UNIT 5 SIMPLE LINEAR REGRESSION

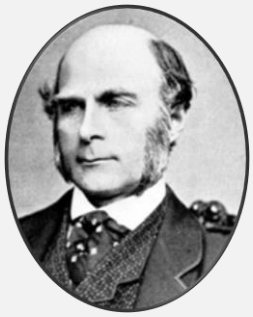
Structure

5.1 INTRODUCTION

In Block 1 of MSTE-004, you have learnt about various methods of statistical analysis methods applied to categorical data. However, we cannot apply those methods when we wish to determine the relationship between two quantitative variables. In this unit, we shall discuss the simple linear regression model, which represents linear relationship between two quantitative variables. In bivariate data, two variables may be related to each other in such a way that one variable *depends* on the other, i.e., the changes in the value of one variable affect the other variable. The variable that affects the other variable is termed the **independent**, **regressor** or **predictor** variable; the variable that is affected by changes in the independent variable, is termed the **dependent** or **response** variable. We will use the terms "**response variable**" and "**regressor variable**" for **dependent** and **independent** variables, respectively, throughout this block.

Regression considers the relationship between two variables. It is to be noted that the dependence or, say, statistical relationship does not imply a cause-and-effect relationship between the two variables.

The idea of regression was first introduced by Sir Francis Galton in 1908 while studying the relationship between the height of fathers and sons. In regression analysis, we propose a mathematical model for ascertaining the linear relationship between two or more variables. We make predictions for the dependent variable using known values of the independent variables through such regression models. If we want to study the relationship of response variable only on **one** regressor variable, for example, the dependence of systolic blood pressure (SBP) on age, then such a study is termed as simple regression analysis. In other words, the term **'simple regression' relates to the fact that there is only one regressor variable in the model**. In the same way, we may predict how the level of blood sugar of a person will be affected if s/he walks for half an hour daily. Here blood sugar level is the dependent variable while the duration of walk is the independent variable. Note that we cannot say that walking is the only variable that affects the blood sugar, but we are considering it as one of the many variables that could. When we have two or more regressor variables, it is known as multiple regression. We shall discuss multiple regression in Unit 7 of this block.



Sir Francis Galton
(1822-1911)
A renowned British biologist who was engaged in the study of heredity.

In this unit, you will study simple linear regression models. In Sec. 5.2, we explain the concept of simple linear regression. We state its underlying assumptions and explain how to construct the scatter plot. Next, we describe how to fit a simple linear regression model using the method of least squares

with the help of several examples (Sec. 5.3). We also discuss how to use a simple linear regression model to predict the value of the response variable. In Sec. 5.4, we describe the properties of estimated regression coefficients.

In the next unit, you will learn about statistical inference in simple linear regression.

Objectives

After studying this unit, you should be able to:

- identify the response and regressor variables in given bivariate data;
- describe the concept of simple linear regression;
- draw a scatter plot for getting an idea of relationship between two variables;
- estimate the parameters of simple linear regression using the method of least squares; and
- fit a simple linear regression model for given data.

5.2 SIMPLE LINEAR REGRESSION

Note that we use the term simple linear regression when we have only one regressor variable in model which is linearly related with the response variable.

Let us begin the discussion with a few examples. Consider the relationship between age and length of arms in human beings. We may take age as the independent variable and arm length as the dependent variable. It is because the arm length may be a function of age, but age may not be determined by arm length. Or consider the relationship between systolic blood pressure (SBP) and the time taken to climb up some stairs (in minutes). Here SBP may be considered as a dependent variable and time taken as an independent variable. In the same way, let us consider the relationship between age and cholesterol levels. Then cholesterol level may be considered as a dependent variable and age as an independent variable.

In conventional notation, we denote the regressor (or independent) variable by X and the response (or dependent) variable by Y. The relationship between X and Y is also useful for prediction. Mathematically, we can define the simple linear relationship of a quantitative response variable (Y) on a quantitative regressor variable (X) as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \ldots (1)$$

where β_0 is the intercept and β_1, the slope when equation (1) is plotted on a graph paper (Fig. 5.1).

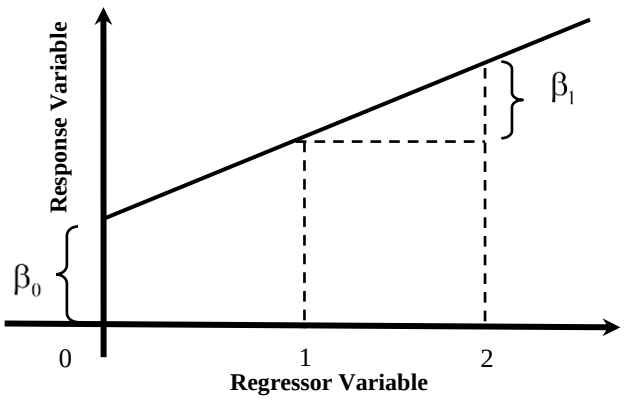


Fig. 5.1: Intercept and slope of the regression line.

The parameters β_0 and β_1 are unknown constants called **regression coefficients**. The slope β_1 is the change in the value of the response variable corresponding to a unit change in the regressor variable. ε is a random error component and it is assumed that it has mean zero and unknown variance σ^2.

In this block, we will frequently use the terms **simple, multiple** and **linear**. If we study the dependence of a response variable only on a single regressor variable, for example, the dependence of systolic blood pressure (SBP) on age, then such a study is termed as **simple regression analysis**. However, if we study the dependence of a response variable on more than one regressor variable, such as the dependence of systolic blood pressure (SBP) on age, weight and many more, it is termed as **multiple regression analysis**. Thus, we have only one regressor variable in simple regression analysis, while in multiple regression, we have more than one regressor variable.

Let us now explain the term "**linear**". You must make sure that you are clear about the meaning of the term **linear** as we shall be using this term throughout the block. Linear relationship between two variables implies that a plot of one variable versus the other is a straight line. In this unit, we shall consider linearity of two types: i) linearity in variables and ii) linearity in parameters. Let us explain what we mean.

If the response variable is a linear function of the regressor variable, then the linearity is termed as '**linearity in variables'**. In this case, the regression curve (or the scatter plot to be discussed in Sec. 5.2.2) will be a straight line.

However, if the response variable is a linear function of the parameters, i.e., β's, then the linearity is termed as '**linearity in parameters'**. If the model is linear in parameters but not in variables, we can use some transformation to make it linear in variables.

In this block, we are considering only those models, which are linear in both parameters and variables. Hence, we can conclude that **simple linear regression** refers to a linear (straight-line) relationship between only two variables. Let us now explain the underlying assumptions for a linear regression model.

5.2.1 Assumptions underlying Linear Regression Model

For applying any statistical method, first of all, we should study the assumptions underlying it. So we shall discuss the assumptions in the context of simple as well as multiple linear regression models. The assumptions given at (i) – (x) given below apply to both simple and multiple regression models while assumption (xi) applies only to multiple regression analysis. So, in order to deal with a valid linear regression analysis, we work with the following assumptions:

i) The regression model is assumed to be linear in parameters.

ii) The values taken by the regressor variables $X_1, X_2, ..., X_k$ are assumed to be fixed (non-stochastic or non-random) in repeated samples.

iii) Variability in X values is assumed, i.e., the values of the regressor variable (X) in a given sample must not be the same.

iv) The error term ε is assumed to be normally distributed with mean 0 and variance σ^2, i.e., $\varepsilon \sim N(0, \sigma^2)$.

v) For a given value of X, the expected value of the random error term (conditional mean of ε_i) is assumed to be zero. Symbolically, we write $E(\varepsilon_i | x_i) = 0$, for each i.

vi) For a given value of X, the variance of the error (ε_i) is assumed to be the same for all observations, i.e., conditional variances of ε_i are identical. Symbolically, we write $\text{var}(\varepsilon_i | x_i) = \sigma^2$. This is known as **assumption of homoscedasticity** or equal variance of ε_i.

vii) For any two given values of regressor variable (X), say, x_i and x_j $(i \neq j)$, the correlation between two corresponding error terms ε_i and ε_j $(i \neq j)$ is assumed to be zero. Symbolically, we write $\text{cov}(\varepsilon_i, \varepsilon_j | x_i, x_j) = 0$. It is known as **assumption of no auto-correlation** between two error terms.

The simple linear regression model can be used to model either the value of the response variable (y) or the mean value of the response variable $(\bar{y})$ as a linear function of the regressor variable.

viii) There should be no correlation between error term and regressor variable. In other words, we assume that the covariance between ε_i and x_i should be zero. Mathematically, we write $\text{cov}(\varepsilon_i, x_i) = 0$.

ix) The number of observations n is assumed to be greater than the number of parameters to be estimated. For example, for the model $Y = \beta_0 + \beta_1 X + \varepsilon$, we need at least three pairs of observation, i.e., (y_i, x_i); i =1, 2 and 3 for estimating two unknown parameters β_0 and β_1.

x) There should not be any specification bias or error in the model. So, the regression model should be correctly specified.

The unit of β_0 is the same as the unit of response variable and the unit of β_1 is the same as the unit of response variable divided by the unit of regressor variable.

xi) There should not be perfect multicollinearity, i.e., the regressor variables must be uncorrected.

Before fitting the regression model, we draw a scatter diagram to get a rough idea of the relationship between response and regressor variables. Let us now explain what a scatter diagram is.

5.2.2 Scatter Diagram

A scatter diagram is a simple two-dimensional plot relating the magnitudes of response and regressor variables. In a scatter diagram, we plot the response and regressor variables to get a rough idea of the relationship between them. We follow the steps given below for plotting a scatter diagram:

1. We draw the horizontal (X) and vertical (Y) axes.
2. We take the regressor variable on the X-axis and the response variable on the Y-axis.
3. We mark points corresponding to the (X, Y) values.

If the points so obtained are approximately scattered around a straight line, the relationship is said to be linear. In this situation, we apply simple linear regression method to study the relationship between the variables. However, if the points do not lie around a straight line, we use transformation or non-linear

regression method, to obtain a best fitted regression model. We shall not discuss transformation and non-linear regression in this course as it is beyond the scope of this course.

Let us consider the following example to learn how to plot a scatter diagram.

Example 1: A sample of 15 women of age group 25-45 years was collected to investigate the effect of age on systolic blood pressure (SBP). Data on the age and systolic blood pressure of 15 women are recorded in Table 1.

Table 1: Data on age and systolic blood pressure

S. No.	Systolic Blood Pressure(mm/hg)	Age(years)
1	124	30
2	134	38
3	135	39
4	121	26
5	122	29
6	119	27
7	128	32
8	118	25
9	120	26
10	123	31
11	129	37
12	117	25
13	131	35
14	126	34
15	134	40

i) Identify the response and regressor variables.

ii) Plot the scatter diagram.

Solution:

i) We know that the SBP depends on age but age does not depend on SBP, so the response (dependent) variable is systolic blood pressure and regressor (independent) variable is age.

ii) To plot a scatter diagram, we take the age on the horizontal (X) axis and the systolic blood pressure on the vertical (Y) axis. The resulting scatter diagram for the data of Table 1 is shown in Fig. 5.2.

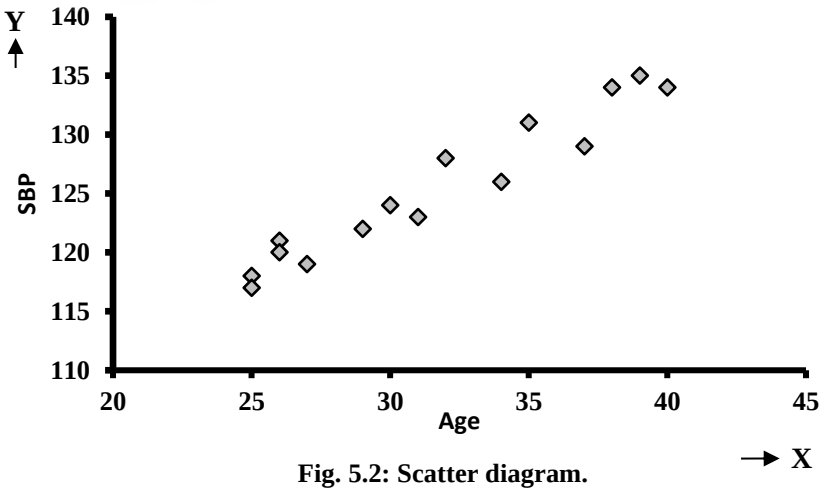


Fig. 5.2: Scatter diagram.

Note that Fig. 5.2 shows an upward linear relationship between age and SBP. Hence, we can conclude that SBP increases as age increases and vice-versa.

Now, you can try the following exercises.

E1) State the assumptions of simple linear regression model.

E2) Differentiate between the following:

i) Response variable and regressor variable

ii) Linearity in parameter and linearity in variable

iii) Simple regression and multiple regression.

E3) To study the impact of duration of daily walk on the sugar level of diabetic patients, the data on the average duration of walk (in minutes) and random blood sugar level (mg/dL) of 15 diabetic patients are given in Table 2.

Table 2: Random blood sugar and duration of walk of 15 diabetic patients

S. No.	Random Blood Sugar	Duration of Walk
1	430	20
2	420	25
3	410	30
4	400	30
5	390	45
6	395	35
7	420	30
8	410	35
9	400	35
10	390	40
11	380	45
12	370	50
13	390	35
14	365	55
15	325	45

i) Identify the response and regressor variables.

ii) Plot the scatter diagram and interpret it.

We now describe the method of least squares for estimating the regression coefficients.

The objective of the method of least squares is to find the estimated values of the parameters by choosing a regression line that is closest to the given data.

5.3 LEAST SQUARES ESTIMATION OF REGRESSION COEFFICIENTS

Suppose, we have two variables X and Y, of which X is the regressor variable and Y, the response variable. Suppose that we have n pairs of observations on X and Y, say, (x_i, y_i); i = 1, 2, 3, ..., n.

We can represent the linear relationship of Y on X defined in equation (1) as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{... (2)}$$

Assuming that the sample observations satisfy the simple linear regression model defined in equation (2), we can rewrite the simple regression model in terms of n pairs of observations in a sample. So for the i^{th} pair (x_i, y_i), we write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \qquad i = 1, 2, \ldots, n \qquad \ldots(3)$$

where ε_i is the i^{th} error component, i.e., the difference between the i^{th} observed and actual values of Y.

For the i^{th} pair, we define the fitted regression model as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i; \qquad i = 1, 2, \ldots, n \qquad \ldots(4)$$

You have learnt in "MST-002: Descriptive Statistics" that the method of least squares leads to a best fitted model. We use the same method to estimate β_0 and β_1. So, the sum of squares of the differences between the given observation, i.e., observed value (y_i) and the estimated value ($\hat{y}_i$) lying on the straight line is minimum.

The corresponding **error sum of squares** is given by

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad \ldots (5)$$

We now equate the derivatives of the **error sum of squares** given in equation (5) with respect to the unknown parameters β_0 and β_1 to zero. It will provide us the least squares estimate of β_0 and β_1, which minimise the error sum of squares, i.e., E. Therefore, differentiating equation (5) with respect to β_0 and equating the result to zero, we get

$$\frac{\partial E}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

or $$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad \ldots (6)$$

In the same way, differentiating equation (5) with respect to β_1 and equating the result to zero, we have

$$\frac{\partial E}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) = 0$$

or $$\sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad \ldots (7)$$

Equations (6) and (7) are known as the **normal equations**.

We can rewrite equation (6) as:

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i$$

or, equivalently

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} \quad \ldots (8)$$

We compute the value of β_1 by substituting the value of β_0 from equation (8) in equation (7) as:

$$\sum_{i=1}^{n} y_i x_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} + \beta_1 \left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}} \quad \ldots (9)$$

We can also rewrite equation (9) as:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_1^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \quad \ldots (10)$$

or $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n \bar{y} \bar{x}}{\sum x_i^2 - n \bar{x}^2} \quad \ldots (11)$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

We determine the value of β_0 from equation (8) as:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} \quad \ldots (12)$$

or $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \ldots (13)$$

You should note that β_0 and β_1 are the least squares estimators of the intercept (β_0) and slope (β_1), respectively.

Thus, the fitted simple linear regression model is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad \ldots (14)$$

We can also write the fitted simple linear regression model in terms of n pairs of data, i.e., for $\{(x_i, y_i); i = 1, 2, ..., n\}$ as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i; \qquad i = 1, 2, \ldots, n \quad \ldots (15)$$

Since the numerator of equation (9) is the corrected sum of cross products of y_i and x_i while the denominator is the corrected sum of square of x_i, we may also rewrite equation (9) in a more compact notations as follows:

$$\hat{\beta}_1 = \frac{SS_{yx}}{SS_x} \quad \ldots (16)$$

where

$$SS_{yx} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i x_i - n\bar{y}\,\bar{x} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \quad \ldots (17)$$

and $$SS_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \ldots (18)$$

Now that you know how to estimate the regression coefficients, you should learn some properties of the fitted regression model.

5.3.1 Properties of Fitted Regression Model

We first need to define the term residual before discussing the properties of the fitted regression model. The **residual** is the difference between the observed and the corresponding predicted (fitted) values of the response variable. It provides the magnitude of the response variable that is not explained by the regression model. If y_i is the i^{th} observed value and $\hat{y}_i$ is the corresponding fitted value of the response variable Y, we can define the i^{th} residual as:

$$r_i = y_i - \hat{y}_i; \qquad i = 1, 2, ..., n \quad \ldots (19)$$

We now substitute the value of $\hat{y}_i$ from equation (15), i.e., $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ in equation (19) and get

$$r_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right); \qquad i = 1, 2, ..., n \quad \ldots (20)$$

Now we can consider some properties of the fitted regression model in accordance with the property of residuals:

1. The sum of the residuals for all given observations is always zero, i.e., $\sum_{i=1}^{n} r_i = 0$. So, we can also say that r_i is the i^{th} residual which has mean zero and variance σ^2.

2. The sum of the predicted values is always equal to the sum of the observed values of the response variable (Y), i.e., $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

3. The fitted least squares linear regression model always passes through the point $(\bar{y}, \bar{x})$, i.e., $\bar{y} = \beta_0 + \beta_1 \bar{x}$

The following example will help you learn how to fit a simple regression line numerically.

Example 2: Let us consider the data of Example 1 given in Sec.5.2, and

i) fit a linear regression model using the method of least squares;

ii) estimate the values of the systolic blood pressure for a woman of age 36 years;

iii) compute the residuals and verify that $\sum_{i=1}^{15} r_i = 0$; and

iv) draw the fitted regression line on a scatter plot.

Solution: (i) As per equation (14), the fitted simple regression model is given as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where X, the age is an a regressor variable and Y, the SBP is a response variable. To determine the regression model, we construct the following table:

Table 3: Computation of $\sum_{i=1}^{15} y_i, \sum_{i=1}^{15} x_i, \sum_{i=1}^{15} x_i^2$ and $\sum_{i=1}^{15} y_i x_i$

S. No	SBP (y_i)	Age (x_i)	x_i^2	$y_i x_i$	$\hat{y}_i$	Residual $r_i = (y_i - \hat{y}_i)$	$r_i^2 = (y_i - \hat{y}_i)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	124	30	900	3720	123.6098*	0.3902	0.1523
2	134	38	1444	5092	132.5610	1.4390	2.0708
3	135	39	1521	5265	133.6799	1.3201	1.7427
4	121	26	676	3146	119.1341	1.8659	3.4814
5	122	29	841	3538	122.4909	–0.4909	0.2409
6	119	27	729	3213	120.2530	–1.2530	1.5701
7	128	32	1024	4096	125.8476	2.1524	4.6330
8	118	25	625	2950	118.0152	–0.0152	0.0002
9	120	26	676	3120	119.1341	0.8659	0.7497
10	123	31	961	3813	124.7287	–1.7287	2.9883
11	129	37	1369	4773	131.4421	–2.4421	5.9637
12	117	25	625	2925	118.0152	–1.0152	1.0307
13	131	35	1225	4585	129.2043	1.7957	3.2247
14	126	34	1156	4284	128.0854	–2.0854	4.3488
15	134	40	1600	5360	134.7988	–0.7988	0.6381
Total	1881	474	15372	59880	1881	0	32.8354

From Columns 1 to 5 of Table 3, we have

$$n = 15,\quad \sum_{i=1}^{15} y_i = 1881,\ \sum_{i=1}^{15} x_i = 474,\ \sum_{i=1}^{15} x_i^2 = 15372 \text{ and } \sum_{i=1}^{15} y_i x_i = 59880$$

We can directly obtain the value of slope (β_1) after substituting the values of $n, \sum_{i=1}^{15} y_i, \sum_{i=1}^{15} x_i, \sum_{i=1}^{15} x_i^2$ and $\sum_{i=1}^{15} y_i x_i$ in equation (10) as:

$$\hat{\beta}_1 = \frac{15 \times 59880 - 1881 \times 474}{15 \times 15372 - (474)^2}$$

$$= \frac{898200 - 891594}{230580 - 224676}$$

$$= \frac{6606}{5904} = 1.1189$$

It is to be noted that all calculations were performed up to 15 fixed decimal places for showing accurate results in this block. For the sake of simplicity, we are showing results up to 4 decimal places only. The results may vary if we carry out the calculations by fixing values at various decimal places.

We calculate the value of intercept (β_0) using equation (12) as:

$$\hat{\beta}_0 = \frac{1881}{15} - 1.1189 \times \frac{474}{15}$$

$$= 125.4 - 1.11890 \times 31.6$$

$$= 125.4 - 35.3573$$

$$= 90.0427$$

After substituting the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (14), we obtain the best fitted regression model using the method of least squares as:

$$\hat{Y} = 90.0427 + 1.1189\ X$$

The intercept term as such has no physical interpretation like the slope term but it remains present in almost all fitted line unless it passes through the origin. If the intercept term is quite large for some data, it is advisable to 'centre' the data to remove the intercept without making it zero.

Interpretation: The value of slope is 1.1189 in the fitted regression model. It means that the average increase in SBP of women with unit change in age (after one year) is approximately 1.12 mm/Hg. This indicates that for every additional year in age you can expect SBP to increase by an average of 1.12 mm/Hg. We can also say that the SBP of a woman is expected to differ by the product of 1.12 with the difference between age. For example, after two years, SBP may increase by $2 \times 1.12 = 2.24$ mm/Hg.

(ii) We can compute the predicted SBP for the given age x = 36 using the fitted model in (i) as follows:

$$\text{predicted SBP}\,(\hat{y}) = 90.0427 + 1.1189\ x$$

$$= 90.0427 + 1.1189 \times 36$$

$$= 90.0427 + 40.2805$$

$$= 130.3232 \approx 130 \text{ mm/Hg}$$

The predicted value of SBP for a woman of 36 years of age is 130 mm/hg on the basis of the given data.

(iii) We now determine the fitted or predicted values, i.e., $\hat{y}_i$ for $i = 1, 2, ..., 15$ of the response variable (y) by substituting the given values of regressor variable (X) in the fitted simple linear regression model obtained in (i) as:

$$*\hat{y}_1 = 90.0427 + 1.1189 \times 30$$

$$= 123.6098$$

Similarly, we can calculate the other values of $\hat{y}_i (i = 2,3,...,n)$ as shown in Table 4. We also calculate the values of residuals using the formula given in equation (19) for all observations given in the data and arrange the predicted SBP and residuals in Columns 6 and 7, respectively, of Table 3.

From Column 7 of Table 3, the sum of the residuals is computed as $\sum_{i=1}^{15} r_i \cong 0$.

Note that it verifies the property of the residuals, i.e., the sum of the residuals is equal to zero.

(iv) If we plot the predicted SBP against age on a scatter diagram shown in Fig. 5.1, the line drawn by joining the plotted points of predicted SBP on the scatter plot is the fitted regression line as shown in Fig. 5.3.

The fitted regression model may not be absolutely correct. The difference between the observed and predicted values is known as the residual. For example, for the first value, the residual is

(124 – 123.6098) = 0.3902.

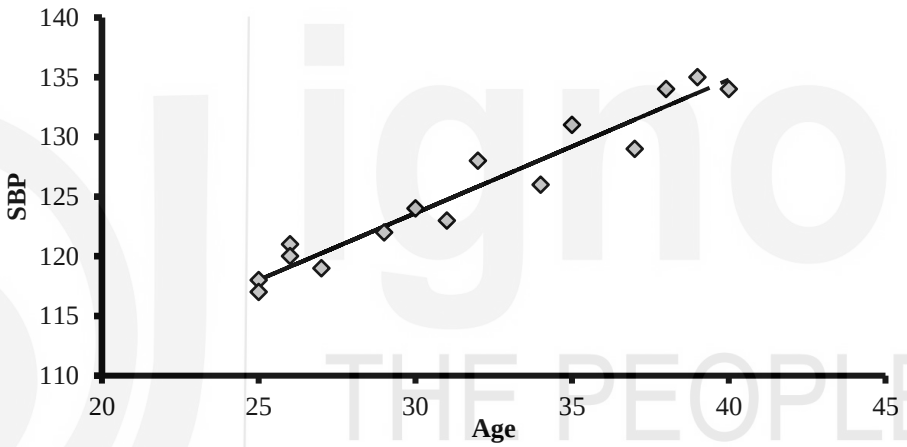


Fig. 5.3: Fitted regression line for Example 2.

You should now solve the following exercises for practice.

E4) For the data given in **E3**,

i) Fit an appropriate regression model to the given data.

ii) Estimate the sugar level of a man if he walks 58 minutes daily.

iii) Draw the fitted regression line on a scatter plot.

iv) Compute the residuals and verify that the sum of the residuals is zero.

E5) The following table provides data on birth weight (in kg) at various gestational age (in weeks) of 12 infants to check the effect of gestational age on birth weight:

S. No.	Birth Weight	Gestational Age
1	2.4	34.0
2	2.3	34.7
3	2.0	29.8
4	2.9	38.2
5	2.7	36.1
6	3.2	42.8

7	3.4	40.8
8	2.8	37.8
9	3.2	38.4
10	3.7	41.3
11	4.0	42.0
12	3.4	40.1

i) Determine the best fitted regression model to the given data.

ii) Predict the birth weight if the gestational age is 40 weeks.

iii) Compute the residuals and verify that $\sum_{i=1}^{15} r_i = 0$.

iv) Draw the fitted regression line on a scatter plot.

We now discuss the properties of the estimated regression coefficients.

5.4 PROPERTIES OF ESTIMATED REGRESSION COEFFICIENTS

In Sec. 5.3, we have obtained the least squares estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) of β_0 and β_1. In this section, we discuss some properties of the least squares estimators of the intercept and slope, respectively.

1. The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimators of the β_0 and β_1, respectively, i.e.,

$$E(\hat{\beta}_0) = \beta_0 \quad \text{... (21)}$$

and $$E(\hat{\beta}_1) = \beta_1 \quad \text{... (22)}$$

2. The variance of the estimated intercept and slope are given as:

$$V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) \quad \text{... (23)}$$

and $$V(\hat{\beta}_1) = \frac{\sigma^2}{SS_x} \quad \text{... (24)}$$

where $SS_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$

But, in general, the value of σ^2 is unknown. So, we use the estimated value of σ^2, i.e., $\hat{\sigma}^2$. We estimate it on the basis of the given data as explained below.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \quad \text{... (25)}$$

where $\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} = 0$

The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ using estimated value of σ^2 can be determined as:

$$V(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) \quad \text{... (26)}$$

and $$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SS_x} \quad \text{... (27)}$$

When σ^2 is unknown, we obtain the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ as:

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)} \quad \text{... (28)}$$

and $$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{SS_x}} \quad \text{… (29)}$$

We now give an example to compute the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ for given data.

Example 3: For the SBP data given in Example 3, compute $\hat{\sigma}^2$, $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.

Solution: From the solution of Example 2, we have

$$n = 15,\ \sum_{i=1}^{15} y_i = 1881,\ \sum_{i=1}^{15} x_i = 474,\ \sum_{i=1}^{15} y_i^2 = 236403, \text{and } \sum_{i=1}^{15} x_i^2 = 15372$$

We have also computed the regression coefficients $(\beta_0 \text{ and } \beta_1)$ as:

$$\hat{\beta}_0 = 90.04268293 \text{ and } \hat{\beta}_1 = 1.11890244$$

Let us now determine the values of $\bar{x}$ and SS_x:

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{15} = 31.6$$

$$SS_x = \sum_{i=1}^{15} x_i^2 - \frac{\left(\sum_{i=1}^{15} x_i\right)^2}{15} = 15372 - \frac{(474)^2}{15} = 15372 - \frac{224676}{15}$$

$$= 15372 - 14978.4 = 393.6$$

It is to be noted that all calculations were performed up to 15 fixed decimal places for showing accurate results in this block. For the sake of simplicity, we are showing results up to 4 decimal places only. The results may vary if we carry out the calculations by fixing values at various decimal places.

From Column 8 of Table 3, we obtain

$$\sum_{i=1}^{15} (y_i - \hat{y}_i)^2 = 32.8354$$

From equation (25), we compute $\hat{\sigma}^2$ as:

$$\hat{\sigma}^2 = \frac{1}{13}\sum_{i=1}^{15} (y_i - \hat{y}_i)^2 = \frac{32.8354}{13} = 2.5258$$

We now compute the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ using equations (26) and (27), respectively, as:

$$V(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) = 2.5258 \times \left(\frac{1}{15} + \frac{(31.6)^2}{393.6}\right)$$

$$= 2.5258 \times [0.0667 + 2.5370]$$

$$= 2.5258 \times 2.6037 = 6.5763$$

and $$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SS_x} = \frac{2.5258}{393.6} = 0.0064$$

The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, are determined by

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{6.5763} = 2.5644$$

$$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{0.0064} = 0.0801$$

Before ending the unit, you may like to solve some exercises.

E6) Explain what is meant by expected values and standard errors of the regression coefficients.

E7) For the data given in **E3** of this unit, determine the estimated variance of error terms and standard errors of regression coefficients.

E8) For the data given in **E5** of this unit, obtain $\hat{\sigma}^2$, $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.

So far, you have learnt how to fit a simple linear regression model using the method of least squares. In the next unit, we shall explain statistical inference in simple linear regression.

We now end this unit by giving a summary of what you have learnt in it.

5.5 SUMMARY

1. Regression analysis is concerned with the study of the dependence of one variable (response variable) on another variable (regressor variable). The focus of such study is to predict the value of response variable using a known value of the regressor variable.
2. If the values of regressor variable and corresponding response variable, when plotted on a scatter diagram, are approximately scattered around a straight line, this relationship is said to be linear. In such a situation, we apply simple linear regression method to study the relationship between the variables. However, if the relationship is not linear, we use transformation or non-linear regression approach.
3. If the regressor variable is denoted by X and response variable is denoted by Y, the linear relationship of Y on X can be defined with the help of a simple linear regression model as:

   $$Y = \beta_0 + \beta_1 X + \varepsilon$$

4. We use the method of least squares to estimate β_0 and β_1 so that the sum of the squares of differences between the observed value (y_i) and the estimated value ($\hat{y}_i$) lying on the straight line is minimum. The values of β_1 and β_0 are given as:

$$\hat{\beta}_1 = \frac{SS_{yx}}{SS_x} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

5. The fitted simple linear regression model is given as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

6. The difference between the observed and corresponding predicted (fitted) values of the response variable is known as residual. It provides the magnitude of the response variable that is not explained by the regression model.

7. The value of σ^2, the variance of ε, is generally unknown. Hence, we estimate it from the given data as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}, \text{ where } \bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} = 0$$

8. The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimators of the parameters β_0 and β_1, respectively, i.e., $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

9. The variances of the estimated intercept and slope are given as:

$$V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) \text{ and } V(\hat{\beta}_1) = \frac{\sigma^2}{SS_x}$$

5.6 SOLUTIONS / ANSWERS

E1) Refer to Sec. 5.2.1.

E2) Refer to Sec. 5.2.

E3) Since the blood sugar level depends upon duration of walking, the response variable is random blood sugar level and regressor variable is duration of walking.

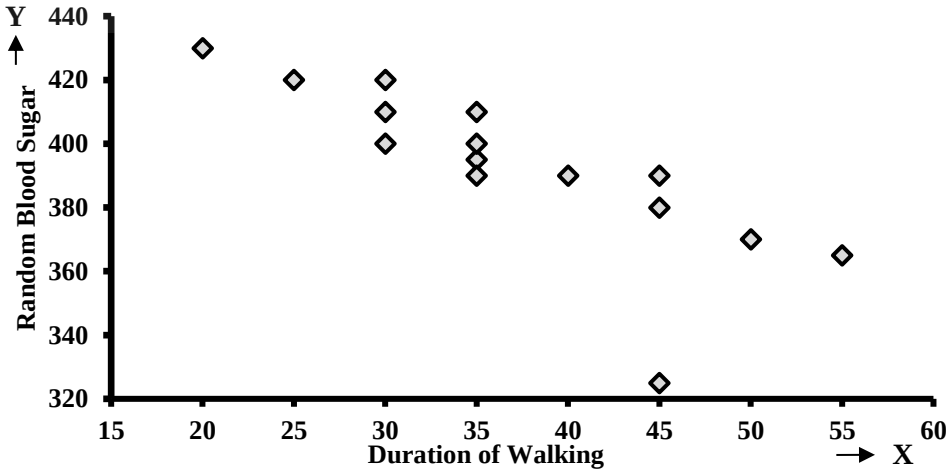


Fig. 5.4: Fitted regression line for E3.

This scatter diagram shows a linear relationship between blood sugar level and duration of walking. Hence, we can conclude that the blood sugar level decreases as duration of walking increases and vice-versa.

E4) We have

$$n = 15,\ \sum_{i=1}^{15} y_i = 5895,\ \sum_{i=1}^{15} x_i = 555,\ \sum_{i=1}^{15} x_i^2 = 21825 \text{ and}$$

$$\sum_{i=1}^{15} y_i x_i = 215275$$

We compute the value of slope (β_1) using equation (11) as:

$$\hat{\beta}_1 = \frac{15 \times 215275 - 5895 \times 555}{15 \times 21825 - (555)^2}$$

$$= -2.2016$$

We calculate the value of intercept (β_0) using equation (8) as:

$$\hat{\beta}_0 = \frac{5895}{15} - (-2.2016) \times \frac{555}{15}$$

$$= 474.4574$$

After substituting the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (13), the best fitted regression model is $\hat{Y} = 474.4574 - 2.2016X$

The value of slope is – 2.2016. It means that the average sugar level decreases approximately 2.2 times with unit increase in duration of walking.

The predicted sugar level of a man who walks 58 minutes daily can be calculated using the fitted regression model as:

$$\hat{y} = 474.4574 - 2.2016 \times 58$$

$$= 346.7674$$

The predicted sugar level against the given durations of walking can be shown in the following scatter diagram:

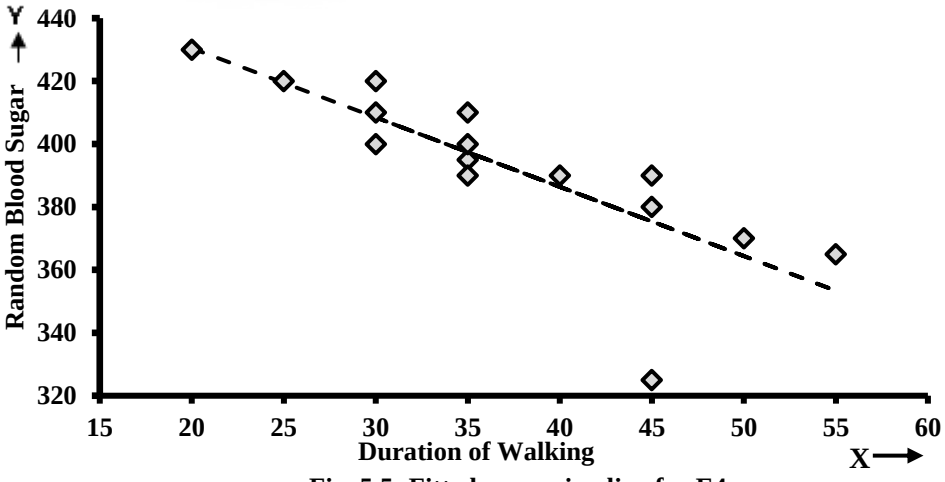


Fig. 5.5: Fitted regression line for E4.

E5) Since birth weight depends upon the gestational age, the gestational age is an independent variable whereas birth weight is a dependent variable. Therefore, we have

$$n = 12,\ \sum_{i=1}^{12} y_i = 36,\ \sum_{i=1}^{12} x_i = 456,\ \sum_{i=1}^{12} x_i^2 = 17488.16 \text{ and}$$

$$\sum_{i=1}^{12} y_i x_i = 1390.81$$

We can obtain the value of slope (β_1) using equation (11) as:

$$\hat{\beta}_1 = \frac{12 \times 1390.81 - 36 \times 456}{12 \times 17488.16 - (456)^2}$$

$$= 0.1424$$

We calculate the value of intercept (β_0) using equation (8) as:

$$\hat{\beta}_0 = \frac{36}{12} - 0.1424 \times \frac{456}{12}$$

$$= -2.4120$$

After substituting the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (13), the best fitted regression model is given by

$$\hat{Y} = -2.4120 + 0.1424\,X$$

The value of slope is 0.1424. It means that the average birth weight increases with unit increase in gestational age is 0.1424.

The predicted birth weight for the gestational age of 40 weeks can be calculated through the fitted regression model as:

$$\hat{y} = -2.4120 + 0.1424 \times 40$$

$$= 3.2848$$

The predicted birth weights against the given gestational ages can be shown in a scatter diagram (Fig. 5.6).

Fig. 5.6: Fitted regression line for E5.

E6) Refer to Sec. 5.4.

E7) From the solution of **E4**, we have

$$n = 15,\ \sum_{i=1}^{15} y_i = 5895,\ \sum_{i=1}^{15} x_i = 555,\ \sum_{i=1}^{15} x_i^2 = 21825 \text{ and}$$

$$\sum_{i=1}^{15} y_i x_i = 215275$$

The fitted regression model is: $\hat{Y} = 474.4574 - 2.2016\,X$

We now determine the values of $\bar{x}$ and SS_x as:

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{555}{15} = 37$$

The value of $(y_i - \hat{y}_i)^2$ is obtained as:

$$\sum_{i=1}^{15} (y_i - \hat{y}_i)^2 = 3387.5969$$

From equation (25), we have $\hat{\sigma}^2 = 260.5844$

$$SS_x = \sum_{i=1}^{15} x_i^2 - \frac{\left(\sum_{i=1}^{15} x_i\right)^2}{15}$$

$$= 21825 - \frac{(555)^2}{15}$$

$$= 21825 - \frac{308025}{15}$$

$$= 21825 - 20535 = 1290$$

We now compute the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ using equations (26) and (27), respectively, as:

$$V(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)$$

$$= 260.5844 \times \left(\frac{1}{15} + \frac{(37)^2}{1290}\right)$$

$$= 293.9149$$

and $$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SS_x} = \frac{260.5844}{1290} = 0.2020$$

The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, are determined by

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)}$$

$$= \sqrt{293.9149} = 17.1439$$

$$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)}$$

$$= \sqrt{0.2020} = 0.4494$$

E8) From the solution of **E5**, we have

$n = 12,\ \sum_{i=1}^{12} y_i = 36,\ \sum_{i=1}^{12} x_i = 456,\ \sum_{i=1}^{12} x_i^2 = 17488.16$ and

$\sum_{i=1}^{12} y_i x_i = 1390.81$

The fitted regression model is: $\hat{Y} = -2.4120 + 0.1424X$

We now determine the values of $\bar{x}$ and SS_x as:

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{36}{12} = 3$$

$$SS_x = \sum_{i=1}^{12} x_i^2 - \frac{\left(\sum_{i=1}^{12} x_i\right)^2}{12}$$

$$= 17488.16 - \frac{(456)^2}{12}$$

$$= 17488.16 - 17328 = 160.16$$

$$\hat{\sigma}^2 = 0.0631$$

We now compute the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ using equations (26) and (27), respectively, as:

$$V(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)$$

$$= 0.0631 \times \left(\frac{1}{12} + \frac{(3)^2}{160.16}\right)$$

$$= 0.5745$$

and $$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SS_x} = \frac{0.0631}{160.16} = 0.0004$$

The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, are determined by

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)}$$

$$= \sqrt{0.5745} = 0.7580$$

$$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)}$$

$$= \sqrt{0.0004} = 0.0199$$