CORRELATION

Contents

If viewing on screen you can click on a contents item above to jump to the page the item is on

1. Introduction

Correlation is concerned with the strength of the relationship between two variables. In clinical assessment practice it is usually the strength of the **linear** relationship between tests or tests and criteria which are of interest, and for which we are likely to have data. The correlation coefficient most frequently encountered is the Pearson Product Moment correlation coefficient. Unless otherwise stated it is this correlation which is referred to as **r** (usually with subscripts, eg r_{xy}) in what follows.

But, of course there can be all sorts of linear and non-linear relationships between variables. Several of these are shown in the graphs below.

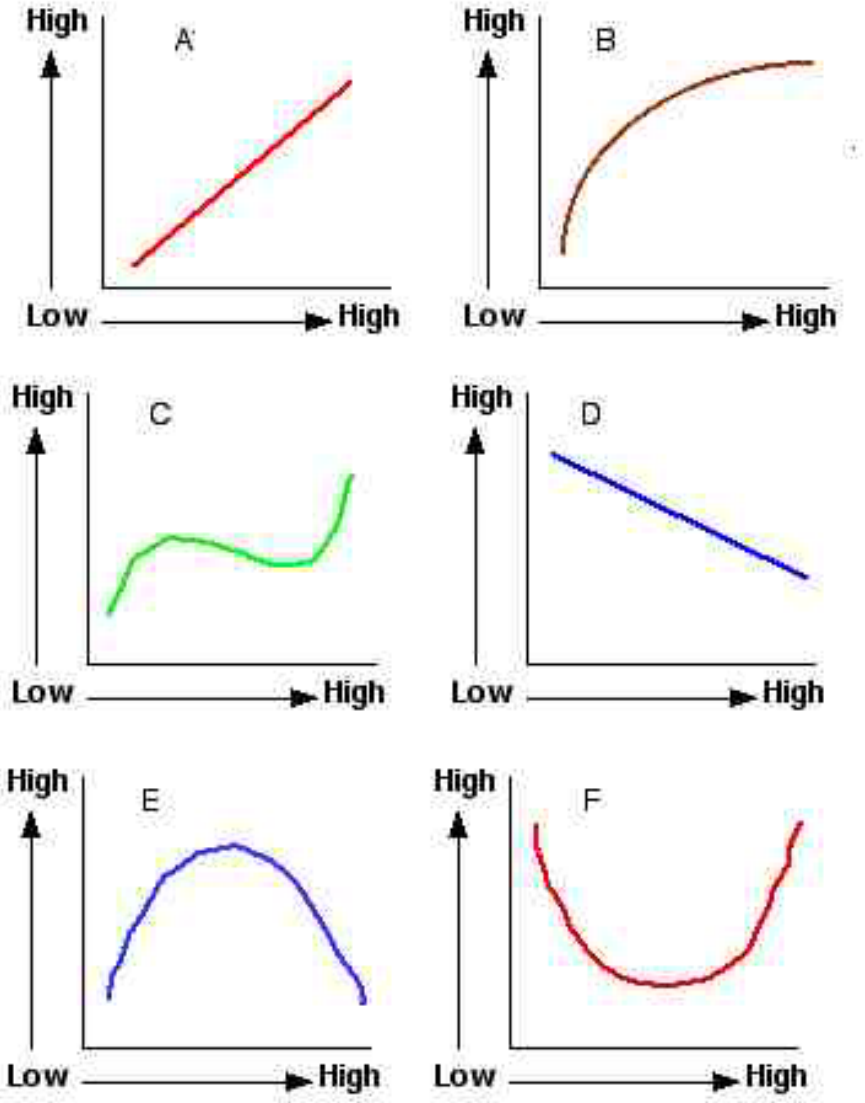
The linear ones are of course Graphs A and D.

Graph A shows a positive linear relationship with both sets of scores rising or falling together.

Graph D shows a negative relationship. As scores on one variable rise, scores on the other variable fall.

Most of the other relationships would be rarely if ever found in clinical assessment, but one of them, the inverted 'U' relationship shown in graph E, is of interest and will be discussed in more detail later.

Some linear and non-linear relationships

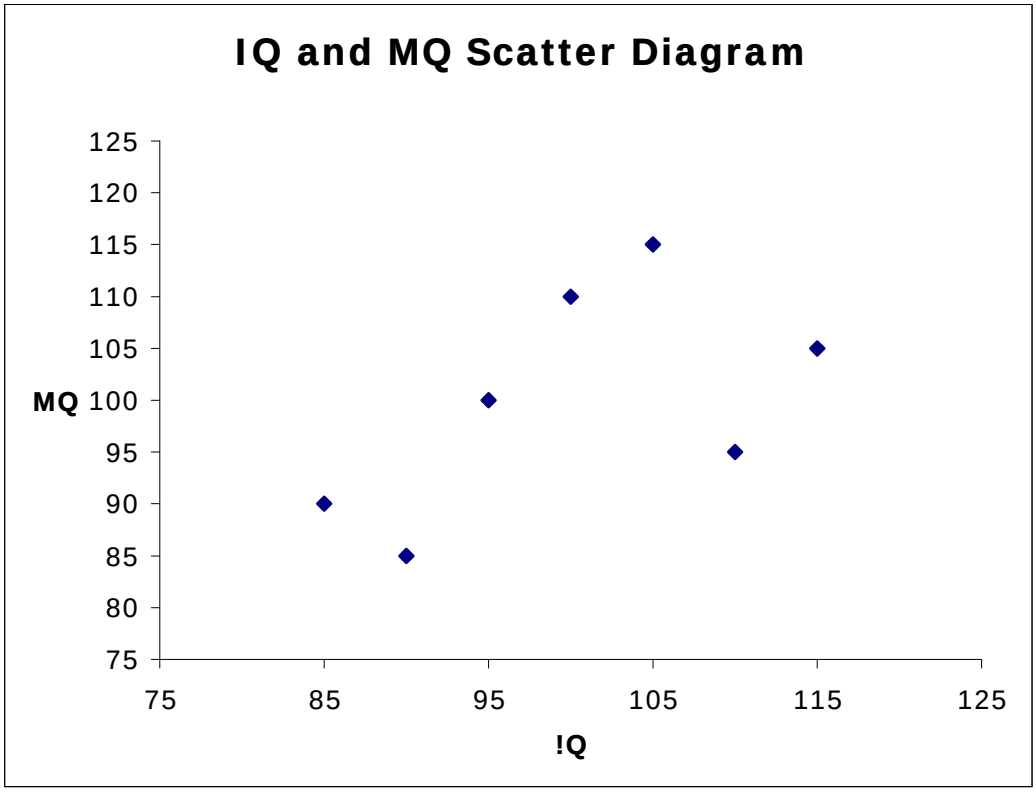


2. Which straight line should be fitted

Below are some hypothetical scores for a small group of people who have been assessed for both IQ and MQ.

Hypothetical IQ and MQ data	
IQ	**Memory Quotient**
85	90
90	85
95	100
100	110
105	115
110	95
115	105

Let's have a look at a scatter diagram of these scores

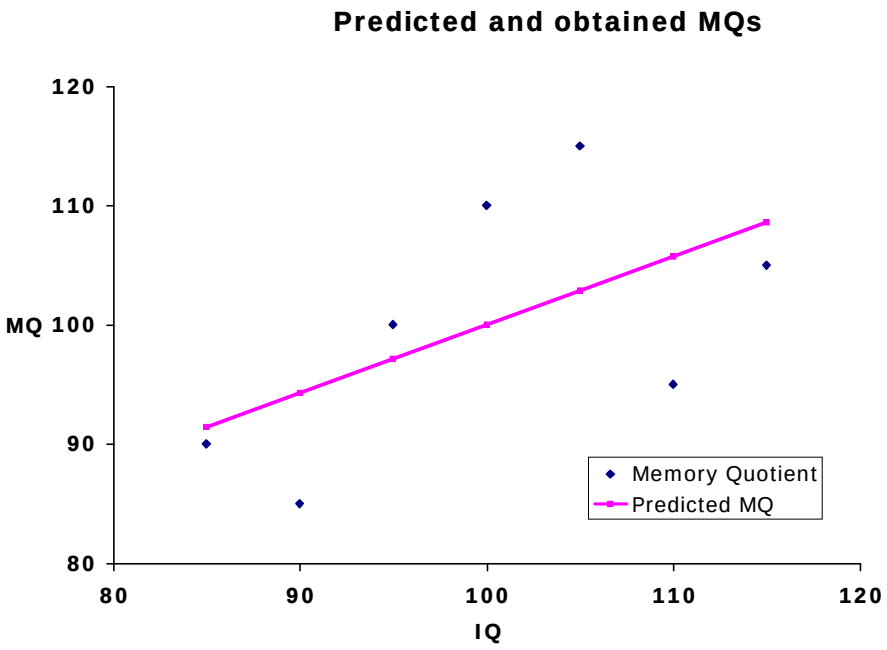


As you can see there is no particularly obvious pair of points between which to draw a straight line.

So how do we decide where the line goes?

The decision is made by using the criterion of **least squares**. That straight line is fitted which leads to the smallest possible value of the sum of squared deviations from it.

Applying this criterion leads to the regression line shown below



The predicted MQ score for someone of a given IQ will be the value shown opposite the point on the regression line perpendicularly above the IQ.

Not that people actually use the graph in this way. To predict we use a formula.

The predicted scores for our hypothetical example are shown below

Hypothetical IQ and MQ data And predicted MQ values		
IQ	**Memory Quotient**	**Predicted MQ**
85	**90**	**91.4**
90	**85**	**94.3**
95	**100**	**97.1**
100	**110**	**100**
105	**115**	**102.9**
110	**95**	**105.7**
115	**105**	**108.6**

The correlation coefficient between IQ and MQ in this example is + .57.

There will also be a regression line for predicting IQ from MQ, thus two regression lines can be fitted to each scatter plot. When the correlation in zero, the two regression lines are at right angles to one another

As the correlation between variables increases the lines get closer and closer together until they coincide with one another when the correlation coefficient equals 1.

The lines intersect at the point where the mean of X and the mean of Y meet on the regression lines.

3. Formulas for the correlation coefficient

The formula that we will use most often is:

$$r_{xy} = \frac{\Sigma ZxZy}{N} \qquad 3.1$$

The raw score formula can easily be derived from (3.1)

$$r_{xy} = \frac{\Sigma ZxZy}{N} = \frac{\sum\left(\left(\frac{X - Mx}{\sigma_x}\right)\left(\frac{Y - My}{\sigma_y}\right)\right)}{N} =$$

$$r_{xy} = \sum \frac{(X - Mx)(Y - My)}{N\sigma_x\sigma_y} \qquad 3.2$$

4. Simple prediction

In clinical practice, one of the main uses of correlation coefficients is for prediction. As you know by now, we assume that the relationship between the variables of interest is linear, i.e. that it can usefully be described by a straight line..

So how do we make the predictions? Quite simply.

Remember the formula for a straight line is:

$$Y = a + bX \qquad 3.3$$

Where:

a = the value of Y when X is zero

and

b = the slope of the line.

If we are using Z scores, the value of Zy when Zx equals zero will be zero as well.

The slope of the line is given by the correlation coefficient $r_{xy,}$ and as the value of ***a*** is zero we do not need to worry about it

So the Z score variant of the formula for predicting Y from X is;

$$\hat{Z}y = r_{xy}Zx \qquad 3.4$$

(The 'cap' on top of *Zy* is there simply to show that it is a predicted score.)

For example, if we wanted to predict performance on a test of verbal memory (mean 10 standard deviation 3) which correlates .5 with Wechsler IQ, what would we expect someone with a IQ of 115 to score on the verbal memory test?

The IQ *Z* is simply (115-100)/15, which equals 1. So the predicted verbal Memory Z will be 1 x .5 = .5.

This translates into a test score of (.5 x 3) +10, which equals 11.5.

Test yourself

A. A test of Neuroticism correlates .6 with a measure of depression. Someone obtains a score at the mean on the Neuroticism Scale. What is there expected Z score on the Depression Scale?

B. A Wechsler intelligence test correlates .7 with a test of reading comprehension (mean 50, standard deviation 10). What is the expected IQ of someone who scores 30 on the reading test?

C. What would be the expected IQ if the reading test score was 80?

Answers

A. zero; B. 79; C. 131.5

Notice in all of these examples that if the **predictor** score is anything other than the mean, and the correlation coefficient is less than unity, the predicted score will always be closer to the mean than the predictor score is.

This phenomenon is known as regression towards the mean..

It also applies in the case of somebody being tested two or more times with the same test – as in before and after assessment to evaluate treatment.

5. The interpretation of correlation coefficients

Given a correlation between X and Y, the variance of Y can be split into two main parts.

Part of the variance will be that which is predictable from X, and the other part is that not predictable from X.

Putting this in terms of a formula gives us:

$$\frac{\Sigma\left(Zy-\bar{Z}y\right)^2}{N}=\frac{\Sigma\left(Zy-\hat{Z}y\right)^2}{N}+\frac{\Sigma\left(\hat{Z}y-\bar{Z}y\right)^2}{N} \qquad 3.5$$

(Don't forget that the variance of Z scores is 1.0, and that the mean of Z scores is zero)

The term on the left is the variance of test Y. As we are dealing with Z scores this will of course equal 1.

The next term is the sum of squared differences from the regression line.

The third term is the sum of squared differences between the predicted score and the mean of Test Y. As we are working with Z scores the mean of test Y will equal zero. So the third term becomes

$$\sum(r_{xy}Zx)^2/N \qquad 3.6$$

In turn this becomes

$$r^2_{xy}\ (\sum Zx^2/N) \qquad 3.7$$

But $(\sum Zx^2/N)$ is the variance of Zx so it equals 1

and the whole expression therefore = r^2_{xy}

Thus the proportion of the variance of Y which is accounted for by X is equal to the square of the correlation coefficient,

If the correlation between X and Y is .5 we can say that X **accounts for** 25 percent of the variance of Y; if *r* is .8 then X accounts for 64 percent of the variance of Y, and so on.

The name given to r^2_{xy} is the **coefficient of determination**.

Returning to our formula;

$$\frac{\Sigma\left(Zy-\bar{Z}y\right)^2}{N}=\frac{\Sigma\left(Zy-\hat{Z}y\right)^2}{N}+\frac{\Sigma\left(\hat{Z}y-\bar{Z}y\right)^2}{N} \qquad 3.8$$

We now know that the first term, being a Z score variance must equal 1.

We also know that the third term = r^2_{xy}

It follows from this that the second term must equal $1 - r^2_{xy}$

This is the variance of the Y scores around the regression line. Its square root $\sqrt{1-r^2_{xy}}$ is called the **standard error of estimate** or the **standard error of prediction**

For the record the raw score formulas for these values are;

(3.7) Coefficient of determination = $r^2_{xy}\,\sigma_y^2$

(3.8) Standard error of estimate = $\sigma_y\sqrt{\left(1-r^2_{xy}\right)}$

As you can see, when r =1 the standards error of estimate will be zero - there will be no error at all in predicting *Y* from *X*. The prediction will be spot on in all cases.

When the correlation is zero the standard error of estimate will equal the standard deviation of *Y*. knowing somebody's *X* score tells you nothing about their *Y* score, which could be anywhere in the distribution of *Y* scores.

There is another measure which is rarely used but worth mentioning. The coefficient of determination tells us how much of the variance is accounted for. Its square root is the correlation coefficient, which is an index of the relationship between *X* and *Y*.

So, why not have another index which indicates the degree of **lack of relationship**? This, by analogy, would be the square root of the variance **not** accounted for.

Such an index exists. It is called the **coefficient of alienation** symbolised as ***k***. Its formula is as follows.

Coefficient of alienation (*k*) = $k = \sqrt{1 - r^2_{xy}}$ 3.9

It is not until r is greater than .7071 that r exceeds k.

Below that value, the lack of relationship is greater than the presence of relationship.

But this index is of little practical value to us, except as a possible way of interpreting the magnitude of a correlation coefficient of a given size.

6. Back to practical prediction

When we use the formula $Zy = r_{xy}\, Zx$ in a prediction problem the answer will of course lie on the fitted regression line.

And we also know that for people with a given Zx there will be a scatter of scores around the regression line. They will not all get the same score. In fact their scores will be normally distributed about the predicted score, and the standard deviation of that distribution will be the standard error of estimate.

So a group of people getting a Zx of 2 on an intelligence test (X), which correlates 0.8 with a test of reading comprehension (Y), would be expected to obtain a mean Zy of 1.6, and their scores would be normally distributed around that point. The standard deviation of that distribution would be (in Z scores) the square root of $(1 - r_{xy}^2)$ which will be 0.6.

Suppose now that somebody has suffered a head injury which it is suspected has lead to impairment of reading ability.

On a Wechsler intelligence test IQ is 145, while on the Reading Test, which also has a mean of 100 and a standard deviation of 15, the Reading score is 135.

The question is whether this represents evidence that reading ability has been impaired.

Using the formula:

$$\hat{Z}y = r_{xy}Zx$$

We expect people of this IQ to obtain a mean Z score of 2.4 on the reading test. In fact the patient's Reading Z is 2.33.

We can also translate this Z of 2.33 into a Z on the distribution of reading scores for people with IQs of 145.

This Z which we will call *Zpsd* (Z for the predicted score distribution) can be calculated from the formula:

$$Z_{psd} = \frac{Zy - \hat{Z}y}{\sqrt{\left(1 - r_{xy}^2\right)}} \qquad 3.10$$

where (as usual)

$$\hat{Z}y = r_{xy}Zx$$

In this example *Zpsd* will equal (2.33 –2.4)/ .6 which equals approximately – 0.12. Looking this up in the table of values for the normal curve or by using the calculator below we can see that about 45 percent of people with an IQ of 145 would obtain lower reading scores than this.

So we would have to conclude that the reading score offers no evidence of decline in reading ability.

Test yourself

A test of abstract reasoning Test X (Mean 50, sd, 10) correlates .8 with Test Y - a similarities subtest (mean 10, sd. 3)

a. If somebody scores 60 on Test X, What would you expect them to score on test Y?
b. What percentage of those who score 70 on Test X would you expect to score below 10 on Test Y?
c. If someone scores 4 on test Y what would you expect them to score on Test X?

Answers

a. 12.4; b. about 0.4 percent; c. 34

7. More on curvilinear relationships

As already stated, the usual correlation coefficients found and used by clinicians are measures of the linear relationships between two variables. For most purposes this assumption of linearity is justified.

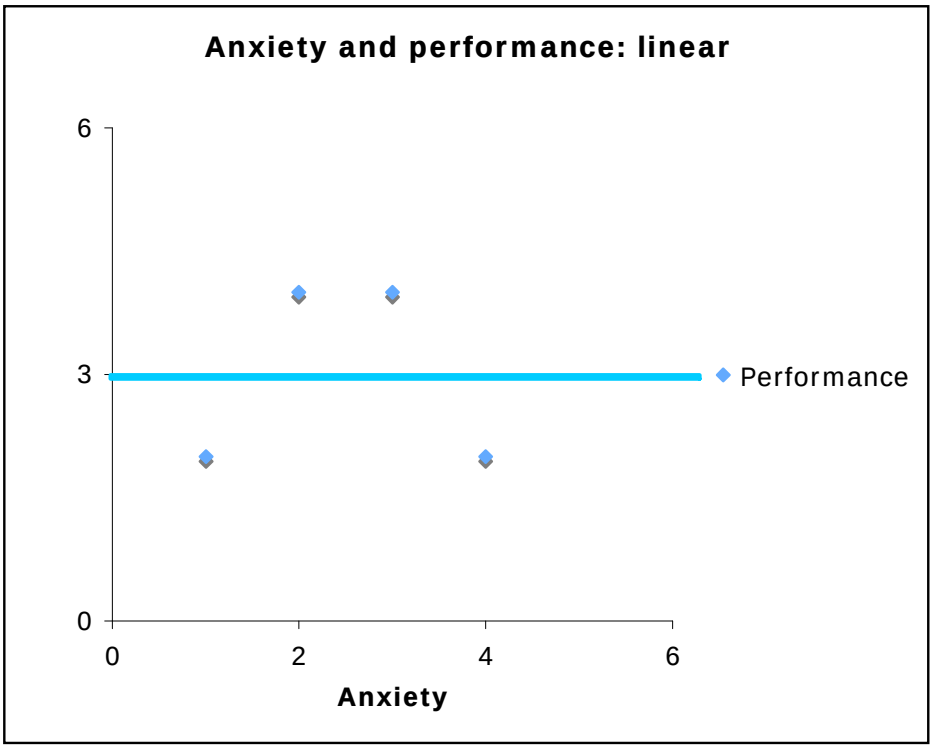
However there is a class of variables subsumed under the heading ‘arousal’ where it cannot automatically be assumed that the relationship is best described by fitting a straight line. Arousal variables often have a curvilinear relationship with measures of performance. This curvilinear relationship sometimes goes under the name of the Yerkes-Dodson Law, so named because it was Yerkes and Dodson who first drew attention to it in 1908.

Essentially it states that there is a curvilinear relationship between ‘arousal’ variables like anxiety or stress and performance, and that this relationship takes the form of an inverted ‘U’.

Performance at low levels and at high levels of arousal is poorer that at some intermediate level. Such a relationship can be seen in the table below.

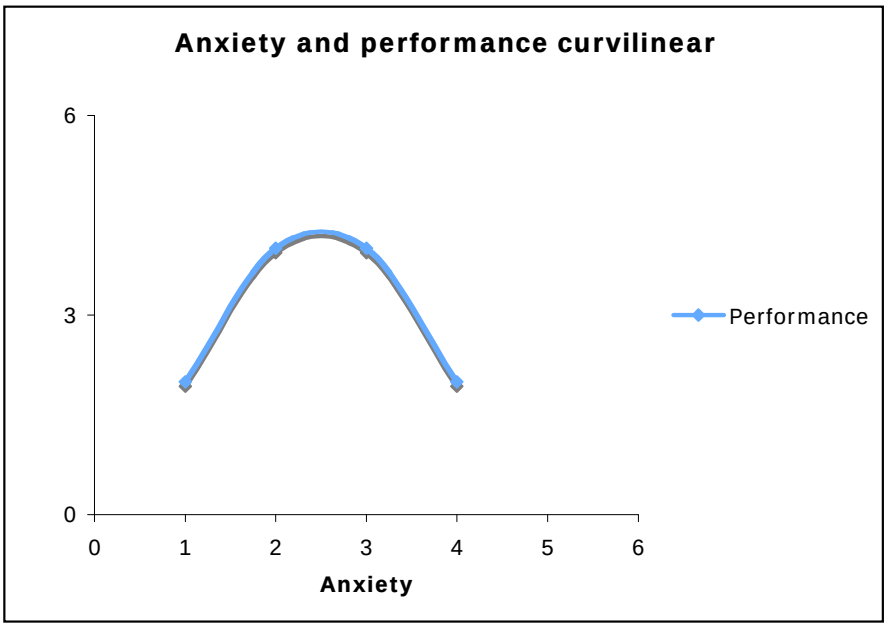
Anxiety level	Performance score
1	2
2	4
2	4
4	2

The graph below shows the best-fit linear regression line. This fit represents a correlation of zero.



This suggests that performance cannot be predicted at all from anxiety scores. In fact, the correlation is zero.

But if we fit a curvilinear regression line to these data, we obtain a correlation of 1. The performance scores can be predicted perfectly from the anxiety scores.



The message is simply this. Whenever correlating arousal variables with performance variables be aware that the relationship might well be curvilinear.

In such circumstances, linear correlation could well give a misleading result

This applies to performance on psychological tests as well as other measures of performance.

For example, a recent investigation by Bierman, Comijs, Jonker, and Beckman (2005) reported curvilinear relationships between anxiety and performance on various measures drawn from the AVLT.

And do not forget that certain drugs can increase arousal as well. A recent example of this was reported by Tipper, Cairo, Woodward, Phillips, *et al* (2005) who found a curvilinear relationship between dose of amphetamine and working memory performance.

There was an inverted U relationship between the amount of amphetamine administered and working memory processing efficiency.

So, with arousal variables curvilinear relationships are not uncommon.

And, remember that they can occur in other circumstances as well.

Finally it is also worth remembering that the optimal level of arousal varies inversely with the difficulty of the task. The harder the task the lower is the level of arousal required for best results.

8. Eta – a general measure of linear and curvilinear relationships

So what do we do if we have a curvilinear relationship and wish to assess the strength of the relationship between two variables?

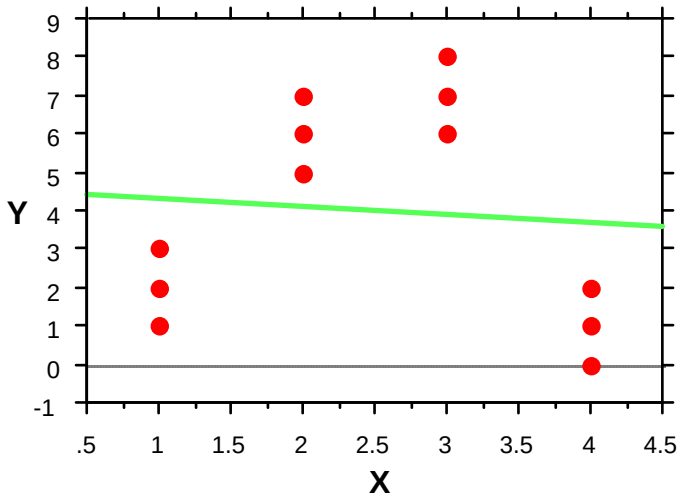
One thing we could do is use **eta**, also known as the **correlation ratio**, which is a useful general purpose measure of relationship between two variables. It makes no assumptions about the linearity of the relationship.

Consider the following data.

Arousal score	Performance score
1	1
1	2
1	3
2	5
2	6
2	7
3	6
3	7
3	8
4	0
4	1
4	2

If we calculate r for these data we obtain the following regression line for performance as predicted from arousal.

The correlation coefficient r = - 0.084



Y = 4.5 - .2 * X; R^2 = .007

What happens if we calculate eta ?

There are a number of ways of arriving at the value of eta. If you have access to a one-way analysis of variance program, you can use the formula:

Eta will equal the square root of (the between sum of squares divided by the total sum of squares)

$$\eta = \sqrt{\frac{BetweenSS}{TotalSS}} \qquad 3.11$$

(If you do not have access to a statistical package, you could try the free one “Smiths Statistical Package” downloadable in Mac and Windows versions from:

http://www.economics.pomona.edu/StatSite/framepg.html

This will enable you to effortlessly carry out a range of statistical procedures)

Below is the result of an analysis of variance carried out on the data above.

To conduct the analysis the data were re-cast as follows.

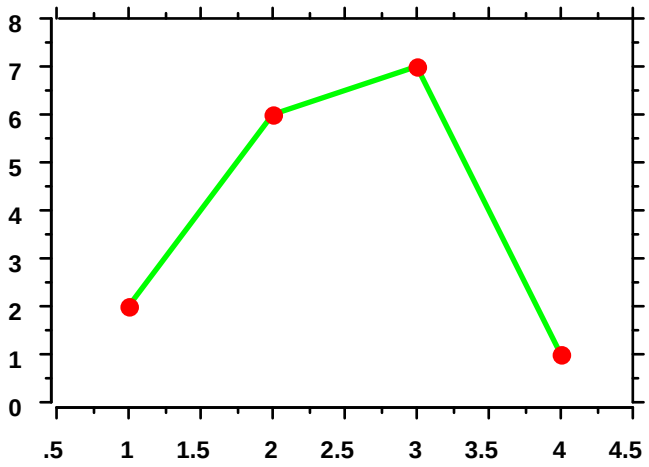
The performance scores have been grouped by using arousal scores as the grouping factor

Arousal Score	1	2	3	4
Performance	*1*	*5*	*6*	*0*
scores	*2*	*6*	*7*	*1*
	3	*7*	*8*	*2*

An analysis of variance of these data, grouped in this way, gave the following results:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	78.000	3	26.000	26.000	.000
Within Groups	8.000	8	1.000		
Total	86.000	11			

There is highly significant association between arousal and performance. This relationship looks like this



To make use of the relationship we simply decide that the best prediction will be the mean of the performance score in the column associated with a given arousal score. So our best bet for someone scoring 1 for arousal is that they will score 2 for performance, and so on.

Applying the formula for eta to these results we get

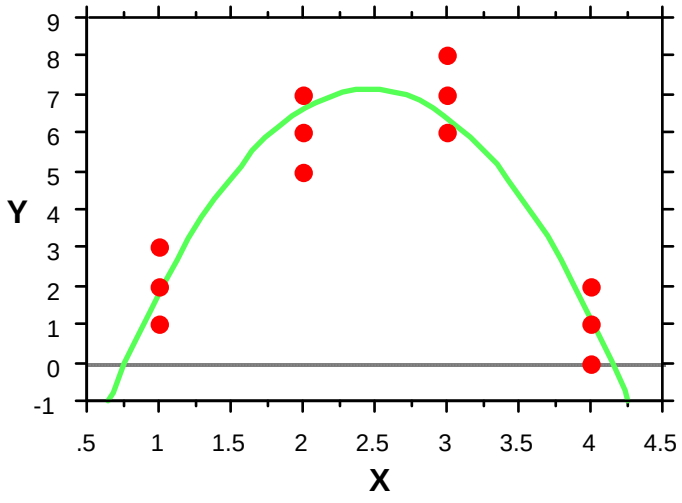
$$\eta = \sqrt{\frac{78}{86}} = .95$$

However, sometimes we can fit a mathematical formula quite closely to the data, and this can often give a more useful and exact prediction.

For the present data the following equation gives a good fit:

$$Y = -8 + 12.3\,X - 2.5\,X^2$$

The graph below shows the resulting curve. The correlation is .938.



Y = -8 + 12.3 * X - 2.5 * X^2; R^2 = .879

You might already have guessed that comparing the values of **eta** and **r** might give some idea of whether a relationship shows significant departure from linearity. One such test of significant departure from linearity is.

$$F = \frac{(\eta^2 - r^2)(N-k)}{(1-\eta^2)(k-2)} \qquad 3.12$$

Where:

N = the number of case

k = the number of categories

Applying this formula to our obtained r and eta gives:

$$F = ((.953^2 - .084^2) \times (11 - 4)) / ((1 - .953^2) \times (4 - 2))$$

Which equals 34.36 which demonstrates a highly significant departure from linearity.

An alternative method involves calculating a chi-square value .

$$\chi^2 = (N-k))\frac{\eta^2 - r^2}{1-\eta^2} \qquad 3.13$$

The main aim of introducing eta is to make you aware of the existence of a reasonably simple measure which is particularly useful with non-linear relationships.