4.6.6 Correlation and regression

When two variables are so related that a change in one is accompanied by a change in the other in such a way that an increase in one is accompanied by an increase or decrease in the other, or decrease in one by a decrease or increase in the other, and the greater the magnitude of change in one, the greater the magnitude of change in the other, then the variables are said to be correlated.

The correlation is said to be positive when an increase or decrease in one results in an increase or decrease in the other. It is said to be negative when an increase in one corresponds to a decrease in the other or vice versa.

Of the two variables, one is said to be a dependent variable while the other is the independent variable. The term regression is used because we regress towards the mean relationship of the variables. If we consider these variables as X and Y, it can be shown that the regression equation of Y on X is given by:

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \tag{7.49}$$

and that of X on Y as

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \tag{7.50}$$

where r is known as the correlation coefficient and is written as

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_X \times \sigma_Y} \tag{7.51}$$

The value of r range between +1 and –1 but is never equal to 1. If r is close to +1 it means a perfect positive correlation and if it is close to –1 it indicates a perfect negative correlation. An example is shown in box 7.15.

Figure 7.7 shows the graph of the regression equation thus obtained. The question now is: could the value of r = - 0.9769 have been obtained solely as a result of chance?

Table 7.5 gives the probability table for a two variable situation. The number of X and Y pairs appear in the left hand column while the probability values for different values of n are listed in the left hand columns. The degree of freedom is taken as n – 2. For instance in our example in box 7.15 the number of datasets are seven so the degrees of freedom is equal to 5. With 5 degrees of freedom and at a probability level of 0.001 the value of r is given as 0.951. That is this value will appear by mere chance about one in a thousand separate analysis. In the from our data r = - 0.9769, is better than this. Hence we write 0.001> P. We are therefore justified in concluding that the relationship is statistically real. The correlation is said to be significant at a better than 0.001 level.

Box 7.15 Problem: Calculate the correlation between temperature and oxygen from the following data:

Temp°C :	0	5	10	15	20	25	30
O_2 mg l^{-1}:	14.6	12.8	11.4	10.2	9.31	8.54	7.86

Solution:

Temp°C X	O_2 mg l^{-1} Y	X^2	Y^2	XY
0	14.6	0	213.16	0
5	12.8	25	163.84	64
10	11.4	100	129.96	114
15	10.2	225	104.04	153
20	9.31	400	86.68	186.2
25	8.54	625	72.93	213.5
30	7.86	900	61.78	235.8
$\sum X = 105$	$\sum Y = 74.71$	$\sum X^2 = 2275$	$\sum Y^2 = 832.9$	$\sum XY = 966.5$

$$\bar{X} = 15 \quad \bar{Y} = 10.67 \quad \overline{X^2} = 325 \quad \overline{Y^2} = 138.07$$

$$\sigma_X = \sqrt{\overline{X^2} - (\bar{X})^2} = \sqrt{325 - 225} = \sqrt{100} = 10$$

$$\sigma_Y = \sqrt{\overline{Y^2} - (\bar{Y})^2} = \sqrt{118.91 - 113.85} = \sqrt{5.07} = 2.25$$

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_X \times \sigma_Y} = \frac{138.07 - 160.05}{10 \times 2.25} = \frac{-21.98}{22.50} \text{ or } r = -0.9769$$

since r is negative it indicates a negative correlation that is the oxygen concentration decreases as temperature increases and vice versa. Our regression equation of the data provided can be formulated as:

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \text{or} \quad Y - 10.67 = -0.9769 \times \frac{2.25}{10}(X - 15)$$

so $Y - 10.67 = -0.2198(X - 15)$ or $Y - 10.67 = -0.2198X + 3.2970$

or $Y = -0.2198X + 13.967$

Regression equation of Y on X is: $Y = a + bx$

The values of a and b can be determined by solving the following normal equations

$$\Sigma Y = Na + b \Sigma X \quad (7.52)$$

and $$\Sigma XY = a \Sigma X + b \Sigma X^2 \quad (7.53)$$

In the data provided in box 7.15, the values for the different variables are $N = 7$; $\Sigma X = 105$; $\Sigma Y = 74.71$; $\Sigma X^2 = 2275$ and $\Sigma XY = 966.5$
The values of a and b can be determined by putting the values in equation 7.52 and 7.53 as follows:

$74.71 = 7a + 105b$ (1)
$966.5 = 105a + 2275b$ (2)
multiplying equation (1) by 105 and (2) by 7 we have

$7844.55 = 735a + 11025b$ (3)
$- 6765.50 = 735a + 15925b$ (4)

deducting equation (4) from (3) we have $b = - 4.541$
putting the value of b in (1) we can equate $a = 78.8$
Therefore the straight line equation can be written as $Y = 78.8 - 4.541X$

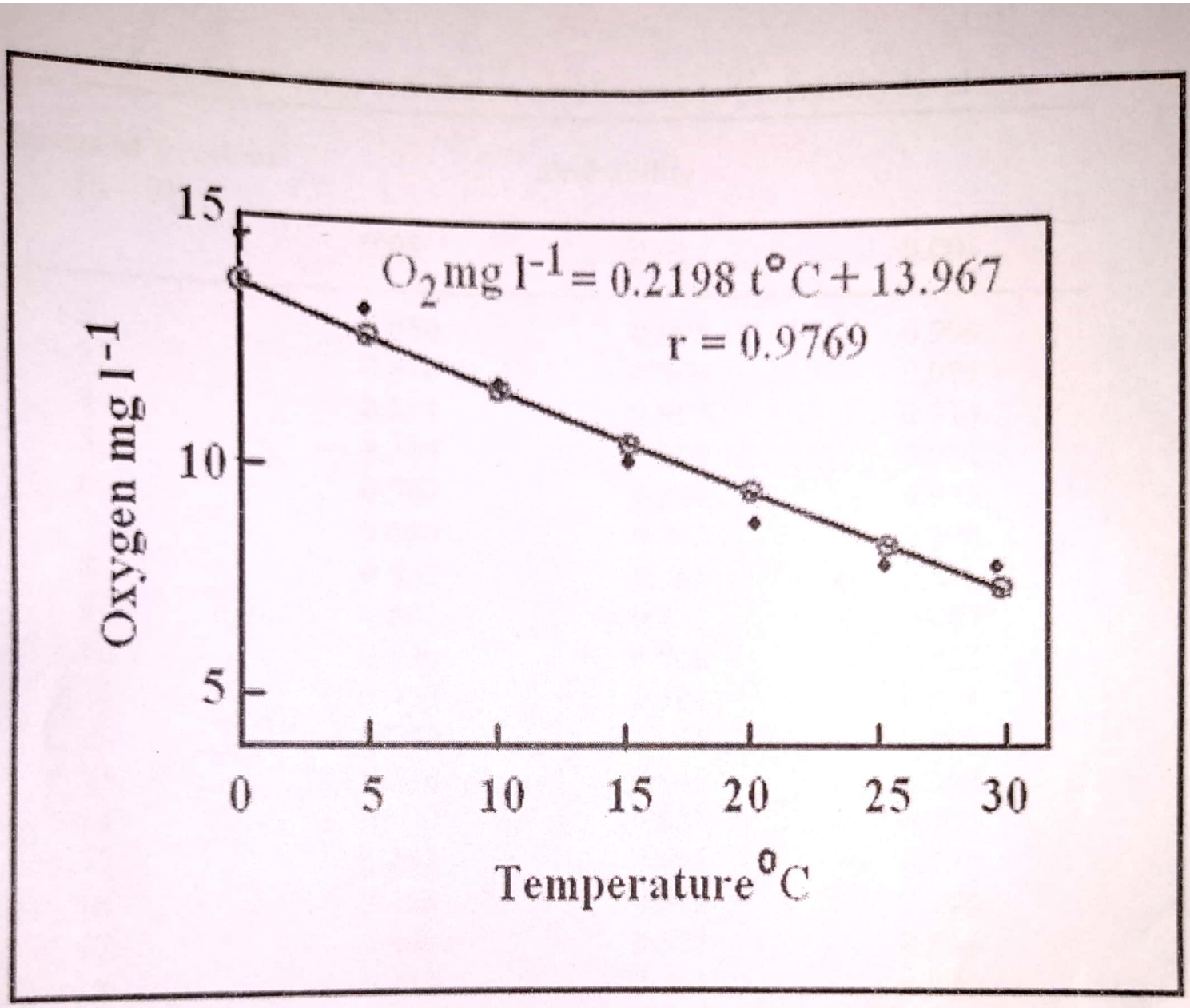

Figure 7.7 Regression line of oxygen versus temperature.