-8-

Correlation and Regression

8.1 Nature of correlation

In many aspects of engineering applications, it is important to decide about the presence of some correlation between two or more variables. This is particularly true when there is an intuitive feeling of the presence of a correlation. For example, a field engineer notices that the incidence of failure of electrical insulators fitted to high voltage wire feeding an electrostatic precipitator increases with an increase in inlet gas temperature. He "feels" that there is some correlation between the two factors. The purpose of the present section is to introduce the concept of correlation and the method of its estimation in case of two or more variables. The simplest case is that of linear correlation between two variables.

The nature of correlation between two variables x and y can follow any of the five schemes shown in Figure (8.1) known as **scatter diagrams**. In the first case (a), is shown a perfect linear correlation, while in (b), data suggest an increasing character of correlation. In (c), the correlation is very poor, while in (d), there is an inverse correlation, and it becomes a perfect linear decreasing correlation in (e).

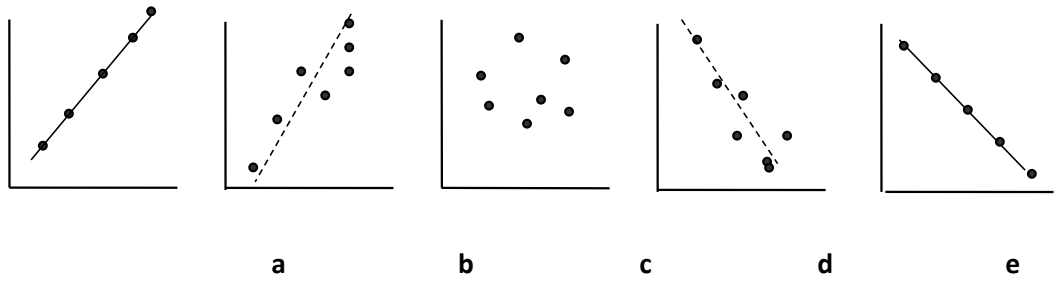

Fig (8.1): Types of linear correlations

8.1.1 The Pearson correlation coefficient

The extent to which the observed data fit to a linear correlation has been quantified by **Pearson**. He first stated that the type of correlation should be independent of the origin chosen and the scale used. So, he suggested that all *x* and *y* values be normalized in a way like that done in the normal distribution by defining:

$$X_i = \frac{x_i - \bar{x}_i}{s_x} \text{ and } Y_i = \frac{y_i - \bar{y}_i}{s_y}$$

He then suggested that the extent of correlation be calculated from the rule:

$$R = \frac{\Sigma\, Xi.Yi}{n}$$

This factor (R) is termed the linear correlation coefficient.

An easier way of computing this coefficient is by expanding the terms of the above definition.

$$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{s_x.s_y} = \frac{\sum_{i=1}^{n} (x_i.y_i - x_i.\bar{y} - \bar{x}y_i + \bar{x}.\bar{y})}{s_x.s_y}$$

Recalling that $\sum_{i=1}^{n} x_i = n.\bar{x}$ and that $\sum_{i=1}^{n} y_i = n.\bar{y}$ as well as the definition of the standard deviation from equation (1.7), we get the following form for the linear correlation coefficient:

$$R = \frac{n.\sum_{i=1}^{n} x_i.y_i - \sum_{i=1}^{n} x_i.\sum_{i=1}^{n} y_i}{\sqrt{n.\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \times \sqrt{n.\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \qquad (8.1)$$

It can be generally proved that: $-1 \le R \le 1$.

We note that the numerator in the definition:

$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{s_x.s_y}$ is equivalent to a covariance term. (See Chapter 5)

The value of R is 1 for perfectly increasing linear correlation (Case a in Figure (8.1), and –1 for a perfectly linear decreasing correlation (Case e). Generally speaking, the more the value of $|R|$ approaches unity, the higher is the extent to which points gather about a straight line. The case of poor correlation (c) would correspond to a value of R close to zero.

Currently, values of correlation coefficient (R) are readily obtained using the EXCEL function "correlation" (= CORELL).

Example 8.1

The following table relates the scores obtained by 16 students in two different exams A and B (Out of 20). Plot the scatter diagram then estimate the correlation coefficient.

A	17	4	8	12	15	13	3	10	18	9	12	5	19	16	12	14
B	14	7	6	16	14	10	3	14	17	5	10	6	20	17	9	15

Solution:

The scatter diagram is shown in Figure (8.2)

The value of R can be directly obtained from the CORELL function that yields ***R* = 0.8715**

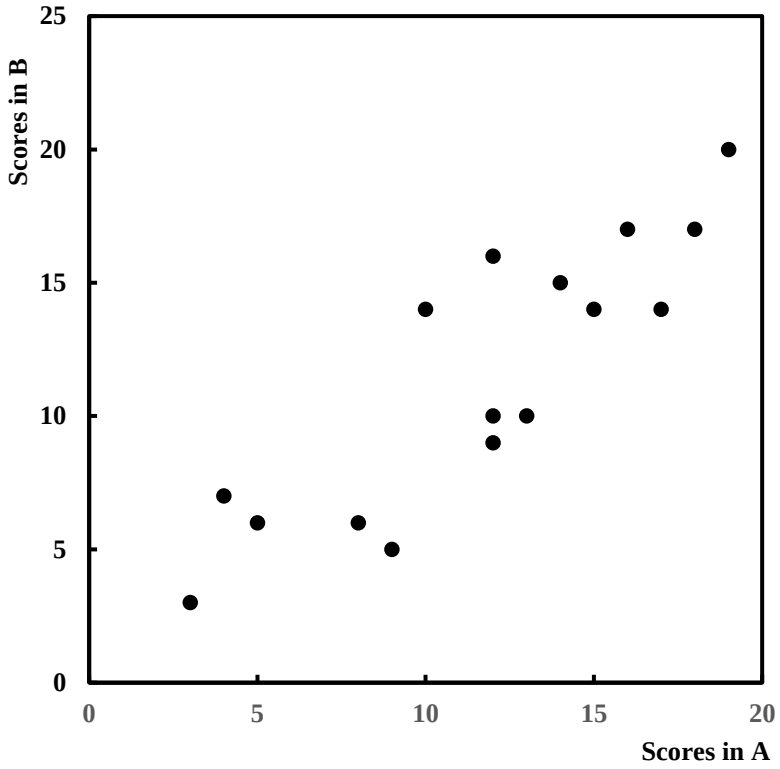

Fig (8.2): Scatter diagram of example (8.1)

8.1.2 Confidence interval of *R*: The Fisher method

The value of *R* calculated in the previous example using equation (8.1) has been obtained from sample data. The question that arises is: How far does this value represent the correlation between *x* and *y* for the whole population. The correlation coefficient for population is designated as ρ. For relatively large sample size ($n > 15$), it may be assumed that errors are normally distributed along samples taken from the population. The confidence interval of ρ is:

$$z_1 = 0.5 ln \frac{1+R}{1-R} - \frac{z_{crit}}{\sqrt{n-3}}. \quad (8.2)$$

$$z_2 = 0.5 ln \frac{1+R}{1-R} + \frac{z_{crit}}{\sqrt{n-3}}. \quad (8.3)$$

Where, z_{crit} is the critical z – value obtained from the function NORM.S.INV at a confidence level $0.5(1 + L)$.

The limits of ρ are then obtained from the following expression:

$$\tanh z_1 < \rho < \tanh z_2 \quad (8.4)$$

For example, if this method is applied on example (8.1) case where $R = 0.826$ and $n = 16$, at significance level = 0.05, we get from equations (8.2) and (8.3):

$$z_1 = 0.5 \ln \frac{1+0.8715}{1-0.8715} - \frac{1.96}{\sqrt{16-3}} = 0.7957$$

$$z_2 = 0.5 \ln \frac{1 + 0.8715}{1 - 0.8715} + \frac{1.96}{\sqrt{16 - 3}} = 3.222$$

From equation (8.4):

tanh 0.7975 < ρ < tanh 3.222

0.663 < ρ < 0.997

8.1.3 Testing hypothesis for correlation coefficient

When the numerical value of R is close to 1 or to zero, then it is easy to make inference about the strength of correlation; that is, if R = 0.95, for example, then one can safely say that there is a strong correlation between the two variables. The same is true if R = 0.1, where practically no correlation exists.

In the event of obtaining inconclusive values of R (such as 0.5 for example), it is necessary to find a way of telling whether the correlation exists or is absent. To this aim, a test is conducted where the null hypothesis involves assuming no correlation at all, that is: H_0: $\rho = 0$

For relatively large sample size (n > 15), the test statistic is:

$$z = \frac{|R|.\sqrt{n}}{\sqrt{1 - R^2}} \qquad (8.5)$$

While for smaller sample size, it is:

$$t = \frac{|R|.\sqrt{n-2}}{\sqrt{1 - R^2}} \qquad (8.6)$$

With n – 2 degrees of freedom.

For instance, in Example (8.1), n = 16 and R = 0.8715. The null hypothesis is H_0: $\rho = 0$

From equation (8.6):

$$t = \frac{0.8715 \times \sqrt{16 - 2}}{\sqrt{1 - 0.8715^2}} = 6.649$$

The critical value of t at α = 0.05 and *d.f.* = 8 as obtained from the function T.INV.2T = 2.306.

Since 6.649 > 2.306 then the null hypothesis is not accepted meaning that the presence of a correlation between the scores in the two exams for the whole population cannot be rejected.

8.2 Linear regression

8.2.1 Equation of the regression line

The interpretation of the value of R obtained is better understood by obtaining the equation of the best straight line passing between the points.

As can be seen from Figure (8.2), the points seem to point out an increasing relation between the two variables. A straight line can be fitted to pass between these points. The best fit is obtained when the sum of squares of differences between the actual and the calculated value of y is a minimum value. The straight line obtained is called **the regression line.**

Let the equation of that line be: $y = a.x + b$

The value of y corresponding to an entry x_i calculated from the above equation is $y_c = a.x_i + b$

If the actual value of y corresponding to x_i is y_i, then the deviation between the actual and the calculated value is:

$D_i = y_i - (a.x_i + b)$

The best line is obtained when:

$\Sigma D_i^2 = \Sigma [y_i - (a.x_i + b)]^2$ is a minimum value. (8.7)

To obtain the values of the constants a and b, the following sets of equations have to be solved together:

$\frac{\partial \Sigma D_i^2}{\partial a} = 0$ and $\frac{\partial \Sigma D_i^2}{\partial b} = 0$

Developing the RHS of equation (8.7) we get:

$\Sigma D_i^2 = \Sigma [y_i^2 + (a.x_i + b)^2 - 2.y_i .(a.x_i + b)]$

$= \Sigma y_i^2 + a^2. \Sigma x_i^2 + 2.a.b.\Sigma x_i + n.b^2 - 2.a. \Sigma x_i .y_i - 2.b.\Sigma y_i$

Performing partial differentiation with respect to a, we get;

$2.a.\Sigma x_i^2 + 2.b.\Sigma x_i - 2.\Sigma x_i .y_i = 0$

Also, performing partial differentiation wrt b we get:

$2.a.\Sigma x_i + 2.n.b - 2.\Sigma y_i = 0$

Solving the above two equations for a and b, we get:

$$a = \frac{n.\sum_{i=1}^{n} x_i.y_i - \sum_{i=1}^{n} x_i . \sum_{i=1}^{n} y_i}{n.\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (8.8)$$

$$b = \frac{\sum_{i=1}^{n} y_i - a.\sum_{i=1}^{n} x_i}{n} \quad (8.9)$$

The EXCEL program readily displays the regression equation through the command: "Add trend line" by right clicking on any point on the scatter diagram of Figure (8.2).

Example 8.3
Find the equation of the regression line of the data in example (8.1)

Solution:
The equation relating the marks obtained in exam A (x_A) to those of exam B (x_B) is directly obtained by the aforementioned EXCEL command.

$$x_B = 0.9052\, x_A + 0.8575$$

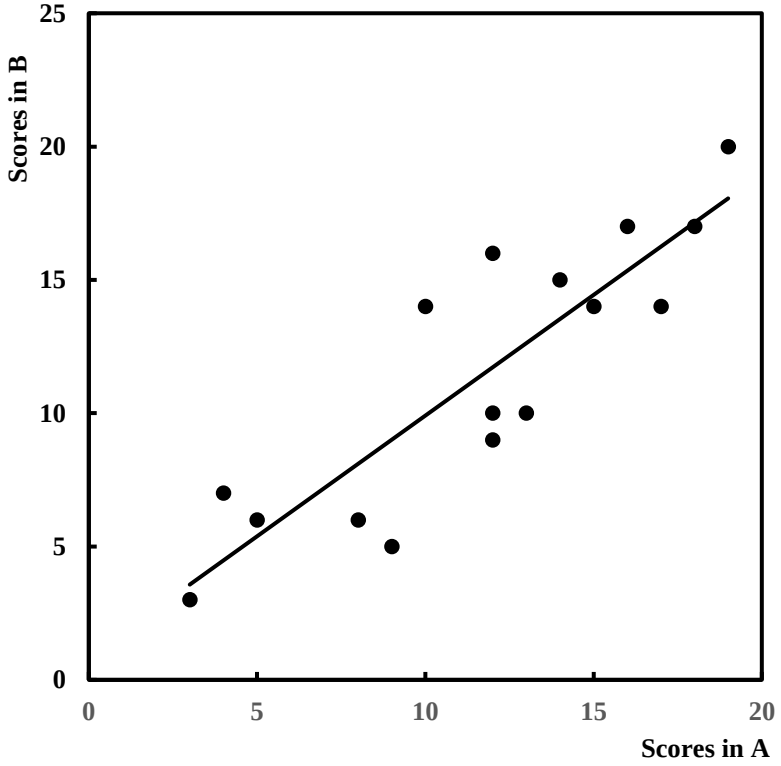

Fig (8.3): Regression line with its equation

8.2.2 The standard error of estimate

The regression line gives an estimate about the average relationship between the two variables for the set of given points. Had another set of points been chosen, the equation would have been different.

In the following section is shown a method to estimate how far does this relation represents the actual relation between variables. In other words, the errors associated with using the regression line equation, rather than the actual values, will be computed.

Let the observed values of y corresponding to values of x_i be y_i, and the calculated values $y_{ci} = a.x_i + b$

Let $\bar{y}$ be the mean value of observed value = $\Sigma\, y_i / n$

There are three types of deviations that can be considered:

- The deviation of the mean from any observed value: $D_i = \bar{y} - y_i$
- The deviation between observed and calculated values, known as **unexplained deviation** or **residual**: $D_{si} = y_i - y_c$

- The deviation between calculated values and the mean value, known as **explained deviation**: $D_{ci} = y_c - \bar{y}$

The following definitions are related to the above differences.

- **The total variation**: $$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (8.10)$$
- **The unexplained variation**: $$\sum_{i=1}^{n} D_{si}^2 = \sum_{i=1}^{n} (y_i - y_c)^2 \qquad (8.11)$$
- **The explained variation**: $$\sum_{i=1}^{n} D_{ci}^2 = \sum_{i=1}^{n} (y_c - \bar{y})^2 \qquad (8.12)$$

Finally, it can be shown that:

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} D_{si}^2 + \sum_{i=1}^{n} D_c^2 \qquad (8.13)$$

The above definitions can be understood in the light of the following discussion: When the value of the independent variable x is higher than the mean values $\bar{x}$ of observations, we expect in an increasing relation that the value of y will also be higher than that of the mean $\bar{y}$. This expected difference between y_c and $\bar{y}$ is hence called explained difference. The difference between observed and calculated values is termed unexplained since it will probably be associated with other factors than the independent variable under consideration.

The square root of the unexplained variance is a type of standard deviation known as the **standard error of estimate**. This is computed from the definition:

$$s_{yx} = \sqrt{\frac{\Sigma(y_c - y_i)^2}{n-2}} \qquad (8.14)$$

The calculation of this quantity is necessary to set the limits of accuracy of the calculated values of y to any desired confidence level. It is assumed that the errors are normally distributed about their mean value.

Let the required significance level be α corresponding to a value of z_{crit} as obtained from Table (6.1) (or from NORMINVS) , then the expected limits of the values of y corresponding to x_i, can be determined from:

$$y_c - z_{crit}.\, s_{yx} < y < y_c + z_{crit}.\, s_{yx} \qquad (8.15)$$

Note that the standard error of estimate can be directly obtained in EXCEL from the function STEYX.

8.2.3 The coefficient of determination

The extent to which the variables x and y are correlated can be estimated by obtaining the ratio of explained variation to the total variation: This is a positive number known as the determination coefficient.

$$R^2 = \frac{\sum_{i=1}^{n}(y_c - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (8.16)$$

For linear regression this value is directly obtained after the line is plotted by clicking on the command "Display R^2 value on chart".

In example (8.4), its value = $R^2 = 0.631$.

This means that 63.1% of the variation in strength is due to the variation in water to cement ratio while the remaining 36.9% are due to other factors including experimental errors.

8.2.5 Limits of accuracy of experimental plots and Error bars

It is common in experimental runs to repeat the experiment more than once and obtain the mean value of each reading. The extent of the accuracy of the mean values is usually determined using bar charts.

Since the number of experimental points is usually less than 30, the following section addresses errors in terms of *t* – distribution rather than normal.

As pointed out to in Chapter 5, the error in determining the mean value of population is $t.\ \frac{s}{\sqrt{r-1}}$ (Equation 5.7).

The value of *t* is obtained from the function T.INV2T (0.05, r – 1), and r represents the number of replicate results under the same experimental conditions. The standard deviation of these results = *s*.

Consider the following data related to the determination of viscosity of a lube oil at 7 different temperatures, each run being repeated three times.

Run	**1**	**2**	**3**	**mean**	***s***	***t***	**error**
***T*°C**	**Viscosity cP**						
20	53	53	50	52	1.732	4.303	5.270
25	48	45	47	46.667	1.5275	4.303	4.647
30	42	40	43	41.667	1.5275	4.303	4.6474
35	40	38	37	38.333	1.5275	4.303	4.6474
40	37	35	35	35.667	1.1547	4.303	3.5131
45	33	34	36	34.333	1.5275	4.303	4.6474
50	29	30	30	28.667	0.5774	4.303	1.7565

To place the error bars on the linear regression curve, we plot the mean value against temperature, then on the "design" option, choose "Add Chart Element" then "Error Bars". We choose "custom" followed by "specify value" then drag the error values for both positive and negative errors. Vertical error bars appear on the chart indicating the upper and lower confidence limits of each run. (Figure 8.6)

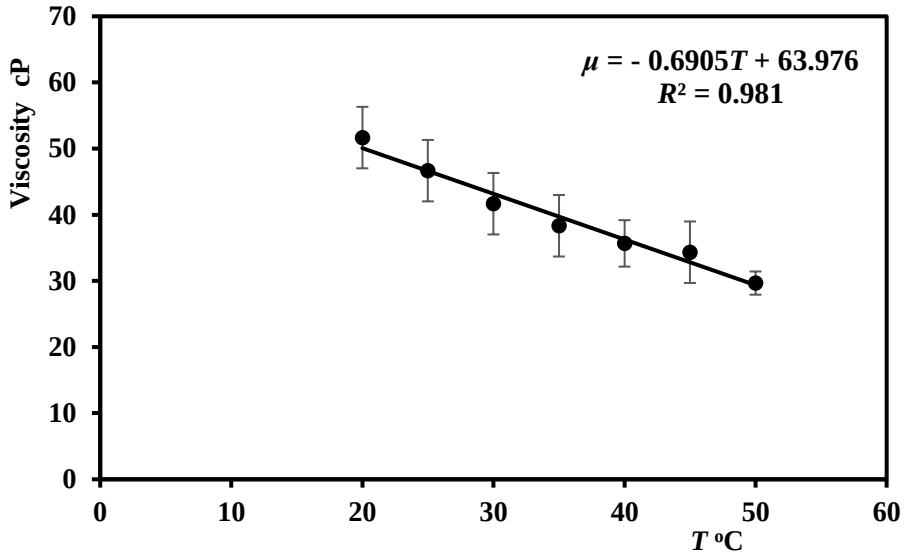

Fig (8.6): Error bars

Finally, to decide whether any of the experimental runs would be considered irrelevant, two straight lines are drawn according to equation (8.15) using the obtained critical t – value instead of z. The standard error of estimate as calculated for experimental viscosity mean values (y) and temperature (x) from the function STEYX = 1.377. With t_{crit} = 4.303, equation (8.15) is written as:

$y_c - 5.92 < y < y_c + 5.92$

This means subtracting and adding to each calculated value of viscosity the number 5.92 to obtain the following table.

T°C	μ calc	min. μ	max. μ
20	50.166	44.246	56.086
25	46.714	40.794	52.634
30	43.261	37.341	48.181
35	38.809	33.889	45.729
40	36.356	30.436	42.276
45	32.904	26.984	38.824
50	28.451	23.531	35.371

The minimum and maximum values of μ at each temperature are plotted so that we get the two straight lines shown in Figure (8.7). Since the upper and lower limits of the error bars lie within the two dotted lines, then there all experimental points are statistically significant.

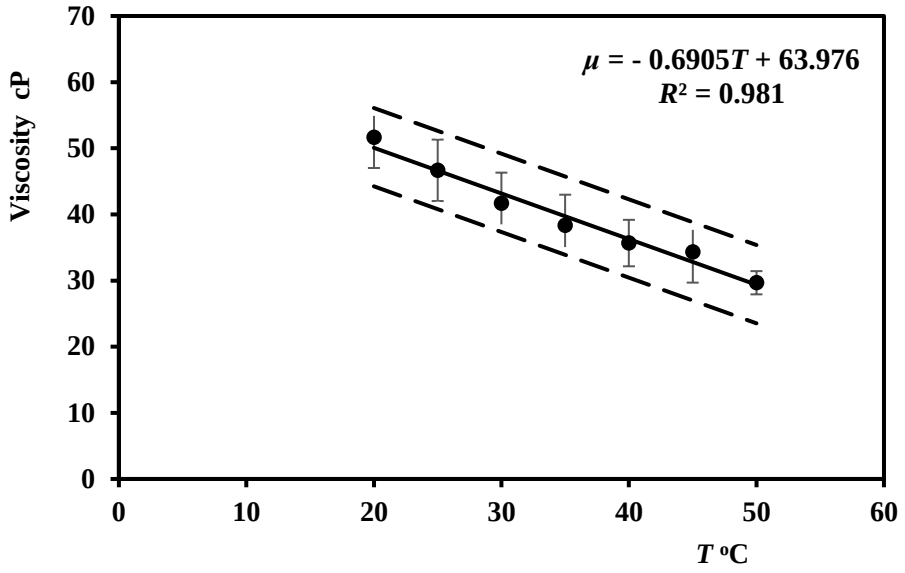

Fig (8.7): Upper and lower confidence limits of experimental points

8.3 Non – linear regression

Regressions obtained through experimentation are not necessary linear. In that case, the methods explained in the previous section are not valid.

- In general, it is always possible to assume a polynomial relation between the variables. This takes the form:

$$y_c = a_0 + a_1.x + a_2.x^2 + \ldots + a_n.x^n = \Sigma\, a_k.x_i^k \qquad (8.17)$$

The values of the coefficients a_0 , a_1 , a_2, ..., a_n are obtained by setting n conditions in the form:
For k = 1 to n:

$\frac{\partial \Sigma D_i^2}{\partial a_k} = 0$ Where $D_i = y_i - \Sigma\, a_k.x_i^k$

This yields a set of n linear equations that can be written in the following matrix form:

$$\boldsymbol{Z} = \boldsymbol{N}.\boldsymbol{A} \qquad (8.18)$$

For example, if the suggested regression equation is a second degree polynomial in the form: $y_c = a_0 + a_1.x_i + a_2.x_i^2$, the form of the matrix N is: $\boldsymbol{N} = \begin{pmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{pmatrix}$

Where $\boldsymbol{A}$ is the column matrix $[a_k] = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$ and $\boldsymbol{Z}$ is the column vector $\begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma x_i^2 y_i \end{bmatrix}$

Hence the coefficients can be obtained from: $\boldsymbol{A} = \boldsymbol{N}^{-1}\boldsymbol{Z}$

The value of the determination coefficient is calculated from the basic definition (8.14)

- If the suggested regression is of **the exponential type** in the form:
  $y_c = a.e^{kxi}$

This can be transformed to a linear form by taking logarithms of both sides:

$\ln y_c = \ln a + k.x_i$

A linear regression will thus be performed between $ln\ y_c$ and x_i.

- If the suggested regression is of **the power type** in the form;
  $y_c = a.x_i^n$

This can be transformed to a linear form by taking logarithms of both sides to get:

$\ln y_c = \ln a + n.\ln x_i$

A linear regression will thus be performed between $\ln y_c$ and $\ln x_i$.

Currently, all common types of non – linear regressions are available through the EXCEL program using the curve fitting module (Insert chart). This gives the best fit between experimental data for any assumed regression, as well as the coefficient of determination.

Example 8.4

A distribution that is often used to relate the mass fraction of suspended solid particles in water to their size is the Rosin – Rammler distribution:

$$\varphi = e^{-\left(\frac{D}{D_m}\right)^n}$$

Where: φ is the fraction having particle size > D, n an empirical constant and D_m a characteristic particle diameter.

The following table illustrates experimental data obtained in this respect using a sedigraph analyzer.

$D\ \mu m$	80	60	50	30	15	10	8	6	5	4	3	2	1
φ	0.005	0.017	0.047	0.277	0.627	0.767	0.887	0.917	0.941	0.963	0.98	0.995	0.999

From a suitable plot, determine the values of the parameters D_m and n.

Solution:

First, the relation should be linearized. Taking logarithms of both sides:

$$\ln \varphi = -\left(\frac{D}{D_m}\right)^n$$

$$\ln(-\ln \varphi) = n \ln D - n \ln D_m$$

Therefore, a plot of $\ln(-\ln \varphi)$ against $\ln D$ should produce a straight line of slope n and intercept $n.\ln D_m$

The table of calculations is shown below together with the corresponding plot.

D μm	80	60	50	30	15	10	8	6	5	4	3	2	1
φ	0.002	0.018	0.048	0.248	0.638	0.798	0.858	0.908	0.932	0.952	0.971	0.987	0.999
ln *D*	4.382	4.094	3.912	3.401	2.708	2.303	2.079	1.792	1.609	1.386	1.099	0.693	0.000
ln ln -*φ*	1.667	1.405	1.118	0.250	-0.762	-1.327	-2.121	-2.446	-2.800	-3.278	-3.902	-5.296	1.667

The regression equation is:

$\ln(-\ln \varphi) = 1.794 \ln D - 5.862$ $(R^2 = 0.984)$

Hence $\boldsymbol{n = 1.794}$ and $n.\ln D_m = 5.862$, which finally yields $\boldsymbol{D_m = 26.25\ \mu m}$

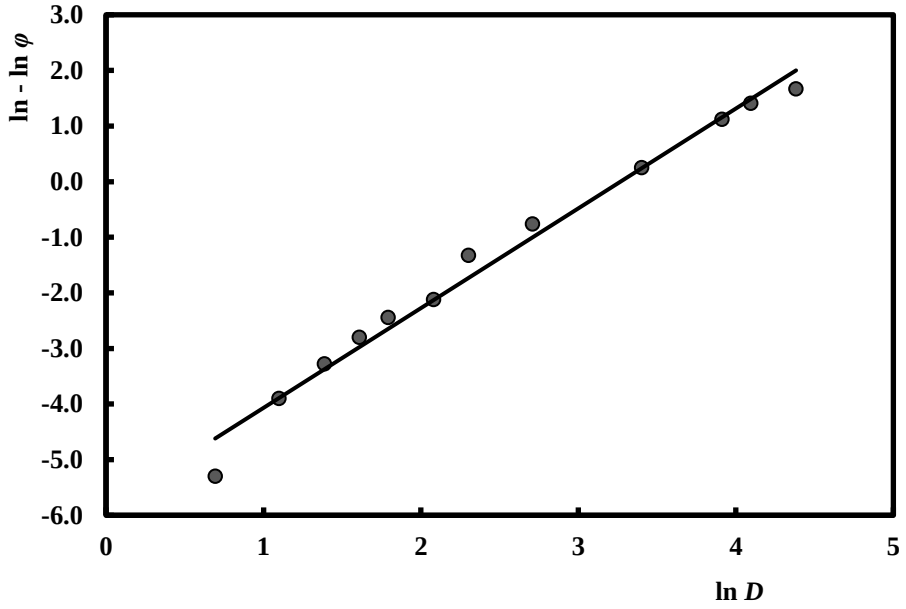

Fig (8.8): Linear relation for data of example (8.4)

Example 8.5

An experiment was conducted to follow up the effect of temperature on the yield strength of a polymer. The results are shown in the following table.

***T* ºC**	25	30	40	50	60	70
Strength MPa	8.82	8.25	7.32	6.2	5.26	0.69

Fit a second order polynomial to represent this correlation and calculate the coefficient of determination.

Solution:

A second order polynomial can be fitted using EXCEL module: Insert chart: 2nd degree polynomial. The plot is shown in Fig (8.9) together with the determination coefficient R^2.

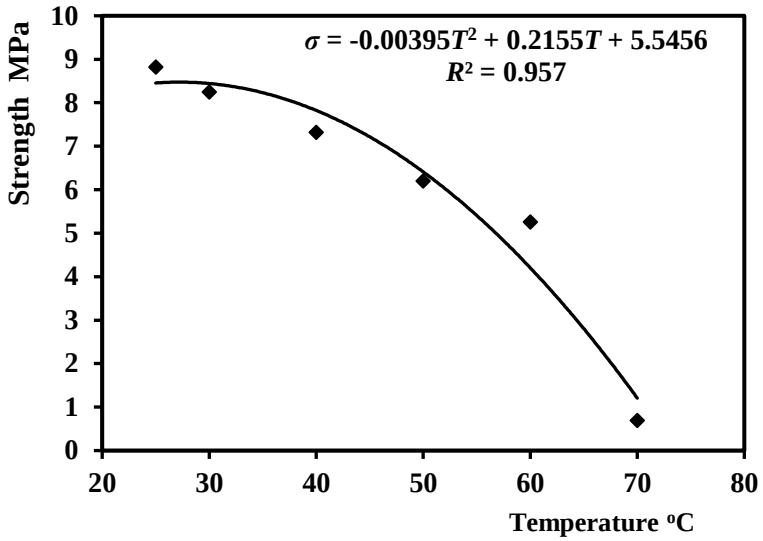

Fig (8.9): Second order polynomial fitting of data of example (8.6)

Example 8.6

The following data were obtained relating the mean bending strength of hardened gypsum board samples to the particle size of gypsum, under a restriction that the maximum particle size that can be used is 1 mm (1000 μm)

Size *D* μm	74	104	147	208	294	416	590
Strength *σ* MPa	7.2	6.4	5.6	4.9	4.5	4.3	4.1

Optimize the relation between strength and particle size by choosing the most suitable regression equation.

Solution:

The relation is plotted and the following trials are performed:

Power function

The regression equation is $\sigma = 22.95\ D^{-0.28}$ with $R^2 = 0.961$ (Fig 8.10)

Polynomial

A third-degree polynomial has been fitted in fig 8.11 and although it yields a value of $R^2 = 0.998$, it cannot be accepted as an extrapolation of the trend indicates that the strength can reach negative values if D > 800 μm, which is illogic.

A direct exponential function fitting of data does not yield a reasonable value of R^2, as exponential functions tend to approach 0 as D approaches ∞ ($R^2 = 0.787$). However, a look at the table reveals that an asymptotic value of about 4 MPa may be assumed as D approaches 1000.

That is why, a plot of $\sigma - 4$ was performed against D as shown in Fig (8.12).

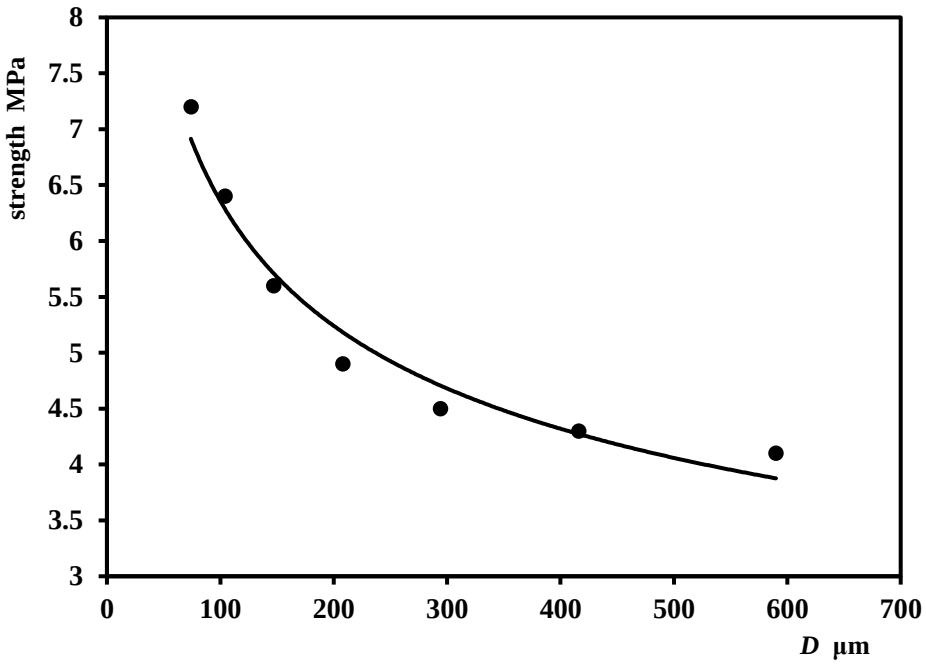

Fig (8.10): Fitted power function for data of Example (8.7)

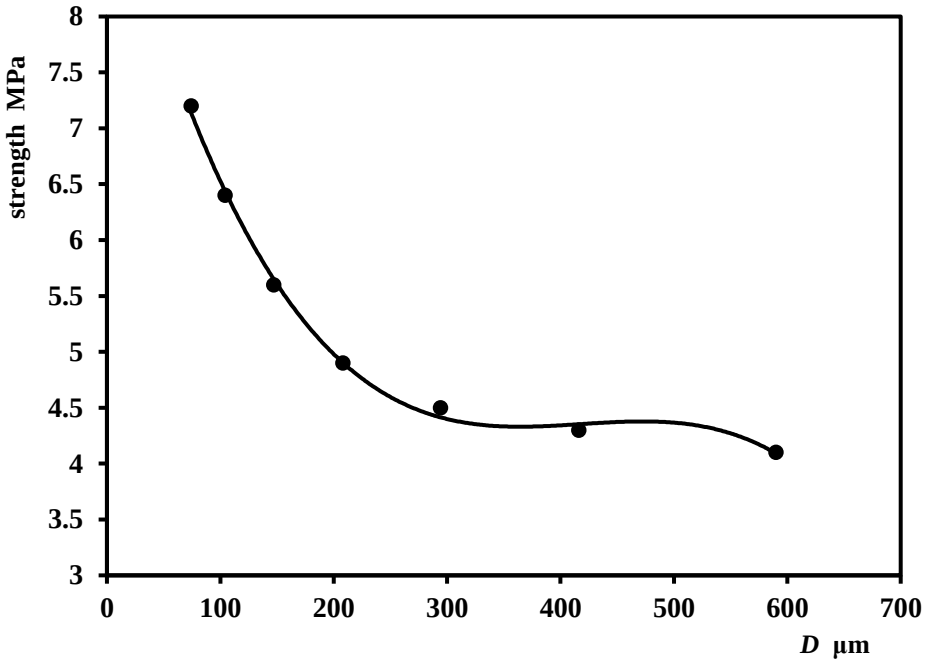

Fig (8.11): Fitted 3rd degree polynomial for data of Example (8.7)

A direct exponential function fitting of data does not yield a reasonable value of R^2, as exponential functions tend to approach 0 as D approaches ∞ ($R^2 = 0.787$). However, a look at the table reveals that an asymptotic value of about 4 MPa may be assumed as D approaches 1000.

That is why, a plot of $\sigma - 4$ was performed against D as shown in Fig (8.12).

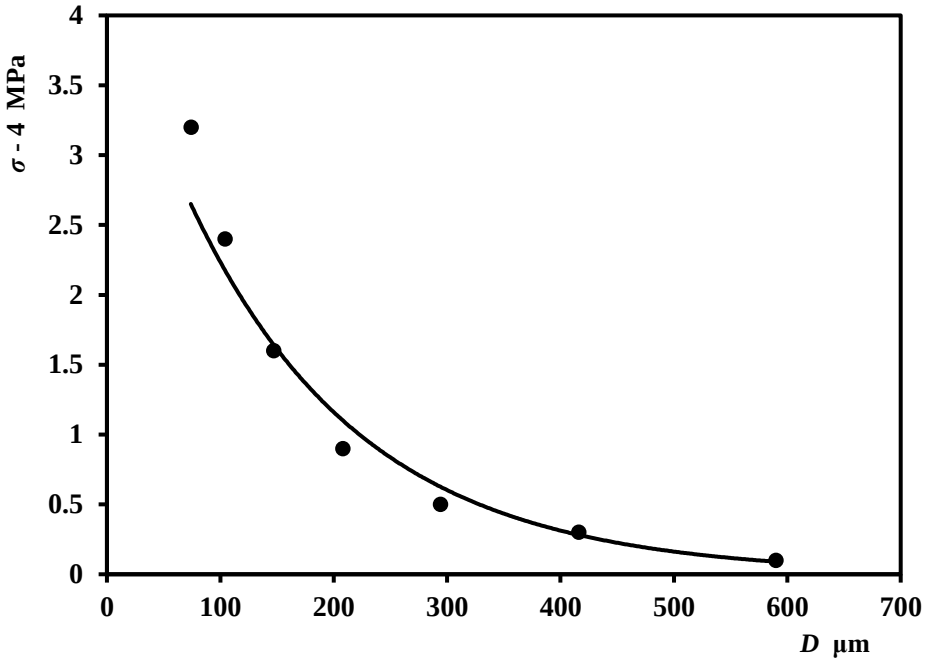

Fig (8.12): Fitted modified exponential function for data of Example (8.7)

The deduced function takes the form: $\sigma = 4 + 4.305e^{-0.0066D}$ with $R^2 = 0.983$

8.4 Exercise problems

(1) During a test with a thermocouple the e.m.f. (mV) was related to temperature through the following table:

T **ºC**	100	200	300	400	500	600	700	800	900	1000
E **mV**	5	5.5	10.5	13.6	18	20.2	26.5	27.3	28.6	35

Obtain a linear regression for temperature as function of e.m.f., and determine the correlation coefficient. Then construct a confidence interval for the population coefficient at significance level = 0.05

(2) The following table shows the results obtained in a poll covering 24 randomly chosen graduate students to relate their scores in the midterm exam in a certain subject A and another subject B. The results were as follows:

A	16	12	11	13	8	14	19	4	5	14	13	20
B	14	10	12	12	10	14	20	7	4	12	13	20
A	12	13	17	11	9	15	12	11	8	20	18	14
B	12	12	15	12	11	14	10	9	5	17	17	12

Prepare a scatter diagram and deduce the regression equation and the correlation coefficient.

(3) The following table shows experimental values for thermal conductivity of insulating boards *K* at different values of porosity. Express *K* as a linear function of porosity and deduce the correlation coefficient. At a confidence level of 0.95, construct a confidence interval for the population determination coefficient using the Fisher method.

Porosity	0.35	0.43	0.28	0.49	0.36	0.50	0.38	0.44	0.52	0.33
***K* W/m.K**	0.89	0.77	0.98	0.78	0.85	0.65	0.85	0.80	0.60	0.81

(4) Compressive strength tests are performed on samples of concrete mortar cubes after 28 days curing as function of cement content per cubic meter concrete. Each sample consists of 3 specimens. The results are as follows:

Cement content	**Specimen 1**	**Specimen 2**	**Specimen 3**
230	28	27	30.2
260	32.3	30.9	33.4
320	36	34.8	33.9
345	38.7	40	38.2
390	41.3	42.5	42.7
420	43.5	42.8	43

Obtain a linear regression equation relating the mean sample strength to the cement content. For a confidence level of 0.95, draw error bars as well as two lines representing the lower and upper boundaries of expected strength values.

(5) The following equation has been suggested to relate specific heat of a solid (J/mol.K) to temperature (K): $c_p = a + b.T + c.T^2$.
Using a proper regression, find the values of the constants *a, b* and *c*. Obtain the determination coefficient.

***T* K**	300	400	500	600	700	800	900
c_p	25	27.5	27.7	28.5	28.8	30.1	30.4

(6) The rate constant of a chemical reaction is known to be related to temperature (K) by the relation: $k = Ae^{-\frac{E}{RT}}$
From the following data relating *k* to temperature, estimate the values of *A* and *E* (J/mole) and estimate the coefficient of determination.

***T* K**	300	340	385	420	455	500
k	0.016	0.02	0.022	0.023	0.024	0.026

(7) The compressibility factor *Z* of a real gas has been related to its molar volume by the relation: $Z = a + b/V + c/V^2$. Using a suitable regression find the values of the constants *a, b* and *c* and estimate the coefficient of determination.

V m^3/mole	0.02	0.03	0.04	0.05	0.06	0.07
Z	1.23	1.15	1.14	1.09	1.07	1.05

(8) The following data were obtained on following the sedimentation of fine silt in water. The height represents the level of interface between clear liquid and suspension.

Time min	0	15	30	45	60	90	120	150	360
Height mm	430	405	385	380	355	345	335	325	290

Find an equation describing the sedimentation operation in the form:
$h = 280 + ke^{-c.t}$ given the constraint that at $t = 0$, $h = 430$.
Write down the coefficient of determination.

(9) An agricultural waste is used for the adsorption of heavy metal ions from wastewater. The equilibrium concentration of ions (q_e mg.L^{-1}) follows the Langmuir model:

$$q_e = \frac{k.c}{1+bc}$$

Where, c is the concentration of the adsorbed phase (mg.L^{-1}).

Prove that the following data are compatible with the above expression using a linear plot:

c	0.0045	0.0087	0.021	0.026	0.092	0.195
q_e	0.026	0.053	0.075	0.082	0.123	0.129

Obtain a correlation coefficient for that relation.

(10) The friction factor in an experiment involving flow of heavy oil in a duct was correlated to the Reynolds number through a relation in the form:

$$f = k.Re^n$$

Obtain the values of the constants k and n by linearization of the above expression and find the correlation coefficient. Draw the error bars at significance level = 0.05.

Re	0.1	1	10	50	100	200	500	1000
# 1	1.37	0.266	0.039	0.0095	0.0068	0.0031	0.0019	0.0011
# 2	1.51	0.241	0.041	0.012	0.0054	0.0035	0.002	0.002
# 3	1.29	0.219	0.029	0.011	0.0076	0.0029	0.0027	0.001

(11) The relation between the mole fraction of a volatile component in vapor phase (y) and its mole fraction in liquid phase (x) is often obtained by the following expression where α is the relative volatility:

$$y = \frac{\alpha.x}{1+(\alpha-1).x}$$

Express the following data in linearized form, then find the value of α from two different parameters obtained from the regression equation.