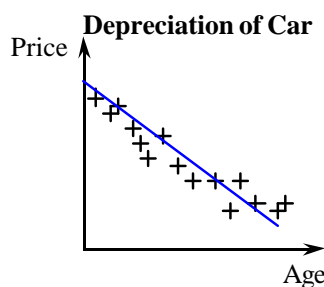


DISCUSS Regression and Correlation

In this activity you will use simulations to help you to understand how regression is used to find a relationship between two variables.



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

First read about the contents of this module and 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.

Read this page - it includes spreadsheets that revise and test the use of $y = mx + c$. Try these if you wish. 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.

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 beyond the scope of Using and Applying Statistics. Move on by clicking:

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. These measures, called the standard error about the line and the coefficient of determination are not mentioned in the specification for Using and Applying Statistics, but considering them leads into the measure of correlation that 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 of these sheets. Instead move on to:

Correlation

Read this important page and 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. 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 associated spreadsheets. 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

The formulae given on this page are beyond the scope of Using and Applying Statistics. Just read the first four lines 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.



Teacher Notes

Unit Advanced Level, *Using and applying statistics*

Notes This module covers much more than is required by the specification for *Using and applying statistics*. The main parts of the module and what students are expected to learn from them are listed below. The UAS column indicates those topics that are included in the specification for *Using and applying statistics* and the Useful column indicates other sections that you may like students to use as extensions. The Omit column identifies the parts that it is recommended that you omit unless they would be useful for the students' other areas of study. The student worksheet leads students through the UAS and Useful topics listed, but can easily be adapted if you wish students to do more or less than this.

Section	Main Points	UAS	Useful	Omit
Introduction	Scatter Graphs Dependent and independent variables. The purpose of scatter plots. Revision of $y = mx + c$.	√		
Examples	A line of best fit explores the relationship between the independent variable x and the dependent variable y . The x value can be used to predict the y value. The relationship is not usually exact – it involves a random unpredictable element. The relationship can be modelled by an equation. The linear model $y = mx + c$ is not necessarily the best.	√		
Lines	The Least Squares Method Regression line passes through mean point. Sum of residuals is zero. Sum of residual squares is minimised. The Excel functions INTERCEPT and SLOPE can be used to find the least squares regression coefficients.	√		
Scaling	What happens to the regression line when a constant is added to each x (or y) value and what happens if each x (or y) value is multiplied by a constant.			√
Criteria	Other ways of finding a linear model may be better in cases where there is an outlier that heavily influences the least squares line.			√
Goodness of Fit	Three statistics that are used to measure how well a line fits the data are the standard error, s , the coefficient of determination, r^2 , and the correlation coefficient, r .		√	
Standard Error	The standard error, s , is the square root of the average of the squared residuals.		√	
R-squared	The coefficient of determination, r^2 , is the % of the variability (as measured by the sum of the squared residuals) that is explained by the regression line.		√	
Correlation	The Pearson correlation coefficient, r . Correlation is perfect when $r = \pm 1$, strong when r is greater than 0.8 in size and weak when r is less than 0.5 in size.	√		
More Correlation	The three common pitfalls when interpreting a correlation coefficient involve causality, linearity and significance. With a small data set it is easy to achieve high correlation. High correlation does not necessarily mean that the relationship is linear. A zero value does not necessarily mean that there is no relationship – it could be non-linear.	√		
Causality	Strong correlation does not necessarily mean that a change in one variable causes a change in the other. The changes may be due to a third variable.	√		
Linearity	A very strong non-linear relationship may give a low value correlation coefficient.	√		



Section	Main Points	UAS	Useful	Omit
Significance	A high value of the correlation coefficient may mean very little if the sample size is very small. Spreadsheets show what a bivariate Normal distribution looks like, what samples from such a distribution look like, how likely different values of the correlation coefficient are when there is no correlation and when there is a particular correlation.	√		√
Assumptions	The relationship between x and y is assumed to be: $y = (\mathbf{a} + \mathbf{b}x) + \mathbf{e}$ where \mathbf{e} represents the unpredictable element in y due to random variation or measurement error. The values of α and β can only be estimated from the intercept, a , and slope, b , of the regression line. To quantify the reliability of any predictions you must assume linearity, independence, constant variance and Normality.			√
Linearity	Non-linearity can be detected by examining the residuals. A plot of residuals against x values can be used to test whether the relationship is linear or non-linear.			√
Independence	Lack of independence of the random errors \mathbf{e} usually leads to a pattern in the residuals.			√
Constant Variance	For simplicity it is usually assumed that the variance of the random errors \mathbf{e} is the same for all values of x i.e. the points are assumed to lie in a band of constant width.			√
Normality	It is assumed that the errors have a mean of 0 and variance σ^2 . (Section still under construction.)			√
Checking	The assumptions of linearity, independence and constant variance can all be checked by a plot of residuals against x values. The assumption of Normality can be checked by drawing a histogram or Normal probability plot of the residuals.			√
Predictions	Making predictions by interpolation is reliable, but extrapolation is not.	√		
Variation	Different samples give different lines and different predictions. The more observations that are used to find the regression line, the more accurate the line and the predictions made from it.	√		
Error Margin	A confidence interval for the mean value of y expected when $x = x_0$ is: $(a + bx_0) \pm t_{\mathbf{a}/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}$ A prediction interval for the individual value y expected when $x = x_0$ is: $(a + bx_0) \pm t_{\mathbf{a}/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}$ where t is calculated from $n - 2$ degrees of freedom with tail area $\mathbf{a}/2$.			√
Non-linear	Using Excel to fit linear, quadratic, power, logarithmic and exponential trend lines.	√		

LINKS Most are useful for *Using and applying statistics*.

