Starting a correlation project

Correlation projects are very popular with students – the work is accessible and on the syllabus, data can be collected from either primary or secondary sources, and simple and further mathematical processes can be demonstrated.

The usual approach is to take two sets of data, assume a linear correlation, and proceed on that basis. However, this approach does not always show that the student has a proper understanding of the techniques being used.

The following is a suggested approach that would better demonstrate a sophisticated understanding of the material.

1. Rather than looking only at two variables and establishing a correlation between these, it is better that the student compares the effect of two independent variables on a third by comparing the correlation coefficients between them. For example, if one were interested in the rate of infant mortality in various countries, one could consider the relative influence of "the number of doctors per 1000 people" compared to "percentage access to potable water".

2. It may well be that a better fit to the data is supplied by a non-linear model, for example an exponential model. Once a correlation coefficient of sufficiently large value has been found and a linear model has been considered, a different model can easily be considered and assessed by the graphic display calculator (GDC).

3. It may well be that the correlation coefficient has such a low value that the conclusion is that there is "no correlation" between the variables. In this case the student should be encouraged to investigate whether the variables are in fact statistically independent and use a chi-squared test to test this hypothesis.

   Note: Should this approach be followed, care must be taken that enough data has been collected to make the chi-squared test a valid option. In order to have two degrees of freedom (to avoid the necessity of using Yates's continuity correction) and enough data in each category (to avoid expected frequencies of less than 5), more than 30 data points are needed. This may well involve some further sampling.

The following approach is suggested to best fit the assessment criteria.

1. Decide upon the factor to be investigated and the two factors that might influence it. At this point write down the title of the project in such a way that the investigation is focussed.

2. Collect the data. If primary data is being used, ensure that enough is collected so that a chi-squared test will be valid (50 in the set). Randomly selecting a subset of this data set for an initial investigation is always an option. If a website is used to collect data, further sampling is easier. The method of sampling – random, stratified – should be stated and justified.

3. Before engaging in any calculations, a scatter diagram should be drawn. Not only does this constitute simple mathematics, relevant to the investigation, but it also gives

some indication as to the direction of the project – two linear models, a different model or independence. The student can also make an initial assessment of the levels of correlation, thus making a conclusion based on a mathematical technique.

4. The correlation coefficient with one of the factors is then calculated.

   There are many different formulas that can be used to calculate the correlation coefficient *r*. It is suggested that the most useful formula is

   $$r = \frac{1}{n}\sum_{k=1}^{n} z_{x_k} z_{y_k}$$

   where *n* is the number of data points used, and

   $$z_x = \frac{x - \overline{x}}{\mathrm{sd}_x} \quad \text{and} \quad z_y = \frac{y - \overline{y}}{\mathrm{sd}_y}$$

   are the standardised scores of each data point.

   The above formula makes use of two pieces of simple mathematics – the mean and the standard deviation (sd) – and so makes relevant their calculation. Use of spreadsheets is envisaged here.

   Agreement with the answer obtained on the GDC can be shown.

   Note: When using a spreadsheet, care must be taken with the above formula to use the appropriate version of the standard deviation as specified in the mathematical studies SL guide.

5. At this stage, the direction of the project becomes apparent. The equation of the regression line *y* on *x* can be calculated if this is appropriate, or the χ^2 test for independence can be used if the correlation coefficient is close to zero.

   It may well be that neither route is applicable. This is part of the student's assessment of the appropriateness of the mathematics being applied.

6. Investigation of the second factor should now be undertaken. Full details of the calculations for the chi-squared value, correlation coefficient or regression line will be required the first time that they occur in the project. Thereafter the graphic display calculator may be utilized.

7. A comparison of the results and final conclusion then complete the project.

A variation on the theme

Using the GDC, other models for the data set can be examined and the relative merits of each assessed. For example, the effect of concentration of reactants and the time to complete a chemical reaction, or the speed of microprocessor chips and the time to complete a given task, might be better modelled as an exponential function.

The method of approach in this case would be to assume a linear model and then try the different types of correlation as part of the validation process. The start of the method is essentially the same as in the traditional approach, differing only at the very end.