Correlation and regression are related but distinct concepts that are frequently confused and misinterpreted in the cardiological literature. To better understand these concepts, consider an example to exemplify how correlation is often used in the literature. Suppose we wish to correlate body mass against height in men. A scatter plot of data obtained from a sample of men is shown in figure 1, a line of ‘best fit’ is also often drawn. We can see that there is an impression that heavier men tend to be taller, but this is by no means consistent. The correlation coefficient, $r$, for these data is 0.35 or 35% with $p<0.0004$. It is then often concluded that height and mass are ‘significantly’ or ‘strongly’ correlated, but what does this mean?

## Correlation

Correlation is a measure of the association between two variables. A positive correlation means that as one variable increases so does the other. A negative correlation means that as one variable increases the other decreases. The correlation coefficient takes values from $+1$ (perfect positive correlation) to $-1$ (perfect negative correlation), a value of zero indicates no correlation. The correlation most frequently used in the literature is the Pearson correlation. This tests the strength of linear association between two variables. Thus data that followed an exponential pattern would have a Pearson correlation coefficient less than one (possibly much less), although there is a perfect association. Sometimes a transformation is undertaken to render the relationship linear. For example for an exponential relationship, a logarithmic transformation will result in a linear relationship. This, however, is unlikely to be of value in cardiological applications. Consequently, Pearson correlation should only be used when the question to be addressed is ‘What is the strength of linear association?’ In addition both variables should be random variables that are approximately normally distributed (a requirement that is rarely commented upon) and any one individual can only contribute one value for each variable. Another important point is the presence of outliers (data points separate from the main body of data on a scatter plot), which can have a disproportionate influence on the value of $r$.

The $p$ value quoted with the value of $r$ arises from the null hypothesis that $r=0$ ie no correlation. It is important to note that the $p$-value depends on both the strength of the correlation and the size of the data set. Thus for very large data sets values of $r$ close to zero may still be significant (ie have a small $p$-value), but be of no clinical relevance. The $p$-value indicates the confidence we can have that $r$ is not zero, it is not a measure of the strength of a correlation. Correlation gives no indication of how one variable may depend on another.

A more general measure of association is the Spearman rank correlation, $r_s$. This does not require that the data are normally distributed or even continuous. For example ordinal data (eg NYHA class) could be used. In addition there is no assumption of linear association between the variables. Thus for data that are exponentially related $r_s = 1$. Whichever method is used, it does not matter in which order the variables are correlated, the resultant $r$ will be the same. It can be helpful to calculate $r^2$, which gives the proportion of the variability explained by the association. For example for $r=0.8$, 64% of the variability is explained by the association. Generally, reporting correlations does not greatly enhance an understanding of data. In the example given in figure 1 there is actually a rather weak correlation (only 12% of the variability of body mass explained by height). It is important not to conflate a significant statistical test for the existence of a correlation with the degree of correlation.

Using correlation to assess agreement is another area where correlation has been misused. For example, suppose there is a new point of care test for BNP and we wish to assess how well this agrees with the standard laboratory result. It is inappropriate to correlate the two measurements and conclude there is good agreement if $r$ is sufficiently close to one. Two variables can be highly correlated but not agree. For example the equation $y=2x$ will give a perfect correlation ($r=1$), but hopeless agreement (all the $y$ values will be twice the $x$ values). Other methods should be used to assess agreement.

## Regression

Regression is generally a more useful approach than correlation. The principle is to use a data set containing a number of independent variables to develop an equation to estimate a dependent variable. The equation can then be used to estimate mean values of the dependent variable from values of the independent variables. The coefficient of determination $R^2$ associated with the equation tells us how successful the regression has been. Specifically $R^2$ gives the proportion of the variability of the dependent variable that is explained by the equation. A $p$-value should be given for the coefficient of each of the independent variables. The $p$-value arises from the null hypothesis that the coefficient of the independent variable is zero. Thus for small $p$-values we can have confidence.

![Figure: Scatter plot of body mass (Kg) against height (m) for healthy men. The line of ‘best fit’ is also shown. This is the regression line; Mass = -87.1 +97.7*height.](image-url)
that a particular independent variable contributes to the variability of the dependent variable. The only assumptions are that the residuals (the difference between the actual value of the dependent variable and that predicted) should be normally distributed and have constant variance. This is rarely commented upon.

To demonstrate how this works in practice we return to the example given in figure 1. Rather than performing a correlation we now develop a regression equation with mass as the dependent variable and height as an independent variable.

\[
\text{Mass} = -87.1 + 97.7 \times \text{height}
\]

\[R^2 = 0.122\%\] (note that this is the square of \(r=0.35\), the correlation coefficient) and \(p<0.0004\). Thus height only contributes 12.2\% to the variability of body mass. There are clearly other important variables explaining body mass. We therefore include other variables for which data are available, for example waist circumference (cm):

\[
\text{Mass} = -117 + 57.9 \times \text{height} + 1.08 \times \text{waist}
\]

Now \(R^2 = 91\%\) and both coefficients have \(p<0.0004\). The equation now explains 91\% of the variability of body mass. In fact using waist circumference alone explains 87\% of the variability of body mass. The software packages that perform these calculations also allow the assumptions (normally distributed residuals with constant variance) necessary for the process to be valid to be checked. In this case they are confirmed. In practice the choice of independent variables is to some extend a matter of trial and error (large numbers can be assessed automatically by software packages), but only variables that are biologically plausible should be considered. In the above example total cholesterol was tested, but found not to be significant.

The above example is unlikely to be of any clinical value, but is given as a demonstration of regression. A practical application of the use of a regression equation (or model) is the estimation of the mean aortic root diameter (AD), for given individual characteristics, for healthy subjects. The important independent variables have been found to be: age, body surface area (BSA \((m^2)\)) and gender.

\[
\text{AD} = 1.52 + 0.09 \times \text{age} + 0.46 \times \text{BSA} - 0.247 \times \text{gender}
\]

Where gender takes the value 1 for men and 2 for women. We can see from this equation that for men and women of the same age and BSA, men will have a mean aortic root diameter 0.25cm greater than that for women. This equation is used in the assessment of patients with suspected Marfan syndrome. This example shows how binary variables (gender) and non-linear variables (BSA - depends on the square of height) can be included as independent variables. When using regression equations it is important not to use values for the independent variables outwith the range used in the construction of the model. For example in the above example the equation should not be used to estimate the aortic root diameter in children, because the model was developed from data obtained from adults, with age and BSA greater than that of children.

Correspondence to:
Professor Andrew Owen
Institute of Medical Sciences
Canterbury Christchurch University.
Rowan Williams Court, 30 Pembroke Court, Chatham Maritime, Kent ME4 4UF
E-mail: Andrewowen@canterbury.ac.uk

References