Linear Correlation

You may be required to find and use a $y\ on\ x$ regression line ($y = ax + b$) or an $x\ on\ y$ regression line ($x = ay + b$)

Dependent and Independent Variables

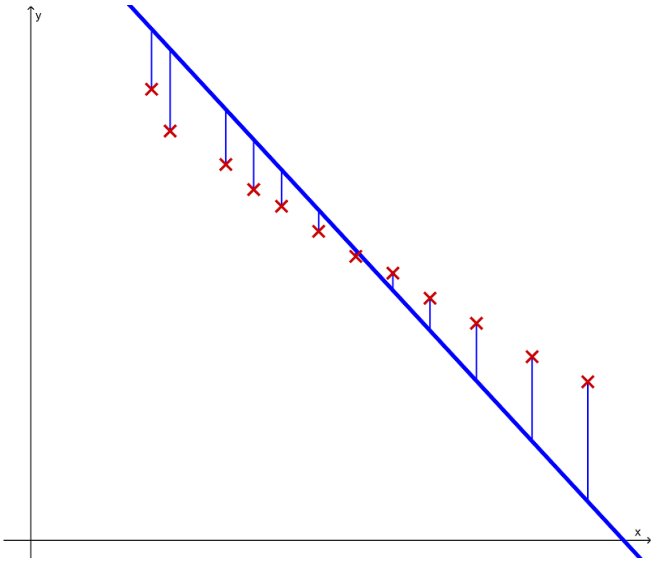
To decide whether you need to use $y\ on\ x$ or $x\ on\ y$, then it is important to understand which of the variables is dependent and which is independent. Here are 2 examples to make it clear to you

- Age (x) and height (y) of children. Height might depend on age; therefore, height is dependent variable and age is the independent variable. $y\ on\ x$ is appropriate
- Length (x) and width (y) of leaves. There is not an obvious dependent variable. $y\ on\ x$ or $x\ on\ y$ are both appropriate.

y on x

When we are finding the $y\ on\ x$ regression line, then the calculator will minimize the vertical residuals (actually it minimizes the sum of the squares of the vertical distances).

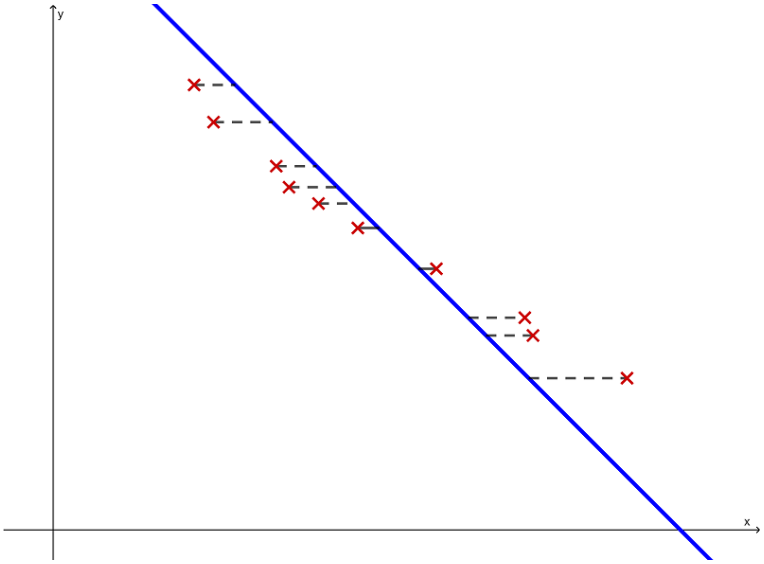
We can use this line to make predictions about y when we know an x value



x on y

When we are finding the $x\ on\ y$ regression line, then the calculator will minimize the horizontal residuals.

We can use this line to make predictions about x when we know an y value



Note that if we plot both the $y\ on\ x$ and the $x\ on\ y$ lines, then they will intersect at the mean point $(\bar{x}, \bar{y})$

Method

You will only ever be asked to use your calculator to do this, so it is important to know how to use the statistics mode in your particular make of calculator.

However, the method should be the same:

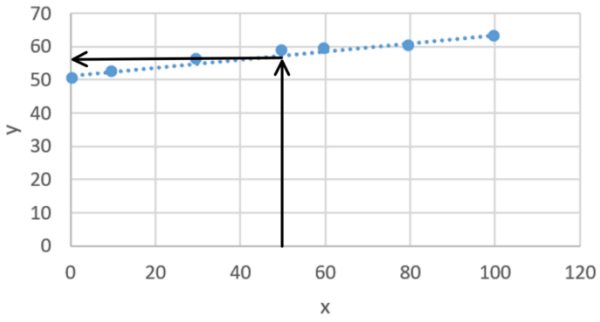
- Put the data into your GDC statistics mode with the independent variable in the first list and dependent variable in the second list
- Check that the data can be modelled by a linear model by plotting a scatter graph. Does it look like a straight line?
- Ensure that diagnostics is switched on (if appropriate) to calculate r (Pearson's Correlation Coefficient).
- Change the degree of accuracy. You are usually required to give answers to 3 significant figures. However, if you are going to use your equation to make a prediction, then a higher degree of accuracy will be needed. On some calculators you can store the equation that your calculator finds.

Interpolation and Extrapolation

Here is a $y\ on\ x$ regression line.

We can use an x value to predict a y value if the value lies in the interval of values given in the data.

For example, I could predict y for an x value of 50 (interpolation).



But, I could NOT predict y for an x value of 120 (extrapolation), as this is outside the interval of values that we have data for.

With $y\ on\ x$ regression line, we should NOT make a prediction about x from a y value. For example, I could NOT predict x if we know that y = 60.