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:

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.

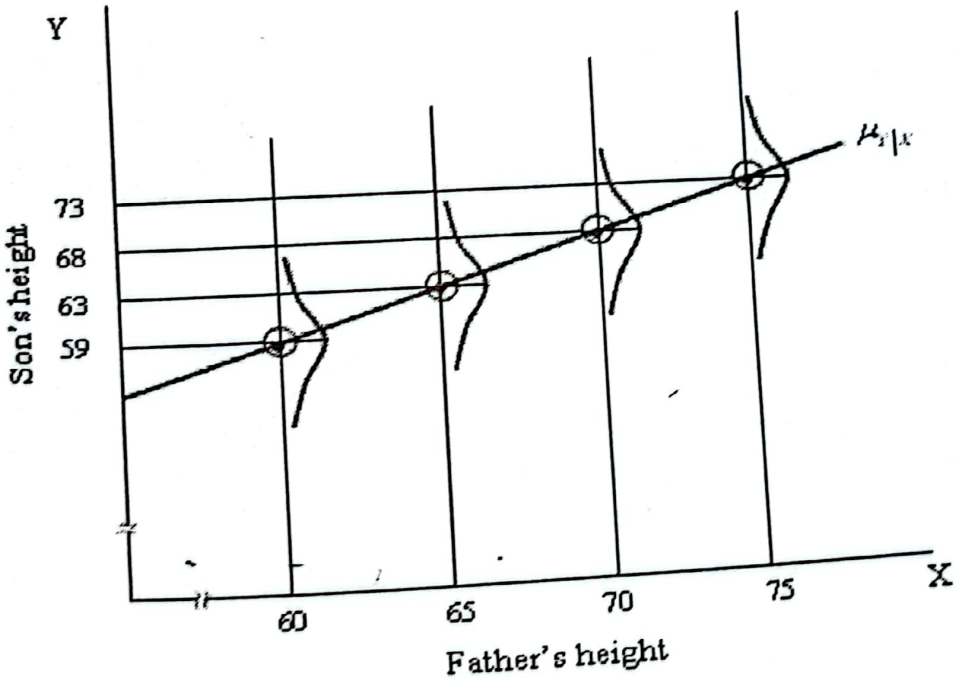

Figure 5.2: Population regression line

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

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.

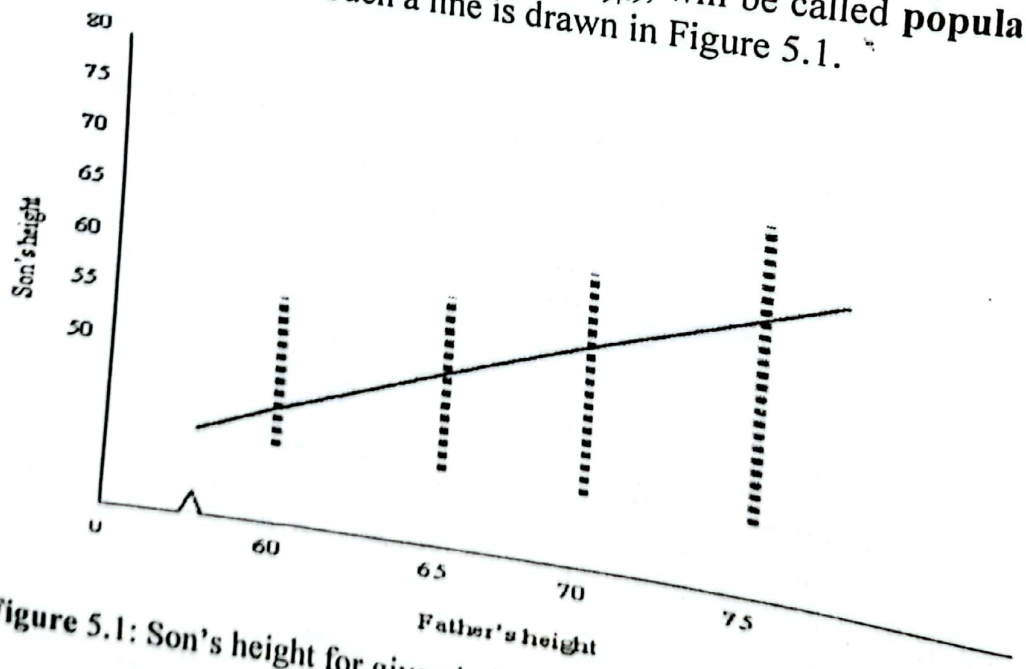

Figure 5.1: Son's height for given height of father: Hypothetical data

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

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:

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

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).

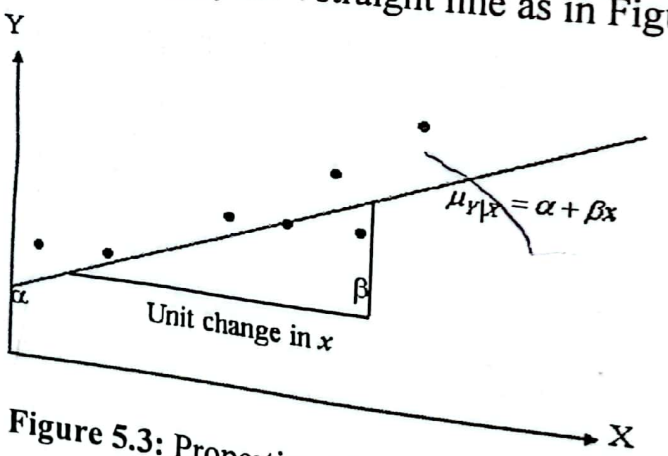

Figure 5.3: Properties of a regression line

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.

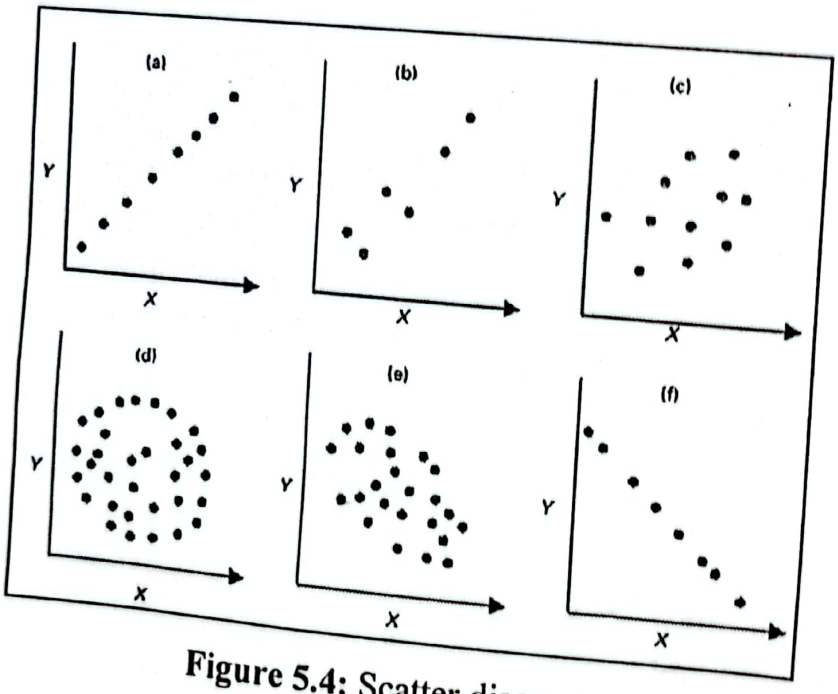

Figure 5.4: Scatter diagram

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

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

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

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

and

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

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

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

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

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

and

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

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:

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

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

$$b = \frac{S_{xx}}{S_x} \qquad \ldots (5.8f)$$

Once b is computed, a can be obtained as

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

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

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

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 $\mu_{y \cdot x} = \alpha + \beta x$.

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:

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

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

and

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 store 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 x= 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 given 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

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

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

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

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:

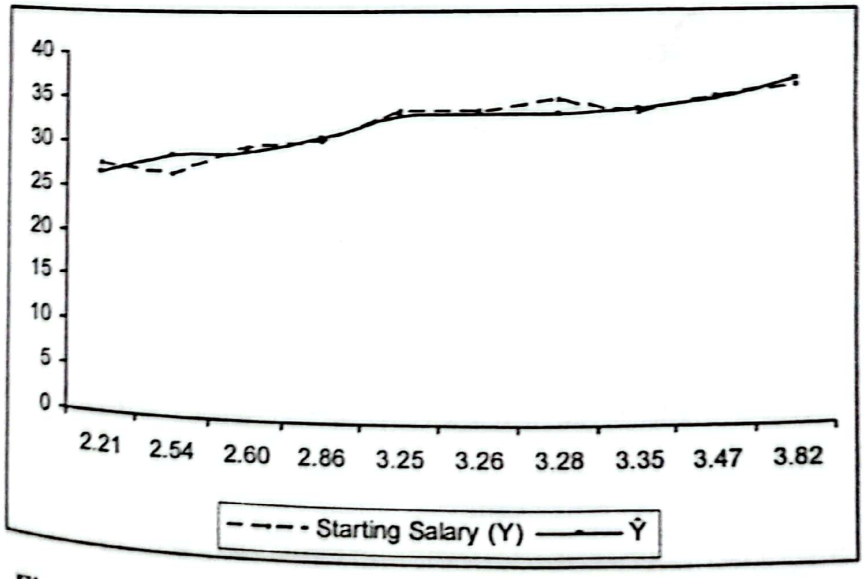

Figure 5.5: Observed and expected values in Example 5.4

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

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

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

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

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

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

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

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

The symbolic representation of the above equation is,

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

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

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

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

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

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

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

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

$$= 4\left\{8128 - \frac{70 \times 1080}{10}\right\}$$

$$= 2272$$

as ought to be

We can now compute r^2 as follows:

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

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:

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

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:

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

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

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

$$b^2 S_{xx} = b(bS_{xx})$$

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

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:

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

Writing in full

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

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.

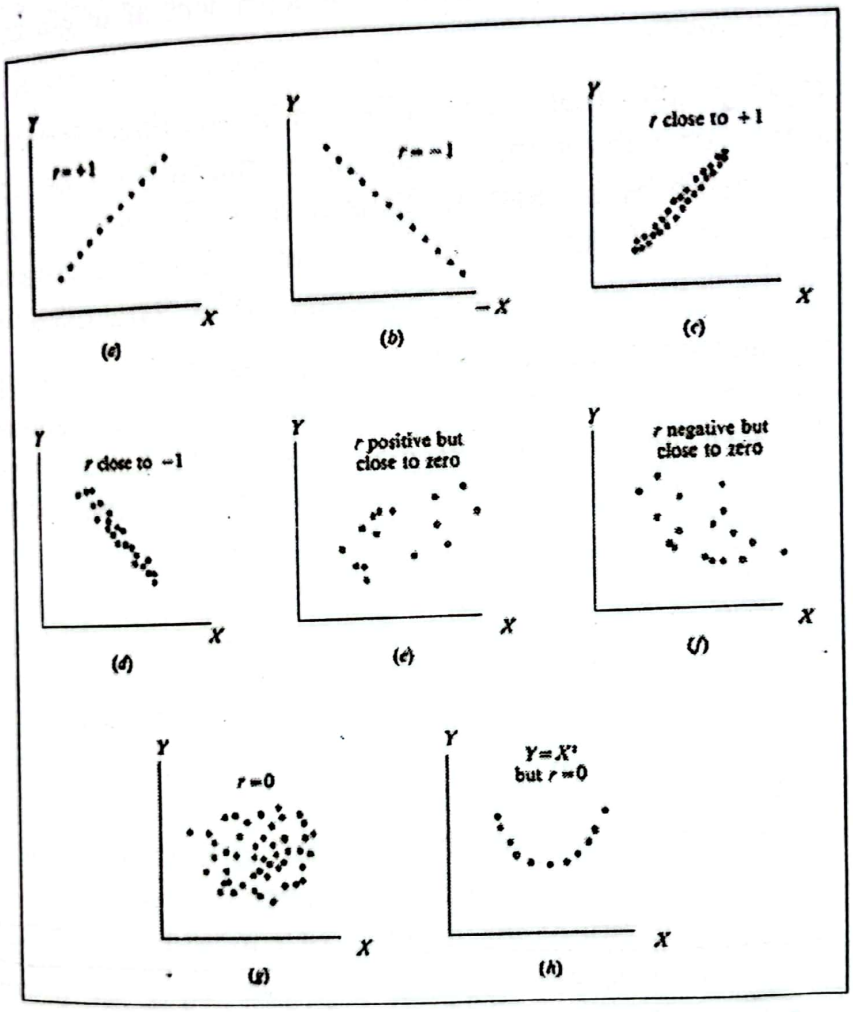

Figure 5.7: Scatter diagrams with varying degrees of r

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:

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

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 $r - \text{PE} \le \rho \le r + \text{PE}$.

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

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

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

Example 5.8: The accompanying table shows the proportions of coa miners who exhibit symptoms of pneumoconiosis to their number of year 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

om the table we have,

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

) that

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

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

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

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):

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

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

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

Hence

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

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

The direct computation yields

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

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

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 way to assess the strength of the monotonic relationship between x and y