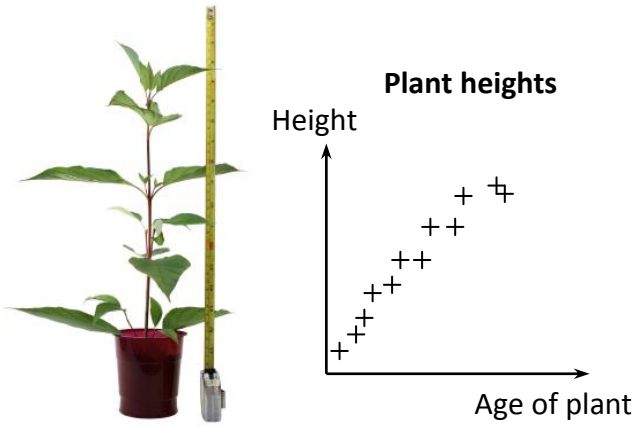
In this activity you will use simulations to help you to understand:

- how regression is used to find a relationship between two variables
- how a correlation coefficient is used to measure the strength of the relationship.



Information sheet

Regression

This is the analysis of the association between a dependent variable and one or more independent variables.

The association is usually given as an equation in terms of the dependent variables, from which values of the independent variable can be predicted.

The simplest case is that of linear regression, in which the regression equation of y on x is written as $y = a + bx$. The parameters, a and b, are called regression coefficients.

Correlation

This indicates the nature and strength of the relationship between two variables. The correlation is positive when each variable tends to increase or decrease as the other does. The correlation is negative when one variable tends to increase as the other decreases. Data pairs that show a close relationship are said to be highly or strongly correlated.

It is important to appreciate that high correlation need not imply a causal relationship. For example, the number of car owners and average daily sales of food in each of a number of cities are likely to be highly correlated, but this may simply be reflecting the size of the populations of these cities.

Try this

Go to http://www.mis.coventry.ac.uk/~nhunt/regress/index.html

First read about the contents of this module. Note that the recommended route is indicated at each stage by the **red** option.

Next go to the top of the page and click on …

Introduction

This page includes spreadsheets that revise and test the use of $y = mx + c$. Try these if necessary.

Then continue on the recommended (red) route by clicking on ...

Examples
Read through this page – it gives real life examples where regression and correlation are useful. Then move on to ...

Lines
This page introduces the 'least squares' method for finding the line of best fit. Click on spreadsheet and work through the exercise that shows why this method is used. You will need to use trial and improvement to find some of the answers. Click on 'Show answers' after you have tried each question.
Then use Back to return to the 'least squares' page and follow the recommended (**red**) route again to ...

Scaling
This part of the programme shows what happens to the regression line if you change the units and origin of the variables x and y. This is time-consuming and not essential, so move straight on to ...

Criteria
The spreadsheet link on this page gives an exercise that tries out other ways of finding a line of best fit. Try this if you wish, but as it is time-consuming and not essential you may prefer to move straight on to ...

Goodness of fit
Read this page (it explains what this section is about), then follow the recommended route by clicking on ...

Standard error then **R-squared**
These pages look at two ideas for measuring how well a line fits data. Considering them leads into the measure of correlation you will be using, namely Pearson's product-moment correlation coefficient. So read through these pages, but do not spend any time on the spreadsheet for investigating the calculation of r^2 on the second sheet. Instead move on to ...

Correlation
Read this important page. Use the **spreadsheet** to see how good you are at recognising positive, negative, strong and weak correlation.
Then click on Back followed by …

More correlation
Use the **spreadsheet** link on this page and try the exercise.
Then click on Back and follow the recommended route by clicking on ...

Causality then **Linearity** then **Significance**
Read these pages (they explain some of the problems associated with interpreting the correlation coefficient), but do not use the spreadsheet links on the last of these pages unless your teacher advises you to do this.

Next click on ...

Assumptions then **Linearity** then **Independence** then **Constant variance** then **Normality** then **Checking**

These pages consider some of the assumptions you make when you use a regression line. Read through each page, but do not spend any time on the spreadsheets on the Checking page unless your teacher asks you to do this.

Instead move on to ...

Prediction
This page shows how a regression line is used to make predictions, and explains interpolation and extrapolation. Read this carefully, then click on ...

Variation
Try the **spreadsheet** on this page – it shows how the accuracy of the regression line depends on the number of observations that were used to find it.

Then click on Back and move on to:

Error margin
Read the first four lines only, then move on to ...

Non-linear
This explores the use of different types of function to model how the value of a car depends on its age. Use the **spreadsheet** to find out how to use Excel to find linear and non-linear regression lines.

Reflect on your work

- Explain what is meant by the terms regression and correlation.
- What is the difference between interpolation and extrapolation?
- If you find that $r = 1$, what can you say about the relationship between the variables? What if $r = -1$?
- What might a value of r near to zero indicate?
- Correlation does not imply causation.
  Explain what this means. Suggest a real-life example.