Lecture-8
Algorithmic Mathematics(CSC545)

Prepared by Asst. Prof. Bal Krishna Subedi

CDCSIT, TU

Regression Analysis

methods of curve fitting for experimental

In many applications, it often becomes necessary to establish a mathematical relationship between experimental values. This relationship may be used for either testing existing mathematical models or establishing new ones. The mathematical equation can also be used to predict or forecast values of the dependent variable. For example, we would like to know the maintenance cost of an equipment (or a vehicle) as a function of age (or mileage) or the relationship between the literacy level and population growth. The process of establishing such relationships in the form of a mathematical equation is known as *regression analysis* or *curve fitting*.

Suppose the values of y for the different values of x are given. If we want to know the effect of x on y, then we may write a functional relationship

$$y = f(x)$$

The variable y is called the *dependent* variable and x the *independent variable*. The relationship may be either linear or nonlinear as shown in Fig. 10.1. The type of relationship to be used should be decided by the experiment based on the nature of scatteredness of data.

It is a standard practice to prepare a *scatter diagram* as shown in Fig. 10.2 and try to determine the functional relationship needed to fit the points. The line should best fit the plotted points. This means that the

average error introduced by the assumed line should be minimum. The parameters a and b of the various equations shown in Fig. 10.1 should be evaluated such that the equations best represent the data.

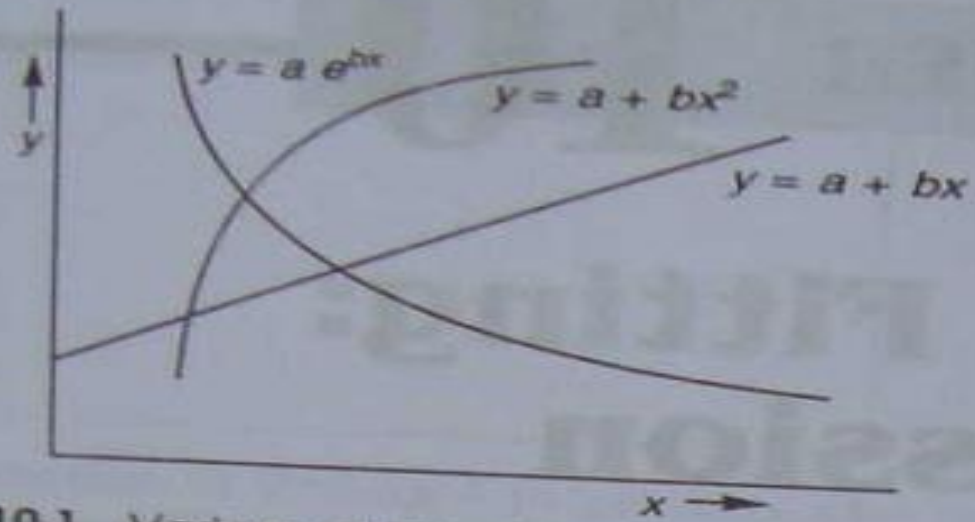

Fig. 10.1 Various relationships between x and y

We shall discuss in this chapter a technique known as *least-squares regression* to fit the data under the following situations:

1. Relationship is linear
2. Relationship is transcendental
3. Relationship is polynomial
4. Relationship involves two or more independent variables

10.2 FITTING

10.2 FITTING LINEAR EQUATIONS

Fitting a straight line is the simplest approach of regression analysis. Let us consider the mathematical equation for a straight line

$$y = a + bx = f(x)$$

to describe the data. We know that a is the intercept of the line and b its slope. Consider a point (x_i, y_i) as shown in Fig. 10.2. The vertical distance of this point from the line $f(x) = a + bx$ is the error q_i. Then,

$$q_i = y_i - f(x_i)$$
$$= y_i - a - bx_i \quad (10.1)$$

There are various approaches that could be tried for fitting a "best" line through the data. They include:

1. Minimise the sum of errors, i.e., minimise

$$\sum q_i = \sum (y_i - a - bx_i) \quad (10.2)$$

2. Minimise the sum of absolute values of errors

$$\sum |q_i| = \sum |(y_i - a - bx_i)| \quad (10.3)$$

3. Minimise the sum of squares of errors

$$\sum q_i^2 = \sum (y_i - a - bx_i)^2 \quad (10.4)$$

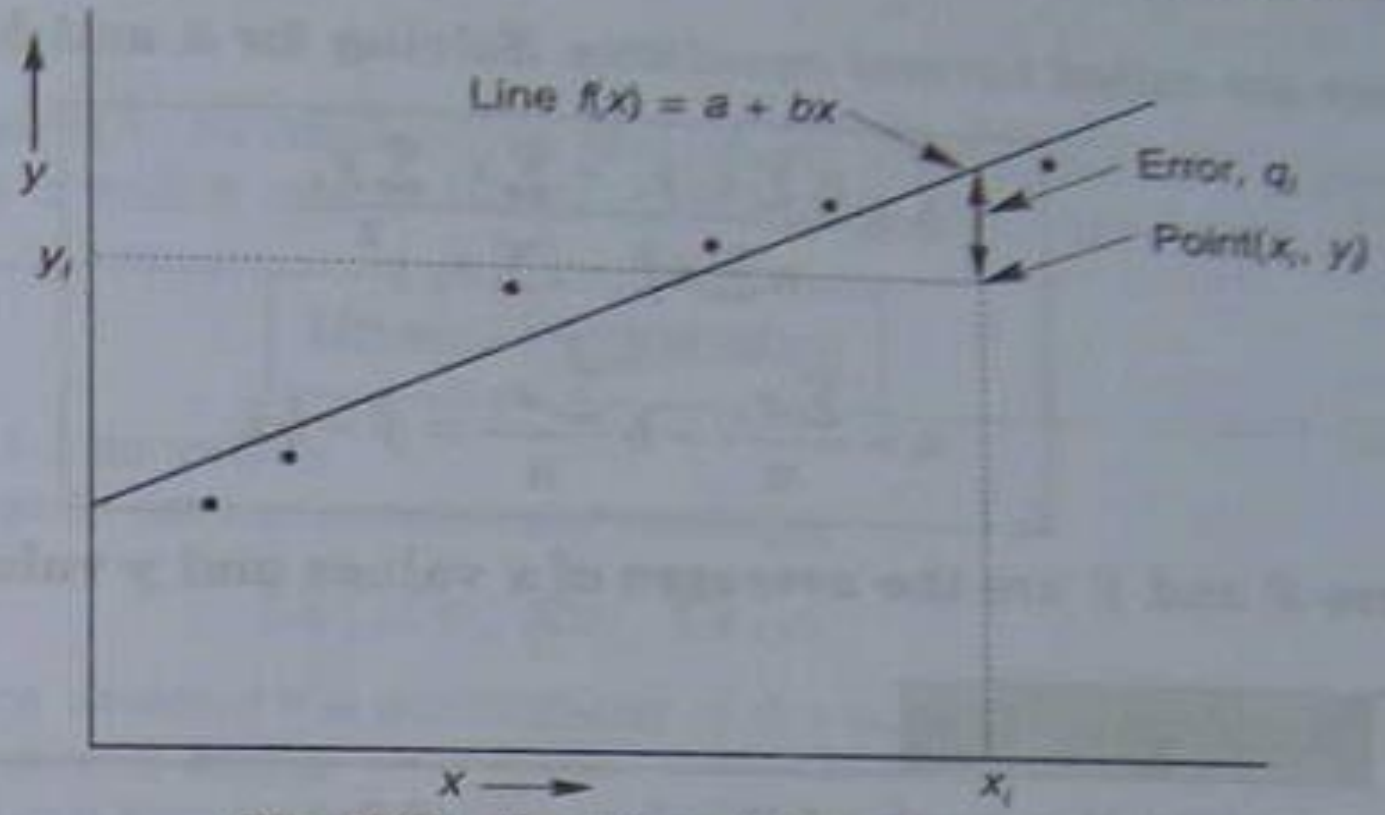

Fig. 10.2 Scatter diagram

It can be easily verified that the first two strategies do not yield a unique line for a given set of data. The third strategy overcomes this problem and guarantees a unique line. The technique of minimising the sum of squares of errors is known as *least squares regression*. In this section we consider the least-squares fit of a straight line.

Least Squares Regression

Let the sum of squares of individual errors be expressed as

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

$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad (10.5)$$

In the method of least squares, we choose a and b such that Q is minimum. Since Q depends on a and b, a necessary condition for Q to be minimum is

$$\frac{\partial Q}{\partial a} = 0 \quad \text{and} \quad \frac{\partial Q}{\partial b} = 0$$

Then

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0 \qquad (10.6)$$

Thus

$$\sum y_i = na + b\sum x_i$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \qquad (10.7)$$

These are called *normal equations*. Solving for a and b, we get

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

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} = \bar{y} - b\bar{x} \qquad (10.8)$$

where $\bar{x}$ and $\bar{y}$ are the averages of x values and y values, respectively.

Fit a straight line to the following set of data

x	1	2	3	4	5
y	3	4	5	6	8

The various summations are given as follows:

	x_i	y_i	x_i^2	$x_i y_i$
	1	3	1	3
	2	4	4	8
	3	5	9	15
	4	6	16	24
	5	8	25	40
Σ	15	26	55	90

Using Eq. (10.8),

$$b = \frac{5 \times 90 - 15 \times 26}{5 \times 55 - 15^2} = 1.20$$

$$a = \frac{26}{5} - 1.20 \times \frac{15}{5} = 1.60$$

Therefore, the linear equation is

$$y = 1.6 + 1.2x$$

$y = 1.6 + 1.2x$

The *regression line* along with the data is shown in Fig. 10.3.

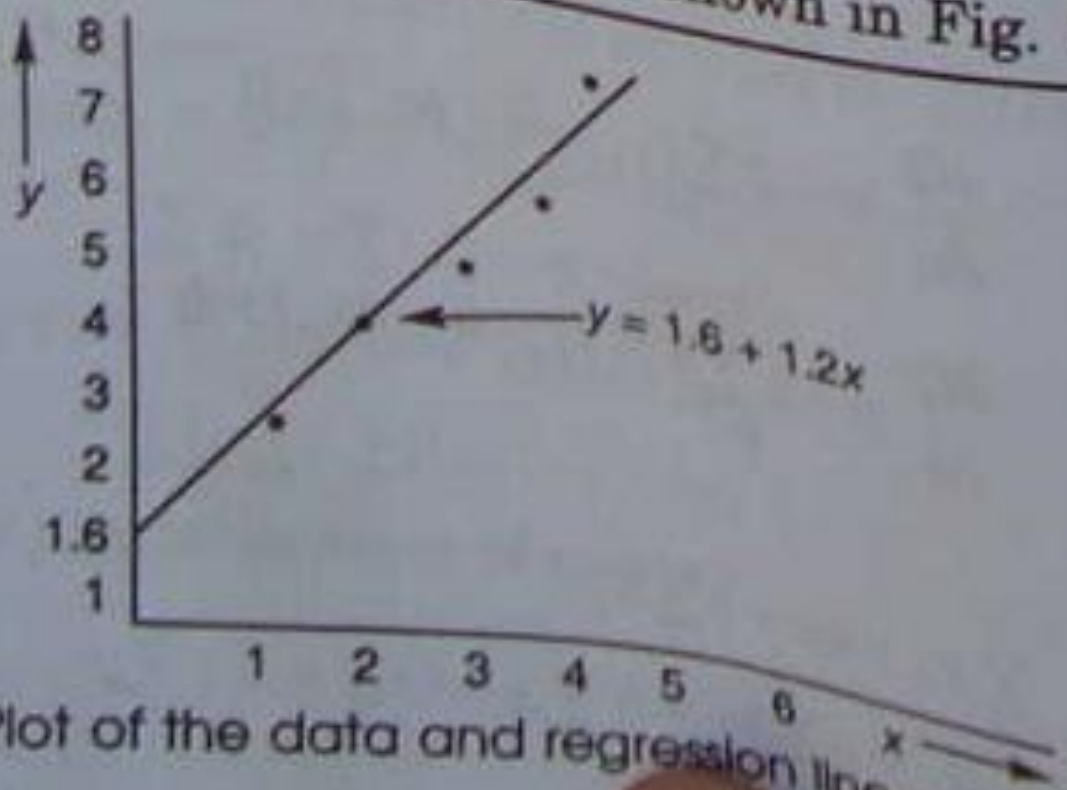

Fig. 10.3 Plot of the data and regression line of example 10.1

Algorithm

It is relatively simple to implement the linear regression on a computer. The coefficients a and b can be evaluated using Algorithm 10.1

Linear Regression

1. Read data values
2. Compute sum of powers and products

$$\Sigma x_i, \Sigma y_i, \Sigma x_i^2, \Sigma x_i y_i$$

3. Check whether the denominator of the equation for b is zero.
4. Compute b and a.
5. Print out the equation.
6. Interpolate data, if required.

Algorithm 10.1

Assignment#8

1. Use least squares regression to fit a straight line to the data.

x	1	3	4	6	8	9	11
y	1	2	4	4	5	7	8

Along with the slope and intercept, also compute the standard error of the estimate.

2. In an organisation, systematic efforts were introduced to reduce the employee absenteeism and results for the first 10 months are shown below:

Months	1	2	3	4	5	6	7	8	9	10
Absentees (per cent)	10	9	9	8.5	9	8	8.5	7	8	7.5

Fit a linear least squares line to the data and from this equation estimate the average weekly reduction in absenteeism.

Thank You

Any Query??