# MODELLING BY FINDING THE EQUATION OF THE LINE OF BEST FIT

There are many different ways to approximate a **line of best fit** for a series of data points. One of the most useful standard procedures in statistics is the method called **least squares regression** analysis. Whenever you are asked to draw a line of best fit using software, you should usually assume that you are finding the least squares regression line.

## Drawing a line of best fit by eye

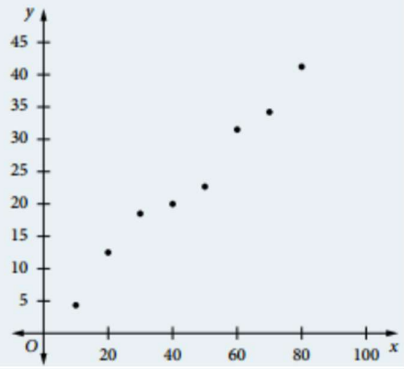
When placing a straight line on a coordinate grid to represent a trend, it is good to aim to have an equal number of points on both sides of the line, with the total distance of points from the line on each side being the same.
**Note:** The line may not necessarily pass through any of the actual points.

Once you have positioned a line of best fit, you can find the gradient, the $y$-intercept and consequently the equation for this line. A line of best fit drawn by eye will be placed in slightly different places by different people, so an answer found in this way can only be considered to be an estimate.

### Example 11

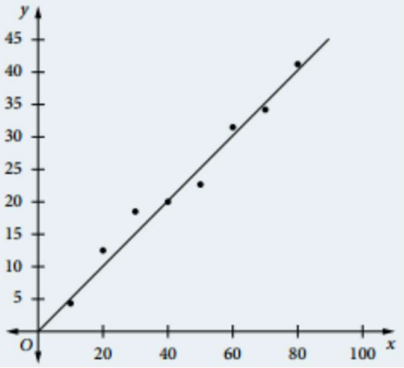
Use the given scatterplot to answer the questions.

**(a)** Place a line of best fit by eye.
**(b)** Calculate the gradient of the line.
**(c)** Find the $y$-intercept.
**(d)** Write the equation of your line of best fit.



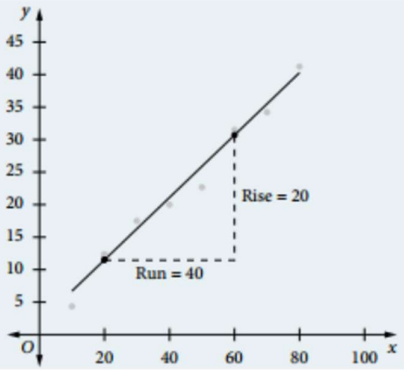

### Solution

**(a)** Use a straight edge to position a line so that there are approximately the same number of points on either side of the line, and so the total distance of the points from the line on both sides is roughly the same.




**(b)** Choose two convenient points from the line and calculate the gradient: Gradient $= \frac{\text{rise}}{\text{run}}$
For this line choose the points (20, 11) and (60, 31).

$$\text{Gradient} = \frac{\text{rise}}{\text{run}}$$
$$= \frac{31-11}{60-20}$$
$$= \frac{20}{40}$$
$$= 0.5$$




**(c)** Remember the general equation for a linear graph is $y = mx + c$.
Use $m = 0.5$ from part **(b)** and substitute either of the points used in the calculation for $m$ to calculate $c$. (20, 11) is used here but (60, 31) would produce the same result.

$$y = mx + c$$
Substitute $m = 0.5$, $x = 20$ and $y = 11$: $11 = 0.5 \times 20 + c$
$$11 = 10 + c$$
$$c = 1$$

If possible, check this result by reading the value of the $y$-intercept directly from the graph:
Extending the line of best fit through to intersect with the $y$-axis, the coordinates of the $y$-intercept are (0, 1).

**(d)** Write the equation by substituting in the values for $m$ and $c$ that have been calculated: $y = mx + c$ becomes $y = 0.5x + 1$.

## Using technology to find the regression equation

You can use technology to find the equation to a line of best fit directly, which will typically use least squares regression analysis. This course does not expect you to understand how a least squares regression line of best fit is calculated, but you should know how to use digital technology to get the equation. Refer to the 'Explore further' activity below for a worked example using digital technology. The least squares regression line is the line that minimises the sum of the squares of the residuals (i.e. the differences between the data points and the line).

### Example 12

The speed and depth of the water flowing at a particular point in a river is recorded at different times.

**(a)** Use software to find the value of $r$.

**(b)** Use software to find the equation of the least squares regression line, $y = a + bx$.

**(c)** Comment on the association using the statistics found.

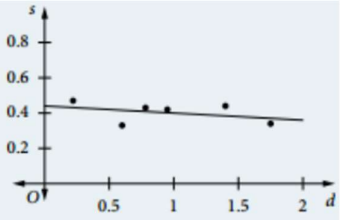
| Depth (m) | Speed ($m\,s^{-1}$) |
|---|---|
| 0.22 | 0.47 |
| 0.6 | 0.33 |
| 0.78 | 0.43 |
| 0.95 | 0.42 |
| 1.40 | 0.44 |
| 1.75 | 0.34 |

### Solution

Using digital technology:

**(a)** $r = -0.3929$

**(b)** Equation of the regression line is $y = 0.4434 - 0.0404x$.
The scatterplot with the line of best fit follows.




**(c)** The depth of water is the independent variable and the speed is the dependent variable. It appears that there is a weak, negative association, $r = -0.39$, linking an increase in water depth to a slowing of the speed.

## Using the regression line to make predictions

The regression equation can be used to model the situation described by the data. This allows you to predict an unknown variable given a known variable.

If the value used for prediction comes from within the boundaries of the original data, then this is called **interpolation** and the prediction should be relatively reliable.

If the regression equation is used to predict a value outside the data boundaries, then this is called **extrapolation** and can sometimes be misleading, as data relationships often do not continue without changing unpredictably at some point.

Generally, the further the regression line prediction is from the original data, the less reliable the result.

To demonstrate this, consider bivariate data collected concerning the shoe size of people under 18 years of age. For this data you can assume a strong, positive linear association: as people get older, their feet get larger.

Using the regression equation to predict the shoe size of a 15-year-old would be interpolation as 15 lies within the age range of the original data, 0–18. Even allowing for individual variation , you would expect this to provide a reasonable level of accuracy.

However, if you were to extrapolate beyond 18 years of age, then the regression line assumes that feet continue to grow larger and larger as people get older, forever—even though, feet generally stop growing before 25 years of age. It would also make no sense to extrapolate the shoe size of a person 500 years old, or to extrapolate anything that involved negative age or negative shoe size, even though the regression line will extend to all these values.

Generally, the further away from the original data the less reliable the prediction will be.

# MODELLING BY FINDING THE EQUATION OF THE LINE OF BEST FIT

## Example 13

It is given that water flow speed = $-0.04 \times$ (depth of water) + 0.44 or $v = -0.04 \times d + 0.44$ where $0.22 \leq d \leq 1.75$. Use this equation to complete the following table.

| Speed ($m\,s^{-1}$) | Depth (m) | Interpolation/Extrapolation |
|---|---|---|
| | 1 | |
| 0.38 | | |
| | 5 | |

## Solution

Find the speed if the depth is 1 m: $s = -0.04 \times d + 0.44$

$$s = -0.04 \times 1 + 0.44$$

$$s = 0.40\ ms^{-1}$$

Depth in the original data ranges from 0.22 m to 1.75 m. As 1 m is within these boundaries, this is interpolation.

Find the depth if the speed is $0.38\ m\,s^{-1}$: $s = -0.04 \times d + 0.44$

$$0.38 = -0.04 \times d + 0.44$$

$$0.38 - 0.44 = -0.04 \times d$$

$$-0.06 = -0.04 \times d$$

$$\frac{-0.06}{-0.04} = d$$

$$d = 1.5\ m$$

Speed in the original data ranges from $0.33\ m\,s^{-1}$ to $0.47\ m\,s^{-1}$. As $0.38\ m\,s^{-1}$ is within these boundaries, this is interpolation.

Find the depth if the speed is $5\ m\,s^{-1}$: $s = -0.04 \times d + 0.44$

$$= -0.04 \times 5 + 0.44$$

$$= 0.24\ m$$

Depth in the original data ranges from 0.22 m to 1.75 m. As 5 m is outside these boundaries, this is extrapolation.

| Speed ($m\,s^{-1}$) | Depth (m) | Interpolation/Extrapolation |
|---|---|---|
| 0.40 | 1 | Interpolation |
| 0.38 | 1.5 | Interpolation |
| 0.24 | 5 | Extrapolation |

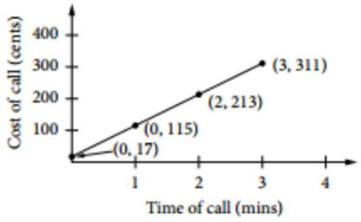
# MODELLING BY FINDING THE EQUATION OF THE LINE OF BEST FIT

## Understanding the gradient and intercept values

In the general linear equation $y = mx + c$, the value of $m$ is the gradient or slope of the line and $c$ represents the $y$-intercept. In modelling situations, these values can often be linked to a practical understanding of the data.

Consider this situation. The cost of phone call with a particular company is a 17 cent connection fee and then an additional fee of 98 cents per minute.

This can be shown graphically as:



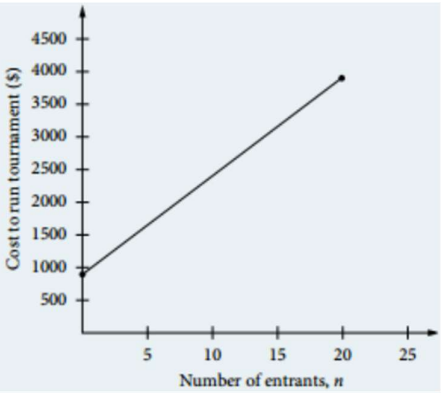

(This assumes that the cost for partial minutes is a proportion of the cost per minute, so that the cost increases steadily as time increases. The graph would be different if the charges were applied in blocks of a minute or 30 seconds, in which case it would look like a step graph.)

The equation of the graph is: Cost of call (cents) = 98 × time of call (minutes) + 17

Note that the $y$-intercept (17) represents the fixed costs and the gradient (98) represents the cost per minute of the call.

### Example 14

Jayde is in charge of arranging a round robin tournament for her local tennis club. Her fixed costs are the cost of hiring the courts and buying the trophies. The extra costs, which depend upon the number of entrants, are for lunch and tennis balls. Due to court space, the maximum number of people who can enter is 48. The graph below shows the total cost of running the event depending upon the number of entrants.




**(a)** Find the cost of hiring the courts and buying the trophies.

**(b)** Find the extra cost for every entrant.

**(c)** Write the equation and find the cost for 25 entrants.

### Solution

**(a)** Read the $y$-intercept from the graph: 900
The fixed costs are $900.

**(b)** Find the gradient: $\text{Gradient} = \dfrac{\text{rise}}{\text{run}}$

$$\text{Gradient} = \frac{2400 - 900}{10 - 0}$$
$$= \frac{1500}{10}$$
$$= 150$$

The extra cost per entrant is $150.

**(c)** Write a linear equation for the situation: $y = mx + c$

$$\text{cost}(\$) = 150 \times (\text{number of entrants}) + \$900$$
$$c = 150n + 900$$

Substitute into the given equation and find the value of the missing variable: $n = 25$

$c = 150n + 900$
$= 150 \times 25 + 900$
$= \$4650$

If there are 25 entrants then Jayde's total cost is $4650.