Understanding the regression line:

Example: A study was done to investigate the relationship between the age x in years of a younger person and the time y in minutes in which the child can run one kilometer. The equation of the regression line was found to be $y = 20 - \frac{1}{2}x$. ← In stats they use $y = b + mx$ (it's the same as $y = mx + b$)

Interpret the slope and y-intercept.

The slope as it relates to this problem is as a child gets one year older their time to run one km decreases by ½ a minute or 30 seconds.

The y-int is not relevant. It means that a person 0 yrs old can run a km in 20 minutes. That is impossible!

> The y-intercept is the height of the line when $x = 0$, and might not always have a meaning. Be careful with you interpretation of the intercept. Sometime the value $x = 0$ is impossible or represents a dangerous extrapolation outside the range of the data.

Example: A biologist wants to study the relationship between the number of trees x per hectare and the number of birds y per hectare. She calculates the equation of the regression line to be $y = 8 + 5.4x$. State the gradient and the y-intercept them.

The gradient says that for every 1 tree per hectare the population of birds increases by 5.4

The y-int says if there are no trees then there are 8 birds per hectare.

> NOTE that all these interpretations follow a pattern:
>
> The **gradient** of the line is the amount by which y increases when x increases by 1 unit. (slope! but always over 1 unit)

Exercise 10d

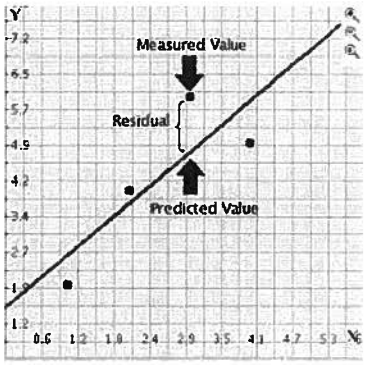
10.3 Least squares regression

The term regression is used in statistics quite differently to other contexts.

- Regression is used for many sorts of curve fitting.
- We can construct a scatter diagram to illustrate the data, find a mean point and draw a line of best fit (regression line) through the mean point.
- Inaccuracies occur because we only have one point to draw the line through and the line of best it drawn "by eye".

There is another way to improve our line, involving residuals.

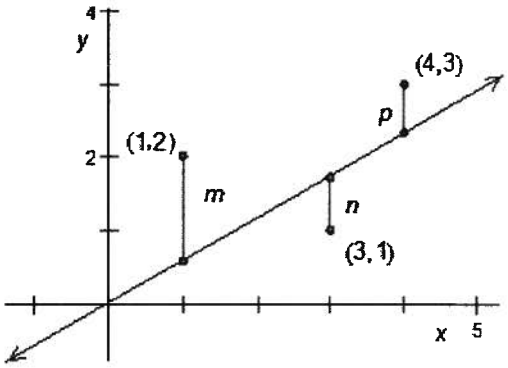
- A **residual** is the vertical distance between a data point and the graph of a regression equation.
- The residual is positive if the data point is above the graph.
- The residual is negative if the data point is below the graph.
- The residual is 0 when the graph passes through the data point.



The equation of the regression line of y on x

The least squares regression line uses our previous formula $y - y_1 = m(x - x_1)$, but now uses the method of least squares to find a suitable value for the slope, m.

- The least squares regression line is the one that has the smallest possible value for the sum of the squares of the residuals.
- In the diagram we aim to make $m^2 + n^2 + p^2$ as close to zero as possible.



- A rather complicated formula emerges:
- The formula for finding the gradient, or slope (m) of a regression line is:

$$m = \frac{S_{xy}}{(S_x)^2}, \text{ where}$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} \text{ and}$$

$$(S_x)^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

Example: Use the least squares regression formula to find the equation of the regression line from the diagram above.

$(\bar{x}, \bar{y}) = (8/3, 2)$

x	y
1	2
3	1
4	3

$m = \frac{S_{xy}}{(S_x)^2} = \frac{1}{\frac{14}{3}} = 3/14$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

$= 17 - \frac{8(6)}{3}$

$S_{xy} = 1$

$(S_x)^2 = \sum x^2 - \frac{(\sum x)^2}{n}$

$26 - \frac{8^2}{3}$

$26 - 21\frac{1}{3}$

$(S_x)^2 = 4\frac{2}{3}$ or $14/3$

$y - \bar{y} = \frac{S_{xy}}{(S_x)^2}(x - \bar{x})$

$y - 2 = 3/14(x - 8/3)$

$y - 2 = 3/14\,x - 4/7$

$y = \frac{3}{14}x + \frac{10}{7}$

* Don't extrapolate (evaluate x's outside your data set)
* Don't predict x's from y's

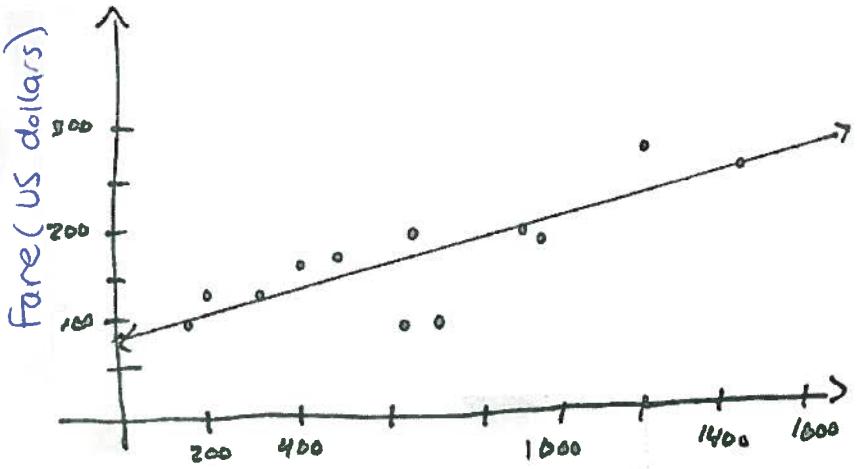
The table shows the distance in km and airfares in US dollars from Changi Airport, Singapore, to twelve destinations.

a. Use your GDC to sketch a scatter diagram of this data with the line of best fit.
b. Write down the equation of your line of best fit.
c. Use your equation to estimate the cost of a 1000 km flight.

Distance	Fare
576	178
370	138
612	94
1216	278
409	158
1502	258
946	198
998	188
189	98
787	179
210	138
737	98

mean pt $712\frac{2}{3}, 166.9\overline{16}$

Distance vs Airfares from Changi Airport, Singapore



b. $y = 0.117x + 83.3$ (3 Sig figs)

c. Cost $= .117(1000) + 83.3$

$= \$200.64$ (exact)

Approx $\$200.30$

* To check that $(\bar{x}, \bar{y})$ is on the line use 2nd calc ① value
Enter $712 + 2/3$
and the y will be $166.9\overline{16}$

Exercise 10E