093$

(b) $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}}} = \dfrac{0.86-(0.65\times0.72)}{\sqrt{1-(0.65)^2}\sqrt{1-(0.72)^2}}$

$= \dfrac{0.86-0.468}{\sqrt{1-0.4225}\sqrt{1-0.5184}}$

$= \dfrac{0.392}{0.7599\times0.6939} = \dfrac{0.392}{0.527} = 0.743$

15.16 $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.8-(0.65\times0.7)}{\sqrt{1-(0.65)^2}\sqrt{1-(0.7)^2}}$

$= \dfrac{0.8-0.455}{\sqrt{1-0.4225}\sqrt{1-0.49}} = \dfrac{0.345}{0.7599\times0.7141}$

$= \dfrac{0.345}{0.5426} = 0.6357$

15.17 $r_{13.2} = \dfrac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.44-(0.98\times0.54)}{\sqrt{1-(0.98)^2}\sqrt{1-(0.54)^2}}$

$= \dfrac{0.44-0.5292}{\sqrt{1-0.9604}\sqrt{1-0.2916}} = \dfrac{-0.0892}{0.199\times0.842}$

$= \dfrac{-0.0892}{0.1676} = -0.5322$

15.18 (a) $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \dfrac{7-(-0.6\times0.9)}{\sqrt{1-(-0.6)^2}\sqrt{1-(0.9)^2}}$

$= \dfrac{0.7+0.54}{\sqrt{1-0.36}\sqrt{1-0.81}} = \dfrac{1.24}{0.8\times0.436} = 3.56$

Since the value of $r_{12.3}$ is 3.56, the data are inconsistent as the value of any partial coefficient of correlation cannot exceed unity or 1.

$$\text{(b) } R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.5)^2 - 2\times0.6\times0.5\times0.8}{1-(0.8)^2}}$$

$$= \sqrt{\frac{0.36 + 0.25 - 0.48}{1 - 0.64}} = 0.6$$

15.19 (a) $$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2\times0.98\times0.44\times0.54}{1-(0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{1 - 0.2916}} = 0.986$$

$$\text{(b) } R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.5)^2 + (0.6)^2 - 2\times0.5\times0.6\times0.7}{1-(0.7)^2}} = 0.622$$

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{(0.5)^2 + (0.7)^2 - 2\times0.5\times0.6\times0.7}{1-(0.6)^2}} = 0.707$$

$$R_{3.13} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2\times0.5\times0.6\times0.7}{1-(0.5)^2}} = 0.757$$

Formulae Used

1. Multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

2. Estimated multiple regression model

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

3. Sample standard error of the estimate

$$S_{y.12} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b_1 \Sigma x_1 y - b_2 \Sigma x_2 y}{n-3}}$$

4. Coefficient of multiple determination based on sample data for two independent variables

$$R_{y.12}^2 = 1 - \frac{S_{y.12}^2}{S_y^2}$$

5. Partial regression coefficients

$$b_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\left(\frac{s_1}{s_2}\right)$$

$$b_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\left(\frac{s_1}{s_3}\right)$$

6. Coefficient of multiple correlation

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

7. Partial correlation coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

Chapter Concepts Quiz

True or False

1. The coefficient of multiple correlation is the square root of the coefficient of multiple determination. (T/F)
2. Partial regression coefficients are also called net regression coefficients. (T/F)
3. A residual plot is useful to investigate the assumption of linearity but not the equality of conditional variance. (T/F)
4. Multiple regression allows us to use more of the information available to us to estimate the dependent variable. (T/F)
5. The closer the coefficient of multiple correlation is to 1, the better the relationship between the variables. (T/F)
6. The error sum of squares for a multiple regression analysis

is equal to the total sum of squares plus the sum of squares of regression. (T/F)

7. The standard error of estimate in multiple regression has $n-k-1$ degrees of freedom. (T/F)
8. In multiple regression, the variables are collectively very significant, but individually not significant. (T/F)
9. In multiple regression analysis, the regression coefficients often become less reliable as the degree of freedom between the independent variable increases. (T/F)
10. Adding additional variables to a multiple regression will always reduce the standard error of estimate. (T/F)

Multiple Choice

11. Given a regression equation: $\hat{y} = 25.26 + 4.78x_1 + 3.09x_2 - 1.98x_3$. The value of b_2 for this equation is
    (a) 25.26 (b) 4.78
    (c) 3.09 (d) −1.98
12. The degrees of freedom for standard error of estimate are $n-k-1$. What does the k stand for?
    (a) number of observations in the sample
    (b) number of independent variables
    (c) mean of the sample values of dependent variable
    (d) none of these
13. The range of partial correlation coefficient $r_{12.3}$ is
    (a) −1 to 1 (b) 0 to ∞
    (c) 0 to 1 (d) none of these
14. If multiple correlation coefficient $R_{1.23} = 1$, then $R_{2.13}$ is
    (a) 0 (b) −1
    (c) 1 (d) none of these
15. If multiple correlation coefficient $R_{1.23} = 1$, then it implies a
    (a) lack of linear relationship
    (b) perfect relationship
    (c) reasonably good relationship
    (d) none of these
16. Which of following relationship is true?
    (a) $R_{1.23} \le r_{12}$ (b) $R_{1.23} \ge r_{12}$
    (c) $R_{1.23} = r_{12}$ (d) $R_{1.23} \ge -1$
17. In the regression equation $y = a + b_1x_1 + b_2x_2$, y is independent of x when
    (a) $b_2 = 0$ (b) $b_2 = 1$
    (c) $b_2 = -1$ (d) none of these
18. Since $r^2 = 1 - \{(y - \hat{y})^2/(y - \bar{y})^2\}$, then r^2 is equal to
    (a) 1 − SSR/SST (b) 1 − SSE/SSR
    (c) 1 − SST/SSE (d) 1 − SSE/SST
19. In regression analysis, the explained deviation of the dependent variable y is given by
    (a) $\Sigma(y - \bar{y})^2$ (b) $\Sigma(\hat{y} - \bar{y})$
    (c) $\Sigma(y - \bar{y})^2$ (d) none of these
20. The relationship between the multiple correlation coefficient of x_1 on x_2 and x_3 and the standard error of estimate is given by the expression
    (a) $R_{1.23} = \sqrt{1 - \frac{S_{1.23}^2}{S_1^2}}$ (b) $R_{1.23} = \sqrt{1 - \frac{S_{1.23}^2}{S_2^2}}$
    (c) $R_{1.23} = \sqrt{1 - \frac{S_{1.23}^2}{S_3^2}}$ (d) none of these
21. Which of the following relationship is true
    (a) $r_{12.3} = \sqrt{b_{12.3} \times b_{21.3}}$ (b) $r_{12.3} = \sqrt{b_{13.2} \times b_{31.2}}$
    (c) $r_{12.3} = \sqrt{b_{23.1} \times b_{32.1}}$ (c) all of these
22. The coefficient of determination in multiple regression is given by
    (a) $R_{y.12}^2 = 1 - (SSR/SST)$ (b) $R_{y.12}^2 = 1 - (SSE/SSR)$
    (c) $R_{y.12}^2 = 1 - (SST/SSE)$ (d) $R_{y.12}^2 = 1 - (SSE/SST)$
23. The standard error of estimate involved to predict the value dependent variable is given by
    (a) $S_{y.12} = \sqrt{SSE/n-2}$ (b) $S_{y.12} = \sqrt{SSE/n-3}$
    (c) $S_{y.12} = \sqrt{SSR/n-2}$ (d) $S_{y.12} = \sqrt{SSR/n-3}$
24. Adjusted multiple coefficient of determination R_a^2 is given by
    (a) $R_a^2 = 1 - \frac{MSE}{SST/(n-1)}$ (b) $R_a^2 = \frac{MSE}{SSE/(n-1)}$
    (c) $R_a^2 = 1 - \frac{MSR}{SSE/(n-1)}$ (d) $R_a^2 = \frac{MSR}{SSE/(n-1)}$
25. The F-ratio used in testing for the existing of a regression relationship between a dependent variable and any of the independent variable is given by
    (a) $F = \frac{R^2}{1-R^2}\left[\frac{n-(k+1)}{n}\right]$
    (b) $F = \frac{R^2}{1-R^2}\left[\frac{n+(k-1)}{n}\right]$
    (c) $F = \frac{R^2}{1-R^2}\left[\frac{n-(k+1)}{k}\right]$
    (d) $F = \frac{R^2}{1-R^2}\left[\frac{n+(k-1)}{k}\right]$

Concepts Quiz Answers

1. T	2. T	3. F	4. T	5. T	6. F	7. T	8. T	9. T
10. F	11. (c)	12. (b)	13. (a)	14. (c)	15. (b)	16. (b)	17. (d)	18. (b)
19. (b)	20. (a)	21. (a)	22. (d)	23. (b)	24. (a)	25. (c)		

Review-Self Practice Problems

15.20 (a) Given $r_{12} = 0.5$, $r_{13} = 0.4$, and $r_{23} = 0.1$, find the values of $r_{12.3}$ and $r_{23.1}$. [*Kerala Univ., MCom1997*]

(b) If $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$, and $S_1 = 1.0$, find $S_{1.23}$, $R_{1.23}$, and $r_{12.3}$.

[*Kurukshetra Univ., MCom, 1998*]

15.21 (a) If $r_{12} = 0.80$, $r_{13} = -0.56$ and $r_{23} = 0.40$, then obtain $r_{12.3}$ and $R_{1.23}$. [*Saurashtra Univ., MCom, 1997*]

(b) In a triavariate distribution, it is found that $r_{13} = 0.6$, $r_{23} = 0.5$, and $r_{12} = 0.8$. Find the value of $r_{13.2}$. [*Madras Univ., MCom, 1999*]

15.22 In a trivariate distribution

$\bar{x}_1 = 28.20$ $\bar{x}_2 = 4.91$ $\bar{x}_3 = 594$

$s_1 = 4.4$ $s_2 = 1.1$ $s_3 = 80$

$r_{12} = 0.80$ $r_{23} = -0.56$ $r_{31} = -0.40$

(a) Find the correlation coefficient $r_{23.1}$ and $R_{1.23}$.

(b) Also estimate the value of x_1 when $x_2 = 6.0$ and $x_3 = 650$.

15.23 The following data relate to agricultural production (x_1) in quintal/hectare, rainfall (x_2) in inches, and use of fertilizers (x_3) in kg/hectare.

x_1:	85	76	82	83	72	93	76	81
x_2:	6	8	14	11	9	16	5	3
x_3:	40	25	5	20	15	10	35	50

Find $R_{1.23}$ and the coefficient of multiple determination and interpret your result.

15.24 In a study of the factors: honours points (x_1), general intelligence (x_2), and hours of study (x_3), which influence 'academic success', a statistician obtained the following results based on the records of 450 students at a university campus.

$\bar{x}_1 = 18.5$, $\bar{x}_2 = 100$, $\bar{x}_3 = 24$

$s_1 = 11.2$, $s_2 = 15.8$, $s_3 = 6$

$r_{12} = 0.60$, $r_{13} = 0.32$, $r_{23} = -0.35$

Find to what extent honours points a related to general intelligence when hours of study per week are held constant. Also find the other partial correlation coefficients.

15.25 In a trivariate distribution, the following data have been obtained

x_1:	3	5	6	8	12	14
x_2:	16	10	7	4	3	2
x_3:	90	72	54	42	30	12

(a) Find a regression equation of x_3 on x_1 and x_2

(b) Estimate the value of x_3 for $x_1 = 10$ and $x_2 = 6$

(c) Find the standard error of estimate of x_3 on x_1 and x_2

(d) Determine the multiple correlation coefficient $R_{3.12}$

Hints and Answers

15.20 (a) $r_{12.3} = 0.504$; $r_{23.1} = -0.126$

(b) $S_{1.23} = 0.688$; $R_{1.23} = 0.726$; $r_{12.3} = 0.267$

15.21 (a) $r_{12.3} = 0.759$; $R_{1.23} = 0.842$

(b) $r_{13.2} = 0.385$

15.22 (a) $r_{23.1} = -0.436$; $R_{1.23} = 0.802$

(b) Regression equation of x_1 on x_2 and x_3

$$x_1 - \bar{x}_1 = \left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)(x_2 - \bar{x}_2) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)(x_3 - \bar{x}_3)$$

$$x_1 - 28.02 = \left\{\frac{0.8 - (-0.4)(-0.56)}{1 - (-0.56)^2}\right\}\left(\frac{4.4}{1.1}\right)(x_2 - 4.91) + \left\{\frac{(-0.4) - (0.8)(-0.56)}{1 - (-0.56)^2}\right\}\left(\frac{4.4}{80}\right)(x_3 - 594)$$

$$x_1 = 9.225 + 3.356x_2 + 0.003x_3$$

For $x_2 = 6$ and $x_3 = 650$, we have $x_1 = 31.896$.

15.23 $R_{1.23} = 0.975$; $R^2_{1.23} = 0.9512$ implies that 95.12 per cent variation in agriculture production is explained

15.24 $r_{12.3} = 0.80$; $r_{13.2} = 0.71$ and $r_{23.1} = -0.721$

15.25 (a) $x_3 = 61.40 - 3.65x_1 + 2.54x_2$

(b) Estimated value of $x_3 = 40$

(c) $S_{3.12} = 3.12$ (d) $R_{3.12} = 0.9927$

Chapter 16

Forecasting and Time Series Analysis

I have but one lamp by which my feet are guided, and that is the lamp of experience. I know of no way of judging the future but the past.

—Patrick Henry

The penguin flies backwards because he does not care to see where he's going, but wants to see where he's been.

—Fred Allen

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand the pattern of the historical data and then extrapolate the pattern into the future.
- understand the different approaches to forecasting that can be applied in business.
- gain a general understanding of time-series forecasting techniques.
- learn how to decompose time-series data into their various components and to forecast by using decomposition techniques.

16.1 INTRODUCTION

The increasing complexity of the business environment together with changing demands and expectations, implies that every organization needs to know the future values of their key decision variables. Forecasting takes the historical data and project them into the future to predict the occurrence of uncertain events. This may help organizations to assess the future consequences of existing decisions and to evaluate the consequences of decisions (actions or strategies). For example, inventory is ordered without certainty of future sales; new equipment is purchased despite uncertainty about the demand for products; investments are made without knowing profits in future; alternative staff mix is made without knowing the increase in the level of service that can be provided, and so on.

Forecasting is essential to make reliable and accurate estimates of what will happen in the future in the face of uncertainty. A flow chart of forecasts and the decision-making process is shown in Fig. 16.1. In general, the decisions are influenced by the chosen strategy with regard to an organization's future priorities and activities. Once decisions are taken, the consequences are measured in terms of expectation to achieve the desired products/services levels.

Decisions are also get influenced by the additional information obtained from the forecasting method used. Such information and the perceived accuracy of the forecasts may also affect the strategy formulation of an organization. Thus an organization needs

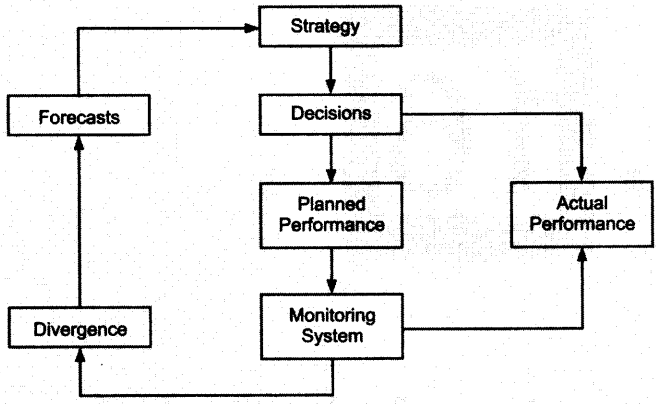


Figure 16.1
Decision-Making Process and Forecasts

to establish a monitoring system to compare planned performance with the actual. Divergence, if any, and no matter what the cause of such divergence between the planned and actual performance, should be fed back into the forecasting process, to generate new forecasts. A few objectives of forecasting are as follows:

(i) The creation of plans of action, because it is not possible to evolve a system of business control without an acceptable system of forecasting.
(ii) Monitoring of the continuing progress of action plans based on forecasts.
(iii) The forecast provides a warning system of the critical factors to be monitored regularly because they might drastically affect the performance of the plan.

16.2 TYPES OF FORECASTS

The objectives of any organization are facilitated by a number of different types of forecasts. These may be related to cash flows, operating budgets, personnel requirement, inventory levels, and so on. However, a broad classification of the types of forecasts is as follows:

Demand Forecasts These are concerned with the predictions of demand for products or services. These forecasts facilitate in formulating material and capacity plans and serve as inputs to financial, marketing, and personnel planning. The forecast itself may be generated in a number of ways, many of which depend heavily upon sales and marketing information.

Environmental Forecasts These are concerned with the social, political, and economic environment of the state and/or the country. Environmental concerns, such as pollution control, are much better managed from an anticipatory rather than an after-the-fact standpoint. Economic forecasts are valuable because they help in predicting inflation rates, money supplies, operating budget, and so on.

Technological Forecasts These are concerned with new developments in existing technologies as well as the development of new technologies. They have become increasingly important to major firms in the computer, aerospace, nuclear, and many other technologically advanced industries.

16.3 TIMING OF FORECASTS

Forecasts are usually classified according to time period and use. The three categories of forecasts are:

Short-Range Forecast This has a time span of upto one year but is typically less than three months. It is normally used in planning purchasing for job scheduling, work force levels, job assignments, production levels, and the like.

Medium-Range Forecast This has a time span from one to three years (typically 3 months to one year). It is used for sales planning, production planning, cash budgeting, and so on.

Long-Range Forecast This has a time span of three or more years. It is used for designing and installing new plants, facility location, capital expenditures, research and development, and so on.

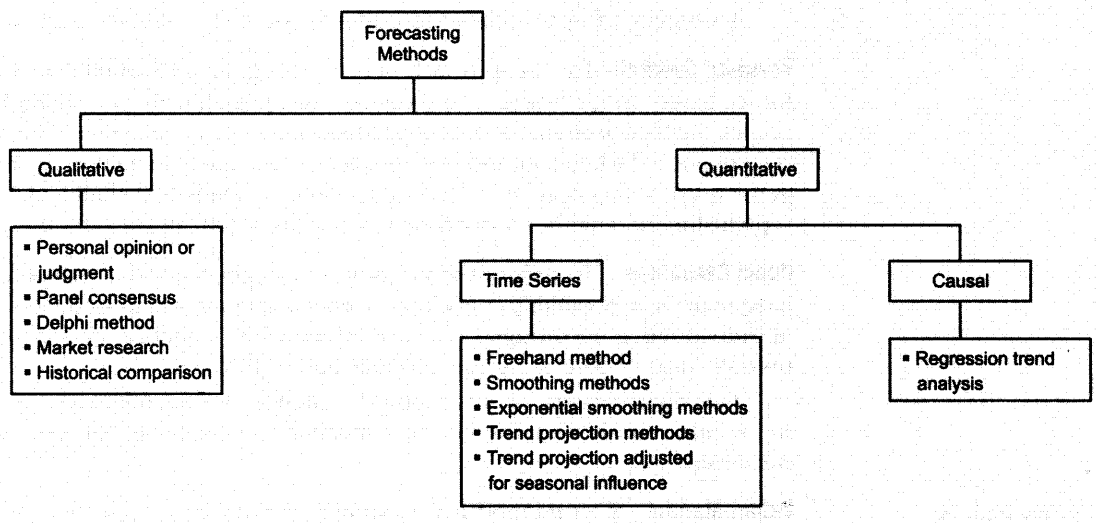
The medium and long-range forecasts differ from short-range forecast on account of following three features:

(i) Medium and long-range forecasts deal with more comprehensive issues and support management decisions regarding design and development of new products, plants, and processes.
(ii) Mathematical techniques such as moving averages, exponential smoothing, and trend extrapolation are used for short-range forecasts.
(iii) The short-range forecasts tend to be more accurate than long-range forecasts. For example, sales forecasts need to be updated regularly in order to maintain their value. After each sales period, the forecast should be reviewed and revised.

16.4 FORECASTING METHODS

Forecasting methods may be classified as either quantitative or qualitative (opinion or judgmental). Figure 16.2 provides an overview of the types of forecasting methods.

Figure 16.2
Forcasting Methods



16.4.1 Quantitative Forecasting Methods

These methods can be used when

(i) past information about the variable being forecast is available,
(ii) information can be quantified, and
(iii) a reasonable assumption is that the pattern of the past will continue into the future.

The quantitative methods of forecasting are further classified into two categories:

Time Series Forecasting Methods A time series is a set of measurements of a variable that are ordered through time. The time variable does not fluctuate arbitrarily. It moves uniformly always in the same direction, from past to future, Thus we can exercise some freedom of choice as to the times at which observations can be made. The time-series data are gathered on a given variable characteristic over a period of time at regular intervals.

The time series forecasting methods attempt to account for changes over a period of time at regular intervals by examining patterns, cycles or trends to predict the outcome for a future time period.

Causal forecasting methods: Forecasting methods that relate a time-series to other variables which are used to explain cause and effect relationship.

Causal Forecasting Methods These methods are based on the assumptions that the variable value which we intend to forecast has a cause-effect relationship with one or more other variables. A linear regression analysis which depends upon the causal relationship or interaction of two or more variables is called causal forcasting method.

16.4.2 Qualitative Forecasting Methods

These methods consist of collecting the opinions and judgments of individuals who are expected to have the best knowledge of current activities or future plans of the organization. For example, knowledge of demand trend and customer plans are often known to marketing executives or product managers. Through regular contact with customers, the marketing and sales personnel are presumably familiar with individual customers or retail market segment. Management usually maintains broader market information on trends by product line, geographic area, customer groups, and so on.

Qualitative forecasting methods have the advantage that they can incorporate subjective experience as inputs along with objectives data. It is the human brain that permits assimilation of all types of information and the ultimate issuance of a prediction.

Since each human being has different knowledge, experience, and perspective of reality, intuitive forecasts are likely to differ from one individual to another. Furthermore, the less they are based upon fact and quantified data, the less they lend themselves to analysis and resolution of differences of opinion. The quantification of data gives them a more precise meaning than words which are inexact and are capable of being misunderstood. Also, if the forecasts prove to be inaccurate there is an objective basis for improvement the next time around.

A number of approaches fall under qualitative methods, and these are as follows:

Personal Opinion In this approach of forecasting, an individual does some forecast of the future based on his or her own judgment or opinion without using a formal quantitative model. Such an assessment can be relatively reliable and accurate. This approach is usually recommended when conditions in the past are not likely to hold in the future. For instance, getting an assessment of whether inventory levels are likely to last until the next replenishment; whether a machine will require repair in the next month, and so on.

Panel Consensus To reduce the prejudices and ignorance that may arise in the individual judgment, it is possible to develop consensus among group to individuals. Such a panel of individuals is encouraged to share information, opinions, and assumptions (if any) to predict future value of some variable under study.

The disadvantage of this method is that it is dependent on group dynamics and frequently requires a facilitator or convenor to coordinate the process of developing a consensus.

Delphi method: A quantitative forecasting method that obtains forecasts through group consensus.

Delphi Method This method is very similar to the panel consensus approach. It uses the collective experience and judgment of a group of experts. In this method, experts may be located in different places and never meet and typically do not know other group members. Each expert is given a questionnaire to complete relating to the area under investigation. A summary is then prepared from all the questionnaires and a copy of it is sent to each expert for revision of responses to the question included in the questionnaire in the light of the summary results. This process of updating the summary results is repeated until the desirable consensus is reached. This method produces a narrow range of forecasts rather than a single view of the future.

Market Research This method is used to collect data based on well-defined objectives and assumptions about the future value of a variable. In this method, a questionnaire is prepared to distribute among respondents. A summery of responses to questions in the questionnaire is prepared to develop survey results.

Historical Comparison Once the data are arranged chronologically, the time-series approach facilitates comparison between one time period and the next. It provides a scientific basis for making comparisons by studying and isolating the effects of various influencing factors on the patterns of variable values. It also helps in making regional comparison amongst data collected on the basis of time.

16.5 STEPS OF FORECASTING

Regardless of the method used to forecast, the following steps are followed:

1. Define objectives and the policies to be achieved, that is, what we are trying to obtain by the use of the forecast. The purpose of forecasting is to make use of the best available present information to guide future activities towards organization's objectives.
2. Select the variables of interest such as capital investment, employment level, inventory level, purchasing of new equipment, which are to be forecasted.
3. Determine the time horizon—short, medium, or long term—of the forecast in order to predict changes which will probably follow the present level of activities.
4. Select an appropriate forecasting model to make projections of the future in accordance to the reasons of past changes which have taken place.
5. Collect the relevant data needed to make the forecast.
6. Make the forecast and implement the results.

These steps present a systematic way of initiating, designing, and implementing a forecasting system. If a particular system is used regularly to generate forecasts, then data should be collected in a routine manner so that computations used to make the forecast can be done automatically using a computer.

16.6 TIME SERIES ANALYSIS

A time series is a set of numerical values of some variable obtained at regular period over time. The series is usually tabulated or graphed in a manner that readily conveys the behaviour of the variable under study. Figure 16.3 presents the export of cement (in tonnes) by a cement company between 1994 and 2004. The graph suggests that the series is time dependent. The management of the company is interested in determining how the series is dependent on time and in developing a means of predicting future levels with some degree of reliability. The nature of the time dependence is often analysed by decomposing the time series into its components.

Year	*Export (tonnes)*
1994	2
1995	3
1996	6
1997	10
1998	8
1999	7
2000	12
2001	14
2002	14
2003	18
2004	19

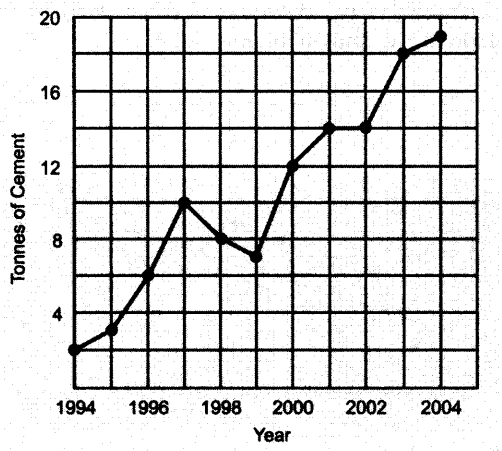


Figure 16.3
Export of Cement

16.6.1 Objectives of Time Series Analysis

1. The assumption underlying time series analysis is that the future will look like the past, that is, the various factors which have already influenced the patterns of change in the value of the variable under study will continue to do so in more or less the same manner in the future. In other words, some underlying pattern exists in historical data. Thus one of the objective of time-series analysis is to identify the pattern and isolate the influencing factors (or effects) for prediction purposes as well as for future planning and control.

2. The review and evaluation of progress made on the basis of a plan are done on the basis of time-series data. For example the progress of our Five-Year Plans is judged by the annual growth rates in the Gross National Product (GNP). Similarly the evaluation of our policy of controlling inflation and price rise is done by the study of various price indices which are based on the analysis of time-series.

16.6.2 Time Series Patterns

We assume that time series data consist of an underlying pattern accompanied by random fluctuations. This may be expressed in the following form:

Actual value of the variable at time t = Mean value of the variable at time t + Random deviation from mean value of the variable at time t

$$\hat{y} = \text{Pattern} + e$$

where $\hat{y}$ is the forecast variable at period t; pattern is the mean value of the forecast variable at period t and represents the underlying pattern, and e is the random fluctuation from the pattern that occurs of the forecast variable at period t.

16.6.3 Components of a Time Series

Time-series: A set of observations measured at successive points in time or over successive periods of time.

Trend: A type of variation in time-series that reflects a long-term movement in time-series over a long period of time.

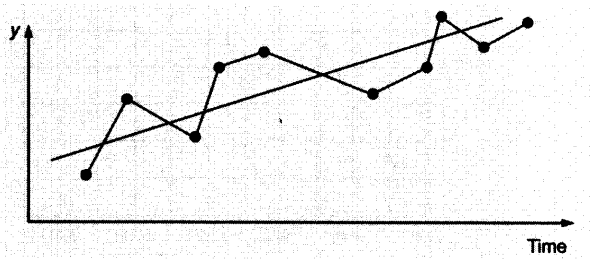
Cyclical variation: A type of variation in time-series, in which the value of the variable fluctuates above and below a trend line and lasting more than one year.

The **time-series** data contain four components: *trend*, *cyclicality*, *seasonality* and *irregularity*. Not all time-series have all these components. Figure 16.4 shows the effects of these time-series components over a period of time.

Trend Sometimes a time-series displays a steady tendency of either upward or downward movement in the average (or mean) value of the forecast variable y over time. Such a tendency is called a trend. When observations are plotted against time, a straight line describes the increase or decrease in the time series over a period of time.

Cycles An upward and downward movement in the variable value about the trend time over a time period are called cycles. A business cycle may vary in length, usually more than a year but less than 5 to 7 years. The movement is through four phases: from *peak* (prosperity) to *contradiction* (recession) to *trough* (depression) to *expansion* (recovery or growth) as shown in Fig. 16.4 .

Figure 16.4
Time-series Effects



Seasonal variation: A type of variation in time-series that shows a periodic pattern of change in time-series within a year; patterns tend to be repeated from year to year.

Seasonal It is a special case of a cycle component of time series in which fluctuations are repeated usually within a year (e.g. daily, weekly, monthly, quarterly) with a high degree of regularity. For example, average sales for a retail store may increase greatly during festival seasons.

Irregular variation: A type of variation in time-series that reflects the random variation of the time-series values which is completely unpredictable.

Irregular Irregular variations are rapid charges or bleeps in the data caused by short-term unanticipated and non-recurring factors. Irregular fluctuations can happen as often as day to day.

16.7 TIME SERIES DECOMPOSITION MODELS

The analysis of time series consists of two major steps:

1. Identifying the various factors or influences which produce the variations in the time series, and
2. Isolating, analysing and measuring the effect of these factors independently, by holding other things constant.

The purpose of decomposition models is to break a time series into its components: Trend (T), Cyclical (C), Seasonality (S), and Irregularity (I). Decomposition of time series aims to isolate influence of each of the four components on the actual series so as to provide a basis for forecasting. There are many models by which a time series can be analysed; two models commonly used for decomposition of a time series are discussed below.

16.7.1 Multiplicative Model

The actual values of a time series, represented by Y can be found by multiplying four components at a particular time period. The effect of four components on the time series is interdependent. The multiplicative time series model is defined as:

$$Y = T \times C \times S \times I \leftarrow \text{Multiplicative model}$$

The multiplicative model is appropriate in situations where the effect of C, S, and I is measured in relative sense and is not in absolute sense. The geometric mean of C, S, and I is assumed to be less than one. For example, let the actual sales for period of 20 months be Y_{20} = 423.36. Further let, this value be broken down into its components as: trend component (mean sales) 400; effect of current cycle (0.90) which decreases sales by 10 per cent; seasonality of the series (1.20) that increases sales by 20 per cent. Thus besides the random fluctuation, the expected value of sales for this period is: 400 × 0.90 × 1.20 = 432. If the random factor decreases sales by 2 per cent in this period, then the actual sales value will be 432 × 0.98 = 423.36.

16.7.2 Additive Model

In this model, it is assumed that the effect of various components can be estimated by adding the various components of a time-series. It is stated as:

$$Y = T + C + S + I \leftarrow \text{Additive model}$$

Here C, S, and I are absolute quantities and can have positive or negative values. It is assumed that these four components are independent of each other. However, in real-life time series data this assumption does not hold good.

Conceptual Questions 16A

1. Briefly describe the steps that are used to develop a forecasting system.
2. What is forecasting? Discuss in brief the various theories and methods of business forecasting. *[Delhi Univ., MBA, 2001]*
3. For what purpose do we apply time series analysis to data collected over a period of time?
4. How can one benefit from determining past patterns?
5. What is the difference between a causal model and a time series model?
6. What is a judgmental forecasting model, and when is it appropriate?
7. Explain clearly the different components into which a time series may be analysed. Explain any method for isolating trend values in a time series.
8. Explain what you understand by time series. Why is time-series considered to be an effective tool of forecasting?
9. Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why? *[Osmania Univ., MBA, 1998]*
10. Identify the four principal components of a time-series and explain the kind of change, over time, to which each applies.
11. What is the advantage of reducing a time series into its four components?
12. Despite great limitations of statistical forecasting, forecasting techniques are invaluable to the economist, the businessman, and the government. Explain.
13. (a) Why are forecasts important to organizations?
    (b) Explain the difference between the terms: seasonal variation and cyclical variation.
    (c) Give reasons why the seasonal component in the time-series is not constant? Give examples where you believe the seasonality may change.
14. Identify the classical components of a time series and indicate how each is accounted for in forecasting.

16.8 QUANTITATIVE FORECASTING METHODS

The quantitative forecasting methods fall into two general categories:

- Time series methods
- Causal methods

The ***time series methods*** are concerned with taking some observed historical pattern for some variable and projecting this pattern into the future using a mathematical formula. These methods do not attempt to suggest why the variable under study will take some future value. This limitation of the time-series approach is taken care by the application of a causal method. The ***causal method*** tries to identify factors which influence the variable is some way or cause it to vary in some predictable manner. The two causal methods, regression analysis and correlation analysis, have already been discussed previously.

A few time series methods such as *freehand curves* and *moving averages* simply describe the given data values, while other methods such as *semi-average* and *least squares* help to identify a trend equation to describe the given data values.

16.8.1 Freehand Method

A freehand curve drawn smoothly through the data values is often an easy and, perhaps, adequate representation of the data. From Fig. 16.3, it appears that a straight line connecting the 1994 and 2004 exports volumes is a fairly good representation of the given data.

The forecast can be obtained simply by extending the trend line. A trend line fitted by the freehand method should conform to the following conditions:

(i) The trend line should be smooth—a straight line or mix of long gradual curves.
(ii) The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line.
(iii) The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible.
(iv) The trend line should bisect the cycles so that area above the trend line should be equal to the area below the trend line, not only for the entire series but as much as possible for each full cycle.

Example 16.1: Fit a trend line to the following data by using the freehand method.

Year	: 1997	1998	1999	2000	2001	2002	2003	2004
Sales turnover (Rs in lakh)	: 80	90	92	83	94	99	92	104.

Solution: Figure 16.5 presents the freehand graph of sales turnover (Rs in lakh) form 1997 to 2004. Forecast can be obtained simply by extending the trend line

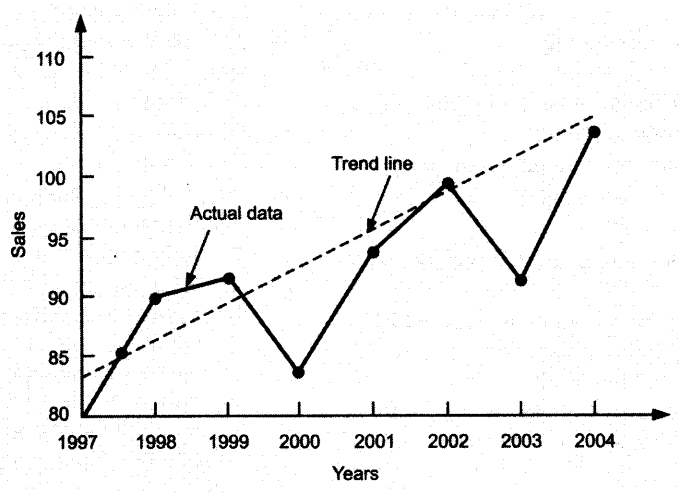


Figure 16.5
Graph of Sales Turnover

Limitations of freehand method

(i) This method is highly subjective because the trend line depends on personal judgment and therefore what happens to be a good-fit for one individual may not be so for another.
(ii) The trend line drawn cannot have much value if it is used as a basis for predictions.
(iii) It is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done.

16.8.2 Smoothing Methods

The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. In this section, we shall discuss three smoothing methods:

(i) Moving averages
(ii) Weighted moving averages
(iii) Semi-averages

The data requirements for the techniques to be discussed in this section are minimal and these techniques are easy to use and understand.

Moving Averages

Moving averages: A quantitative method of forecasting or smoothing a time-series by averaging each successive group of data values.

If we attempt to observe the movement of some variable values over a period of time and try to project this movement into the future, then it is essential to smooth out first the irregular pattern in the historical values of the variable, and later use this as the basis for a future projection. This can be done by using the technique of **moving averages**.

This method is a subjective method and depends on the length of the period chosen for calculating moving averages. To remove the effect of cyclical variations, the period chosen should be an integer value that corresponds to or is a multiple of the estimated average length of a cycle in the series.

The moving averages which serve as an estimate of the next period's value of a variable given a period of length n is expressed as:

$$\text{Moving average, } MA_{t+1} = \frac{\Sigma\{D_t + D_{t-1} + D_{t-2} + \ldots + D_{t-n+1}\}}{n}$$

where t = current time period
D = actual data which is exchanged each period
n = length of time period

In this method, the term 'moving' is used because it is obtained by summing and averaging the values from a given number of periods, each time deleting the oldest value and adding a new value.

The major *advantage* of a moving average is the opportunity it provides to focus on the long-term trend (and cyclical) movements in a time series without the obscuring effect of short-term 'noise' influences.

The *limitation* of this method is that it is highly subjective and dependent on the length of period chosen for constructing the averages. Moving averages have the following three limitations:

(i) As the size of n (the number of periods averaged) increases, it smoothens the variations better, but it also makes the method less sensitive to real changes in the data.
(ii) It is difficult to choose the optimal length of time for which to compute the moving average. Moving averages can not be found for the first and last $k/2$ periods in a k-period moving average.
(iii) Moving averages cannot pick-up trends very well. Since these are averages, it will always stay within past levels and will not predict a change to either a higher or lower level.

(iv) It causes a loss of information (data values) at either end of the original time series.

(v) Moving averages do not usually adjust for such time-series effects as trend, cycle or seasonality.

Example 16.2: Shown is production volume (in '000 tonnes) for a product. Use these data to compute a 3-year moving average for all available years. Also determine the trend and short-term error.

Year	*Production (in '000 tonnes)*	*Year*	*Production (in '000 tonnes)*
1995	21	2000	22
1996	22	2001	25
1997	23	2002	26
1998	25	2003	27
1999	24	2004	26

Solution: The first average is computed for the first 3 years as follows:

$$\text{Moving average (year 1–3)} = \frac{21 + 22 + 23}{3} = 22$$

The first 3-year moving average can be used to forecast the production volume in fourth year, 1998. Because 25,000 tonnes production was made in 1998, the error of the forecast is $\text{Error}_{1998} = 25{,}000 - 22{,}000 = 3000$ tonnes.

Similarly, the moving average calculation for the next 3 years is:

$$\text{Moving average (year 2–4)} = \frac{22 + 23 + 25}{3} = 23.33$$

A complete summary of 3-year moving average calculations is given in Table 16.1.

Table 16.1 Calculation of Trend and Short-term Fluctuations

Year	*Production* y	*3-Year Moving Total*	*3-Yearly Moving Average (Trend values)* $\hat{y}$	*Forecast Error* $(y - \hat{y})$
1995	21	—	—	—
1996	22	(21 + 22 + 23) = 66	66/3 = 22.00	0
1997	23	(22 + 23 + 25) = 70	70/3 = 23.33	−0.33
1998	25	(23 + 25 + 24) = 72	72/3 = 24.00	1.00
1999	24	71	23.67	0.33
2000	22	71	23.67	−1.67
2001	25	73	24.33	0.67
2002	26	78	26.00	0
2003	27	(26 + 27 + 26) = 79	79/3 = 26.33	0.67
2004	26	—	—	—

Odd and Even Number of Years When the chosen period of length n is an odd number, the moving average period is centred on i (middle period in the consecutive sequence of n periods). For instance with $n = 5$, $MA_3(5)$ is centred on the third year, $MA_4(5)$ is centred on the fourth year..., and $MA_9(5)$ is centred on the ninth year.

No moving average can be obtained for the first $(n - 1)/2$ years or the last $(n - 1)/2$ year of the series. Thus for a 5-year moving average, we cannot make computations for the just two years or the last two years of the series.

When the chosen period of length n is an even numbers, equal parts can easily be formed and an average of each part is obtained. For example, if $n = 4$, then the first moving average M_3 (placed at period 3) is an average of the first four data values, and the second moving average M_4 (placed at period 4) is the average of data values 2 through 5. The average of M_3 and M_4 is placed at period 3 because it is an average of data values for period 1 through 5.

Example 16.3: Assume a four-year cycle and calculate the trend by the method of moving average from the following data relating to the production of tea in India:

Year	*Production (million lbs)*	*Year*	*Production (million lbs)*
1987	464	1992	540
1988	515	1993	557
1989	518	1994	571
1990	467	1995	586
1991	502	1996	612

[*Madras, Univ., MCom, 1997*]

Solution: The first 4-year moving average is:

$$MA_3(4) = \frac{464 + 515 + 518 + 467}{4} = \frac{1964}{4} = 491.00$$

This moving average is centred on the middle value, that is, the third year of the series.

Similarly,

$$MA_4(4) = \frac{515 + 518 + 467 + 502}{4} = \frac{2002}{4} = 500.50$$

This moving average is centred on the fourth year of the series.

Table 16.2 presents the data along with the computations of 4-year moving averages.

Table 16.2 Calculation of Trend and Short-term Fluctuations

Year	*Production (mn lbs)*	*4-Yearly Moving Totals*	*4-Yearly Moving Average*	*4-Yearly Moving Average Centred*
1987	464	—	—	—
1988	515	—	—	—
		1964	491.00	
1989	518			495.75
		2002	500.50	
1990	467			503.62
		2027	506.75	
1991	502			511.62
		2066	516.50	
1992	540			529.50
		2170	542.50	
1993	557			553.00
		2254	563.50	
1994	571			572.50
		2326	581.50	
1995	586	—	—	—
1996	612	—	—	—

Weighted Moving Averages

In moving averages, each observation is given equal importance (weight). However, it may be desired to place more weight (importance) on certain periods of time than on others. So *a moving average in which some time periods are weighted differently than others is called a weighted*

Weighted moving average: A quantitative method of forecasting or smoothing a time-series by computing a weighted average of past data values; sum of weights must equal one.

moving average. In such a case different values may be assigned to compute a weighted average of the most recent n values. Choice of weights is somewhat arbitrary because there is no set formula to determine them. In most cases, the most recent observation receives the most weightage, and the weight decreases for older data values.

A weighted moving average is computed as:

$$\text{Weighted moving average} = \frac{\Sigma(\text{Weight for period } n)\,(\text{Data value in period } n)}{\Sigma \text{Weights}}$$

Example 16.4: Vacuum cleaner sales for 12 months is given below. The owner of the supermarket decides to forecast sales by weighting the past three months as follows:

Weight Applied	*Month*
3	Last month
2	Two months ago
1	Three months ago
6	

Months :	1	2	3	4	5	6	7	8	9	10	11	12
Actual sales : (in units)	10	12	13	16	19	23	26	30	28	18	16	14

Solution: The results of 3-month weighted average are shown in Table 16.3

$$\bar{x}_{\text{weighted}} = 3M_{t-1} + 2M_{t-2} + 1M_{t-3}$$

$$= \frac{1}{6}[3 \times \text{Sales last month} + 2 \times \text{Sales two months ago} + 1 \times \text{Sales three months ago}]$$

Table 16.3 Weighted Moving Average

Month	*Actual Sales*	*Three-month Weighted Moving Average*
1	10	—
2	12	—
3	13	—
4	16	$\frac{1}{6}[(3 \times 13) + (2 \times 12) + (1 \times 10)] = \frac{121}{6}$
5	19	$\frac{1}{6}[(3 \times 16) + (2 \times 13) + (1 \times 12)] = \frac{141}{3}$
6	23	$\frac{1}{6}[(3 \times 19) + (2 \times 16) + (1 \times 13)] = 17$
7	26	$\frac{1}{6}[(3 \times 23) + (2 \times 19) + (1 \times 16)] = \frac{201}{2}$
8	30	$\frac{1}{6}[(3 \times 26) + (2 \times 23) + (1 \times 19)] = \frac{235}{6}$
9	28	$\frac{1}{6}[(3 \times 30) + (2 \times 26) + (1 \times 23)] = \frac{271}{2}$
10	18	$\frac{1}{6}[(3 \times 28) + (2 \times 30) + (1 \times 26)] = \frac{289}{3}$
11	16	$\frac{1}{6}[(3 \times 18) + (2 \times 28) + (1 \times 30)] = \frac{231}{3}$
12	14	$\frac{1}{6}[(3 \times 16) + (2 \times 18) + (1 \times 28)] = \frac{182}{3}$

Example 16.5: A food processor uses a moving average to forecast next month's demand. Past actual demand (in units) is shown below:

Month :	43	44	45	46	47	48	49	50	51
Actual demand :	105	106	110	110	114	121	130	128	137

(a) Compute a simple five-month moving average to forecast demand for month 52.
(b) Compute a weighted three-month moving average where the weights are highest for the latest months and descend in order of 3, 2, 1.

Solution: Calculations for five-month moving average are shown in Table 16.4.

Table 16.4 Five-month Moving Average

Month	*Actual Demand*	*5-month Moving Total*	*5-month Moving Average*
43	105	—	—
44	106	—	—
45	110	545	109.50
46	110	561	112.2
47	114	585	117.0
48	121	603	120.6
49	130	630	126.0
50	128	—	—
51	137	—	—

(a) Five-month average demand for month 52 is

$$\frac{\Sigma x}{\text{Number of periods}} = \frac{114+121+130+128+137}{5} = 126 \text{ units}$$

(b) Weighted three-month average as per weights is as follows:

$$\bar{x}_{\text{weighted}} = \frac{\Sigma \text{ Weight} \times \text{Data value}}{\Sigma \text{weight}}$$

where

Month	Weight	×	Value	=	Total
51	3	×	137	=	411
50	2	×	128	=	256
49	1	×	130	=	130
	6				797

$$\bar{x}_{\text{weighted}} = \frac{797}{6} = 133 \text{ units.}$$

Semi-Average Method

The semi-average method permits us to estimate the slope and intercept of the trend line quite easily if a linear function will adequately describe the data. The procedure is simply to divide the data into two parts and compute their respective arithmetic means. These two points are plotted corresponding to their midpoint of the class interval covered by the respective part and then these points are joined by a straight line, which is the required trend line. The arithmetic mean of the first part is the intercept value, and the slope is determined by the ratio of the difference in the arithmetic mean of the number of years between them, that is, the change per unit time. The resultant is a time series of the form : $\hat{y} = a + bx$. The $\hat{y}$ is the calculated trend value and a and b are the intercept and slope values respectively. The equation should always be stated completely with reference to the year where $x = 0$ and a description of the units of x and y.

The semi-average method of developing a trend equation is relatively easy to commute

and may be satisfactory if the trend is linear. If the data deviate much from linearity, the forecast will be biased and less reliable.

Example 16.6: Fit a trend line to the following data by the method of semi-average and forecast the sales for the year 2002.

Year	*Sales of Firm (thousand units)*	*Year*	*Sales of Firm (thousand units)*
1993	102	1997	108
1994	105	1998	116
1995	114	1999	112
1996	110		

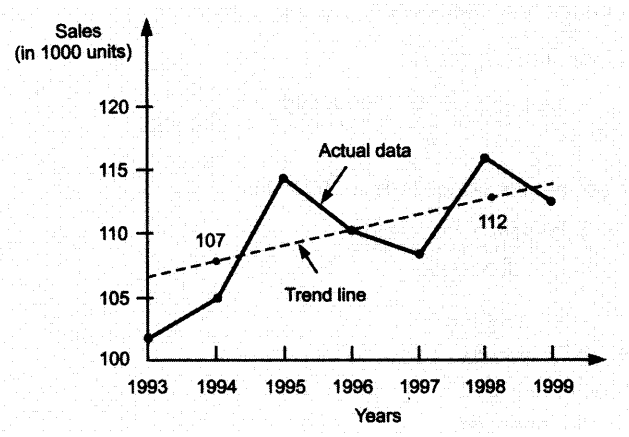
Solution: Since number of years are odd in number, therefore divide the data into equal parts (A and B) of 3 years ignoring the middle year (1996). The average of part A and B is

$$\bar{y}_A = \frac{102+105+114}{3} = \frac{321}{3} = 107 \text{ units}$$

$$\bar{y}_B = \frac{108+116+112}{3} = \frac{336}{3} = 112 \text{ units}$$

Part A is centred upon 1994 and part B on 1998. Plot points 107 and 112 against their middle years, 1994 and 1998. By joining these points, we obtain the required trend line as shown Fig. 16.6. The line can be extended and be used for prediction

Figure 16.6
Trend Line by the Method of Semi-Average



To calculate the time-series $\hat{y} = a + bx$, we need

$$\text{Slope} = b = \frac{\Delta y}{\Delta x} = \frac{\text{change in sales}}{\text{change in year}}$$

$$= \frac{112-107}{1998-1994} = \frac{5}{4} = 1.25$$

$$\text{Intercept} = a = 107 \text{ units at } 1994$$

Thus, the trend line is: $\hat{y} = 107 + 1.25x$

Since 2002 is 8 year distant from the origin (1994), therefore we have

$$\hat{y} = 107 + 1.25(8) = 117$$

16.8.3 Exponential Smoothing Method

Exponential smoothing method: A quantitative forecasting method that uses a weighted average of past time-series values to arrive at new time-series values.

Exponential smoothing is a type of moving-average *forecasting technique which weighs past data from previous time periods with exponentially decreasing importance in the forecast so that the most recent data carries more weight in the moving average.* Simple exponential smoothing makes no explicit adjustment for trend effects whereas adjusted exponential smoothing does take trend effects into account (see next section for details).