Linear Regression Analysis: More Review and Examples

Simple linear regression describes the relationship between the dependent and independent variables by means of a linear equation. This is the simplest form of equation between two variables. This type of relationship can at least approximate many business situations.

Least Squares Method

If the scatter diagram shows that there is a linear relationship between the two variables, we can use the diagram to predict values of the dependent variable for selected values of the independent variable by drawing a straight line roughly through the points to find a linear model of a business situation.

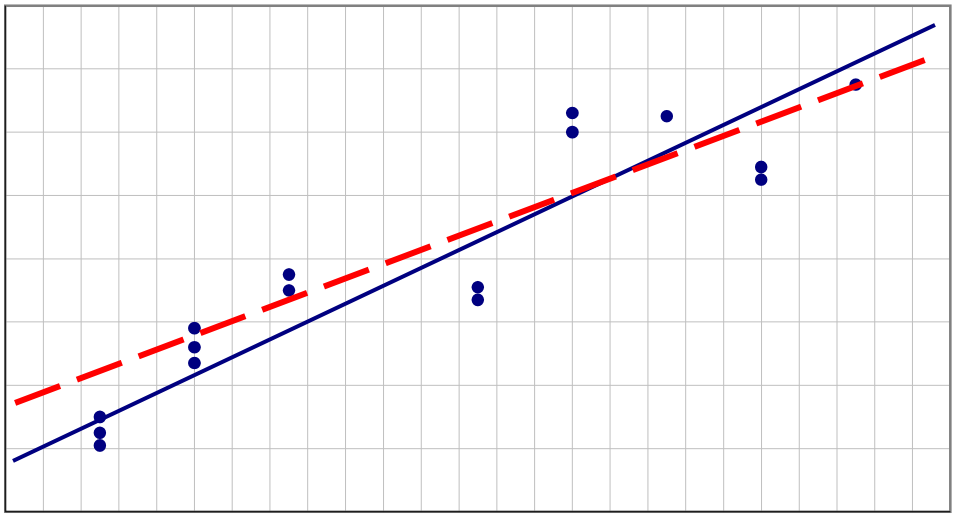
We know that the equation of a straight line is

$$y = ax + b$$

where a is the slope of the line and b is the y-intercept.

Using a ruler we could measure the y-intercept b and the slope a.

However, this method is not very accurate, and lines drawn to fit the set of points by different people might give different values for the intercept and the slope of the line.



The least squares method provides a mathematical process for determining the equation of a straight line that best fits the data. The equation is called the *regression equation*. The graphical representation of the equation is known as the *regression line*.

The basic principle of this method is that the sum of the squares of the vertical distances from the data points to a particular line is as small as possible. For this reason the regression line is also called the *line of best fit*.

The general form of the regression equation is given by

$$y = ax + b$$

where b is the value of y when $x = 0$ (the y-intercept)
a is the slope of the regression line and shows the change in the dependent variable y for a change of one unit in the independent variable x
x is the selected value of the independent variable
y is the predicted value of the dependent variable for a selected value of x

The values of regression coefficients a and b can be computed from the following formulas:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Follow these steps to use the least squares method:

Step 1 Determine the dependent variable.

Step 2 Construct a scatter diagram.

Step 3 If the scatter diagram shows a reasonable linear relationship between the two variables, use the two formulas to determine a and b. It is advised to set out the calculations in tabular form.

Step 4 Determine the regression equation.

Step 5 Plot the regression line on the scatter diagram.

Example 1

We want to analyze the relationship between production and manufacturing expenses. A sample of nine companies gives the following data.

Production, x (in thousands of units)	40	42	48	55	65	79	88	100	120
Manufacturing expenses, y (in thousands of $)	150	140	160	170	150	162	185	165	190

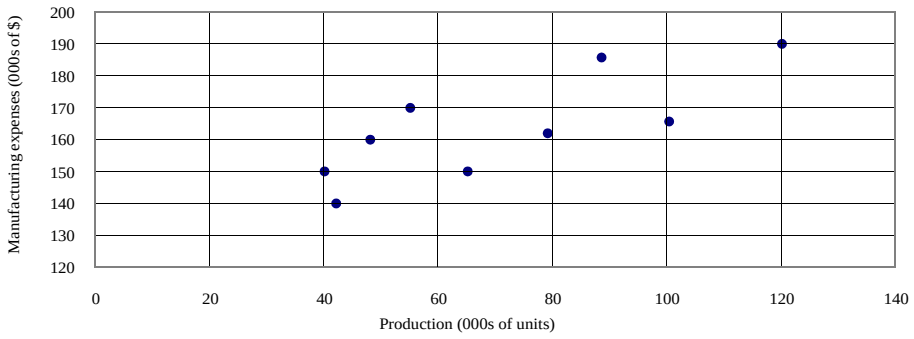
Step 1 "Manufacturing expenses" is the dependent variable, y, and "production" is the independent variable, x. Manufacturing expenses change as the volume of production changes. The value of y depends on the value of x.

Step 2 The following figure shows the scatter diagram for this data.

Step 3 **The scatter diagram shows a reasonable linear relationship between production and manufacturing expenses.**

Step 4 The regression equation is $y = ax + b \Rightarrow y = 0.4589x + 131.08$

Step 5 Plot the regression line on the scatter diagram.



$n = 9$ (There are 9 pairs of data values)

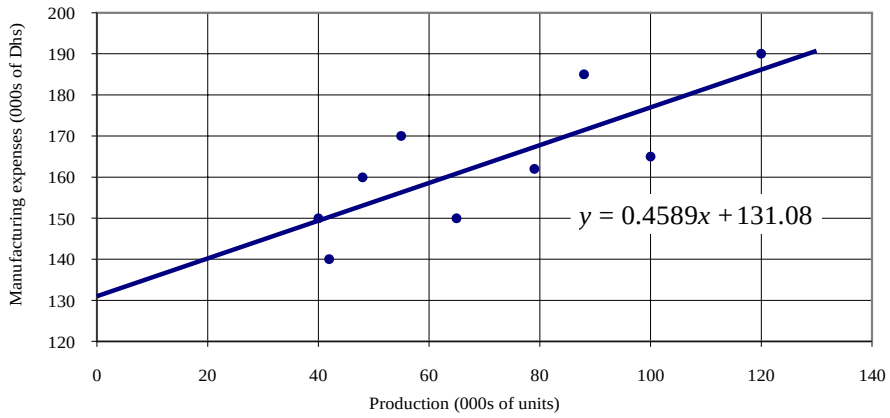
x	y	x^2	$x \cdot y$
40	150	1600	6000
42	140	1764	5880
48	160	2304	7680
55	170	3025	9350
65	150	4225	9750
79	162	6241	12798
88	185	7744	16280
100	165	10000	16500
120	190	14400	22800
$\sum x = 637$	$\sum y = 1472$	$\sum x^2 = 51\ 303$	$\sum xy = 107\ 038$

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{9(107\,038) - (637)(1472)}{9(51\,303) - (637)^2} = 0.4589$$

The slope of the regression line is 0.4589 (positive slope). The manufacturing expenses will increase by approximately 0.46 $s for an increase of one unit in production.

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(51\,303)(1472) - (107\,038)(637)}{9(51\,303) - (637)^2} = 131.08$$

The y-intercept is 131.08 (in thousands of dollars) that means the manufacturing expenses are $ 131 080 (131.08 × 1000) when there is no production (due to fixed costs).


Manufacturing expenses (000s of Dhs)
200
190
180
170
160
150
140
130
120
0
20
40
60
80
100
120
140
Production (000s of units)
y = 0.4589x + 131.08

Example 2

Adam Suits wants to know the relationship between the age and annual maintenance costs of sewing machines. A sample of 16 machines shows the following ages and maintenance costs during the past year.

Age (years), x	8	3	1	9	5	7	5	2	1	3	6	2	1	2	6	8
Maintenance costs, y ($)	109	75	21	135	67	125	71	52	25	70	126	58	30	47	120	105

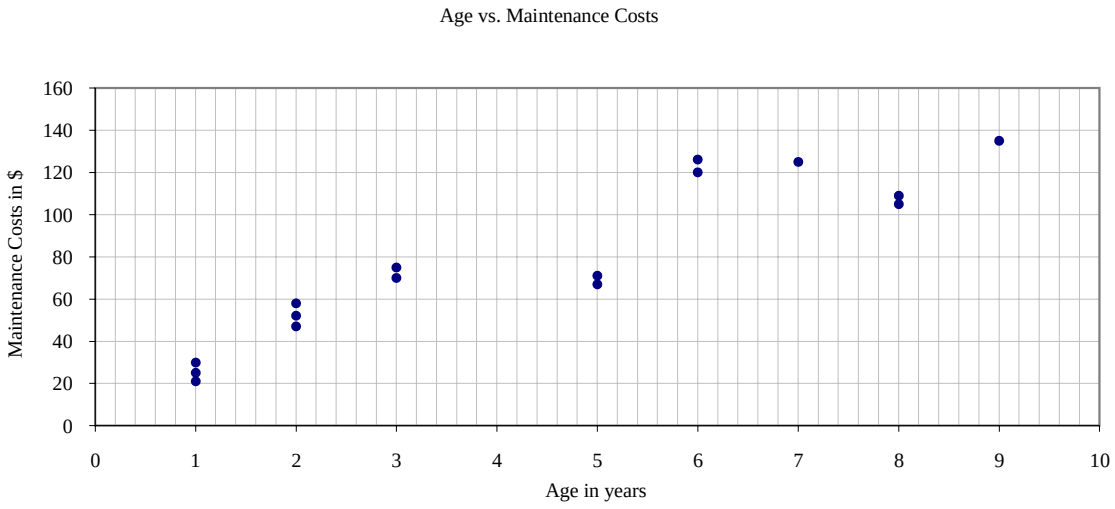
Step 1 "Maintenance costs" is the dependent variable, y, and "age" is the independent variable, x. Maintenance costs change as the machine gets older. The value of y depends on the value of x.

Step 2 The following figure shows the scatter diagram for this data.

Step 3 The scatter diagram suggests a reasonable linear relationship between maintenance costs and age.

Step 4 The regression equation is $y = ax + b \Rightarrow y = 12.67x + 22.6$

Step 5 Plot the regression line on the scatter diagram.



x	y	x^2	$x \cdot y$
8	109	64	872
3	75	9	225
1	21	1	21
9	135	81	1215
5	67	25	335
7	125	49	875
5	71	25	355
2	52	4	104
1	25	1	25
3	70	9	210
6	126	36	756
2	58	4	116
1	30	1	30
2	47	4	94
6	120	36	720
8	105	64	840
$\sum x = 69$	$\sum y = 1236$	$\sum x^2 = 413$	$\sum xy = 6793$

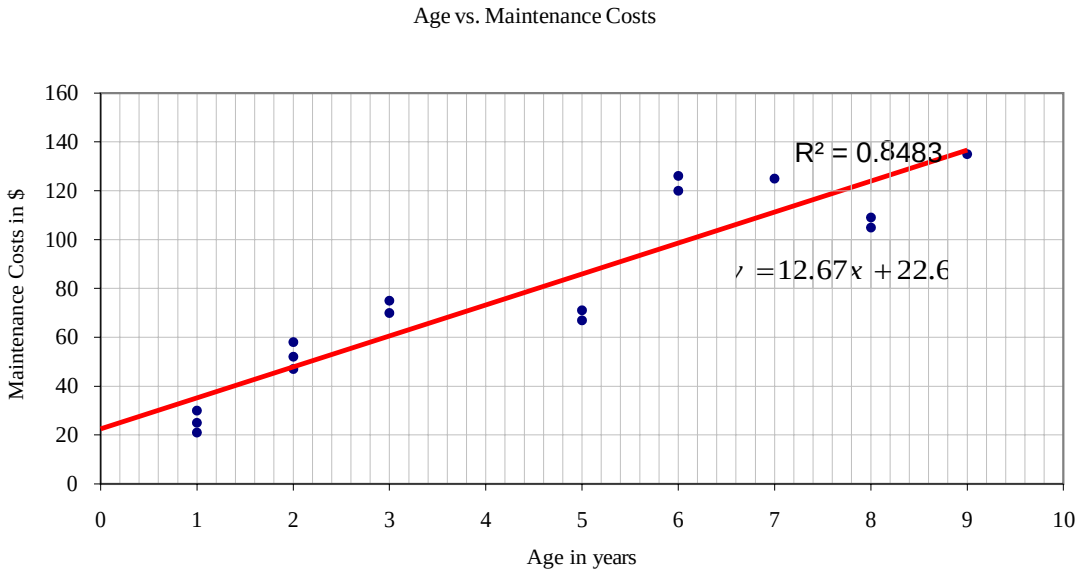
n = 16 (There are 16 pairs of data values)

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{16(6793) - (69)(1236)}{16(413) - (69)^2} = 12.67$$

The slope of the regression line is 12.67 (positive slope). The maintenance expenses will increase by approximately 12.67 dollarss per year.

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(413)(1236) - (6793)(69)}{16(413) - (69)^2} = 22.6$$

The y-intercept is 22.6 that means the maintenance expenses are \$ 22.6 when t = 0.


Age vs. Maintenance Costs
Maintenance Costs in $
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10
Age in years
R² = 0.8483
y =12.67x +22.6

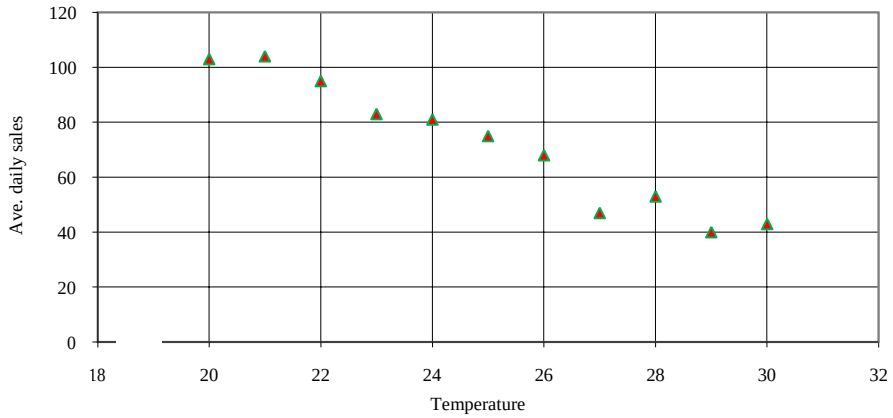
Example 3

Over a five-month period the average daily temperatures and the average daily sales of a certain product for each temperature are given in the following table.

Temperature, x (°C)	20	21	22	23	24	25	26	27	28	29	30
Average daily sales, y	103	104	95	83	81	75	68	47	53	40	43

Step 1 "Average daily sales" is the dependent variable, y, and "temperature" is the independent variable, x. Average daily sales of a certain product changes as the temperature changes. The value of y depends on the value of x.

Step 2 The following figure shows the scatter diagram for this data.



Step 3 The scatter diagram suggests a linear relationship between average daily sales and temperature. The scatter diagram indicates that the average daily sales will be less at higher temperatures.

x	y	x^2	$x \cdot y$
20	103	400	2060
21	104	441	2184
22	95	484	2090
23	83	529	1909
24	81	576	1944
25	75	625	1875
26	68	676	1768
27	47	729	1269
28	53	784	1484
29	40	841	1160
30	43	900	1290
$\sum x = 275$	$\sum y = 792$	$\sum x^2 = 6985$	$\sum xy = 19\ 033$

$n = 11$ (There are 16 pairs of data values)

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{11(19\,033) - (275)(792)}{11(6985) - (275)^2} = -6.973$$

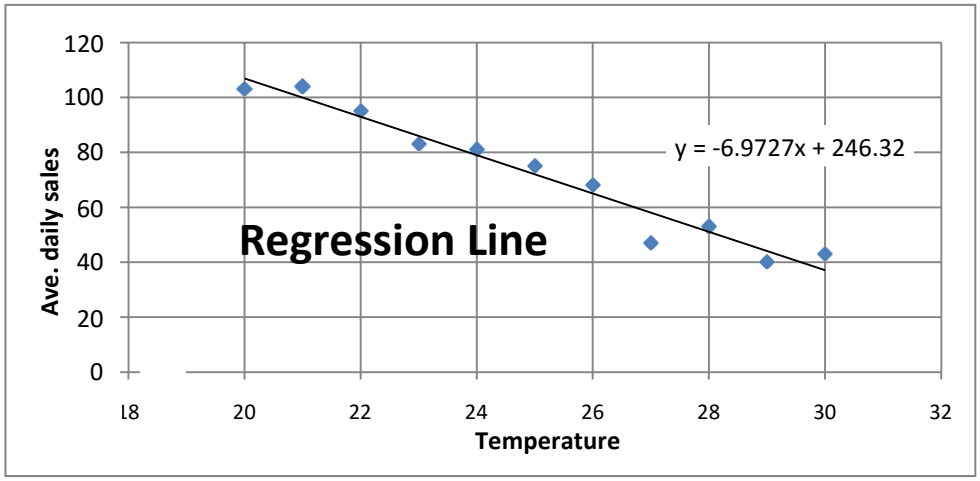
The slope of the regression line is –6.973 (negative slope). The average daily sales will decrease by 6.073 for each unit increase in the temperature.

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum xy\right)\left(\sum x\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(6985)(792) - (19\,033)(275)}{11(6985) - (275)^2} = 246.318$$

The y-intercept is 246.318 that means the average daily sales will reach around 246 when the temperature drops to 0°C.

Step 4 The regression equation is $y = ax + b \Rightarrow y = -6.973x + 246.318$

Step 5 Plot the regression line on the scatter diagram.



The horizontal axis does not start at 0°C. We cannot see the *y*-intercept from the diagram.

B. Correlation Coefficient

The correlation coefficient gives a measure of the strength of the linear relationship between two variables, x and y. The letter r is the symbol of the correlation coefficient. Unlike simple linear regression it does not require dependent and independent variables. For that reason it can be used in more applications. For example, it could be used to measure the strength of the linear relationship between height and weight without taking one of them as the dependent variable.

The formula for r is:

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Example 4

A sample of nine companies gives the following data.

Production, x (in thousands of units)	40	42	48	55	65	79	88	100	120
Manufacturing expenses, y (in thousands of $)	150	140	160	170	150	162	185	165	190

x	y	x^2	y^2	$x \cdot y$
40	150	1600	22500	6000
42	140	1764	19600	5880
48	160	2304	25600	7680
55	170	3025	28900	9350
65	150	4225	22500	9750
79	162	6241	26244	12798
88	185	7744	34225	16280
100	165	10000	27225	16500
120	190	14400	36100	22800
$\sum x = 637$	$\sum y = 1472$	$\sum x^2 = 51\,303$	$\sum y^2 = 242\,894$	$\sum xy = 107\,038$

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

$$r = \frac{9(107\,038) - (637)(1472)}{\sqrt{[(9)(51\,303) - (637)^2][(9)(242\,894) - (1472)^2]}} = 0.782131$$

The correlation coefficient is 0.782131.

Example 5

Over a five-month period the average daily temperatures and the average daily sales of a certain product for each temperature are given in the following table.

Temperature, x (°C)	20	21	22	23	24	25	26	27	28	29	30
Average daily sales, y	103	104	95	83	81	75	68	47	53	40	43

x	y	x^2	y^2	$x \cdot y$
20	103	400	10 609	2060
21	104	441	10 816	2184
22	95	484	9025	2090
23	83	529	6889	1909
24	81	576	6561	1944
25	75	625	5625	1875
26	68	676	4624	1768
27	47	729	2209	1269
28	53	784	2809	1484
29	40	841	1600	1160
30	43	900	1849	1290
$\sum x = 275$	$\sum y = 792$	$\sum x^2 = 6985$	$\sum y^2 = 62\,616$	$\sum xy = 19\,033$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}}$$

$$r = \frac{11(19\,033) - (275)(792)}{\sqrt{[(11)(6985) - (275)^2][(11)(62\,616) - (792)^2]}} = -0.97795$$

The correlation coefficient is – 0.97795.

As we can see from the examples above the correlation coefficient is always between –1 and +1. Its value within this range indicates the strength of the linear relationship between *x* and *y*.

Example 6

Zoho Service Company is trying to see how the number of shifts affects production. The following data has been collected. What conclusion can the company reach from these data?

Shifts	6	5	7	8	2	7	9	4	3	8
Production	2810	2620	3080	4200	1500	3160	4680	2330	1780	3910

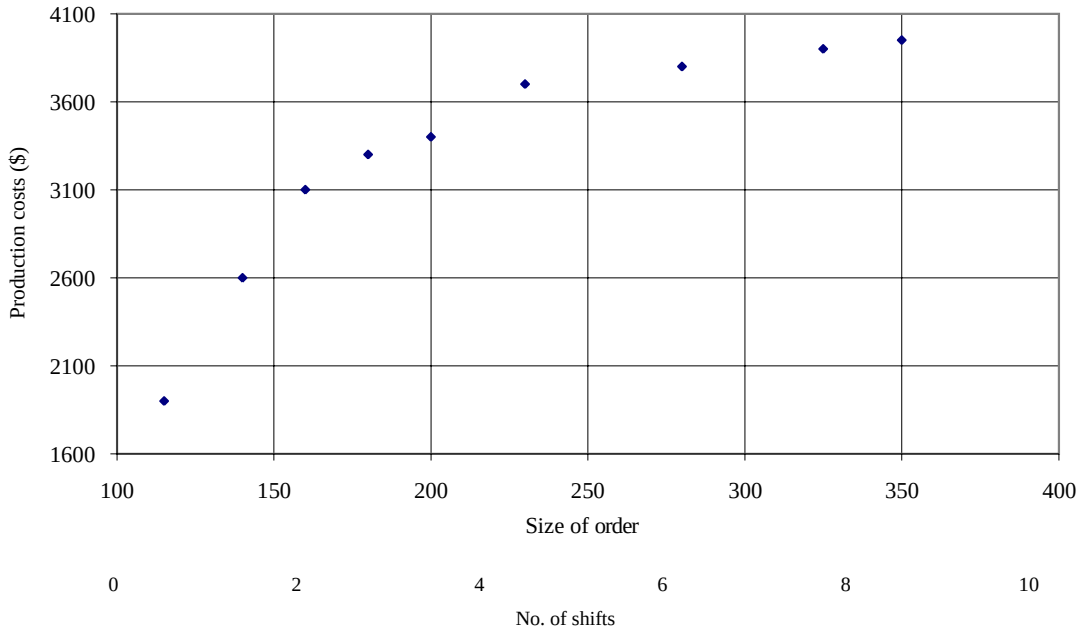
x	y	x^2	y^2	$x \cdot y$
6	2810	36	7 896 100	16 860
5	2620	25	6 864 400	13 100
7	3080	49	9 486 400	21 560
8	4200	64	17 640 000	33 600
2	1500	4	2 250 000	3000
7	3160	49	9 985 600	22 120
9	4680	81	21 902 400	42 120
4	2330	16	5 428 900	9320
3	1780	9	3 168 400	5340
8	3910	64	15 288 100	31 280
$\sum x = 59$	$\sum y = 30\ 070$	$\sum x^2 = 397$	$\sum y^2 = 99\ 910\ 300$	$\sum xy = 198\ 300$

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

$$r = \frac{10(198\,300) - (59)(30\,070)}{\sqrt{\left[(10)(397) - (59)^2\right]\left[(10)(99\,910\,300) - (30\,070)^2\right]}} = 0.9696$$

The correlation coefficient is 0.9696. It is almost +1. This indicates quite a strong positive correlation between the number of shifts and production. As the number of shifts increases the production will also increase.

The scatter diagram below shows the strong positive correlation between the number of shifts



and production.

Note that a non-linear relationship could have a high correlation, so the scatter diagram must be used together with the correlation.

Example 7

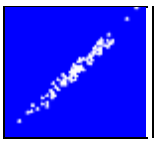
A company wants to investigate the relationship between the size of order of a certain product and the cost of producing the order. The following data have been recorded for different sized orders.

Size of order	115	140	160	180	200	230	280	325	350
Production costs ($)	1900	2600	3100	3300	3400	3700	3800	3900	3950

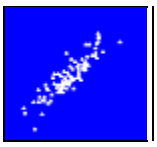
The correlation for the data is $r = 0.8782$. Although this indicates a strong positive correlation the scatter diagram below shows a non-linear relationship.

Example 8

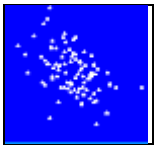
The scatter diagrams and their correlation coefficients are given below.



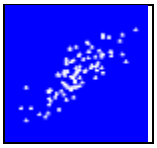
$r = 0.98$ Very strong positive linear relationship (correlation).



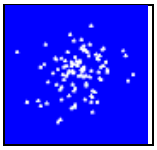
$r = 0.89$ Strong positive linear relationship (correlation).



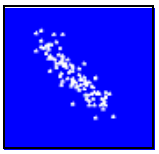
$r = -0.48$ Some negative linear relationship (correlation).



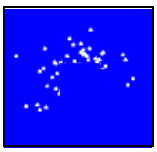
$r = 0.79$ Fairly strong positive linear relationship (correlation).



$r = 0.19$ Weak positive linear relationship (correlation).



$r = -0.87$ Strong negative linear relationship (correlation).



Non-linear relationship.

C. Coefficient of Determination

Before a regression equation can be used effectively to predict the values of the dependent variable, it is necessary to decide how well it fits the data. Another measure that gives you this information is the *coefficient of determination*.

The coefficient of determination measures the proportion of the variation in the dependent variable explained by the variation in the independent variable. It is given by r^2 or R^2 as used in Excel, which is the square of the correlation coefficient.

$$\boxed{0 \le r^2 \le 1}$$

One significant advantage the correlation coefficient has over the coefficient of determination is that it shows whether the relationship between the variables is positive or negative.

$$\boxed{-1 \le r \le 1}$$

Example 9

(a) A sample of nine companies gives the following data.

Production, x (in thousands of units)	40	42	48	55	65	79	88	100	120
Manufacturing expenses, y (in thousands of $)	150	140	160	170	150	162	185	165	190

From the calculations in Example 4 of section 4.2 we know that the value of the correlation coefficient is 0.782131.

Coefficient of determination $r^2 = (0.782131)^2 = 0.6117$

0.6117 or around 61 % of the variation in manufacturing expenses is explained by the production volume. Alternatively, 39 % of the variation is not explained by the production volume.

Note: R^2 can be obtained directly from the graph. Look at our Excel solution for Example 3 for the steps.

(b) Over a five-month period the average daily temperatures and the average daily sales of a certain product for each temperature are given in the following table.

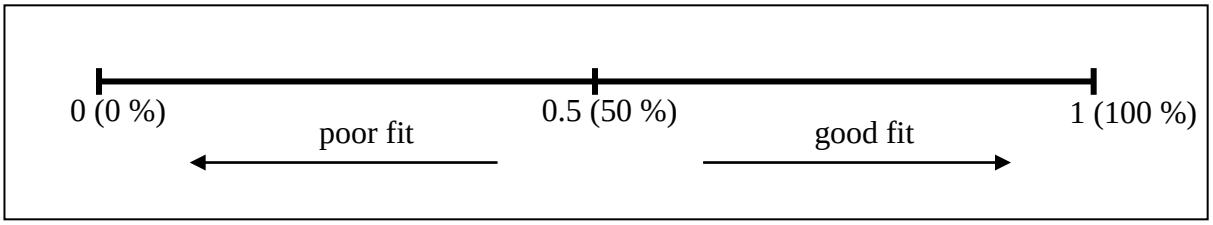
Temperature, x (°C)	20	21	22	23	24	25	26	27	28	29	30
Average daily sales, y	103	104	95	83	81	75	68	47	53	40	43

As we have calculated in Example 5 of section 4.2 the correlation coefficient is – 0.97795.

Coefficient of determination $r^2 = (-0.97795)^2 = 0.9564$

0.9564 or around 96 % of the variation in average daily sales is explained by the temperature. Alternatively, 4 % of the variation is not explained.

If a coefficient of determination had a value of 0.9, we would know that 90 % of the variation in the dependent variable, y, is explained by the regression and 10 % is unexplained. Normally, any value for the coefficient of determination above about 0.5 is considered a good fit. If the coefficient of determination is low at, say, 0.2, then 80 % of the variation is not explained by the regression and some other factors should be considered. A value close to 1 (100 %) shows a good fit, while a value close to 0 (0 %) shows a poor fit.



A general way of tackling linear regression problems:

- Decide which variable is the dependent variable.
- Draw a scatter diagram of the data and see whether it suggests a linear relationship.
- Find the regression equation and draw the regression line.
- Calculate the correlation coefficient and the coefficient of determination to see how well this line fits the data.
- Only if there is a good fit, we can use the regression equation to predict values of the dependent variable for selected values of the independent variable.