Correlation

SCAN ME

REVISE THIS TOPIC

1 The manager at a car dealership investigates the relationship between a car's age, A years, and its value, £V.

In the past 12 months the car dealership sold 1270 cars.
The manager takes a sample of 50 cars from all cars that the dealership has sold in the past 12 months.

(a) Describe how the manager could produce a sample of size 50 using simple random sampling. **(3)**

For the sample of 50 cars the manager finds the equation of the regression line of V on A to be

$$V = 30000 - 2500A$$

(b) Describe the correlation between the age of the car and its value. **(1)**

(c) Give an interpretation of the gradient of this regression line. **(1)**

(d) Describe one limitation of this regression model. **(1)**

(a) Number each of the cars sold from 1 to 1270.

Randomly generate numbers between 1 and 1270 e.g. using a calculator

Ignore any repeated numbers and continue until 50 different numbers have been selected.

Include the cars corresponding to these 50 randomly generated numbers in the sample.

(b) Negative correlation.

(c) For each extra year of age the car's value decreased by £2500

The car's value decreases by £2500 per year.

(d) For cars with an age of greater than 12 years the value would be negative, which does not make sense.

(Total for Question 1 is 6 marks)

2 Josh investigates the relationship between a person's age, A years, and their reaction time, R seconds.

Josh takes a sample of 60 people from his town. He asks their age and tests their reaction time.
To ensure he has a range of ages he samples 10 people from each of the following age groups.

20 – 29 years, 30 – 39 years, 40 – 49 years, 50 – 59 years, 60 – 69 years, 70 – 79 years

If the person is from an age group that already has 10 people sampled, he does not include them.
If the person is aged below 20 years or above 79 years, he does not include them.

(a) State the sampling technique used by Josh. **(1)**

Josh uses a linear regression model to model his data.

(b) State, giving a reason, which variable would be the explanatory variable. **(1)**

For the sample of 60 people Josh finds the equation of the regression line of R on A to be

$$R = 0.15 + 0.005A$$

(b) Describe the correlation between the age of the person and their reaction time. **(1)**

(c) Give an interpretation of the gradient of this regression line. **(1)**

Josh uses this model to estimate the reaction time of a person aged 10.

(d) Comment, giving a reason, on the reliability of this estimate. **(1)**

(a) Quota sampling.

(b) Age is the explanatory variable as reaction time is likely to depend on age.

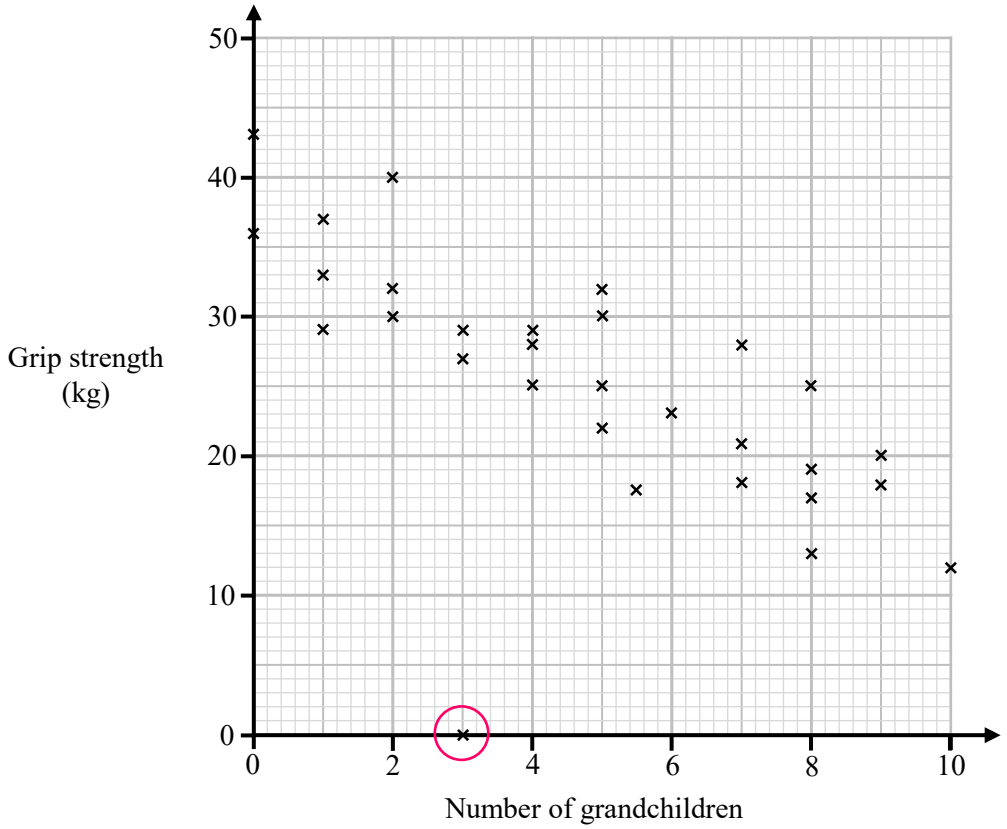
(c) For each extra year of age the reaction time increases by 0.005 seconds.
Reaction time increases by 0.005 seconds per year.

(d) The ages in the sample are from 20 to 79 years.
10 years is outside of the range of the sample (too low) therefore this is extrapolation.
This makes the estimate unreliable.

(Total for Question 2 is 5 marks)

3 Kareem is investigating whether there is a linear relationship between grip strength, in kilograms and the number of grandchildren that a person has.

Kareem selects a sample of 30 people and draws the scatter diagram below using the data.



Kareem says: "Having more grandchildren causes a reduction in grip strength".

(a) Comment on Kareem's claim. **(1)**

One of the points on the scatter diagram is an anomalous result.

(b) Circle the point and explain how you know that it is anomalous. **(1)**

Before doing further calculation Kareem decides to clean the data.

(c) Explain what is meant by cleaning the data. **(1)**

(a) The presence of correlation does not imply causation.

Both are likely to correlate well as they would both also correlate with age.

(b) It does not make sense to have 0 kg grip strength.

(c) Removing anomalous values from the dataset.

(Total for Question 3 is 3 marks)

4 A tyre manufacturer investigates the lifespan of its tyres by collecting data on the distance driven, x miles, and the tyre tread depth, y mm.

The tyre manufacturer samples 1000 tyres and uses a linear regression model to mode the data.

(a) State, giving a reason, which variable would be the response variable. **(1)**

For the sample, the tyre manufacturer finds the equation of the regression line of y on x to be

$$y = 8.0 - 0.00016x$$

(b) Describe the correlation between the distance driven and the tyre tread depth. **(1)**

(c) Give an interpretation of the gradient of this regression line. **(1)**

(d) Give an interpretation of the y-intercept of this regression line. **(1)**

(e) With reference to the equation, describe the effect that driving an extra 500 miles may have, on average, on the tread depth of the tyre. **(1)**

(f) Describe one limitation of this regression model. **(1)**

(a) Tyre tread depth (y) is the response variable as it is likely to depend on the number of miles driven.

(b) Negative correlation.

(c) For each extra mile driven the tread depth decreases by 0.00016 mm

Tread depth decreases by 0.00016 mm per mile driven.

(d) A tyre with 0 miles driven will have a tread depth of 8.0 mm.

A new tyre has tread depth 8.0 mm

(e) $500 \times 0.00016 = 0.08$

Driving an additional 500 miles will result in a decrease of 0.08 mm of tread on the tyre.

(f) Tyres that are driven over 50 000 miles will have a negative tread depth, which does not make sense.

(Total for Question 4 is 6 marks)

5 A athletics coach records the age (x years), 100-metre sprint time (y seconds), and shot-put throw distance (z metres) for a sample of 20 athletes at their athletics club.

The coach uses the data to draw the two scatter diagrams shown below in **Figure 1** and **Figure 2**.

Figure 1

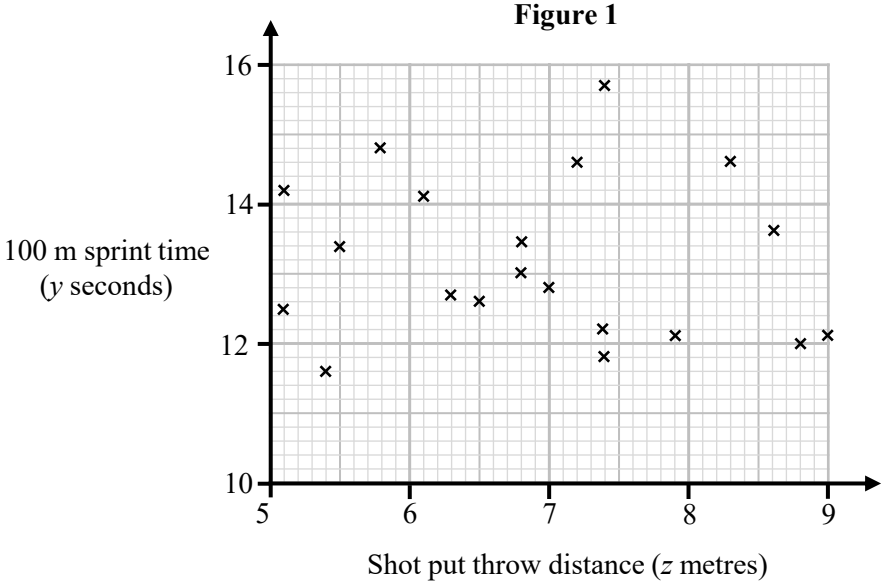
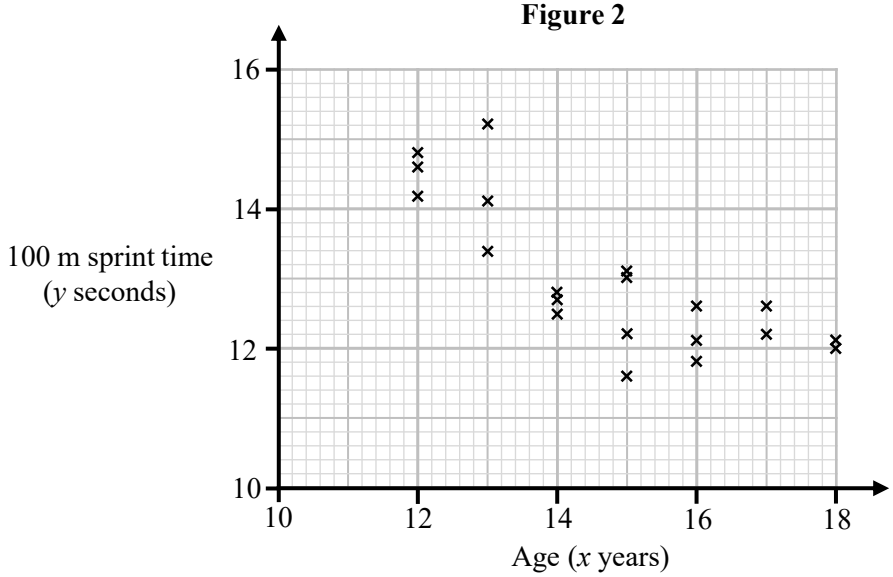


Figure 2



To sample the 20 athletes the coach selected the first 20 athletes who arrived to training.

(a) State the sampling technique used by the coach. **(1)**

(b) Describe the correlation shown in Figure 1. **(1)**

(c) Interpret the correlation shown in Figure 2. **(1)**

The coach uses a linear regression model to model the data for age (y) and 100-metre sprint time (x).

For the 20 athletes sampled the coach finds the equation of the regression line of y on x to be

$$y = 19.5 - 0.45x$$

(d) State, with reason, which variable would be the response variable. **(1)**

(e) Give an interpretation of the gradient of this regression line. **(1)**

(f) State the units of the gradient of this regression line. **(1)**

The coach uses the model to estimate the 100-metre sprint time of an athlete at aged 23.

(g) Comment, giving a reason, on the reliability of this estimate. **(1)**

(a) Opportunity sampling (convenience sampling)

(b) No correlation.

(c) As age increases the 100 m sprint times decrease.

(d) The 100m sprint time is the response variable as it is likely to depend on age.

(e) For each extra year of age the 100 m sprint time decreases by 0.45 seconds.
100 m sprint time decreases by 0.45 seconds per year.

(f) seconds per year.

(f) The ages in the sample are from 12 to 18 years (as seen on the scatter diagram).
23 years is outside of the range of the sample (too high) therefore this is extrapolation.
This makes the estimate unreliable.

(Total for Question 5 is 7 marks)

6 The town of Chippenham is close to Junction 17 of the M4 motorway.

Chris believes that petrol stations that are located nearer to the motorway junction charge more for petrol.

He records the price (y pence/litre) of petrol and distance from the motorway junction (x miles) at all 14 petrol stations within 7 miles of Junction 17.

For the sample of 14 petrol stations Chris finds the equation of the regression line of y on x to be

$$y = 159 - 4.1x$$

(a) State, giving a reason, which variable would be the explanatory variable. **(1)**

(b) Give an interpretation of the gradient of this regression line. **(1)**

(c) State the units of the gradient of this regression line. **(1)**

A new petrol station is going to be built 4 miles away from Junction 17.
Chris uses this model to estimate the price of petrol at this new petrol station.

(d) Comment, giving a reason, on the reliability of this estimate. **(1)**

The town of Royal Wootton Bassett is 9.3 miles away from Junction 17.
Using the model Chris estimates the price of petrol in Royal Wootton Bassett to be 120.9 pence/litre.
The actual price of petrol in Royal Wootton Bassett is 139.9 pence/litre.

(e) Give a possible reason for the difference between the estimate from the model and the actual price of petrol in Royal Wootton Bassett. **(1)**

(a) Distance to the motorway junction is the explanatory variable as the price will likely depend on the distance to the motorway junction.

(b) For each extra mile away from the motorway junction the price of petrol decreases by 4.1 pence/litre.
The price of petrol/litre decreases by 4.1 pence per mile away from the motorway junction.

(c) pence/litre per mile.

(d) This estimate is reliable as the data in the sample ranges from 0 to 7 miles away.

(e) Whilst Royal Wootton Bassett is further away from Junction 17 there may be other variables contributing to the higher than expected price (compared to the model) e.g. it may be nearer to a **different** motorway junction.

(Total for Question 6 is 5 marks)

7 Sumira investigates the relationship between the number of ice cream sales, (x), and number of reported cases of sunburn, (y), at a beach during July 2024.

Rather than taking a sample from the days in July 2024 Sumira decides to complete a census.

(a) Explain what is meant by the term census. **(1)**

For the data Sumira finds the equation of the regression line of y on x to be

$$y = 0.15x - 9$$

(b) Describe the correlation between the number of ice cream sales and the number of reported cases of sunburn. **(1)**

(c) Describe one limitation of this regression model. **(1)**

Sumira says:

"Since there is a strong positive correlation, ice cream sales cause sunburn cases"

(d) Comment on Sumira's claim. **(1)**

(a) When the entire population is included in the sample.

(b) Positive correlation.

(c) If fewer than 60 ice creams are sold the number of sunburn cases is negative, which makes no sense.

(d) The presence of correlation does not imply causation.

Both are likely to correlate well as they would both correlate temperature, which may "cause" both.

(Total for Question 7 is 4 marks)

8 A vet investigates the relationship between the mass of a kitten and the age of the kitten.

The vet records the mass of the kitten every 0.5 months from birth until the kitten is 6 months old.

Using a linear regression model for the data the vet finds the equation of the regression line of M on A to be

$$M = d + cA$$

where

M is the mass of the kitten in kilograms.
A is the age of the kitten in months.
c and d are constants with $c > 0$ and $d > 0$

Using the model, the vet concludes that, on average

- The mass of the kitten increased by 158 grams every 0.5 months
- The mass of a 6-month-old kitten is 2 kg

(a) (i) Find the value of c **(2)**

(ii) Find the value of d **(2)**

(b) State the units of the gradient of the regression line. **(1)**

(c) Explain why it may not be reliable to use this model to estimate the mass of a kitten aged 10 months. **(1)**

(a) (i) $158 \times 2 = 316$ g per month.

$316 \div 1000 = 0.316$ kg per month.

$c = 0.316$

(a) (ii) $2 = d + 0.316 \times 6$

$d = 2 - 0.316 \times 6$

$d = 0.104$

(b) kg/month

(c) The ages of kittens in the sample are 0 to 6 months.

10 months is outside of the range of the sample (too high) therefore this is extrapolation.

This makes the estimate unreliable (for instance the growth rate may slow down as the kitten matures)

(Total for Question 8 is 6 marks)