Fitting a least squares regression line to numerical data

Year 12 General Maths
Units 3 and 4

www.maffsguru.com

Learning Objectives

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

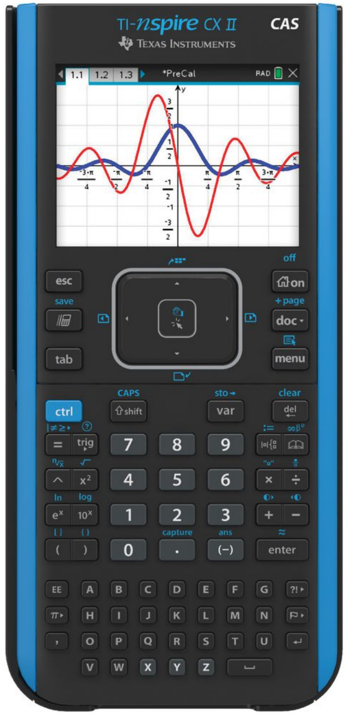
- To be able to define linear regression.
- To be able to define a residual.
- To introduce the least squares line of best fit.
- To be able to find the equation of the least squares line using summary statistics.
- To be able to find the equation of the least squares line using technology.

Recap

This is the first video in this particular section of the course but it builds on the work we have been doing in the previous chapter.

This section of work features prominently in SACs. It is important that you totally understand how to do what will be presented in this lesson and chapter.

It will require the use and full understanding of how to use your CAS.



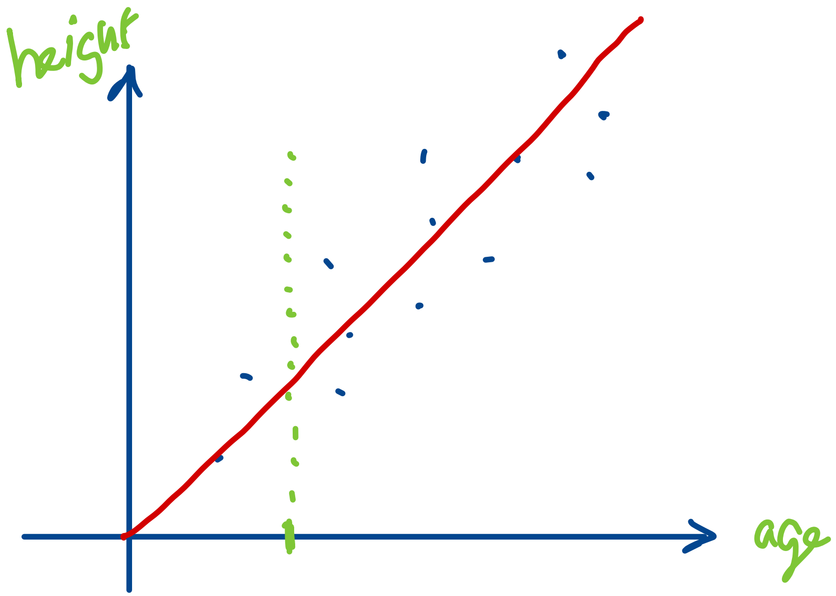
What is linear regression

We have seen, in previous chapters, that we can plot numerical data onto a scatter plot.

When we do this, many times, there will be some form of association between the two variables we have plotted. If we are lucky, there will be a perfect **correlation** between the two variables and the data points will all fall on a straight line.

Other times, there might be some scatter but the points are still generally close to a line.

We can try and model these associations using a process called **linear regression.**



The form of a regression line

We have been told throughout our school journey that straight lines can always be written to look like the following equation:

$$y = mx + c$$

Where *m* stands for the gradient of the line and *c* stands for the y-axis intercept.

In General Maths we write the formula a little differently (but it still has the same meaning):

$$y = a + bx$$

Where *a* is the y-axis intercept and *b* is the gradient of the line

gradient/slope

$$y = a + bx$$

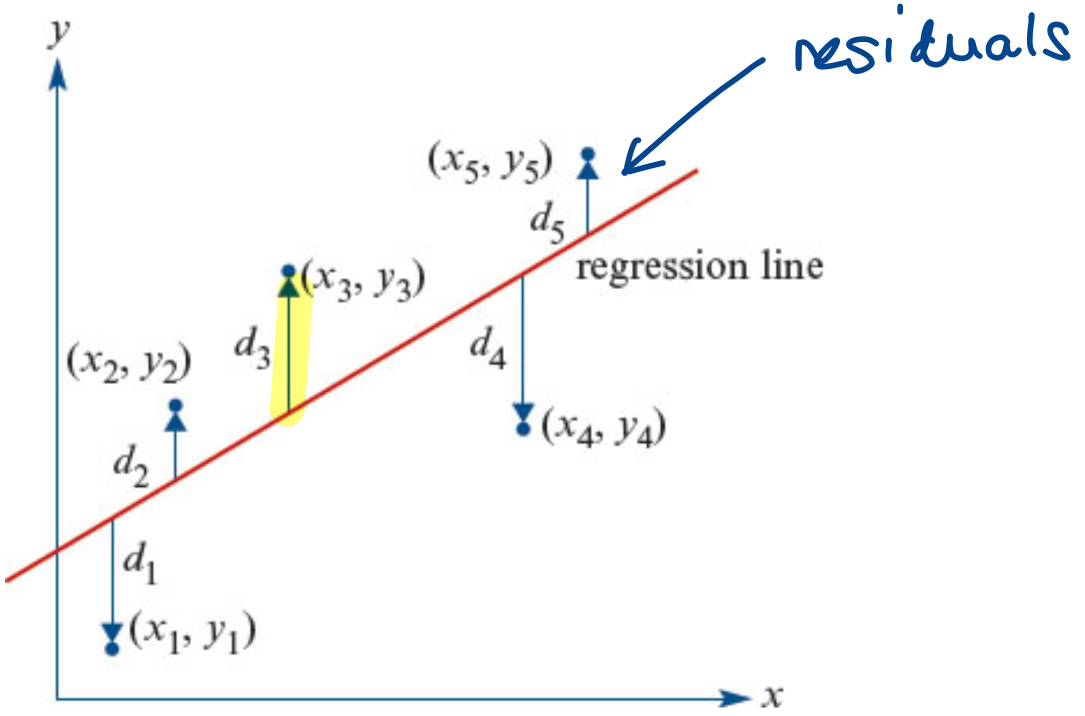
y axis int

The least-squares method

Language
Residual
Positive residual
Negative residual
Zero residual

This is best explained with the use of a diagram as to how the CAS finds the least-squares regression line.

The **least squares line** is the line that minimises the sum of the squares of the residuals.



Assumptions for fitting a least squares line

I have seen this on too many exams and SACs.

For a least squares line to be fitted to a data set we assume the following three things:

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

Finding the equation of the least squares line

The equation of a least squares line can be found using the following two equations:

$$slope, b = \frac{rs_y}{s_x}$$

$$intercept, a = \bar{y} - b\bar{x}$$

Exams have also been know to try and trick you and ask for the value of ‘r’. Hence we can use the following:

$$r = \frac{bs_x}{s_y}$$

Notation:
The following explains each of the letters.

$b\ =\ slope$
$r\ =\ correlation\ coefficient$
$s_y = standard\ deviations\ of\ the\ y\ values$
$s_x = standard\ deviation\ of\ the\ x\ values$
$\bar{x} = mean\ of\ x\ values$
$\bar{y} = mean\ of\ y\ values$

Warning:
You must get the explanatory and response variables the right way around or this will not work!

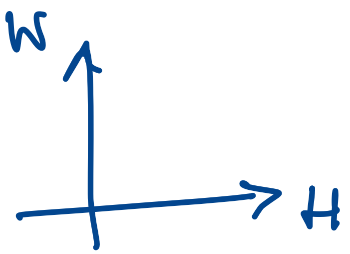
$y = a + bx$

Example

The height and weight of 11 people have been recorded, and the values of the following statistics determined:

	height (x)	weight (y)
mean	173.3 cm	65.45 kg
standard deviation	7.444 cm	7.594 kg
correlation coefficient	$r = 0.8502$	

Use the formula to determine the equation of the least squares regression line that enables weight to be predicted from height. Calculate the values of the slope and intercept rounded to two decimal places.



$y = a + bx$

$b = \dfrac{r . s_y}{s_x}$

$b = \dfrac{0.8502 \times 7.594}{7.444}$

$= \underline{\underline{0.87}}$

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

$= 65.45 - 0.87 \times 173.3$

$= \underline{\underline{-85.32}}$

Weight $= -85.32 + 0.87 \times$ Height

Using the CAS to find and graph the equation of least squares regression line

This is a CAS course so it must be possible to do this using the CAS!

The following data give the height (in cm) and weight (in kg) of 11 people.

Height (x)	177	182	167	178	173	184	162	169	164	170	180
Weight (y)	74	75	62	63	64	74	57	55	56	68	72

$h = 160$

$w = 53.92$

Determine and graph the equation of the least squares regression line that will enable weight to be predicted from height. Write the intercept and slope rounded to three significant figures.

$\therefore$ weight $= -84.8 + 0.867 \times$ height

VCAA Questions

VCAA 2021 Further Maths
Exam 1

Question 10

Oscar walked for nine consecutive *days*. The *time*, in minutes, that Oscar spent walking on each day is shown in the table below.

Day	1	2	3	4	5	6	7	8	9
Time	46	40	45	34	36	38	39	40	33

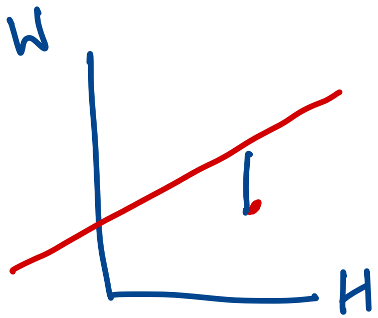
A least squares line is fitted to the data.

The equation of this line predicts that on day 10 the time Oscar spends walking will be the same as the time he spent walking on

A. day 3
B. day 4
C. day 6
D. day 8
E. day 9

$time = 44 - 1.day$

day 10 $\quad time = 44 - 10$
$= 34$



Question 11

The table below shows the *weight*, in kilograms, and the *height*, in centimetres, of 10 adults.

Weight **(kg)**	***Height*** **(cm)**
59	173
67	180
69	184
84	195
64	173
74	180
76	192
56	169
58	164
66	180

A least squares line is fitted to the data.

The least squares line enables an adult's *weight* to be predicted from their *height*.

The number of times that the predicted value of an adult's *weight* is greater than the actual value of their *weight* is

A. 3
B. 4
C. 5
D. 6
E. 7

VCAA 2020 Further Maths
Exam 1

Question 13

A least squares line of the form $y = a + bx$ is fitted to a scatterplot.

Which one of the following is always true?

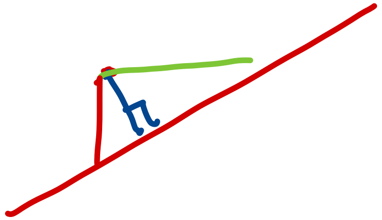
~~A.~~ As many of the data points in the scatterplot as possible will lie on the line.

~~B.~~ The data points in the scatterplot will be divided so that there are as many data points above the line as there are below the line.

~~C.~~ The sum of the squares of the shortest distances from the line to each data point will be a minimum.

D. The sum of the squares of the horizontal distances from the line to each data point will be a minimum.

E. The sum of the squares of the vertical distances from the line to each data point will be a minimum.



VCAA Questions

VCAA 2020 Further Maths
Exam 1

Question 11

A study was conducted to investigate the effect of drinking *coffee* on sleep.

In this study, the amount of *sleep*, in hours, and the amount of *coffee* drunk, in cups, on a given day were recorded for a group of adults.

The following summary statistics were generated.

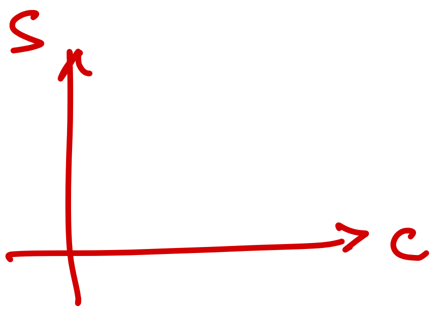
	y ***Sleep* (hours)**	x ***Coffee* (cups)**
Mean	7.08	2.42
Standard deviation	1.12	1.56
Correlation coefficient (r)	–0.770	

On average, for each additional cup of *coffee* drunk, the amount of *sleep*

A. decreased by 0.55 hours.
B. decreased by 0.77 hours.
C. decreased by 1.1 hours.
D. increased by 1.1 hours.
E. increased by 2.3 hours.

$y = a + bx$

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



$b = \frac{-0.770 \times 1.12}{1.56}$

$= -0.5528 \ldots$

www.maffsguru.com