36

Linear regression

36.1 Introduction to linear regression

Regression analysis, usually termed **regression**, is used to draw the line of 'best fit' through co-ordinates on a graph. The techniques used enable a mathematical equation of the straight line form $y = mx + c$ to be deduced for a given set of co-ordinate values, the line being such that the sum of the deviations of the co-ordinate values from the line is a minimum, i.e. it is the line of 'best fit'. When a regression analysis is made, it is possible to obtain two lines of best fit, depending on which variable is selected as the dependent variable and which variable is the independent variable. For example, in a resistive electrical circuit, the current flowing is directly proportional to the voltage applied to the circuit. There are two ways of obtaining experimental values relating the current and voltage. Either, certain voltages are applied to the circuit and the current values are measured, in which case the voltage is the independent variable and the current is the dependent variable; or, the voltage can be adjusted until a desired value of current is flowing and the value of voltage is measured, in which case the current is the independent value and the voltage is the dependent value.

36.2 The least-squares regression lines

For a given set of co-ordinate values, (X_1, Y_1), $(X_2, Y_2), \ldots, (X_N, Y_N)$ let the X values be the independent variables and the Y-values be the dependent values. Also let $D_1, \ldots, D_N$ be the vertical distances between the line shown as PQ in Fig. 36.1 and the points representing the co-ordinate values. The least-squares regression line, i.e. the line of best fit, is the line which makes the value of $D_1^2 + D_2^2 + \cdots + D_N^2$ a minimum value.

The equation of the least-squares regression line is usually written as $Y = a_0 + a_1X$, where a_0 is the Y-axis intercept value and a_1 is the gradient of the line (analogous to c and m in the equation $y = mx + c$). The values of a_0 and a_1 to make the sum of the 'deviations squared' a minimum can be

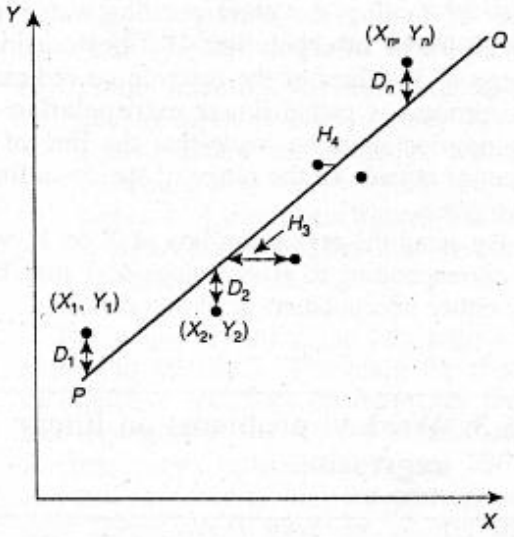

Figure 36.1

obtained from the two equations:

$$\sum Y = a_0N + a_1 \sum X \tag{1}$$

$$\sum (XY) = a_0 \sum X + a_1 \sum X^2 \tag{2}$$

where X and Y are the co-ordinate values, N is the number of co-ordinates and a_0 and a_1 are called the **regression coefficients** of Y on X. Equations (1) and (2) are called the **normal equations** of the regression lines of Y on X. The regression line of Y on X is used to estimate values of Y for given values of X. If the Y-values (vertical-axis) are selected as the independent variables, the horizontal distances between the line shown as PQ in Fig. 36.1 and the co-ordinate values (H_3, H_4, etc.) are taken as the deviations. The equation of the regression line is of the form: $X = b_0 + b_1Y$ and the normal equations become:

$$\sum X = b_0N + b_1 \sum Y \tag{3}$$

$$\sum (XY) = b_0 \sum Y + b_1 \sum Y^2 \tag{4}$$

where X and Y are the co-ordinate values, b_0 and b_1 are the regression coefficients of X on Y and

N is the number of co-ordinates. These normal equations are of the regression line of X on Y, which is slightly different to the regression line of Y on X. The regression line of X on Y is used to estimated values of X for given values of Y. The regression line of Y on X is used to determine any value of Y corresponding to a given value of X. If the value of Y lies within the range of Y-values of the extreme co-ordinates, the process of finding the corresponding value of X is called **linear interpolation**. If it lies outside of the range of Y-values of the extreme co-ordinates than the process is called **linear extrapolation** and the assumption must be made that the line of best fit extends outside of the range of the co-ordinate values given.

By using the regression line of X on Y, values of X corresponding to given values of Y may be found by either interpolation or extrapolation.

36.3 Worked problems on linear regression

Problem 1. In an experiment to determine the relationship between frequency and the inductive reactance of an electrical circuit, the following results were obtained:

Frequency (Hz)	Inductive reactance (ohms)
50	30
100	65
150	90
200	130
250	150
300	190
350	200

Determine the equation of the regression line of inductive reactance on frequency, assuming a linear relationship.

Since the regression line of inductive reactance on frequency is required, the frequency is the independent variable, X, and the inductive reactance is the dependent variable, Y. The equation of the regression line of Y on X is:

$$Y = a_0 + a_1 X,$$

and the regression coefficients a_0 and a_1 are obtained by using the normal equations

$$\sum Y = a_0 N + a_1 \sum X$$

and $$\sum XY = a_0 \sum X + a_1 \sum X^2$$

(from equations (1) and (2))

A tabular approach is used to determine the summed quantities.

Frequency, X	Inductive reactance, Y	X^2	XY	Y^2
50	30	2500	1500	900
100	65	10000	6500	4225
150	90	22500	13500	8100
200	130	40000	26000	16900
250	150	62500	37500	22500
300	190	90000	57000	36100
350	200	122500	70000	40000
$\sum X = 1400$	$\sum Y = 855$	$\sum X^2 = 350000$	$\sum XY = 212000$	$\sum Y^2 = 128725$

The number of co-ordinate values given, N is 7. Substituting in the normal equations gives:

$$855 = 7a_0 + 1400a_1 \quad (1)$$

$$212000 = 1400a_0 + 350000a_1 \quad (2)$$

1400 × (1) gives:

$$1197000 = 9800a_0 + 1960000a_1 \quad (3)$$

7 × (2) gives:

$$1484000 = 9800a_0 + 2450000a_1 \quad (4)$$

(4) − (3) gives:

$$287000 = 0 + 490000a_1$$

from which, $a_1 = \dfrac{287000}{490000} = 0.586$

Substituting $a_1 = 0.586$ in equation (1) gives:

$$855 = 7a_0 + 1400(0.586)$$

i.e. $$a_0 = \frac{855 - 820.4}{7} = 4.94$$

Thus the equation of the regression line of inductive reactance on frequency is:

$$Y = 4.94 + 0.586X$$

Problem 2. For the data given in Problem 1, determine the equation of the regression line of frequency on inductive reactance, assuming a linear relationship.

In this case, the inductive reactance is the independent variable X and the frequency is the dependent variable Y. From equations 3 and 4, the equation of the regression line of X on Y is:

$$X = b_0 + b_1 Y,$$

and the normal equations are

$$\sum X = b_0 N + b_1 \sum Y$$

and $$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

From the table shown in Problem 1, the simultaneous equations are:

$$1400 = 7b_0 + 855b_1$$

$$212000 = 855b_0 + 128725b_1$$

Solving these equations in a similar way to that in Problem 1 gives:

$$b_0 = -6.15$$

and $b_1 = 1.69$, correct to 3 significant figures

Thus the equation of the regression line of frequency on inductive reactance is:

$$X = -6.15 + 1.69Y$$

Problem 3. Use the regression equations calculated in Problems 1 and 2 to find (a) the value of inductive reactance when the frequency is 175 Hz and (b) the value of frequency when the inductive reactance is 250 ohms, assuming the line of best fit extends outside of the given co-ordinate values. Draw a graph showing the two regression lines.

(a) From Problem 1, the regression equation of inductive reactance on frequency is $Y = 4.94 + 0.586X$. When the frequency, X, is 175 Hz, $Y = 4.94 + 0.586(175) = 107.5$, correct to 4 significant figures, i.e. the inductive reactance is **107.5 ohms** when the frequency is 175 Hz

(b) From Problem 2, the regression equation of frequency on inductive reactance is $X = -6.15 + 1.69Y$. When the inductive reactance, Y, is 250 ohms, $X = -6.15 + 1.69(250) = 416.4$ Hz, correct to 4 significant figures, i.e. the frequency is **416.4 Hz** when the inductive reactance is 250 ohms.

The graph depicting the two regression lines is shown in Fig. 36.2. To obtain the regression line of inductive reactance on frequency the regression line equation $Y = 4.94 + 0.586X$ is used, and X (frequency) values of 100 and 300 have been selected in order to find the corresponding Y values. These values gave the co-ordinates as (100, 63.5) and (300, 180.7), shown as points A and B in Fig. 36.2. Two co-ordinates for the regression line of frequency on inductive reactance are calculated using the equation $X = -6.15 + 1.69Y$, the values of inductive reactance of 50 and 150 being used to obtain the co-ordinate values. These values gave co-ordinates (78.4, 50) and (247.4, 150), shown as points C and D in Fig. 36.2.

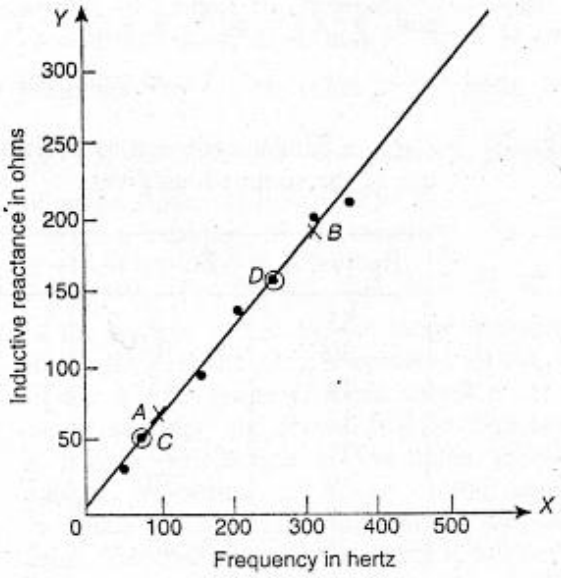

Figure 36.2

It can be seen from Fig. 36.2 that to the scale drawn, the two regression lines coincide. Although it is not necessary to do so, the co-ordinate values are also shown to indicate that the regression lines do appear to be the lines of best fit. A graph showing co-ordinate values is called a **scatter diagram** in statistics.

Problem 4. The experimental values relating centripetal force and radius, for a mass travelling at constant velocity in a circle, are as shown:

Force (N)	5	10	15	20	25	30	35	40
Radius (cm)	55	30	16	12	11	9	7	5

Determine the equations of (a) the regression line of force on radius and (b) the regression line of radius on force. Hence, calculate the force at a radius of 40 cm and the radius corresponding to a force of 32 newtons.

Let the radius be the independent variable X, and the force be the dependent variable Y. (This decision is usually based on a 'cause' corresponding to X and an 'effect' corresponding to Y).

(a) The equation of the regression line of force on radius is of the form $Y = a_0 + a_1X$ and the constants a_0 and a_1 are determined from the normal equations:

$$\sum Y = a_0N + a_1\sum X$$

and $$\sum XY = a_0\sum X + a_1\sum X^2$$

(from equations (1) and (2))

Using a tabular approach to determine the values of the summations gives:

Radius, X	Force, Y	X^2	XY	Y^2
55	5	3025	275	25
30	10	900	300	100
16	15	256	240	225
12	20	144	240	400
11	25	121	275	625
9	30	81	270	900
7	35	49	245	1225
5	40	25	200	1600
$\sum X = 145$	$\sum Y = 180$	$\sum X^2 = 4601$	$\sum XY = 2045$	$\sum Y^2 = 5100$

Thus $180 = 8a_0 + 145a_1$

and $2045 = 145a_0 + 4601a_1$

Solving these simultaneous equations gives $a_0 = 33.7$ and $a_1 = -0.617$, correct to 3 significant figures. Thus the equation of the regression line of force on radius is:

$$\boldsymbol{Y = 33.7 - 0.617X}$$

(b) The equation of the regression line of radius on force is of the form $X = b_0 + b_1Y$ and the constants b_0 and b_1 are determined from the normal equations:

$$\sum X = b_0N + b_1\sum Y$$

and $$\sum XY = b_0\sum Y + b_1\sum Y^2$$

(from equations (3) and (4))

The values of the summations have been obtained in part (a) giving:

$$145 = 8b_0 + 180b_1$$

and $$2045 = 180b_0 + 5100b_1$$

Solving these simultaneous equations gives $b_0 = 44.2$ and $b_1 = -1.16$, correct to 3 significant figures. Thus the equation of the regression line of radius on force is:

$$\boldsymbol{X = 44.2 - 1.16Y}$$

The force, Y, at a radius of 40 cm, is obtained from the regression line of force on radius, i.e. $y = 33.7 - 0.617(40) = 9.02$,

i.e. **the force at a radius of 40 cm is 9.02 N.**

The radius, X, when the force is 32 newtons is obtained from the regression line of radius on force, i.e. $X = 44.2 - 1.16(32) = 7.08$,

i.e. **the radius when the force is 32 N is 7.08 cm.**