Reliability of regression predictions

Grade 9 | Worked solutions

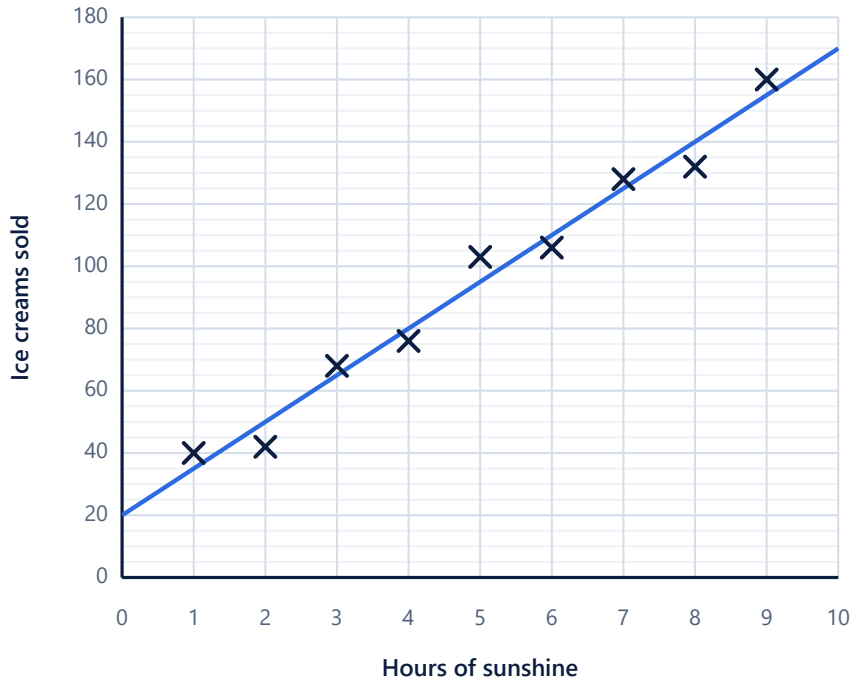
GCSE Statistics topic practice

1 Elif owns an ice cream van. On each of 9 days she records the number of hours of sunshine, x, and the number of ice creams she sells, y. The scatter diagram shows her data and the regression line.

The equation of the regression line is $y = 20 + 15x$

a. Use the equation to estimate the number of ice creams Elif sells on a day with 6 hours of sunshine. (1)

b. Elif also uses the equation to estimate her sales on a day with 12 hours of sunshine. Which of her two estimates is more reliable? Give a reason for your answer. (2)



(3 marks)

a) Put $x = 6$ into the equation

$y = 20 + 15 \times 6$

$y = 20 + 90$

$y = 110$

b) The estimate for 6 hours is more reliable

6 hours is inside the range of her data (1 to 9 hours)

This is interpolation

12 hours is outside the range of her data

This is extrapolation, the pattern may not continue

a) 110 b) 6 hours, it is inside the data range

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.

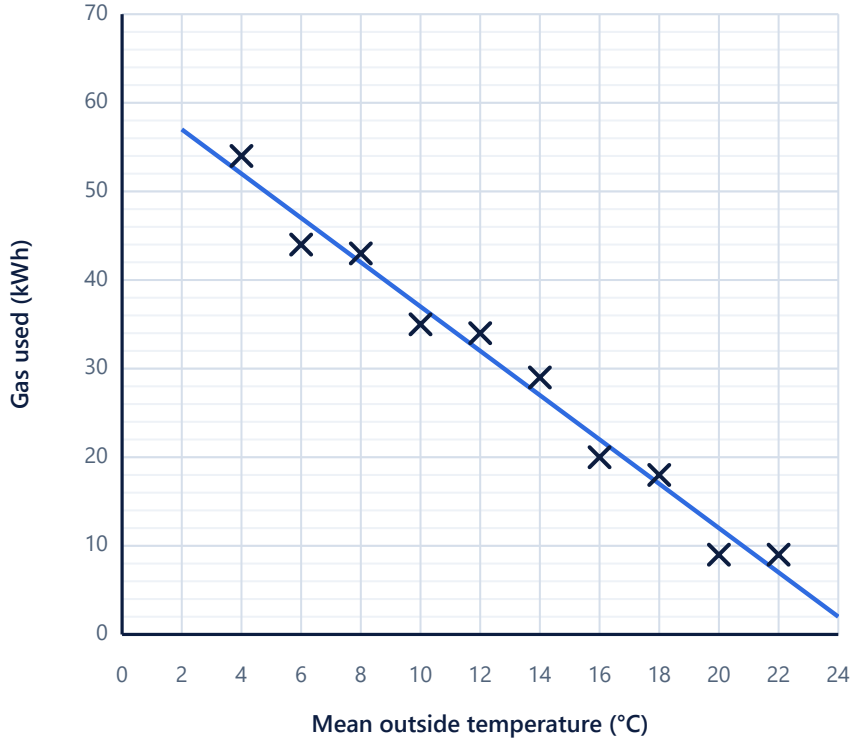
2 Imran records the mean outside temperature, x °C, and the amount of gas used to heat his house, y kWh, on each of 10 days. The scatter diagram shows his data and the regression line.

The equation of the regression line is $y = 62 - 2.5x$

a. Use the equation to estimate the gas used on a day when the mean outside temperature is 15 °C. (1)

b. Interpret the value −2.5 in the equation of the regression line. (1)

c. Imran uses the equation to predict the gas used on a day when the mean outside temperature is 30 °C. Work out his prediction and give two reasons why it is not reliable. (2)



(4 marks)

a) Put $x = 15$ into the equation

$y = 62 - 2.5 \times 15$

$y = 62 - 37.5$

$y = 24.5$ kWh

b) −2.5 is the gradient

For each 1°C rise in outside temperature

the gas used goes down by 2.5 kWh, on average

c) Put $x = 30$ into the equation

$y = 62 - 2.5 \times 30$

$y = 62 - 75$

$y = -13$ kWh

Reason 1

He cannot use a negative amount of gas

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.

Reason 2

30°C is outside the range of his data (4°C to 22°C)

This is extrapolation, the pattern may not continue

a) 24.5 kWh b) 2.5 kWh less per 1°C c) -13 kWh

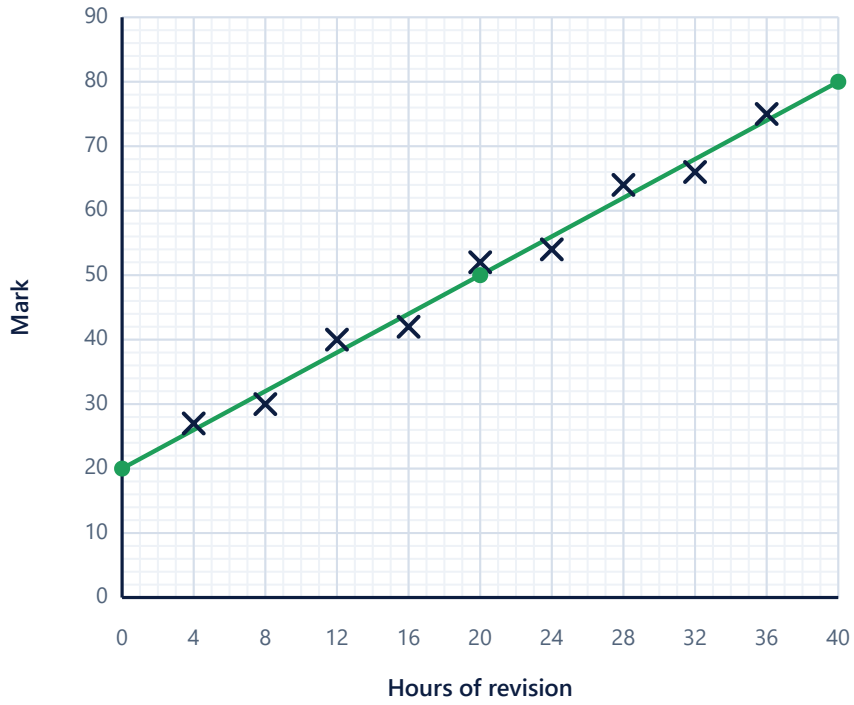
3 A teacher records the number of hours of revision, x, and the mark in a test, y, for each of 9 students. The scatter diagram shows the data. The mean number of hours is 20 and the mean mark is 50.

The equation of the regression line is $y = 20 + 1.5x$

a. Show that the regression line passes through the point of the means. (1)

b. Draw the regression line on the scatter diagram. (2)

c. The teacher uses the regression line to predict the mark of a student who revises for 25 hours. Give two reasons why this prediction is likely to be reliable. (1)



(4 marks)

a) Put the mean number of hours, $x = 20$, into the equation

$y = 20 + 1.5 \times 20$

$y = 20 + 30$

$y = 50$

This is the mean mark, so the line goes through (20, 50)

b) Work out two more points on the line

When $x = 0$

$y = 20$

When $x = 40$

$y = 20 + 1.5 \times 40$

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.

$y = 80$

Plot (0, 20), (20, 50) and (40, 80)

Join them with a straight line

c)

25 hours is inside the range of the data (4 to 36 hours)

The points are all close to the line

a) $y = 50$ when $x = 20$ b) Line drawn c) in range, close fit

4 A sports scientist measures the age, x years, and the reaction time, y milliseconds, of each of 30 adults. The adults are aged from 20 to 60. She finds that the equation of the regression line is $y = 180 + 1.2x$

a. Interpret the value 1.2 in the equation of the regression line. (1)

b. Interpret the value 180 in the equation of the regression line, and explain why this interpretation is not sensible. (2)

c. Use the equation to predict the reaction time of an adult aged 45. (1)

d. Spearman's rank correlation coefficient for the data is 0.35. Comment on the reliability of your prediction in part (c). (1)

(5 marks)

a) 1.2 is the gradient

For each extra year of age, the reaction time

goes up by 1.2 milliseconds, on average

b) 180 is the value of y when $x = 0$

It is the predicted reaction time at age 0

Age 0 is far outside the range of the data (20 to 60)

and a baby could not do the test, so it is not sensible

c) Put $x = 45$ into the equation

$y = 180 + 1.2 \times 45$

$y = 180 + 54$

$y = 234$ milliseconds

d) 0.35 is a weak positive correlation

The points are not close to the line,

so the prediction is not very reliable

a) 1.2 ms more per year b) time at age 0 c) 234 ms d) weak

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.

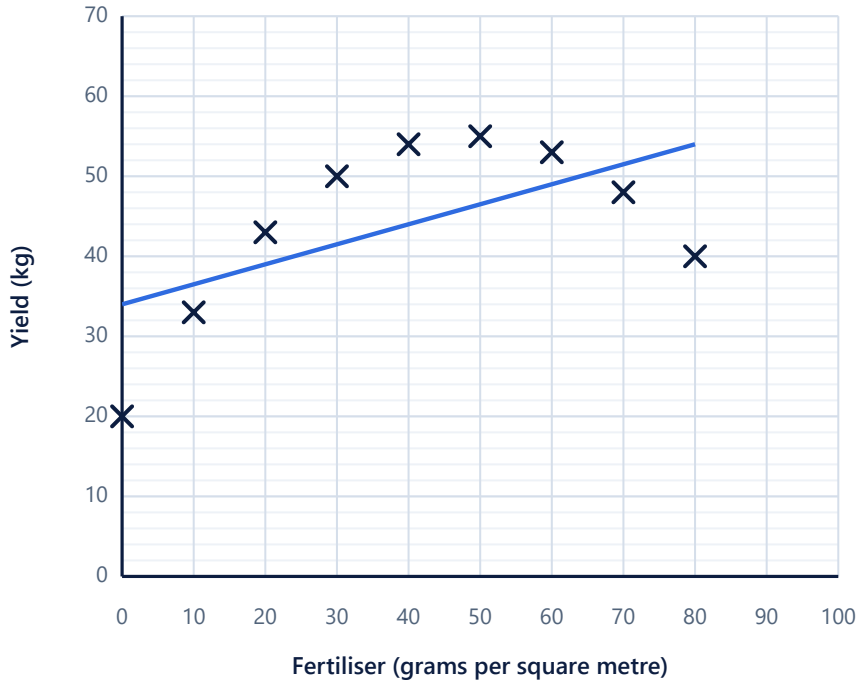
5 A farmer puts different amounts of fertiliser, x grams per square metre, on 9 plots of land and records the yield of wheat, y kg, from each plot. The scatter diagram shows his data and the regression line.

The equation of the regression line is $y = 34 + 0.25x$

a. Use the equation to predict the yield from a plot given 100 grams of fertiliser per square metre. (1)

b. Give two reasons, using the scatter diagram, why this prediction is not reliable. (2)

c. The farmer says, "The regression line shows that the more fertiliser I use, the more wheat I get." Use the scatter diagram to comment on his statement. (2)



(5 marks)

a) Put $x = 100$ into the equation

$y = 34 + 0.25 \times 100$

$y = 34 + 25$

$y = 59$ kg

b) Reason 1

100 is outside the range of the data (0 to 80)

so this is extrapolation

Reason 2

The points follow a curve, not a straight line

so a straight line is not a suitable model

c) He is only right for small amounts of fertiliser

The yield goes up until about 50 grams per square metre

For 60, 70 and 80 grams the yield goes down

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.

as more fertiliser is used

a) 59 kg b) extrapolation, curved pattern c) not always true

Free for classroom and home use: print, photocopy and share. Not for resale or re-hosting.