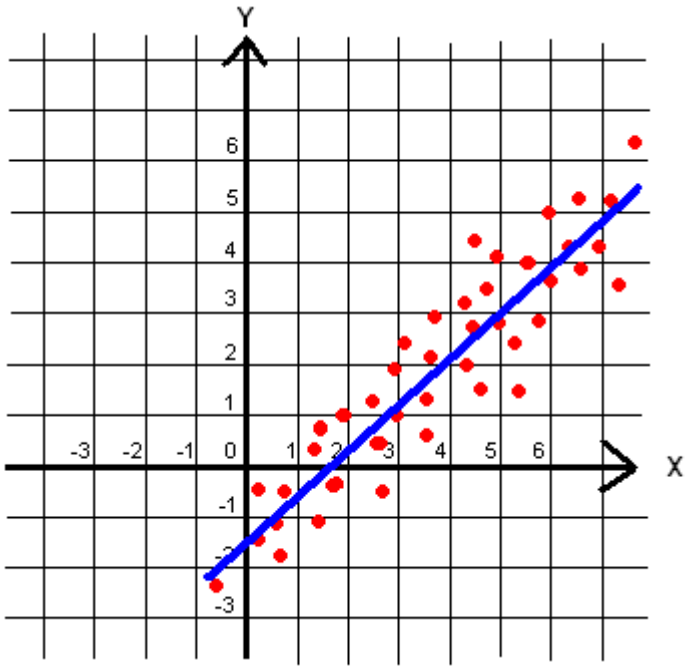
APPENDIX 5: Determining the line of best fit by least-squares regression

The relationship between two variables – for example between ***temperature*** and ***time*** – is commonly displayed graphically as a *scatter plot*. Often it is useful to attempt to represent the relationship, between an independent variable ***x*** and a dependent variable ***y***, by a straight line summarising the data points; such a straight line enables us to describe the general direction of an association between the two variables and to predict values that may not be displayed on the plot. A straight line of this kind is called the "line of best fit." It may also be called a "trendline".



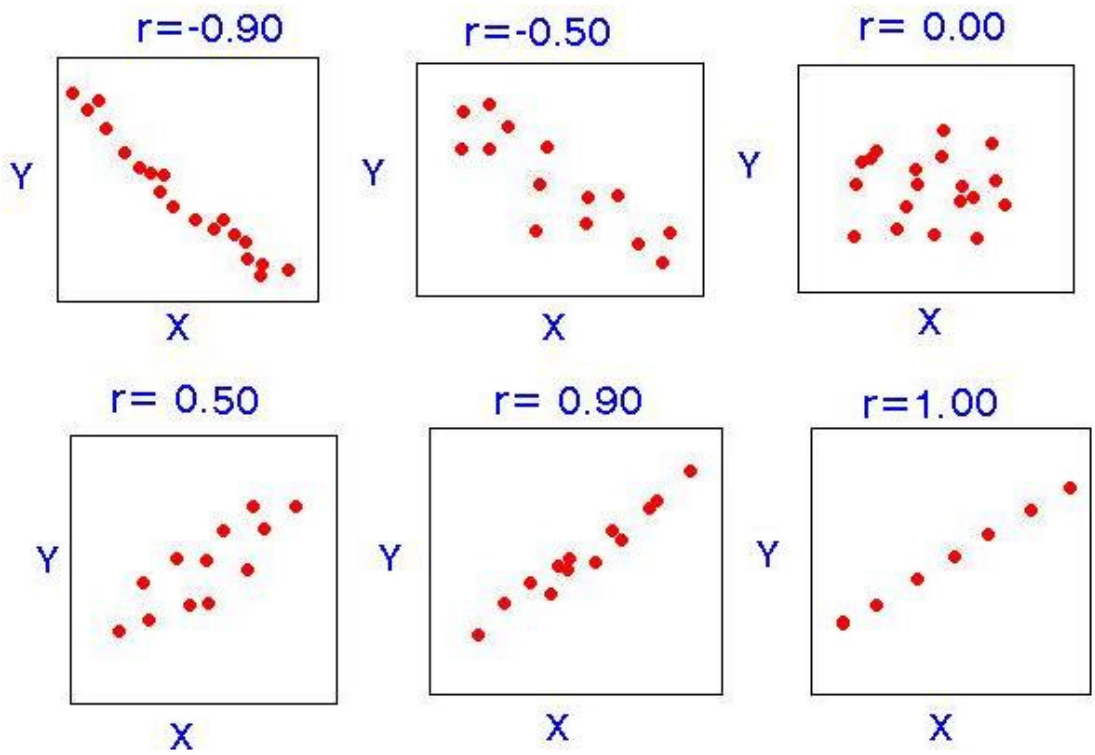
y = mx + b

The line of best fit to the data is a straight line that most nearly represents the data on a scatter plot by passing as close to as many of the data points as possible. Note that this line may pass through some of the points, all of the points or none of the points. The line of best fit will give us the best estimates of the gradient (slope) and intercept of the trend in the data. The equation for the straight line of best fit is of the form ***y= mx + b*** where ***y*** is the dependent variable, ***x*** is the independent variable, ***m*** is the slope and ***b*** is the intercept (i.e. the value of ***y*** when ***x*** = 0).

There is an objective procedure, the method of **least squares regression** (which we will not describe here), for determining the line of best fit through a set of data points**.** The least squares regression analysis not only provides a way of estimating the most appropriate values for the parameters '***m***' and '***b***' in the equation ***y = mx +b***, but it also provides for an estimate of the closeness of fit between this linear equation and the raw data. This estimate of the 'degree of fit' is known as the R^2 value, which for a straight line fit is equal to the square of the so-called 'correlation-coefficient' ***r***.

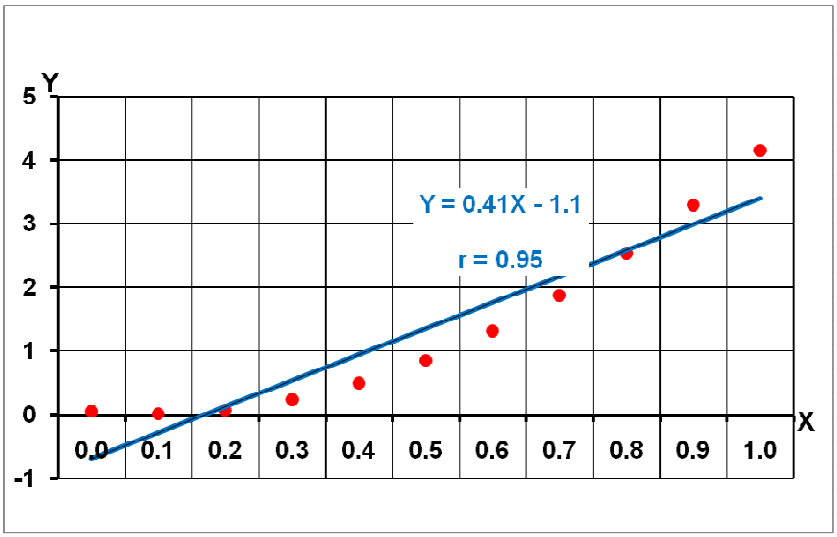
The correlation coefficient, ***r***, is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1. If there is a positive correlation, then whenever one variable (e.g. **x**) has a high value, so does the other (e.g. y) and when it (**x)** has a low value so does the other (**y**). If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1. If there is negative correlation, then

whenever one variable has a high value, the other has a low value and vice versa. A correlation coefficient of 0 means that there is no linear relationship between the variables.

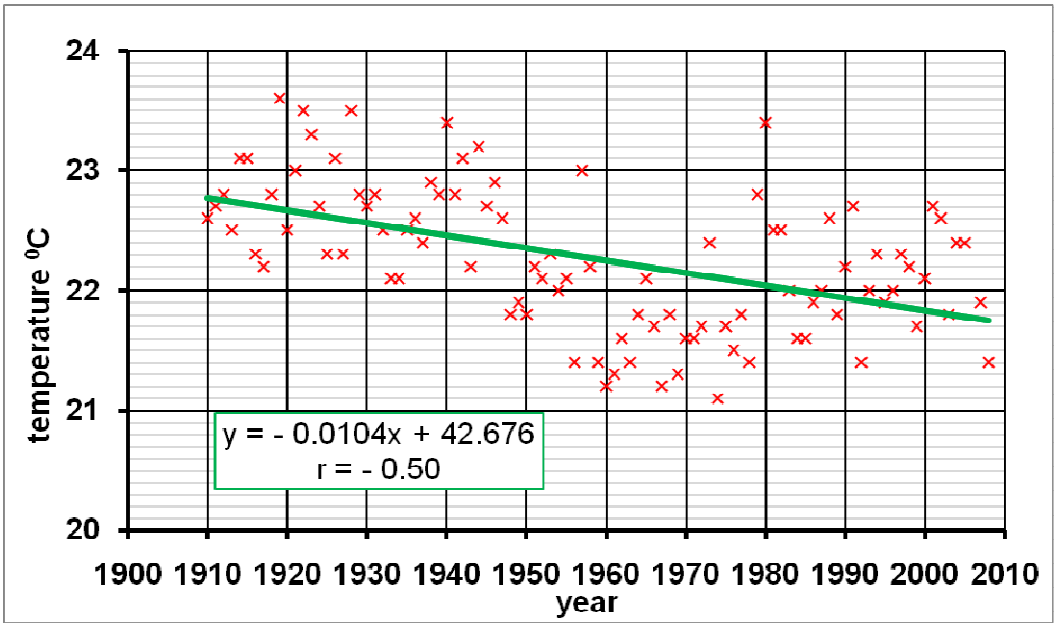


A scatter plot of data is linear if the pattern in its data points resembles a line. A linear trendline usually shows that something is increasing or decreasing at a steady rate. Even if there is no theoretical reason to assume that a scatter plot between two variables will closely resemble a straight line, in the absence of any theoretical reasons to the contrary, it is a reasonable starting point to attempt to fit a straight line trendline since this is one of the simplest of the possible relationships between two variables.

As the graph below shows it is possible to draw a straight line of best fit through a set of data points even when the data is obviously not linear. Notice how far some of the points are from the line. In fact the relationship between the two variables in this particular case is best described by a parabola rather than a straight line. Yet even so, in the first instance, the line of best fit gives a good approximation of the broad association between the two variables within the range of data points plotted. However, it would be a great mistake to attempt to use this linear trend to try to predict values of **y** for values of **x** beyond this range

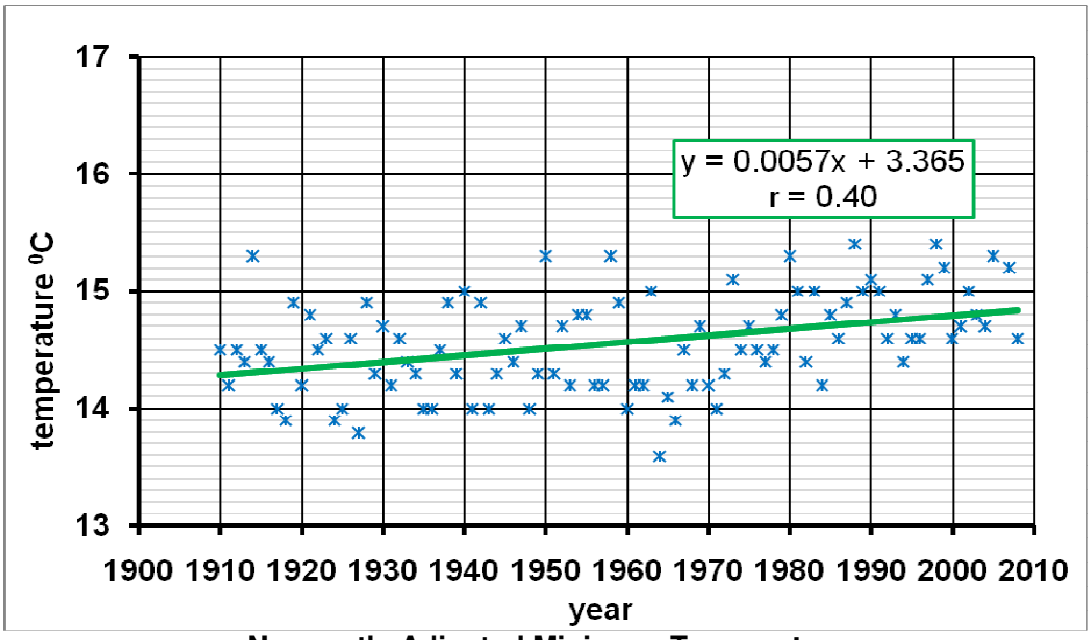


The application of the least squares regression to the High Quality temperature time-series is demonstrated below for the case of the Newcastle dataset *(BOM, 2009 c).*



Newcastle Adjusted Maximums Temperatures

CASE STUDY: Global Warming – the forest from the trees



Newcastle Adjusted Minimum Temperatures

An examination of the linear regression equations (displayed on the graphs) for the Newcastle time series reveals:

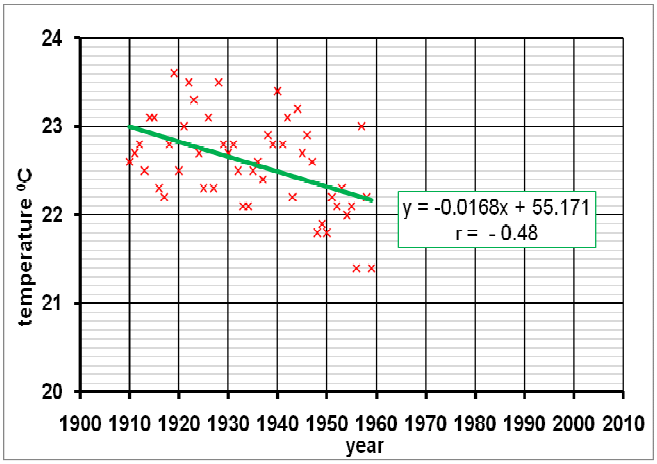
- The line of best fit for the Maximums had a downward slope as shown by the negative gradient
  m = - 0.0104 suggesting that on average the maximum temperatures declined over 1910 to 2008.
- The line of best fit for the Minimums had a moderate upwards slope as shown by the positive gradient **m = + 0.0057** indicating a modest upward trend in temperatures.
- For both maximums and minimums the there was quite a significant amount of scatter about the lines of best fit as indicated by the relatively low magnitudes of the correlation coefficients
  (**r = - 0.50** for the maximums and **r = 0.40** for the minimums).

CASE STUDY: Global Warming – the forest from the trees

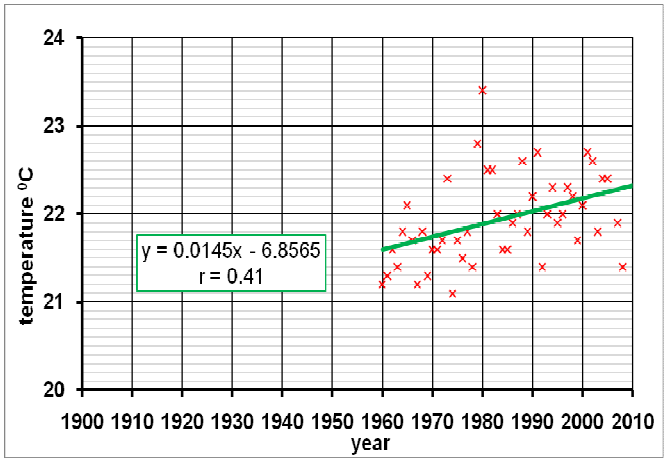
One important note of caution that arises from the Newcastle example is that it is very risky to use linear equations of this type to try to predict temperatures (**y** values) for years (**x** values) outside the range of data points included in the analysis (known as **extrapolating** from the data). For example, it would be absurd to use this equation to try to estimate what the temperatures were back in say the year 500 AD; we would get an estimate for the maximum temperature of **37.5 ^{0}C** by substituting **500** for **x** in the regression equation for the maximums (**-0.0104x500 + 42.676**) and of **6.2 ^{0}C** for the minimums (**0.0057x500 + 3.365**) – highly unrealistic estimates.

Even when using the line of best fit for **interpolation** (i.e. estimating a temperature **(y) within** the range of data points), great care is needed; for example, if we used the regression equation to get a theoretical estimate of the temperature in say 1960 we would calculate a temperature of **22.3 ^{0}C** (obtained from **-0.0104x1960 + 42.676**) whereas the actual observed annual temperature for that year was 21.2 ^{0}C, a full 1.1 degrees below the estimate.

How could this arise? Well if you look closely at the data for Newcastle maximums you will notice that for a decade or more in the 1940s and 50s maximum temperatures took a plunge. This was a wide spread phenomenon at that time, especially in eastern Australia where it was associated with a period of high rainfall. This highlights the fact that the line of best fit, as a first approximation to a trend, often obscures the real complexity in the relationship between two variables. This is demonstrated below by splitting the Newcastle maximum data into two equal periods, the years from 1910 to 1959 and the years from 1960 to 2008.



Newcastle Maximum Temperatures 1910-1959



Newcastle Maximum Temperatures 1960-2008

You will notice that the slopes of the trend lines are in completely opposite directions. What is less apparent is that despite the fact that the number of data points in the separate analyses has been halved (from 98 to 49), the magnitude of the correlation coefficients has only declined slightly from what it was in the total analysis; this suggests that applying the two separate regression lines has accounted for a much higher amount of the variability in the maximum temperature time series. However we have to be careful in interpreting correlation coefficients. A high correlation coefficient does not of necessity imply - indeed, **rarely** implies - anything about the physical cause of the mathematical association; it simply says that the two sets of numbers (in this case, year and temperature values) happen to vary in a particular way in relation to one another. The reality of whether there is a physical connection or not between the two variables is determined by the science, not from the mathematics. Great care is needed in interpreting regressions and correlations.