*Archives of Disease in Childhood* 1992; **67**: 1306–1309

STATISTICS FROM THE INSIDE

# 7. Regression and correlation

M J R Healy

Some of the most important statistical analyses are those which deal with the *relationship* between a variate $y$ and another quanitity $x$ which is recorded on the same items. We may for example wish to relate the response to a drug ($y$) to the dose administered ($x$); or the head circumference of a baby ($y$) to its weight ($x$). A possible simple model for a relationship of this kind is to suppose that we can write

$$y=\alpha+\beta x+\varepsilon \qquad (1)$$

where the $\varepsilon$'s are random quantities with mean zero.

To explore the implications of this, I shall need some notation. Suppose we consider all the items in the population which have a certain specific value of $x$, $x=x_0$ say. In our examples, these would be all the responses to a particular dose, or all the babies with a particular weight. These selected items constitute a population in their own right, a *subpopulation* as it is called. This subpopulation will have a mean value and this in general will depend upon which value of $x$ we have chosen (mean response will depend upon dose, mean head circumference will depend upon weight).

A mathematician's term for a mean is 'expectation' or 'expected value' and the corresponding notation $E(y)$ is sometimes used to denote the population mean of $y$. Here for the mean of the subpopulation at $x=x_0$ I shall write $E(y|x=x_0)$ where the vertical bar should be read as 'given that', or in technical language 'conditional upon'. More simply, I can write $E(y|x_0)$. This mean value of a particular subpopulation is called a *conditional mean*, and in principle I can obtain the conditional mean of $y$ for any given value of $x$.

Suppose now that I calculate the conditional means of $y$ for a whole set of values of $x$ and that I plot these means against their $x$'s. Suppose too that when I do this the result is a straight line whose equation can be written as

$$E(y|x)=\alpha+\beta x \qquad (2)$$

This is just another way of writing the relationship (1) above. The relationship (2) is called a *regression equation*, and $\alpha$ and $\beta$ are the *regression coefficients*. The coefficient $\alpha$ is the *intercept*, that is the mean of the subpopulation at $x=0$, and the coefficient $\beta$ is the *slope*, the amount by which the mean of $y$ increases for a unit increase in $x$.

The regression equation (2) is only a partial description of the population of ($x$, $y$) pairs; it tells us the mean of the subpopulation at any particular value of $x$, but we still need to specify the variability and other properties. As well as a mean, each subpopulation will have a standard deviation and in the simplest case we assume that this standard deviation is constant and does not depend upon $x$. This standard deviation can be written as $\sigma_{y|x}$ to distinguish it from the *unconditional* standard deviation $\sigma_y$ which applies to the whole population of $y$'s ignoring the $x$'s. It is sometimes useful in addition to assume that each of the subpopulations has a Normal distribution.

With these assumptions, the usual textbook formulas or computer programs can be used to obtain estimates $a$ and $b$ of $\alpha$ and $\beta$ from sample data consisting of ($x$,$y$) pairs, and also an estimate $s_{y|x}$ of $\sigma_{y|x}$, the conditional standard deviation which measures the scatter of the y values around the regression line. From this, the standard errors of $a$ and $b$ can be obtained. Simple $t$ tests can then be use to test hypotheses about $\alpha$ and $\beta$ and to provide confidence intervals for them. For a reason to be clarified below, the degrees of freedom in these tests will be two less than the number of data pairs in the sample.

With regard to the $x$'s, it is useful to distinguish two situations.

(1) Consider the example in which children are treated with one out of (say) four doses of a drug. What we have is a set of four subpopulations, and each subpopulation carries an $x$ value which is a quantitative label—here, the dose. In this type of situtation, the $x$'s are *not* random variables; if we think about drawing further samples of data, these will come from the same subpopulations with the same values of $x$. The $x$'s are better described as values of a *variable* rather than of a variate.

(2) Now consider the other example, in which the head circumference of a baby ($y$) is regressed upon its weight ($x$). Now we have a *bivariate* population whose items are the babies and which has two variates recorded upon each item, head circumference and weight. Provided that our assumptions about the $y$'s hold good, we can use exactly the same model for the regression of head circumference upon weight, embodied in equation (2), and we can estimate the parameters of the model using exactly the same methods. However, a new feature emerges. It is now quite reasonable to investigate another regression equation,

$$E(x|y)=\gamma+\delta y \qquad (3)$$

This is the regression of weight upon head circumference, which relates the conditional means of $x$ to values of $y$. It is important to

**23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD**

Correspondence to: Professor Healy.

No reprints available.

realise that this represents a quite different line from that given by equation (2). If we wish to know the average head circumference of a baby of a certain weight, then equation (2) with the coefficients $\alpha$ and $\beta$ estimated from a sample will enable us to do so; if (for some reason) we wish to know the average weight of babies with a particular head circumference, the equation (3) must be used.

Suppose that we have a sample of $(x, y)$ pairs and that from them we obtain the estimated version of equation (2),

$$\hat{y}=a+bx \qquad (3)$$

where $\hat{y}$ denotes the estimate of the conditional mean $E(y|x)$. For each actual pair of observations $(x,y)$ we can calculate the residual $(y-\hat{y})$, the discrepancy between the observed value of $y$ and the estimate of its mean. The estimate of the variance about the regression line, $s^2_{y|x}$, is found by forming the sum of squares of the residuals and dividing this by its degrees of freedom. It can be shown that there are two exact relationships between the residuals—the sum of the residuals is exactly zero, and so is the sum of the residuals each multiplied by its value of $x$. This means that, if you are told all but two of the residuals (and all the $x$'s), you can calculate the last two without looking at the data. This is why the degrees of freedom are two less than the sample size. The true conditional standard deviation $\sigma_{y|x}$ is just the standard deviation of the true residuals $(y-E(y|x))$, and it and its estimate are often referred to as the *residual standard deviation*.

Calculating and using the residual standard deviation carries with it the assumption that the residuals constitute a random sample and exhibit no systematic features. Checking this assumption should be an important part of any regression investigation. The most useful checks are graphical. For example, a plot of the residuals against the values of $x$ should show no exceptional features, such as outlying values or curvilinear trends.

A presentational point arises with the estimated regression equation (3) above. In this form, the intercept $\alpha$ represents the mean of the subpopulation of $y$'s at $x=0$. If $x$ is a variable such as weight or stature this may be a quite meaningless quantity far outside the range of the actual data. As a result, the estimate $a$ will be meaningless too, and because of the extrapolation it will have a very large standard error. It is preferable to express the equation in the form

$$\hat{y}=a+b\,(x-X) \qquad (4)$$

where $X$ is a convenient round number somewhere near the centre of the data values. Note that this does not affect the value of the slope.

A situation that requires a good deal of care is that in which there are two levels of variation (see note 6 of this series), such as when repeated measurements of both $x$ and $y$ are made upon a number of subjects. Data of this kind are common when the investigation relates to the growth of children or the time course of the response to a drug. It will then be possible to calculate a regression line for each subject separately, describing the relationship of $y$ to $x$ within the subject, and an average of the within-subject slopes can be obtained. It will also be possible to calculate the means of $x$ and $y$ for each of the subjects and to estimate a between-subject regression based upon these means. It is important to realise that the within-subject and between-subject slopes will generally be quite different—they may not even have the same sign. It is possible, for example, to measure some quantity on a number of newborn babies and to calculate the regression of this quantity (as $y$) on gestational age (as $x$). But it is quite unsafe to assume that this between-subject regression can serve to describe what happens to a single fetus in utero as its gestational age increases.

Regression can often be a useful method of allowing for a factor whose effect upon the main variate of interest cannot be tightly controlled. As a simple example, suppose we wish to compare the head circumferences of male newborns with those of females. There is no great difficulty in obtaining samples from the two populations and the means can be compared by way of an unpaired $t$ calculation. However, head circumference at birth is related to gestational age and it is unlikely that the two samples have exactly the same mean gestational age. Any difference that we observed between the two mean head circumferences may thus be in part a reflection of a difference in mean gestational age. Suppose then that we calculate the regression of head circumference (as $y$) on gestational age (as $x$) for each of the two groups. The two lines permit us to read off the mean head circumferences at a particular gestational age (perhaps 40 weeks), and these two conditional means can be compared by a modification of the usual $t$ procedure. This technique has two advantages. Not only does it compensate for any imbalance between the gestational ages of the two samples, it also increases the precision of the comparison by utilising the standard deviations about the regression lines which will be smaller than those of the head circumferences considered in isolation. Note that the technique is most useful when the two lines are parallel; in this case the difference between the two conditional means will not depend upon the particular

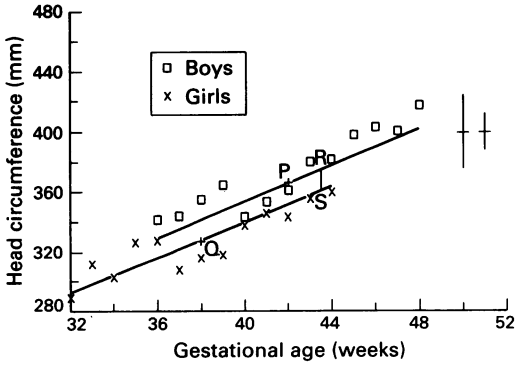


*Figure 1 Taken in isolation, the mean difference in head circumference between girls and boys is the vertical distance between the means P and Q. After adjustment for gestational age, the difference is that between the parallel regression lines, RS. The bars at the right of the diagram are twice the unconditional and conditional standard deviations.*

$x$ value that has been chosen. The parallelism of the regression lines can itself be tested by a straightforward $t$ calculation using the standard errors of the two estimated slopes. This method is sometimes called an *analysis of covariance*, the quantity $x$ being referred to as a *covariate*. I have illustrated it in fig 1, where I have for clarity exaggerated the difference between the $x$ values in the two groups.

Closely linked to the idea of regression is that of *correlation*. Suppose that both $x$ and $y$ are random variables and let $\sigma_y^2$ be the variance of $y$ ignoring the values of $x$ and $\sigma_{y|x}^2$ the conditional variance of the $y$'s at a particular $x$ value. Then the *correlation coefficient* $\varrho$ is defined by the equation

$$\varrho^2 = 1 - \frac{\sigma_{y|x}^2}{\sigma_y^2} \qquad (5)$$

with $\varrho$ taking the same sign as the regression slope. This quantity is sometimes called the *Pearson* or *product-moment* correlation, to distinguish it from other similar quantities. There appears to be no good historical reason for adding the name of Bravais, as is sometimes done.

It can be seen that $\varrho^2$ is a measure of the information about $y$ that is provided by a knowledge of $x$. Our uncertainty about a future value of $y$, considered in isolation, can be measured by the standard deviation $\sigma_y$ or its square, the variance $\sigma_y^2$. If we are then told the corresponding value of $x$, the variance is reduced to $\sigma_{y|x}^2$, so that the percentage reduction in variance due to the knowledge of the value of $x$ is just $100\varrho^2$. Incidentally, this shows that $\varrho^2$ lies between 0 and 1 so that $\varrho$ must lie in the range $-1$ to $+1$.

A point which is not made clear by this definition is that the correlation coefficient is symmetric as between $y$ and $x$. This distinguishes it sharply from the regression coefficient and shows that correlation is only relevant when we are dealing with a bivariate population so that both $x$ and $y$ are random variables. We speak of the regression *of y on x* (or *of x on y*), but of the correlation *between x and y*.

The correlation coefficient is often described as a measure of the *association* between $x$ and $y$, and this is true in the sense described above. It has to be stressed that it may be a rather misleading measure. Consider for example a correlation of 0·5—this is the correlation between the heights of fathers and their adult sons and sounds quite a high value. Yet knowledge of a father's height reduces the variance associated with his son's height by only $100\times0{\cdot}5^2=25\%$, and this translates into a reduction of no more than 13% in the standard deviation. In the same vein, the correlation coefficient vividly illustrates the difference between statistical significance and practical importance. With a sample of 50 $(x,y)$ pairs, an estimated correlation of 0·35 is highly significantly differerent from zero, with $p<0{\cdot}01$; but a plot of some actual data points (fig 2) shows that this degree of association between the two variates is so weak as to be of little practical value in most circumstances.

It can be shown quite easily that the correlation coefficient, whether of the sample or the population, is equal to the geometric mean of the two regression slopes—in symbols, $\varrho=\sqrt{\beta_{y|x}.\beta_{x|y}}$. It follows that the correlation is equal to zero only if the regression coefficients are equal to zero. The test of significance of the correlation coefficient against zero is thus exactly the same $t$ test as would be used for either of the regression slopes, though as usual confidence limits are likely to be much more interesting than the result of the test in both instances.

The method of calculating confidence limits for a correlation coefficient, which can also be adapted to compare the values of two correlations, is of some general interest. A formula exists for the standard error of a sample correlation, but this is not useful for two reasons—the formula involves the unknown correlation, and in addition the distribution of the sample coefficient is liable to be far from Normal. This latter finding follows from the fact that the sample coefficient cannot go outside the range $-1$ to $+1$; thus if the true value of the correlation is (say) 0·8, the sample value can be considerably smaller than this but not very much larger (fig 3 shows the distribution for a true value of 0·8 and a sample of 20 pairs).

Both these difficulties can be avoided by *transforming* the sample coefficient (r, say) to a different mathematical form. The procedure involves two steps:

(1) Calculate $p=\frac{1}{2}(1+r)$. This quantity lies in the range 0 to 1.

(2) Calculate $z=\log\{p/(1-p)\}$ using 'natural' logarithms to the base e. This quantity is known as the *logit* of p and is unlimited in both directions.

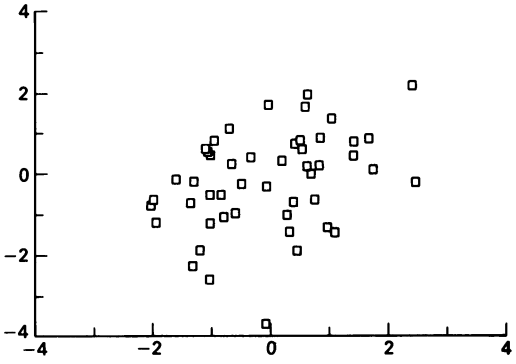


*Figure 2 Data points with a correlation of 0·35 ($p<0{\cdot}01$).*

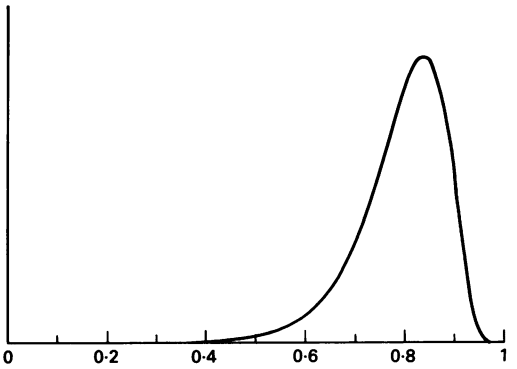


*Figure 3 The distribution of the sample correlation coefficient, true value $\varrho=0{\cdot}8$, sample size 20.*

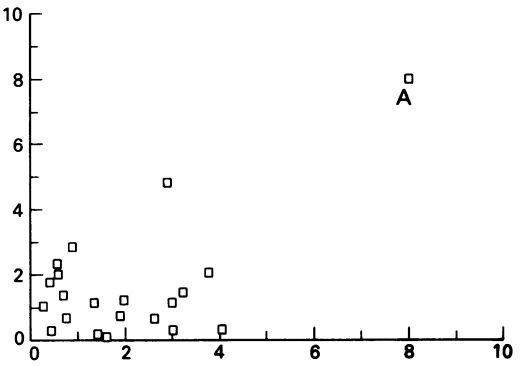


*Figure 4 A sample with one outlying point.*

It turns out that z is Normally distributed to a close approximation, with a standard deviation given by $\sqrt{4/(n-3)}$ where n is the sample size. Suppose then that we find a sample correlation of $r=+0{\cdot}8$ in a sample of 20 pairs. Then $p=0{\cdot}9$, $z=2{\cdot}197$ with a standard error of $\sqrt{4/17}=0{\cdot}485$. The 95% confidence limits for z are given by $2{\cdot}197\pm1{\cdot}96\times0{\cdot}485=1{\cdot}246$ to $3{\cdot}148$. These limits can be converted back to the correlation scale by calculating successively $p=e^z/(1+e^z)$ and $r=2p-1$. Corresponding to the limits $1{\cdot}246$ and $3{\cdot}148$, the values of p are $0{\cdot}777$ and $0{\cdot}959$, and the confidence limits for $\rho$ are thus $0{\cdot}553$ to $0{\cdot}918$. This trick of transforming quantities to make them better behaved statistically is useful in many contexts, and I hope to return to it in a subsequent article.

For all its prominence in the textbooks, there are in fact rather few situations in clinical and laboratory medicine in which the correlation coefficient is a useful statistic. It is often quoted alongside the equation of a regression line, probably because it has been automatically provided by a computer program; but it is far more informative to provide the standard error of the slope and also the residual standard deviation which measures the variability of the data points about the fitted line. When $x$ is an ordinary variable (such as the dose of a drug in a dose-response study) rather than a random variate, the correlation coefficient is usually not a meaningful quantity. It should also be noted that the correlation coefficient can be heavily influenced by the presence in a sample of one or two outlying points. Figure 4 illustrates this. As it stands, the sample there has a correlation of $0{\cdot}61$; but if the single point marked A is removed, this falls to no more than $0{\cdot}02$. As with most statistical analyses, a plot of the data is an almost essential part of the interpretive process.

One situation in which the correlation coefficient has no place whatsoever, and in which regression must be used with caution, is that in which two methods of measuring the same quantity are to be compared. The null hypothesis $\rho=0$ is obviously silly; nobody would suppose that two methods purporting to measure the same thing could be completely unrelated. If one method is a gold standard and the other a new technique, it may be of some interest to form the regression of the gold standard (as $y$) on the new technique (as $x$). The standard deviation about the regression line then provides information about the range of 'true' values which are consistent with a particular reading obtained by the new technique. Note however that for technical reasons the slope of the regression is *not* expected to be equal to $1{\cdot}0$, but to be rather less than this. A full discussion of the method comparison problem would take us too far afield at this point; a useful article is that by D G Altman and J M Bland, *Lancet* 1986;i:307–17.