Health Sciences M.Sc. Programme

Applied Biostatistics

Week 8: Correlation and Regression

The correlation coefficient

Correlation coefficients are used to measure the strength of the relationship or association between two quantitative variables. For example, Table 1 shows height, muscle strength and age in 41 alcoholic men. We will begin with the relationship between height and strength. Figure 1 shows a plot of strength against height. This is a **scatter diagram**. Each point represents one subject. If we look at Figure 1, it is fairly easier to see that taller men tend to be stronger than shorter men, or, looking at the other way round, that stronger men tend to be taller than weaker men. It is only a tendency, the tallest man is not the strongest not is the shortest man the weakest. Correlation enables us to measure how close this association is.

The correlation coefficient is based on the products of differences from the mean of the two variables. That is, for each observation we subtract the mean, just as when calculating a standard deviation. We then multiply the deviations from the mean for the two variables for a subject together, and add them. We call this the **sum of products about the mean**. It is very like the sum of squares about the mean used for measuring variability.

To see how correlation works, we can draw two lines on the scatter diagram, a horizontal line through the mean strength and a vertical line through the mean height, as shown in Figure 2. Because large heights tend to go with large strength and small heights with small strength, there are more observations in the top right quadrant and the bottom left quadrant than there are in the top left and bottom right quadrants. In the top right quadrant, the deviations from the mean will be positive for both variables, because each is larger than its mean. If we multiply these together, the products will be positive. In the bottom left quadrant, the deviations from the mean will be negative for both variables, because each is smaller than its mean. If we multiply these two negative numbers together, the products will also be positive. In the top left quadrant, the deviations from the mean will be negative for height, because the heights are all less than the mean, and positive for strength, because strength is greater than its mean. The product of a negative and a positive number will be negative, so all these products will be negative. In the bottom right quadrant, the deviations from the mean will be positive for height, because the heights are all greater than the mean, and negative for strength, because the strengths are less than the mean. The product of a positive and a negative number will be negative, so all these products will be negative also. When we add the products for all subjects, the sum will be positive, because there are more positive products than negative ones. Further, subjects with very large values for both height and strength, or very small values for both, will have large positive products. So the stronger the relationship is, the bigger the sum of products will be. If the sum of products positive, we say that there is a **positive correlation** between the variables.

Table 1. Height, quadriceps muscle strength, and age in 41 male alcoholics (data of Hickish *et al.*, 1989)

Height (cm)	Quadriceps muscle strength (N)	Age (years)		Height (cm)	Quadriceps muscle strength (N)	Age (years)
155	196	55		172	147	32
159	196	62		173	441	39
159	216	53		173	343	28
160	392	32		173	441	40
160	98	58		173	294	53
161	387	39		175	304	27
162	270	47		175	404	28
162	216	61		175	402	34
166	466	24		175	392	53
167	294	50		175	196	37
167	491	35		176	368	51
168	137	65		177	441	49
168	343	41		177	368	48
168	74	65		177	412	32
170	304	55		178	392	49
171	294	47		178	540	41
172	294	31		178	417	42
172	343	38		178	324	55
172	147	31		179	270	32
172	319	39		180	368	34
172	466	53				

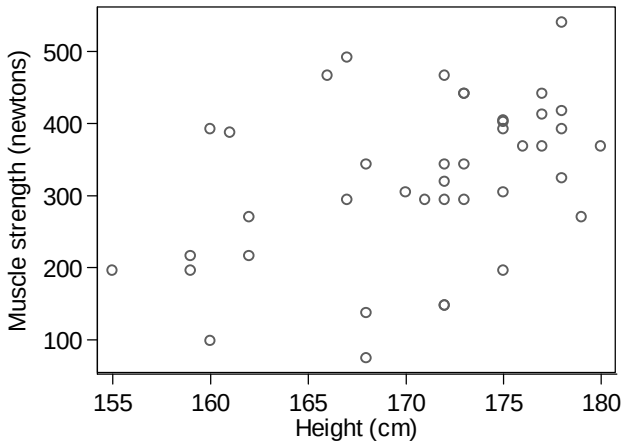


Figure 1. Scatter diagram showing muscle strength and height for 41 male alcoholics

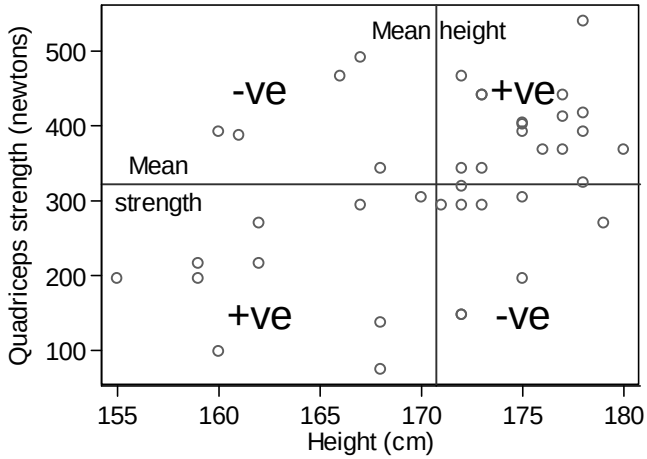


Figure 2. Scatter diagram showing muscle strength and height for 41 male alcoholics, with lines through the mean height and mean strength

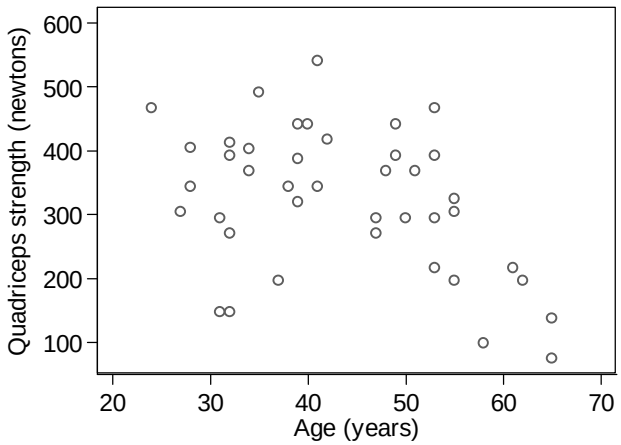


Figure 3. Scatter diagram showing muscle strength and age for 41 male alcoholics

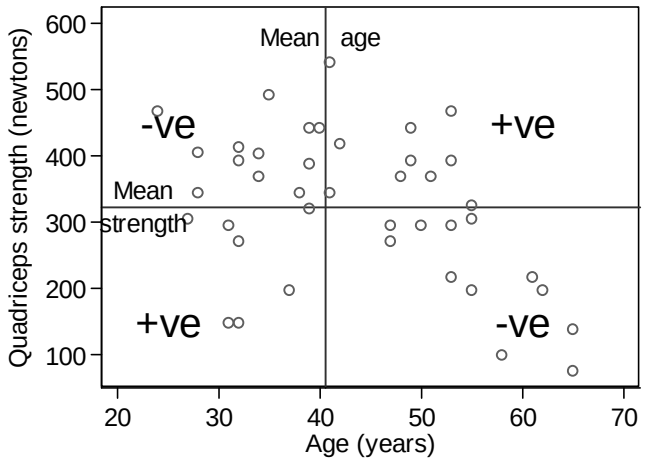


Figure 4. Scatter diagram showing muscle strength and age for 41 male alcoholics, with lines through the mean.

Figure 3 shows the relationship between strength and age in Table 1. Strength tends to be less for older men than for younger men. Figure 4 shows lines through the means, as in Figure 2. Now there are more observations in the top left and bottom right quadrants, where products are negative, than in the top left and bottom right quadrants, where products are positive. The sum of products will be negative. When large values of one variable are associated with small values of the other, we say we have **negative correlation**.

The sum of products will depend on the number of observations and the units in which they are measured. We can show that the maximum possible value it can have is the square root of the sum of squares for height multiplied by the square root of the sum of squares for strength. Hence we divide the sum of products by the square roots of the two sums of squares. This gives the **correlation coefficient**, usually denoted by *r*.

Using the abbreviation '*r*' looks very odd. Why '*r*' and not '*c*' for correlation? This is for historical reasons and it is so ingrained in statistical practice that we are stuck with it. If you see an unexplained '*r* =' in a paper, it means the correlation coefficient. Originally, 'r' stood for 'regression'.

Because of the way *r* is calculated, its maximum value = 1.00 and its minimum value = –1.00. We shall look at what these mean later.

The correlation coefficient is also known as **Pearson's correlation coefficient** and the **product moment correlation coefficient**. There are other correlation coefficients as well, such as Spearman's and Kendall's, but if it is described simply as 'the correlation coefficient' or just 'the correlation', the one based on the sum of products about the mean is the one intended.

For the example of muscle strength and height in 41 alcoholic men, $r = 0.42$. This a positive correlation of fairly low strength. For strength and age, $r = –0.42$. This is a negative correlation of fairly low strength.

Figure 5 shows the correlations between several simulated variables. Each pair of variables was generated to have the correlation shown above it. The first panel in Figure 5 shows a perfect correlation. The points lie exactly on a straight line and we could calculate Y exactly from X. In fact, Y = X; they could not be more closely related. $r = +1.00$ when large values of one variable are associated with large values of the other and the points lie exactly on a straight line. The second panel shows a strong but not perfect positive relationship. The third panel also shows a positive relationship, but less strong. The size of the correlation coefficient clearly reflects the degree of closeness on the scatter diagram. The correlation coefficient is positive when large values of one variable are associated with large values of the other.

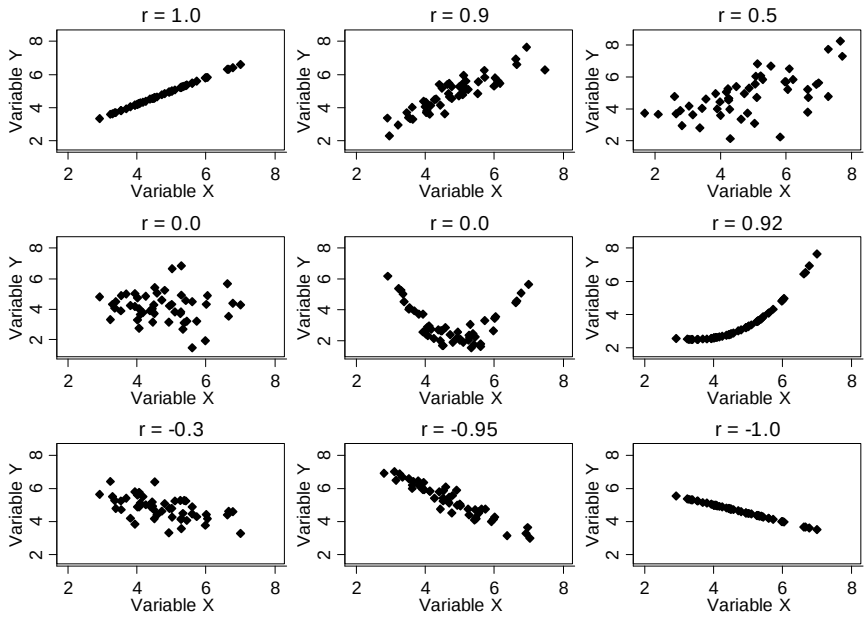


Figure 5. Simulated data from populations with different relationships between the two variables, and the population correlation coefficient

The fourth panel In Figure 5 shows what happens when there is no relationship at all, $r = 0.00$. This is the not only way r can be equal to zero, however. The fifth panel shows data where there is a relationship, because large values of Y are associated with small values of X and with large values of X, whereas small values of Y are associated with values of X in the middle of the range. The products about the mean will be positive in the upper left and upper right quadrants and negative in the lower left and lower right quadrants, giving a sum which is zero. It is possible for r to be equal to 0.00 when there is a relationship which is not linear. A correlation $r = 0.00$ means that there is no ***linear*** relationship, i.e. that there is no relationship where large values of one variable are consistently associated either with large or with small values of the other, but not both. The sixth panel shows another perfect relationship, but not a straight line. The correlation coefficient is less than 1.00. r will not equal –1.00 or +1.00 when there is a perfect relationship unless the points lie on a straight line. Correlation measures closeness to a ***linear*** relationship, not to any perfect relationship.

The correlation coefficient is negative when large values of one variable are associated with small values of the other. The seventh panel in Figure 5 shows a rather weak negative relationship, the eighth a strong one, and the ninth panel a perfect negative relationship. $r = –1.00$ when large values of one variable are associated with small values of the other and the points lie on a straight line.

Test of significance and confidence interval for *r*

We can test the null hypothesis that the correlation coefficient in the population is zero. This is done by a simple t test. The distribution of r if the null hypothesis is true, i.e. in the absence of any relationship in the population, depends only on the number of observations. This is often described in the terms of the degrees of freedom for the t test, which is the number of observations minus 2. Because of this,

it is possible to tabulate the critical value for the test for different sample sizes. Bland (2000) gives a table.

For the test of significance to be valid, we must assume that:

- at least one of the variables is from a Normal distribution,
- the observations are independent.

Large deviations from the assumptions make the P value for this test very unreliable.

For the muscle strength and height data of Figure 1, $r = 0.42$, P = 0.006. Computer programmes almost always print this when they calculate a correlation coefficient. As a result you will rarely see a correlation coefficient reported without it, even when the null hypothesis that the correlation in the population is equal to zero is absurd.

We can find a confidence interval for the correlation coefficient in the population, too. The distribution of the sample correlation coefficient when the null hypothesis is not true, i.e. when there is a relationship, is very awkward. It does not become approximately Normal until the sample size is in the thousands. We use a very clever but rather intimidating mathematical function called **Fisher's z transformation**. This produces a very close approximation to a Normal distribution with a fairly simple expression for its mean and variance (see Bland 2000 if you really want to know). This can be used to calculate a 95% confidence interval on the transformed scale, which can then be transformed back to the correlation coefficient scale. For the strength and height data, $r = 0.42$, and the approximate 95% confidence interval is 0.13 to 0.64. As is usual for confidence intervals which are back transformed, this is not symmetrical about the point estimate, r.

For Fisher's z transformation to be valid, we must make a much stronger assumption about the distributions than for the test of significance. We must assume that ***both*** of the variables are from Normal distributions. Large deviations from this assumption can make the confidence interval very unreliable.

The use of Fisher's z is tricky without a computer, approximate, and requires a strong assumption. Computer programs rarely print this confidence interval and so you rarely see it, which is a pity.

Regression

Regression is the rather strange name for a set of statistical methods which we use to predict on variable from another. Inspection of Figure 1 suggests that muscle strength increases with height. Can we estimate the mean strength for men with a given height? Or estimate the strength of an individual man from his height? Regression analysis seeks to do this.

We usually represent the relationship by a line on the scatter diagram. The simplest line is a straight one, but more complicated relationships can be examined. As we are predicting one variable from the other, we must choose which we want to predict. In this case, we will predict strength from height. Strength is the **outcome**, **dependent**, **y**, or **left hand side** variable. Height is the **predictor**, **explanatory**, **independent**, **x**, or **right hand side** variable. All these names are used.

A straight line or **linear** relationship takes the form:

$$\text{strength} = \text{intercept} + \text{slope} \times \text{height}$$

Figure 6. Muscle strength against height for 41 alcoholic men, showing several possible prediction lines

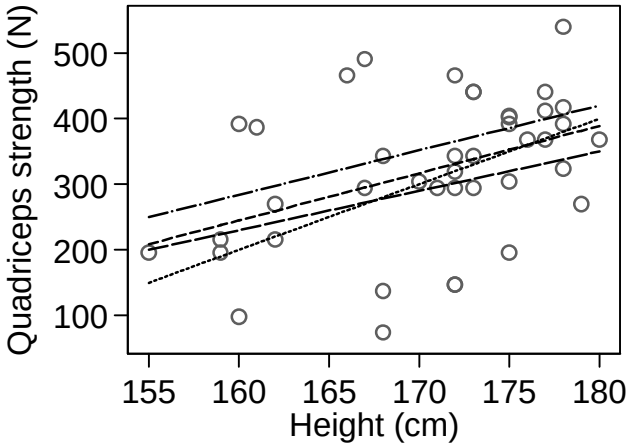


Figure 7. Deviations from the line in the direction of the outcome variable

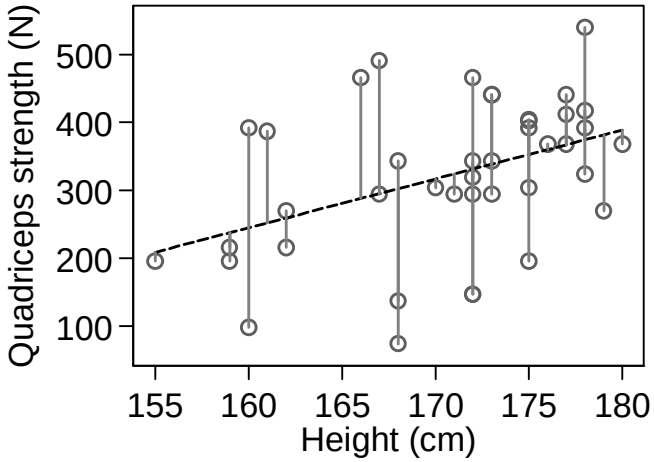
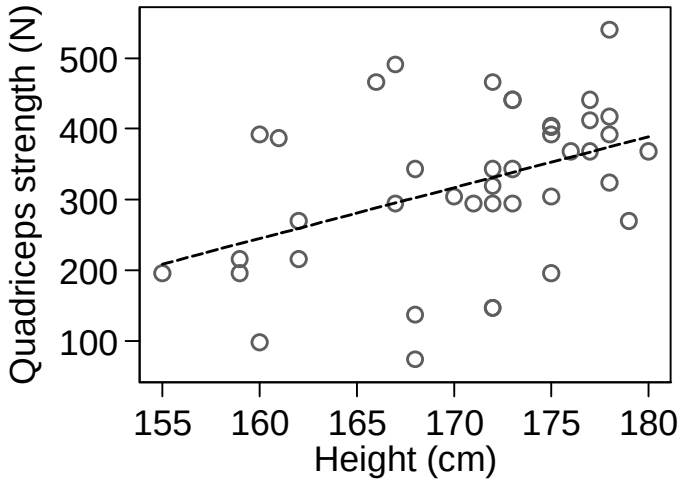


Figure 8. The regression of muscle strength on height



The **intercept** is the value strength would have for a person of zero height. The **slope** or **gradient** is the difference in strength associated with a difference of one unit of height. The slope is amount that the mean strength would differ between men whose height differed by one cm.

Strength will not be predicted exactly from height. There will be other factors which we don't know about. We call the other variation in the outcome variable **error**, or wandering, and our regression model of the data is

$$\text{strength} = \text{intercept} + \text{slope} \times \text{height} + \text{error}$$

The method of least squares

If the points all lay along a line and there is no random variation, it would be easy to draw a line on the scatter diagram. In Figure 1 this is not the case. There are many possible values of intercept and slope which could represent the data (Figure 6) and we need a criterion for choosing the best line. Figure 7 shows the deviation of a point from the line, the distance from the point to the line in the direction of the outcome variable. The line will fit the data well if the deviations from it are small, and will fit badly if they are large. These deviations represent the error, that part of the strength not explained by height. One solution to the problem of finding the best line is to choose that which leaves the minimum amount of the variability of strength unexplained, by making the variance of the error a minimum. This will be achieved by making the sum of squares of the deviations about the line a minimum. This is called the **method of least squares** and the line found is the **least squares line**.

The method of least squares is the best method if the deviations from the line follow a Normal distribution with uniform variance along the line. This is likely to be the case, as the regression tends to remove from outcome variable the variability between subjects and leave the measurement error, which is likely to be Normal. We observed the same process in the paired t method. We shall deal with deviations from this assumption later.

The equation of the line which minimises the sum of squared deviations from the line in the outcome variable is found quite easily, but the calculations are always done by computer. The regression equation of muscle strength on height is

$$\text{Strength} = -908 + 7.20 \times \text{height}$$Figure 8 shows the line drawn on the scatter diagram. The intercept and slope are call **coefficients**. The slope of the line is sometimes called **the regression coefficient**, with the emphasis on the 'the'.

Unlike the correlation coefficient, these coefficients have units. They can take any value. There is no maximum or minimum value which they can have. In the example, strength is measured in newtons and height in centimetres.

$$\text{Strength in newtons} = -908 \text{ newtons} + 7.20 \text{ newtons per cm} \times \text{height in cm}$$

The intercept has the same units as the outcome variable, the slope is in the outcome variable units per unit of the predictor variable.

Figure 9. Histogram and Normal plot of the residuals for the strength data

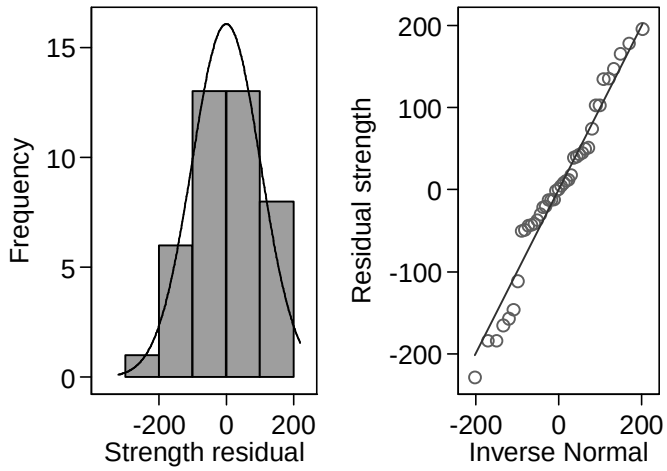
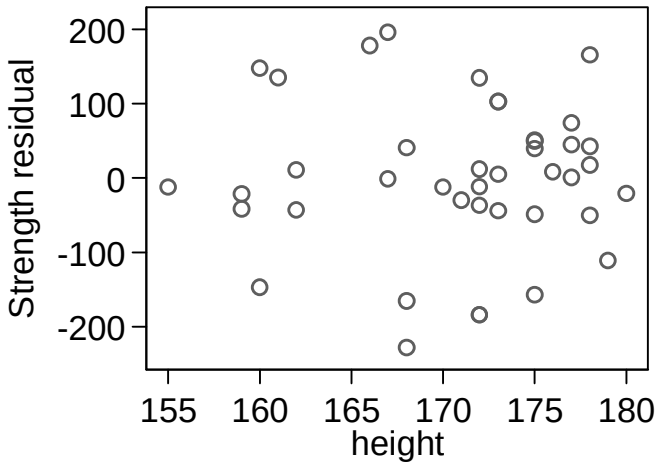


Figure 10. Scatter plot of residuals against predictor variable



Confidence intervals and significance tests in regression

We can find confidence intervals and P values for the coefficients subject to assumptions. These are that:

1. observations are independent, as usual for basic statistical methods,
2. deviations from line (as shown in Figure 7) should have a Normal distribution,
3. deviations from line should have uniform variance, that is, the variability should be the same for all values of the predictor.For the example, the slope = 7.20, the 95% CI = 2.15 to 12.25 newtons/cm. If we test the null hypothesis that in the population the slope = 0, we get P=0.006. Hence the data are inconsistent with the null hypothesis and the data provide good evidence that a relationship exists in the population from which these men come. The intercept = –908, 95% CI = –45 to –1771 newtons. Computer programs almost always test the null hypothesis that the intercept is zero as well, but this is very rarely of interest. The tests and confidence intervals use the t distribution.

Deviations from assumptions in regression

We call the deviations from the line, as shown in Figure 7, the **residuals**, what is left after the effect of height on strength has been removed. Both the appropriateness of the method of least squares and the confidence intervals and tests of significance depend on the assumption that the residuals are Normally distributed.

This assumption is easily met, for the same reasons that it is in the paired t test. The removal of the variation due to height tends to remove some of the variation between individuals, leaving the measurement error. Problems can arise, however, and it is always a good idea to plot the original scatter diagram and the residuals to check that there are no gross departures from the assumptions of the method.

We can check the assumption of a Normal distribution by histogram and Normal plot. These are shown in Figure 9. In this case the distribution looks approximately Normal.

We can check the assumption of uniform variance by plotting the residuals against the predictor variable. Figure 10 shows this and the spread of the residuals looks very similar along the graph.

Regression methods are fairly robust and small deviations from the assumptions should not cause problems. If there is an obvious departure from the assumptions, we can try a transformation of the data.

Correlation or regression?

Correlation and regression provide two different ways to look at the relationship between two quantitative variables. Correlation measures how closely they are related and makes no distinction between outcome and predictor. Regression measures what the relationship is and has direction. The regression of height on strength is not the same as the regression of strength on height, We must choose. The tests of significance are identical, however, for both regressions and for correlation.

Correlation and regression are closely related. If we calculate the sum of squares about the mean for the outcome variable and the sum of squares of the deviations from the regression line, then

$$\frac{\text{SS of deviations}}{\text{SS about mean}} = 1 - r^2$$

We call r^2 the proportion of variability explained by the regression. This is often written as R^2.

J. M. Bland
29 February 2012

References

Bland M. (2000) An Introduction to Medical Statistics. Oxford University Press.

Hickish T, Colston K, Bland JM, Maxwell JD. (1989) Vitamin D deficiency and muscle strength in male alcoholics. *Clinical Science* 77, 171-176.