UNIT 3 REGRESSION ANALYSIS

3.0 INTRODUCTION

In many problems there are two or more variables that are inherently related and it may be necessary to explore the nature of their relationship. Regression analysis is a statistical technique for modeling and investigating the relationship between two or more variables. For example in a chemical process suppose that the yield of the product is related to the process operating temperature. Regression analysis can be used to build a model that expresses yield as a function of temperature. This model can be used to predict yield at a given temperature level. It can also be used for process optimization or process control purposes.

In general, suppose that there is a single dependent variable or response variable y and that is related to k independent or regressor variables say $x_1,\ldots\ldots..,x_k$. The response variable y is a random variable and the regressor variables $x_1,\ldots\ldots,x_k$ are measured with negligible error. The relationship between y and $x_1,\ldots\ldots..x_k$ is characterized by a mathematical model and it is known as the regression model. It is also known as the regression of y on $x_1,\ldots\ldots..,x_k$. This regression model is fitted to a set of data. In many situations the experimenter knows the exact form of the functional relationship between y and $x_1, ..., x_\mathrm{k}$, say $\varphi(x_1, ..., x_\mathrm{k})$, except for a set of unknown parameters. When the functional form is unknown, it has to be approximated on the basis of past experience or from the existing information. Because of its tractability, a polynomial function is popular in the literature.

In this unit we will be mainly discussing the linear regression model and when $k = 1$, that is only one regressor variables. We will be discussing in details how to estimate the regression line and how it can be used for prediction purposes from a given set of data. We will also discuss briefly how we can estimate the function ϕ, if it is not linear.

3.1 OBJECTIVES

After reading this unit, you should be able to

- Decide how two variables are related.
- Measure the strength of the linear relationship between two variables.
- Calculate a regression line that allows to predict the value of one of the variable if

the value of the other variable is known.

- Analyze the data by the method of least squares to determine the estimated regression line to be used for prediction.
- Apply the least squares methods to fit different curves and use it for prediction purposes.

3.2 SIMPLE LINEAR REGRESSION

We wish to determine the relationship between a single regressor variable *x* and a response variable *y(note: The linear regression with one independent variable is referred to as simple linear regression).* We will refer to y as the dependent variable or response and x as the independent variable or regressor. The regressor variable *x* is assumed to be a continuous variable controlled by the experimenter. You know that it is often easy to understand data through a graph. So, let us plot the data on *Scatter diagram* (a set of points in a 2-D graph where horizontal axis is regressor and vertical axis is response) Suppose that the true relationship between *y* and *x* is straight line. Therefore, each observation y can be described by the following mathematical relation (model)

$$y = \beta_0 + \beta_1 x + \in \qquad (1)$$

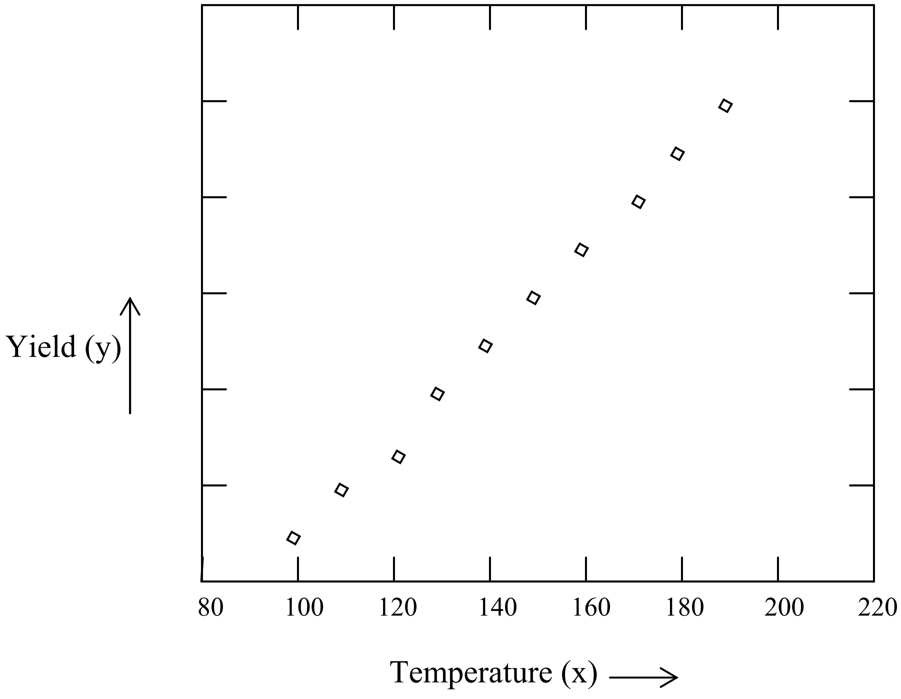

Figure 1: Scatter diagram of yield versus temperature

where $\in$ is a random variable with mean 0 and variance σ^2. The $\in$ is known as the error component and it is assumed to be small. If the error $\in$ was absent then it was a perfect relation between the variables *y* and *x* which may not be very practical. Let us look at the following example.

Example 1: A chemical engineer is investigating the effect of process operating temperature on product yield. The study results in the following data.

Temperature °C (x)	100	110	120	130	140	150	160	170	180	190
Yield,%(y)	45	51	54	61	66	70	74	78	85	89

The scatter diagram between the temperature and the yield is presented in the Figure 1 above. From the Figure 1 it is clear that there is a linear relationship between yield and temperature but clearly it is not perfect. For example we can not write the relationship between y and x as follows

$$y = \beta_0 + \beta_1 x$$

Clearly the presence of the error $\in$ is needed. Moreover the error $\in$ is a random variable because it is not fixed and it varies from one temperature to another. It may also vary when two observations are taken at the same temperature. If there was a perfect linear relationship between y and x we would have required just two points to find the relationship. Since the relationship is not perfectly linear it is usually required much more than two data points to find their relationship. Our main objective is to find the relationship between them from the existing information (data points). Since it is assumed that the relationship between x and y is linear therefore the relationship can be expressed by the equation (1) and finding the relationship basically boils down finding the unknown constants β_0 and β_1 from the observations.

Let us discuss this concept of linear regression by one more illustration/collection of data described in the table 1 given below. This table encloses the data of 25 samples of cement, for each sample we have a pair of observation (x,y) where x is percentage of SO_3, a chemical and y is the setting time in minutes. These two components are strongly related; it is the percentage of SO_3 which influences the setting time of any cement sample, the recorded observations are given in table 1 below.

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	1.64	178
16	1.19	170
17	1.56	160
18	1.53	160

19	0.96	140
20	1.7	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 the table 1, you see that setting time y increases as percentage of SO_3 increases. Whenever you find this type of increasing (or decreasing) trend in a table , same will be reflected in the scatter diagram ,and it indicates that there is a linear relationship between x and y. By drawing the scatter diagram you can 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 *x* 0.96). Let (x_i, y_i) denote the values of (x, y) for the i^{th} sample. From Table-1, notice that x_{12}=x_{19}=0.96, whereas y_{12} = 138 and y_{19} = 140. Let $\hat{y} = 90 + 50x_i$. That is, $\hat{y}_i$ **is the predicted value of y** (then using y = 90 + 50x for the i^{th} sample. Since, x_{12}=x_{19}=0.96, both $\hat{y}_{12}$ and $\hat{y}_{19}$ are equal to 138. Thus the difference $\hat{e}_i = y_i - \hat{y}_i$, **the error in prediction**, also called residual is observed to be $\hat{e}_{12}$ = 0 and $\hat{e}_{19}$ = 2. The formula we have considered above, y = 90 + 50x, is called a **simple linear regression equation** , we will study these terms in detail in our successive sections.

3.2.1 Least squares estimation

Suppose that we have *n* pairs of observations, say (x_1 , y_1),…….(x_n , y_n). It is assumed that the observed y_i and x_i satisfy a linear relation as given in the model (1). These data can be used to estimate the unknown parameters β_0 and β_1. The method we are going to use is known as the method of least squares, that is, we will estimate β_0 and β_1 so that the sum of squares of the deviations from the observations to the regression line is minimum. We will try to explain it first using a graphical method in Figure 2. For illustrative purposes we are just taking 5 data points (*x*, *y*) = (0.5, 57), (0.75, 64), (1.00, 59), (1.25, 68), (1.50,74). The estimated regression line can be obtained as follows. For any line we have calculated the sum of the differences (vertical distances) squares between the *y* value and the value, which is obtained using that particular line. Now the estimated regression line is that line for which the sum of these differences squares is minimum.

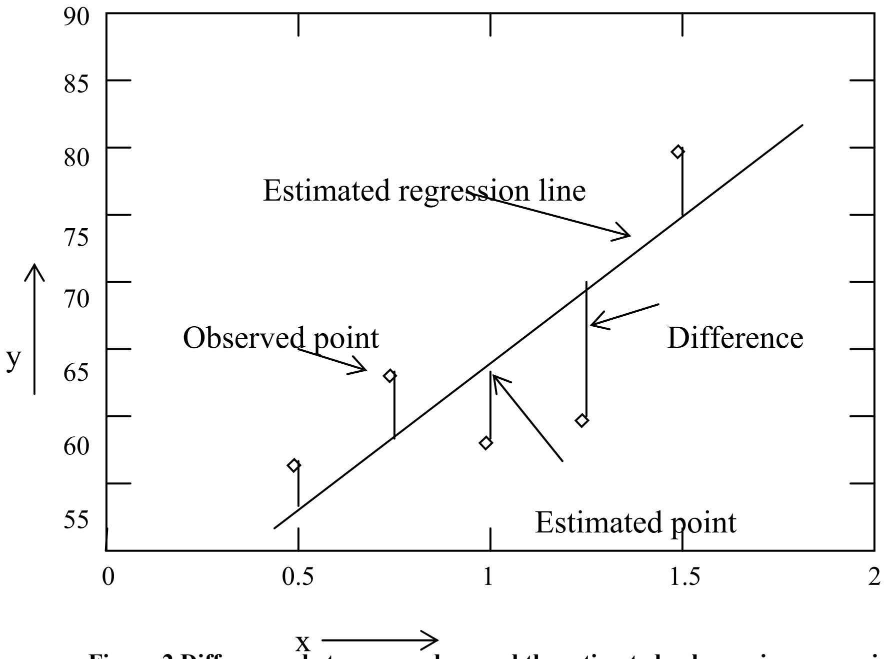

Figure 2 Differences between y values and the estimated values using regression line

Matematically the sum of squares of the deviations of the observations from the regression line is given by

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

The least squares estimators of β_0 and β_1, are $\hat{\beta}_0$ and $\hat{\beta}_1$ which can be obtained by solving the following two linear equations.

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

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

Simplifying these two equations yields

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

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (3)$$

Solving (2) and (3) $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained as follows:

$$\boxed{\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \qquad (4) \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} yi\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} xi\right)^2} \qquad (5) \end{aligned}}$$

where $\overline{y} = \sum_{i=1}^{n} y_i$ and $\overline{x} = \sum_{i=1}^{n} x_i$. Therefore, the fitted simple linear regression line between these n points is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

Note : Linear Correlation and Regression

linear correlation and regression are very similar. One uses the correlation coefficient to determine whether two variables are linearly related or not. The correlation coefficient measures the strength of the linear relationship. Regression on the other hand is used when we want to answer question about the relationship between two variables.

Some Properties for Linear Correlation and Regression

(1) The line of regression of y on x always passes through($\overline{x}$, $\overline{y}$) where $\overline{x}$, and $\overline{y}$ are the mean of x and y values.

(2) There are always two line of regression one of y on x and other of x on y. i.e., y = a_1 + b_{yx} x or x = a_2 + b_{xy} y

where byx = Regression coeff of y on x = $r\frac{\sigma_y}{\sigma_x}$

bxy = Regression coeff of y on x = $r\frac{\sigma_x}{\sigma_y}$

Correlation can be obtained by the following formula also,

$r = \sqrt{b_{xy} * b_{yx}}$ (-1≤ r ≤ 1)

Angle between lines of regression is given by,

$$\theta = \tan^{-1}\left\{\frac{r^2-1}{r}\left(\frac{\sigma_x * \sigma_y}{\sigma^2{}_x + \sigma^2{}_y}\right)\right\}$$

Where r = correlation coeff between x and y

σ_x = standard deviation of variable x

σ_y = standard deviation of variable y

So, now, Regression equation of y on x can be rewritten as

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

And Regression equation of x on y as,

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

Example 1 (contd.) Now we will compute the estimates of β_0 and β_1 for the data points given in Example 1. In this case it is observed

$$n = 10, \qquad \sum_{i=1}^{10} x_i = 1450, \qquad \sum_{i=1}^{10} yi = 673, \quad \overline{x} = 145, \quad \overline{y} = 67.3$$

$$\sum_{i=1}^{10} x_i^2 = 218{,}500,\quad \sum_{i=1}^{10} y_i^2 = 47{,}225,\quad \sum_{i=1}^{10} x_i y_i = 101{,}570$$

Therefore,

$$\hat{\beta}_1 = \frac{101{,}570 - 10 \times 1450 \times 673}{218{,}500 - 10 \times 1450^2} = 0.483,$$

and

$$\hat{\beta}_0 = 673 - 0.483 \text{ X } 145 = - 2.739.$$

The fitted simple linear regression line through those 10 points is

$$\hat{y} = -2.739 + 0.483x \qquad (7)$$

The best fitted line along with the data points are plotted in the Figure 3. Note that the best fitted line can be used effectively for prediction purposes also. For example suppose we want to know the expected yield when the temperature is 170°C, for which the data is not available. We can use the (7) for this purpose as follows.

$$\hat{y} = -2.739 + 0.483 \times 170 = 79.371\,.$$

Therefore, the best fitted line shows that the expected yield at 170°C is 79.371.

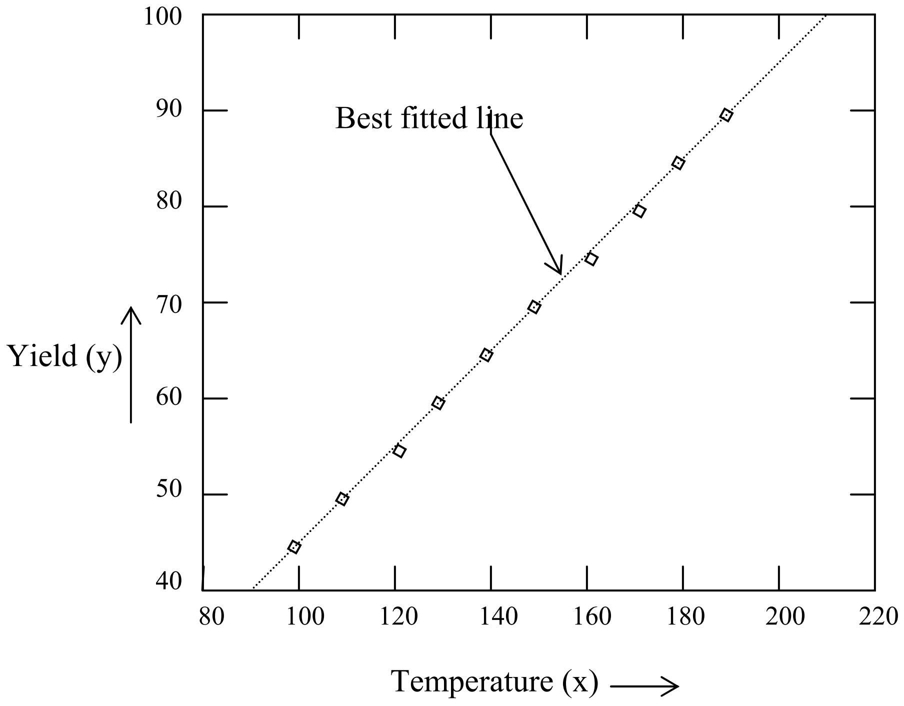

Figure 3: Data points and the best fitted regression line passing through these points

Soon in this section only, we will discuss the technique consisted of few steps; which can be used to fit a line in best way, such that the error is minimum. In crux we will study the technique to determine the best equation, that can fit a line in the data such that the error is minimum. But before that lets see one more example.

Example 2: A survey was conducted to relate the time required to deliver a proper presentation on a topic , to the performance of the student with the scores he/she receives. The following Table shows the matched data:

Table 2

Hours (x)	Score (y)
0.5	57
0.75	64
1.00	59
1.25	68
1.50	74
1.75	76
2.00	79
2.25	83
2.50	85
2.75	86
3.00	88
3.25	89
3.50	90
3.75	94
4.00	96

(1) Find the regression equation that will predict a student's score if we know how many hours the student studied.

(2) If a student had studied 0.85 hours, what is the student's predicted score?

Solution. We will arrange the data in the form of a chart to enable us to perform the computations easily.

Table 3

x	y	x^2	xy
0.5	57	0.25	28.5
0.75	64	0.56	48.0
1.00	59	1.00	59.0
1.25	68	1.56	85.0
1.50	74	2.25	111.0
1.75	76	3.06	133.0
2.00	79	4.00	158.0
2.25	83	5.06	186.75
2.50	85	6.25	212.5
2.75	86	7.56	236.5
3.00	88	9.00	246.0
3.25	89	10.56	289.25
3.50	90	12.25	315.0
3.75	94	14.06	352.50
4.00	96	16.00	384.0
33.75	1188	93.44	2863

In this case n = 15, therefore

$$\hat{\beta}_1 \frac{15\times 2863 - 33.75\times 1188}{15\times 93.44 - 33.75^2} = 10.857\ ,\ \hat{\beta}_0 = \frac{1}{15}[1188 - 10.857\times 33.75] = 54.772$$

Therefore, the prediction line becomes:

$$\hat{y} = 54.772 + 10.857x$$

Now the predicted score when $x = 0.85$, is

$$\hat{y} = 54.772 + 10.857 \times 0.85 = 64.00$$

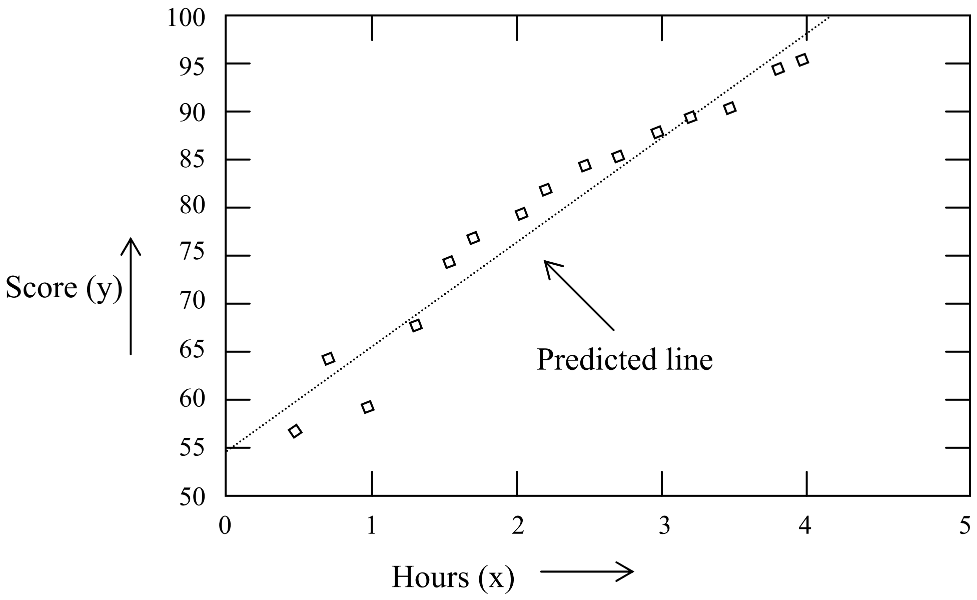

Figure 4: Hours studied and the corresponding score with the best fitted regression line passing through these points

Thus the predicted score of the student who had studied 0.85 hours is approximately 64.00.

We have plotted the individual scores and the hours studied with the best fitted prediction line in the Figure 4. It shows the hours studied by the student and the corresponding score follows a linear pattern and the predicted line can be used quite effectively to predict the score of a student if we know how many hours the student had studied.

Now, it's the time to discuss the technique for determining the best equation, i.e., the equation which fits the line in a way that the overall error is minimized.

From above illustrations and examples 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.

This means that whatever straight line we use, it is not possible to make all $\hat{e_i}s$ zero , where $\hat{e_i} = y_i - \hat{y_i}$ (the difference). 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 β_0 and β_1 in the formula $y = \beta_0 + \beta_1 x + e \;\; (s.t. \;\; e □ 0)$. Let us see how we do this.

Now our aim is to find the values β_0 and β_1 so that the error $\hat{e_i s}$ are minimum. For that we state here four steps to be done.

1) Calculate a sum S_{xx} defined by

$$S_{xx} = \sum_{i=1}^{n} x^2{}_i - n\bar{x}^2 \tag{8}$$

where x_i's are given value of the data and $\bar{x} = \frac{\Sigma x_i}{n}$ 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

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \tag{9}$$

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 $\frac{S_{xy}}{S_{xx}} = \beta_1$ say. That is

$$\beta_1 = \frac{S_{xy}}{S_{xx}} \tag{10}$$

4) Find $\bar{y} - \beta_1 \bar{x} = \beta_0$, say.

Let us now compute these values of the data in Table 1: Data on SO_3 and Setting Time, we get

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

Substituting these values in (10) and (11), we get

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = 54.943 \text{ and } \beta_0 = 168.68 - 54.943 \text{ x } 1.5616 = 82.88 \tag{11}$$

Therefore, the best linear prediction formula is given by
y = 82.88 + 54.943x.

After drawing this line on the scatter diagram, you can find that this straight lines is close to more points, and hence it is the best linear prediction.

Example 3: 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 the following table.

Table 4

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

(a) Find the scatter diagram for the data given above.
(b) Find the regression equation when the monthly output is the dependent variable (x) and monthly cost is the independent variable (y).
(c) Use this regression line to predict the firm's monthly cost if they decide to produce 4 tons per month.
(d) Use this regression line to predict the firm's monthly cost if they decide to produce 9 tons per month.

Solution

a) Suppose that x_i denote the output for the ith month and y_i denote the cost for the i^{th} month. Then we can plot the graph for the pair (x_i, y_i) of the values given in Table . Then we get the scatter diagram as shown in Figure below.

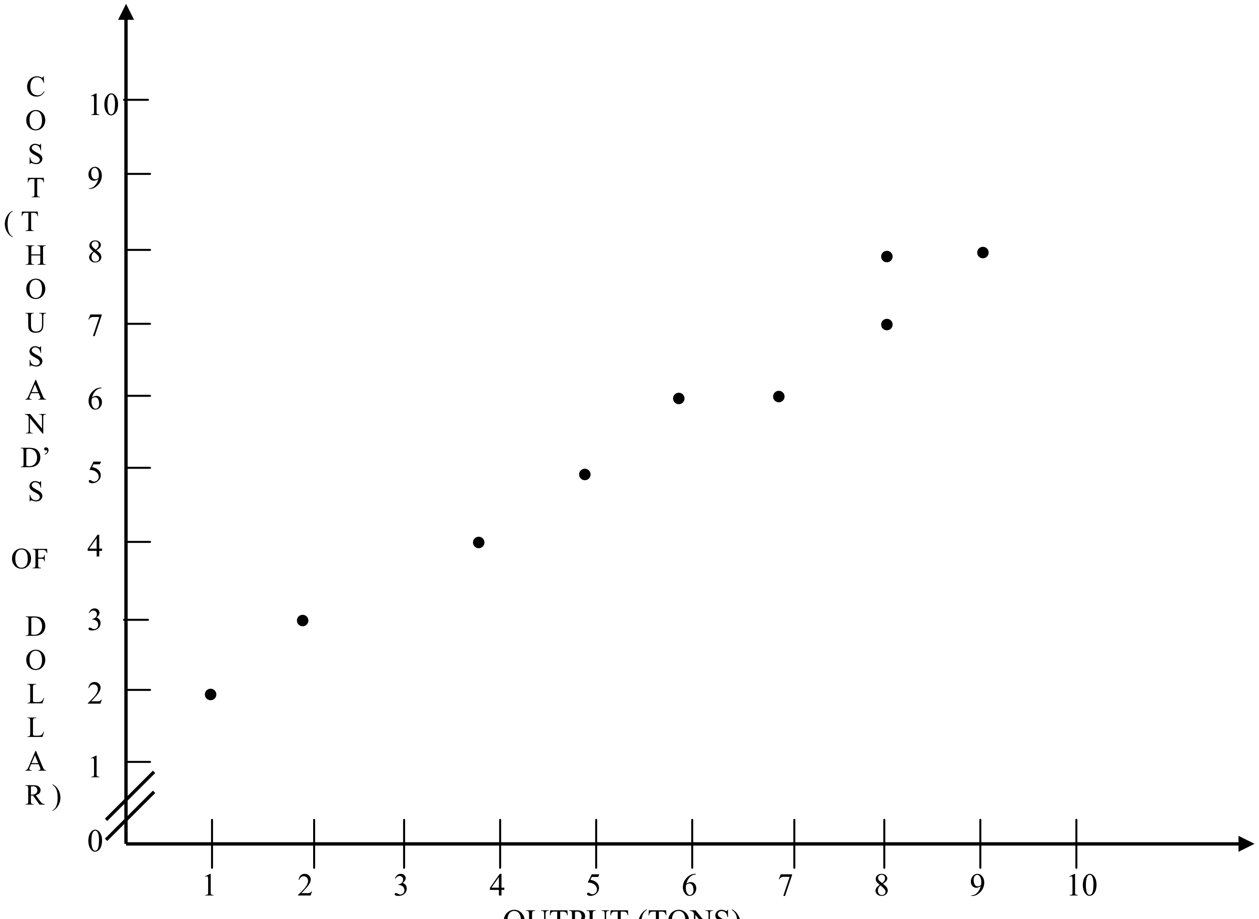

Figure 5: Scatter Diagram

b) Now to find the least square regression line, we first calculate the sums S_{xx} and S_{xy} from Eqn.(8) and (9).

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

Note that from Table(4) we get that

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{9}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{49}{9}$$

$$\sum x^2{}_i = 340$$

$$\sum y^2{}_i = 303$$

and $\sum x_i y_i = 319$

Therefore, we get that

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_1{}^2 - n\bar{x}^2}$$

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

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

Correspondingly, we get

$$\beta_0 = \frac{49}{9} - (0.752)\ x\ \frac{50}{9}$$

= 1.266

Therefore, the best linear regression line is

y = 1.266 + (0.752)x

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

1.266 + (0.752) x 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.

d) If the firms decides to produce 9 tons per month, then one can predict that its cost would be 1.266 + (0.752) x 9=8.034

Since the costs are measured in thousands of dollars, this means that the total costs would be expected to be $8034.

Check your progress 1

E 1: In partially destroyed laboratory record of an analysis of correlation data, the following results only are legible.

Variance of x = 9

Regression equations : 8x – 10y + 66 = 0
40x – 18y – 214 = 0

what were (1) the mean values of x and y,
(2) the correlation coeff between x and y
(3) the standard deviation of y

E2 Since humidity influences evaporation, the solvent balance of water reducible paints during sprayout is affected by humidity. A controlled study is conducted to examine the relationship between humidity (X) and the extent of evaporation (Y) is given below in table 5. Knowledge of this relationship will be useful in that it will allow the painter to adjust his or her spray gun setting to account for humidity. Estimate the simple linear regression line and predict the extent of solvent evaporation (i.e loss of solvent ,by weight)when the relative humidity is 50%

Table 5

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1
19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

3.2.2 Goodness of Fit

We have seen in the previous subsection that the regression line provides estimates of the dependent variable for a given value of the independent variable. The regression line is called the best fitted line in the sense of minimizing the sum of squared errors. The best fitted line shows the relationship between the independent (x) and dependent (y) variables better than any other line. Naturally the question arises “How good is our best fitted line?”. We want a measure of this goodness of fit. More precisely we want to have a numerical value which measures this goodness of fit.

For developing a measure of goodness of fit, we first examine the variation in y. Let us first try 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’s is accounted by the variation in x’s.

Actually, we can mathematically show that the total variation in *y*'s can be split up as follows:

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{S^2{}_{xy}}{S_{xx}} + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad (12)$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 ; S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \hat{y}_i)$$

Now if we divide (12) by S_{yy} on both sides, we get

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

Since the quantities on the right hand side are both non-negative, none of them can exceed one. Also if one of them is closer to zero the other one has to be closer to one. Thus if we denote

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

then

$$R^2 = \frac{S^2{}_{xy}}{S_{xx} S_{yy}}$$

Since R^2 must be between 0 and 1, R must be between – 1 and 1. It is clear that if $R^2 = 1$, then

$$\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{S_{yy}} = 0 \text{ or } \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0 \qquad \text{or} \qquad y_i = \hat{y}_i \qquad \text{for all } i.$$

Again when R^2 is close to 1, $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is close to zero. When *R* is negative, it means that *y* decreases as *x* increase and when *R* is positive *y* increases when *x* increases. Thus R gives a measure of strength of the relationship between the variables *x* and *y*.
Now let us compute the value of *R* for Example 1. For calculating the numerical value of *R*, the following formula can be used;

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

Therefore, for Example 1, the value of R becomes;

$$\mathrm{R} = \frac{101,570 - 10 \times 145 \times 67.3}{\sqrt{218,500 - 10 \times 145^2 \sqrt{47225 - 10 \times 67.3^2}}} = \frac{3985}{\sqrt{8250 \sqrt{1932.1}}} = 0.9981$$

and $R^2 = 0.9963$.

Therefore, it is clear from the value of *R* or from R^2 that both of them are very close to one. From the figure also it is clear that the predicted line fits the data very well.

Moreover R is positive means, there is a positive relation between the temperature and yield. As the temperature increases the yield also increases.

Now the natural question is how large this R or R^2 will be to say that the fit is very good. There is a formal statistical test based on F-distribution which can be used to test whether R^2 is significantly large or not. We are not going into that details. But as a thumb rule we can say that if R^2 is greater that 0.9, the fit is very good, if it is between 0.6 to 0.8, the fit is moderate and if it is less than 0.5 it is not good.

Check your progress 2

E1) For the data given in the table below compute R and R^2

Table 6: $\hat{y_i}$ 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 $\hat{e_i} = y_i - \hat{y_i}$

3.2.3 Residual Analysis

Fitting a regression model to a set of data requires several assumptions. For example estimation of the model parameters requires the assumptions that the errors are uncorrelated random variables with mean zero and constant variance. If these assumptions are not satisfied, then using the simple least squares method may not produce the efficient regression line. In fitting a linear model, it is also assumed that the order of the model is correct, that is if we fit a first order polynomial, then we are assuming that phenomenon actually behave in a first order manner. Therefore, for a practitioner it is important to verify these assumptions and the adequacy of the model.

We define the residuals as $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, where y_i is an observation and $\hat{y}_i$ is the corresponding estimated value from the best fitting regression line.

Analysis of the residuals is frequently helpful in checking the assumption that errors are independent and identically distributed with mean zero and finite variance and in determining `whether the additional terms in the model would be required not. It is advisable to plot the residuals

a) in time sequence (if known),
b) against the $\hat{y}_i$ or
c) against the independent variable x. These graphs will usually look like one of the four general patterns as shown in the Figures 6 – 9.

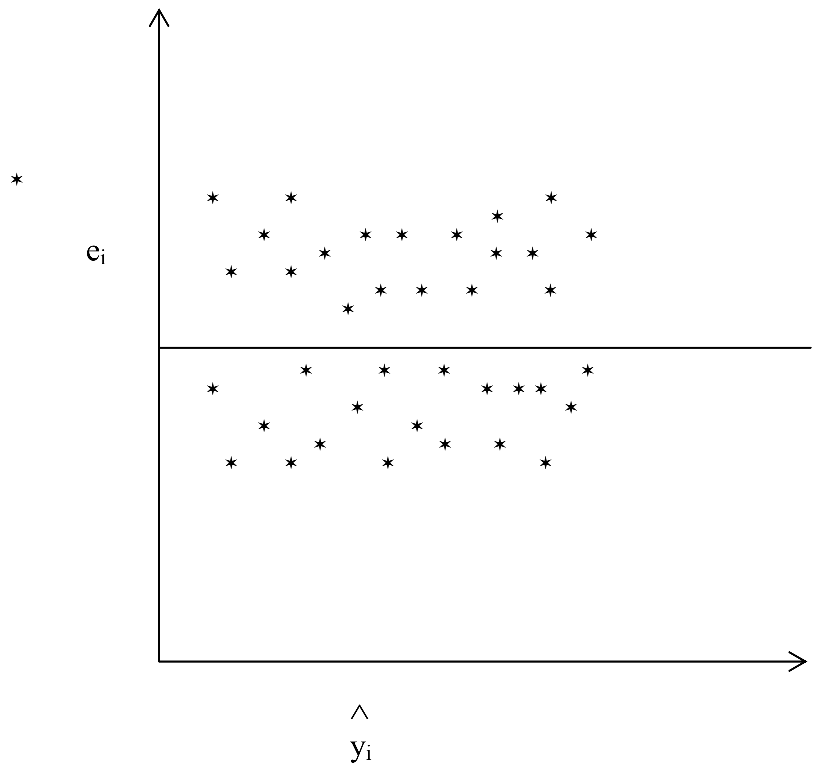

Figure 6 Pattern of the residual plot; satisfactory.

The Figure 6 indicates that the residuals are behaving in satisfactory manner and the model assumptions can be assumed to be correct. The Figures 7 – 9 given below indicate unsatisfactory behaviour of the residuals. The Figure 7 clearly indicates that the variances are gradually increasing. Similarly the Figure 8 indicates that the variances are not constant. If the residuals plot is like the Figure 9, then it seem the model order is not correct, that means, the first order model is not the correct assumption. We should look for higher order models.

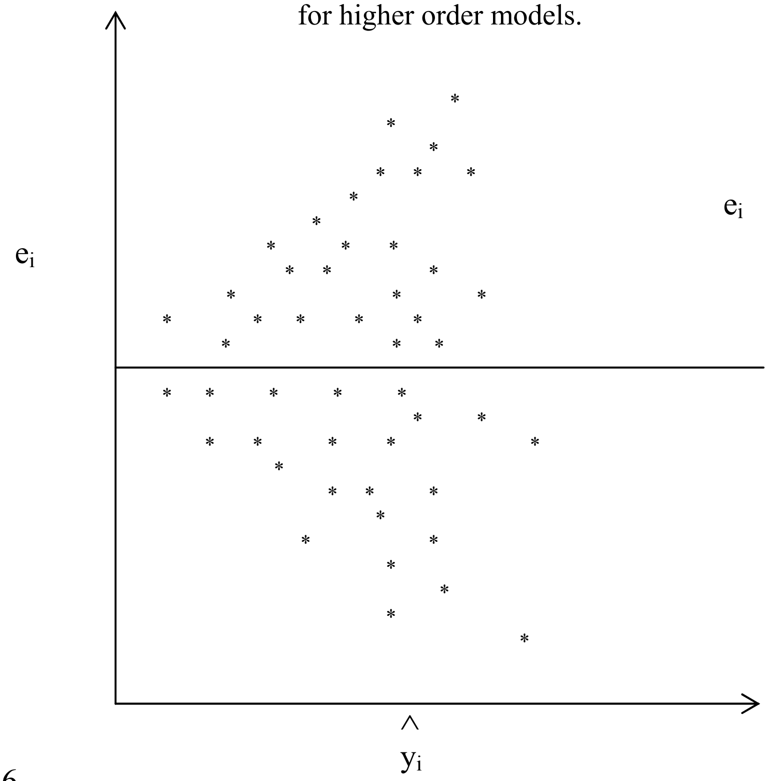

Figure 7 Pattern of the residual plot; indicates the variance is gradually increasing this case

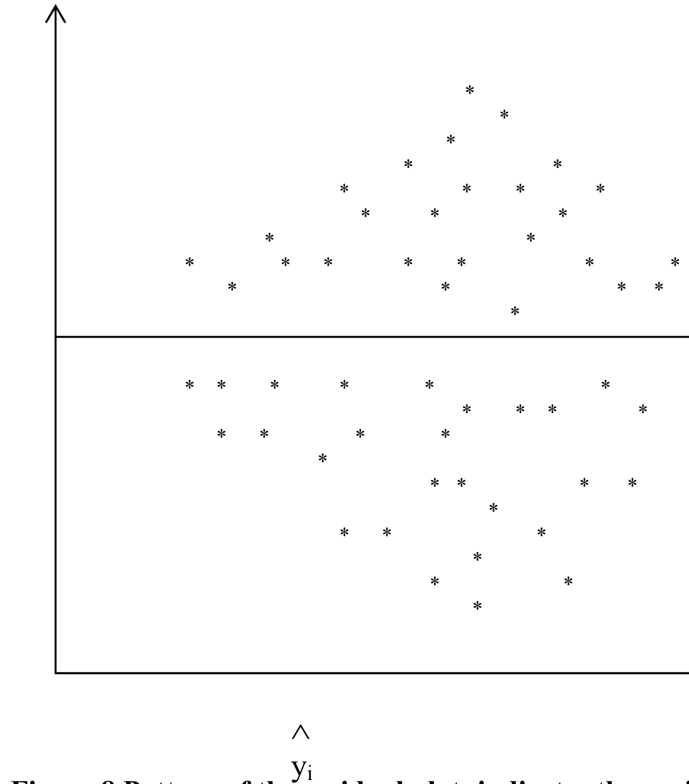

Figure 8 Pattern of the residual plot; indicates the vari not constant.

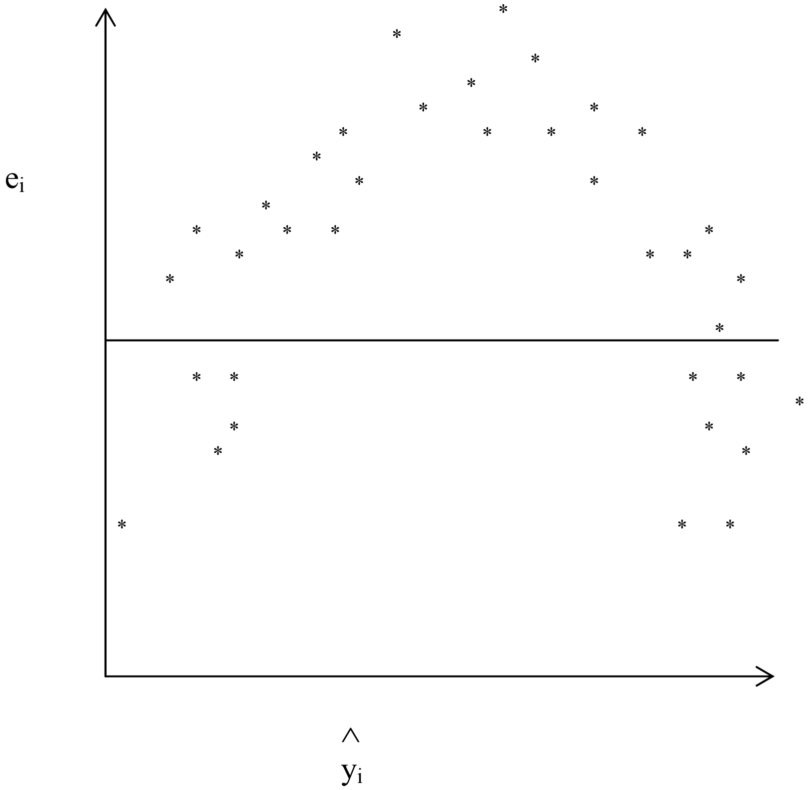

Figure 9 Patent of the residual plot; indicates the model order is not correct.

Example 4: Now we provide the residual plots of the data given in Example 1. We have plotted $\hat{y}_i$ vs. e_i. It is provided in the Figure 10. From the Figure 10, it is quite clear that the residuals plot is quite satisfactory and apparently all the model assumptions are satisfied in Figure 10 here.

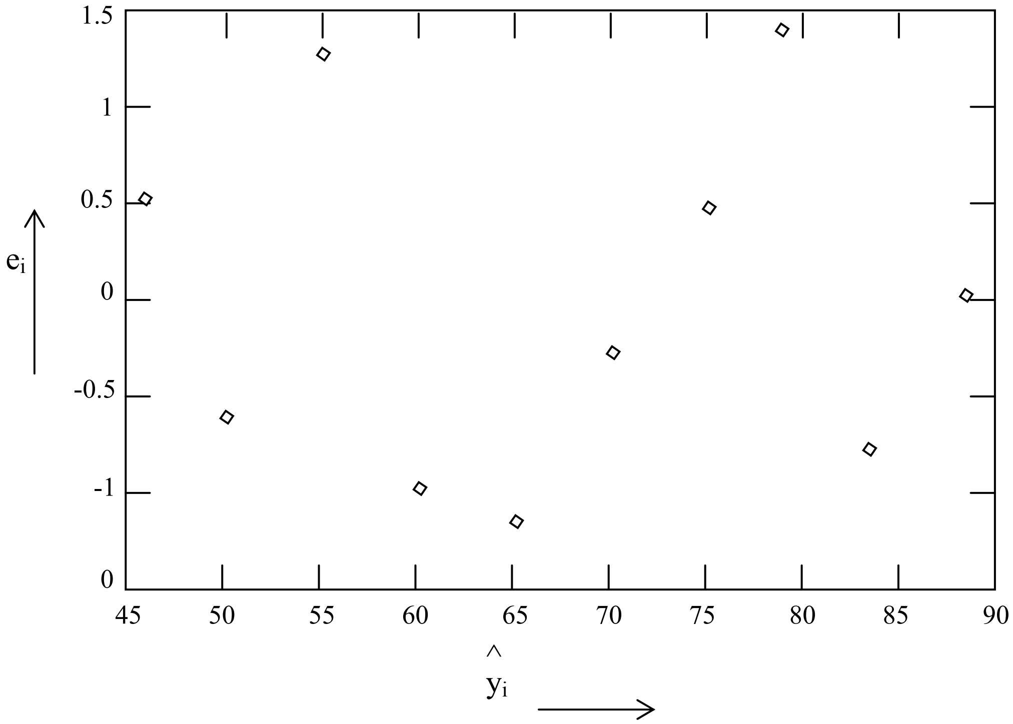

Figure10 Pattern of the residual plot; satisfactory.

Check your progress 3

E1 What is the utility of residual plots? what is the disadvantage of residual plots?

3.3 NON-LINEAR REGRESSION

Linear regression is a widely used method for analyzing data described by models which are linear in parameters. However, in many practical situations, people come across with data where the relationship between the independent variable and the dependent variable is no more linear. In that case definitely one should not try to use a linear regression model to represent the relationship between the independent and dependent variable. Let us consider the following example.

Example 5 Data on the amount of protein generated by a certain experiment were counted and reported over time. They are presenting below in Table 7:

Table 7

Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)
0	0.844	80	0.818	160	0.580	240	0.457
10	0.908	90	0.784	170	0.558	250	0.448
20	0.932	100	0.751	180	0.538	260	0.438
30	0.936	110	0.718	190	0.522	270	0.431
40	0.925	120	0.685	200	0.506	280	0.424
50	0.908	130	0.685	210	0.490	290	0.420
60	0.881	140	0.628	220	0.478	300	0.414
70	0.850	150	0.603	230	0.467	310	0.411

We present the Time vs. Protein generated in the Figure 11.

From Figure 11 it is clear that the relationship between the time and protein generated is not linear. Therefore, they can not be explained by a linear equation. In a situation like this we may often go for a non-linear model to explain the relationship between the independent and dependent variables and they are called the non-linear regression model.

A non-linear regression model can be formally written as

$$y = f(x,\theta) + \in, \tag{13}$$

where $f(x,\theta)$ is a known response function of k-dimensional vector of explanatory variable x and p-dimensional vector of unknown parameter θ. Here also $\in$ represents the error component and it is assumed that it has mean zero and finite variance. Therefore, it is clear that the non-linear regression model is a generalization of the linear regression model. In case of linear regression model $f(x, \theta)$ is a linear function, but there it can be any non-linear function also. Similar to the linear regression model, here also our problem is same, that is, if we observe a set of n, $\{(x_1, y_1),\ldots\ldots, (x_n, y_n)\}$ how to estimate the unknown parameters θ, when we know the functional form of $f(x, \theta)$.

3.3.1 LEAST SQUARES ESTIMATION

Similar to the linear regression method here also to estimate the unknown parameters, we adopt the same method. We find the estimate of θ by minimizing the residual sums of squares, that is minimize

$$Q(\theta)=\sum_{i=1}^{n}[y_i - f(x_i,\theta)]^2, \tag{14}$$

with respect to the unknown parameters. The idea is same as before, that is we try to find that particular value of θ for which the sum of squares of the distance between the points y_i and $f(x_i,\theta)$ is minimum. Unfortunately in this case the minimum can not be performed as easily as before. We need to adopt some numerical technique to minimize the function $Q(\theta)$. This minimization can be performed iteratively and one technique that can be used to accomplish this is the Gauss-Newton method.

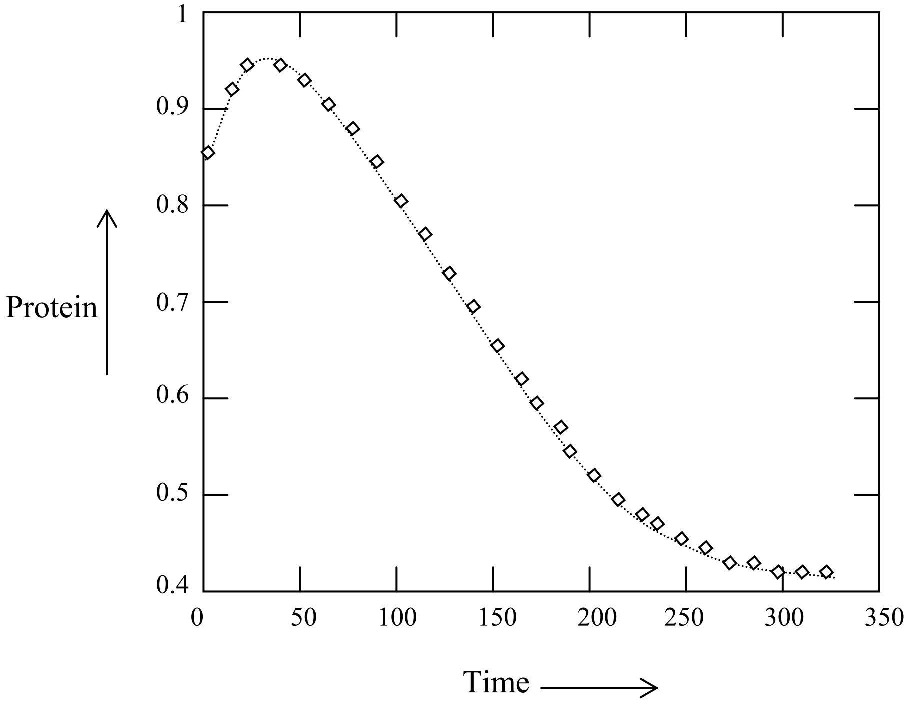

Figure 11 Time vs. Protein generated in an experiment.

You have already learned about the Gauss-Newton method in details before, we just give a brief description for your convenience. We use the following notations below:

$$\theta=(\theta_1,.....,\theta_p), \qquad \theta^{(k)}=(\theta_1^{(k)},........,\theta_p^{(k)}) \tag{15}$$

Expand the function $f(x,\theta)$ using a Taylor series expansion about the starting point $\theta^{(0)}$ and using the only the first order expansion, we get:

$$f(x_i,\theta)\approx f(x_i,\theta^{(0)})+v_{i1}(\theta_1-\theta_1^{(0)})+.....+v_{ip}(\theta_p-\theta_p^{(0)})$$

where

$$v_{ij}=\left.\frac{\partial f(x_i,\theta)}{\partial\theta_j}\right|_{\theta=\theta^{(0)}} \qquad \text{for} \qquad j=1,.........,p.$$

Let $\eta(\theta) = (f(x_1,\theta),.....,f(x_n,\theta))'$ and $y =$ $y = (y_1,......y_n)'$ then in the matrix notation we can write (15)

$$\eta(\theta) \approx \eta(\theta^{(0)}) + V^{(0)}(\theta - \theta^{(0)}),$$

where $V^{(0)}$ is the $\eta \times p$ derivative matrix with elements v_{ij}. Therefore, to compute the first estimates beyond the starting value is to compute

$$b_0 = [V^{(0)'}V^{(0)}]^{-1}[y - \eta(\theta^{(0)})]$$

and then solve for the new estimate $\theta^{(1)}$ as

$$\theta^{(1)} = b_0 + \theta^{(0)}.$$

This procedure is repeated then with $\theta^{(0)}$ is replaced by $\theta^{(1)}$ and $V^{(0)}$ by $V^{(1)}$ and this produces a new set of estimates. This iterative procedure continues until convergence is achieved. Naturally these computations can not performed by hands, we need calculators or computers to perform these computations.

Example 5 (Contd). In this case it is observed (theoretically) that the following model (16) can be used to explain the relationship the time and yield generated y_i where

$$y_t = \alpha_0 + \alpha_1 e^{\beta_1 t} + \alpha_2 e^{\beta_2 t} + \in_t. \qquad (12)$$

Note that as we have mentioned for the general non-linear regression model, in this case also the form of the non-linear function namely $\alpha_0 + \alpha_1 e^{\beta_1 t} + e^{\beta_2 t}$ is known, but the parameters of the model, that is, $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ is unknown. Given the data as provided in Example 5, we want to estimate the unknown parameters.

We use the following initial guess $\alpha_0 = 0.5, \alpha_1 = 1.5, \alpha_2 = -1.0, \beta_1 = -0.01, \beta_2 = -0.02$, and finally using the Gauss-Newton algorithm we obtain the estimates of the parameters as follows:

$\hat{\alpha}_0 = 0.375, \quad \hat{\alpha}_1 = 1.936 \quad \hat{\alpha}_2 = 1.465, \quad \hat{\beta}_0 = -0.013 \quad \hat{\beta}_1 = -0.022$

We have plotted the points and also the best fitted regression line, namely

$$\hat{y} = 0.375 + 1.936e^{-0.013t} - 1.465e^{-0.013t} \qquad (17)$$

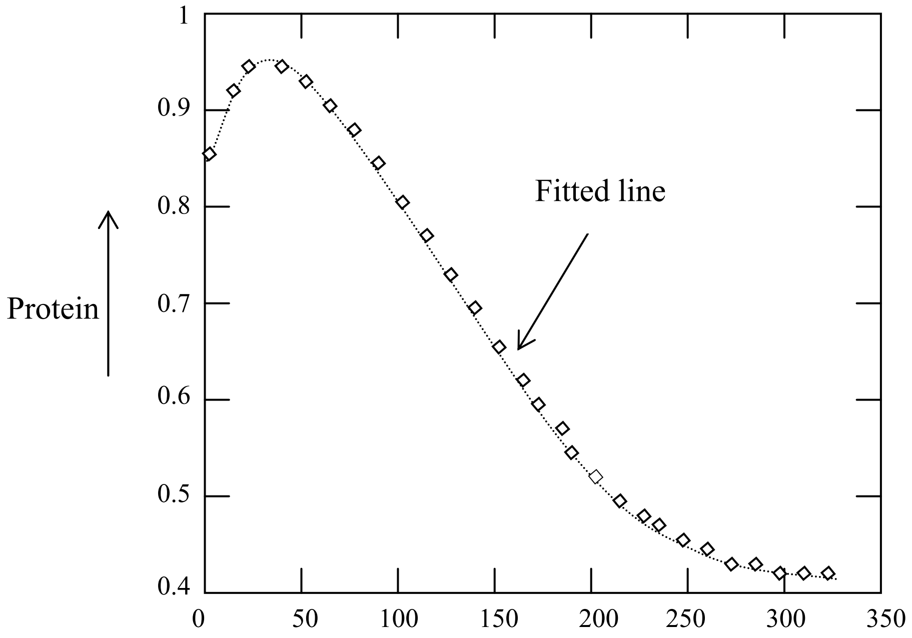

Figure 12 Time vs Protein generated in an experiment and the best fitted curve

in the Figure 12. The Figure 12 indicates that the fitted regression curve provides a very good relationship between the time and protein generated in that experiment. As before the prediction curve, that is, the curve (17) can be used easily for prediction purposes also. For example suppose we want to estimate the expected protein generation at the time 115 minutes after the experiment, then using (17), we obtain

$$\hat{y} = 0.375 + 1.936e^{-0.013\times 115} - 1.465e^{-0.013\times 115} = 0.698.$$

Therefore, at the 115 minutes the expected protein generation is 0.698 gms.

Some points we should remember about the non-linear regression model is that we have to know the functional form of the relationship, but the parameters involved are unknown. Usually the functional forms are obtained from the physical nature of the process and they are available. If they are completely unknown it is difficult to use the non-linear regression model. In that case we need to try with some guess models but they are not very easy and they are not pursued here. Another important issue is the choice of the guess value of the iterative process. This is also a difficult problem. Usually from the prior experimental results we may have to try some trial and error method to find the initial guess values.

Check your progress 4

E1 Data on the amount of heat generated by friction were obtained by Count Rumford in 1798. A bore was fitted into a stationery cylinder and pressed against the bottom by means of a screw. The bore was turned by a team of horses for 30 minutes and Rumford measured the temperature at small in intervals of time. They are provided in the Table below:

Table 8

Time (min)	Temperature ($^\circ F$)	Time (min)	Temperature ($^\circ F$)
4	126	24	115
5	125	28	114
7	123	31	113
12	120	34	112
14	119	37.5	111
16	118	41	110
20	116		

(1) Plot the time versus temperature and convince yourself that the linear regression model is not the correct model in this case.

(2) A model based on Newton's law of cooling was proposed as

$$f(t,\theta) = 60 + 70e^{-\theta t}$$

Using an initial guess of $\theta^{(0)} = 0.01$, find the least squares estimate of θ.

(3) Based on the fitted least squares regression line find the expected temperature at the time 15th minute after starting the experiment.

3.4 SUMMARY

In this unit you have seen :

- that regression analysis is an important technique, which can be used to verify the results of any experiment.
- How to determine the relationship between a dependent and an independent variable by using the Scatter diagram
- that by knowing the technique of regression you have an edge to analyse the results in an organized way. Further this analysis is smoothened by application of the concepts like least square estimation, goodness to fit and residual analysis.
- that many times the data obtained by conducting an experiment does not follow the linear relation. So, to handle such aspects we have also discussed the concept of non linear regression, under we have emphasized least square estimation technique.
- Formulas and applications of following topics:
    - Simple Linear Regression
        - Least Squares Estimation
        - Goodness to Fit
        - Residual Analysis
    - Non-Linear Regression
        - Least Squares Estimation

3.5 SOLUTIONS

Check you r progress 1

E 1:

(1) Since both the regression lines pass through the point ($\bar{x}$, $\bar{y}$), we have

$$8\bar{x} - 10\ \bar{y} + 66 = 0$$

$$40\bar{x} - 18\ \bar{y} - 214 = 0$$

Solving we get, $\bar{x} = 13$

$\bar{y} = 17.$

Let 8x – 10y + 66 = 0 and 40x – 18y – 214 = 0
Be the lines of regression of y and x and x on y respectively. Now, we put them in the following form.

$$= \frac{8}{10}x + \frac{66}{10} \text{ and } x = \frac{18}{40}\text{y} + \frac{214}{40} \qquad (4)$$

$\therefore$ byx = regression coeff of y on x = $\frac{8}{10} = \frac{4}{5}$

bxy = regression coeff of x on y = $\frac{18}{40} = \frac{9}{20}$

Hence, r^2 = bxy. byx = $\frac{4}{5} \; \frac{9}{20} \;\; = \;\; \frac{9}{25}$

So r = $\pm \; \frac{3}{5} \;\; = \;\; \pm \;\; 0.6$

Since, both the regression coeff are +ve, we take r = + 0.6

(3) We have, byx = $r\frac{\sigma_y}{\sigma_x} \qquad \Rightarrow \qquad \frac{4}{5} = \frac{3}{5} \; x \; \frac{\sigma_y}{3}$

$\therefore \;\; \sigma_y = 4$

Remarks (i) had we taken 8x – 10y + 66 = 0 as regression equation of x on y and 40x – 18y = 214, as regression equation of y on x.

Then bxy = $\frac{10}{8}$ and byx = $\frac{40}{18}$

or r^2 = bxy byx = $\frac{10}{8} \; x \; \frac{40}{18} = 2.78$

so r = ± 1.66

Which is wrong as r lies between ± 1.

E 2

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1

19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

Summary statistics for these data are

$n = 25$	$\Sigma\, x = 1314.90$	$\Sigma\, y = 235.70$
$\Sigma\, x^2 = 76.308.53$	$\Sigma\, y^2 = 2286.07$	$\Sigma\, xy = 11824.44$

To estimate the simple linear regression line, we estimate the slope β_1 and intercept $\beta_{0.}$ these estimates are

$$=\beta_1 \Rightarrow \hat{\beta}_1 = b_1 = \frac{n\sum xy - [(\sum x)(\sum y)]}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{25(11{,}824.44) - [(1314.90)(235.70)]}{25(76{,}308.53) - (1314.90)^2}$$

$$= -\ .08$$

$$\beta_{0.} \Rightarrow \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$$

$$= 9.43 - (-\ .08)\ (52.60) = 13.64$$

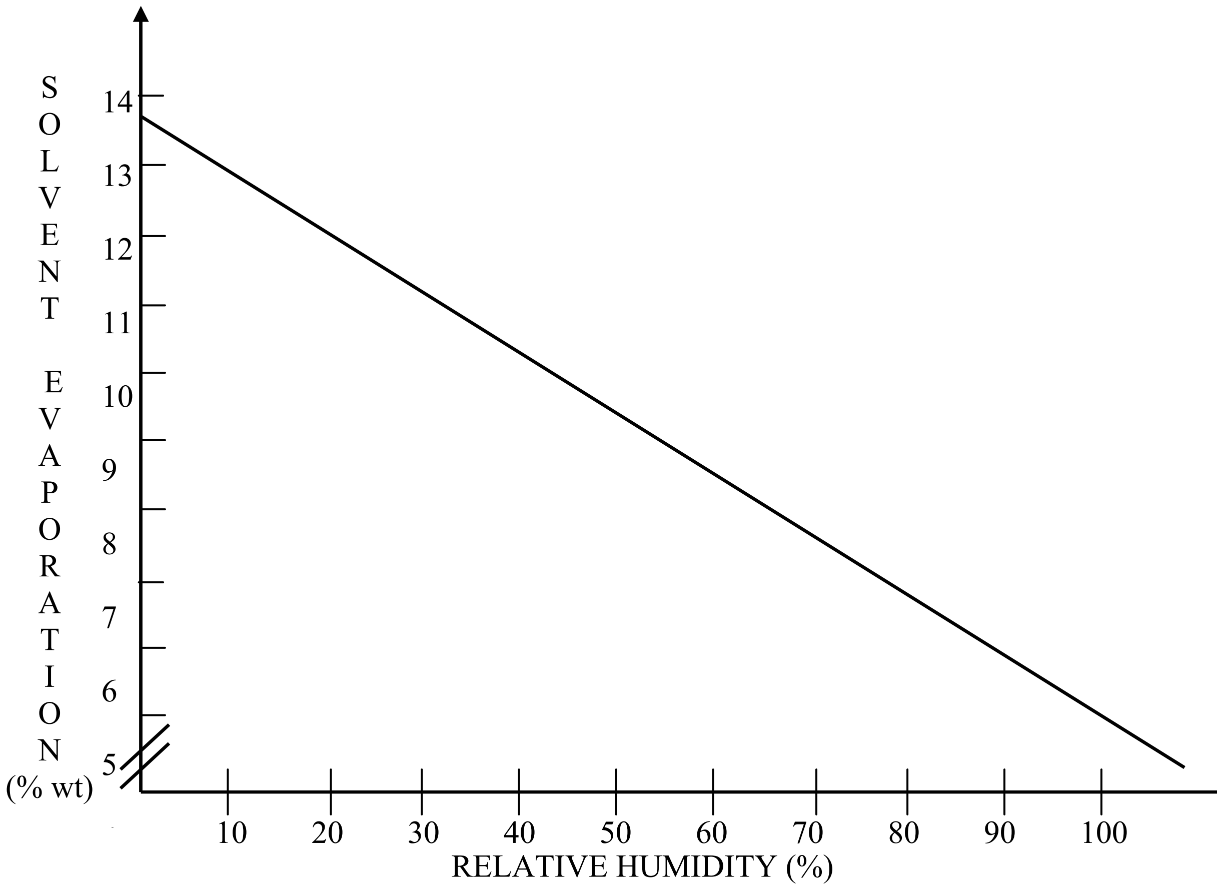

Figure: A graph of the estimated line of regression of Y, the extent of evaporation, on X, the relative humidity

Hence, the estimated regression equation is

$$\hat{\mu}_{Y\,|\,x} = \hat{y} = 13.64 - .08x$$

The graph of this equation is shown in Figure above. To predict the extent of solvent evaporation when the relative humidity is 50 %, we substitute the value 50 for *x* in the equation.

$$\hat{y} = 13.64 - .08x$$

to obtain $\hat{y} = 13.64 - .08(50) = 9.64$. That is, when there relative humidity is 50 % we predict that 9.64% of the solvent, by weight, will be lost due to evaporation.

Recall from elementary calculus that the slope of a line gives the change in *y* for a unit change in *x*. If the slope is positive, then as *x* increases so does *y*; as *x* decreases, so does *y*. If the slope is negative, things operate in reverse. An increase in *x* signals a decrease in *y*, whereas a decrease in *x* yields an increase in *y*.

Check your progress 2

E1 $R = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ = (138.076) / [(2.6929 * 7839.93)$^{1/2}$] = 0.9504

So, R^2 = 0.9033

Check your progress 3

E1 Residual plots are helpful in spotting potential problems. How ever theyare not always easy to interpret. Residual patterns are hard to spot with small data set except in extreme cases, residual plots are most useful with fairly large collection of data.

Check your progress 4

E1 Refer to example solved in the section 3.3.1