Least-squares regression line

Tuesday, 26 February 2019 5:57 PM

★ By the end of the lesson I would hope that you have an understanding and be able to apply to questions the following concepts:

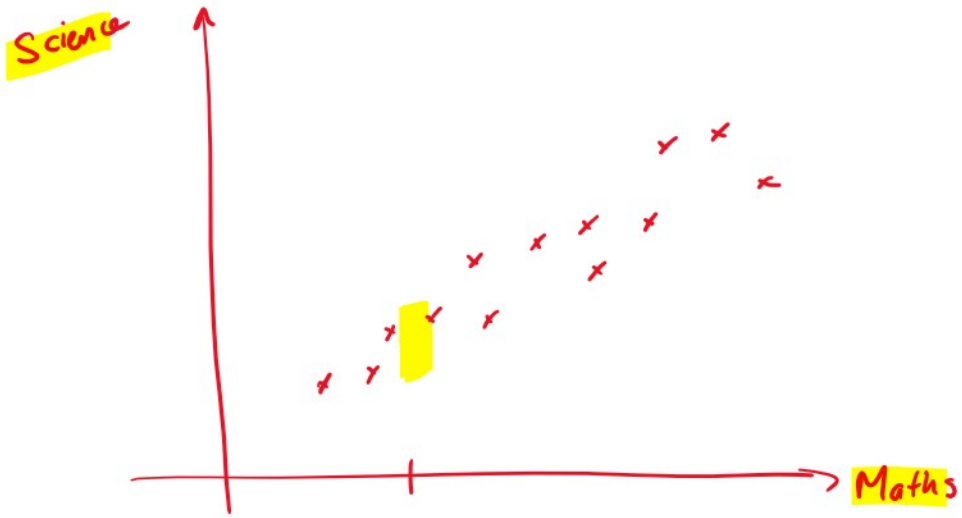
- Understand what it means by the term linear regression
- Know what the least squares method is
- Understand the term residual
- Know what the least squares line is

RECAP:

This is a new topic which, whilst it builds on the previous one, takes the learning to a whole new level.

Line of best fit

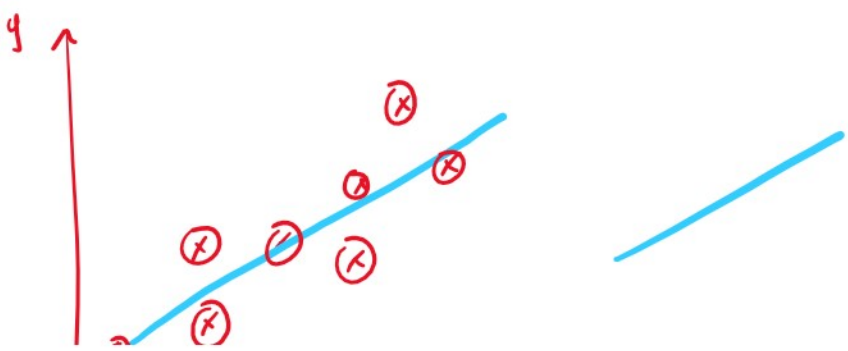
We have talked, in past lessons, about how we might be able to use a scatter plot to show an association between a classes **Mathematics Scores** and their **Science Scores.**

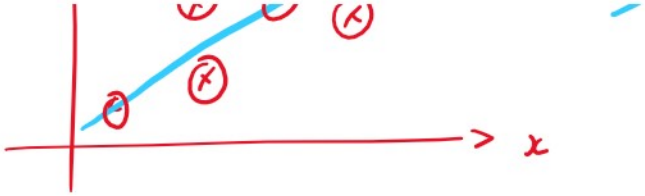


It is **highly unlikely** that we will have a correlation coefficient, r, of 1 for this data. There are too many external variables which affect test scores. But, we see there is probably a distinct relationship between the two.

To help us **predict** Science scores from Mathematics scores, we can draw a **line of best fit** through the data and use it to help us predict scores.

Drawing lines of best fit is challenging at the best of times! It's also not particularly accurate.



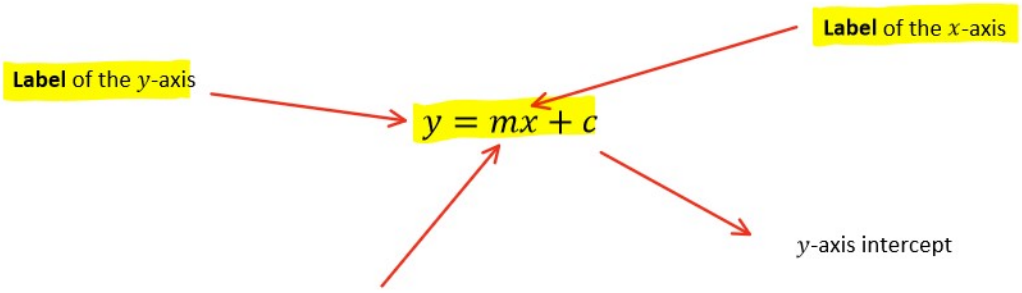


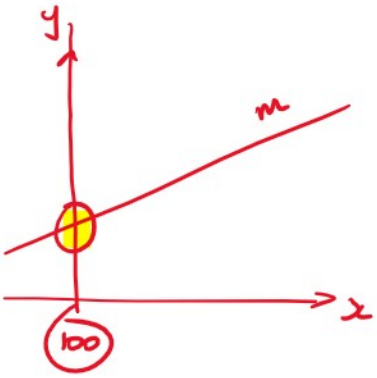
Hence, using the CAS, we can come up with something a little more accurate.

But more on this a bit.

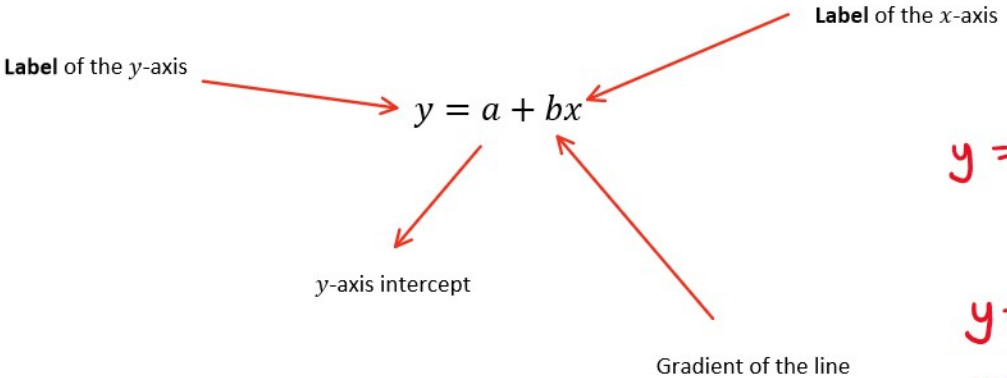
Lines of best fit are LINES and hence we can use $y = mx + c$

We have, in Years 9 and 10, looked at the equation of a straight line.
We know that all straight lines have equations $y = mx + c$





Further Maths tries to trick us!
They use the same **idea** of the equation of a straight line but move the things around ...
Sigh!



$y = a + bx$

$y = mx + c$

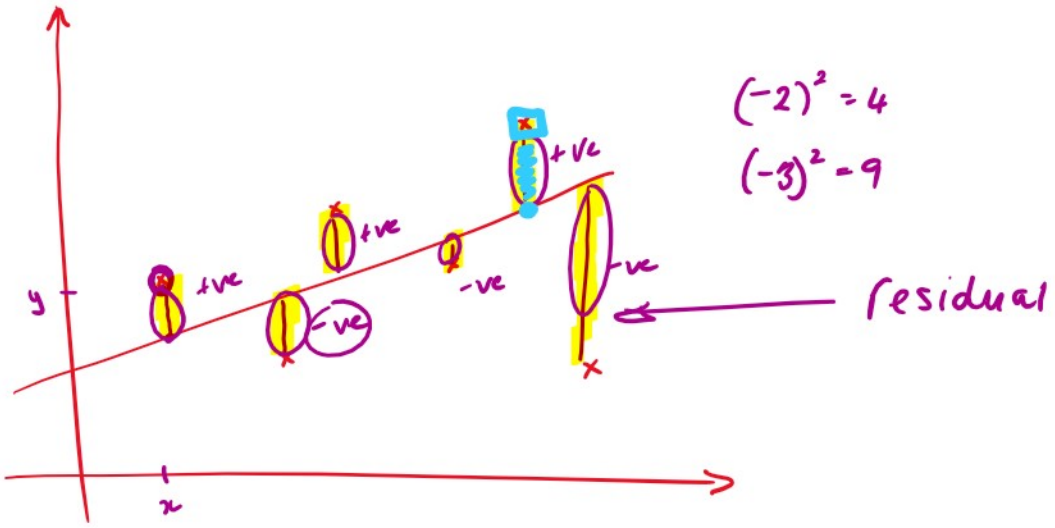
$y = c + mx$

$y = a + bx$

Creating the line of best fit (now known as the least squares line)

So, we now call the line of best fit the **least squares line.**

To be able to create this you, and the CAS, need to know about residuals.



Residual = Actual value - Predicted value

The vertical distances between each data points and the least squares line is called a **residual**.

The least squares line of the line that minimises the squares of the residuals and assumes the following:

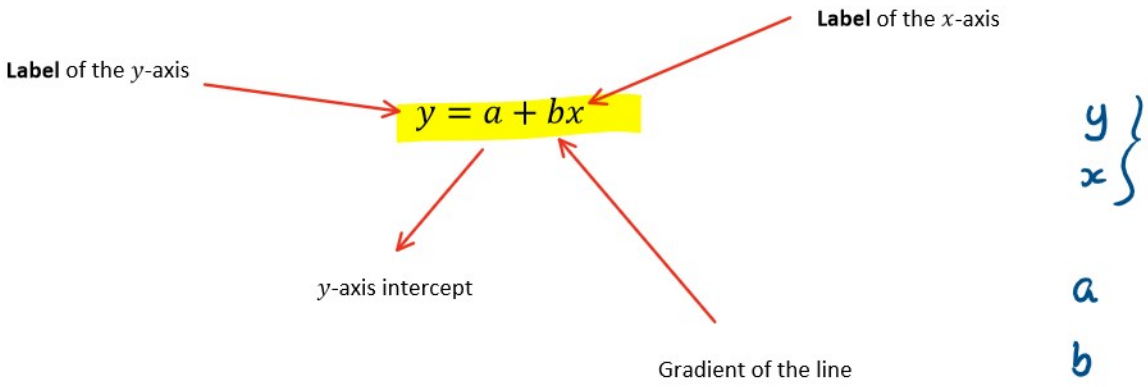
- The data is numerical
- The association is linear
- There are no clear outliers

Finding the equation of the least squares line

Thankfully you do not need to draw graphs, find residuals, square all the values and do all that Maths!
This is Further Mathematics so there must be an easier way!

There is ... and it's done using two formulae:

Remember, we are trying to get the data to create a line.
We use the equation of a straight line from above:



$$b = \frac{r \times s_y}{s_x}$$

Slope of the line

r: Pearsons correlation coefficient
s_x: Standard deviation of x
s_y: Standard deviation of y

$$a = \bar{y} - b \times \bar{x}$$

y-axis intercept

Slope of line from previous equation

$\bar{x}$: Mean of the x values
$\bar{y}$: Mean of the y values

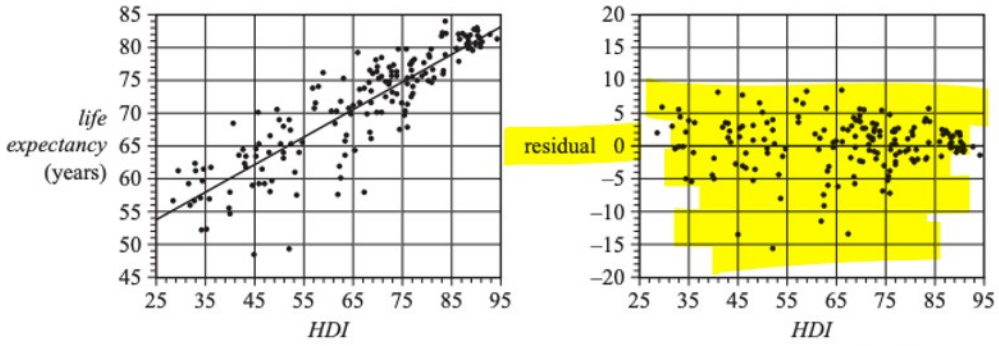
Warning Will Robinson

You are using the value you calculate of b in the second equation. If you get this value wrong, then your equation will also be wrong! You need to be very careful here to ensure you get the values correct.

VCAA Exam Question on this concept
2016 Paper 1

The scatterplot below shows life expectancy in years (*life expectancy*) plotted against the Human Development Index (*HDI*) for a large number of countries in 2011.

A least squares line has been fitted to the data and the resulting residual plot is also shown.



Data: Gapminder

The equation of this least squares line is

$$life\ expectancy = 43.0 + 0.422 \times HDI$$

The coefficient of determination is $r^2 = 0.875$

Question 9

Given the information above, which one of the following statements is **not** true?

A. The value of the correlation coefficient is close to 0.94

B. 12.5% of the variation in life expectancy is not explained by the variation in the Human Development Index.

C. On average, life expectancy increases by 43.0 years for each 10-point increase in the Human Development Index.

D. Ignoring any outliers, the association between life expectancy and the Human Development Index can be described as strong, positive and linear.

E. Using the least squares line to predict the life expectancy in a country with a Human Development Index of 75 is an example of interpolation.

VCAA Exam Question on this concept
2018 Paper 1

Question 10

In a study of the association between a person's *height*, in centimetres, and *body surface area*, in square metres, the following least squares line was obtained.

$$body\ surface\ area = -1.1 + 0.019 \times height$$

Which one of the following is a conclusion that can be made from this least squares line?

A. An increase of 1 m^2 in *body surface area* is associated with an increase of 0.019 cm in *height*.

B. An increase of 1 cm in *height* is associated with an increase of 0.019 m^2 in *body surface area*.

C. The correlation coefficient is 0.019

D. A person's *body surface area*, in square metres, can be determined by adding 1.1 cm to their *height*.

E. A person's *height*, in centimetres, can be determined by subtracting 1.1 from their *body surface area*, in square metres.

VCAA Exam Question on this concept
2018 Paper 1

Question 14

A least squares line is fitted to a set of bivariate data.

Another least squares line is fitted with response and explanatory variables reversed.

Which one of the following statistics will **not** change in value?

A. the residual values

B. the predicted values

C. the correlation coefficient r

D. the slope of the least squares line

E. the intercept of the least squares line