Linear Regression

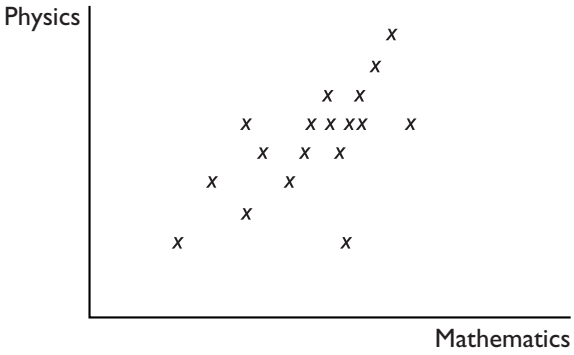
Junior high school

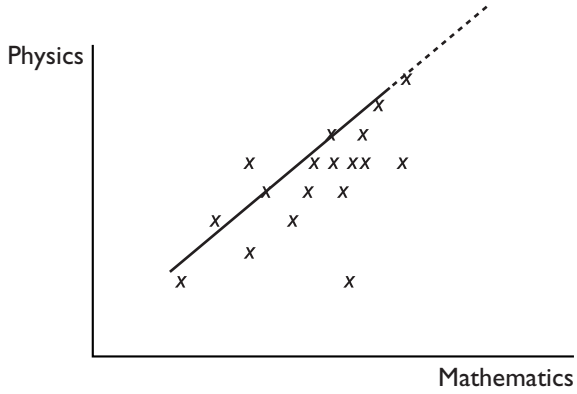
Objective

To explore how a bivariate (two variables) series of data could be used to estimate the value of one of the variables for a given value of the other.

Explanation of the activity

For many bivariate distributions there appears to be a connection between the two items of data. For example, if we collect the marks in a science examination for a class of students and we collect the marks for those same students in a mathematics examination, then we may see that a student with good marks in science could have good marks in mathematics. There appears to be a *positive correlation* between these marks. If the marks for each student are plotted on a graph, then the line that we can draw that best fits the spread of the marks can be used to estimate the value of one mark given the other. The plotting of the examination marks gives us a *scatter diagram*.





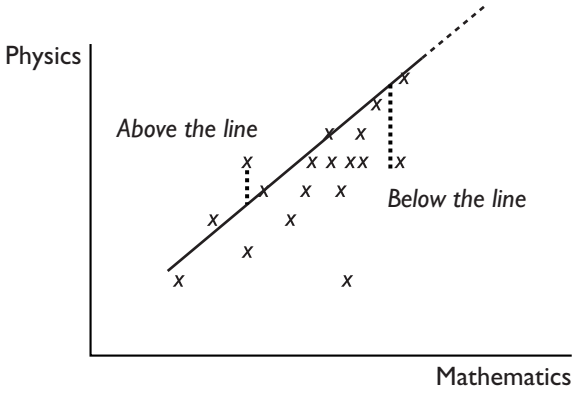
The line is called a line of linear regression, and we can use the calculator to find its equation.

If we look at the previous example of the line drawn on a scatter diagram, it will have a well known equation for a straight line:

$y = \mathbf{a} + \mathbf{b}\,x$

The position of the line on the scatter diagram can be determined by drawing the line such that, when we square *all* individual differences above the line and *all* individual differences below the line, the sum of all these squares comes to the least value.

This known as the **method of least squares** and the line is the **line of regression of y on x**. This is the process which the calculator uses.



When a series of bivariate data has been entered correctly, then the calculator can be used to find the values of **a** and **b**, to give the equation of the line of regression from which we can make estimates of the variable *y*, if we know a value for the other variable *x*.

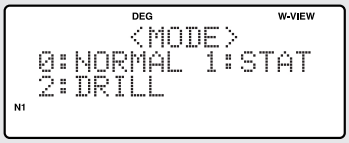
Linear Regression

Junior high school

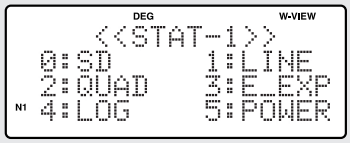
Using the calculator

Set the calculator in Statistics mode.

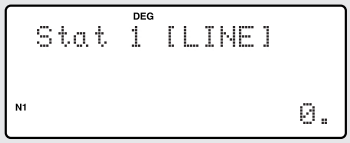
MODE key



Numeric key 1



Numeric key 1

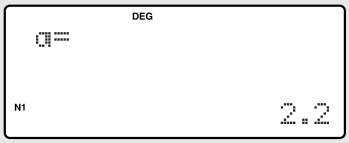


The calculator is now set for bivariate data entry; each pair p, q being input as p [(x, y) key] q [DATA key]. For example, if the bivariate pair 0, 2 were the first item of a set of data, they would appear on the calculator as:

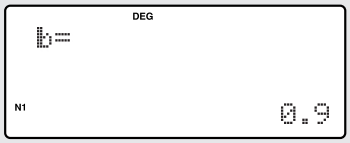
For the following table, enter the bivariate data shown. The calculator screens are shown for the whole process

p	0	1	2	3	4
q	2	3	5	4	6

Having entered the data, the values of the **a** and **b** in the equation of the line of regression of y on x, y = **a** + **b** x, can be found using the calculator. The ALPHA key identifies **a** and **b** as extra functions on the bracket keys



ALPHA **a** =



ALPHA **b** =

The line of regression of y on x is therefore y = 2.2 + 0.9 x

Estimates of the value of y can be found by substituting given values of x in the equation.

Points for students to discuss

Since the line of regression of y on x is not a perfect fit for all the points on the scatter diagram, it is often referred to as a line of best fit. Hence, only estimates of the value of y can be made for any given x.

The concept of two variables having a connection between them leads to ideas of positive correlation (both increasing or decreasing together) and negative correlation (one increasing whilst the other decreases).

A correlation coefficient is a way of attempting to measure this connection numerically.