Had we been working with the regression of x on y (rather than y on x), the *coefficient of determination* (r^2) would have been:

$$\frac{sum\ of\ cross\ products^2}{sum\ of\ squares\ of\ y} \text{ divided by } sum\ of\ squares\ of\ x$$

The *coefficients of determination* of both regressions turn out to be identical, since both can be re-written as:

$$\frac{sum\ of\ cross\ products^2}{sum\ of\ squares\ of\ x \times sum\ of\ squares\ for\ y}$$

So our *correlation coefficient (r)*, the square root of the *coefficient of determination* (r^2), has the property we expect (see first sentence of this section) that it involves no assumptions about whether y regresses on x or x on y.

r can therefore be calculated directly from data, without a preceding regression analysis, from the above formula, which can also be "square-rooted" to read:

$$\frac{sum\ of\ cross\ products}{\sqrt{sum\ of\ squares\ of\ x \times sum\ of\ squares\ for\ y}}$$

The maximum value of r is 1 (like the coefficient of determination), representing 100% fit of the points to the line, but like the regression coefficient (though unlike the coefficient of determination, which is a squared value) it can be positive or negative depending on the sign of the sum of cross products. This has to be checked when deriving r from any expression using the square of the sum of cross products, as this will be positive whether the un-squared sum is positive or negative.

The statistical significance of r can be checked in statistical tables for $n-2$ d.f., i.e. 2 less than the number of data points. Just as with regression, a straight line accounts for 1 d.f. As d.f. for all the data points is 1 less than the number of points, $n-2$ is left for the residual variation from the line.

An example of correlation

We might wish to test how far there is a relationship between the length and breadth of leaves of the leaves on a cultivar of *Azalea* based on measuring 12 leaves. There is no reason why one dimension should be designated the variable which is dependent on the other independent one; they are pairs of characteristics "of equal status" from individual leaves. The data and some of the calculations we require are shown in Box 17.4. From

BOX 17.4

Data of length and width (both in cm) of Azalea leaves.

	Length (y)	Width (x)	$y \times x$
	11.0	3.8	41.80
	11.7	5.3	62.01
	12.1	4.6	55.66
	8.0	4.2	33.60
	4.1	1.6	6.56
	7.0	2.8	19.60
	8.4	3.3	27.72
	6.8	3.0	20.40
	9.5	3.7	35.15
	6.2	3.2	19.84
	7.4	3.1	22.94
	10.9	4.0	43.66
Total	103.1	42.6	388.88
Mean	8.59	3.55	

Added squares

$$\sum y^2 = 953.17 \quad \sum x^2 = 161.16$$

Sum of squares

$$\sum y^2 - \frac{(\sum y)^2}{n} = 953.17 - \frac{(103.1)^2}{12} = 67.37$$

$$\sum x^2 - \frac{(\sum x)^2}{n} = 161.16 - \frac{(42.6)^2}{12} = 9.93$$

Sum of cross products

$$\sum (y \times x) - \frac{(\sum y \times \sum x)}{n} = 388.88 - \frac{(103.1 \times 42.6)}{12} = 22.88$$

these calculations:

$$r = \frac{sum\ of\ cross\ products}{\sqrt{sum\ of\ squares\ of\ x \times sum\ of\ squares\ for\ y}}$$

$$= \frac{22.88}{\sqrt{9.93 \times 67.37}} = 0.885$$

This represents a high degree of correlation between leaf length and breadth, and this is also shown by the plot of the points in Fig. 17.16a, which shows as the points forming a clear ellipse.

Is there a correlation line?

If r is 0.885, then surely we can square it to get to an r^2 of 0.783, and r^2 is the *coefficient of determination* in regression – from which we can deduce that the line which the points fit accounts for 78.3% of the variation of the points from their mean. In regression, the mean in question is that of the dependent variable (y in a regression of y on x).

But what is the correlation line, and which variable does the mean relating to the r^2 refer to? There are always two lines (Figure 17.16b), depending on which way round we hold the paper (compare Fig. 17.16b and c), and so which axis forms the horizontal treads of our "staircase" (page 247). These two lines are of course the lines for the regressions of y on x and x on y. The only time there is only one line is when there is perfect correlation, when the two regressions fall along the same line. Otherwise, although both regression lines pass through both mean x and mean y, their slope differs increasingly as the level of correlation deteriorates. The calculations for both regressions in our *Azalea* leaf example are given in Box 17.5, and you will see that – although the slopes (b) and intercepts (a) differ considerably (b is 2.304 for y on x and much smaller at 0.340 for x on y; a is respectively 0.413 and 0.629) – the *coefficient of determination* (r^2) for the two regressions is identical at 0.782.

To avoid confusion, I suggest that you calculate the *coefficient of determination* (r^2) only for regressions, and the *correlation coefficient* (r) only when quantifying the association between variables which have no dependence on each other. This is not to say that you should not carry out regression analysis with such variables for "prediction" purposes. In the *Azalea* leaf example, you might wish to measure only the width and use this to convert to length – I can't think why, but you might if an insect had eaten chunks out of some leaves, making measurement of many of the widths difficult?

Extensions of regression analysis

There are some very useful extensions of regression, which are beyond the scope of this elementary text, but it is important that you should know about them. After having worked through this book, you should have the confidence to be able to use at least some of the relevant "recipes" given in

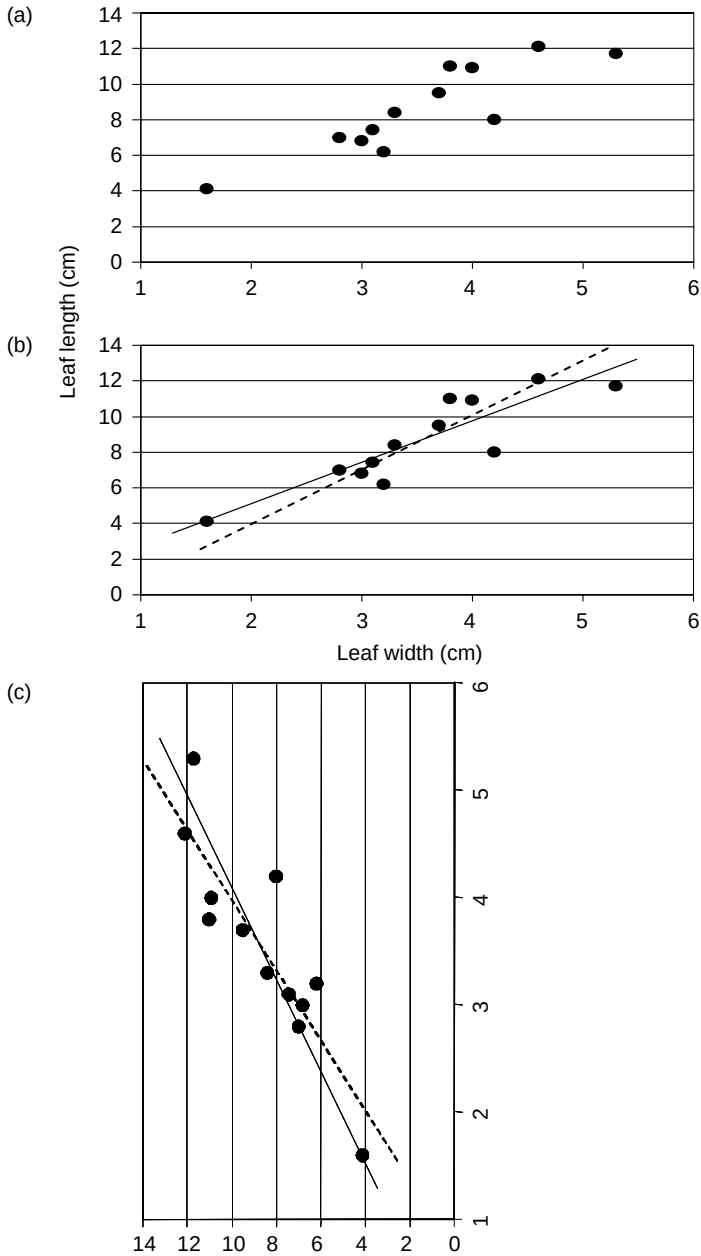
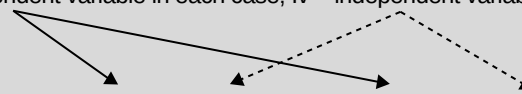


Fig. 17.16 (a) The correlation between length (y) and width (x) of *Azalea* leaves. (b) Regression lines fitted to these data; the solid line shows regression of y on x and the broken line the regression of x on y. (c) Graph (b) re-oriented to make y the horizontal axis.

BOX 17.5

Regressions using *Azalea* leaf data from Box 17.4, and using the statistics already calculated there. Dv = dependent variable in each case; Iv = independent variable.



Regression	**Length (*y*) on width (*x*)**	**Width (*x*) on length (*y*)**
Regression coefficient (*b*):		
$\frac{\textit{Sum of cross products}}{\textit{Sum of squares for Iv}}$	$\frac{22.88}{9.93} = 2.304$	$\frac{22.88}{67.37} = 0.340$
Intercept (*a*):		
Mean Dv – (*b* × *mean* Iv)	8.592 – (2.304 × 3.550) = 0.413	3.550 – (0.340 × 8.592) = 0.629
Regression equation:		
Dv = *a* + (*b* × *mean* Iv)	$y = 0.413 + (2.304 \times x)$	$x = 0.629 + (0.340 \times y)$

Regression analysis table:

Source of variation	**d.f.**	**Sum of squares**	**Mean square**	***F***	**Sum of squares**	**Mean square**	***F***
Regression ($b^2 \times$ *sum of squares for Iv*)	1	52.71 ($2.30^2 \times 9.93$)	52.71	35.99	7.79 ($0.340^2 \times 67.37$)	7.79	35.41
Residual *(by subtraction)*	10	14.66	1.47		2.15	0.22	
Total *(sum of squares for* Dv*)*	11	67.37			9.93		
Coefficient of determination (r^2)*		$\frac{2.88^2}{67.37 \times 9.93} = 0.782$			$\frac{2.88^2}{9.33 \times 67.37} = 0.782$		

* $\frac{\textit{sum of cross products}}{\textit{sum of squares for Dv} \times \textit{sum of squares for Iv}}$