Paper C03 (also relevant to P1, P3) Fundamentals of Business Mathematics

Correlation analysis is a useful forecasting tool, but it's important not to jump to the conclusion, if a strong correlation between a pair of factors is found, that variations in one are the result of variations in the other

By Bob Scarlett

Two variables are said to correlate if a change in one of them is accompanied by a predictable change in the other. The concept of correlation is commonly encountered in a range of techniques used in business forecasting and modelling.

If both of the variables in question are numerical, a technique known as the Pearson method can be used to calculate the degree to which they correlate. The result is expressed as a correlation coefficient, otherwise known as the Pearson coefficient or r score. If one or both of the variables are not given in a suitable quantitative form, an alternative approach can be used to measure the degree of correlation, which is expressed in such cases as Spearman's rank correlation coefficient.

The basic mathematics behind the Pearson method can be illustrated using the simple case of a class of students. There are six people in the class, each of whom sits a maths exam and then an English exam the following week. Suppose that each student achieves exactly half of the mark in their English exam that they scored in their maths paper: in this case the correlation between their maths and English scores is perfect and the Pearson coefficient derived from comparing the two sets of results is 1:

Case 1: perfect positive (linear) correlation

Student	Maths mark	English mark
1	80%	40%
2	60%	30%
3	44%	22%
4	26%	13%
5	70%	35%
6	64%	32%

Pearson coefficient: 1.000

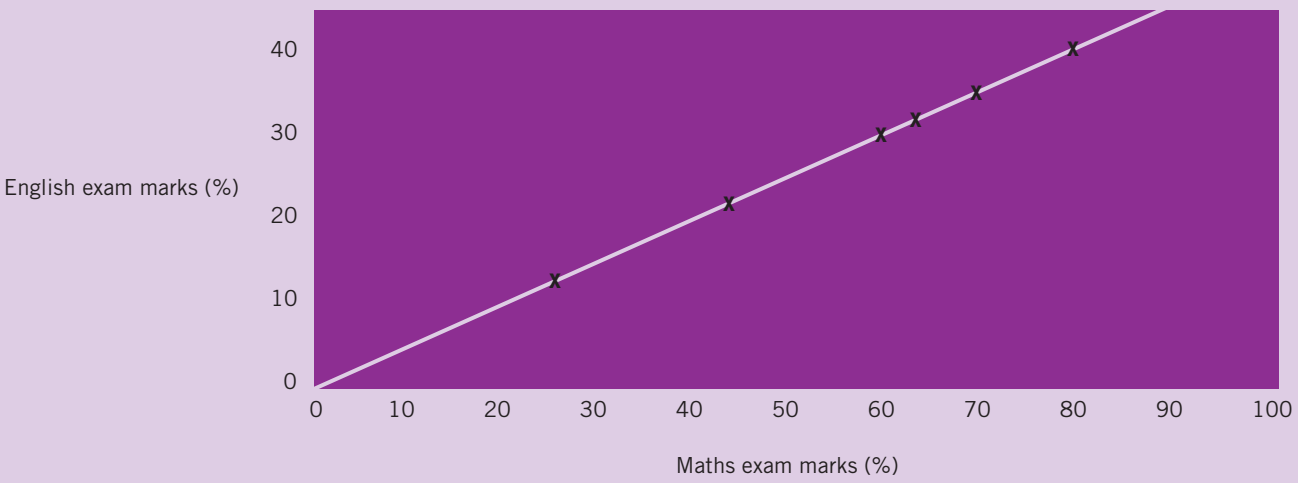
This is a plausible finding, given that performance in exams is an expression of academic ability. An able student should score relatively highly in both exams, while a weak student should score lower marks in both.

The r score is calculated using a formula that measures the range of dispersion of the number of points around a mean average value. Microsoft's Excel spreadsheet software has a Pearson function: if you arrange the two sets of figures in columns A and B, and then type "=pearson(a1:a6,b1:b6)" into a cell, the r score will appear there. When the r score is 1, it indicates a perfect positive correlation, which can be represented graphically in the diagram below. The six points have been plotted on the graph (known as a scatter diagram) and then joined by a straight line. It is good practice to draw a scatter diagram to ascertain whether or not there's a linear relationship between the two variables.

The correlation between the maths and English exam marks is perfect and this appears as a straight line on the graph. All of the six points observed in the data lie exactly on that line.

The fact that the two sets of figures correlate suggests a relationship or causal link between them, but says nothing about the amount of change in the first that corresponds to a given change in the second. To determine that, an exercise in regression analysis is required. The relationship between the English and maths marks can be represented by the regression equation E = ßM, where ß is known as the regression coefficient. In this >

Correlation of maths and English exam marks in case 1



case ß is evidently 0.5. Compare this equation with the equation for the linear regression of y on x, which is given in the list of formulas required for C03 as y = a + bx. Here the value of a is zero, which is because the y intercept is zero.

One obvious use of regression analysis is that it enables us to forecast what result a student should obtain in the English exam as soon as their result in the maths paper is known. If they score 68 per cent in maths, for example, we can forecast that they will achieve 34 per cent in English.

A correlation can also be negative. For example, if we alter the English exam results of our class of six, the following outcome might occur:

Case 2: perfect negative (linear) correlation

Student	Maths mark	English mark
1	80%	20%
2	60%	40%
3	44%	56%
4	26%	74%
5	70%	30%
6	64%	36%

Pearson coefficient: -1.000

An r score of -1 means that all of the points will again lie on a straight line when plotted on a graph, but this time the line has a negative gradient. We can still use this knowledge to make forecasts. If a student obtains 68 per cent in maths, we can predict that they will score 32 per cent in the English exam. This is an inherently implausible situation, but it serves to illustrate the point.

The r score must lie between 1 and -1. A high degree of correlation can be either positive or negative (i.e. close to 1 or -1). The following set of exam results demonstrate a high, although not perfect, degree of correlation:

Case 3: high positive correlation

Student	Maths mark	English mark
1	80%	53%
2	60%	41%
3	44%	28%
4	26%	18%
5	70%	48%
6	64%	41%

Pearson coefficient: 0.995

E ÷ M average (ß): 0.666

In this case the English marks average 0.666 of the maths marks, but there are small variations around that average in individual cases. Despite this, the degree of correlation may be considered high enough for forecasting purposes. If a student achieves a mark of 68 per cent in maths, say, we can predict that they will score 45 per cent in the English exam.

In this case we are adopting a ß of 0.666. That figure can be obtained with varying degrees of mathematical refinement. It can be derived from a simple inspection of the data or by plotting six observations on a graph and drawing the line of nearest fit among them. Alternatively, more sophisticated models may be used – calculating the least-squares regression line for example.

Clearly, a high (positive or negative) r score for a pair of variables gives us more assurance that the correlation between them is meaningful and no mere coincidence. Also, a higher number of observations (N) in the data will, all other things being equal, give a higher level of assurance that the correlation is significant. Note that, although it has been useful for illustrative purposes, the number of observations in the cases I have cited (N = 6) is rather small.

Mathematical models have been used to develop tables of significance, which give the probability of significance at different correlation levels and with different numbers of observations. What is deemed an acceptable level of assurance that a correlation is significant depends on the circumstances of each case.

The fact that two sets of data seem to correlate does not automatically mean that they relate to two linked variables. The correlation could have no significance and may be pure coincidence, or it may be the result of some third factor. For instance, there may be a high positive correlation between the number of accidents in public swimming pools (variable A) and sales of ice cream (variable B). Does this mean that A causes B or vice versa? Of course not. There is most likely to be a third factor (known as the confounding variable): the ambient daytime temperature that causes A and B to change together. The relationship between A and B in such a case is sometimes described as a spurious correlation.

Correlation and regression analysis has numerous possible business applications. Say, for example, we are trying to forecast the number of widgets a store will sell in the coming year. One possible approach to this exercise is to consider its sales figures over the previous decade and check for correlations with those factors that might influence the sale of widgets. Such factors might include the average summer temperature, the number of local residents aged 15 to 17, the number of new homes built in the area and the number of divorces.

Getty Images

Let's say that we find a significant positive correlation between widget sales and the number of local people aged 15 to 17 and no significant correlation with any of the other factors. Such a finding would be plausible if the widget was a youth-oriented product. So, if it is known that the number of people in that age group will be 8 per cent greater in the coming year than it was in the previous one, then we would forecast widget sales to rise by a similar proportion. This logic incorporates the assumption that correlation is a proxy for causality – i.e. if A correlates with B, then A must cause B.

The idea that causality follows directly from correlation underpins a great deal of modern research in both business and social sciences. For example, the public debate about cannabis use has been influenced by scientific research that makes observations such as: "The first thing to know about this topic is that it is indisputable that there is a correlation between the repeated use of cannabis and a variety of mental health issues."[1]

The more often that someone uses cannabis, the more likely they are to suffer from mental health problems. The obvious inference is that the heavy use of cannabis causes mental health problems. Anti-drugs campaigners have often cited this observation as a justification for imposing stronger legal controls on cannabis use. But is their inference correct? Perhaps the direction of causation is the other way around: maybe people with incipient mental health problems use cannabis more readily as a form of self-medication. If this alternative theory is correct, then the case for prohibition is much less clear.

The general point is that correlation should not imply causality. Factors A and B may correlate with one another, but we should not assume that A necessarily causes B. This is because the following three possibilities exist:

- B may cause A.
- Some third factor may cause A and B to vary together.
- The correlation observed between A and B may be mere coincidence.

Nevertheless, correlation analysis remains a significant tool in business research. For example, one recent study was described by its authors as follows: "We develop a model using financial data for 311 publicly listed retail firms for the years 1987 to 2000 to investigate the correlation of inventory turnover with gross margin, capital intensity and sales surprise (the ratio of actual sales to expected sales for the year). The model explains 66.7 per cent of the within-firm variation and 97.2 per cent of the total variation (across and within firms) in inventory turnover."[2]

The fact that two variables correlate may be a significant observation that offers some insight into a situation. But further research is required before it can be said that movements in one variable cause or explain movements in another.

Bob Scarlett is a management accountant and consultant.

References and further reading

1. "Cannabis and psychosis: a guide to current research about cannabis and mental health", *Erowid Extracts*, June 2005 (bit.ly/ResearchCannabis).

2. "An econometric analysis of inventory turnover performance in retail services", *Management Science*, February 2005 (bit.ly/ManSciInvTurn).

CIMA Official Study Text – C03 Fundamentals of Business Mathematics, CIMA Publishing, 2012.