CHAPTER 5

SIMPLE REGRESSION AND CORRELATION

5.1 REGRESSION ANALYSIS: AN INTRODUCTION

In social sciences, we frequently encounter variables that are associated in some functional way. For instance, the amount of money spent in advertising a new product may be related to the first month's sales figures for that product, or the height of a father may be related to that of his son. Although such a functional relation of two variables implies nothing about cause and effect, it nevertheless enables us to predict the value of one variable on the condition that we have prior information about the other. This leads to an important topic in statistics namely the **regression analysis**. If two variables are involved, the variable that is the basis of estimation, is conventionally called the **independent variable** and the variable whose value is to be estimated, is called the **dependent variable**. In statistical literature, the dependent variable is variously known as **explained variable, predictand, regressand, response** or **endogenous variable**, while the independent variable is known as **explanatory variable, predictor, regressor, control variable** or **exogenous** variable. In the first example above, advertising budget, which is the basis for sales figures, is the independent variable, while sales figure is the dependent variable. Although it is a matter of personal choice and tradition, we will use the 'dependent variable–explanatory variable' terminology in this text in most of the time.

The term **regression** was first coined in the nineteenth century to describe a biological phenomenon, namely that the progeny of exceptional

individuals tends on average to be less exceptional than their parents and more alike their more distant ancestors. Francis Galton, a cousin of Charles Darwin, studied this phenomenon. He opined that the mean value of a child's characteristic (such as height) was not equal to his or her parent's height but rather was between this value and the average value of the entire population. Thus, for instance, the height of the offspring of very tall people (called by Galton, people "taller than mediocrity") would tend to be shorter than their parents. Similarly, the offspring of those shorter than mediocrity would tend to be taller than their parents. Galton called this phenomenon 'regression to mediocrity', while we call it 'regression to the mean'. More often, the term "regression" is synonymous with "least-squares curve fitting".

With this introduction, we are now in a position to define what a regression analysis is:

Definition* 5.1**: *Regression analysis is a statistical technique that serves as a basis for studying the dependence of one variable, called* ***dependent variable, *on one or more other variables, called* ***explanatory variables.***

The primary objective of a regression analysis is to build a simple regression equation to

a) Estimate the relationship that exists, on the average, between the dependent variable and the explanatory variables.
b) Determine the effect of each of the explanatory variables on the dependent variable, controlling the effects of all other explanatory variables.
c) Predict the value of the dependent variable for a given value of the explanatory variables.

Given below are some situations where regression analysis is appropriate:

1) A company might wish to improve its marketing process. After collecting data on the demand for a product, the product's price, and the advertising expenditure incurred in promoting the product, the company might use regression analysis to develop an equation to predict the future demand on the basis of price and advertising.

2) A real estate company fixes the selling price of its apartments, as it claims, on the basis of size of the apartments measured in terms of square footage of living space. A sample of 20 apartments was chosen and the apartment owners were asked to report the size of their apartments and the price they paid. Given this information, a

regression analysis may be undertaken to see if there is any basis of such claim of the company and to make prediction of the price for a specified floor space.

3) From the knowledge of economics, it is known that, other things remaining the same, the higher the rate of inflation, the lower is the proportion of their incomes that people would want to hold in the form of money. A regression analysis of this relationship will enable the economist to predict the amount of money, as a proportion of their income that people would want to hold at various rates of inflation.

4) A physician collected blood sample from 50 infants on pulmonary blood flow (PBF) and pulmonary blood volume (PBV) to examine if there is any relationship between PBF and PBV. A linear regression analysis seems appropriate for the purpose to see if there is any such relationship.

To see how a regression analysis works in an actual setting, consider a hypothetical example as described below:

Example 5.1: A population consists of 28 families. We are interested to predicting the average height of adult sons knowing the heights of their fathers. To this end, we record the heights in inches of these sons (y) along with the heights of their fathers (x). Here we assume the explanatory variable x to represent the father's height, while the dependent variable y is assumed to represent the son's height. The accompanying table shows the recorded data.

x	y	x	y	x	y	x	y
60	55	70	68	70	69	70	71
65	60	60	58	75	72	75	74
70	65	65	63	60	65	70	72
75	65	70	68	65	65	75	75
60	56	75	70	70	70	75	76
65	62	60	61	75	73	75	77
70	67	65	64	65	66	75	78

When we organize the sons' heights by the heights of their fathers, we obtain a summary table of the following form:

Father's height (x)	Corresponding son's height (y)	Total	Mean
60	55, 56, 58, 61, 65	295	59
65	58, 62, 63, 64, 65, 66	378	63
70	64, 65, 67, 68, 69, 71, 72	476	68
75	66, 69, 70, 72, 73, 74, 75, 76, 77, 78	730	73

Notice that the fathers' heights have been arranged in 4 groups in the first column (from 60 to 75) and the sons' heights have been placed against these groups so that we have 4 fixed values of x and their corresponding y values thereby constituting 4 sub-populations corresponding to each x value. This is because, ordinarily, not all sons, whose fathers have the same height, also have the same height. For example, corresponding to a father's height of 60 inches, we have 5 sons, with respective heights of 55, 56, 58, 61 and 65 inches. Similarly, we have 6 fathers with a common height of 65 inches, which corresponds to heights of 6 sons ranging between 58 and 66 inclusive. Thus for a given x, there is a frequency distribution, which has its own mean and variance. This distribution is known as the **conditional distribution** of y for a fixed value of x. The mean of this distribution is the **conditional mean** ($\mu_{y|x}$). In the above example, the conditional mean of y for a given height of 60 inches of father is 59 inches, i.e. $\mu_{y|60}=59$. If the conditional means of y's for other different values of x are computed and plotted against x, then the equation of the line passing through these points $(x, \mu_{y|x})$, will be called **population regression line** of y on x. Such a line is drawn in Figure 5.1.

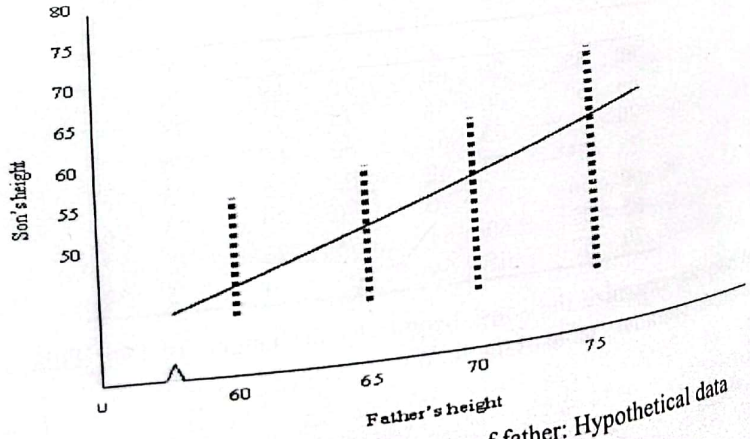

Figure 5.1: Son's height for given height of father: Hypothetical data

This line simply shows how the average height of sons increases with the father's height. Since for each fixed value of x, the height of the regression line represents the arithmetic mean of a theoretically infinite number of y values, the line is also referred to as the **line of conditional means**

The conditional distributions referred to above are assumed to be normal with the same variance, i.e. $\sigma^2_{y|x} = \sigma^2$. In regression analysis, we are primarily interested to study the relationship between $\mu_{y|x}$ and x and the resulting regression equation of $\mu_{y|x}$ on x is more generally called **regression curve**. This curve is simply the locus of the conditional means $\mu_{y|x}$. More simply, it is the curve that connects the means of the sub-populations of y shown against the values of the regressor x. Such a curve is depicted in Figure 5.2. This figure shows that for each x value, there is a population of y values that are spread around the conditional mean of those values.

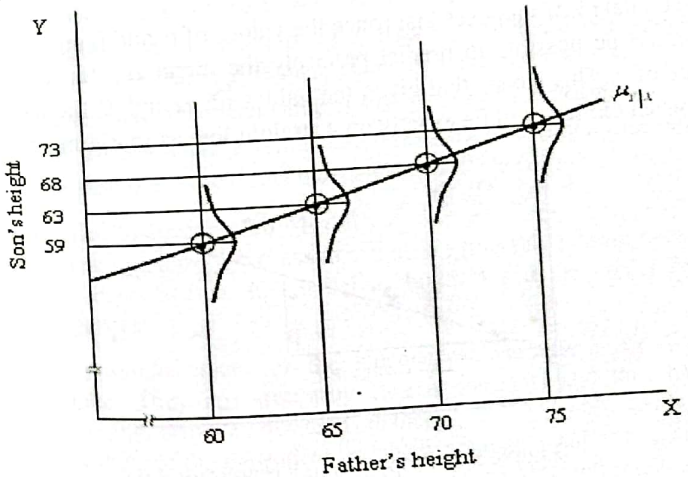

Figure 5.2: Population regression line

5.2 SIMPLE LINEAR REGRESSION MODEL

From the preceding discussion, it seems obvious that each conditional mean $\mu_{y|x}$ is a function of the variable x. Symbolically

$$\mu_{y|x} = f(x) \quad \ldots (5.1)$$

Equation (5.1) is known as the **population regression function** (*prf*). It states merely that the mean of the distribution of y for given x is functionally related to x. In the present context, we assume that the *prf* is a linear function of x, so that it can be represented by an equation of the following form:

$$\mu_{y|x} = \alpha + \beta x \quad \ldots (5.2)$$

where α and β are the unknown constants of the regression function. This function represents a **mathematical model** rather than a statistical model, because it does not allow for any error in predicting $\mu_{y|x}$ as a function of x. By this we mean that $\mu_{y|x}$ always takes the value $\alpha+\beta x_0$ whenever $x=x_0$. Because of this nature, the equation (5.2) represents a **deterministic model**. In regression terminology, the function in (5.2) is referred to as the line of regression of y on x, and β is called the **regression coefficient** of y on x. The properties of this line are shown in Figure 5.3.

The model (5.2) supposes that (once the values of α and β are determined), it would be possible to predict precisely the mean $\mu_{y|x}$ for any specified value of x. This means that given the values of α and β, the mean values $\mu_{y|x}$ when plotted, will lie exactly on a straight line as in Figure (5.3).

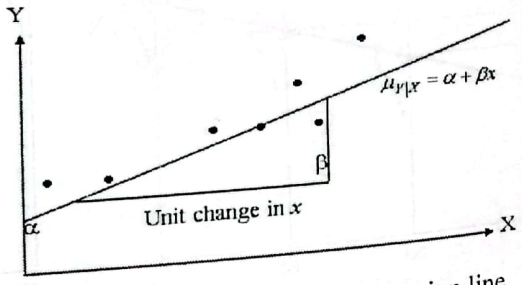

Figure 5.3: Properties of a regression line

In practice, however, such precision is almost never attainable. The observed values will tend to deviate from the $\mu_{y|x}$ values and the most that one can expect that the equation (5.2) is valid subject to some random error. This tells us that the deterministic model is not an exact representation of the relationship between the two variables in question. To represent this phenomenon, we use the simple linear regression model of the type

$$y = \mu_{y|x} + \varepsilon = \alpha + \beta x + \varepsilon \quad \ldots (5.3)$$

Here ε is a **stochastic error term** describing the discrepancy between the observed y and the mean $\mu_{y|x}$:

$$\varepsilon = y - \mu_{y|x} = y - (\alpha + \beta x) \quad \ldots (5.4)$$

The equation (5.3) is a **probabilistic model** which accounts for the random behavior of y exhibited in Figure 5.3 and provides a more accurate description of reality than the deterministic model described by the equation (5.2). This is a model of what we believe to represent the observed situation. The interpretation of the underlying model is as follows:

(a) $\mu_{y|x}$: The mean value of the dependent variable y when the value of the independent variable is x.
(b) α : The y–intercept. It is the mean value of y when x equals 0.
(c) β : The slope. It measures the change (amount of increase or decrease) in the mean value of y, associated with a one-unit increase in x. If β is positive, the mean value of y increases as x increases. If β is negative, the mean value of y decreases as x increases.
(d) ε : A stochastic error term that describes the effects of all factors on y other than the value of the independent variable x.

5.2.1 Properties of Regression Model

In linear regression model we assume that the true relationship between x and y can be described by the model as in (5.3) and the model has the following properties:

(a) The possible values of the independent variable x are fixed in advance. They are arbitrarily chosen constants and thus have no observation errors associated with them.
(b) The values of the dependent variable y are dependent on the values of x. The variable y possesses a random property; it is left free to take on any value that may possibly be associated with a given value of x.
(c) The ε 's are uncorrelated and normally distributed random variables with a mean zero and constant variance σ^2.
(d) The distribution of y values corresponding to a pre-determined x value is normal with mean $\mu_{y|x}$ (the mean of y for a given value of x).
(e) The conditional probability distribution of y has the same variance and thus the same standard deviation for each of the possible values of x.
(f) The y values are statistically independent of each other.

(g) The mean values will lie on a straight line, which is the population regression line. An alternative way of stating this assumption is that the linear model is correct.

The equation (5.3) is variously known as **linear population regression, population regression model** or simply **linear regression equation.** In regression analysis, our interest is in estimating this line, i.e. in estimating the values of the unknowns α and β on the basis of the observations on y and x. This part of the job will be undertaken in section 5.5 of this chapter.

5.3 TYPES OF REGRESSION ANALYSIS

Although infinitely many different statistical models can be used to represent the mean value of the dependent variable y as a function of one or more explanatory variables, we will concentrate on what we call **linear statistical models**. If y is a dependent variable and x is a single explanatory variable, it may be reasonable in some situations to use the model $\mu_{y|x} = \alpha + \beta x$ for unknown parameters α and β. If the model relates $\mu_{y|x}$ as a linear function of α and β only, the model is called a **simple linear regression model**. If more than one explanatory variable, say $x_1, x_2, \ldots, x_k$ are of interest, and we model $\mu_{y|x}$ by

$$\mu_{y|x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

the model is called a **multiple linear regression** model. This is discussed in Chapter 6. With one independent variable, it is frequently assumed that the regression function is a polynomial in the independent variable. This type of regression is known as **polynomial regression**. In such cases, we model $\mu_{y|x}$ by

$$\mu_{y|x} = \alpha + \beta_1 x + \beta_2 x^2$$

which is a second degree polynomial function of the independent variable x with $x_1 = x$ and $x_2 = x^2$. This model would be appropriate for a response that traces a segment of a parabola over the experimental region.

5.3.1 Linearity in the Model

A regression analysis may involve a linear model or a nonlinear model. The term **linearity** is used to describe two aspects of the relationship between the response and a set of independent variables, namely, (i) linearity with respect to the variables and (ii) linearity with respect to the parameters.

Consider the following models which relate the mean of y to two independent variables x_1 and x_2 :

(a) $\mu_{y|x} = \alpha + \beta_1 x_1 + \beta_2 x_2$

(b) $\mu_{y|x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 x_2$

(c) $\mu_{y|x} = \alpha + \beta_1 x_1 + \beta_1^2 x_2$

(d) $\mu_{y|x} = \alpha x_1^{\beta_1} x_2^{\beta_2}$

(i) **Linearity in the parameters:** Linearity in the parameter implies that the conditional mean $\mu_{y|x}$ is a linear function of the parameters and it may or may not be linear in the variable x. Models (a) and (b) are linear in the parameters, since the parameters α, β_1 and β_2 appear linearly. Model (c) is linear in α but non-linear in β_1 since the coefficient of x_2 is β_1^2. Model (d) is linear in the parameter α but non-linear in β_1 and β_2, which appear as exponents.

(ii) **Linearity in the variables:** By linearity in variable, we mean that the conditional mean is a linear function in x. The regression curve in this case represents a straight line. Models (a) and (c) are linear in variables x_1 and x_2, while the models (b) and (d) are non-linear in the variables, since they include non-linear functions of x_1 and x_2.

Of the two interpretations, linearity in the parameters is relevant in the regression analysis. Therefore, for our purpose, the term **linear** will always mean a regression that is linear in the parameters; it may or may not be linear in the explanatory variable. Thus, $\mu_{y|x} = \alpha + \beta x$, which is linear both in the parameter and variable, is representation of a linear regression model, and so is $\mu_{y|x} = \alpha + \beta x^2$, which is linear in parameter but not linear in variable.

Our discussion in this chapter will be restricted to simple linear regression only with two variables x and y, in which case the equation describing the relationship between x and y is assumed to be linear and can be graphically represented by a straight line. When variables are found to be related, we often want to know how close the relationship is. The degree or closeness of the relationship is commonly referred to as the **correlation** between the variables. The problem of correlation is intimately associated with that of regression and is an integral part of bivariate analysis. This topic will be taken up later in this chapter.

5.4 SCATTER DIAGRAM

In simple regression analysis, we assume that the relationship between the dependent variable, denoted y, and the explanatory variable, denoted x, can be approximated by a straight line equation. We can tentatively decide whether there is an approximate straight-line relationship between y and x by drawing a diagram called **scatter diagram** (also called scatter plot) of y versus x. Such a diagram gives us a visual impression of the relationship involved and suggests the type of model that may best fit the data.

The conventional procedure in constructing a scatter diagram is to have the independent (explanatory) variable x scaled on the horizontal axis and the dependent variable y on the vertical axis. A point representing a pair of observations of x and y is plotted, the resulting graph of all the points thus plotted for all the pairs of x and y values in the sample, is the scatter diagram. If the y values tend to increase or decrease in a straight-line fashion, as the x values increase, and if there is a scattering of the (x, y) points along a straight line, then it is reasonable to describe the relationship between x and y by using a simple regression model. Such a typical diagram appears in Figure 5.4 below.

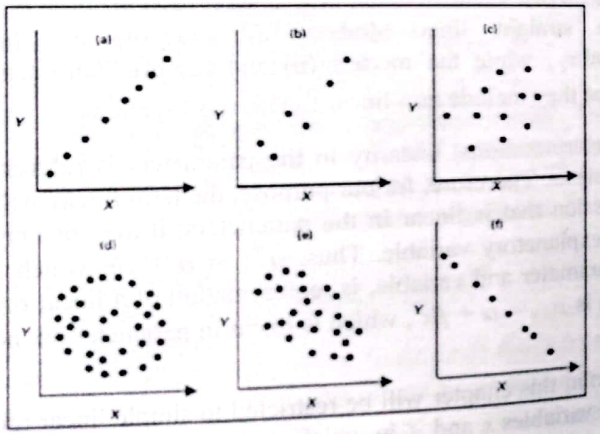

Figure 5.4: Scatter diagram

Once we are reasonably assured that a **linear relation** exists between the two variables, our next task is to estimate the true relationship. The simplest and crudest way of doing this is the so-called **freehand method**. This method involves drawing a straight line freehand near or through the points so that the line appears to best describe the relationship. The

principal drawback of such a method is, of course, the absence of precision in the measurement of any prediction based on such a line. It is because of this reason, we use a refined method that takes care of this limitation. The method, so employed, is called the **least-squares method** and is discussed in the following section.

5.5 THE LEAST-SQUARES METHOD

The least-squares method is a powerful procedure used for estimating parameters particularly in regression analysis by minimizing the difference between the observed response and the value predicated by the model. For example, if the mean value of the response variable y is of the form

$$\mu_{y|x} = \alpha + \beta x \qquad \ldots (5.5)$$

then the least-squares estimators a and b of the parameters α and β may be obtained from n pairs of the sample values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ by minimizing the sum of squares of the vertical deviations from the fitted line. With the estimators a and b, the i^{th} predicted y value when $x=x_i$ is

$$\hat{y}_i = a + bx_i \qquad \ldots (5.6)$$

The difference $y_i - \hat{y}_i$ between the observed and the estimated values of y at $x = x_i$ is called the residual or error corresponding to y_i and the quantity $\sum (y_i - \hat{y}_i)^2$ is called the sum of squares of residuals or **error sum of squares (SSE)**.

Given the observations (x_i, y_i), different pairs of values of a and b will yield different values of this sum of squares. The method of least squares is a neat solution to this problem, which estimates a and b in such a manner that this sum of squares is a minimum. The resulting estimators a and b are called the **least squares estimators** of α and β and the line $\hat{y}_i = a + bx_i$ is the least-squares line which is completely defined if a and b are known. Thus the least-squares line is the line that minimizes

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (a + bx_i)]^2 \qquad \ldots (5.7)$$

where e_i is the deviation of the observed value of y from the $\hat{y}$ line. Since the least-squares line minimizes the sum of squared deviations from the conditional means, the least squares fit is usually regarded as the **best** fit.

Our problem now is to compute the values of a and b that make the sum of

squares of e_i as small as possible. One method of doing this is to set the partial derivatives of $\sum e_i^2$ with respect to both a and b equal to zero and solve the resulting equations. Thus differentiating first with respect to a and equating to zero

$$\frac{\partial \sum e_i^2}{\partial a} = \frac{\partial \left[\sum (y_i - (a + bx_i))\right]^2}{\partial a}$$

$$= -\sum 2[y_i - (a + bx_i)]$$

$$= -2\left(\sum y_i - na - b\sum x_i\right) = 0 \quad \ldots (5.8a)$$

and

$$\frac{\partial \sum e_i^2}{\partial b} = \frac{\partial \left\{\sum [y_i - (a + bx_i)]^2\right\}}{\partial b}$$

$$= -\sum 2[y_i - (a + bx_i)]x_i$$

$$= -2\left(\sum x_i y_i - a\sum x_i - b\sum x_i^2\right) = 0 \quad \ldots (5.8b)$$

From (5.8a) and (5.8b), we arrive at

$$\sum y_i = na + b\sum x_i \quad \ldots (5.8c)$$

and

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \ldots (5.8d)$$

The equations (5.8c) and (5.8d) are known as the **normal equations.** These equations are linear in a and b and hence can be solved simultaneously. The solution for b is:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \ldots (5.8e)$$

The quantity in the numerator of (5.8e) is known as the sum of product of x and y, while the quantity in the denominator is known as the sum of squares of x. In all subsequent discussions, we will denote these quantities by S_{xy} and S_x respectively so that b can be written as

$$b = \frac{S_{xy}}{S_x} \quad \ldots (5.8f)$$

Once b is computed, a can be obtained as

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} = \bar{y} - b\bar{x} \quad \ldots (5.8g)$$

For all computational purposes, b can be expressed as follows:

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} \quad \ldots (5.8h)$$

Since $a = \bar{y} - b\bar{x}$, the estimated or fitted regression line is thus

$$\hat{y}_i = a + bx_i = \bar{y} + b(x_i - \bar{x}) \quad \ldots (5.8i)$$

The line represented by the above equation is our estimate of the population regression equation $\mu_{y/x} = \alpha + \beta x$.

5.5.1 Properties of Sample Regression Line

1. The regression line passes through the mean values of y and x.
2. The mean value of the estimated y (i.e. $\hat{y}$) is equal to the mean value of the observed (actual) y. This is proved as follows:

$$\hat{y}_i = a + bx_i = (\bar{y} - b\bar{x}) + bx_i = \bar{y} + b(x_i - \bar{x})$$

Summing both sides of the above equation and dividing throughout by the sample size n, and noting that $\sum (x_i - \bar{x}) = 0$, we find that

$$\frac{\sum \hat{y}_i}{n} = \frac{\sum \bar{y}}{n} \Rightarrow \bar{\hat{y}} = \bar{y}$$

3. The sum and hence the mean of the residual e_i is zero.

$$\sum e_i = \sum (y_i - \hat{y}_i) = \sum y_i - \sum \hat{y}_i = n\bar{y} - n\bar{\hat{y}} = 0, \text{ since } \bar{\hat{y}} = \bar{y}$$

4. The residuals e_i's are uncorrelated with $\hat{y}_i$'s. That is $\sum \hat{y}_i e_i = 0$
5. The residuals e_i's are uncorrelated with x_i's. That is $\sum x_i e_i = 0$.

The proof of the property 5 above follows from the partial derivative of $\sum e_i^2$ when set to zero. By virtue of (5.8b), we have

$$-2\sum (y_i - a - bx_i)x_i = -2\sum (y_i - \hat{y}_i)x_i = 0 \Rightarrow \sum e_i x_i = 0$$

Example 5.2 A department store has the following statistics on sales (y) for a period of last one year of 10 salesmen, who have varying years of sales experience (x).

(i) Find the regression line of y on x

(ii) Predict the annual sales volume of persons who have 12 and 15 years of sales experience

Table 5.1: Sales figures and years of experience

Salesperson (i)	Years of experience	Annual sales (in '000 taka)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

The required computations are shown in the accompanying table

Salesperson	x_i	y_i	x_i^2	$x_i y_i$
1	1	80	1	80
2	3	97	9	291
3	4	92	16	368
4	4	102	16	408
5	6	103	36	618
6	8	111	64	888
7	10	119	100	1190
8	10	123	100	1230
9	11	117	121	1287
10	13	136	169	1768
Total	70	1080	632	8128

Calculation of $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{70}{10} = 7 \text{ and } \bar{y} = \frac{\sum y_i}{n} = \frac{1080}{10} = 108$$

Calculation of a and b:

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{8128 - \frac{70 \times 1080}{10}}{632 - \frac{(70)^2}{10}} = 4$$

$$a = \bar{y} - b\bar{x} = 108 - 4(7) = 80$$

Thus the regression line estimated by employing least-squares method is

$$\hat{y}_i = 80 + 4x_i$$

Note that the slope b is positive. This implies that as the average years of experience (x) increases, so does the annual sales (y). Thus we would say that for our sample data, there appears to have a positive association between x and y.

The estimate of α is of little significance. Its only importance lies in the fact that it locates the regression line at the point when $x = 0$. Thus, if the firm employs a person without any experience (i.e. x=0), the average increase in sales volume will be almost Tk. 80 thousand.

On the other hand, the slope b is of great significance. It represents an estimate of the average change in the value of the dependent variable y for each unit change in the independent variable x. In this particular example, the value of $b = 4$ means that for an average increase of one year sales experience of a salesperson, the sales volume would increase on the average by Tk. 4 thousand.

We will now use the values of a and b to estimate the sales for x=12 and 15 years. Putting a =80 and b = 4 in the estimated equation, we obtain

(i) Estimated sales for x =12 is $\hat{y}_{(12)} = 80 + 4\,(12) =$ Tk. 128

(ii) Predicted sales for x =15 is $\hat{y}_{(15)} = 80 + 4(15) =$ Tk. 140

In some situations, the slope b could be negative, indicating that as x increases, y decreases, in which case there exists a negative relationship between x and y. The following example illustrates this phenomenon.

Example 5.3: A bank is planning to introduce a new word processing system to its secretarial staff. To learn about the amount of training that is needed to effectively implement the new system, the bank chose 8 employees of roughly equal skill. These employees were trained for varying durations of time and were then individually put to work on a certain project. The following data indicate the training times and the resulting times (both in hours) that it took each employee to complete the project.

Employee #	Training time (x)	Time to complete the project (y)
1	22	18.4
2	18	19.2
3	30	14.5
4	16	19.0
5	25	16.6
6	20	17.7
7	10	24.4
8	14	21.0
Total	155	150.8

(a) Estimate the least-square line.
(b) Predict the amount of time it would take a worker who receives 28 hours of training to complete the project.
(c) Predict the amount of time it would take a worker who receives 50 hours of training to complete the project.

Solution: (a) The quantities to be computed for the purpose of fitting a regression line are a and b. These are

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{2796.4 - \frac{155 \times 150.8}{8}}{3285 - \frac{(155)^2}{8}} = -0.44$$

and

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} = \frac{150.8}{8} - (-.44)\frac{155}{8} = 27.38.$$

so that the estimated least-squares line is

$$\hat{y}_i = 27.38 - 0.44x_i$$

A close examination of the data reveals that as training time increases, less and less time is required to complete the training. This is reflected in the negative value of b (= –.44). This implies that an one hour enhanced training will reduce the time to complete the project work by 0.44 hours, i.e. 26.4 minutes.

(b) The best prediction of the completion time for a training duration of 28 hours on average is

$$\hat{y}_{(28)} = 27.38 - 0.44x_i = 27.38 - 0.44(28) = 15.06 \text{ hours.}$$

(c) The input value (=50) is far away from the range of values observed. We must therefore be cautious to make prediction in such cases, although the estimated equation can fairly make such prediction. Keeping this limitation in view, we estimate the completion time for a training duration of 50 hours:

$$\hat{y}_{(50)} = 27.38 - 0.44x_i = 27.38 - 0.44(50) = 5.38 \text{ hours.}$$

In situations where x and y are dependent on each other, we obtain two lines of regression. In the case when x is assumed to be independent variable and y as dependent variable, the regression is said to be regression of y on x and the estimating regression line is of the form $\hat{y}_i = a + bx_i$ as we have discussed above. When x acts as dependent variable and y as independent, we have a regression of x on y and the resulting regression line is of the form

$$\hat{x}_i = c + dy_i \qquad \ldots (5.9)$$

The formulae for estimating c and d follow the same procedure as in the case of estimating a and b by least-squares method. Thus the formulae for d (the regression coefficient of x on y) and c are

$$d = \frac{S_{xy}}{S_{yy}} \qquad \ldots (5.10)$$

Having obtained d, we obtain c as follows:

$$c = \frac{\sum y_i}{n} - d\frac{\sum x_i}{n} = \bar{y} - d\bar{x} \qquad \ldots (5.11)$$

When do we expect two lines of regression: regression of y on x, and regression of x on y? If the straight line is so chosen that the sum of squares of deviations parallel to the axis of y is minimized, we get the **line of regression of y on x** and it will give the best estimate of y for any given value of x.

If on the other hand, the sum of squares of the deviations parallel to the x axis is minimized, the resulting straight line is the **line of regression of x on y** and it gives the best estimate for any given value of y.

Example 5.4: The chairman of a marketing department at a large private university undertakes a study to relate starting salary (y) after graduation for marketing majors to grade point average (GPA) in major courses. To do this, records of 10 recent marketing graduates are randomly selected. The GPA (x) and the corresponding starting salary were as follows:

GPA (x)	Observed salary (y)	Estimated salary ($\hat{y}$)
3.26	33.8	33.5
2.60	29.8	29.2
3.35	33.5	34.1
2.86	30.4	30.9
3.82	36.4	37.2
2.21	27.6	26.6
3.47	35.3	34.9
3.28	35.0	33.6
2.54	26.5	28.8
3.25	33.8	33.4

(a) Estimate the least squares prediction equation of y on x.
(b) Find the point prediction of starting salary corresponding to each of the GPAs 2.75 and 3.75.
(c) Compare the observed and the estimated salary graphically.

Solution: (a) Let the prediction model be

$$y = \alpha + \beta x + \varepsilon$$

where α and β are the parameters of the model and ε is the random error.

The least squares estimates of the parameters are a and b where

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

and

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

You can easily check that

$$\sum x = 30.64, \sum y = 322.10, \sum x^2 = 96.08, \sum xy = 1001.33$$

so that

$$b = \frac{1001.33 - \frac{(30.64)(322.10)}{10}}{96.08 - \frac{(30.64)^2}{10}} = 6.55$$

and

$$a = \frac{322.10}{10} - (6.55)\left(\frac{30.64}{10}\right) = 12.14.$$

Hence the estimated regression equation is

$$\hat{y} = 12.14 + 6.55x.$$

(b) The estimated starting salaries for GPAs corresponding to 2.75, and 3.75 are respectively

$$\hat{y}_{(2.75)} = 12.14 + 6.55(2.75) = 30.15$$

and

$$\hat{y}_{(3.75)} = 12.14 + 6.55(3.75) = 36.70$$

(c) The estimated values of the staring salaries are shown in the last column of the table above and the resulting graphs of the observed and the expected values are displayed in the figure below:

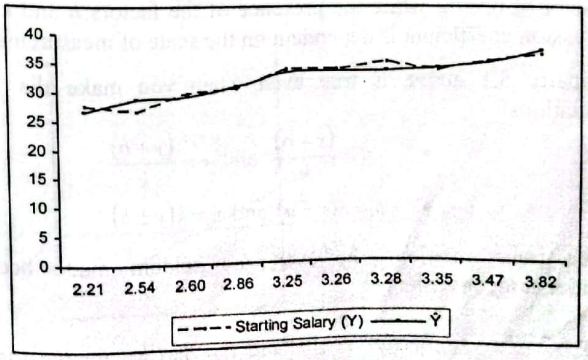

Figure 5.5: Observed and expected values in Example 5.4

5.6 PROPERTIES OF REGRESSION COEFFICIENT

***Property* 5.1:** *Regression coefficient is independent of origin but dependent on the scale of measurement.*

Proof: For n pairs of values $(x_1, y_1), (x_2, y_2)\ldots(x_n, y_n)$ of the variables x and y, the regression coefficient of y on x, b as defined earlier is

$$b_{y|x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \ldots(*)$$

Now let us change the variable x to u and y to v where

$$u = \frac{x-a}{h} \text{ and } v = \frac{y-b}{k} \qquad [h>0, k>0]$$

so that for the ith pair of the variables

$$x_i = a + hu_i \text{ and } y_i = b + kv_i$$

The corresponding mean values are

$$\bar{x} = a + h\bar{u} \text{ and } \bar{y} = b + k\bar{v}$$

Setting these values in (*)

$$b_{y|x} = \frac{hk\sum(u_i - \bar{u})(v_i - \bar{v})}{h^2\sum(u_i - \bar{u})^2} = \left(\frac{k}{h}\right) b_{v|u}$$

The absence of the factors a and b proves that regression coefficient is independent of origin, while the presence of the factors h and k confirms that regression coefficient is dependent on the scale of measurement.

The property 5.1 above is true even when you make the following transformations:

- $u = \dfrac{(x+a)}{h}$ and $v = \dfrac{(y+b)}{k}$
- $u = h(x \pm a)$ and $v = k(y \pm b)$

But such transformations, however, are seldom made because of computational inconvenience.

***Property* 5.2**: *Two regression coefficients $b_{y|x}$ and $b_{x|y}$ for the same set of data cannot simultaneously exceed 1. This means that if $b_{y|x} \geq 1$, then $b_{x|y}$*

must be less than or equal to 1 and vice versa such that the product of $b_{y|x}$ and $b_{x|y}$ is less than unity, i.e. $b_{y|x} \times b_{x|y} < 1$. The proof of this property is left until we study correlation in a latter section.

***Property* 5.3**: *The regression coefficient lies between $-\infty$ to $+\infty$. In other words,*

$$-\infty \leq b_{y|x} \leq \infty$$

This property is obvious and hence needs no proof.

***Property* 5.4:** *Two regression coefficients $b_{Y|X}$ and $b_{X|Y}$ for the same set of data cannot have opposite signs. This means that if $b_{Y|X}$ is positive, $b_{X|Y}$ is also positive and if $b_{Y|X}$ is negative, $b_{X|Y}$ is also negative. The proof of this property will be obvious from the section on correlation that follows.*

5.7 PARTITIONING THE TOTAL VARIATION IN REGRESSION

Although the regression line is a useful summary of the relationship between two variables, the values of the slope and the intercept alone do little to indicate how well the line actually fits the data. A goodness of fit index is needed to come to a conclusion. In this section, we introduce such a statistical index for measuring or describing how good the estimated regression is. To begin with, we need to decompose the total variation into its meaningful components: explained variation and unexplained variation.

The difference between the individual y_i values and $\bar{y}$, when no regression line is fitted, is called the **total variation**. As indicated in Figure 5.6 below, the total deviation $y_i - \bar{y}$ can be partitioned into the **explained variation** $\hat{y}_i - \bar{y}$ and the **unexplained variation** $y_i - \hat{y}_i$. That is

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \qquad \ldots(5.12)$$

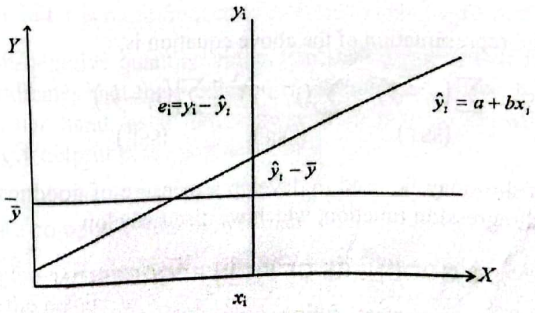

Figure 5.6: Partitioning of total variation $(y_i - \bar{y})$

The deviation $\hat{y}_i - \bar{y}$ is called **explained** because it is regarded as the amount of error that is removed by fitting the regression line to the data. The resulting sum of squares is commonly referred to as the **regression sum of squares** or **sum of squares due to regression** abbreviated SSR. That is

$$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2 \quad \ldots (5.13)$$

The deviation $y_i - \hat{y}_i$ is called **unexplained** because it is the amount of error that still remains after the regression line has been fitted. Note that the differences between y_i and $\hat{y}_i$ actually represent the error in using $\hat{y}_i$ as the estimate of y_i. Thus the resulting sum of squares is referred to as the **error sum of squares** or sum of squares due to error (SSE). That is

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 \quad \ldots (5.14)$$

The sum of squared deviations of the actual values about the mean $\bar{y}$ before the regression analysis is $\sum(y_i - \bar{y})^2$. This value is commonly known as the **total** or **corrected sum** of squares (SST) about the mean. That is

$$\text{SST} = \sum(y_i - \bar{y})^2 \quad \ldots (5.15)$$

The relations among SSE, SST and SSR form the basis of one of the most significant results in applied statistics. In general, this result states that the total sum of squares of the observations about their mean (SST) can be partitioned into two components: SSE and SSR. That is

$$\text{SST} = \text{SSR} + \text{SSE} \quad \ldots (5.16)$$

The symbolic representation of the above equation is,

$$\underset{(\text{SST})}{\sum(y_i - \bar{y})^2} = \underset{(\text{SSR})}{\sum(\hat{y}_i - \bar{y})^2} + \underset{(\text{SSE})}{\sum(y_i - \hat{y}_i)^2} \quad \ldots (5.17)$$

This relationship may be used to develop a measure of goodness of fit of an estimated regression function, which we discuss below.

5.8 GOODNESS OF FIT IN REGRESSION

We would have a perfect fitting estimated regression line if every observation happened to lie on a straight line. In such a case, our least-

squares estimated regression line would pass through each point and thus SSE = 0. For a perfect fit, then SST = SSR and hence SSR/SST = 1. On the other hand, a poorer fit to the observed data results in a larger SSE and hence worst fit. Since SST = SSE + SSR, the worst fit would occur when SSE = SST implying that SSR=0. If this is the case, the estimated regression line does not help to predict y.

If we want to use the ratio SSR/SST to evaluate how good the estimated regression line is, we would have a measure that would take on values between 0 and 1. Values of this ratio closer to 1 would imply a better fitting estimated regression line. The ratio SSR/SST is commonly known as the **coefficient of determination** and is denoted by r^2.

Thus

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$$= \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \quad \ldots (5.18)$$

The quantity r^2 measures the proportion or percentage of the total variation in the dependent variable explained by the regression model. More precisely, it is a summary measure that tells us how well the sample regression line fits the observed data. In a two-variate case, it is the square of the simple correlation coefficient that we study in section 5.8 that follows. In the regression context, r^2 is a more meaningful measure than r though the latter is more frequently referred to than the former.

r^2 is a non-negative quantity and its limits are $0 \leq r^2 \leq 1$. If it is closed to zero, it indicates that the prediction is not much improved by knowing x. On the other hand, as it moves away from 0 to 1, knowing x will be increasingly helpful in the prediction of y.

We now illustrate below how the different components of total sum of squares are computed from sample data.

Example 5.5 Compute SST, SSR, SSE and r^2 for data in Example 5.2 and interpret the result

Solution: The accompanying table shows the required computations.

Sl.	x_i	y_i	$\hat{y}_i$	$(y_i-\bar{y})^2$	$(\hat{y}_i-\bar{y})^2$	$(y_i-\hat{y}_i)^2$
1	1	80	84	784	–24	16
2	3	97	92	121	–16	25
3	4	92	96	256	–12	16
4	4	102	96	36	–12	36
5	6	103	104	25	–4	1
6	8	111	112	9	4	1
7	10	119	120	121	12	1
8	10	123	120	225	12	9
9	11	117	124	81	16	49
10	13	136	132	784	24	16
Total	70	1080	–	2442	2272	170

$$\bar{y}=\frac{1080}{10}=108$$

From the tabular values, the SST, SSR and SSE can now be computed.

$$\text{SST}=\sum(y_i-\bar{y})^2=2442\,,\ \text{SSR}=\sum(\hat{y}_i-\bar{y})^2=2272.\ \text{and}$$

$$\text{SSE}=\sum(y_i-\hat{y}_i)^2=170$$

It is easy to verify that SSE and SSR make up the total sum of squares (SST). Knowing SST and SSR, we can also obtain SSE as the difference of SST and SSR.

There are a number of ways to compute SSR. One such formula, which we derive later, is as follows:

$$\text{SSR}=b^2\sum(x_i-\bar{x})^2 \qquad \text{... (a)}$$

which has an alternative form too:

$$\text{SSR}=b\sum(x_i-\bar{x})(y_i-\bar{y}) \qquad \text{... (b)}$$

We can numerically verify that both these formula yield the same result. In the given problem, b=4 and

$$\sum(x_i-\bar{x})^2=\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}=(1^2+3^2+\ldots+13^2)-\frac{70^2}{10}=632-490=142$$

so that

$$\text{SSR}=b^2\sum(x_i-\bar{x})^2=16\,(142)=2272\,,\ \text{as before}$$

To numerically verify (b), we compute the sum of product of x and y:

$$\sum x_iy_i=(1\times80)+(3\times97)+\ldots+(13\times136)=8128$$

This gives

$$\text{SSR}=b\sum(x_i-\bar{x})(y_i-\bar{y})$$

$$=b\left\{\sum x_iy_i-\frac{\sum x_i\sum y_i}{n}\right\}$$

$$=4\left\{8128-\frac{70\times1080}{10}\right\}$$

$$=2272$$

as ought to be

We can now compute r^2 as follows:

$$r^2=\frac{\text{SSR}}{\text{SST}}=\frac{2272}{2442}=0.93$$

The r^2 value implies that 93% of the variations in annual sales volume are explained by the variations in the experience of the sales persons. Since r^2 can at most be 1, the observed r^2 suggests that the sample regression line fits the data reasonably well.

5.8.1 Standard Error of the Estimate

The standard error of the estimate is a measure that indicates how precise the prediction of y is based on x or, conversely, how inaccurate the prediction might be. The standard error of the estimate is the same concept as the standard deviation we discussed earlier in chapter 4. The standard deviation measures the dispersion about an average, while the standard error of the estimate measures the dispersion about an average line, the regression line.

The standard error of the estimate s_e is computed as follows:

$$s_e=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}} \qquad \text{... (5.19)}$$

The following example illustrates how the standard error of the estimate can be computed for a given data set.

Example 5.6: In a nutrition study, a sample of 10 children under 5 years of age was weighed and their daily family incomes in '000 US$ were recorded. The results are shown in the accompanying table. Calculate the standard error of the estimate.

Family	Family income	Weight in kg
1	13	15
2	20	19
3	34	21
4	24	16
5	16	12
6	30	16
7	36	18
8	11	18
9	8	13
10	27	20
Total	219	168

Solution: By definition, the standard error of the estimate is

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

The computational steps are shown in the accompanying table:

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	y_i^2	$x_i y_i$	$(y_i - \bar{y}_i)^2$
13	15	15.2	.04	225	195	3.24
20	19	16.4	6.76	361	380	4.84
34	21	19.0	4.00	441	714	17.64
24	16	17.2	1.44	256	384	0.64
16	12	15.7	13.69	144	192	23.04
30	16	18.2	4.84	256	480	0.64
36	18	19.3	1.69	324	648	1.44
11	18	14.8	10.24	324	198	1.44
8	13	14.3	1.69	169	104	14.44
27	20	17.7	5.29	400	540	10.24
219	168	–	49.68	2900	3835	76.96

The standard error of the estimate is thus

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{49.68}{8}} = 2.50$$

In most cases, the standard error is computed using an alternative formula

$$s_e = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i y_i}{n-2}}$$

To make use of this formula, we need an additional column containing the product of x and y. This column is shown in the above table. With the estimated values a=3.58 and b=.40, the alternative formula yields

$$s_e = \sqrt{\frac{2900 - 12.88(168) - .1789(3835)}{8}} = 2.50$$

as ought to be.

5.8.2 Relationship between r^2 and Standard Error of Estimate

To demonstrate the aforesaid relationship, we recall that total variation in the regression set up can be partitioned as follows:

$$\text{SST} = \text{SSR} + \text{SSE}$$

The coefficient of determination, r^2 and the standard error of the estimate can now be shown to be related as follows:

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{(n-2)s_e^2}{\text{SST}}$$

To compute r^2 from the nutrition data, we incorporate an additional column for SST in the above table so that

$$r^2 = 1 - \frac{(10-2)(2.50)^2}{76.96} = 0.350$$

5.8.3 Some Important Theorems on Regression

***Theorem* 5.1:** *Show that* $\text{SSE} = S_{yy} - bS_{xy}$ *where* S_{yy} and S_{xy} *are the sum of squares of y and sum of product of x and y respectively.*

Proof: By definition

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Since $\hat{y}_i = a + bx_i$

$$\text{SSE} = \sum (y_i - a - bx_i)^2 = \sum \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2$$
$$= \sum (y_i - \bar{y})^2 + b^2 \sum (x_i - \bar{x})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum (y_i - \bar{y})^2 + b\sum (x_i - \bar{x})(y_i - \bar{y}) - 2b\sum (x_i - \bar{x})(y_i - \bar{y})$$
$$= \sum (y_i - \bar{y})^2 - b\sum (x_i - \bar{x})(y_i - \bar{y})$$
$$= S_{yy} - bS_{xy} \text{ (Proved)}$$

***Theorem* 5.2:** *Prove that* $\text{SSE} = \sum y_i^2 - a\sum y_i - b\sum x_i y_i$

We have proved earlier in Theorem 5.1 that

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

Since $$\sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

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

$$\text{SSE} = \sum y_i^2 - n\bar{y}^2 - b\sum x_i y_i + bn\bar{x}\,\bar{y}$$
$$= \sum y_i^2 - n\bar{y}(\bar{y} - b\bar{x}) - b\sum x_i y_i$$
$$= \sum y_i^2 - an\bar{y} - b\sum x_i y_i$$
$$= \sum y_i^2 - a\sum y_i - b\sum x_i y_i \text{ (Proved)}$$

This result leads to a convenient form of the formula for standard error of the estimate $\hat{y}_i$:

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i y_i}{n-2}}$$

A short-cut proof of the result is as follows:

Let the residual be denoted by e_i so that $e_i = y_i - \hat{y}_i$. Then

$$\sum e_i^2 = \sum e_i e_i = \sum e_i (y_i - \hat{y}_i)$$
$$= \sum y_i e_i - \sum e_i \hat{y}_i$$

Since $\hat{y}_i = a + bx_i$, we have

$$\sum e_i^2 = \sum y_i e_i - a\sum e_i - b\sum x_i e_i$$

By assumption $\sum e_i = 0$ and e_i and x_i are uncorrelated so that $\sum x_i e_i = 0$.

Hence

$$\sum e_i^2 = \sum y_i e_i = \sum y_i (y_i - \hat{y}_i)$$
$$= \sum y_i^2 - \sum y_i (a + bx_i)$$
$$= \sum y_i^2 - a\sum y_i - b\sum x_i y_i$$

This leads to the desired result.

***Theorem* 5.3**: *Show that sum of squares of regression (SSR) and sum squares for error (SSE) add to total sum of squares (SST)*

Proof: By definition

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \text{ and } \text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Adding these two quantities

$$\text{SSR} + \text{SSE} = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

When expanded, the expressions in the right hand side become

$$\sum (\hat{y}_i - \bar{y})^2 = \sum (a + bx_i - a - b\bar{x})^2 = b^2 \sum (x_i - \bar{x})^2 \quad \ldots \text{(a)}$$

and

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$
$$= \sum \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2$$
$$= \sum (y_i - \bar{y})^2 + b^2 \sum (x_i - \bar{x})^2 - 2b\sum (x_i - \bar{x})(y_i - \bar{y}) \ldots \text{(b)}$$

Since $b = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_i - \bar{x})^2$, the last term simplifies to $-2b^2 \sum (x_i - \bar{x})^2$.

Hence adding (a) and (b),

$$\text{SSR} + \text{SSE} = \sum (y_i - \bar{y})^2 = \text{SST}$$

***Theorem* 5.4**: *Show that* $SSR = b^2 S_{xx}$ or bS_{xy}

Proof: We have by definition

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$
$$= \sum (a + bx_i - a - b\bar{x})^2$$
$$= b^2 \sum (x_i - \bar{x})^2 = b^2 S_{xx}$$

Again

$$b^2 S_{xx} = b(bS_{xx})$$
$$= b\left[\left(\frac{S_{xy}}{S_{xx}}\right)S_{xx}\right] = bS_{xy}$$

This completes the proof.

5.9 CORRELATION ANALYSIS

The second major part of bivariate analysis is the problem of correlation or relatedness of variables. When variables are found to be related, we often want to know how close the relationship is. For example, we may be interested in measuring the relationship between the

a) Amount of fertilizer used and wheat production
b) Ages of husbands and their wives
c) Volume of sales and years of experience of sales persons
d) Heights and weights of a group of people
e) Income earned and income saved.

The study of this relationship is accomplished through what is referred to as the **correlation analysis**.

Correlation analysis is intimately related but conceptually very much different from regression analysis. The primary objective of correlation analysis is to measure the strength or degree of linear association between two or more variables. In regression analysis, however, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of the other variables.

In addition to the differences indicated above, the techniques of regression and correlation have some more fundamental differences. In the regression set up of the type $\mu_{y|x} = \alpha + \beta x$, x is not a random variable, since its values are fixed or pre-assigned; while the dependent variable y is a random variable because the observation is randomly selected from the probability distribution on the condition that x has occurred. In contrast both x and x in the correlation analysis are random variables.

The foregoing discussions lead to make the following distinctions between correlation analysis and regression analysis:

1. In correlation analysis, we are primarily interested in the measurement of the strength or degree of linear relationship between two or more variables. The regression analysis, on the other hand, does not assess such relationship.
2. Correlation analysis provides a means of measuring the goodness of fit of the estimated regression line to the observed data. The regression analysis, on the other hand, does not provide any means to measure the goodness of fit; rather it tells us the average amount of change in the dependent variable to a unit change in the independent variable.
3. In regression analysis, there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variable here is stochastic or random variable, while the explanatory variable is fixed. In correlation analysis, on the other hand, we consider any two variables symmetrically. This means that you can interchange between the dependent variable and explanatory variable. This distinction makes the correlation coefficient between x and y the same as that between y and x

The correlation analysis, which we shall undertake in this chapter, involves two variables. The associated quantity, that measures the strength of linear association between these two variables, will be referred to as the **correlation coefficient**. We will denote this measure by the small letter 'r'. Here is a working definition of correlation coefficient:

Definition **5.2**: *Correlation coefficient r is a statistical measure that quantifies the linear relationship between a pair of variables.*

We assume that the measurement of this coefficient is based on the sample values, so that r denotes **sample** correlation coefficient. The corresponding population correlation is usually denoted by the Greek letter ρ.

5.10 MEASURING THE CORRELATION

For n pairs of sample observations (x_1, y_1), (x_2, y_2),...,(x_n, y_n), the correlation coefficient r can be computed employing the following formula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad \ldots (5.20a)$$

Writing in full

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad \ldots (5.20b)$$

For computational purposes, either of the following two formulae for r may be used

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \quad \cdots (5.20c)$$

or

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}} \quad \cdots (5.21)$$

As we will see later, r can also be obtained just by extracting the square root of the coefficient of determination:

$$r = \pm\sqrt{r^2} \quad \cdots (5.22)$$

5.10.1 Interpretation of r

Because of the ways in which it is defined, values of the correlation coefficient always lie between −1 and +1. The absolute value of r indicates the strength of linear relationship. As the reliability of the estimate of r largely depends upon the closeness of the relationship, it is imperative that utmost care be taken while interpreting the value of the coefficient, otherwise fallacious conclusion may be drawn. However, the following general rules would help in interpreting the value of r:

1. A value of +1 indicates that x and y are perfectly related in a positive linear sense. In this case, all the points in a scatter diagram lie on a straight line that has a positive slope (Fig. 5.6a).
2. A value of −1 for r indicates that x and y are perfectly related in a negative linear sense. That is, all the points lie on a straight line that has a negative slope (Fig. 5.6b).
3. Values of r lying between −1 and +1 indicate varying degrees of linear association as is evident from Fig. 6.c through Fig. 6.6h below.
   (a) Data sets exhibiting no linearity produce r=0. (Fig. 6g).
   (b) Values of r close to 1 indicate a strong linear relationship with positive slope (Fig. 5.6c).
   (c) Positive values of r close to 0 indicate a weak linear association with positive slope (Fig. 5.6e).
   (d) Values of r close to −1 indicate a strong linear relationship with negative slope (Fig. 5.6d) and negative values close to 0 indicate a weak linear relationship with negative slope (Fig.5.6f).

(e) For curvilinear relationship between the variables, r tends to zero (Fig. 6.6h).

A common mistake in interpreting r is to assume that correlation implies causation. No such conclusion is automatic. As Kendall and Stuart narrate:

> A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other.

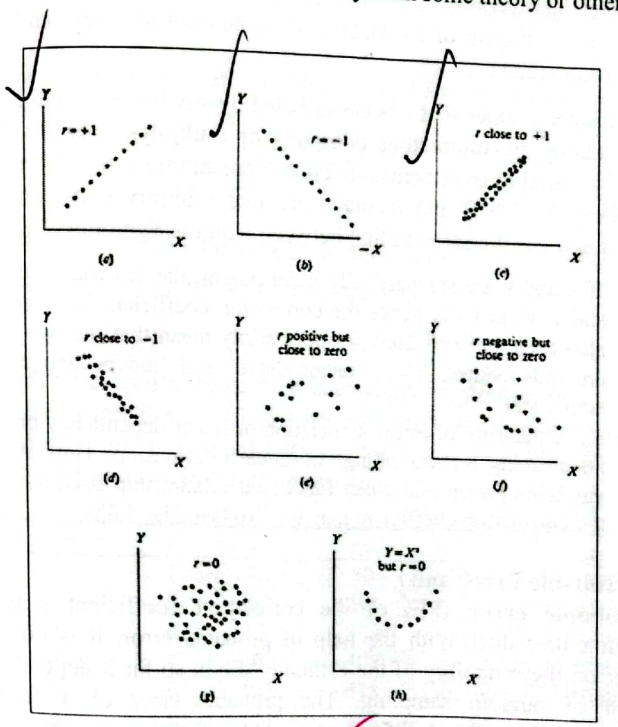

Figure 5.7: Scatter diagrams with varying degrees of r

5.10.2 Some Properties of r

The coefficient of correlation has some appealing properties. These appear below:

(a) The correlation coefficient is a symmetric measure.

This means that interchanging the two variables x and y in the formula does not change the results. Thus, if the correlation coefficient between x

and y is denoted by r_{xy} and that between y and x by r_{yx}, then this symmetric property states that $r_{xy} = r_{yx}$.

(b) The correlation coefficient will be negative or positive depending on whether the sign of the numerator of the formula (5.20b) is positive or negative.

(c) The correlation coefficient lies between – 1 and +1.

(d) The correlation coefficient is a dimensionless quantity, implying that it is not expressed in any units of measurement.

(e) The coefficient of correlation is independent of origin and scale of measurement.

The last property states that r is not affected by any linear transformations, such as adding or subtracting constants or multiplying or dividing all values of a variable by a constant. Thus if we define $u = a + bx$ and $v = c+dy$, where $b, d > 0$ and a and c are two arbitrary constants, then r between x and y is the same as that between u and v. Symbolically, $r_{xy}=r_{uv}$.

(f) If x and y are stochastically independent, the covariance between x and y is zero and hence the correlation coefficient between them is also zero, but r=0 does not necessarily mean that the two variables are independent. Thus, 'uncorrelated' and 'independence' are not equivalent

(g) r is a measure of linear association or linear dependency only; it has no meaning for describing non-linear relationship. Thus even if the variables possess an exact functional relationship such as $y=x^2$, yet the correlation coefficient may be zero (see Fig. 5.6h).

5.10.3 Probable Error and *r*

The **probable error** (PE) of the correlation coefficient r helps in interpreting its value. With the help of probable error, it is possible to comment on the reliability of the estimate of r in so far it depends on the condition of random sampling. The probable error of a correlation coefficient is obtained as follows:

$$PE = 0.6745\left(\frac{1-r^2}{\sqrt{n}}\right) \quad \ldots (5.23)$$

where r is the coefficient of correlation and n is the number of pairs of values on which the value of r is based. The second factor in the right hand side of PE is the standard error of r. The reason for taking the factor 0.6745 is that in a normal distribution, the range $\mu \pm 0.6745\sigma$ covers 50 percent of the total area.

The empirical rules for interpreting r are as follows:

- If the value of r is less than the probable error, there is no evidence of correlation between the variables.
- If the value of r is more than six times the probable error, the existence of correlation is practically certain.
- An approximate interval, within which the value of the coefficient in the population is expected to lie, can be constructed if the probable error is known. Thus if ρ stands for the correlation coefficient in the population, then $r - PE \le \rho \le r + PE$.

Thus with a sample estimate of r=0.6 for n=64 pairs of observations, the probable error is

$$PE = 0.6745\left(\frac{1-.6^2}{\sqrt{64}}\right) = 0.054$$

Hence the limits within which the population correlation (ρ) coefficient is expected to lie, are 0.6 ± 0.054. Or in other words

$$0.546 \le \rho \le 0.650$$

Example 5.7: Let us use the data on sales volume presented in Table-5.1 to compute the correlation coefficient between the years of experience of the salespersons (x) and the annual sales volume (y). The calculations required to compute r are shown in the accompanying table:

Salesperson	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	1	80	80	1	6400
2	3	97	291	9	9409
3	4	92	368	16	8464
4	4	102	408	16	10404
5	6	103	618	36	10609
6	8	111	888	64	12321
7	10	119	1190	100	14161
8	10	123	1230	100	15129
9	11	117	1287	121	13689
10	13	136	1768	169	18496
Total	70	1080	8128	632	119082
	Σx_i	Σy_i	Σx_iy_i	Σx_i^2	Σy_i^2

Employing formula (5.20c) and using the summary values of the above table, we get

$$r = \frac{10(8128) - 70(1080)}{\sqrt{[10(632) - 70^2]}\sqrt{[10(119082) - 1080^2]}} = 0.96$$

If we compare this value with the maximum value of r, which is +1, we conclude that there exists a strong positive correlation between the years of experience of the salespersons and the annual sales volume of the department store. Based on the values of r^2 (=0.92), we assert that 92% of the variations in sales is accounted for by the sales experience of the sales force. The probable error is 0.017 implying an approximate interval for the population correlation coefficient $0.94 \le \rho \le 0.98$.

Example 5.8: The accompanying table shows the proportions of coal miners who exhibit symptoms of pneumoconiosis to their number of years of working in coal mines.

Years	5	10	15	20	25	30	35	40	45	50
Proportions	0	.01	.02	.07	.15	.17	.18	.21	.35	.45

a) Calculate the regression line of proportion with pneumoconiosis (y) on working years (x).
b) Obtain the standard error of the estimate and hence find the coefficient of correlation between x and y.
c) Calculate the correlation coefficient directly from the formula and compare the same with the one obtained in (b)
d) Use your fitted regression line to estimate the proportion of coal miners developing pneumoconiosis who have worked for 42 years and 52 years.

Solution: The accompanying table shows the necessary computations for arriving at the estimates.

x_1	y_i	x_i^2	y_i^2	$x_1 y_i$
05	0	25	.00	.00
10	0.01	100	.00	.10
15	0.02	225	.00	.30
20	0.07	400	.00	1.40
25	0.15	625	.02	3.75
30	0.17	900	.03	5.10
35	0.18	1225	.03	6.30
40	0.21	1600	.04	8.40
45	0.35	2025	.12	15.75
50	0.45	2500	.20	22.50

From the table we have,

$$\sum x_i = 275, \sum y_i = 1.61, \sum x_i^2 = 9625, \sum y_i^2 = 0.46, \sum x_i y_i = 63.60$$

so that

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{63.60 - \frac{275 \times 1.61}{10}}{9625 - \frac{(275)^2}{10}} = 0.00937$$

$$a = \bar{y} - b\bar{x} = 0.161 - .00937(27.5) = -.09667$$

The fitted regression line is thus

$$\hat{y} = a + bx = -.09667 + .00937x$$

Based on this regression line the estimated proportions for 42 and 52 years are

$$\hat{y}_{42} = -.09667 + .00937(42) = 0.29687$$

and

$$\hat{y}_{52} = -.09667 + .00937(52) = 0.39057$$

The standard error of the estimate is obtained as

$$s_e = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i y_i}{n-2}}$$

$$= \sqrt{\frac{.46 + (.09667)(1.61) - (.00937)(63.6)}{8}}$$

$$= 0.04960$$

To calculate r, we compute the total sum of squares (SST) and sum of squares of error (SSE):

$$SST = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 0.46 - \frac{(1.61)^2}{10} = 0.20079$$

$$SSE = \sum y_i^2 - a\sum y_i - b\sum x_i y_i$$

$$= .46 + (.09667)(1.61) - (.00937)(63.6) = 0.01971$$

Hence

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{.01971}{.20079} = 0.90184$$

so that $r = \sqrt{.90184} = 0.94965$

The direct computation yields

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}$$

$$= \frac{63.6 - \frac{275 \times 1.61}{10}}{\sqrt{9625 - \frac{(275)^2}{10}}\sqrt{.46 - \frac{(1.61)^2}{10}}} = 0.94964$$

which agrees quite well with the earlier result.

5.11 RANK CORRELATION

We introduced the concept of correlation in the previous section as a measure of linear association for data that attain at least an interval level of measurement. Furthermore, it was noted that the two variables had a joint normal distribution and that the conditional variance of one variable given the other was the same. In situations where the truth of these assumptions is doubtful, we may use other technique generally known as the **rank correlation** method. Rank correlation method is applied when the rank-order data are available or when each variable can be ranked in some order. The measure based on this method is known as **rank correlation coefficient**. It is, in essence, a **non-parametric** counterpart of the conventional correlation coefficient r. In this text, we will present two methods of computing correlation coefficient based on rank-ordered data, of which one is due to Spearman and the other is due to Kendall.

5.11.1 Spearman Rank Correlation

Spearman rank correlation, also known as the **Spearman's *rho***, is after Karl Spearman, who proposed it in 1904. Spearman ***rho*** is a convenient [illegible] the strength of the monotonic relationship between x and y

without being concerned whether or not the relationship is linear. The relationship between the variables is analyzed according to the ranks of each of the variables. As a result, the method can be applied to situations in which exact numerical measurements are not available. For instance, it may be very difficult for a judge to measure beauty of a group of women reporting for beauty contest or for a manager to measure exactly the sincerity or honesty of each employee working in a factory, but it would be rather a simple job for the judge or the manager to rank each of them on the basis of some pre-determined criteria. If two sets of ordered scores are available for each individual from two judges, in the case of beauty competition and two managers in the case of the factory, then the statistical question involves here in determining r is that whether or not there is any agreement between the rankings of the two judges (or managers) so far as their assessment is concerned. To help us answer this question, we will compute the Spearman **rank-correlation coefficient**, which we denote by r_s.

In order to compute rank correlation coefficient, the data are comprised of ordinal numbers 1^{st}, 2^{nd},...,n^{th}. These are then replaced by cardinal numbers 1, 2,...,n for the purpose of calculation. The substitution of cardinal numbers for ordinal numbers always assumes equality of intervals. The difference between the first and the second is assumed equal to the difference between the second and the third and so on.

In rank ordering, we first rank the x values among themselves, giving rank 1 to the largest value (or the smallest), rank 2 to the second largest (second lowest) value, and so forth, then we similarly rank the y values among themselves and calculate r_s. In sum, the rank correlation method is recommended when

1. The values of the variables are available in rank-ordered form
2. The data are qualitative in nature and can be ranked in some order
3. The data were originally quantitative in nature but because of smallness of the sample size or for convenience in fitting the requirements of analytical techniques, were converted into ranks.

The last point needs clarification. Suppose that the age of a group of individuals were recorded in years as 0, 1,...,90. One may categorize these individuals as being children who are under 9 and assign a rank 1, between 10 and 19 as adolescents and a rank 2, between 20 and 64 as adults and a rank 3 and finally above age 64 as old and assign a rank 4. This scheme of ordering is shown below:

Category	Age	Rank
Children	Under 10	1
Adolescent	10–19	2
Adult	20-64	3
Old	65 and over	4

The height of the same individuals, for example, may be assigned ranks of different order in a similar manner and a rank correlation may be calculated between the ranks of the age and height. Such ranks, however, lack much of the information contained in the original data and as such the correlation coefficient obtained from these ranks is not expected to be the same as would have been obtained from original data before categorization.

5.11.2 Computing Rank Correlation

The Spearman rank correlation coefficient r_s is just the ordinary sample correlation coefficient r applied to the rank order data. That is, we could compute r_s by treating the ranks as the values of the variables and then using the formula developed for simple correlation coefficient. The method calls for computing the **sum of the squared differences** between each pair of ranks, after each of the two variables to be correlated is arranged in order of ranks. Then, if no *tie* in ranks exists, we can apply the following formula for computing r_s:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \ldots (5.24)$$

where d_i is the difference between ranks of the *ith* pair and n is the number of pairs included. To derive the formula (5.24), let x_i denote the rank of item or individual i according to the first criterion, and y_i, the rank of item or individual according to the second criterion, then $d_i = x_i - y_i$ for any ordered pair (x_i, y_i).

Assume that no two items are awarded the same rank by either criterion. Each of the variables x and y under this condition takes on values 1, 2, 3, ..., n. In deriving the formula for r_s, we will make use of the formula for the simple correlation coefficient r:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{ns_x s_y}$$

from which

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

Since both x and y assume the same values (i.e. 1. 2, 3, ...,n), they have identical means and identical variances, i.e. $\bar{x} = \bar{y}$ and $s_x^2 = s_y^2$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+2+\ldots+n}{n} = \frac{n(n+1)}{2n} = \frac{(n+1)}{2} = \bar{y}$$

and

$$ns_x^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{[n(n+1)]^2}{4n}$$

$$= \frac{n(n^2-1)}{12} = ns_y^2$$

Thus $s_x = s_y$

Since $\bar{x} = \bar{y}$, the difference d_i can be expressed as follows:

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

Squaring and adding,

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

$$= ns_x^2 + ns_y^2 - 2nrs_x s_y = 2ns_x^2(1-r) \qquad [\text{since } s_x = s_y]$$

$$= \frac{2n(n^2-1)(1-r)}{12} \qquad \left[\text{since } ns_x^2 = \frac{n(n^2-1)}{12}\right]$$

Rearranging the above expression and replacing r by r_s

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \ldots (5.25)$$

***Corollary* 5.1**: *Show that an alternative formula for computing rank correlation coefficient as defined in (5.25) is*

$$r_s = \frac{\sum x_i y_i - a}{\sum x_i^2 - a} \qquad \ldots (5.26)$$

where $a = \frac{n(n+1)^2}{4}$

and further show that

$$r_s = \frac{12\left[\sum x_i y_i - a\right]}{n(n^2-1)} \qquad \ldots (5.27)$$

where $$a = \frac{n(n+1)^2}{4}$$

To prove the corollary, we start with the definition of simple correlation coefficient:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}}$$

Recall that in dealing with ranks, we have $\Sigma x_i = \Sigma y_i$ and $\Sigma x_i^2 = \Sigma y_i^2$, so that

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

$$= \frac{\sum x_i y_i - \frac{n(n+1)^2}{4}}{\sum x_i^2 - \frac{n(n+1)^2}{4}} = \frac{\sum x_i y_i - a}{\sum x_i^2 - a} = r_s$$

This proves the corollary.

To prove (5.26), we recall that

$$\sum x_i^2 = \frac{n(n+1)(2n+1)}{6}$$

Substituting this in (d), and simplifying,

$$r_s = \frac{\sum x_i y_i - a}{\sum x_i^2 - a} = \frac{\sum x_i y_i - \frac{n(n+1)^2}{4}}{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}}$$

$$= \frac{\sum x_i y_i - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}}$$

$$= \frac{12\left[\sum x_i y_i - a\right]}{n(n^2-1)} \qquad \text{[proved]}$$

We caution here that r_s should not be interpreted as a measure of linear association between two variables; it is a measure of linear association between the ranks of the variables. For example, a value of r_s near 1 may indicate that two variables are related in highly non-linear (but monotonically increasing) fashion.

In many instances, it is not recommended to develop ranks from the given quantitative values. It is because of the reason that whenever quantitative values are available, there is no justification of computing rank correlation. Thus only when rank ordered values are available, we directly apply rank correlation formula to compute r_s.

We now illustrate below how the Spearman rank correlation coefficient can be calculated from rank order data.

Example 5.9: Ten students appearing at an examination were evaluated by two independent examiners out of 100 marks. We wish to determine whether the marks given by these examiners are correlated. Table below shows these marks.

Examiner	Marks assigned									
Examiner 1	65	70	76	75	80	78	83	84	85	90
Examiner 2	30	25	35	40	38	42	48	50	55	45

(a) Rank the data and hence compute the rank correlation coefficient;
(b) Use Corollary (5.27) to compute the rank correlation;
(c) Calculate the simple correlation coefficient (r) using original data;
(d) Comment on the results.

Solution: To facilitate the computation, we construct the following table. Further, we use x and y to denote the original marks and u and v to denote

Student	Examiner–1 Mark	Examiner–1 Rank	Examiner–2 Mark	Examiner–2 Rank	d_i	d_i^2
1	65	10	30	9	+1	1
2	70	9	25	10	−1	1
3	76	7	35	8	−1	1
4	75	8	40	6	+2	4
5	80	5	38	7	−2	4
6	78	6	42	5	+1	1
7	83	4	48	3	+1	1
8	84	3	50	2	+1	1
9	85	2	55	1	+1	1
10	90	1	45	4	−3	9

Solution: (a) Computation of rank correlation.
From the table $\sum d_i^2 = 24$ and $n=10$, so that

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(24)}{10(100-1)} = 0.85$$

(b) To make use of the Corollary (5.27) to calculate the rank correlations, we calculate the sum of product of u and v and a:

$$\sum u_i v_i = 373, \quad a = \frac{n(n+1)^2}{4} = 302.5$$

Hence

$$r = \frac{12\left[\sum u_i v_i - a\right]}{n(n^2-1)} = \frac{12(373-302.5)}{10(100-1)}$$

$$= 0.85 \text{ as before}$$

Use of (5.26) gives

$$r = = \frac{\sum u_i v_i - a}{\sum u_i^2 - a} = \frac{373-302.5}{385-302.5} = 0.85$$

which exactly agrees with our previous results.

(c) Employing the original data,

$$\sum x_i = 786, \quad \sum y_i = 408,$$

$$\sum x_i y_i = 32585, \quad \sum x_i^2 = 62280, \quad \sum y_i^2 = 17412.$$

so that

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]}\sqrt{\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]}}$$

$$= \frac{32585 - \frac{786\times 408}{10}}{\sqrt{\left(62280 - \frac{786^2}{10}\right)}\sqrt{\left(17412 - \frac{408^2}{10}\right)}} = 0.83.$$

The result implies that the rank correlation coefficient is unlikely to be equal to the simple correlation simply because of the fact that two sets of data are different in magnitude as well as well as in nature, the former being based on the ordinal data, while the later being based on the ratio level of data.

Example 5.10: Two interviewers E_1 and E_2 ranked 10 students independently who appeared in an oral examination for a position in a company. Their ranks appear in the accompanying table.

E_1	E_2	d_i	d_i^2
3	6	−3	9
5	4	+1	1
8	9	−1	1
4	8	−4	16
7	1	+6	36
10	2	+8	64
2	3	−1	1
1	10	−9	81
6	5	+1	1
9	7	+2	4
Total	–	–	214

Compute the rank correlation coefficient and comment.

Solution: Here $\sum d_i^2 = 214$, $n=10$. Hence

$$r_s = 1 - \frac{6(214)}{10(10^2-1)} = -0.297$$

The negative value of the correlation coefficient indicates that the two interviewers differ in their opinion in ranking the candidates.

5.11.3 Computing r_s for Repeated Ranks

The use of the formula (5.25) assumes that no two observations would have the identical or equal ranks. If, however, there are ties (i.e. when two observations of the same variable are identical), some adjustment in the formula (5.25) is needed to compute r_s. In such cases, we assign to each of the tied observations, the **mean** of the ranks which they jointly occupy. Thus if the third and the fourth largest values are identical, we assign each the rank (3+4)/2=3.5, and if the fifth, sixth and the seventh largest values are identical, we assign each the rank (5+6+7)/3=6.

To illustrate the technique, let us consider the data in the first two columns of the following table, which pertain to the number of hours which ten students studied for an examination and the scores they received:

Hours studied (x_i)	Scores (y_i)	Rank of x_i	Rank of y_i	x_i^2	y_i^2	x_iy_i
8	56	6.5	7	42.25	49	45.5
5	44	8.5	9	72.25	81	76.5
11	79	4	3	16	9	12
13	72	3	4	9	16	12
10	70	5	5	25	25	25
5	54	8.5	8	72.25	64	68
18	94	1	1	1	1	1
15	85	2	2	4	4	4
2	33	10	10	100	100	100
8	65	6.5	6	42.25	36	39
Total	-	-	-	384	385	383

An alternative formula is then used to compute the rank correlation coefficient incorporating the above adjustment for tie ranks. The adjustment consists in adding a factor $m(m-1)/12$ to $\sum d_i^2$, where m is the number of times an item has been repeated. If there is more than one such group of items that has been repeated, this factor is added as many times as the number of such groups. The formula thus can be written as follows:

$$r_s = 1 - \frac{6\left\{\sum d_i^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \ldots\right\}}{n(n^2-1)} \quad \ldots (5.28)$$

In the above example, note that each of the values 6.5 and 8.5 has been repeated twice. Thus the factor to be added to $\sum d_i^2$ is

$$\frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) = \frac{2}{12}m(m^2-1) \quad \ldots (5.29)$$

You can verify that $\sum d_i^2 = 3$. Hence with m=2 and $\sum d_i^2 = 3$, we find that

$$r_s = 1 - \frac{6\left\{\sum d_i^2 + \frac{2}{12}m(m^2-1)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6(3+1)}{10(10^2-1)} = 0.97$$

5.11.4 Properties of r_s

Like simple correlation, the Spearman correlation coefficient also ranges from −1 to +1, with an interpretation similar to that for the sample correlation coefficient r. It is a measure of monotonicity of a relationship. Spearman's r_s is considered to be a measure of the increasing or decreasing relationship between two variables, while the correlation coefficient r measures the strength of the linear relationship between two variables.

When the ranks of x completely agree with the ranks of y, i.e. (x, y)=(1, 1) (2, 2),…,(n, n), r_s=1. The value of r_s is −1 when there is complete disagreement in the ranks, in which case (x, y) = (1, n) (2, n-1),….,(n, 1). Further, the r_s can be +1 or −1 without r being +1 or −1, but the converse is not true.

We show this feature of the rank correlation below:

Case 1: When n is odd such that n=2m+1, for m=1, 2, 3, ….

Let us construct a difference table of the following form

x_i	y_i	$d_i = x_i - y_i$	$d_i^2 = (x_i - y_i)^2$
1	n	1−n	$4m^2$
2	n−1	3−n	$4(m-1)^2$
3	n−2	5−n	$4(m-2)^2$
…	…	…	…
…	…	…	−2
…	…	…	0
…	…	…	2
…	…	…	…
…	…	…	$4(m-2)^2$
n−2	3	n−5	$4(m-1)^2$
n−1	2	n−3	$4m^2$
n	1	n−1	

The sum of d_i^2 values can now be expressed as follows

$$\sum d_i^2 = 2\{4m^2 + 4(m-1)^2 + 4(m-2)^2 + ... + 4(2)^2 + 4(1)^2\}$$
$$= 8\{m^2 + (m-1)^2 + (m-2)^2 + ... + 2^2 + 1^2\}$$
$$= \frac{8m(m+1)+(2m+1)}{6} \quad \text{... (a)}$$

Now recall that

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{... (b)}$$

Substituting (a) in (b)

$$r_s = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2-1\}}$$
$$= 1 - \frac{8m(m+1)}{4m(m+1)} = -1$$

Hence the proof.

Case I1: When n is even such that $n=2m$, for $m=1, 2, 3, ...$

x_i	y_i	$d_i = x_i - y_i$	$d_i^2 = (x_i - y_i)^2$
1	n	$1-n = -(2m-1)$	$(2m-1)^2$
2	$n-1$	$3-n = -(2m-3)$	$(2m-3)^2$
3	$n-2$	$5-n = -(2m-5)$	$(2m-5)^2$
...	...	...	...
...	...	...	-3
...	...	...	-1
...	...	...	1
...	...	...	...
...	...	...	...
$n-2$	3	$n-5 = (2m-5)$	$(2m-5)^2$
$n-1$	2	$n-3 = (2m-3)$	$(2m-3)^2$
n	1	$n-1 = (2m-1)$	$(2m-1)^2$

The sum of d_i^2 values can be expressed as follows

$$\sum d_i^2 = 2[(2m-1)^2 + (2m-3)^2 + (2m-5)^2 + ... + (3)^2 + (1)^2]$$
$$= 2\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + ... + (2)^2 + (1)^2\}$$
$$-2\{(2m)^2 + (2m-2)^2 + (2m-4)^2 + ... + (4)^2 + (2)^2\}$$

$$= 2\{1^2 + 2^2 + ... + (2m-1)^2 + (2m)^2\}$$
$$-8\{m^2 + (m-1)^2 + (m-2)^2 + ... + (2)^2 + (1)^2\}$$
$$= 2\left\{\frac{2m(2m+1)(4m+1)}{6}\right\} - 8\left\{\frac{m(m+1)(2m+1)}{6}\right\}$$
$$= \frac{2m(4m^2-1)}{3} \text{ (on simplification)} \quad \text{... (c)}$$

Substituting (c) in (b)

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Hence the proof.

We demonstrate below with the help of simple examples the cases when $r_s=1$, $r_s=-1$ and $r_s=0$.

(a) When the ranks are in complete agreement:

Ranks of x	1	2	3	4	5
Ranks of y	1	2	3	4	5
d_i^2	0	0	0	0	0

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(0)}{(5)(24)} = 1$$

(b) When the ranks are in complete disagreement:

Ranks of x	1	2	3	4	5
Ranks of y	5	4	3	2	1
d_i^2	16	4	0	4	16

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(40)}{(5)(24)} = -1$$

(c) When the ranks are in random order:

Ranks of x	1	2	3	4	5
Ranks of y	4	3	2	1	5
d_i^2	9	1	1	9	0

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(20)}{(5)(24)} = 0$$

5.11.5 Advantages of r_s over r

Some advantages in using r_s, rather than r are apparent. For instance,

a) We no longer assume the underlying relationship between x and y to be linear and therefore, when the data possess a curvilinear relationship, the rank correlation coefficient is likely to be more reliable than the conventional correlation coefficient.
b) No assumption of normality is made concerning the distributions of x and y in the case of rank correlation.
c) When no numerical measurements of the variable x and y are possible, rank correlation is the only way to assess the relationship between the two variables.
d) The computation of rank correlation is much simpler than the product moment correlation.

5.11.6 Kendall's Rank Correlation

Kendall's rank correlation coefficient is computed using the similar type of data for which the Spearman rank correlation coefficient is applicable. Like r_s, the Kendall's rank correlation coefficient, denoted by τ, is a measure of the degree of association between two sets of ranks. Suppose that two examiners I and II are to rank 5 students, A, B, C, D, and E on the basis of their performance in an examination and that the following rankings were obtained:

Table I

Examiner	A	B	C	D	E
I	2	3	1	5	4
II	3	1	2	4	5

Table II

Examiner	C	A	B	E	D
I	1	2	3	4	5
II	2	3	1	5	4

The following steps are then involved in computing τ:

1. Rearrange the rankings of the first examiner in natural order. The natural ranks correspond to CABED.
2. Set the rankings of the second examiner corresponding to the order CABED. These are 2, 3, 1, 5, and 4. The resulting table is marked labeled Table II above
3. Start with the first number 2 of step 2 and count the number of

ranks to the right that are greater. There are *three* such ranks: 3, 5, and 4.

4. Subtract from 3 the number of ranks to the right that are smaller. Since there is only one such rank that is 1, the net result is +3 −1=+2.
5. Move to the second number on the right. This is 3. There are two ranks to its right that are greater than 3 (5, 4) and one rank to its right that is smaller (1). The net contribution is therefore +2−1=+1.
6. The next rank is 1. This has no rank that is smaller than this, but two ranks (5, 4) that are greater. The net contribution is 0+2=+2. Finally, the rank 5 has only one rank following it that is smaller. The net contribution is −1+0=−1. The total score is S, where

$$S=+2+1+2-1=+4.$$

The following formula is then used to compute the Kendall's rank correlation coefficient τ:

$$\tau = \frac{2S}{n(n-1)}$$

As of the Pearson's product moment correlation coefficient, the τ ranges from − 1 to + 1. In the above illustration, n=5. Thus we have

$$\tau = \frac{2S}{n(n-1)} = \frac{2\times 4}{5(5-1)} = 0.4$$

5.12 SOME USEFUL RESULTS

a) If $b_{y|x}$ is the regression coefficients of y on x and r is the correlation coefficient between x and y, then

$$r = b_{y|x}\left(\frac{s_x}{s_y}\right)$$

where s_x and s_y are the standard deviations of x and y respectively.

To prove this, recall that

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}}\sqrt{\frac{S_{xx}}{S_{yy}}} = b_{y|x}\left(\frac{s_x}{s_y}\right) \quad \text{... (i)}$$

b) If $b_{x|y}$ is the regression coefficient of x on y, we can show that

$$r = b_{x|y}\left(\frac{s_y}{s_x}\right) \quad \text{... (ii)}$$

Multiplying (i) and (ii), we arrive at the following identity:

$$r^2 = b_{y|x} \times b_{x|y} \quad \text{... (iii)}$$

The implication of the relation (iii) is that

$$r = \pm\sqrt{b_{y|x} \times b_{x|y}} \quad \text{... (iv)}$$

which says that coefficient of correlation is equal to the geometric mean of two regression coefficients whenever two regression coefficients are available for a single data set.

c) We further state that the two regression coefficients $b_{y|x}$ and $b_{x|y}$ (if exist) must of the same sign. The correlation coefficient r bears the same sign as do the regression coefficients. That is if b is positive, r is also positive and if b is negative, r is also negative. Thus if $b_{y|x}$=–.90 and $b_{x|y}$= –.40, then $r = -\sqrt{.90 \times .40} = -0.6$

d) Equation (iv) furthers says that the product of two regression coefficients cannot exceed unity:

$$b_{y|x} \times b_{x|y} \leq 1$$

Thus if one regression coefficient exceeds 1, the other one must be less than 1 so as to make the product of the two regression coefficients less than or equal to 1.

***Theorem* 5.5:** *The coefficient of correlation r, between two variables is independent of origin and scale of measurement.*

Proof: To prove the theorem, we define the correlation coefficient between x and y as follows:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad \text{... (a)}$$

We now transform x and y to u and v as follows:

$$u_i = \frac{x_i - a}{h} \text{ and } v_i = \frac{y_i - b}{k}$$

which yields

$x_i = a + hu_i$ and $y_i = b + kv_i$ so that $\bar{x} = a + h\bar{u}$, and $\bar{y} = b + k\bar{v}$. Substituting these values in (a)

$$r_{xy} = \frac{\sum(a + hu_i - a - h\bar{u})(b + kv_i - b - k\bar{v})}{\sqrt{\sum(a + hu_i - a - h\bar{u})^2 \sum(b + kv_i - b - k\bar{v})^2}}$$

$$= \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum(u_i - \bar{u})^2 \sum(v_i - \bar{v})^2}} = r_{uv} \quad \text{... (b)}$$

which is an expression that is independent of a, b, h and k. Clearly the expression (b) above represents the correlation coefficient between u and v. This implies that transformation of the variables x and y to u and v does not have any effect on the value of r. That is

$$r_{xy} = r_{uv} \quad \text{... (c)}$$

This proves the theorem.

Example 5.11: Given the following values of x and y. Verify theorem 5.2.

x:	5	10	15	20	25
y:	2	6	10	18	26

Solution: From the given data

$$\sum x = 75, \sum y = 62, \sum x^2 = 1375, \sum y^2 = 1140, \sum xy = 1230$$

Hence

$$r_{xy} = \frac{1230 - \frac{75 \times 62}{5}}{\sqrt{\left(1375 - \frac{75^2}{5}\right)\left(1140 - \frac{62^2}{5}\right)}} = 0.9848$$

Now define two new variables u and v with a=15, h=5, b=10 and k=4 as follows

$$u = \frac{x - 15}{5}$$

and

$$v = \frac{y - 10}{4}$$

The transformed values of u and v and the necessary computations in terms of the new origin and scale appear in the following table:

u	v	u^2	v^2	uv
−2	−2	4	4	4
−1	−1	1	1	1
0	0	0	0	0
1	2	1	4	2
2	4	4	16	8

$$\sum u = 0, \sum v = 3, \sum u^2 = 10, \sum v^2 = 25, \sum uv = 15$$

Putting these values in (b)

$$r_{xy} = \frac{15 - 0}{\sqrt{(10 - 0)\left(25 - \frac{9}{5}\right)}} = 0.9848$$

which completely agrees with the value of r_{xy} as obtained before. This verifies the theorem numerically.

***Theorem* 5.6**: *The coefficient of correlation between x and y lies between* -1 *and* $+1$, *i.e.* $-1 \le r \le +1$.

Proof: Let us consider the following expression

$$\left(\frac{x_i - \bar{x}}{s_x} \pm \frac{y_i - \bar{y}}{s_y}\right) \quad \text{... (a)}$$

which, when squared is always positive. In other words

$$\left(\frac{x_i - \bar{x}}{s_x} \pm \frac{y_i - \bar{y}}{s_y}\right)^2 \ge 0 \quad \text{... (b)}$$

Performing the square and summing

$$\frac{\sum (x_i - \bar{x})^2}{s_x^2} + \frac{\sum (y_i - \bar{y})^2}{s_y^2} \pm \frac{2\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \ge 0 \quad \text{... (c)}$$

Now recall that

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = r n s_x s_y, \quad n s_x^2 = \sum (x_i - \bar{x})^2, \quad n s_y^2 = \sum (y_i - \bar{y})^2$$

Substituting these values in (c)

$$\frac{n s_x^2}{s_x^2} + \frac{n s_y^2}{s_y^2} + \frac{2 n r s_x s_y}{s_x s_y} \ge 0$$

from which $1 \pm r \ge 0$

When $1 + r \ge 0$, $r \ge -1$, and when $1 - r \ge 0$, $r \le +1$, so that $-1 \le r \le 1$. This proves the theorem.

The theorem can also be proved as follows:

We have established earlier that S_{ee}, sum of squares due to error (SSE) can be expressed as

$$S_{ee} = S_{yy} - b S_{xy} \quad \text{... (d)}$$

where $S_{xy} = \text{SP}(x, y) = \sum (x_i - \bar{x})(y_i - \bar{y})$ and $S_{yy} = \text{SS}(y) = \sum (y_i - \bar{y})^2$

The right hand side of (e) can be expanded as follows:

$$S_{yy} - b S_{xy} = S_{yy}\left(1 - b\frac{S_{xy}}{S_{yy}}\right) = S_{yy}\left[1 - \left(\frac{S_{xy}}{S_{xx}}\right)\left(\frac{S_{xy}}{S_{yy}}\right)\right]$$

$$= S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx} S_{yy}}\right) = S_{yy}(1 - r^2)$$

That is $S_{ee} = S_{yy}(1 - r^2)$ .. (e)

Since S_{ee} cannot be negative, it follows from (e) that $r^2 \le 1$, or

$$-1 \le r \le 1$$

***Theorem* 5.7**: *For two variables x and y connected by the equation* $y_i = a + bx_i$ *where a and b are two arbitrary constants, show that* $r = +1$ *if* $b > 0$ *and* $r = -1$ *if* $b < 0$.

Proof: Since $y_i = a + bx_i$, we have

$$\bar{y} = a + b\bar{x}, \quad s_y^2 = b^2 s_x^2$$

and hence

$$s_y = |b| s_x .$$

Subtracting $\bar{y}$ from y_i, $y_i - \bar{y} = b(x_i - \bar{x})$.

Let us now define

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum (x_i - \bar{x}) b (x_i - \bar{x})}{n}$$

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

But

$$r = \frac{\text{Cov}(x, y)}{s_x s_y} = \frac{b s_x^2}{s_x |b| s_x} = \frac{b}{|b|}$$

It follows that

$$r = +1 \text{ if } b > 0 \text{ and } r = -1 \text{ if } b < 0.$$

***Theorem* 5.8**: *The coefficient of regression is independent of origin but in general, depends on the scale of measurement.*

Proof: Consider the regression coefficient of y on x as defined below

$$b_{y|x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r\left(\frac{s_y}{s_x}\right)$$

where s_x and s_y are the standard deviations of x and y respectively.

Let u and v be the two variables defined as follows:

$$u = \frac{x - a}{h} \text{ and } v = \frac{y - b}{k} \qquad [h > 0,\ k > 0]$$

Here a and b are the origins and h and k are the scale factors

The transformation above suggests that for the ith values of the variables $x_i = a + hu_i$ and $y_i = b + kv_i$. Consequently

$$s_x^2 = h^2 s_u^2 \text{ and } s_y^2 = k^2 s_v^2, \text{ leading to } s_x = h s_u \text{ and } s_y = k s_v$$

$$b_{y|x} = r\left(\frac{s_y}{s_x}\right) = r\left(\frac{k s_v}{h s_u}\right) = \frac{k}{h} r\left(\frac{s_v}{s_u}\right) = \frac{k}{h} b_{v|u}$$

The presence of the factors k and h shows that regression coefficient is dependent on the scale of measurement but independent of origin (a, b).

Example 5.12: Given the following pairs of values of x and y. Verify Theorem 5.8.

x:	5	10	15	20	25
y:	2	6	10	15	20

Solution: It is easy to verify that

$$\sum x = 75, \sum y = 53, \sum x^2 = 1375, \sum y^2 = 765, \sum xy = 1020$$

and

$$\sum u = 0, \sum v = 0.75, \sum u^2 = 10, \sum v^2 = 12.8125, \sum uv = 11.25$$

The regression coefficient of y on x is

$$b_{y|x} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{1020 - \frac{75 \times 53}{5}}{1375 - \frac{75^2}{5}} = 0.90$$

while

$$b_{v|u} = \frac{\sum u_i v_i - \frac{\sum u_i \sum v_i}{n}}{\sum u_i^2 - \frac{\left(\sum u_i\right)^2}{n}} = \frac{11.25 - \frac{0 \times .75}{5}}{10 - \frac{0^2}{5}} = 1.125$$

Hence

$$\frac{k}{h} b_{v|u} = \frac{4}{5}(1.125) = 0.90 = b_{y|x}$$

This numerically verifies the theorem.

5.13 PROPERTIES OF LEAST SQUARES ESTIMATORS

To use the least squares estimators in statistical inferences, we require to understand their properties. In this section we show that the least squares estimators a and b for the parameters in the simple linear model

$$y = \alpha + \beta x + \varepsilon \qquad \text{... (a)}$$

are unbiased estimators of their respective parameter values. We also derive the variances and covariance of these estimators. In our derivation, we assume as before that ε is a random variable with $E(\varepsilon)=0$ and that $V(y) = V(\varepsilon) = \sigma^2$. Assume that n independent observations are to be made on the model (a) so that, before sampling, we have n independent random variables of the form

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{... (b)}$$

With reference to the model (a), we will show that

(i). The estimators a and b are unbiased

(ii) $V(a) = \dfrac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}\sigma^2$

(iii) $V(b) = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}$

(iv) (iv) $\text{Cov}(a,b) = \dfrac{-\bar{x}}{\sum (x_i - \bar{x})^2}\sigma^2$

Proof: By definition

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

which can be written as

$$b = \frac{\sum (x_i - \bar{x})y_i - \bar{y}\sum (x_i - \bar{x})}{S_{xx}} = \frac{\sum (x_i - \bar{x})y_i}{S_{xx}} \quad \ldots (c)$$

since $\sum (x_i - \bar{x}) = 0$

Taking expectation

$$E(b) = E\left[\frac{\sum (x_i - \bar{x})y_i}{S_{xx}}\right] = \frac{\sum (x_i - \bar{x})E(y_i)}{S_{xx}}$$

$$= \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i)}{S_{xx}}$$

$$= \alpha\frac{\sum (x_i - \bar{x})}{S_{xx}} + \beta\frac{\sum (x_i - \bar{x})x_i}{S_{xx}}$$

$$= 0 + \beta\frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{S_{xx}}$$

$$= \beta\frac{\sum (x_i - \bar{x})^2}{S_{xx}} = \beta\left(\frac{S_{xx}}{S_{xx}}\right) = \beta$$

Thus, b is an unbiased estimator of β.

To prove that $E(a) = \alpha$, we recall that

$$a = \frac{\sum y_i}{n} - b\bar{x}$$

Taking expectation

$$E(a) = \frac{\sum E(y_i)}{n} - \bar{x}E(b)$$

$$= \frac{\sum (\alpha + \beta x_i)}{n} - \beta\bar{x}$$

$$= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

showing that a is an unbiased estimator of α.

To find $V(b)$, we recall that $y_1, y_2, \ldots, y_n$ are all independent with constant variance σ^2. Then by virtue of (c)

$$V(b) = V\left[\frac{\sum (x_i - \bar{x})y_i}{S_{xx}}\right]$$

$$= \left(\frac{1}{S_{xx}}\right)^2 \sum V[(x_i - \bar{x})y_i]$$

$$= \left(\frac{1}{S_{xx}}\right)^2 \sum (x_i - \bar{x})^2 V(y_i) = \left(\frac{1}{S_{xx}}\right)^2 S_{xx}V(y_i)$$

$$= \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \quad \text{[Proved]}$$

To find the variance of a, we recall that

$$b = \frac{\sum (x_i - \bar{x})y_i}{S_{xx}}$$

We now rewrite the above expression for b as follows:

$$b = \sum c_i y_i \text{, where } c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

so that $$\sum c_i = 0, \ \sum c_i^2 = \frac{S_{xx}}{(S_{xx})^2} = \frac{1}{S_{xx}}$$

We now write the expression for a

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

$$= \frac{\sum y_i}{n} - \sum c_i y_i \bar{x}$$

$$= \sum \left(\frac{1}{n} - \bar{x}c_i \right) y_i$$

The variance of a is thus

$$V(a) = \sum \left(\frac{1}{n} - \bar{x}c_i \right)^2 V(y_i)$$

$$= \sum \left(\frac{1}{n} - \bar{x}c_i \right)^2 \sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \sigma^2$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \sigma^2$$

$$= \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2 = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$$

We now turn to find the covariance between a and b as follows:

$$\text{Cov}(a,b) = E[\{ a - E(a)(b - E(b)) \}]$$

$$= E[\{ a - E(a)(b - \beta) \}] \quad \ldots \text{(d)}$$

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

$$E(a) = \bar{y} - \bar{x}\beta,$$

so that

$$a - E(a) = -\bar{x}(b - \beta) \quad \ldots \text{(e)}$$

Substituting (e) in (d)

$$\text{Cov}(a, b) = -\bar{x}E(b - \beta)$$

$$= -\bar{x}E\{b - E(b)\}^2$$

$$= -\bar{x}V(b) = \frac{-\bar{x}}{S_{xx}} \sigma^2$$

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

***Theorem* 5.9**: *The variance of the estimated y values is r^2 times the variance of the observed y values. That is*

$$V(\hat{y}) = r^2 V(y), \text{ that is } s_{\hat{y}}^2 = r^2 s_y^2$$

Proof: By definition

$$S_{\hat{y}\hat{y}} = \sum (\hat{y}_i - \bar{y})^2 = n s_{\hat{y}}^2 \text{ and } S_{yy} = \sum (y_i - \bar{y})^2 = n s_y^2 \quad \ldots (**)$$

Now

$$n s_{\hat{y}}^2 = \sum (\hat{y}_i - \bar{y})^2 = \sum (a + bx_i - \bar{y})^2$$

$$= \sum [\bar{y} + b(x_i - \bar{x}) - \bar{y}]^2 = b^2 \sum (x_i - \bar{x})^2$$

$$= \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} S_{yy} = r^2 S_{yy} = n r^2 s_y^2$$

Hence

$$s_{\hat{y}}^2 = r^2 s_y^2 \text{ (proved)}$$

***Theorem* 5.10:** *The correlation coefficient between the observed y and the estimated y is $|r|$, where r is the correlation coefficient between x and y. That is*

$$r_{y\hat{y}} = r_{xy}$$

Proof: By definition

$$r_{y\hat{y}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy} S_{\hat{y}\hat{y}}}}$$

where

$$S_{yy} = n s_y^2, \; S_{\hat{y}\hat{y}} = n s_{\hat{y}}^2 = n r^2 s_y^2$$

$$S_{y\hat{y}} = \sum (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum (y_i - \bar{y})[\bar{y} + b(x_i - \bar{x}) - \bar{y}]$$

$$= b \sum (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{S_{xy}}{S_{xx}} \right) S_{xy}$$

$$= \frac{(S_{xy})^2}{S_{xx} S_{yy}} (S_{yy}) = r_{xy}^2 S_{yy} = r_{xy}^2 n s_y^2 = n s_{\hat{y}}^2$$

Hence

$$r_{y\hat{y}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy} S_{\hat{y}\hat{y}}}} = \frac{n s_{\hat{y}}^2}{\sqrt{n s_y^2 \, n s_{\hat{y}}^2}} = \frac{s_{\hat{y}}}{s_y} = \frac{|r_{xy}| s_y}{s_y} = |r_{xy}|$$

***Theorem* 5.11:** *The correlation coefficient between x and the estimated y is −1 or +1 depending on whether r (correlation coefficient between x and y) is negative or the sum of product of x and y is positive or negative.*

Proof: The correlation coefficient between x and the estimated y (i.e $\hat{y}$) is defined as

$$r_{x\hat{y}} = \frac{S_{x\hat{y}}}{\sqrt{S_{xx}S_{\hat{y}\hat{y}}}} \quad \ldots (*)$$

By definition

$$S_{xx} = ns_x^2, \; S_{\hat{y}\hat{y}} = nr^2 s_y^2$$

And

$$S_{x\hat{y}} = \sum (x_i - \bar{x})(\hat{y}_i - \bar{y}) = \sum (x_i - \bar{x})\, b(x_i - \bar{x}) = b\, S_{xx} = S_{xy} = nrs_x s_y$$

Now substituting these values in (*)

$$r_{x\hat{y}} = \frac{S_{x\hat{y}}}{\sqrt{S_{xx}S_{\hat{y}\hat{y}}}} = \frac{nrs_x s_y}{\sqrt{ns_x^2\, nr^2 s_y^2}} = \frac{r}{|r|} = \pm 1$$

This proves the theorem.

5.14 MORE PROBLEMS ON CORRELATION AND REGRESSION

Example 5.13: If x_1, x_2, and x_3 are three uncorrelated variables with equal variance s^2, then show that the correlation coefficient between $x_1 + x_2$ and $x_2 + x_3$ is $\frac{1}{2}$

Solution: We are given that $V(x_1) = V(x_2) = V(x_3) = s^2$. Since x_1, x_2 and x_3 are uncorrelated, $\mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_1, x_3) = \mathrm{Cov}(x_2, x_3) = 0$. Let us suppose that

$$u = x_1 + x_2 \text{ and } v = x_2 + x_3, \text{ so that}$$

$$V(u) = V(x_1) + V(x_2) + 2\mathrm{Cov}(x_1, x_2) = s^2 + s^2 + 0 = 2s^2$$

and similarly

$$V(v) = V(x_2) + V(x_3) + 2\mathrm{Cov}(x_2, x_3) = 2s^2$$

The correlation coefficient between the variables u and v is

$$r_{uv} = \frac{\frac{1}{n}\sum (u - \bar{u})(v - \bar{v})}{\sqrt{V(u)V(v)}} \quad \ldots (**)$$

The numerator of the above expression can be written as

$$\frac{1}{n}\sum (u - \bar{u})(v - \bar{v}) = \frac{1}{n}\left[\sum \{(x_1 + x_2) - (\bar{x}_1 + \bar{x}_2)\}\{(x_2 + x_3) - (\bar{x}_2 + \bar{x}_3)\}\right]$$

$$= \frac{\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n} + \frac{\sum (x_2 - \bar{x}_2)^2}{n} + \frac{\sum (x_1 - \bar{x}_1)(x_3 - \bar{x}_3)}{n} + \frac{\sum (x_2 - \bar{x}_2)(x_3 - \bar{x}_3)}{n}$$

$$= \mathrm{Cov}(x_1, x_2) + V(x_2) + \mathrm{Cov}(x_1, x_3) + \mathrm{Cov}(x_2, x_3)$$

$$= 0 + s^2 + 0 + 0 = s^2$$

Thus from (**)

$$r_{uv} = \frac{s^2}{\sqrt{2s^2 \times 2s^2}} = \frac{1}{2} \quad \text{(Proved)}$$

Example 5.14: In the following data set, x represents the dosage of a drug and y represents the corresponding time in hours until a particular response is observed for which the drug is administered.

x	y	x	y	x	y
1.5	1.70	5.0	0.80	8.5	0.50
2.0	1.32	5.5	0.59	9.0	0.27
2.5	1.48	6.0	0.35	9.5	0.30
3.0	1.00	6.5	0.61	10.0	0.18
3.5	1.41	7.0	0.70	10.5	0.15
4.0	1.19	7.5	0.40		
4.5	1.23	8.0	0.33		

(i) Compute the correlation coefficient between response time and dose and interpret your result.

(ii) Obtain an estimate of the regression of y on x

(iii) Using your estimated regression line, predict the average response time for a dosage of 8.25.

Solution: Before proceeding with the calculations, we note that the data can be simplified considerably with some adjustments in the location and scale. Let us transform the variables x and y to u and v as follows:

$$u = \frac{x - 6}{0.5} = 2(x - 6) \text{ and } v = \frac{y - 1}{0.01} = 100(y - 1)$$

The table in the following page incorporates the transformed values and other calculations.

(i) Using the appropriate column totals, we compute the coefficient of

correlation between u and v as

$$r = \frac{\sum u_i v_i - \frac{\sum u_i \sum v_i}{n}}{\sqrt{\left\{\sum u_i^2 - \frac{\left(\sum u_i\right)^2}{n}\right\} \times \left\{\sum v_i^2 - \frac{\left(\sum v_i\right)^2}{n}\right\}}}$$

$$= \frac{-4625 - \frac{0 \times (-445)}{19}}{\sqrt{\left(570 - \frac{0}{19}\right) \times \left(53829 - \frac{(-445)^2}{19}\right)}} = -0.93$$

Since r is independent of origin and scale of measurement, the r value between x and y is the same as between u and v. The value of r being close to -1, there exists a strong negative linear relationship between dosage and response time.

Table outlining the computational steps

x_i	u_i	y_i	v_i	u_i^2	$u_i v_i$	v_i^2
1.5	−9	1.70	70	81	−630	4900
2.0	−8	1.32	32	64	−256	1024
2.5	−7	1.48	48	49	−336	2304
3.0	−6	1.00	0	36	0	0
3.5	−5	1.41	41	25	−205	1681
4.0	−4	1.19	19	16	−76	361
4.5	−3	1.23	23	9	−69	529
5.0	−2	0.80	−20	4	40	400
5.5	−1	0.59	−41	1	41	1681
6.0	0	0.35	−65	0	0	4225
6.5	1	0.61	−39	1	−39	1521
7.0	2	0.70	−30	4	−60	900
7.5	3	0.44	−56	9	−168	3136
8.0	4	0.33	−67	16	−268	4489
8.5	5	0.50	−50	25	−250	2500
9.0	6	0.27	−73	36	−438	5329
9.5	7	0.30	−70	49	−490	4900
10.0	8	0.18	−82	64	−656	6724
10.5	9	0.15	−85	81	−765	7225
Total	0		−445	570	−4625	53829

(ii) To estimate the relationship $v=a+bu$ for the transformed variables, we calculate

$$b = \frac{\sum u_i v_i - \frac{\sum u_i \sum v_i}{n}}{\sum u_i^2 - \frac{\left(\sum u_i\right)^2}{n}} = = \frac{-4625 - \frac{0 \times (-445)}{19}}{570 - \frac{0^2}{19}} = -8.11$$

$$a = \frac{\sum v_i}{n} - b\frac{\sum u_i}{n} = \frac{-445}{19} - (-8.11)(0) = -23.42$$

The estimated regression line is thus

$$v = -23.42 - 8.11u$$

The estimated regression line in terms of x and y is thus

$$100\,(\hat{y}_i - 1) = -23.42 - 8.11 \times 2(x_i - 6)$$

from which we obtain

$$\hat{y}_i = 1.74 - 0.1622x_i$$

(iii) For a given dosage of 8.25, the predicted average response time is

$$\hat{y}_{(18.25)} = 1.74 - (0.1622 \times 8.25) = 0.402$$

Example 5.15: If the variables x and y are uncorrelated, and $v(x)=s_x^2$ and $v(y)=s_y^2$, show that the correlation coefficient between $x+y$ and $x-y$ is

$$\frac{s_x^2 - s_y^2}{s_x^2 + s_y^2}$$

Proof: Let $u = x + y$ and $v = x - y$. Then $\bar{u} = \bar{x} + \bar{y}$ and $\bar{v} = \bar{x} - \bar{y}$

Also

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

$$= \frac{\sum [(x_i + y_i - \bar{x} - \bar{y})]\,[(x_i - y_i - \bar{x} + \bar{y})]}{n}$$

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

Also

$$V(u)=V(x+y)=V(x)+V(y)+2\text{Cov}(x, y)=s_x^2+s_y^2$$

$$V(v)=V(x-y)=V(x)+V(y)-2\text{Cov}(x, y)=s_x^2+s_y^2$$

since $\text{cov}(x,y)=0$

Hence

$$r_{uv}=\frac{\text{Cov}(u, v)}{\sqrt{v(u)}\sqrt{v(v)}}=\frac{s_x^2-s_y^2}{\sqrt{(s_x^2+s_y^2)}\sqrt{(s_x^2+s_y^2)}}=\frac{s_x^2-s_y^2}{s_x^2+s_y^2}\quad\text{(Proved)}$$

5.15 PEARSON'S *r* FROM BIVARIATE FREQUENCY TABLE

When the number of pairs of observation is large, it becomes a laborious job to compute correlation coefficient using the usual method. What is needed in this case is to form a **bivariate frequency table** and then attempt to compute the correlation coefficient from this table making a simple modification in the formula. The table so constructed is also known as the **correlation table.**

The construction of a correlation table is straightforward and follows the same procedure as in the case of a univariate table. In its construction, both the variables are assumed to be measured on an interval scale. Suitable class limits are set up for each of the variates, one along the row and the other along the column. The following is a simple example of a bivariate frequency table where the X variable assumes the values x_1 and x_2 and the Y variable assumes the values y_1 and y_2. The f_{11}, for example, is the joint frequency when X assumes the value x_1 and Y assumes value y_1.

Y ↓ X →	x_1	x_2	Total
y_1	f_{11}	f_{12}	f_{1+}
y_2	f_{21}	f_{22}	f_{2+}
Total	f_{+1}	f_{+2}	n

The correlation coefficient from a table of this type may be calculated using the formula

$$r_{xy}=\frac{\sum fxy-\frac{1}{n}\sum fx\sum fy}{\sqrt{\left\{\sum fx^2-\frac{1}{n}(\sum fx)^2\right\}\left\{\sum fy^2-\frac{1}{n}(\sum fy)^2\right\}}}\qquad(5.29)$$

Since correlation coefficient is independent of origin and scale of measurement, the x and y variates may be suitably changed for ease of calculation.

The following example is designed to illustrate the use of the above formula to compute r from bivariate frequency distribution.

Example 5.16: Two variables x and y have the following bivariate distribution. Compute the coefficient of correlation.

	y values					
x values	7	10	13	16	19	Row total
1.5	3	2	3	7	2	17
3.5	3	2	3	3	1	12
5.5	–	–	1	–	–	1
7.5	1	3	1	–	1	6
9.5	–	–	–	1	–	1
11.5	–	–	2	–	–	2
13.5	–	–	–	–	–	0
15.5	–	–	–	1	–	1
Column total	7	7	10	12	4	40

The table that follows illustrates how the correlation coefficient is calculated. The entries in the table under reference are straightforward to calculate. For example, the first entry in the fx column (i.e.25.5) along the first row against the value x=1.5 is calculated as the product of f and x: 17×1.5=25.5. Likewise, the second value 42 is the product of 12 and 3.5, i.e. 42=12×3.5. Thus $\sum fx$=25.5+42.0+...+13.5=166.0. The entries in fx^2 column are obtained by simply multiplying the fx values by the corresponding values of x. Thus, the entry 38.25 is obtained as the product of 25.5 and 1.5, i.e. 38.25=25.5×1.5.

The fxy values are obtained by multiplying f by y values with a fixed value of x. Thus the first entry (345.0) is obtained as: (3×1.5×7)+(2×1.5×10)+....+(2×1.5×19) = 345.0. The quantities in the rows fy, fy^2 and fyx are similarly obtained.

The correlation coefficient now may be obtained by applying the usual formula (5.29) developed in earlier.

The above computational procedure could be made more easier through a transformation of the variables x and y in their origin and scale. Let us define two new variables as follows:

$$u = \frac{x - 7.5}{2}$$

and

$$v = \frac{y - 13}{3}$$

The new variable u will thus take on values $-3, -2, -1, 0, 1, 2, 3$ and 4, while the variable v will assume values $-2, -1, 0, 1$ and 2. Thus, replacing the values of x and y by the values of u and v, one can easily compute the correlation coefficient between u and v, and arrive at the same correlation coefficient as between x and y.

$$r_{xy} = \frac{\sum fxy - \frac{\sum fx \sum fy}{n}}{\sqrt{\left\{\sum fx^2 - \frac{\left(\sum fx\right)^2}{n}\right\}\left\{\sum fy^2 - \frac{\left(\sum fy\right)^2}{n}\right\}}}$$

Setting the computed values in the formula (see table in the next page), we obtain

$$= \frac{2147.5 - \frac{166 \times 517}{40}}{\sqrt{\left\{1148 - \frac{166^2}{40}\right\}\left\{7249 - \frac{517^2}{40}\right\}}}$$

$$= \frac{2147.5 - 2145.55}{\sqrt{\{1148 - 688.9\}\{7249 - 6682.22\}}}$$

$$= \frac{1.95}{\sqrt{459.1 \times 566.78}}$$

$$= \frac{1.95}{510} = 0.004$$

Table for Computation of Correlation Coefficient from Bivariate Data

	y												
x	7	10	13	16	19	f	fx	fx^2	fxy
1.5	3	2	3	7	2	17	25.5	38.25	345.0
3.5	3	2	3	3	1	12	42.0	147.0	514.5
5.5	–	–	1	–	–	1	5.5	30.25	71.5
7.5	1	3	1	–	1	6	45.0	337.5	517.5
9.5	–	–	–	1	–	1	9.5	90.25	152.0
11.5	–	–	2	–	–	2	23.0	264.5	299.0
13.5	–	–	–	–	–	0	0	0	0
15.5	–	–	–	1	–	1	15.5	240.25	248.0
f	7	7	10	12	4	40	166.0	1148.0	2147.5
fy	49	70	130	192	76	517			
fy^2	343	700	1690	3072	1444	7249			
fyx	157.5	325.0	663.0	736.0	266.0	2147.5			

5.16 CORRELATION RATIO

In a correlation table, data are grouped in columns and rows. The data represented in any column in such table constitute a **vertical array** or an array of x's, because in each such array, x assumes different values for a fixed value of y. Similarly, the data in any row constitute a **horizontal array** or an array of y's. The nature of the regression in such a table is estimated by the best fitting straight line. Sometimes, the line through the column means might show a well-marked departure from linearity, indicating a curvilinear relationship between variables. An appropriate measure that takes this curvilinearity into consideration is known as the **correlation ratio** and is denoted by η (**eta**).

When the relationship between variables is linear, the extent of association, as we know, is measured by the Pearson's product moment correlation coefficient r. Therefore, r measures the correlation of points about the best fitting straight line, while the correlation ratio measures the concentration of points about the best fitting curve.

In contrast to correlation coefficient, the correlation ratio is not a symmetric measure. This implies that the correlation ratio of y on x is different from that of x on y.

5.16.1 Computing Correlation Ratio

Let $\bar{y}_i$ stand for the mean of the $i-th$ array consisting of all y's and $\bar{y}$, the mean of y in the whole distribution. Then the sum of squares of deviations for the whole distribution within the vertical array is denoted by NS_y^2, so that

$$S'_{yy} = NS_y^2 = \sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2 \quad \text{... (a)}$$

where

$$f_i = \sum_j f_{ij} \text{ and } \bar{y}_i = \frac{1}{f_i}\sum_j f_{ij} y_{ij} .$$

We denote the variance of the y's in the distribution by σ_y^2, where

$$N\sigma_y^2 = S_{yy} = \sum_i \sum_j f_{ij}(y_{ij} - \bar{y})^2 \quad \text{... (b)}$$

where

$$\bar{y} = \frac{\sum_i \sum_j f_{ij} y_{ij}}{\sum_i \sum_j f_{ij}} = \frac{\sum_i \sum_j f_{ij} y_{ij}}{N} = \frac{\sum_i f_i \bar{y}_i}{N}$$

Now by analogy with (5.18), the correlation ratio η_{yx} of y on x is defined by

$$\eta_{yx}^2 = 1 - \frac{S_y^2}{\sigma_y^2} \quad \text{... (c)}$$

From (c), we find that

$$S_y^2 = \sigma_y^2(1 - \eta_{yx}^2) \quad \text{... (d)}$$

which when multiplied by N assumes the form

$$S'_{yy} = S_{yy}(1 - \eta^2{}_{yx}) \quad \text{... (e)}$$

an expression that is analogous to

$$S_{ee} = S_{yy}(1 - r^2) \quad \text{... (f)}$$

where S_{ee} is the sum of squares of deviations from the line of regression of y on x. Since the sum of squares of the deviations in any array being least, when they are measured from the mean of the array,

$$S'_{yy} \le S_{ee} \quad \text{... (g)}$$

That is $(1 - \eta_{yx}{}^2) \le (1 - r^2)$, from which it follows that

$$\eta^2{}_{yx} \ge r^2 \quad \text{... (h)}$$

From (e), it is also evident that

$$\eta^2{}_{yx} \le 1 \quad \text{... (i)}$$

Thus

$$1 \ge \eta^2{}_{yx} \ge r^2 \quad \text{... (j)}$$

In general, η^2 is greater than r^2, and the greater the departure of the line of column means from linearity, the greater the difference. A non-zero value of $\eta^2 - r^2$ is associated with the non-linearity of the regression of one variable over the other. When $\eta_{yx}^2 = r^2$, the regression of y on x is linear and the means of the arrays will lie on the regression line.

To obtain a convenient expression for the correlation ratio, we attempt to express the variance of the y's (σ_y^2) into two parts as follows:

$$N\sigma_y^2 = \sum_i \sum_j f_{ij}(y_{ij} - \bar{y})^2$$
$$= \sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2$$
$$= \sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2 + \sum_i f_i(\bar{y}_i - \bar{y})^2 \quad \ldots (k)$$

The product term becomes zero, since the product term disappears for each array. Hence on dividing both sides of (k) by N, we have

$$\sigma_y^2 = \frac{\sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2}{N} + \frac{\sum_i f_i(\bar{y}_i - \bar{y})^2}{N} \quad \ldots (l)$$

The second term on the right hand side of (k) is the variance of the means of the vertical arrays, each mean being weighted with the frequency f_i of that array. This is denoted by σ_{my}^2 so that

$$\sigma_y^2 = S_y^2 + \sigma_{my}^2 \quad \ldots (m)$$

This shows that the variance of y is expressible as the sum of the two components: the variance within array and $\sigma^2{}_{my}$ is the variance between means of arrays. Comparing (d) and (m), we see that

$$\eta_{yx}^2 = \frac{\sigma_{my}^2}{\sigma_y^2} \quad \ldots (n)$$

Thus the square of the correlation ratio, as appears from (m) is the ratio of the 'variance of between arrays' to the 'total variance'.

The correlation ratio of y on x is

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y} \quad \ldots (o)$$

The correlation ratio η_{yx} is therefore the ratio of the standard deviation of the weighted means of the arrays of y's to the standard deviation of all the y's of the distribution.

The correlation ratio of x on y can similarly be obtained:

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x} \quad \ldots (p)$$

$$\eta_{yx}^2 = \frac{\sigma_{my}^2}{\sigma_y^2} = \frac{\sum f_i(\bar{y}_i - \bar{y})^2}{\sum\sum f_{ij}(y_{ij} - \bar{y})^2}$$
$$= \frac{\frac{\sum f_i \bar{y}_i^2}{N} - \left(\frac{\sum f_i \bar{y}_i}{N}\right)^2}{\frac{\sum f_i y_i^2}{N} - \left(\frac{\sum f_i y_i}{N}\right)^2} \quad \ldots (q)$$

Example 5.13. Calculate the correlation ratio of y on x and that of x on y from correlation table of Example 5.12

Solution: By definition, correlation ratio of y on x is $\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$

From the given table, we calculate σ_y :

$$\sigma_y^2 = \frac{\sum f_i y_i^2}{N} - \left(\frac{\sum f_i y_i}{N}\right)^2 = \frac{7249}{40} - \left(\frac{517}{40}\right)^2 = 14.17$$

so that $\sigma_y = 3.76$.

Now to compute σ_{my}, we construct the following table

Mean of rows ($\bar{y}_i$)	f_i	$f_i\bar{y}_i$	$f_i\bar{y}_i{}^2$
13.53	17	230	3112.00
12.25	12	147	1800.75
13.0	1	13	169.00
11.5	6	69	793.50
16.0	1	16	256.00
13.0	2	26	338.00
0.0	0	0	0
16.0	1	16	256.00
Total	40	517	6725.25

Now the variance between the mean of arrays is

$$\sigma_{my}^2 = \frac{\sum f_i \bar{y}_i^2}{N} - \left(\frac{\sum f_i \bar{y}_i}{N}\right)^2 = \frac{6725.25}{40} - \left(\frac{517}{40}\right)^2 = 1.07$$

Hence

$$\sigma_{my} = 1.03 \text{ and thus } \eta_{yx} = \frac{\sigma_{my}}{\sigma_y} = \frac{1.03}{3.76} = 0.27$$

To calculate the correlation ratio of x on y, a similar table may be constructed. The required calculations are as follows:

Mean of columns ($\bar{x}_i$)	f_i	$f_i\bar{x}_i$	$f_i\bar{x}_i^2$
3.21	7	22.5	72.22
4.64	7	32.5	150.80
5.10	10	51.0	260.10
3.83	12	46.0	176.18
3.50	4	14.0	49.0
Total	40	166	708.30

Now the variance between the mean of arrays is

$$\sigma_{mx}^2 = \frac{\sum f_i\bar{x}_i^2}{n} - \left(\frac{\sum f_i\bar{x}}{n}\right)^2 = \frac{708.3}{40} - \left(\frac{166}{40}\right)^2 = 0.485$$

Hence $$\sigma_{mx} = 0.70$$

Also

$$\sigma_x^2 = \frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2 = \frac{1148}{40} - \left(\frac{166}{40}\right)^2 = 11.48$$

so that σ_x=3.39

Hence $$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x} = \frac{0.70}{3.39} = 0.21$$

5.17 FURTHER ASPECTS OF CORRELATION

In some situations, the variables that we deal with are neither measured on interval scale nor are in rank-order. For example, in the case of two variables, one variable may be nominally measured (e.g. sex), the other may be continuous (e.g. age). In other situations, both the variables may be nominal (e.g. sex and color). The continuous (interval) variable may also sometimes be forced to be dichotomous (e.g. yes/no). In such situations, conventional methods of correlation fail to be appropriate to measure the association between the variables of interest. Depending on the type of variables or level of measurements, the relationships between variables can be measured using some unconventional types of correlation coefficients. We make a brief survey of these coefficients in this section without any

derivation or computational details.

5.17.1 Intra-class Correlation

Intra-class correlation refers to within-class correlation. In some situations, especially in biological and agricultural studies, we may be interested to know how the pairs of the members of the same class or families are correlated among themselves with respect to some common characteristics. Thus, we may be interested in the correlation of heights of brothers or weights of sisters or the lengths of leaves in trees or of yields of certain crops of one or more experimental blocks. The coefficient of correlation obtained in such cases is known as the **intra-class correlation coefficient.**

In obtaining the intra-class correlation, we observe that the relation between two members of the same family is a reciprocal one; for, if P belongs to the same family as Q, then Q belongs to the same family as P. Each pair of members, P and Q will therefore contribute two entries to the correlation table. In one of them, x will be the measure of the characteristic for P and y the measure for Q. In the other, x will be the measure for Q, and y for P. The table will thus be symmetrical.

Let us consider the simplest case in which each of the h families has the same size k. Then there are $k(k-1)$ pairs of values for each family, and consequently, $N=hk(k-1)$ gives the total numbers of pairs of values in the correlation table.

Let x_{ij} denote the measure of the characteristic for the jth member of the ith family (i=1, 2, ...,h ; j=1, 2, ...,k). The hk values of x's and y's of the correlation table occur in different orders. Any one value x_{ij}, occurring as an x in the table will have as its y each of the other $k-1$values for the same family. Thus each value x_{ij} occurs $k-1$ times as an x_i; and mean $\bar{x}$ for the bivariate distribution is therefore

$$\bar{x} = \frac{(k-1)\sum\sum x_{ij}}{hk(k-1)} = \frac{\sum\sum x_{ij}}{hk} \quad \ldots (5.30)$$

and, because of symmetry, this equals $\bar{y}$. The variance of x's is

$$\sigma_x^2 = \frac{\sum\sum(x_{ij} - \bar{x})^2}{hk} \quad \ldots (5.31)$$

and, because of symmetry, the variance of y will be equal to the variance of x, i.e. $\sigma_x^2 = \sigma_y^2 = \sigma$ (a common variance). The coefficient of intra-class

correlation may now be given by the usual formula

$$r_{(\text{intra})} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_x}$$

It can be shown that

$$\begin{aligned} N\text{Cov}(x,y) &= k\sum(\bar{x}_i - \bar{x})(k\bar{x}_i - k\bar{x}) - \sum\sum(x_{ij} - \bar{x})^2 \\ &= k^2\sum(\bar{x}_i - \bar{x})^2 - hk\sigma^2 \\ &= hk^2\sigma_m^2 - hk\sigma^2 \end{aligned}$$

where σ_m^2 is the variance of the means of the families. Hence

$$\begin{aligned} \text{Cov}(x,y) &= \frac{hk^2\sigma_m^2 - hk\sigma^2}{N} \\ &= \frac{hk\left(k\sigma_m^2 - \sigma^2\right)}{N} \\ &= \frac{k\sigma_m^2 - \sigma^2}{k-1} \end{aligned}$$

Since $\sigma_x = \sigma_y = \sigma$, we have

$$r_{(\text{intra})} = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_x} = \frac{k\sigma_m^2 - \sigma^2}{(k-1)\sigma^2}$$

Since $\sigma_m^2 \geq 0$, the intra-class correlation cannot be less than $-1/(k-1)$ though it may attain a value +1.

The method of computing the intra-class correlation coefficient is illustrated below with an example.

Example 5.18: A survey of five families each with three brothers was conducted and their weights were measured in kilogram. The following table displays the results of this investigation. Compute the intra-class correlation coefficient.

	Family				
	1	2	3	4	5
Brother 1	60	65	62	45	64
Brother 2	40	56	60	55	64
Brother 3	50	50	64	65	64
Mean	50	57	62	55	64

$$\bar{x} = \frac{\sum\sum x_{ij}}{hk} = \frac{864}{15} = 57.6$$

$$\begin{aligned} \sigma^2 &= \frac{\sum\sum(x_{ij} - \bar{x})^2}{hk} \\ &= \frac{(60-57.6)^2 + (65-57.6)^2 + \ldots + (64-57.6)^2}{15} \\ &= 59.84 \end{aligned}$$

$$\begin{aligned} \sigma_m^2 &= \frac{\sum(\bar{x}_i - \bar{x})^2}{h} \\ &= \frac{(50-57.6)^2 + (57-5.7.6)^2 + \ldots + (64-57.6)^2}{5} \\ &= 25.04 \end{aligned}$$

The intra-class correlation is thus

$$r_{(\text{intra})} = \frac{k\sigma_m^2 - \sigma^2}{(k-1)\sigma^2} = \frac{3(25.04) - 59.84}{2(59.84)} = 0.1276$$

5.17.2 Tetrachoric Correlation

The word tetrachoric refers to a 2×2 contingency table. The tetrachoric correlation coefficient is an alternative to the Pearson's r and measures the relationship between two variables, which are continuous in nature, but both have been turned into dichotomous variable through a suitable recoding procedure. Suppose we have n objects of two attributes A and B classified as follows:

	A	$\overline{A}$	Total
B	f_{11}	f_{12}	f_{1+}
$\overline{B}$	f_{21}	f_{22}	f_{2+}
Total	f_{+1}	f_{+2}	n

It is assumed that the attributes A and B are measurable quantitatively and are normally distributed. The coefficient used to measure the degree of linear association between the two attributes is called the **tetrachoric correlation coefficient.**

There is no straightforward method of computing tetrachoric correlation coefficient. Davidoff and Gohen (1953), however, developed a set of

correlation may now be given by the usual formula

$$r_{(\text{intra})} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_x}$$

It can be shown that

$$\begin{aligned} N\text{Cov}(x, y) &= k\sum(\bar{x}_i - \bar{x})(k\bar{x}_i - k\bar{x}) - \sum\sum(x_{ij} - \bar{x})^2 \\ &= k^2\sum(\bar{x}_i - \bar{x})^2 - hk\sigma^2 \\ &= hk^2\sigma_m^2 - hk\sigma^2 \end{aligned}$$

where σ_m^2 is the variance of the means of the families. Hence

$$\begin{aligned} \text{Cov}(x, y) &= \frac{hk^2\sigma_m^2 - hk\sigma^2}{N} \\ &= \frac{hk(k\sigma_m^2 - \sigma^2)}{N} \\ &= \frac{k\sigma_m^2 - \sigma^2}{k-1} \end{aligned}$$

Since $\sigma_x = \sigma_y = \sigma$, we have

$$r_{(\text{intra})} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_x} = \frac{k\sigma_m^2 - \sigma^2}{(k-1)\sigma^2}$$

Since $\sigma_m^2 \geq 0$, the intra-class correlation cannot be less than $-1/(k-1)$ though it may attain a value +1.

The method of computing the intra-class correlation coefficient is illustrated below with an example.

Example 5.18: A survey of five families each with three brothers was conducted and their weights were measured in kilogram. The following table displays the results of this investigation. Compute the intra-class correlation coefficient.

	Family				
	1	2	3	4	5
Brother 1	60	65	62	45	64
Brother 2	40	56	60	55	64
Brother 3	50	50	64	65	64
Mean	50	57	62	55	64

$$\bar{x} = \frac{\sum\sum x_{ij}}{hk} = \frac{864}{15} = 57.6$$

$$\begin{aligned} \sigma^2 &= \frac{\sum\sum(x_{ij} - \bar{x})^2}{hk} \\ &= \frac{(60-57.6)^2 + (65-57.6)^2 + \ldots + (64-57.6)^2}{15} \\ &= 59.84 \end{aligned}$$

$$\begin{aligned} \sigma_m^2 &= \frac{\sum(\bar{x}_i - \bar{x})^2}{h} \\ &= \frac{(50-57.6)^2 + (57-5.7.6)^2 + \ldots + (64-57.6)^2}{5} \\ &= 25.04 \end{aligned}$$

The intra-class correlation is thus

$$r_{(\text{intra})} = \frac{k\sigma_m^2 - \sigma^2}{(k-1)\sigma^2} = \frac{3(25.04) - 59.84}{2(59.84)} = 0.1276$$

5.17.2 Tetrachoric Correlation

The word tetrachoric refers to a 2×2 contingency table. The tetrachoric correlation coefficient is an alternative to the Pearson's r and measures the relationship between two variables, which are continuous in nature, but both have been turned into dichotomous variable through a suitable recoding procedure. Suppose we have n objects of two attributes A and B classified as follows:

	A	$\overline{A}$	Total
B	f_{11}	f_{12}	f_{1+}
$\overline{B}$	f_{21}	f_{22}	f_{2+}
Total	f_{+1}	f_{+2}	n

It is assumed that the attributes A and B are measurable quantitatively and are normally distributed. The coefficient used to measure the degree of linear association between the two attributes is called the **tetrachoric correlation coefficient.**

There is no straightforward method of computing tetrachoric correlation coefficient. Davidoff and Gohen (1953), however, developed a set of

correlation coefficients corresponding to the ratios

$$\frac{f_{11} \times f_{22}}{f_{12} \times f_{21}} \text{ and } \frac{f_{12} \times f_{21}}{f_{11} \times f_{22}}.$$

The computation of correlation coefficient first involves the calculation of these ratios and then with reference to the tabulated values of these ratios, the tetrachoric correlation is read off from the table. Two sets of values are given in the table. The first set is used if $f_{11} \times f_{22} > f_{12} \times f_{21}$, otherwise the second set is used. A detailed exposition of the method is found in Davidoff and Gohen (1953).

5.17.3 Phi-coefficient

When the categories of the dichotomous attributes can be coded 0 and 1, a convenient formula may be developed to obtain the product moment correlation coefficient (r). The formula is as follows:

$$r_{\phi} = \frac{f_{11} f_{22} - f_{12} f_{21}}{\sqrt{f_{1+} f_{2+} f_{+1} f_{+2}}}$$

The square of r_{ϕ}, in essence, is a measure of association in a 2×2 contingency table and is known as phi-squared coefficient (ϕ^2) varying between 0 and 1. Obviously, phi-squared coefficient is commonly interpreted as the percent of variation in the dependent variable that is explained by the independent variable.

Example 5.19: In a community survey, 1218 males and 973 females were classified as having good health or poor health. For the attribute 'sex', we assign a value '1' for male and '0' for female. Similarly, for health status, the values '1' and '0' are assigned for the good health and poor health respectively. With these labels, the following table is constructed.

	Health status		
Sex	Good	Poor	Total
Male	1103	115	1218
Female	207	766	973
Total	1310	881	2191

Compute the coefficient of correlation between sex and health status.

Solution: Here f_{11}=1103, f_{22}=766, f_{12}=115, f_{21}=207, $f_{1\cdot}$=1218, $f_{2\cdot}$=973, $f_{\cdot 1}$=1310 and $f_{\cdot 2}$=881. Hence using the above formula

$$r_{\varphi} = \frac{(1103)(766) - (115)(207)}{\sqrt{(1310)(881)(1218)(973)}} = 0.702$$

From the formula, it is apparent that r_{ϕ}=1, if f_{12}=f_{21}=0 and r_{ϕ}=−1, if f_{11}=f_{22}=0. In this sense, the correlation coefficient gives both the direction and strength of association. Since r_{ϕ}^2=0.493, about 49% of the variations in health status is explained by sex.

5.17.4 Point Bi-serial Correlation

Point bi-serial correlation provides a measure of the linear relation between a continuous variable, such as scores on a test, and two-categorized, or dichotomous responses of another variable, such as "pass" or "fail" on a test, "male" or "female", "malnourished" or "well nourished". Such types of data are frequently found in psychology, education and public health.

The data when arranged in the form of a frequency distribution display a table comprised of R rows and 2 columns. Point bi-serial correlation is essentially a Pearson's product-moment correlation. If we assign a value 1 to individuals in one category and 0 to individuals in the other category and calculate the product-moment correlation coefficient, the result is a point bi-serial correlation coefficient. Points other than 0 and 1 may also be assigned.

The following formula is used to compute point bi-serial correlation coefficient:

$$r_{\text{p-bi}} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\sqrt{pq}$$

where

s_x=standard deviation of scores on the continuous variable;
p=proportion of individuals in the first category of the dichotomous variable;
q=proportion of individuals in the second category of the dichotomous variable such that q=1−p;

$\bar{x}_p$ and $\bar{x}_q$ are the mean scores on the continuous variable of individuals corresponding to the two categories.

Example 5.20: The following table shows the scores on a test and on a test item for a group of 14 students. Compute point bi-serial correlation coefficient. (The last two columns were not given in the original problem).

Individual	Inventory score (x)	Item score	x_p	x_q
1	6	0	-	6
2	8	1	8	-
3	8	0	-	8
4	11	0	-	11
5	16	1	16	-
6	25	0	-	25
7	27	0	-	27
8	31	0	-	31
9	31	1	31	-
10	39	0	-	39
11	44	0	-	44
12	50	1	50	-
13	56	1	56	-
14	68	1	68	-
Total	420	-	229	191

Solution: In the item score column, a '1'stands for an individual's passing the test, while '0' stands for his failing the test. Thus mean scores for those who pass and those who fail are respectively

$$\bar{x}_p = \frac{\sum x_p}{6} = \frac{229}{6} = 38.17 \text{ and } \bar{x}_p = \frac{\sum x_q}{8} = \frac{191}{8} = 23.88.$$

The mean of the scores is

$$\bar{x} = \frac{\sum x}{n} = \frac{420}{14} = 30$$

and the standard deviation of the scores is computed as

$$s_x^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{(6-30)^2 + (8-30)^2 + \ldots + (68-30)^2}{13} = 379.4704,$$

so that $s_x = 19.48$. The p and q values are p=6/14=0.43, q=8/14=0.57 respectively.

Hence

$$r_{\text{p-bi}} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\sqrt{pq} = \frac{38.17 - 23.88}{19.48}\sqrt{.43 \times .57} = 0.36$$

An alternative method for calculating point bi-serial correlation coefficient is to use the formula

$$r_{\text{p-bi}} = \frac{\bar{x}_p - \bar{x}}{s_x}\sqrt{\frac{p}{q}}$$

where $\bar{x}$ is the mean of all scores on the continuous variable, the other quantities are as defined before. It is easy to check that for the given data, the alternative formula also gives a value of 0.36 for bi-serial correlation coefficient:

$$r_{\text{p-bi}} = \frac{\bar{x}_p - \bar{x}}{s_x}\sqrt{\frac{p}{q}} = \frac{38.17 - 30}{19.48}\sqrt{\frac{.43}{.57}} = 0.36$$

Point bi-serial correlation is not independent of the proportions in the two categories. When p=q=0.5, its maximum and minimum values will differ from those which would be obtained when, say p=0.20 and q=0.80. The maximum value of $r_{\text{p-bi}}$ never reaches +1 and its minimum value never attains –1.

5.17.5 Bi-serial Correlation

The bi-serial correlation is designed to measure the correlation in a situation where one variable is quantitatively measured and the other variable is dichotomous. It is assumed that the variable underlying the dichotomy is continuous and normal. The formula for computing r_{bi} is as follows:

$$r_{\text{bi}} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\left(\frac{pq}{y}\right) \quad \ldots (5.32)$$

where s_x= standard deviation of the continuous scores, defined as for point bi-serial correlation, $\bar{x}_p$ and $\bar{x}_q$ =mean test scores for those who 'pass' and 'fail' the item respectively, p and q = proportions who pass and fail the item and y = height of the ordinate of the unit normal curve at the point of division between the p and q proportions of cases.

If a bivariate table comprised of R rows and C columns is reduced to a table of R rows and 2 columns, bi-serial correlation will provide a more accurate estimate of the correlation based on the R×C table than the point bi-serial correlation. Like simple correlation coefficient, the point bi-serial correlation assumes values between –1 and +1. Values of the coefficient greater than unity might occur under gross departure of the continuous variable from normality.

Thus for the above table illustrating the calculation of point bi-serial

correlation with p=.43 and q=.57, the height of the ordinate of the unit normal curve at the point of dichotomy is y=.393. And for $\bar{x}_p = 38.17$ and $\bar{x}_q = 23.88$, s_x=19.48,

$$r_{bi} = \frac{\bar{x}_p - \bar{x}_q}{s_x}\left(\frac{pq}{y}\right) = \frac{38.17 - 23.88}{19.48}\left(\frac{.43 \times .57}{.393}\right) = 0.46$$

An alternative formula for computing bi-serial correlation is

$$r_{bi} = \frac{\bar{x}_p - \bar{x}}{s_x}\left(\frac{p}{y}\right) \quad \ldots (5.33)$$

where $\bar{x}$ is the overall mean score in the sample. The above example yields a mean score of the total sample $\bar{x} = 30$. Applying the alternate formula,

$$r_{bi} = \frac{\bar{x}_p - \bar{x}}{s_x}\left(\frac{p}{y}\right) = \frac{38.17 - 30}{19.48}\left(\frac{.43}{0.393}\right) = 0.46 \text{ ,as before.}$$

EXERCISES 5

1. Distinguish between the following concepts:
   (a) Univariate and bivariate distributions, (b) Simple and multiple regressions, (c) Linear and curvilinear regressions and (d) (d) Curvilinear and polynomial regressions.
2. What do you mean by a regression model? How do you determine the order of a model? Distinguish between a linear model and a non-linear model. Give examples. Explain the assumptions underlying a linear regression model.
3. How do you distinguish correlation analysis from regression analysis? Give examples of their practical utility.
4. What do you mean by a regression equation? Why is the regression line often referred to as the line of conditional means?
5. What are the parameters of a simple regression model? Why do you need these parameters to be estimated? How do you estimate these parameters?
6. What is least-squares method? What is its importance? Describe this method for estimating the parameters of a simple regression model.
7. What is scatter diagram? What purpose does it serve? Discuss its importance in correlation and regression analysis.
8. What is the meaning of regression line? Why are there two regression lines in a bivariate distribution? What is the difference between an observed regression line and a true regression line?

9. What is coefficient of correlation? What does it measure? Show that the coefficient of correlation is invariant to the changes in origin and scale of measurement.
10. What does a coefficient of correlation measure? Discuss the situations when r=+1, r=–1, r=0.
11. What is a regression coefficient? How does it differ from a correlation coefficient? Explain positive correlation and negative correlation. When do you call a correlation to be perfect? Show that correlation coefficient lies between –1 and +1.
12. Discuss the principles of least-squares method. Show that a least-squares line always passes through the point $(\bar{X}, \bar{Y})$. Show that the least-squares line Y=a+bX can also be expressed as $Y = \bar{Y} + b(X - \bar{X})$. Define sum of squares due to regression and sum of squares of deviations. Show that these two sums of squares add up to the total sum of squares of the dependent variable.
13. What is coefficient of determination? What does it measure? Why do you need an adjusted value for this coefficient?
14. Discuss the properties of a regression coefficient. In what way is it related to correlation coefficient? Discuss the important uses of regression coefficient.
15. Suppose that we wish to study the distributions of the weights of college students with relation to their heights. We divide the population of college students into groups of approximately equal heights and for any selected height there is a distribution of weights. Let x be the measurement of heights in inches and y the measurement of weights in pounds. The observations are shown as follows:

Heights (x)	Weights (y)	Heights (x)	Weights (y)
60	115	66	135
60	120	66	170
60	130	66	140
60	125	66	155
62	130	68	150
62	140	68	160
62	120	68	175
64	135	70	180
64	130	70	160
64	145	70	175

   (a) Plot the data as a scatter diagram.
   (b) Draw a freehand line on the diagram to fit these data.
   (c) Construct an equation to represent the freehand line.