UNIT 9 REGRESSION ANALYSIS

Structure **Page No.**

9.1 INTRODUCTION

Decision making is an important activity of everybody's life. It requires knowledge and information. For example, for a farmer to decide whether to grow paddy or not in a particular field, he should know the soil conditions, availability of water, etc. An experienced farmer will be able to predict the yield of paddy by examining the soil features. In this example, the yield of paddy, y, is *associated* with the fertility of the soil, x. It would be nice if there is a reasonably good mathematical relationship that can be used to predict the yield, y, at least approximately, for a given value of x, so that the farmer can work out the yield and decide whether it is worth growing paddy or not. *Regression Analysis* is a powerful statistical technique useful in building such mathematical relationships.

To make you appreciate the subject, here are some interesting practical applications of the technique. These are compiled from project studies carried out in Indian industries.

An Application in a Sugar Factory

The yield of sugar depends on the time at which the sugar cane is harvested. During the harvesting period, the yield of sugar that can be obtained from the sugar cane gradually increases with time up to a certain period and starts falling beyond that period. To get maximum yield, the sugar factory has to decide when to harvest the crops. This decision making requires the relationship between the amount of yield and the time. Using past data, a regression equation was developed which formed the basis for arriving at a decision making procedure. This project was carried out in a sugar factory situated in Andhra Pradesh. As a result of this project, the factory's profits went up by several lakhs of rupees by way of increased turnover.

An Application in an Electronic Industry

In one semiconductor manufacturing unit, about 41% of the transistors produced were getting rejected due to various faults. Rejection data analysis revealed that 39% of the 41% rejections were due to low *β-value* while the remaining 2% were caused by other faults. The β-value, which should be between 30 and 100 according to specifications, was influenced by three process parameters x_1, x_2, x_3 (these are resistances at certain positions). It was possible to control these process parameters at desired levels. From the past data, the following formula was obtained using regression analysis.

$$\beta - \text{value} = 32.67 + 0.64x_1 - 1.18x_2 - 20.98x_3.$$

Using this equation, the levels of x_1, x_2, x_3, which result in increased β-values, were identified and maintained in the process. As a result of this study, the rejections in transistors came down from 41% to 20%.

Prediction of Tensile Strength of Castings

For a particular grade of castings produced by a foundry, customers' specifications were in terms of hardness and tensile strength. While the hardness and chemical composition were being tested in the foundry, tensile strength was being tested by an outside laboratory. The results used to come after a time lag of 45 days and as such were useless from the view point of process control. Though it was known that tensile strength depends on the chemical composition and hardness, no relationship between these variables was available. Using the past data, the following regression equation was developed.

$$y = 19.366 - 8.359x_1 + 2.854x_2 + 10.675x_3 + 0.096x_4,$$

where y is tensile strength, x_1, x_2, x_3 are carbon, silicon and manganese percentages respectively, and x_4 is hardness. This prediction formula was discussed with foundry management who decided to use it for day to day process control.

The term regression was first used as a statistical concept by an English Statistician Sir Francis Galton in 1877. Galton made a study that showed that the heights of the children born to unusually tall or short parents tends to move back or "regress" towards the mean height of the population

Thus, you have seen three applications of regression analysis technique. It is one of the most widely applied statistical techniques. In this unit, you will learn this technique through illustrative examples. Through a simple example, we will learn how to build a simple *linear regression equation* to predict a given variable y based on another associated variable x. We will look at some simple methods to examine how good such an equation is. After this, we will consider the situations where we have two associated variables and learn how to build the formula and assess it. For example, you can think of predicting yield of paddy y, based on two associated variables: (i) x, the soil fertility and (ii) u, the quantity of chemical fertiliser applied. Finally, we will consider the case where three associated variables are involved.

Objectives

After reading this unit, you should be able to

- identify linear relationship between two variables through scatter diagram (a graphical aid),
- specify the simple linear regression model and build it from data,
- compute the coefficient of correlation and evaluate the regression formula through this coefficient,
- specify multiple linear regression models with two independent variables and build them from data,

9.2 SIMPLE LINEAR REGRESSION MODEL

In this section, we will study the simple linear regression model. Many of the concepts that you learn here will be useful when we deal with multiple regression models in Section 9.4.

Let us start with an example.

When the floor of a house is plastered with cement, you know that one should not walk on it until it gets set. The setting time is an important characteristic of cement and it is mandatory that any cement manufacturing company should adhere to certain specification limits. Once the cement is produced, its setting time cannot be changed. So the manufacturers should be in a position to predict the setting time well before it is produced. Fortunately, it is possible to do this because the setting time depends, to a large extent, on the chemical composition of the raw materials used to produce cement.

In Table-1, you will find data on 25 samples of cement. For each sample, we have a pair of observations (x, y), where x is percentage of SO_3, a chemical, and y is the setting time in minutes. Our aim is to study how y depends on x. We will refer to y as the **dependent variable** or **response**, and x as **independent variable** or **regressor**. You know that it is often easy to understand data through a graph. So, let us plot the data on a *scatter diagram*. A scatter diagram is a simple two-dimensional graph in which the horizontal axis represents x and the vertical axis represents y. Each pair of points is plotted on the graph. See the graph in Figure 1 given in the next page.

Table-1: Data on SO_3 and Setting Time

S. No. i	Percentage of SO_3 x	Setting Time y (in minutes)
1	1.84	190
2	1.91	192
3	1.90	210
4	1.66	194
5	1.48	170
6	1.26	160
7	1.21	143
8	1.32	164
9	2.11	200
10	0.94	136
11	2.25	206
12	0.96	138
13	1.71	185
14	2.35	210
15	1.65	178
16	1.19	170
17	1.56	160
18	1.53	160
19	0.96	140
20	1.67	168
21	1.68	152
22	1.28	160
23	1.35	116
24	1.49	145
25	1.78	170
Total	39.04	4217
Sum of Squares	64.446	726539

From this figure, you see that y increases as SO_3 increases. Whenever you find this type of increasing (or decreasing) trend in a scatter diagram, it indicates that there is a linear relationship between x and y. You may observe that the relationship is not perfect in the sense that a straight line cannot be drawn through all the points in the scatter diagram. Nevertheless, we may approximate it with some linear equation. What formula shall we use? Suppose we use the formula $y = 90 + 50x$ to predict y based on x. To examine how good this formula is, we need to compare the actual values of y with the corresponding predicted values. When $x = 0.96$, the predicted y is equal to $138 (= 90 + 50 \times 0.96)$. Let (x_i, y_i) denote the values of (x, y) for the ith sample. From Table-1, notice that $x_{12} = x_{19} = 0.96$ where as $y_{12} = 138$ and $y_{19} = 140$.

Let $\hat{y}_i = 90 + 50x_i$. That is, $\hat{y}_i$ is the predicted value of y (using $y = 90 + 50x$) for the ith sample. Since, $x_{12} = x_{19} = 0.96$, both $\hat{y}_{12}$ and $\hat{y}_{19}$ are equal to 138. The difference $\hat{e}_i = y_i - \hat{y}_i$, the **error in prediction**, is called the **residual**. Observe that $\hat{e}_{12} = 0$ and $\hat{e}_{19} = 2$. The formula we have considered above, $y = 90 + 50x$, is called a **simple**

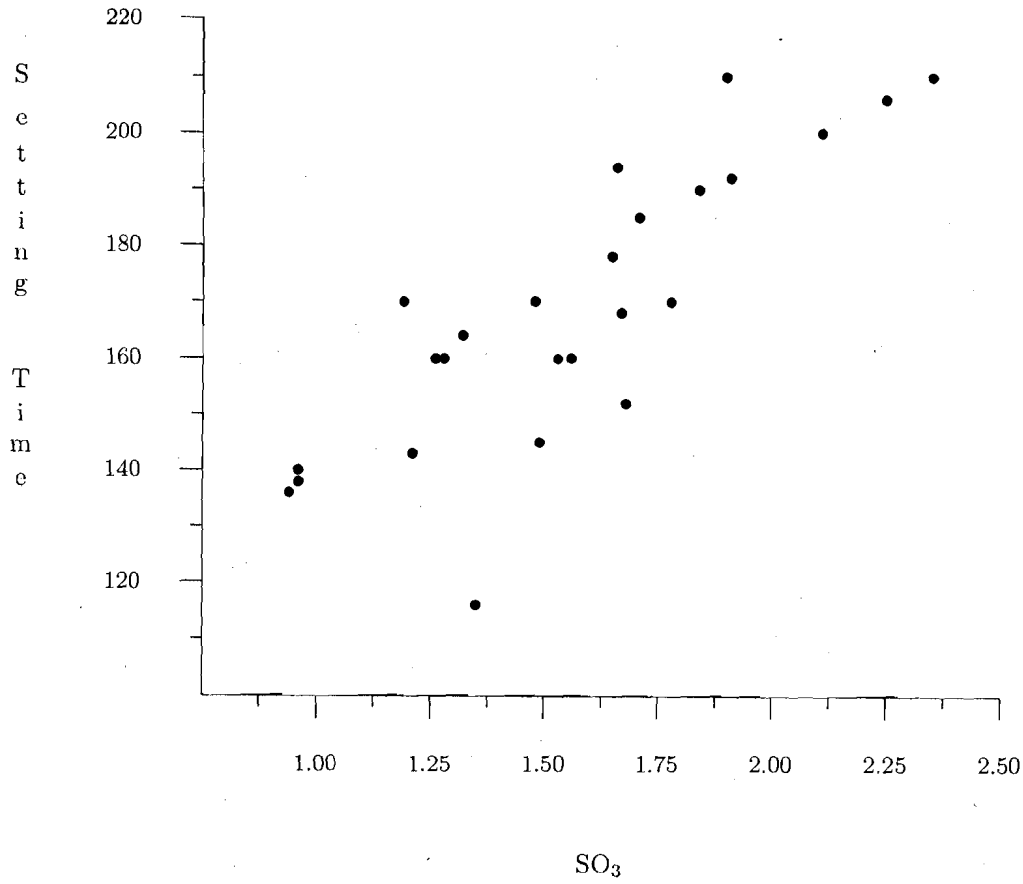


Fig. 1: Scatter Diagram of Setting Time vs SO_3

linear regression equation. If you recall from your study of coordinate geometry, this equation is the equation of a straight line. You can easily see that $\hat{y}_i$ and $\hat{e}_i$ depend on the straight line we choose to predict y. That is, if we change the equation, let us say, from $y = 90 + 50x$ to $y = 100 + 45x$, then the $\hat{y}_i$ and $\hat{e}_i$ will be different. To fix these ideas in your mind, try these exercises now.

E1) Using the regression equation $y = 90 + 50x$, fill up the values in the table below.

Table-2: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138				
$\hat{e}_i$	0				

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

E2) Using $y = 100 + 45x$, fill up the values in the table below.

Table-3: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	143.2				
$\hat{e}_i$	-5.2				

Note: $\hat{y}_i = 100 + 40x$ and $\hat{e}_i = y_i - \hat{y}_i$

E3) Compare the $\hat{y}_i$ values of Table-2 with the corresponding ones in Table-3. Do this comparison with respect to $\hat{e}_i$s also.

From the exercises above, you might have noticed that different equations give us different residuals. What will be the best equation? Obviously, the choice will be that equation for which $\hat{e}_i$s are small.

From samples 12 and 19, we have found that for the same value of $x = 0.96$, the y values are different. This means that whatever straight line we use, it is not possible to make all the $\hat{e}_i$s zero. However, we would expect that the errors are positive in some cases and negative in the other cases so that, on the whole, their sum is close to zero. So, our job is to find out the *best* values of a and b in the formula $y = a + bx$. Let us see how we do this.

Now our aim is to find the values a and b so that the error e_is are minimum. For that we state here four steps to be done.

1) Calculate a sum S_{xx} defined by

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{1}$$

where x_i's are the given values of the data and $\bar{x} = \frac{\sum x_i}{n}$ is the mean of the observed values and n is the sample size.

The sum S_{xx} is called the corrected sum of squares.

2) Calculate a sum S_{xy} defined by

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} \tag{2}$$

where x_i's and y_i's are the x-values and y-values given by the data and $\bar{x}$ and $\bar{y}$ are their means.

3) Calculate $\frac{S_{xy}}{S_{xx}} = b$, say. That is

$$b = \frac{S_{xy}}{S_{xx}} \tag{3}$$

4) Find $\bar{y} - b\bar{x} = a$, say.

Let us now compute these values for the data in Table-1, we get

$$\bar{x} = 1.5616, \bar{y} = 168.68, S_{xx} = 3.4811, \text{ and } S_{xy} = 191.2328.$$

Substituting these values in (3) and (4), we get

$$b = \frac{S_{xy}}{S_{xx}} = 54.943 \text{ and } a = 168.68 - 54.943 \times 1.5616 = 82.88. \tag{4}$$

Therefore, the best linear prediction formula is given by

$$y = 82.88 + 54.943x. \tag{5}$$

So far, we have come across three prediction formulae, namely, $y = 90 + 50x$, $y = 100 + 45x$ and $y = 82.88 + 54.943x$. While the first two of these were proposed in an ad hoc manner, the third one was obtained objectively *using data*. Let us draw these three lines on the scatter diagram and see how the picture looks like. From Figure 2 on page number 28, you can see that compared to straight lines in (a) and (b), the one in (c) is close to more points.

In the exercise below, we have another set of observations on (x, y). Now do this exercise to make sure that you have understood the procedure to get best linear regression formula.

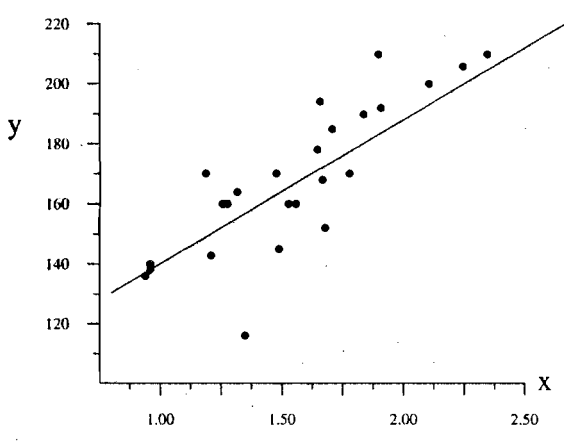
E4) Work out the best linear regression formula from the following data.

Table-4: Data on SO_3 and Setting Time

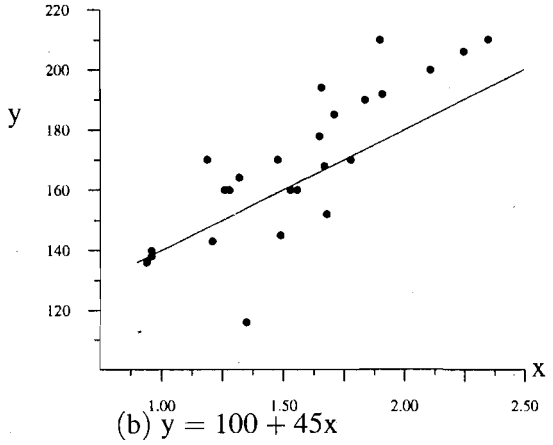
S. No. (i)	Percentage SO_3 (x)	Setting Time y (in minutes)	S. No. (i)	% SO_3 (x)	Setting Time y (in minutes)
1	2.25	206	9	1.67	168
2	0.96	138	10	1.28	160
3	1.71	185	11	1.78	170
4	2.35	210	12	0.78	130
5	1.65	178	13	1.26	146
6	1.56	160	14	1.50	180
7	1.53	160	15	1.35	148
8	0.96	140			

$\sum x_i = 22.59,\ \sum x_i^2 = 36.7135,\ \sum y_i = 2479,\ \sum y_i^2 = 511637.$

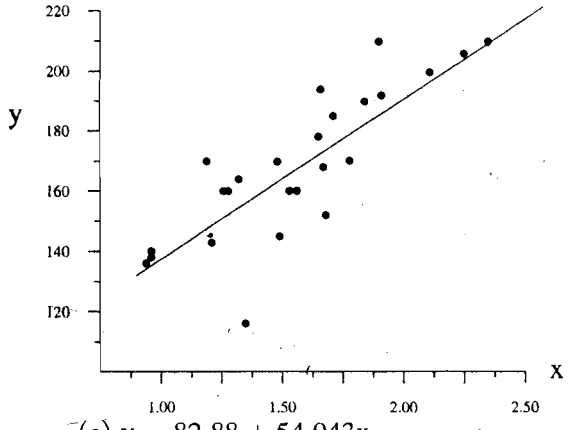
If you have done the above exercise, you must have got y = 88.05 + 51.273 x as your best linear regression formula. Compare this with what we have got earlier (see Eqn. (5)). Have you noticed that the values of a and b have changed. Well, this is bound to happen because the set of observations we have used there was different from the one



(a) y = 90 + 50x



(b) y = 100 + 45x



(c) y = 82.88 + 54.943x

Fig. 2: Three Different Regression Lines For Setting Time Data

we have used in (E4). The point that you should learn from this exercise is that the values of a and b depend on the sample of observations. Therefore, it would have been proper for us to have used the symbols $\hat{a}$ and $\hat{b}$ in Eqns.(4) and (6) to indicate that these are the estimated values obtained from the sample of observations.

By now you must have understood how to calculate the best linear regression line and how will we use it to make some predictions about the problems that are analysed. Let us see an example.

Problem 1: : A hosiery mill wants to estimate how its monthly costs are related to its monthly output rate. For that the firm collects a data regarding its costs and output for a sample of nine months as given in Table 5 below.

Table 5

Output (tons)	Production cost (thousands of dollars)
1	2
2	3
4	4
8	7
6	6
5	5
8	8
9	8
7	6

1) Construct a scatter diagram for the data given above.
2) Calculate the best linear regression line, where the monthly output is the dependent variable and the monthly cost is the independent variable.
3) Use this regression line to predict the firm's monthly costs if they decide to produce 4 tons per month.

Solution:

1) Suppose that x_i denote the output for the ith month and y_i denote the cost for the ith month. Then we can plot the graph for the pair (x_i, y_i) of the values given in Table 5. Then we get the scatter diagram as shown in Fig. 3 in the next page.
2) Now to find the least square regression line, we first calculate the sums S_{xx} and S_{xy} from Eqn.(1) and (2).

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}$$

Note that from Table (5) we get that

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{9} \\ \bar{y} &= \frac{\sum_{i=1}^{n} y_i}{n} = \frac{49}{9} \\ \sum x_i^2 &= 340 \\ \sum y_i^2 &= 303 \\ \text{and } \sum x_i y_i &= 319 \end{aligned}$$

Therefore we get that

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

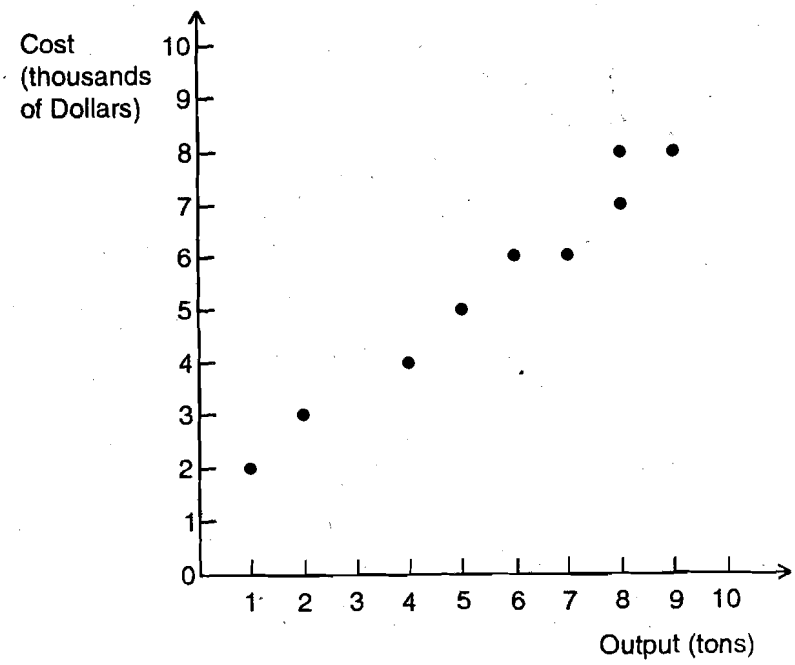


Fig. 3: Scatter Diagram

$$= \frac{9 \times 319 - 50 \times 49}{9 \times 340 - 50^2}$$

$$= \frac{421}{560} = 0.752$$

Correspondingly, we get

$$\hat{a} = \frac{49}{9} - (0.752) \times \frac{50}{9}$$

$$= 1.266$$

Therefore the best linear regression line is

$$y = 1.266 + (0.752)x$$

3) If the firms decides to produce 4 tons per month, then one can predict that its cost would be

$$1.266 + (0.752) \times 4 = 4.274$$

Since the costs are measured in thousands of dollars, this means that the total costs would be expected to be \$4, 274.

———— ×————

The above example illustrates that the regression line can be of great practical importance. You will be more clear about this, when you try the following exercise.

E5) An economist wants to estimate the relationship in a small community between a family's annual income and the amount that the family saves. The following data from nine families are obtained:

Annual income (thousands of dollars)	Annual savings (thousands of dollars)
12	0.0
13	0.1
14	0.2
15	0.2
16	0.5
17	0.5
18	0.6
19	0.7
20	0.8

Calculate the least-squares regression line, where annual savings is the dependent variable and annual income is the independent variable, and interpret your results.

In the next section we shall consider some procedures to measure how good our best fit is.

9.3 MEASURES OF GOODNESS OF FIT

You have seen that regression line provides estimates of the dependent variable for a given value of the independent variable. The regression line is called the line of best fit. It shows the relationship between x and y better than any other line.

Let us consider the question "How good is our best linear regression formula?" We would like to able to measure how good our best fit is. More precisely we want to have some measures for goodness of fit.

To develop a measure of goodness of fit, we first examine the variation in y. Let us try to examine the variation in the response y. Since y depends on x, if we change x, then y also changes. In other words, a part of variation in y_is is accounted by the variation in x_is. Actually, we can mathematically show that the total variation in y_is can be split up as follows:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}} + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{6}$$

Dividing this equation by S_{yy} on both sides, we get

$$1 = \frac{S_{xy}^2}{S_{xx}S_{yy}} + \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}} \tag{7}$$

Since the quantities on the right hand side are both nonnegative, none of them can exceed one. Also, if one of the two is closer to one, then the other has to be closer to zero. Let us use the notation

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{8}$$

Then, from what we have just now argued, $R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$ must be between 0 and 1. This, in turn, will imply that R will be between -1 and +1.

Supposing $R^2 = 1$, then from (7) we see that $\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}} = 0$ or $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$

or $y_i = \hat{y}_i$ for all i. Again, when R^2 is close to 1, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is close to zero. R **is called the correlation coefficient between** x **and** y. When R is negative, it means that y decreases as x increases; and when R is positive, y increases as x increases. Thus, R gives a measure of the strength of the relationship between the variables x and y.

Let us compute R and R^2 for the data in Table-1.

$$R = \frac{191.2328}{\sqrt{3.4807 \times 15216.7776}} = 0.8309 \text{ and } R^2 = 0.6904.$$

By now, you must have understood that the large R^2 (i.e., closer to 1) should mean that our regression formula is more reliable in the sense that predicted values from such a regression equation will be close to the actual values.

Let us now look at the value of the correlation coefficient R obtained for the data in Table 1. Since, in this case, $R^2 = 0.69$ is not zero, we can conclude on the basis of our sample data that there is a relationship between x and y.
Now the question is: How large should R^2 be. We can answer this question by carrying out a statistical test. But before we go onto that, try this exercise.

E6) For the data in Table 2, compute R and R^2. Compare this with the above result.

In order to carry out statistical tests, we shall assume that e_is are independent and normally distributed. Coming to the question of what values of R^2 can be considered as large, it is logical to compare R^2 with $1 - R^2$ (in the light of the fact that R^2 can be at most equal to 1). In fact, it has been shown that $\frac{(n-2)R^2}{1-R^2}$ follows F-distribution with 1 and $(n-2)$ degrees of freedom (df). For the data in Table-1, this value is given by

$$\frac{(n-2)R^2}{(1-R^2)} = \frac{23 \times 0.6904}{1-0.6904} = 51.29.$$

From the table of F-distribution given in the Appendix, we get that the tabulated F-value at 5% level of significance (i.e.$\alpha = 0.05$) with 1 and 23 df is equal to 4.28. Therefore in this example, R^2 is significantly large.

So far in this unit we have considered only one independent variable and a dependent variables. Many situations require the use of more than one independent variable to explain the values that the dependent variable may take. In the next section we shall discuss this.

9.4 MULTIPLE LINEAR REGRESSION

In the previous section we considered problems where the response y depended on a single independent variable x. But, often responses depend on several independent variables. For example, the yield of paddy not only depends on soil fertility but also on the amount of water used, the quantity of chemical fertilisers applied and so on. In this section we will consider problems in which response depends on 2 or 3 independent variables. Many of the ideas and concepts that we have learnt in simple linear regression model are extended to multiple linear regression as well. The only change is that the computations will be a little more involved. Here we shall explain how the techniques are extended in the multiple case.

First we shall give some preliminaries for formulating the regression line.

9.4.1 Preliminaries

We shall learn the multiple regression technique with the help of a study that was carried out in a spring manufacturing company in Bangalore a few years ago.

Example 1: One of the types of springs manufactured by the company is known as flat type springs which are used in textile machines. These springs are produced in batches and after production a spring is selected at random from each batch and is subjected to vibration test. The spring has to withstand 1800 vibrations, failing which the batch will

be sent for rework. At one time the company was facing severe rework problem. Most of the batches were getting rejected in the vibration test, and hence, were being reworked.

Tempering is one of the operations in the manufacture of springs in which springs are heated to a particular temperature, known as tempering temperature, and are kept at that temperature for a specified amount of time, known as tempering time. After tempering the springs are soaked in a chemical solution for a specified amount of time which is known as soaking time. We shall denote the tempering temperature by x_1, tempering time by x_2 and soaking time by x_3.

After preliminary investigations it was found out that there was lack of control in x_1, x_2 and x_3. It was not clear whether the variation in these variables was causing failures in the vibration test. To study the problem, data were collected for 15 batches, and for each batch the values of x_1, x_2, x_3 and y, the number of vibrations (in hundreds) withstood by the spring selected at random from the batch. These data are presented in Table-6.

Table-6: Flat Springs Data

Sl. No.	Tempering		Soaking Time (x_3)	Number of Vibrations (in 100s) (y)
	Temperature (x_1)	Time (x_2)		
1	321	59	26	19.01
2	339	69	24	12.66
3	334	63	27	17.45
4	329	70	20	12.32
5	325	58	23	19.15
6	331	56	24	19.09
7	324	61	27	17.98
8	321	69	29	18.50
9	337	59	22	17.48
10	335	68	20	12.93
11	320	69	28	16.80
12	331	64	22	11.79
13	329	67	30	20.53
14	336	60	21	14.32
15	339	58	25	19.21

We shall now analyse the data and see how we will resolve the problem of high rework. In this process, we shall learn the multiple linear regression technique with two and three independent variables. To ease your learning process, it is better to compute certain quantities before going into multiple regression. Let us know what these quantities are and how to compute them. In Table-6 we have 15 observations on (x_1, x_2, x_3, y). The i^{th} observation on x_1 will be denoted by x_{1i}. Similar notation will be used for x_2, x_3 and y. Each of the 15 observations on (x_1, x_2, x_3, y) is called a data point. Thus, (325, 58, 23, 19.15) is the fifth data point.

The corrected sum of products between the variables x_1 and x_2 is denoted by S_{12} and is defined by

$$S_{12} = \sum_{i=1}^{15} x_{1i}x_{2i} - \frac{(\sum_{i=1}^{15} x_{1i})(\sum_{i=2}^{15} x_{2i})}{15} \qquad (9)$$

The **corrected sum of products** between the variables x_i and y is denoted by S_{1y} and is defined as

$$S_{1y} = \sum_{i=1}^{15} x_{1i}y_i - \frac{(\sum_{i=1}^{15} x_{1i})(\sum_{i=2}^{15} y_i)}{15}$$

The **corrected sum of squares** of x_1 is denoted by S_{11} and is defined by

$$S_{11} = \sum_{i=1}^{15} x_{1i}^2 - \frac{(\sum_{i=1}^{15} x_{1i})^2}{15}$$

* * *

Note that when only one variable is involved, we call it sum of squares, and if two variables are involved, we call it sum of products. It will be easy for you to do the below exercise.

E7) Write down the notation for the following and define them for the data given in Table 6.

(i) corrected sum of products between x_2 and x_3

(ii) corrected sum of products between x_3 and y

(iii) corrected sum of squares of y.

E8) Explain the following: (i) S_{13}, (ii) S_{33}.

Example 2: Let us now compute the quantities, S_{11}, S_{12}, S_{22}, S_{1y} and S_{2y} for the data given in Table 6.

Forming Table-7 and Table-8 will make our computations easier.

Table-7 : Computation of Sum of Squares and Products

S.No.	x_1	x_2	x_1^2	x_2^2	x_1x_2
1	321	59	103041	3481	18939
2	339	69	114921	4761	23391
3	334	63	111556	3969	21042
4	329	70	108241	4900	23030
5	325	58	105625	3364	18850
6	331	56	109561	3136	18536
7	324	61	104976	3721	19764
8	321	69	103041	4761	22149
9	337	59	113569	3481	19883
10	335	68	112225	4624	22780
11	320	69	102400	4761	22080
12	331	64	109561	4096	21184
13	329	67	108241	4489	22043
14	336	60	112896	3600	20160
15	339	58	114921	3364	19662
Total	4951	950	6134775	60508	313493

Table-8 : Computation of Sum of Squares and Products

S.No.	x_1	x_2	y	y^2	x_1y	x_2y
1	321	59	19.01	361.3801	6102.21	1121.59
2	339	69	12.66	160.2756	4291.74	873.54
3	334	63	17.45	304.5025	5828.30	1099.35
4	329	70	12.32	151.7824	4053.28	862.40
5	325	58	19.15	366.7225	6223.75	1110.70
6	331	56	19.09	364.4281	6318.79	1069.04
7	324	61	17.98	323.2804	5825.52	1096.78
8	321	69	18.50	342.2500	5938.50	1276.50
9	337	59	17.48	305.5504	5890.76	1031.32
10	335	68	12.93	167.1849	4331.55	879.24
11	320	69	16.80	282.2400	5376.00	1159.20
12	331	64	11.79	139.0041	3902.49	754.56
13	329	67	20.53	421.4809	6754.37	1375.51
14	336	60	14.32	205.0624	4811.52	859.20
15	339	58	19.21	369.0241	6512.19	1114.18
Total	4951	950	249.22	4264.168	82160.97	15683.11

From Table-7,

$$S_{11} = 1634775 - \frac{(4951)(4951)}{15} = 614.9333,$$

$$S_{12} = 313493 - \frac{(4951)(950)}{15} = -70.3333,$$

$$S_{1y} = 82160.97 - \frac{(4951)(249.22)}{15} = -98.2446.$$

* * *

You can now easily do the following exercises.

E9) Compute S_{yy} and S_{2y}.

E10) Guess what S_{21} will be. Is it same as S_{12}? What can you say about S_{1y} and S_{y1}?

E11) Compute S_{13}, S_{23}, S_{33} and S_{3y}. To save your time the following quantities are already computed for you.

$$\sum x_{1i}x_{3i} = 121313, \sum x_{2i}x_{3i} = 23335, \sum x_{2i}y_i = 15683.11,$$

$$\sum x_{3i} = 368, \sum x_{3i}y = 6206.03, \sum x_{3i}^2 = 9174.$$

You are now in a position to learn the multiple regression technique easily. Here, we shall look at multiple regression with two independent variables only.

9.4.2 Regression With Two Independent Variables

Recall the spring rework problem. For the time being let us ignore the soaking time and examine the effect of tempering temperature (x_1) and tempering time (x_2) on vibrations, y. This is done by fitting a linear regression formula for y on x_1 and x_2. The form of the linear regression function is given by

$$y = a_0 + a_1x_1 + a_2x_2 + e,$$

where a_0, a_1 and a_2 are unknown constants and e is the random error with mean zero and standard deviation σ. We have 15 observations on this model, namely, the 15 data points (x_{1i}, x_{2i}, y_i) presented in Table 8 (note that x_{3i}s are omitted as we are not considering x_3). As we have done in the one independent variable case, a_0, a_1 and a_2 are estimated by minimising the error sum of squares
$SSE = \sum_{i=1}^{15} e_i^2 = \sum_{i=1}^{15}(y_i - a_0 - a_1x_{1i} - a_2x_{2i})^2$.

The constants a_0, a_1, a_2 and σ^2 are estimated from the data. Let these estimates be denoted by $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$ and $\hat{\sigma^2}$. The resulting estimates $\hat{a}_1$ and $\hat{a}_2$ are obtained by solving the following equations:

$$\begin{aligned} S_{11}a_1 + S_{12}a_2 &= S_{1y} \\ S_{21}a_1 + S_{22}a_2 &= S_{2y} \end{aligned}$$

Above, we have already computed S_{11}, S_{12} etc., from the data. Substituting these values, we get

$$\begin{aligned} 614.9333\,a_1 - 70.3333\,a_2 &= -98.2447 \\ -70.3333\,a_1 + 341.3333\,a_2 &= -100.8230 \end{aligned}$$

We can solve these equations for a_1 and a_2 to obtain $\hat{a}_1$ and $\hat{a}_2$ respectively. However, there is a simple formula for $\hat{a}_1$ and $\hat{a}_2$ when $\Delta = S_{11}S_{22} - S_{12}^2 \neq 0$ which is the case in all most all the examples that we come across in practice. In our example,

$$\Delta = 614.9333 \times 341.3333 - (-70.3333)^2 = 204950.44 \neq 0$$

The estimates of $\hat{a}_1$ and $\hat{a}_2$ are given by

$$\hat{a}_1 = C_{11}S_{1y} + C_{12}S_{2y} \quad (10)$$

$$\hat{a}_2 = C_{21}S_{1y} + C_{22}S_{2y}, \quad (11)$$

where $C_{11} = S_{22}/\Delta$, $C_{12} = C_{21} = -S_{12}/\Delta$ and $C_{22} = S_{11}/\Delta$. Let us compute these quantities now.

$$C_{11} = \frac{341.3333}{204950.44} = 0.0017, \; C_{22} = \frac{614.9333}{204950.44} = 0.0030,$$

$$C_{12} = C_{21} = -\frac{-70.3333}{204950.44} = 0.0003.$$

Substituting C_{ij}s and S_{iy}s in Equations (10) and (11) we get

$$\begin{aligned} \hat{a}_1 &= -0.1982 \\ \hat{a}_2 &= -0.3362, \end{aligned}$$

To get the regression equation we need to obtain $\hat{a}_0$. This is given by the formula

$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2, \quad (12)$$

where $\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$ are the averages of y, x_1 and x_2 computed from the data. Thus,

$$\hat{a}_0 = 16.615 - (-0.1982) \times 330.07 - (-0.3362) \times 63.33 = 103.33.$$

Therefore, our linear regression formula for vibrations y based on x_1 and x_2 is given by

$$y = 103.33 - 0.1982\, x_1 - 0.3362\, x_2. \quad (13)$$

From the regression equation, we find that the coefficients of both x_1 and x_2 are negative. This means that number of vibrations can be increased by reducing the tempering temperature and tempering time or both.

Try this exercise now.

E12) Suppose we maintain x_1 at 320°C and x_2 at 60 minutes, then find the expected number of vibrations that a spring will withstand.

With this, we bring this unit to a close. Let us go back and recall the points covered in it.

9.5 SUMMARY

In this unit you have seen

1. that, regression analysis is a very useful technique by looking at some case applications,
2. how to identify linear relationship between a dependent variable and an independent variable by examining the scatter diagram,
3. the simple linear regression formula and how to build it from the data using the method of least squares principle,
4. the correlation coefficient and its significance in evaluating the regression formula,
5. the multiple linear regression formula with two independent variables and how to build them from data.

9.6 SOLUTIONS/ANSWERS

E1) **Table-2: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected i**

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138	154	172.5	182	207.5
$\hat{e}_i$	0	6	5.5	8	2.5

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

E2) **Table-3: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected i**

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	143.2	157.6	174.25	182.8	205.75
$\hat{e}_i$	-5.2	2.4	3.75	7.2	4.25

Note: $\hat{y}_i = 100 + 40x$ and $\hat{e}_i = y_i - \hat{y}_i$

E3) Most of the errors are on the positive side with both the formulae.

E4) $S_{xx} = 36.7135 - \frac{(22.59)^2}{15} = 2.6929$, $\sum x_i y_i = 3871.45$, $S_{xy} = 138.0760$,
$S_{yy} = 417533 - \frac{(2479)^2}{15} = 7836.93$.

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{138.076}{2.6929} = 51.2729,$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 165.2666 - 51.2729 \times 1.506 = 88.0496.$$

Therefore, the best linear regression formula is

$$y = 88.0496 - 51.2729\, x.$$

E5) Letting x_i be the income (in thousands of dollars) of the ith family, and y_i be the saving (in thousands of dollars) of the ith family, we find that

$$\sum_{i=1}^{9} x_i y_i = 63.7, \sum_{i=1}^{9} y_i = 3.6 \quad \bar{y} = 0.4,$$

$$\sum_{i=1}^{9} x_i^2 = 2364, \quad \sum_{i=1}^{9} x_i = 144, \quad \bar{x} = 16.$$

Thus, substituting these values in the alternate formula for $\hat{b}$, we obtain

$$\hat{b} = \frac{9(63.7) - (144)(3.6)}{9(2364) - 144^2} = \frac{573.3 - 518.4}{21.276 - 20.736} = 0.1017.$$

Consequently,

$$\hat{a} = \bar{y} - b\bar{x} = 0.4 - 0.1017(16) = -1.2272.$$

Thus, the regression line is

$$y = -1.2272 + 0.1017x,$$

where both x and y are measured in thousands of dollars.

The interpretation of this regression line is as follows: On the average, families with zero income would be expected to save -1,227.20. (Negative saving means that a family's consumption expenditure exceeds its income.) An increase in family income of 1,000 is associated with an increase in family saving of 101.70.

E6) $R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{138.076}{\sqrt{2.6929 \times 7839.93}} = 0.9504$

$\therefore R^2 = 0.9033$

There is a substantial increase in R^2.

E7) i) The notation for the corrected sum of products between x_2 and x_3 is S_{23} and this is defined by

$$S_{23} = \sum_{i=1}^{15} x_{2i}x_{3i} - \frac{(\sum_{i=1}^{15} x_{2i})(\sum_{i=1}^{15} x_{3i})}{15}.$$

ii) The notation for the corrected sum of products between x_3 and y is S_{3y} and this is defined by

$$S_{3y} = \sum_{i=1}^{15} x_{3i}y_i - \frac{(\sum_{i=1}^{15} x_{3i})(\sum_{i=1}^{15} y_i)}{15}.$$

iii) The notation for the corrected sum of squares of y is S_{yy} and this is defined by

$$S_{yy} = \sum_{i=1}^{15} y_i^2 - \frac{(\sum_{i=1}^{15} y_i)^2}{15}.$$

E8) (i) S_{13} is called the corrected sum of products between x_1 and x_3,

(i) S_{33} is called the corrected sum of squares of x_3.

E9) $S_{yy} = 4264.168 - \frac{(249.22)^2}{15} = 123.461,$

$$S_{2y} = 15683.11 - \frac{(950)(249.22)}{15} = -100.8233.$$

E10) S_{12} and S_{21} are equal as $x_{1i}x_{2i} = x_{2i}x_{1i}$. Similarly, $S_{1y} = S_{y1}$.

E11) $S_{13} = 151.533$, $S_{23} = 28.3333$, $S_{33} = 145.7333$ and $S_{3y} = 91.8326$.

E12) The required number is obtained by substituting the values $x_1 = 320°C$ and $x_2 = 60$ min. in Eqn.(13). Then we get

$$103.33 - 0.1982 \times 320 - 0.3362 \times 62 = 19.62.$$