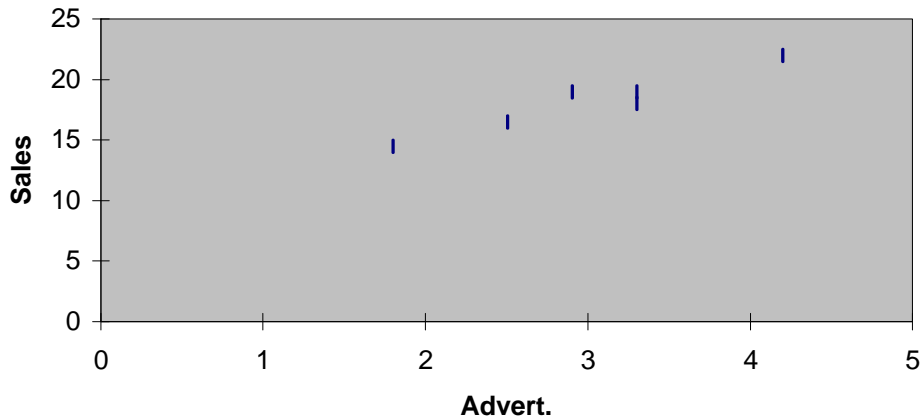


## Regression.

### Example 1:

The marketing manager of a company examines the advertising figures and sales figures for a number of different years. The values (in £1000) examined are tabulated below.

Advert	3.3	2.5	1.8	4.2	2.9	3.3
Sales	18	16.5	14.5	22	19	19



### Notes:

1. This graphical representation is called a scatter diagram. Pairs of points (advertising, sales) are plotted as you would do in coordinate geometry but do not necessarily form a smooth curve.
2. The scatter diagram gives an indication as to how the two variables advertising and sales are related. In this example it seems clear that as the level of advertising rises, the sales levels also tend to rise.

However we can make no absolute claim in this respect.

If we could quantify in some way the relationship between the two variables, sales and advertising, we could use this analysis to make predictions such as:

“What sales might be expected if £3,000 was invested in advertising?”

(The answer is likely to be about £18,000-£20,000. Why?)

We do this by finding a smooth curve (possibly a straight line) which approximately describes the behavior of our data and use this curve to make predictions.

It is easy to understand what happens graphically (see graph).

In practice we find such a curve by inputting our “bivariate” data into formulae and from this determining our curve.

The determination of such curves and their usage is known as regression.

**Notes:**

1. The example above is known as **bivariate regression** as there are only two variables involved.

If the marketing manager had decided to consider other factors then the problem would involve **multivariate regression**.

2. In situations where the smooth curve obtained is a line, the problem involves **linear regression**

We will now consider linear bivariate regression.

3. When considering a problem we need to identify the dependent and independent variables.

In the problem above, the advertising values are the independent variables because the company may choose to invest any amount they wish on advertising in a given year. i.e. These values are not dependent on anything.

The sales values are the dependent values because the company cannot decide what their sales are going to be in a given year.

These values depend on the advertising levels.

The independent variable is plotted horizontally on the scatter diagram while the dependent variable is plotted vertically.

In some problems, it is not always clear which variable is independent and which is dependent, e.g. a data set recording a number of peoples heights and weights.

**Problem**

Find the “best line” which describes the relationship between the advertising figures and sales figures given above.

**Solution**

Recall that the equation of any line is  $y = mx + c$  where  $m$  is the slope and  $c$  is the y-intercept (where the lines cuts the y-axis).

In regression we use

$a$  instead of  $c$

and

$b$  instead of  $m$ .

So the regression line is

$$Y = bX + a = a + bX.$$

Thus we need only find  $a$  and  $b$ . For this problem we find that

$$a = 9.175 \sim 9.2 \text{ and } b = 2.996 \sim 3.$$

This is done by means of the following formulae.

$$b = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\frac{\sum x^2}{n} - \bar{x}^2} \quad a = \bar{y} - b\bar{x}$$

The formula for the slope is quite difficult to work with and so it is common to split it into its numerator and denominator. Therefore we let

$$S_{xy} = \frac{\sum xy}{n} - \bar{x}\bar{y} \quad S_{xx} = \frac{\sum x^2}{n} - \bar{x}^2$$

$$S_{yy} = \frac{\sum y^2}{n} - \bar{y}^2$$

We will need the last of these three expressions later on. Correspondingly we get that

$$b = \frac{S_{xy}}{S_{xx}}$$

**Note**

1. It is necessary to find the slope (b) before finding the y-intercept (a) as b is used in the calculation of a.
2. In these formulae, x refers to the independent variable and y to the dependent variable.
3.  $\bar{X}$  refers to the sample mean of x and similarly with y.
4.  $\Sigma x$  means the sum of all data values x. Thus  $\bar{X} = \Sigma x / n$ .
5.  $\Sigma x^2$  means that all x values should be squared and then summed.
6.  $(\Sigma x)^2$  means that all x values should be summed and then squared.  
**N.B.**  $\Sigma x^2$  is not the same as  $(\Sigma x)^2$ .
7.  $\Sigma xy$  means each x value is multiplied by its corresponding y value and then the results are summed.

We get the values of 9.175 and 2.996 for a and b respectively by means of the following table.

x	y	x <sup>2</sup>	y <sup>2</sup>	x y
3.3	<b>18</b>			
2.5	<b>16.5</b>			
1.8	<b>14.5</b>			
4.2	<b>22</b>			
2.9	<b>19</b>			
3.3	<b>19</b>			

$$\begin{array}{lll} \bar{x} = 3 & \Sigma x^2 = 57.32 & \Sigma xy = 336.95 \\ \bar{y} = 18.167 & \Sigma y^2 = 2012.5 & n = 6 \end{array}$$

**Note:** You should use the STATS mode on your calculator to evaluate these.

So

$$Y = 9.2 + 3X$$

or

$$\text{Sales} = 9.2 + 3(\text{Advertising}).$$

Lets find two points on the line.

$$X = 2 \Rightarrow Y = 9.2 + 3(2) = 15.2 \quad (X, Y) = (2, 15.2)$$

$$X = 4 \Rightarrow Y = 9.2 + 3(4) = 21.2 \quad (X, Y) = (4, 21.2)$$

We can now plot these points (and hence draw the line) on our scatter diagram.

Note that the line gives a “good” indication as to how the variables are related.

We now use the equation of this line and not the original data to make predictions.

### Problem

What sales might be expected if £3,000 is invested in advertising?

### Solution 1.

Find 3 on the advertising axis, draw a vertical line to meet the regression line, draw a horizontal line to the sales axis and read of the result.

Answer ~ 18 = £18,000.

### Solution 2.

Put  $X = 2$  in the regression line  $Y = 9.2 + 3X$  to get

$$\text{sales} = 18.2 = \text{£}18,200$$

**Note** The second method is more accurate and also is quicker as we do not need to draw the scatter diagram.

### Exercise 2:

Estimate sales for this problem if advertising is

- (a) £3,300
- (b) £5,000
- (c) £1,200

### Note

If the advertising values lie between £1,800 and £3,300 then an estimation like the above is called **interpolation** and is reasonably accurate. Otherwise it is called **extrapolation** and such results should be used with caution.

The difference between them is that in the latter case we must extend our line outside the area for which we have data to make our prediction.

We can not be sure that the curve behaves in the expected way in such cases.

**Example 3:**

A chain of stores invests varying amounts in security from store to store. Losses due to theft are recorded in each of these stores for one year.

Store	1	2	3	4	5	6	7
Security	£5000	£6500	£3000	£7000	£2500	£4500	£6000
Theft	£4500	£2500	£6500	£2000	£7500	£5000	£3000

(Hint: For ease of calculation it is a good idea to re-tabulate the data above in units of £1000. i.e. £6500 = 6.5 etc.)

Plot a rough scatter diagram of the data.

Assuming that the two variables are related in a linear way, find the least squared error regression line.

Estimate expected losses if the amount spent in advertising in one store is

[a] £3500

[b] £2000

[c] £8000

and comment on your estimates.

**Note**

The above example results in a negative slope ( $b < 0$ ).

Also the scatter diagram indicates that the dependent values (theft losses) tend to decrease as the independent variable (security expenditure) increases.

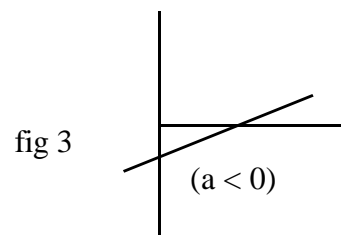
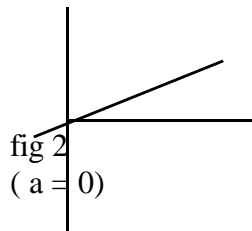
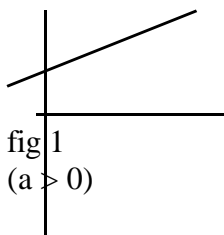
**Recall**

In coordinate geometry, a line which makes an acute angle (i.e.  $<90$  degrees) with the **negative** side of the x-axis has negative slope. Clearly we would hope to get a negative value for b in this problem.

We can anticipate the type of values (i.e. positive, negative or zero) we might obtain for a and b by considering the nature of the problem.

**Recall**

'a' is the y intercept which is the value of the dependent variable when the independent variable = 0. 'b' is the slope of the line.



### Note

In the 3 examples above we have drawn the line with positive slope, (i.e.  $b > 0$ ). In this example we are only interested in 'a'.

### Examples 4:

**a > 0:** Sales v advertising:

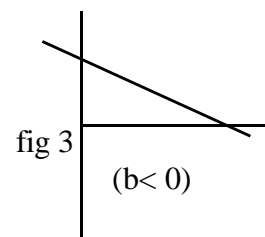
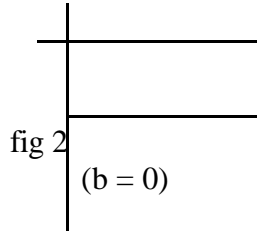
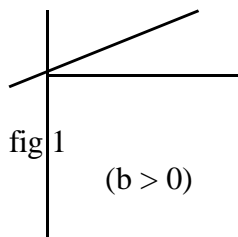
We might expect  $a > 0$  here because we expect some sales would be made if the advertising level was equal to zero.

**a = 0:** A shops income v number of customers in the shop (for 1 day):

If no customers come in to the shop, there would be zero income.

**a < 0:** Profit made by a company selling a product v price of product:

We would expect  $a < 0$  here because if the product price = 0 (the product is being given away free) the company would make a loss (profit < 0)



### Examples 5:

**b > 0:** Sales v advertising:

As the advertising levels rise, we would expect the sales levels to rise.

**b < 0:** Theft levels v security expenditure:

As security expenditure rises, we would expect theft levels to fall.

**b = 0:** If  $b = 0$ , then as the variable plotted on the horizontal axis increases, the variable plotted on the vertical does not change, i.e. the variables are **independent**.

e.g. rainfall levels vs unemployment rates

### Note

In situations where the two variables being examined are not related we would hope to find  $b$  equals (or close to) zero and vice versa. However there are (unfortunate) examples of situations where totally unrelated variables have been observed to exhibit a linear relationship.

### Example 6:

Data once collected in England showed that the number of TV licenses issued and the number of admissions to mental institutes followed a good linear relationship.

### Note

When calculating the regression line, a pair of formulae is used to determine the parameters  $a$  and  $b$ . Clearly these formulae can be used for any bivariate data set even if there is no evidence that they are related.

**Example 7:**

A personnel officer in a company is examine if the average monthly absentee rate is somehow related to the number of rainy days for the month in question. To investigate this, the following data is obtained.

Rainy days	23	15	24	12	19	20	22
Absentees	<b>11</b>	<b>18</b>	<b>13</b>	<b>15</b>	<b>16</b>	<b>14</b>	<b>22</b>

**Exercise 8:**

Show that

$$\bar{x} = 19.28$$

$$\Sigma x^2 = 2719$$

$$\bar{y} = 15.57$$

$$\Sigma y^2 = 1775$$

$$n = 7$$

$$\Sigma xy = 2083$$

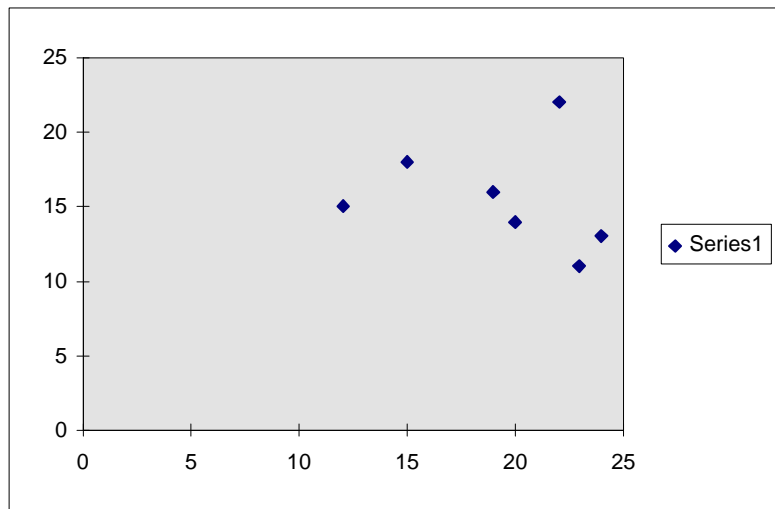
$$a = 18.77$$

$$b = -0.17$$

$$(Y = 18.77 - 0.17X)$$

This line could then be used to make predictions as before.

However, if we plot the data on a scatter diagram, we get the following



Clearly there is no real relationship between the two variables. Note that the value for b we obtained (-0.17) was very small.

This is to be expected when there is no relationship between the variables.

**N.B.**

We can not decide that two variables are unrelated by examining b. If there is no relationship, we would expect b to be small.

But if b is small that does not necessarily mean that there is no relationship.

Alternatively we use another quantity to measure the strength of the relationship between two variables.

## The Correlation Coefficient & Coefficient of Determination.

The correlation coefficient (  $r$  ) is used to measure the strength of the relationship between 2 (or more variables).

The variable  $r$  may take any value between -1 and +1 but it is  $r^2$  (**the coefficient of determination**) which has a more meaningful interpretation.

In any bivariate data set the Y values (dependent variables) will exhibit variation (i.e. not all Y values will be the same).

Some of that variation will be accounted for by variation in the independent variable. (i.e. If X increases, then (say) Y decreases)

However other variation in Y will exist. This is evident from the fact that some of the points (most in fact) on the scatter diagram do not fall on the regression line.

Thus the variation in Y can be divided into two parts:

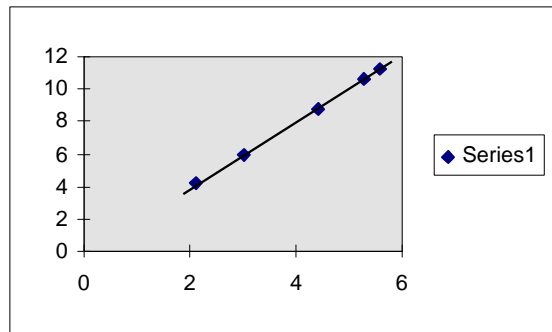
1. Variation due to regression line
2. Variation not due to regression line

The value of  $r^2$  is the proportion of the total variation which is explained by the regression line.

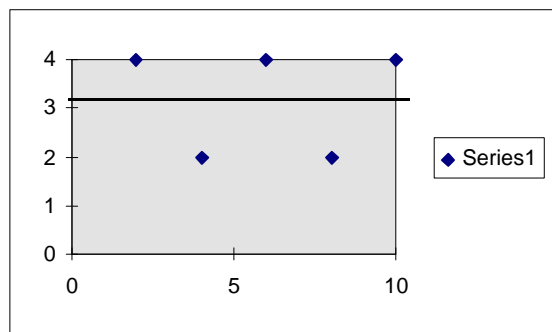
### **Example 9:**

If  $r = 0.8$ , then  $r^2 = 0.64$  so 64% of the variation in Y is due to the variation in X and 36% is not due to X.

Below is shown two extreme cases



$r^2 = 1$  (all the variation in Y is explained by X)



$r^2 = 0$  (none of the variation in Y is explained by X)

In most cases ( $0 < r^2 < 1$ )

To calculate the correlation coefficient  $r$  (and thus  $r^2$ ) we use the following formula

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

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

Using the notation we introduced earlier we get

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

**Exercise 10:**

Show that for the last example (rainy days v absentees) the value of  $r = -0.202$  and so  $r^2 = 0.04$

Thus only 4% of the variation experienced in the average number of absentees in a month can be explained by the number of rainy days.

Thus we might well decide that this is insignificant.

Alternatively we might decide that the average number of absentees in a month is determined by a number of factors all of which need not have a very large effect and the number of rainy days might be one factor.

The decision we make very much depends on the problem in question. It is easier to draw conclusions for large values of  $r$ .

**Recall**

The correlation coefficient can be used to measure the strength of evidence to support two variables being related with a linear curve.

**Example 11:**

If empirical data for the amount spent on advertising (X) by a firm and the corresponding resultant sales (Y) yields a coefficient of determination of say 0.92 then we could expect a strong linear trend on the scatter diagram.

Also we could use the determined regression equation with confidence.

**Note:** However this does not establish cause and effect, i.e. we cannot say that a change in Y (effect) is due to a change in X (cause).

**Example 12:**

Recall that records in Britain show that a strong correlation exists between the number of TV licenses purchased in a given year and the number of admissions to mental institutes!

Could we conclude that purchasing a TV license has an effect on people's mental health?

The following table gives the figures for the number of TV licenses (1,000's) purchased in a given year during the 60's and the corresponding number of admissions (100's) to mental institutes.

Year	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
TV's	21	32	37	43	51	62	70	75	79	81
M.I.'s	2.95	4.33	4.95	5.7	6.7	8.08	9.07	9.7	10.2	10.45

A plausible explanation for the phenomena above goes as follows.

The records may have been taken over a period of years when the population of Britain was changing.

Thus if the population increased by say 5%, it would be reasonable to expect the number of TV licenses AND the number of admissions to mental institutes to increase by about 5%.

Thus a sensible reasoning might see population change as the cause and the number of TV licenses as the effect. And similarly with the number of admissions to mental institutes.

**Problem**

How can a cause and effect relationship be established between two variables?

**Solution** Using controlled experiments.

**Example 13:**

A new drug is introduced to the market which claims to be effective at treating a particular ailment. Does the drug have an effect?

- A sample of people suffering from this ailment are randomly selected and divided into two groups.
- One group is treated with the drug while the other group are not
- After the treatment period is over the % of patients who received the drug and recovered is compared with the % of patients who did not receive the drug and recovered.
- If the first % is significantly larger than the second, then we can deduce that we have cause and effect.

**Problem**

Do statistics prove that smoking is bad for your health?

No! There is strong correlation between smoking and various health problems but this does not establish cause and effect.

It is plausible (but most unlikely) that some other (unknown) reason has the dual effect of adversely affecting peoples health and also of enticing them to smoke.

To prove it statistically, we would need to carry out an experiment similar to the one described above.

i.e. We would need to take a sample of non-smokers and require half the sample to spend their life smoking!

Clearly this cannot be done.

**Problem**

Is smoking bad for your health?

**Answer**

Almost certainly!

**Exercise 14:**

Show that  $r = 0.96$  in the very first example we dealt with involving advertising levels and sales levels.

Thus  $r^2 = 0.92$  which means that 92% of the variation in the sales values can be explained by the variation in the advertising levels.

**Exercise 15:**

The following data relates the number of years that students spent studying French in school and the grade the received in a proficiency test.

Years	3	4	4	1	5	3	4	5	3	2
Grade	<b>57</b>	<b>78</b>	<b>72</b>	<b>58</b>	<b>89</b>	<b>63</b>	<b>73</b>	<b>84</b>	<b>75</b>	<b>48</b>

- (a) Plot a scatter diagram for the data.
- (b) Find the least squared error regression line relating the two variables.
- (c) Comment on the strength of evidence that relates the two variables.
- (d) Estimate the grade that a student might receive if the number of years they had spent studying French in school was

[1] 3 years [2] 6 years [3] 0 years

Comment on your estimates.

**Exercise 16:**

The following data shows the yields of wheat in bushels per acre recorded for a corresponding rainfall level in inches.

Rain	12.9	7.2	11.3	18.6	8.8	10.3	15.9	13.1
Yield	<b>62.5</b>	<b>28.7</b>	<b>52.2</b>	<b>80.6</b>	<b>41.6</b>	<b>44.5</b>	<b>71.3</b>	<b>54.4</b>

- (a) Draw a (rough) scatter diagram of the data.
- (b) Find the parameters of the regression line.
- (c) Estimate the expected yield if the rainfall level is  
[1] 15 inches [2] 18 inches [3] 19 inches  
and comment on your estimates.
- (d) “The yield in any particular year depends on several factors but rainfall levels is the most significant factor”  
Comment on this statement.