EDEXCEL ANALYTICAL METHODS FOR ENGINEERS H1

UNIT 2 - NQF LEVEL 4

OUTCOME 4 - STATISTICS AND PROBABILITY

TUTORIAL 3 – LINEAR REGRESSION

Tabular and graphical form**:** data collection methods; histograms; bar charts; line diagrams; cumulative frequency diagrams; scatter plots

Central tendency and dispersion**:** the concept of central tendency and variance measurement; mean; median; mode; standard deviation; variance and interquartile range; application to engineering production

Regression, linear correlation**:** determine linear correlation coefficients and regression lines and apply linear regression and product moment correlation to a variety of engineering situations

Probability**:** interpretation of probability; probabilistic models; empirical variability; events and sets; mutually exclusive events; independent events; conditional probability; sample space and probability; addition law; product law; Bayes' theorem

Probability distributions**:** discrete and continuous distributions, introduction to the binomial, Poisson and normal distributions; use of the Normal distribution to estimate confidence intervals and use of these confidence intervals to estimate the reliability and quality of appropriate engineering components and systems

INTRODUCTION

This section is mainly concerned with the collection of data that is thought to have a directly proportional relationship. Suppose there are two measured variables x and y and it is thought that they are related by a formula y = mx + C (the straight line law). Often it is fairly apparent from the x, y plot if such a relationship exists but often there is enough scatter in the points to make it doubtful and difficult to decide what the best values of m and C (slope and intercept) should be.

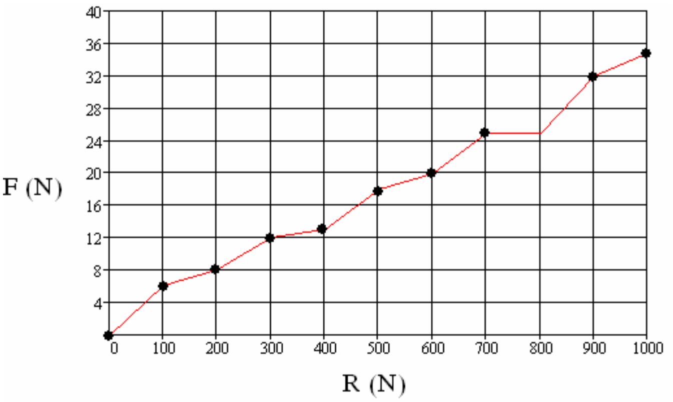
FINDING THE BEST STRAIGHT LINE

Much statistical work has been done to try and find a way of fitting a mathematical model (equation) to obtain the best possible representation of data found from experiments or sampling.

Let's use as an example, the results of an experiment to find the coefficient of friction between materials. The law of dry friction (Coulomb's Laws) applied to sliding objects states that F = μ R where F is the friction force, μ is a constant called the coefficient of friction and R is the force squeezing the two surfaces together. If we conduct an experiment we might have small errors in the instruments for measuring the weights and forces and we might have variations in the surface texture but overall we expect the results to follow the law.

We might get a graph like this.

Just joining the points is not satisfactory. We can see that it looks as though it should be a straight line graph so the question is how to draw the best straight line that places the data as close to the line as possible.

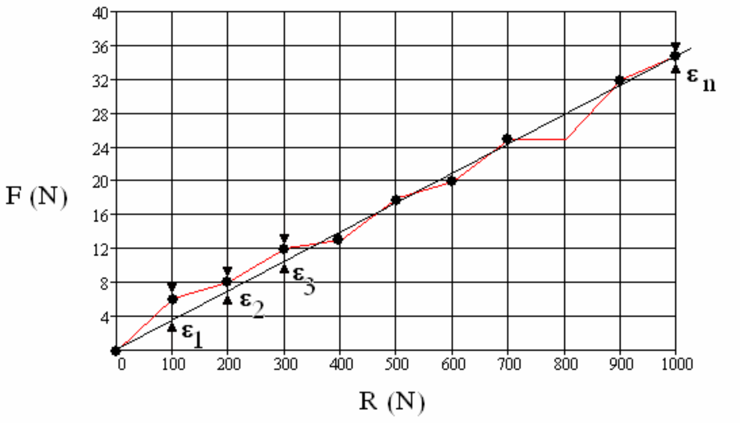


For linear relations, the most widely used method is the method of least squares also called the linear regression method.

For any given point (R) the difference between the raw data (F) and the exact point predicted by the law is $\varepsilon = F - \mu R$

Research has shown that the best fit occurs when the values of ε^2 is a minimum. This is the method of least squares. The reason for using ε^2 and not ε is based around the fact that a negative value squared is positive and produces a better result. For n points the total error is $(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 .... + \varepsilon_n^2)$ and the object is to make $\sum\varepsilon^2$ a minimum.

The problem now is to find the value of the constant μ (the gradient) that produces the minimum value of ε^2 and this is a big problem. It would be very laborious to guess at values and then work out all the values of ε^2 and repeat until we have a value of μ that gives the minimum. That is basically what regression is.



Fortunately these days we have computer programs that will do the work for us and produce the best fit. You will find one such aid at this web address
http://www.shodor.org/unchem/advanced/lls/leastsq.html

The following is the way to use Microsoft Excel to do it.

The following set of instructions enables you to create a best fit and display the best equation for any data plot using Excel.

1. Create your raw data plot in two columns (ideally with the y values in the first column).

2. Highlight all the cells by clicking and dragging. If the data you want to highlight is not in two adjacent columns highlight them separately by holding down the Ctrl key as you drag first one then the other.

3. Click on **Insert** on the menu bar.

4. Click on **Chart....**

5. Under **Standard Types**, **Chart type:** click on **XY (Scatter)**.

6. Under **Chart sub-type:** click on the chart with only data markers and no lines.

7. Click on **Next>**.

8. Click on **Next>**.

9. Under **Titles**, Click in the text box under **Chart title:** and enter a title for the graph.

10. Click in the text box under **Category (X) axis:** and enter a title for the x-axis. Click in the text box under **Value (Y) axis:** and enter a title for the y-axis.

11. Click on the **Gridlines** tab. Click in the checkboxes to turn gridlines on or off.

12. Click on the **Legend** tab. Click in the checkbox next to **Show legend** to turn the legend on or off. **Placement** of the legend on the graph can also be selected here.

13. Click on the **Data Labels** tab. Click on the radio buttons to turn data labels on or off.

14. Click on the **Data Table** tab. Click in the checkboxes to turn a data table on or off.

15. Click on **Next**.

16. Under **Place chart:** click on the button next to **As new sheet.** Enter a name for the graph.
17. Click on **Finish**.

Having created a data plot the following instructions are to create a best fit graph and print the function.

1. Move the mouse cursor to *any data point* and left click. All of the data points should now be highlighted. Right click and click on **Add Trendline**.

2. From within the **Add Trendline** window, under **Type**, click on the box with the type of fit you want (e.g., **Linear**).

3. Click on **Options** at the top of the **Add Trendline** window.

4. Click in the checkbox next to **Display equation on chart** and the checkbox next to **Display R-squared value on chart**.

5. Click on **OK**.

WORKED EXAMPLE No. 1

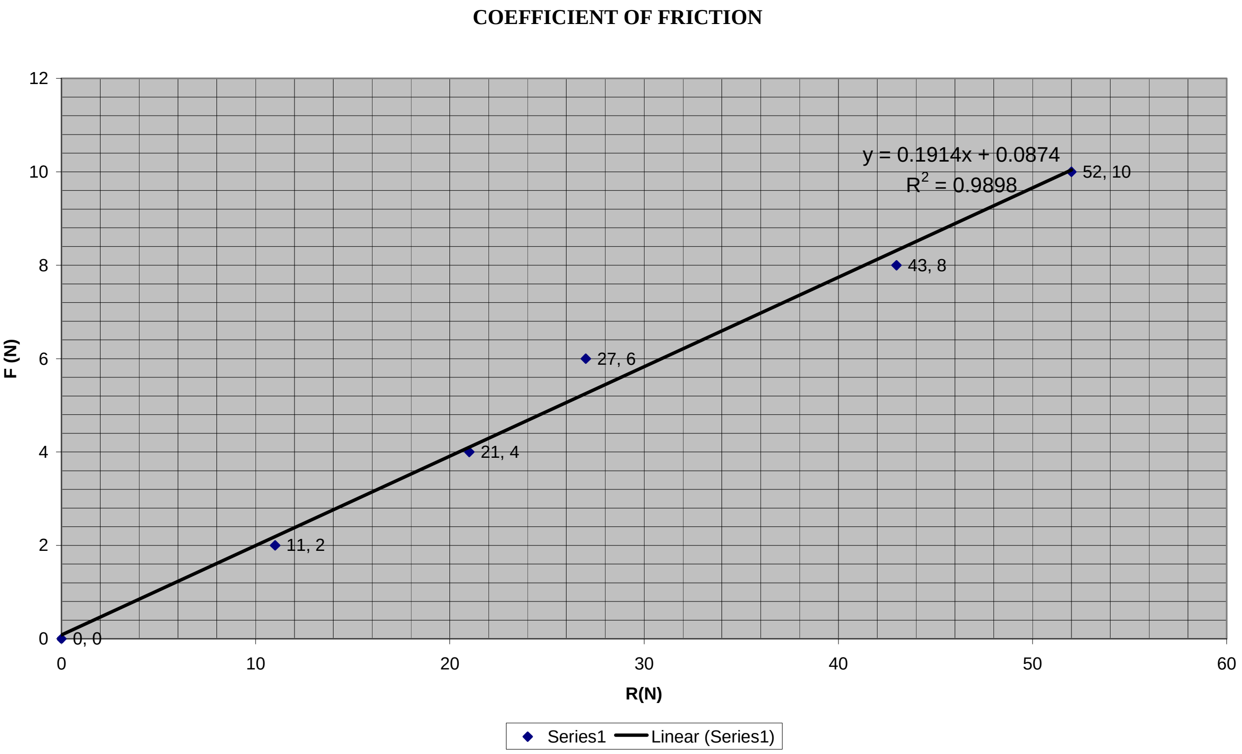
In an experiment to find the coefficient of friction, the force F and normal load R was measured and the following results obtained.

F	0	2	4	6	8	10
R	0	11	21	27	43	52

Using excel, produce the best linear graph and the function produced. What is the coefficient of friction?

SOLUTION

Plotting according to the above instructions we get the following.



The function is F = 0.1914R + 0.0874

The best plot does not pass through zero but this would have been an option in the last instruction. The coefficient of friction is the gradient 0.1914

UNDER LAYING THEORY OF LINEAR REGRESSION EQUATIONS

If you don't want to understand the theory, go straight to the next page.

Normally a straight line function is of the form $y = mx + C$ where C is the intercept and m the gradient. If we were trying to fit the data to this law, we would have two constants m and C to determine so the work becomes even more complex.

$\varepsilon = y- (mx + C)$ where y is the raw data.

$$\varepsilon^2 = (y - mx - C)^2 = y^2 - 2\ mxy - 2Cy - m^2x^2 + 2Cmx + C^2$$

$$\Sigma(\varepsilon)^2 = \Sigma(y^2 - 2\ mxy - 2Cy - m^2x^2 + 2Cmx + C^2)$$

$$\Sigma(\varepsilon)^2 = \Sigma\ y^2\ - 2m\Sigma xy - 2C\Sigma y - m^2\Sigma x^2 + 2Cm\ \Sigma x + \Sigma C^2$$

$$\Sigma(\varepsilon)^2 = \Sigma\ y^2 - 2m\Sigma xy - 2C\Sigma y - m^2\Sigma x^2 + 2Cm\ \Sigma x + NC^2$$

N is the number of data points.

The problem now is to find the minimum value.

$$\Pi = \Sigma\ y^2 - 2m\Sigma xy - 2C\Sigma y - m^2\Sigma x^2 + 2Cm\ \Sigma x + NC^2$$

For the value of m that makes Π a minimum $\frac{d\Pi}{dm} = 0 = -2\Sigma xy - 2m\Sigma x^2 + 2C\ \Sigma x \quad = 0$

$$m = \frac{\sum xy - C\sum x}{\sum x^2} \ldots\ldots\ldots\ldots\ldots(1)$$

For the value of C that makes Π a minimum $\frac{d\Pi}{dC} = 0\ = -\ 2\Sigma y + 2m\ \Sigma x + 2NC$

$$C = \frac{\sum y - m\sum x}{N} \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Substitute (2) into (1)

$$m = \frac{\sum xy - \frac{\sum y - m\sum x}{N}\sum x}{\sum x^2} = \frac{\sum xy}{\sum x^2} - \frac{\sum x\sum y}{N\sum x^2} + \frac{m\{\sum x\}^2}{N\sum x^2}$$

$$m - \frac{m\{\sum x\}^2}{N\sum x^2} = \frac{\sum xy}{\sum x^2} - \frac{\sum x\sum y}{N\sum x^2} \qquad m\left\{1 - \frac{\{\sum x\}^2}{N\sum x^2}\right\} = \frac{\sum xy}{\sum x^2} - \frac{\sum x\sum y}{N\sum x^2}$$

$$m = \frac{\frac{\sum xy}{\sum x^2} - \frac{\sum x\sum y}{N\sum x^2}}{\left\{1 - \frac{\{\sum x\}^2}{N\sum x^2}\right\}} = \frac{N\sum xy - \sum x\sum y}{\sum x^2\left\{N - \frac{\{\sum x\}^2}{\sum x^2}\right\}} = \frac{N\sum xy - \sum x\sum y}{N\sum x^2 - \{\sum x\}^2}$$

With further work (2) gives $C = \frac{\sum y\sum x^2 - \sum x\sum xy}{N\sum x^2 - (\sum x)^2}$ but is easier to use (2) once m is found.

The method of linear regression for the best straight line gives:

Gradient $m = \dfrac{N\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2}$ Intercept by $C = \dfrac{\sum y \sum x^2 - \sum x \sum xy}{N\sum x^2 - \left(\sum x\right)^2}$

Here is a little exercise you don't find in text books that explains the next simplification.

WORKED EXAMPLE No. 2

Prove that $N\sum(x - \bar{x})(y - \bar{y}) = N\sum xy - \sum x \sum y$

SOLUTION

$$N\sum(x - \bar{x})(y - \bar{y}) = N\sum\left[xy - x\bar{y} - \bar{x}y + \overline{xy}\right]$$

$$N\sum(x - \bar{x})(y - \bar{y}) = N\left[\sum xy - \sum x\bar{y} - \sum \bar{x}y + \sum \overline{xy}\right]$$

$$N\sum(x - \bar{x})(y - \bar{y}) = N\left[\sum xy - \bar{y}\sum x - \bar{x}\sum y + N\overline{xy}\right]$$

The mean value of the x and y samples are respectively $\bar{x} = \dfrac{\sum x}{N}$ and $\bar{y} = \dfrac{\sum y}{N}$ Substitute these

$$N\sum(x - \bar{x})(y - \bar{y}) = N\left[\sum xy - \frac{\sum y \sum x}{N} - \frac{\sum x \sum y}{N} + \frac{N\sum x \sum y}{N^2}\right]$$

$$N\sum(x - \bar{x})(y - \bar{y}) = N\sum xy - \sum y \sum x - \sum x \sum y + \sum x \sum y$$

$$N\sum(x - \bar{x})(y - \bar{y}) = N\sum xy - \sum x \sum y$$

WORKED EXAMPLE No. 3

Prove that $N\sum(x - \bar{x})^2 = N\sum x^2 - \left(\sum x\right)^2$

SOLUTION

$$N\sum(x - \bar{x})^2 = N\sum\left[x^2 + \bar{x}^2 - 2x\bar{x}\right]$$

$$N\sum(x - \bar{x})^2 = N\sum x^2 + N\sum \bar{x}^2 - N\sum 2x\bar{x}$$

$$N\sum(x - \bar{x})^2 = N\sum x^2 + N^2\bar{x}^2 - 2N\bar{x}\sum x$$

Substitute $\bar{x} = \dfrac{\sum x}{N}$

$$N\sum(x - \bar{x})^2 = N\sum x^2 + \left(\sum x\right)^2 - 2\left(\sum x\right)^2$$

$$N\sum(x - \bar{x})^2 = N\sum x^2 - \left(\sum x\right)^2$$

Using the above simplification we now have $m = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$

WORKED EXAMPLE No. 4

Check out the answers from worked example No.1 using the formulae to calculate m and C.

Prove that $m = \dfrac{N\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2}$ gives the same result as $m = \dfrac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$

F	0	2	4	6	8	10
R	0	11	21	27	43	52

Using excel, produce the best linear graph and the function produced. What is the coefficient of friction?

SOLUTION

n	y	x	xy	x^2	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	0	0	0	0	-25.667	-5	128.33	658.77
2	2	11	22	121	-14.667	-3	44	215.11
3	4	21	84	441	-4.667	-1	4.667	21.778
4	6	27	162	729	1.333	1	1.333	1.778
5	8	43	344	1849	17.333	3	52	300.444
6	10	52	520	2704	26.333	5	131.667	693.444
Total	30	154	1132	5844	0.0	0.0	362	1891.333
Mean	5	25.67						

$$m = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2} = \frac{6(1132) - (154)(30)}{6(5844) - 154^2} = \frac{2172}{11348} = 0.1914$$

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{362}{1891.33} = 0.1914$$ Both formulae give the same answer.

$$C = \frac{\sum y - m\sum x}{N} = \frac{30 - (0.1914)(154)}{6} = 0.087$$

Or

$$C = \frac{\sum y \sum x^2 - \sum x \sum xy}{N\sum x^2 - \left(\sum x\right)^2} = \frac{(30)(5844) - (154)(1132)}{(6)(5844) - 154^2} = \frac{992}{11348} = 0.0874$$

SELF ASSESSMENT EXERCISE No. 1

1. In an experiment to verify Ohm's Law, the voltage and current acting on a resistor was measured and the following results obtained.

V	0	2	4	6	8	10
I	0	0.19	0.42	0.62	0.75	0.92

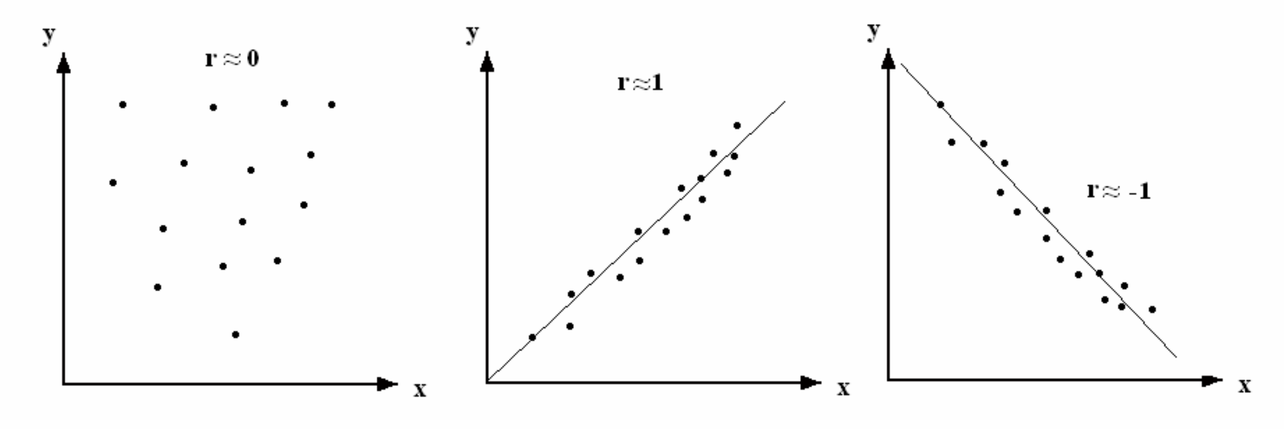
Assuming the law should be linear, find the best fit straight line law and check the answers with using Excel. (Calculation gives m =10.73, C = 0.184 Excel gives same answers)

LINEAR CORRELATION COEFFICIENT

The linear correlation coefficient is a way to measure how good a linear relationship would be without gouing through all the preceding work. Let's suppose we are examining a scatter plot and we are trying to work out the best straight line to use. The linear correlation coefficient (r or R) is a number between -1 and 1 which measures how close to a straight line a set of points falls. A zero value means the widest scatter with no relationship. A value of 1 is a perfect correlation with a positive slope. A value of -1 is a perfect correlation with a negative slope. Other values in between represent various degrees of scatter.

Note that the R^2 value can be produced by excel using the option tab and this is printed on the plot in worked example 1.

The diagrams illustrate correlation.



Without proof the formula for r is $r = \dfrac{N\sum xy - \sum x \sum y}{N\sqrt{\left(\sum x^2 - \left(\sum x\right)^2\right)\left(\sum y^2 - \left(\sum y\right)^2\right)}}$

This is more easily understood as $r = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ where σ_x is the standard deviation of the x values, σ_y is the standard deviation of the y values. σ_{xy} is the product moment.

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{N}} \qquad \sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{N}} \qquad \sigma_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{N}$$

This works best when x and y are both normally distributed. The following explains the process.

CALCULATING THE CORRELATION COEFFICIENT:

Suppose we have N sets of data, (x_1,y_1), (x_2,y_2) .... (x_n,y_n)

STEP 1. Calculate the average of the x and y vales $\bar{x}$ and $\bar{y}$

$$\bar{x} = (x_1 + x_2 + x_3 ...+ x_N)/N$$

$$\bar{y} = (y_1 + y_2 + y_3 ...+ y_N)/N$$

STEP 2. Calculate the standard deviations for x and y (σ_x and σ_y).

$$\sigma_x = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + ...(x_n-\bar{x})^2}{N}}$$

$$\sigma_y = \sqrt{\frac{(y_1-\bar{y})^2 + (y_2-\bar{y})^2 + (y_3-\bar{y})^2 + ...(y_n-\bar{y})^2}{N}}$$

STEP 3 Calculate the covariance between the two data sets:

$$\sigma_{xy} = \frac{\left[(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y}) + ...(x_n - \bar{x})(y_n - \bar{y})\right]}{N}$$

STEP 4 The correlation coefficient is then defined as $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Note some text give N-1 for the calculations of standard deviations but since all three terms contain the same it makes no difference to the answer.

WORKED EXAMPLE No. 5

Calculate the correlation coefficient for the following x and y data sets.

n	1	2	3	4	5
x	1	2	3	4	5
y	4	4	3	2	1

SOLUTION

Calculate $\bar{x}$ and $\bar{y}$

$\bar{x} = (1 + 2 + 3 + 4 + 5)/5 = 3.0$

$\bar{y} = (4 + 4 + 3 + 2 + 1)/5 = 2.8$

Calculate σ_x and σ_y,

$$\sigma_x = \sqrt{\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}} = 1.41$$

$$\sigma_y = \sqrt{\frac{(4-2.8)^2 + (4-2.8)^2 + (3-2.8)^2 + (2-2.8)^2 + (1-2.8)}{5}} = 1.17$$

Calculate the covariance σ_{xy}

$$\sigma_{xy} = \frac{\left[(1-3)(4-2.8) + (2-3)(4-2.8) + (3-3)(3-2.8) + (4-3)(2-2.8) + (5-3)(1-2.8)\right]}{5}$$

$$\sigma_{xy} = \frac{\left[(-2)(1.2) + (-1)(1.2) + (0)(0.2) + (1)(-0.8) + (2)(-1.8)\right]}{5}$$

$$\sigma_{xy} = \frac{\left[-2.4 - 1.2 + 0 - 0.8 - 3.6\right]}{5} = -1.6$$

Calculate r, the correlation coefficient.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-1.6}{(1.41)(1.17)} = -0.97$$

SELF ASSESSMENT EXERCISE No. 2

1. Calculate the correlation coefficient for the following x and y data sets and establish the best straight line law.

n	1	2	3	4	5
x	0	5	10	20	30
y	2	4.5	6.5	11.5	17.5

(Answer $r = -0.998$ $y = 0.511x + 1.75$)

2. Calculate the correlation coefficient for the following x and y data sets and establish the best straight line law.

n	1	2	3	4	5	6
x	0	5	10	15	20	25
y	8	22	40	53	66	90

(Answer $r = 0.996$ $y = 3.171x - 9$)

2. Calculate the correlation coefficient for the following x and y data sets and establish the best straight line law.

n	1	2	3	4	5	6
x	0	2	4	6	8	10
y	16	10	5	8	5	0

(Answer $r = -0.906$ $y = -1.314x + 13.9$)