

Summary of key points

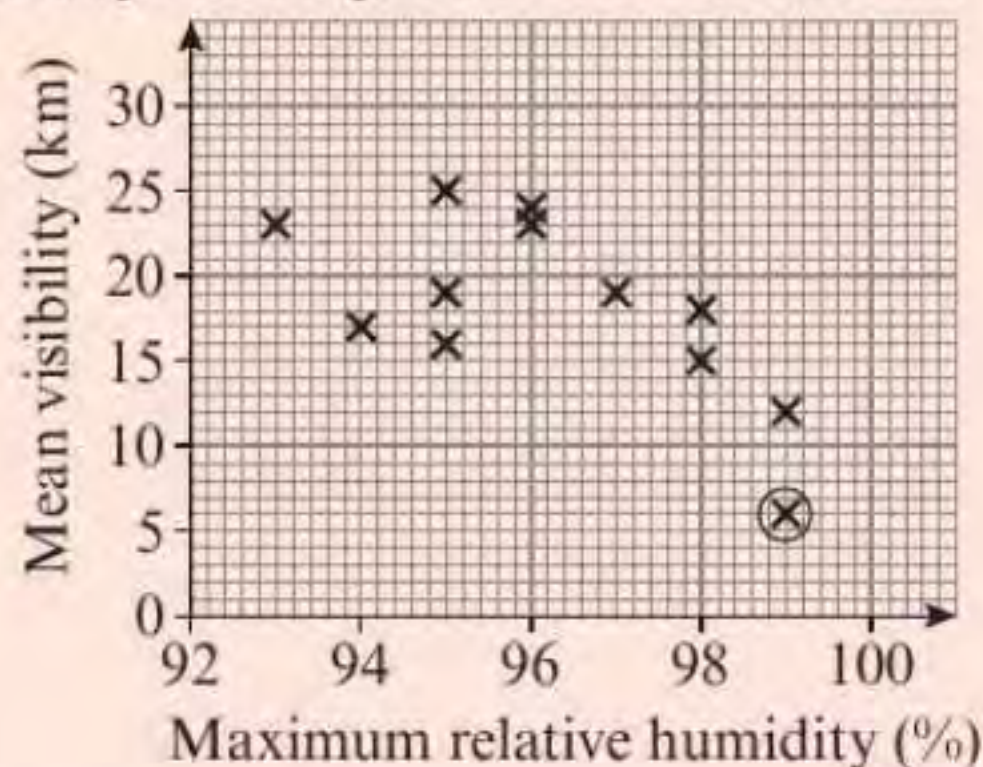
- 1 Bivariate data** is data which has pairs of values for two variables.
- 2 Correlation** describes the nature of the linear relationship between two variables.
- 3** When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.
- 4** The **regression line** of y on x is written in the form $y = a + bx$.
- 5** The coefficient b tells you the change in y for each unit change in x .
 - If the data is positively correlated, b will be positive.
 - If the data is negatively correlated, b will be negative.
- 6** 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.

Correlation and cleaning data

You can use correlation to describe a **linear relationship** between two variables. The closer the data approximates to a linear relationship, the **stronger** the correlation.

Worked example

The scatter diagram shows the maximum relative humidity, and mean visibility for the first 12 days in July, 2015 in Camborne.



- (a) Describe and interpret any correlation shown on the scatter diagram. (2 marks)

Quite strong negative correlation. As humidity increases, visibility decreases.

- (b) From your knowledge of the large data set, explain whether the relationship between humidity and visibility might be causal. (2 marks)

High humidity can give rise to misty and foggy conditions, which can reduce visibility, so the relationship could be causal.

The visibility for the circled data point is found to be an outlier.

- (c) Justify the inclusion of this outlier when analysing the data. (1 mark)

A visibility of 6 km is not unlikely, and follows the trend of the data.

Cause and effect

Watch out! Correlation doesn't always mean that two variables are related.



Bottled water doesn't cause bee stings but, when the weather is hotter, bottled water sales and bee stings both increase. This is an example of a non-causal relationship.

You need to understand what the different variables in the large data set represent, and how they might be related.

There is more on the large data set on page 60.

Problem solved!

Removing obviously incorrect data values is sometimes called **cleaning data**. You might want to **keep** an outlier if:

- it appears to be genuine
- the data is from a reliable source, or has already been cleaned.

You might want to **remove** an outlier if:

- it is obviously a mistake
- you want to intentionally omit extreme data values

You will need to use problem-solving skills throughout your exam – **be prepared!**



Now try this

Michelle is training a group of young people to use image-editing software. She records the time, y hours, it takes each person in the group to reach a certain standard of proficiency.

Group member	A	B	C	D	E	F	G	H	I	J
Age (x years)	15	19	81	16	20	17	19	21	18	15
Time (y hours)	14	13	10	12	7	18	9	6	9	11

Michelle determines that $x = 81$ is an outlier, and decides to omit the data for group member C from her results.

- (a) Comment on the validity of Michelle's decision to omit the outlier. (2 marks)
- (b) Draw a scatter diagram for the remaining 9 results. (3 marks)
- (c) Describe the correlation shown on your scatter diagram, and interpret this in the context of the model. (2 marks)

Regression

A **regression line** on a scatter diagram is a type of line of best fit. It can be used as a **linear model** for the relationship between the two variables.

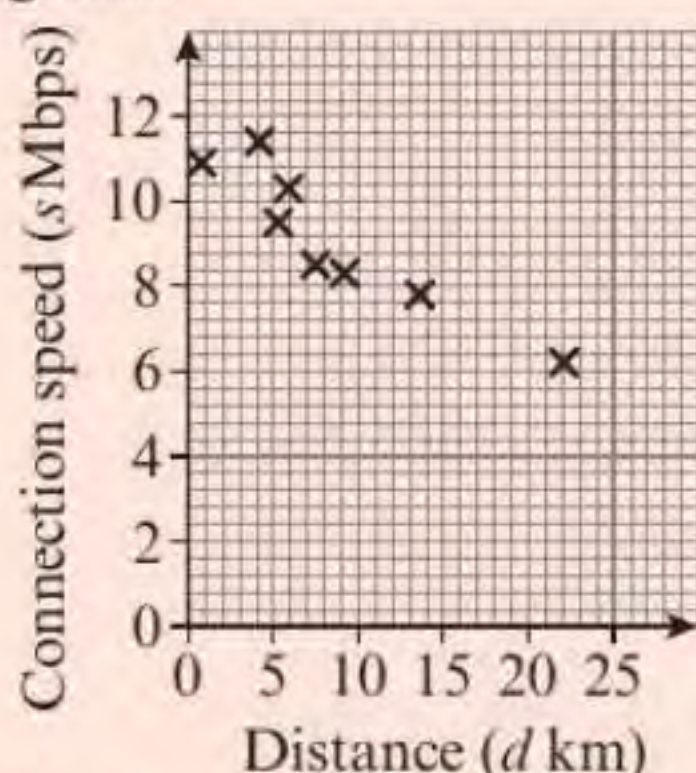
You don't need to be able to calculate regression lines in your exam, but you do need to be able to **interpret** their equations. The regression line (or **least squares regression line**) of y on x is written as:

$$y = a + bx$$

The coefficient b tells you the **gradient** of the regression line – this is how much the variable y increases for each unit increase in x .

Worked example

An internet provider advertises an internet connection speed of 12 Mbps. Dhevan records the speed of 8 internet connections, s Mbps, at different distances, d km from the phone exchange. He records his results on a scatter diagram:



The equation of the regression line of s on d for these data is $s = 11.2 - 0.242d$

- (a) Give an interpretation of the gradient of the regression line. (1 mark)

According to this model, for each additional km from the phone exchange, your connection speed will reduce by 0.242 Mbps.

- (b) Explain why a regression model of the form $s = a + bd$ is supported for these data. (1 mark)

The data show strong negative correlation. Hence a linear model is suitable.

You could also consider whether the relationship is likely to be causal. Distance could affect connection speed, so this supports a relationship between the variables.

Dependent and independent

The order of the variables in a regression equation is important.

In the regression model $y = a + bx$, x is the **independent** (or **explanatory**) variable, and y is the **dependent** (or **response**) variable. In an experiment, x would be the variable you changed, and y would be the variable you recorded. You always plot the explanatory variable on the **horizontal axis**.

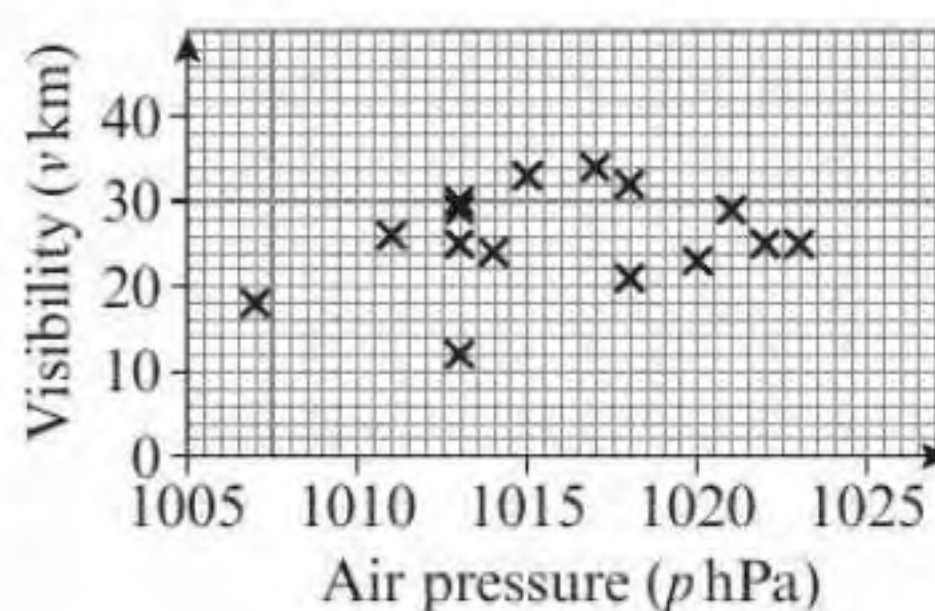
If the value of b (the gradient) in the regression equation is **negative**, then the regression line slopes **downwards**. This means that as one variable increases, the other decreases.

Validity

The **strength of the correlation** between two variables tells you how accurately a **linear model** represents the relationship between them. The stronger the correlation, the more closely the regression line models the data.

Now try this

The scatter diagram shows the daily mean visibility, v km, and daily mean air pressure p hPa, in Hurn for the first 15 days of August 2015.



The equation of the regression line of v on p for these data is $v = -353.5 + 0.373p$

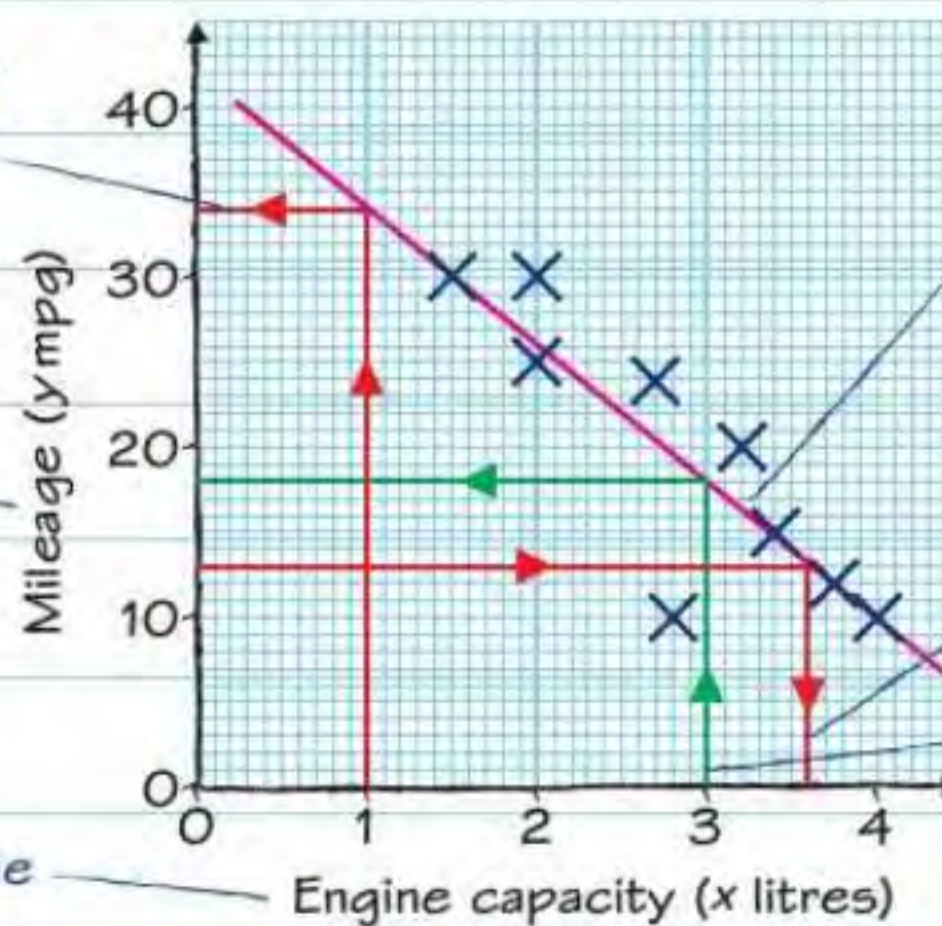
- (a) Interpret the value 0.373 in this model. (1 mark)
- (b) Comment on the validity of a linear regression model for these data. (1 mark)

Using regression lines

If you know the value of the **independent** variable, you can use a regression line (or its equation) to **estimate** or **predict** the corresponding value of the **dependent** variable. You won't need to make predictions like this in your exam, but you do need to know when they are **reliable**. This scatter diagram shows the mileage in miles per gallon (mpg) of 9 cars, and their engine capacities.

X $x = 1$ is **outside the range** of the data. This is called **extrapolation** and it produces an **unreliable estimate**.

Dependent (response) variable



The equation of the regression line of y on x is $y = 42.7 - 8.21x$.

X You can only estimate a value of the **dependent** variable. You can't use the regression line of y on x to estimate a value of x given a value of y .

✓ $x = 3$ is **within the range** of the data. This is called **interpolation** and it produces a **reliable estimate**.

Independent (explanatory) variable

Worked example

The age, x years, and shell size, y mm, of 8 Dungeness crabs were recorded.

y	151.8	150.4	140.3	133.4	155.9	153.3	141.6	127.7
x	3.3	3.0	2.4	2.3	3.3	3.0	2.7	2.2

The equation of the regression line of y on x for these data is $y = 83.17 + 22.03x$.

(a) Give an interpretation of the gradient of the regression line. **(1 mark)**

A typical crab shell will grow about 22 mm per year.

(b) Comment on the reliability of using this regression equation to estimate the shell size of a Dungeness crab with age:

(i) 1.6 years (ii) 2.5 years **(2 marks)**

(i) 1.6 years is **outside the range** of the data, so this is an example of **extrapolation**. The estimate will be **unreliable**.

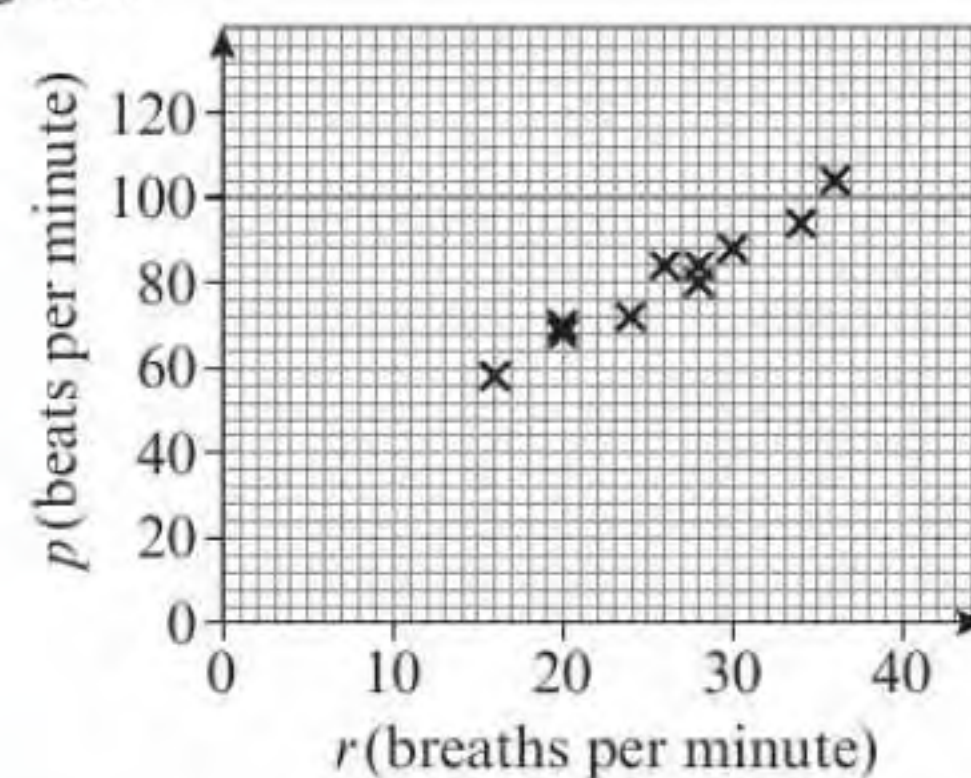
(ii) 2.5 years is **inside the range** of the data, so this is an example of **interpolation**. The estimate will be **reliable**.

(d) Give a reason why this regression model would not be suitable for estimating the age of a Dungeness crab with a shell size of 145 mm. **(1 mark)**

Age is the independent variable in this model. You should only use a regression model to estimate values of the **dependent** variable.

Now try this

In a study, participants exercise for 5 minutes then record their breath rate (r breaths per minute) and pulse (p beats per minute). The data for 10 participants are shown in the scatter diagram.



The equation of the regression line of p on r for these data is $p = 25.5 + 2.09r$

(a) Give a reason why a linear regression model is likely to be suitable for these data. **(1 mark)**

(b) The equation of the regression line is used to estimate the pulse rate of a participant making

(i) 22 breaths per minute

(ii) 10 breaths per minute.

Comment on the reliability of these estimates. **(2 marks)**

You would need to use the regression line of x on y to estimate age given shell size.