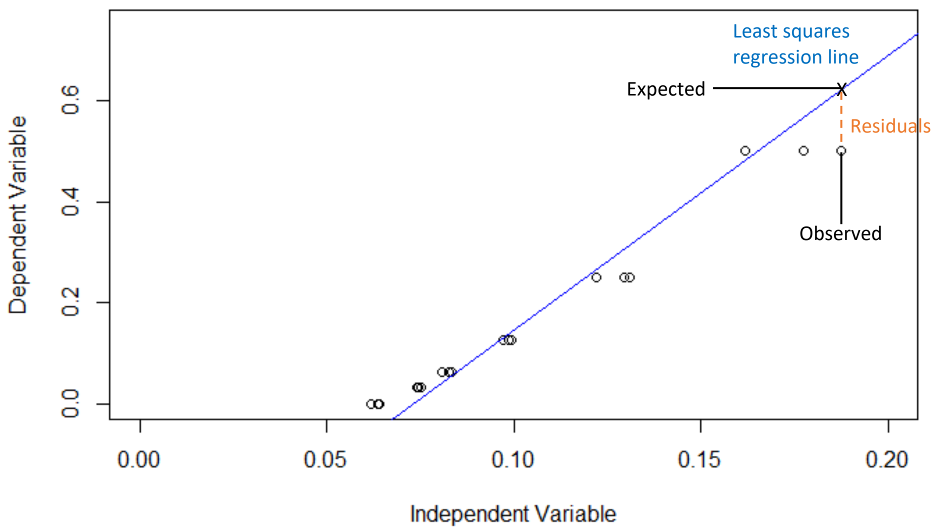
Linear Regression Cheat Sheet

Least Squares Linear Regression

The least squares regression line is a best-fit line which minimises the sum of the residuals squared. A residual is the difference between the observed value and predicted value of the dependent variable.

Example:



The least squares linear regression can be used to estimate the corresponding value of the dependent variable for a value of the independent variable. It can be written in the form:

$$y = ax + b$$

where y = dependent variable
a = gradient of the regression line
x = independent variable
b = y-intercept

a and b can be calculated using the following:

$$b = \frac{S_{xy}}{S_{xx}}$$
$$a = \bar{y} - b\bar{x}$$
$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$
$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

Note: the above formulae are all given in the formula booklet, so you don't need to memorise them, but you do need to know how to use them! You will also need to know how to interpret the equation of your regression line in the context of the question.

Interpolation and Extrapolation

- Interpolation is when you predict the dependent variable for a value of independent variable which is in the range of the data available.
- Extrapolation is when predictions for a value outside the range of the data available.
- Extrapolation should not be used with least squares regression as the prediction will not be reliable.

Example 1: A researcher has carried out an experiment to determine how the rate of a reaction is affected by temperature. The temperature, x, and the time taken for the reaction to be completed, y, are recorded in the table below.

Temperature (°C)	22.0	23.0	24.0	25.0	26.0	27.0	28.0	29.0
Time taken (s)	43.2	41.6	39.9	38.1	36.5	33.6	31.7	30.1

You can use $\Sigma x = 204$, $\Sigma x^2 = 5244$, $\Sigma y = 294.7$, $\Sigma y^2 = 11012.53$ and $\Sigma xy = 7434$.

i) Find the equation of the regression line.

Find S_{xx}.	$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$ $S_{xx} = 5244 - \frac{(204)^2}{8}$ $= 42$
Find S_{xy}.	$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$ $S_{xy} = 7434 - \frac{204 \times 294.7}{8}$ $= -80.85$
Find b.	$b = \frac{S_{xy}}{S_{xx}} = \frac{-80.85}{42} = -1.925$
Find $\bar{x}$ and $\bar{y}$.	$\bar{x} = \frac{\Sigma x}{n} = \frac{204}{8} = 25.5$ $\bar{y} = \frac{\Sigma y}{n} = \frac{294.7}{8} = 36.8375$
Find a.	$a = \bar{y} - b\bar{x}$ $= 36.8375 - (-1.925)25.5$ $= 85.925$
Write the regression equation in the form of $y = a + bx$.	$y = 85.9 - 1.93x$

ii) Interpret what the values of a and b means.

For $a = 85.9$	When temperature (x) is 0, the time required for reaction to be completed is 85.9 seconds.
For $b = -1.93$	For every increase of 1°C, the time taken for reaction to complete decreases by 1.93 second.

iii) Estimate the time needed for the reaction to be completed at 25.5°C and 30°C. Explain whether your predictions are reliable.

For $x = 25.5$	$y = 85.9 - 1.93(25.5)$ $= 36.685$ Answer: 36.7 s
For $x = 30.0$	$y = 85.9 - 1.93(30)$ $= 28$ Answer: 28 s
Explain whether the predictions are reliable.	The prediction for 25.5°C is reliable as it is intrapolated and within the range of the given data. 30°C is outside the range of the given data so prediction is unreliable.

Coded Data

Sometimes data may contain huge numbers. They can be coded so to make calculations easier and less messy. You will be given the coding formula and you will need to know how to transform the original and coded regression line from one another.

Example 2: A data set has been coded using the formulae $x = 3 - s$ and $y = -16t$. The regression line for y on x is $y = 3x + 8$.
I
i) What is the corresponding value of t for a given value of $s = 216$?

Find the value of x when $s = 219$.	$x = 3 - 219$ $= -216$
Find the value of y using the regression line.	$y = 3x + 8$ $= 3(-216) + 8$ $= -640$
Find the value of t.	$y = -16t$ $t = \frac{y}{-16}$ $= \frac{-640}{-16}$ $= 40$

ii) Find the regression line of t on s.

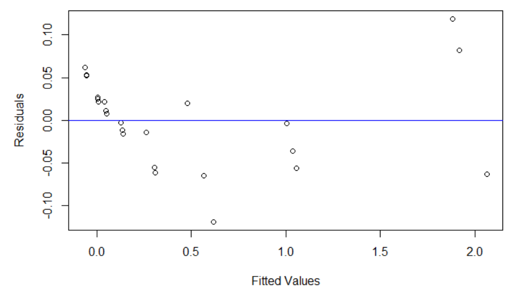
Substitute y and x in the regression line.	$y = 3x + 8$ $-16t = 3(3 - s) + 8$ $-16t = 9 - 3s + 8$ $t = \frac{-102 + 18s}{16}$ $t = 1.125s - 6.375$

Residuals

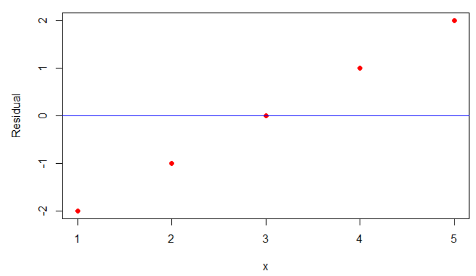
Residuals can be used to assess whether the linear regression is a suitable model and identify any outliers. The equation for the residual of a data point is $y_i - a + (bx_i)$, where x_i and y_i are co-ordinates of the data point and a and b are values from the regression equation. The sum of residuals of all the data points should always add up to 0.

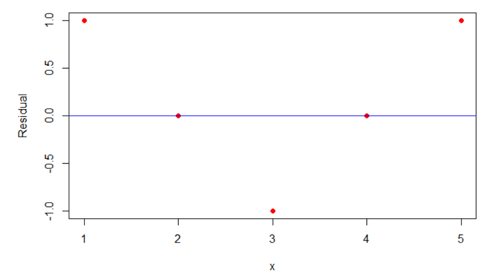
The residuals can be plotted on a graph and visually assessed if a linear model is appropriate. A residual plot with a random distribution near 0 indicates a good fit. If the residuals follow a non-random distribution, it shows that the data is not suitably modelled by linear regression.

Example of a residual plot which shows that the linear regression is suitable:



Examples of residual plots which show that the linear regression may not be suitable:





Residual Sum of Squares

The residual sum of squares (RSS) is used to assess whether the data is appropriately modelled by a linear fit. A small RSS indicates that the linear model is a good fit. RSS can be calculated using the formula:

$$\text{RSS} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

Example 3: i) Calculate the residual sum of squares for the data given in Example 1.

From previous calculations:	$S_{xy} = -80.85$ $S_{xx} = 42$
Find S_{yy}.	$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$ $S_{yy} = 11012.53 - \frac{(294.7)^2}{8}$ $= 156.51875$
Find RSS using the formula.	$\text{RSS} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$ $= 156.51875 - \frac{(-80.85)^2}{42}$ $= 0.8825$

ii) A second dataset has an RSS value of 0.43. Explain which is more suitable for a linear model.

State which is more suitable.	The second data set
Explanation.	$0.43 < 0.8825$ A smaller RSS value indicates that the linear model is a good fit.