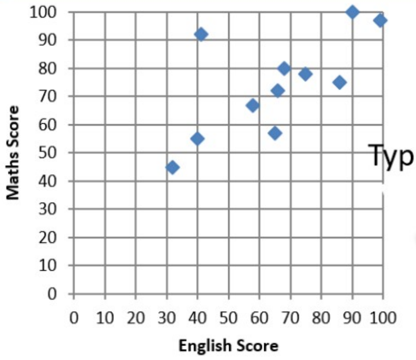
Chapter 4: Correlation

Recap of GCSE

Correlation gives the **strength of the relationship** (and the type of relationship) between two variables. Data with two variables is known as **bivariate data**.

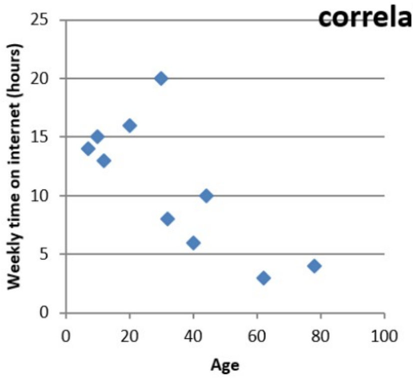


Type of correlation:

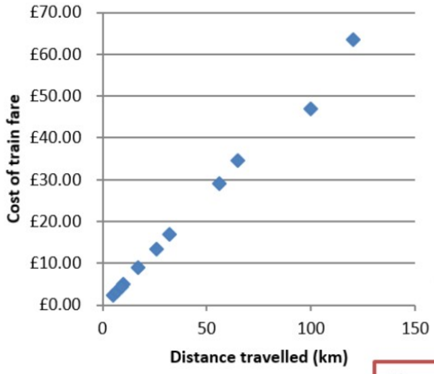
correlation

strength

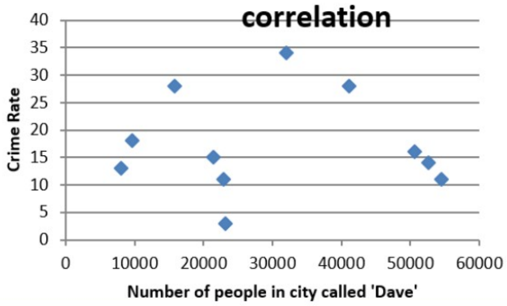
type



correlation



correlation



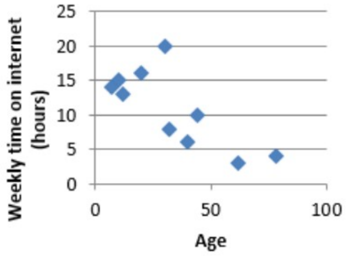
correlation

The vertical-axis variable usually **depends** on the horizontal-axis value. For this reason distance would be the **independent/explanatory variable** and cost the **dependent/response variable**.

Important correlation concepts

Important Point 1

To **interpret** the correlation between two variables is to give a worded description in the context of the problem.

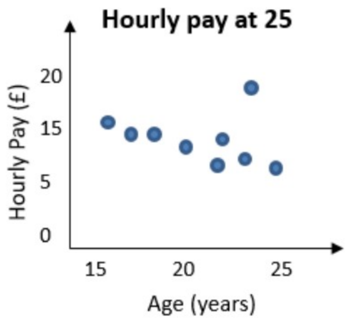


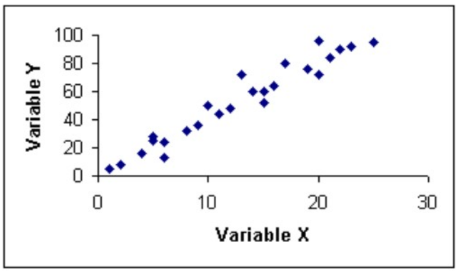
a) State the correlation shown.
b) Describe/interpret the relationship between age and weekly time on the internet.

Important Point 2

Two variables have a **causal relationship** if a change in one variable directly causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

Hideko was interested to see if there was a relationship between what people earn and the age which they left education or training. She says her data supports the conclusion that more education causes people to earn a lower hourly rate of pay. Give one reason why Hideko's conclusion might not be valid.





a) Describe the type of correlation shown.

b) The scatter diagram shown uses the following data: number of fast food restaurants in a town, X, and number of serious road accidents in a town, Y. Interpret the correlation between X and Y.

c) Comment on your answer to part b)

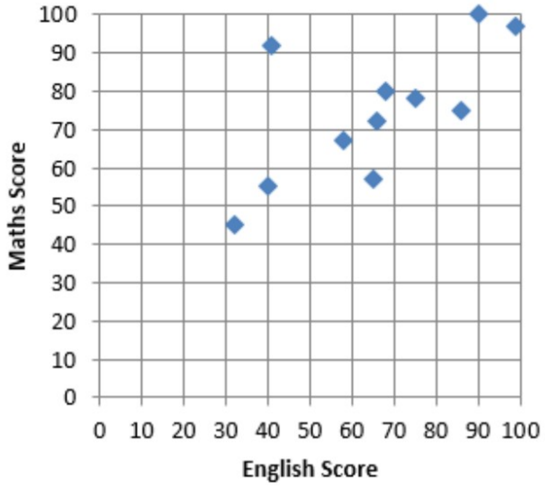
What is regression?

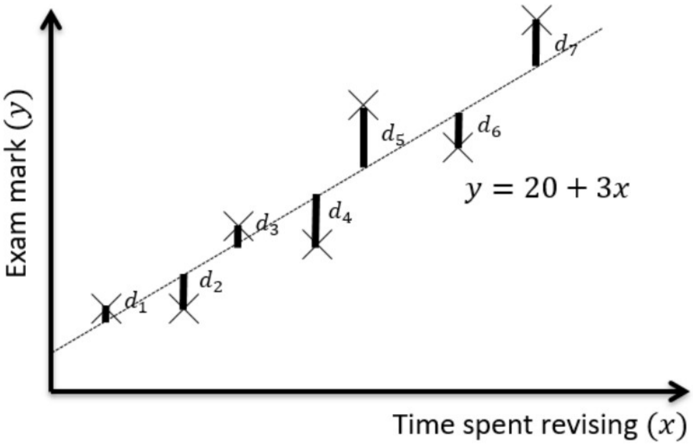
We usually draw a **line of best fit** on a scatter diagram to estimate other values. This line is called the 'least squares regression line', or just the '**regression line**'.

At A-Level, we don't 'guess' where the line of best fit goes at GCSE – instead, it is calculated (the method is no longer in the specification, but our calculators can still do it!).

The regression line will be in the form $\boldsymbol{y = a + bx}$

GCSE style





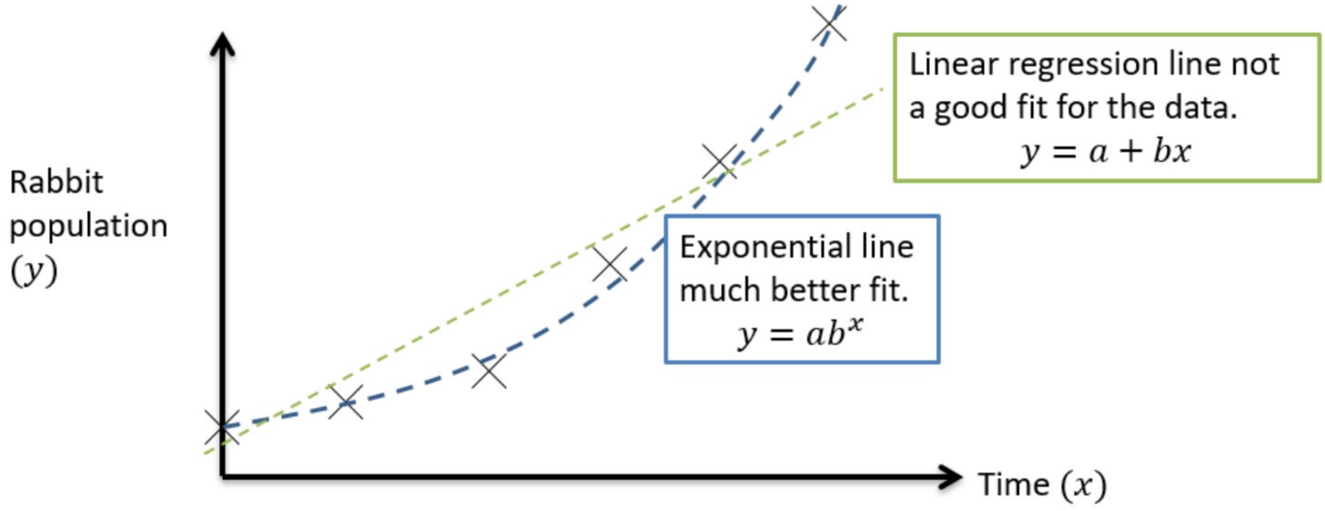
What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set a and b such that the resulting y value matches the actual exam marks as closely as possible.

The 'regression' bit is the act of setting the parameters of our model (here the gradient and y-intercept of the line of best fit) to best explain the data.

One type of line of best fit is the **least squares regression line**. This minimises the sum of the square of these 'errors', i.e.

$$d_1^2 + d_2^2 + \cdots = \Sigma d_i^2$$

Part of the reason we square these errors is so that each distance is treated as a positive value.



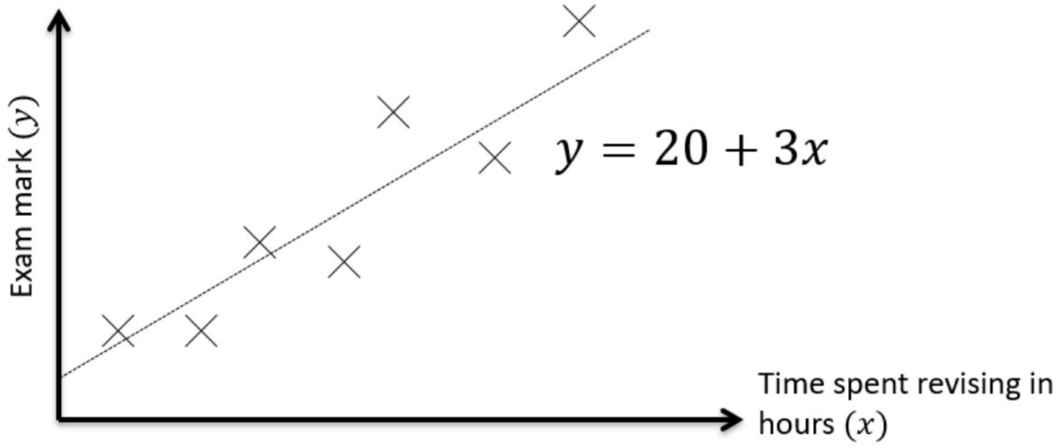
Linear regression line not a good fit for the data. $y = a + bx$

Exponential line much better fit. $y = ab^x$

In this chapter we only cover **linear regression**, where our chosen model is a straight line. If the data does not appear to be in a straight line, then we should not use a linear model.

In general we could use any model that might best explain the data. Population tends to grow exponentially rather than linearly, so we might make our model $y = a \times b^x$ and then try to use regression to work out the best a and b to use. **You will do exponential regression in Chapter 14 of Pure Year 1 and Chapter 1 of Applied Year 2**

Interpreting a and b



How do we **interpret** the y-intercept of 20?

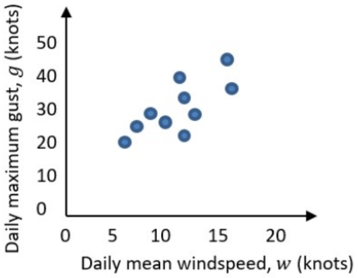
How do we **interpret** the gradient of 3?

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

© Met Office

The data was plotted on a scatter diagram.



(a) Describe the correlation between daily mean windspeed and daily maximum gust.

The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

(b) Give an interpretation of the value of the gradient of this regression line.

(c) Justify the use of a linear regression line in this instance.

The stronger the (linear) correlation, the more suitable a linear regression line is.

Interpolating and Extrapolating

Estimating a value inside the data range is known as **interpolating**.
Estimating a value outside the data range is known as **extrapolating**

You should only use the regression line to make predictions for values **of the dependent variable** that are **within the range of the given data**.

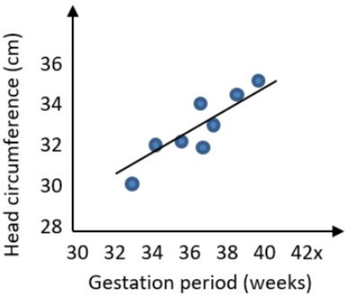
WARNING! **Extrapolation** is **unreliable** and should not be done!

The head circumference, y cm, and gestation period, x weeks, for a random sample of eight newborn babies at a clinic are recorded.
The scatter graph shows the results.
The equation of the regression line of y on x is $y = 8.91 + 0.624x$. The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.

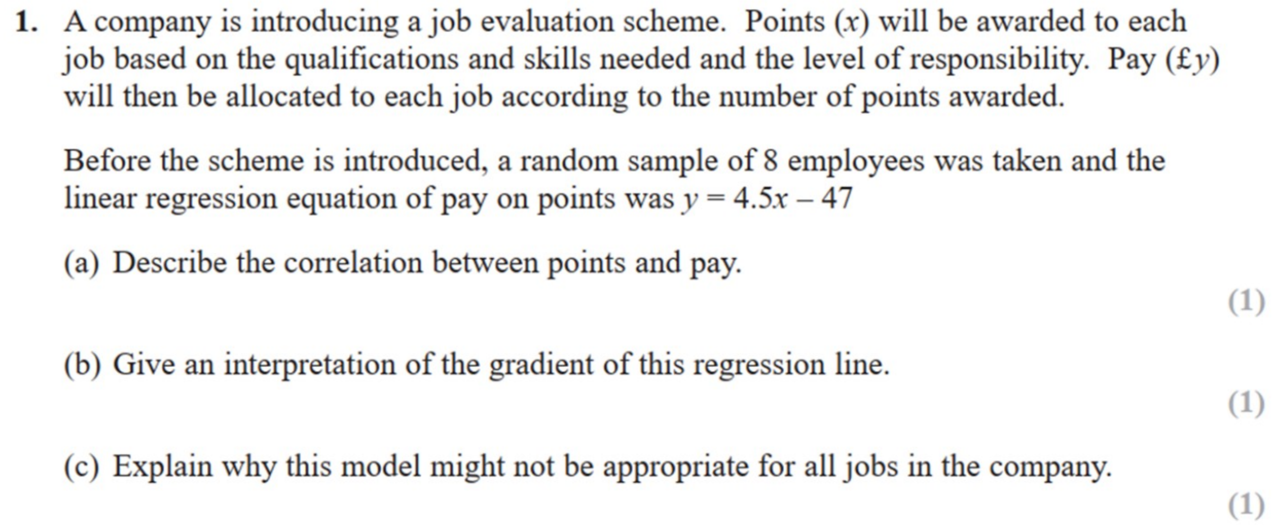
(a) Comment on the reliability of these estimates.

A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.

(b) Explain why the regression equation given above is not suitable for this estimate.



How does this topic come up in exams?



1. A company is introducing a job evaluation scheme. Points (x) will be awarded to each job based on the qualifications and skills needed and the level of responsibility. Pay (£y) will then be allocated to each job according to the number of points awarded.

Before the scheme is introduced, a random sample of 8 employees was taken and the linear regression equation of pay on points was $y = 4.5x - 47$

(a) Describe the correlation between points and pay.

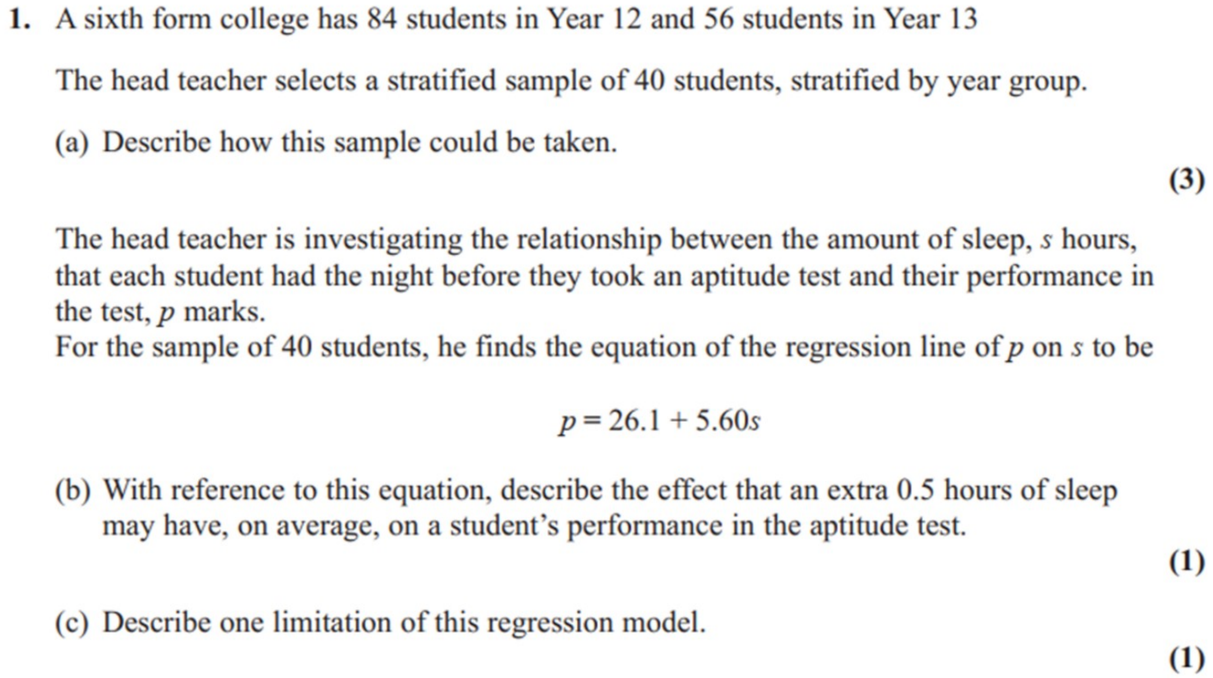
(1)

(b) Give an interpretation of the gradient of this regression line.

(1)

(c) Explain why this model might not be appropriate for all jobs in the company.

(1)



1. A sixth form college has 84 students in Year 12 and 56 students in Year 13

The head teacher selects a stratified sample of 40 students, stratified by year group.

(a) Describe how this sample could be taken.

(3)

The head teacher is investigating the relationship between the amount of sleep, s hours, that each student had the night before they took an aptitude test and their performance in the test, p marks.
For the sample of 40 students, he finds the equation of the regression line of p on s to be

$$p = 26.1 + 5.60s$$

(b) With reference to this equation, describe the effect that an extra 0.5 hours of sleep may have, on average, on a student's performance in the aptitude test.

(1)

(c) Describe one limitation of this regression model.

(1)

Large Data Set example

3. Pete is investigating the relationship between daily rainfall, w mm, and daily mean pressure, p hPa, in Perth during 2015. He used the large data set to take a sample of size 12. He obtained the following results.

p	1007	1012	1013	1009	1019	1010	1010	1010	1013	1011	1014	1022
w	102.0	63.0	63.0	38.4	38.0	35.0	34.2	32.0	30.4	28.0	28.0	15

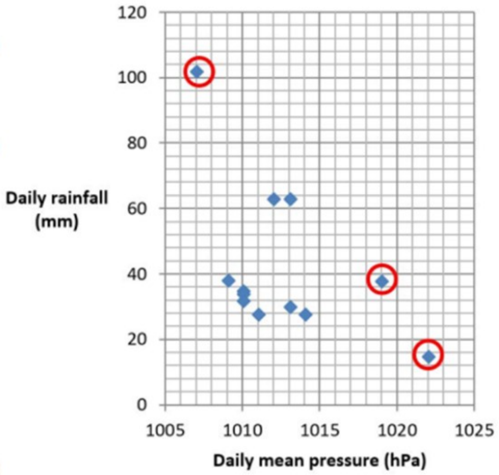
	Q_1	Q_2	Q_3
p	1010	1011.5	1013.5
w	29.2	34.6	50.7

Pete drew the following scatter diagram for the values of w and p and calculated the quartiles.

An outlier is a value which is more than 1.5 times the interquartile range above Q3 or more than 1.5 times the interquartile range below Q1.

(a) Show that the 3 points circled on the scatter diagram above are outliers.

(2)



(b) Describe the effect of removing the 3 outliers on the correlation between daily rainfall and daily mean pressure in this sample.

(1)

John has also been studying the large data set and believes that the sample Pete has taken is not random.

(c) From your knowledge of the large data set, explain why Pete's sample is unlikely to be a random sample.

John finds that the equation of the regression line of w on p, using all the data in the large data set, is

$$w = 1023 - 0.223p$$

(d) Give an interpretation of the figure –0.223 in this regression line.

(1]

John decided to use the regression line to estimate the daily rainfall for a day in December when the daily mean pressure is 1011 hPa.

(e) Using your knowledge of the large data set, comment on the reliability of John's estimate.

(1)