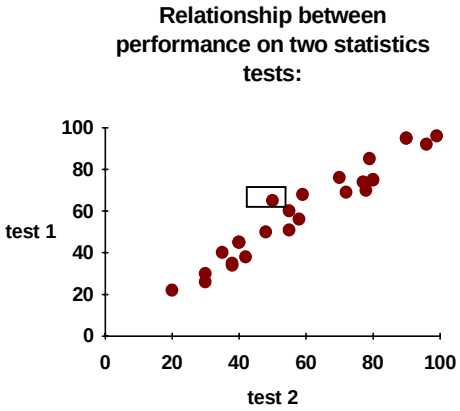
LINEAR REGRESSION:

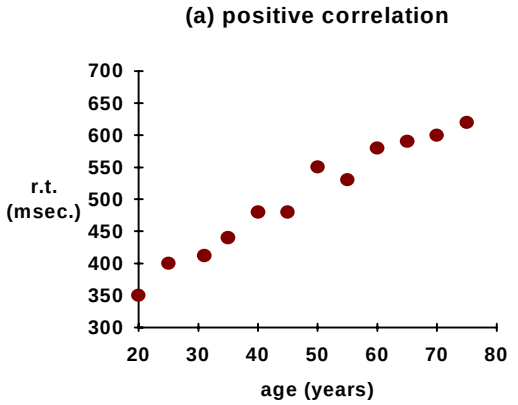
Often in psychology, we are interested in the relationship between two independent variables. For example, we might want to know if there a relationship between age and I.Q., or between people's Extraversion scores and their performance on some memory test.

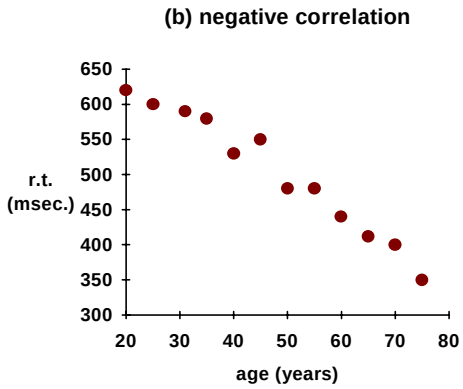
The relationship between two variables can be described graphically, with a **scatterplot:**

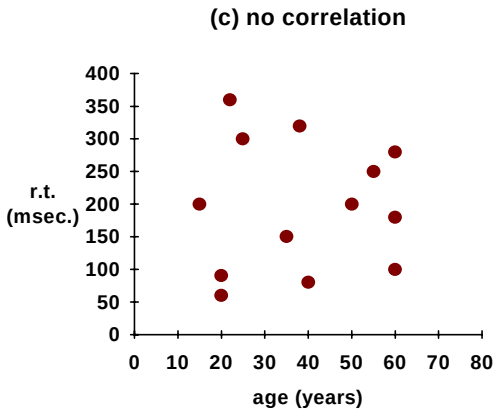


Each point on the graph represents the data for one individual, the permutation of their scores on the two variables that you are concerned with. In the case of the graph above, each point represents a particular person's performance on two separate statistics. The point in the rectangle represents a person with a score of 65 on test one and 55 on test two.

Often in psychology, we are interested in seeing whether or not a particular kind of relationship exists for our two variables: a *linear* relationship. Often, looking at the scatterplot will tell us what kind of relationship exists. In graph (a) below, there is a strong positive linear relationship: as values on one of the variables increase, so too do values on the other variable. Graph (b) shows an equally strong relationship, but this one is a *negative* relationship: an *increase* in the value of one of the variables is associated with a *decrease* in the value of the other. Sometimes, we might get a scatterplot like that in graph (c), suggesting that there is essentially no relationship whatsoever between the two variables in question!







If we find a reasonably strong linear relationship between two variables, we might want to fit a straight line to the scatterplot. There are two reasons for wanting to do this:

(a) for *description*: it would be nice to be able to summarise the relationship between the two variables with a straight line. This would act as a succinct description of the "idealised" relationship between our two variables, a relationship which we assume the real data reflect somewhat imperfectly.

(b) for *prediction*: we could use the line to obtain estimates of values for one of the variables, on the basis of knowledge of the value of the other variable. Thus if we produced a straight line which reflected the relationship between height and weight, given knowledge of a person's height, we could predict what their weight might be. On the basis of our knowledge of the relationship between the two variables (as reflected by the straight line), we would be able to go beyond the data that we actually possess.

There are various ways of actually fitting a straight line to our obtained data. You could do it by eye, drawing a line that goes through the set of data points; however, this method would be rather subjective and open to different interpretations (my idea of a "best" line might not coincide with yours). This is where **linear regression**, using the method of "least squares", comes in.

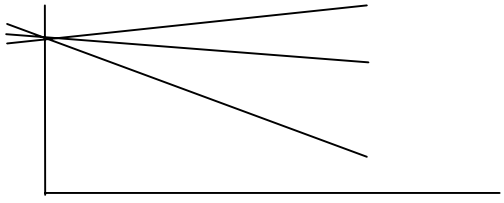
A brief algebraic digression:

To use linear regression, you need to know a tiny bit of algebra: the formula for a straight line. Any straight line can be draw if you know just two things: the slope of the line, and the point at which the line intercepts the vertical axis of the graph (a point known as the "intercept" of the line). Thus

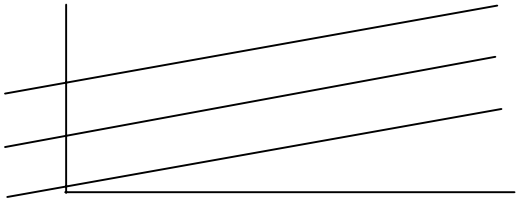
$$Y = a + b * X$$

where Y is a value on the vertical (Y) axis;
a is the point at which the line intersects the vertical axis of the graph;
b is the slope of the line; and
X is a value on the horizontal (X) axis.

Any particular straight line will have a particular slope and a particular intercept. Here are a set of lines which have the same intercept, but different slopes:



And here are lines which have the same slope, but different intercepts:



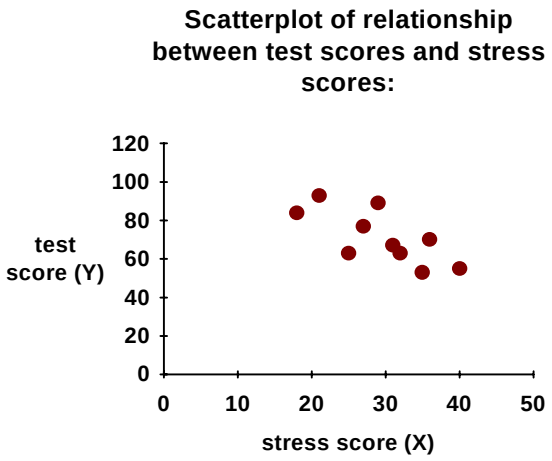
If we know a and b, for any particular value of X that we care to use, a value of Y will be produced. Linear regression involves finding values for a and b that will provide us with a straight line that goes through, or close to, as many of our data points as possible. The way that the "least squares" method does this is to minimise the vertical distance between the line and each point in the set of points to which the line belongs.

Linear regression step-by-step:

1. Suppose we gave two tests to each of ten individuals. One test is a measure of stress, and the other is a statistics test. We want to investigate the relationship between stress and statistics performance. Here are the raw data:

subject:	stress (X)	test score (Y)
A	18	84
B	31	67
C	25	63
D	29	89
E	21	93
F	32	63
G	40	55
H	36	70
I	35	53
J	27	77

2. If we draw a scatterplot of these data, we get the following:



It seems reasonably clear that there is a fairly strong negative relationship between stress score and statistics test performance: subjects who scored high on the statistics test tended to have low stress levels, and subjects who scored low on the statistics test tended to have high stress levels. So, the next step is to fit a regression line to these data.

3. To calculate the regression line, we need to find "a" (the intercept) and "b" (the slope) of the line. *Work out "b" first, and "a" second*.

4. To calculate "b", use the following formula:

$$b = \frac{\sum XY - \frac{\left(\sum X\right) * \left(\sum Y\right)}{N}}{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}$$

5. For our data, this is as follows:

subject:	**X**	**X^2**	**Y**	**XY**
A	18	$18^2 = 324$	84	18 * 84 = 1512
B	31	$31^2 = 961$	67	31 * 67 = 2077
C	25	$25^2 = 625$	63	25 * 63 = 1575
D	29	$29^2 = 841$	89	29 * 89 = 2581
E	21	$21^2 = 441$	93	21 * 93 = 1953
F	32	$32^2 = 1024$	63	32 * 63 =2016
G	40	$40^2 = 1600$	55	40 * 55 =2200
H	36	$36^2 = 1296$	70	36 * 70 =2520
I	35	$35^2 = 1225$	53	35 * 53 = 1855
J	27	$27^2 = 729$	77	27 * 77 = 2079
	$\Sigma X = 294$	**$\Sigma X^2 = 9066$**	**$\Sigma Y = 714$**	**$\Sigma XY = 20368$**

We also need the following:
N = the number of pairs of scores, = 10 in this case.
$(\Sigma X)^2$ = "the sum of X squared" = 294 * 294 = 86436. (Notice the difference between $(\Sigma X)^2$ and ΣX^2. The former means "square the sum of X"; in other words, add together all of the X values to get a total, and then square this total. The latter means "sum the squared X values"; in other words, square each X value, and then add together all of these squared X values to get a total. The difference between the two instructions is logical, if you bear in mind that you have to do things in brackets first).

6. Now we have all the quantities we need, we can work though the formula for b:

$$b = \frac{20368 - \left(\frac{294 * 714}{10}\right)}{9066 - \left(\frac{86436}{10}\right)} = \frac{20368 - 20991.60}{9066 - 8643.60} = \frac{-623.60}{422.40} = -1.476$$

b = -1.476. (b is negative, because the regression line slopes *downwards* from left to right: remember, that as values on the horizontal axis *increase* in size, so values on the vertical axis *decrease* in size).

7. Now, work out "a".

$a = \overline{Y} - b * \overline{X}$

$\overline{Y}$ is the mean of the Y scores: 71.4 in this instance.
$\overline{X}$ is the mean of the X scores: 29.4 in this case.

Therefore a = 71.4 - (-1.476 * 29.4) = 114.80.

8. The complete regression equation is therefore:

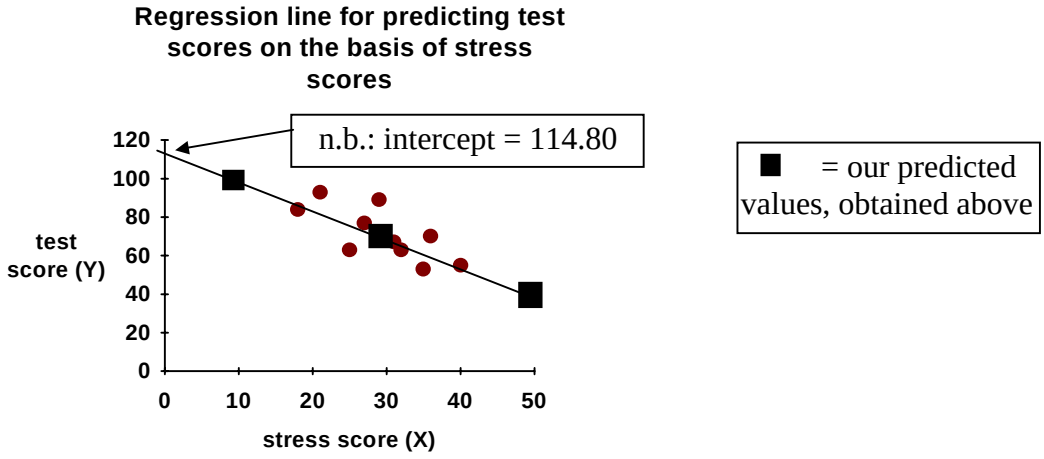
Y' = 114.80 + (-1.476 * X)

9. To draw our regression line, all we have to do now is to put some values of X into this equation, in order to obtain some predicted Y values. Then use these pairs of values (i.e., actual X and predicted Y) as coordinates to be joined up with a straight line on your scatterplot. It doesn't matter which particular values of X you use: any three different ones will do. (Use three as a check that your calculations are correct. Any *two* points on a graph can always be connected by a straight line, so if you just worked out two predicted Y values, you would not be able to guarantee that your equation was correct. If you work out three predicted Y values, it is unlikely that you would be able to connect them with a straight line unless your regression equation was right). As an additional check that you have the right equation, note that your regression line should go through the cluster of dots on the scatterplot. If not, then you have made a mistake with the calculations!

So, let's use our regression equation for arbitrarily-picked X values of 10, 30 and 50.

For X = 10, Y' = 114.80 + (-1.476 * 10) = 100.04.
For X = 30, Y' = 114.80 + (-1.476 * 30) = 70.52.
For X = 50, Y' = 114.80 + (-1.476 * 50) = 41.00.

Here's the original scatterplot, with the regression line drawn on:



One final, but important, point. We have worked out the regression line for predicting *test* score on the basis of knowledge of a person's *stress* score; this is the "regression of Y on X". If we want to predict *stress* score on the basis of knowledge of *test* score, the "regression of X on Y", we can't use this regression line to do it! The simplest way around this problem is to go back to the start; swap the column labels (so that the "X" values are now the "Y" values, and vice versa); and re-do the calculations. (Thus ΣX now becomes ΣY and vice versa; ΣX^2 becomes ΣY^2 , etc.) The result will be a regression line that is *different* to the one you originally obtained, and one that enables you to predict stress scores on the basis of knowledge of test scores. The following graph shows the two regression lines for the data in our example. The regression formula for predicting stress score on the basis of knowledge of test score is:

Y' = 55 - (0.359 * X).

where Y' is predicted Y score (i.e., stress score) and X is actual test score.

It may seem rather odd that the regression line for predicting Y from X is different from the regression line for predicting X from Y, but it stems from the fact that the least-squares method of fitting the regression line has a different goal in each case. To predict test scores (Y) on the basis of knowledge of stress scores (X), we want a regression line which is as close to all of our actual Y scores as possible. The least-squares method achieves this end by making the difference between the actual Y values and the regression line (i.e., our predicted Y values) as small as possible. However, if we are trying to predict X scores from known Y scores, we are concerned with a *different task*. The least-squares method is now being used with the intention of minimising a *differen*t set of discrepancies, making the difference between the actual X values and the regression line (i.e., our predicted X values) as small as possible.