Correlation and regression

IB Mathematics: Analysis and Approaches — SL & HL

Section 4.4 | 50 shared questions + 10 HL challenges

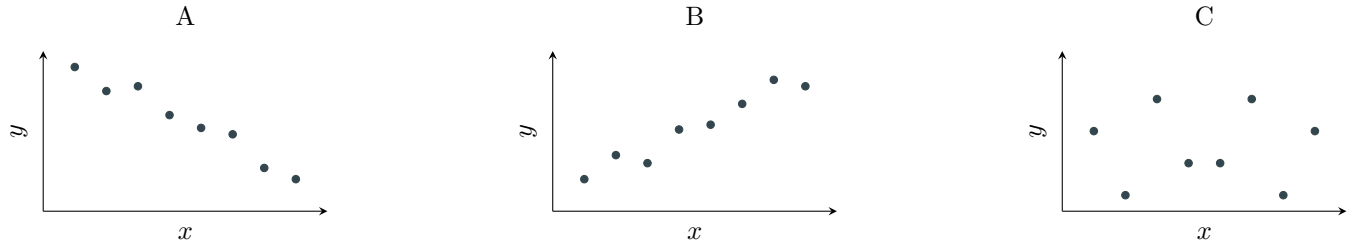
Learning goal	Read scatter diagrams, use linear regression and judge the reliability of predictions.
Syllabus link	AA SL 4.4, shared by SL and HL: correlation, Pearson's r, mean point, lines of best fit and regression of y on x.
Structure	Q1–15 short response; Q16–30 extended response; Q31–50 exam-style; Q51–60 HL-exclusive challenge. Answers follow all questions.
Technology	Use a GDC where needed in Q1–30 and Q51–60. Q31–50 are non-calculator: any required regression results are supplied.

Instructions. Show clear working and include units. Give exact answers where possible; otherwise use 3 significant figures unless stated. Keep full calculator precision until the final answer. For hand-drawn lines, aim to have as many points above as below your line, and pass through the mean point. Use separate paper for longer working. All questions are original.

1. Short response Q1–15

1. A coach wants to use weekly training time to predict a runner's 5 km race time. State which variable should be placed on each axis of the scatter diagram.

2. Describe the direction and strength of the linear correlation in each scatter diagram.



3. Three data sets have correlation coefficients -0.92, 0.81 and -0.34. List them in order from the weakest to the strongest linear correlation.

4. State which of the following could be values of Pearson's correlation coefficient:

$$-1.08, \qquad -0.65, \qquad 0, \qquad 0.99, \qquad 1.12.$$

5. The correlation between hours of practice and a test score is $r = 0.88$. Describe this association in context. Does $r = 0.88$ mean that the score rises by 88% for every extra hour?

6. Find the mean point $(\bar{x}, \bar{y})$ for the following data.

x	3	5	6	8	13
y	9	11	14	18	23

7. The regression line $y = 2.5x + b$ passes through the mean point $(12, 38)$. Find b.

8. A regression model is $y = -1.5x + 26$. Use the model to estimate y when $x = 6$.

9. A plant's height, h cm, is modelled by $h = 1.8t + 5$, where t is the number of days after planting. Interpret the gradient and find the predicted increase in height from day 6 to day 10.

10. A taxi fare is modelled by $C = 2.2d + 4.8$, where C is in euros and d is the distance in kilometres. Interpret the intercept in context.

11. A regression line is fitted using data with $6 \leq x \leq 24$. Classify a prediction at each of $x = 4$, $x = 18$ and $x = 30$ as interpolation or extrapolation.

12. A data set has $r = 0$. State what this tells you about linear correlation. What type of relationship could still be present?

13. A data set has $r = -1$. Describe how its points are arranged on a scatter diagram. Must the gradient of its regression line be -1?

14. Ice-cream sales and the number of cases of sunburn have a positive correlation. Explain why this does not show that eating ice cream causes sunburn. Suggest a factor that could affect both variables.

15. The regression line of y on x is $y = 3x + 4$. A student says, "The correlation coefficient is positive and equals 3." Identify the correct part of the statement and explain the error.

2. Extended response — Q16–30 | GDC where needed

16. The table shows the number of practice sessions completed by six students and their scores on the same test.

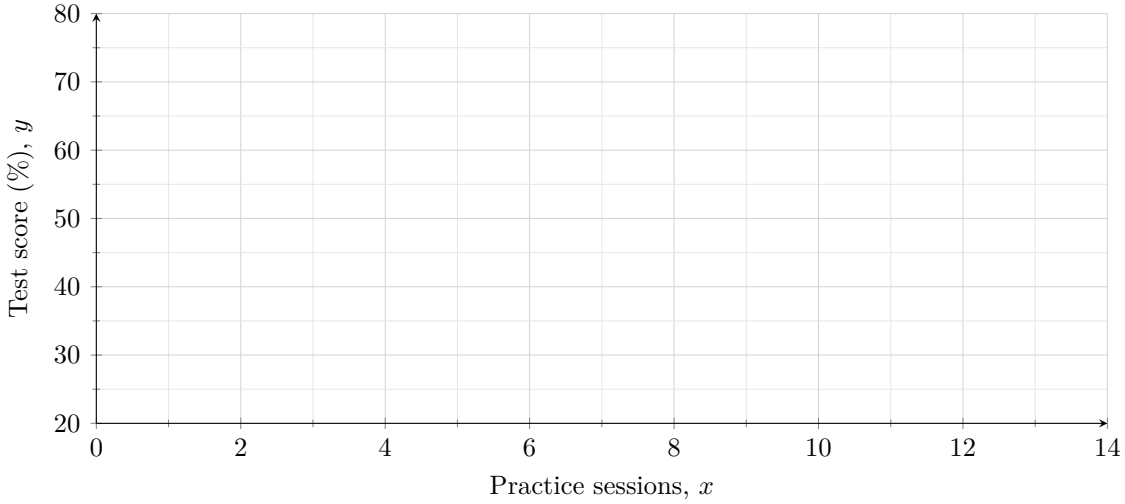
Practice sessions, x	2	4	6	8	10	12
Test score (%), y	36	42	45	55	59	63

(a) Plot the six points on the axes below.

(b) Calculate the mean point $M(\bar{x}, \bar{y})$ and mark it on the diagram.

(c) Draw a line of best fit by eye through M. Aim to have approximately as many points above the line as below it.

(d) Use your line to estimate the score for a student who completed nine practice sessions.



17. Use the data from Question 16.

(a) Use technology to find Pearson's r and the regression line of y on x.

(b) Use the regression line to predict the score at $x = 9$. Compare this with your answer to Question 16.

(c) Interpret the gradient in context.

(d) Explain why the correlation does not, by itself, prove that extra practice sessions caused the higher scores.

18. A café records the afternoon temperature, $x\,^\circ\text{C}$, and the number of hot drinks sold, y, on six days.

x	8	12	16	20	24	28
y	92	81	76	60	51	42

(a) Find Pearson's r and describe the linear correlation.

(b) Find the regression line of y on x.

(c) Predict the number of hot drinks sold when $x = 22$. Give your answer to the nearest whole number.

(d) Explain why a prediction at $x = 40$ should be treated with caution.

19. The battery charge, $y\%$, of a tablet is recorded after x minutes of continuous use.

x (minutes)	0	10	20	30	40	50
y (%)	96	89	80	74	63	57

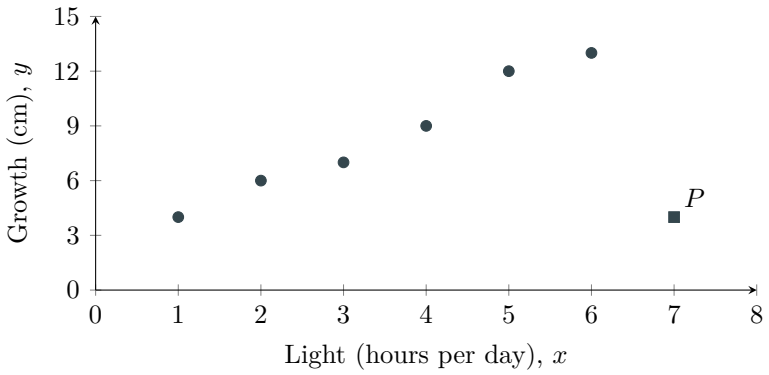
(a) Find the regression line of y on x.

(b) Predict the charge after 35 minutes.

(c) Find the value predicted by this line after 160 minutes. Explain what this tells you about using the model far beyond the observed times.

20. Seven plants of the same species are grown with different daily amounts of light. Their light exposure, x hours per day, and weekly growth, y cm, are shown below. The final observation is labelled P.

Plant	A	B	C	D	E	F	P
x	1	2	3	4	5	6	7
y	4	6	7	9	12	13	4



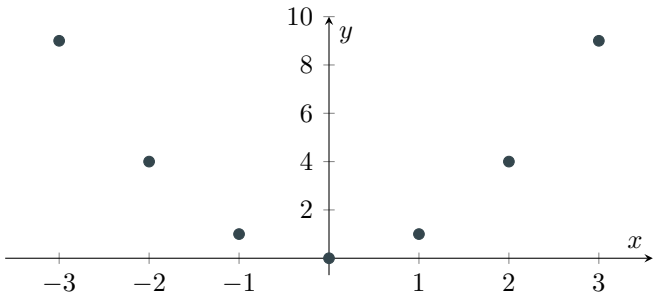
(a) Find r using plants A–F only.

(b) Find r using all seven plants. Describe the effect of including P.

(c) Give one sensible check to make before deciding whether to exclude P.

21. A set of paired values is shown in the table and scatter diagram.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9



(a) Use technology to find Pearson's r.

(b) Explain why this value does not mean that there is no relationship between x and y.

(c) Explain why a straight-line model is unsuitable for these data.

22. Five paired observations are shown below. Their regression line of y on x is $y = 1.65x + 5.7$.

x	2	4	6	8	10
y	9	12	k	19	22

(a) Find $\bar{x}$, then use the regression line to find $\bar{y}$.

(b) Find k.

(c) Use the line to estimate y when $x = 7$ and state whether this is interpolation or extrapolation.

23. For two groups of students, x is the weekly number of hours of practice and y is the test score, as a percentage. The following results were obtained, using $0 \leq x \leq 10$ in both groups.

Group	Regression line of y on x	r
A	$y = 8x + 10$	0.72
B	$y = 2x + 40$	0.95

(a) Which group has the stronger linear correlation? Explain.

(b) Find the predicted increase in score for two additional hours of practice in each group.

(c) Explain why a larger gradient does not necessarily mean a stronger correlation.

24. For a sample of adults, height x cm and arm span y cm have $r = 0.88$. The regression line of y on x is $y = 0.8x + 34$, based on heights from 150 cm to 185 cm.

(a) Predict the arm span of an adult whose height is 175 cm.

(b) A student substitutes $y = 180$ into this equation and solves for x. Find the estimate the student obtains.

(c) Explain why reversing this equation is not generally the most suitable method for predicting height from arm span.

25. A shop records the outdoor temperature, $x\,^\circ\text{C}$, and the volume of bottled water sold, y litres, on four days.

x	5	10	15	20
y	18	27	34	43

(a) Find r and the regression line of y on x.

(b) Predict the volume sold at $12\,^\circ\text{C}$ and at $35\,^\circ\text{C}$.

(c) Which prediction is better supported by the observed data? Explain, despite the very high value of r.

26. A temperature sensor is tested against known temperatures. Let $x\,^\circ\text{C}$ be the actual temperature and $y\,^\circ\text{C}$ be the sensor reading.

x	0	10	20	30	40
y	2	11	23	31	43

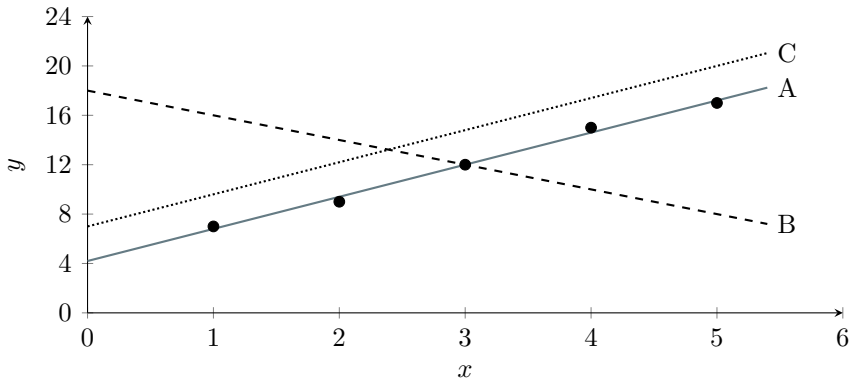
(a) Find the regression line of y on x.

(b) Predict the sensor reading when the actual temperature is $25\,^\circ\text{C}$.

(c) A residual is the observed y-value minus the predicted y-value. Find the residual for the observation with $x = 20$, and state whether that point is above or below the regression line.

27. Five observations are $(1, 7)$, $(2, 9)$, $(3, 12)$, $(4, 15)$ and $(5, 17)$. Three possible lines are drawn on the scatter diagram:

A: $y = 2.6x + 4.2$, B: $y = -2x + 18$, C: $y = 2.6x + 7$.



(a) Find the mean point.

(b) Which of the three lines pass through the mean point?

(c) Choose the best-fitting line and explain why the other line through the mean point is unsuitable.

28. Six students report their weekly hours of practice and their scores on the same test. Each column is one student's paired observation.

Student	A	B	C	D	E	F
Hours, x	1	2	3	4	5	6
Score, y	44	55	49	64	58	72

(a) Find Pearson's r for the original paired data.

(b) A student keeps the x-list unchanged but sorts the y-list into $44, 49, 55, 58, 64, 72$. Find the new r.

(c) Explain why the result in part (b) is not a valid correlation coefficient for the original students.

29. For a small business, x is weekly advertising spending in thousands of euros and y is weekly revenue in thousands of euros. The fitted model is $y = 4.5x + 18$, using data with $1 \leq x \leq 8$.

(a) Find the predicted revenue when advertising spending is €4500. Give your answer in euros.

(b) Interpret the gradient in context.

(c) Let X and Y be the same spending and revenue measured in euros. Write the model relating Y to X.

30. A study records weekly revision time x minutes and a test score y%. For the observed range $20 \leq x \leq 100$, the regression line is $y = 0.4x + 42$ and $r = 0.91$. A report claims: *Revising for 150 minutes guarantees a score of 102%. The high correlation proves that extra revision caused the higher scores.*

(a) Use the model to predict a score for 90 minutes of revision.

(b) Confirm the calculation behind the report's prediction of 102%.

(c) Give three reasons why the report's conclusion is not justified.

3. IB exam-style practice Q31–50 | No calculator

These are original Paper 1-style questions. Use the results supplied; you do not need to calculate Pearson's r or fit a regression line by hand. Marks are shown in brackets.

31. [Maximum mark: 7]

The following four paired observations are recorded.

x	1	2	3	4
y	3	5	4	8

The regression line of y on x is $y = 1.4x + 1.5$.

(a) Find the mean point. [2]

(b) Verify that the regression line passes through the mean point. [2]

(c) Estimate y when $x = 3.5$. [2]

(d) State whether this estimate is interpolation or extrapolation. [1]

32. [Maximum mark: 7]

A café records the temperature t (°C) and the number of smoothies sold, s, on several days. The observed temperatures satisfy $14 \leq t \leq 26$. The regression model is $s = 5t - 30$, with $r = 0.91$.

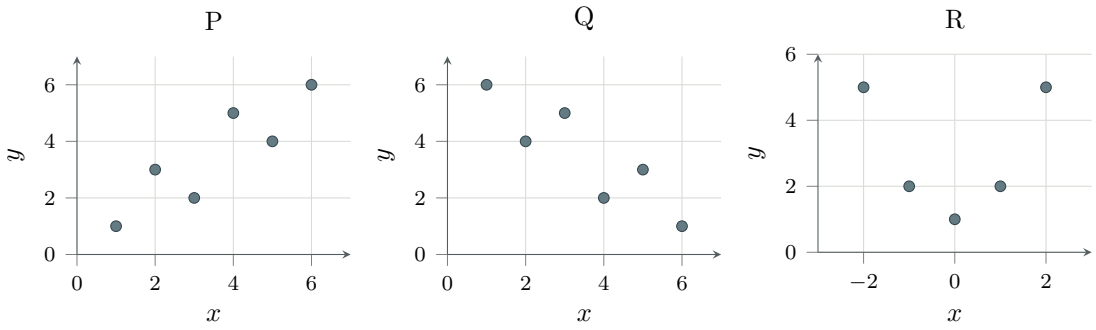
(a) Interpret the gradient of this line in context. [2]

(b) Estimate the number of smoothies sold when $t = 22$. [2]

(c) The model gives $s = -5$ when $t = 5$. Explain two reasons why this is not a sensible prediction. [3]

33. [Maximum mark: 5]

The scatter diagrams show three data sets. Their correlation coefficients, in a different order, are -0.886, 0 and 0.886.



(a) Match each diagram to its correlation coefficient. [3]

(b) Explain why $r = 0$ does not mean that the variables in diagram R have no relationship. [2]

34. [Maximum mark: 6]

Four paired observations have the following values.

x	2	4	6	8
y	5	7	k	13

The regression line of y on x is $y = 1.4x + 2$.

(a) Use the mean-point property to find k. [3]

(b) Estimate y when $x = 7$. [2]

(c) State which variable this regression line is designed to predict. [1]

35. [Maximum mark: 5]

For eight paired observations, $\bar{x} = 12$ and the regression line is $y = -0.75x + 22$.

(a) Find $\bar{y}$. [1]

(b) Find the sum of the eight y-values. [2]

(c) A student states that $r = -0.75$ because the gradient is -0.75. Explain why this conclusion is not justified. [2]

36. [Maximum mark: 6]

A factory studies maintenance time x (hours per week) and the number of faulty items d produced per week. For $2 \leq x \leq 8$, the model is $d = 24 - 2x$.

(a) In a week with $x = 5$, the observed number of faulty items is 18. Find how far the observation is above or below the model's prediction. [2]

(b) Find the predicted change in d when maintenance time increases by 1.5 hours. [2]

(c) Explain why the model alone does not prove that an increase in maintenance time causes the predicted change. [2]

37. [Maximum mark: 6]

For a sample of adults, height x (cm) and mass y (kg) have regression line $y = 0.8x - 70$ and $r = 0.68$. The observed heights range from 150 cm to 190 cm.

(a) Estimate the mass of an adult of height 180 cm. [2]

(b) A student substitutes $y = 70$ and rearranges the equation to estimate height. Find the value obtained. [2]

(c) Explain why this calculation is not generally the appropriate regression method for predicting height from mass. [2]

38. [Maximum mark: 6]

For eight paired observations, $\sum x = 64$ and $\sum y = 152$. The regression line $y = ax + b$ also passes through the point $(2, 7)$.

(a) Find the mean point. [2]

(b) Hence find a and b. [3]

(c) Use the model to estimate y when $x = 10$. [1]

39. [Maximum mark: 6]

The data below have $r = 0$. Their regression line of y on x is $y = 3$.

x	-2	-1	0	1	2
y	5	2	1	2	5

(a) Find the mean point and verify that it lies on the regression line. [2]

(b) Write down a simple quadratic equation satisfied by all five observations. [2]

(c) Explain why the regression line gives a poor description of the pattern, even though it passes through the mean point. [2]

40. [Maximum mark: 7]

Students attend x revision sessions and obtain a test score y out of 100. The observed values satisfy $0 \le x \le 5$. The regression model is $y = 8x + 35$.

(a) Estimate the score of a student attending four sessions. [2]

(b) Find the score predicted for ten sessions. [2]

(c) Explain two reasons why this second prediction should not be used. [3]

41. [Maximum mark: 6]

A study of different beaches finds $r = 0.87$ between the number of lifeguards employed and the number of rescues made during a summer.

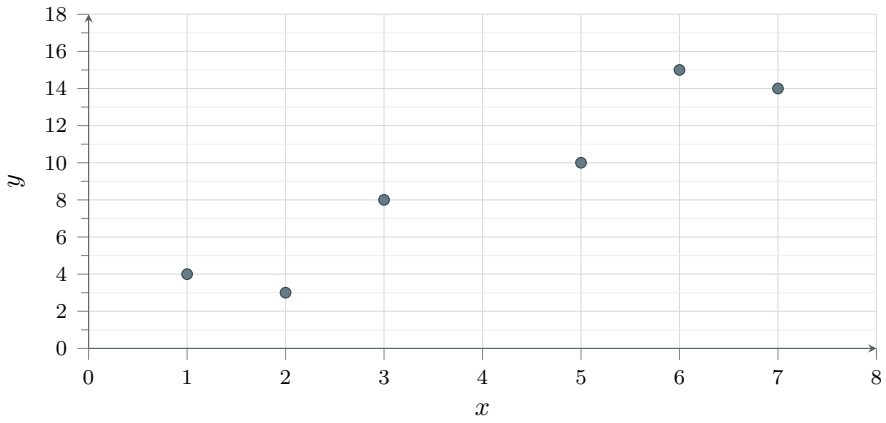
(a) Describe this linear correlation. [2]

(b) Explain why the study does not show that employing more lifeguards causes more people to need rescuing. [2]

(c) Suggest a third variable that could help explain the observed association, and explain its role. [2]

42. [Maximum mark: 6]

The mean point of the observations shown is $M(4, 9)$.



(a) Mark M on the diagram. [1]

(b) Draw a reasonable line of best fit by eye through M. [2]

(c) Use two points on your line to find its equation. [2]

(d) Use your line to estimate y when $x = 4.5$. [1]

43. [Maximum mark: 6]

A student copies three regression outputs. In each case both variables vary.

Data set	Reported r	Reported regression line
A	0.82	$y = -2x + 5$
B	−1.06	$y = -x + 3$
C	−0.91	$y = -0.4x + 7$

(a) Explain why the output for A cannot be correct. [2]

(b) Explain why the output for B cannot be correct. [1]

(c) Describe the linear correlation for C. [2]

(d) State the predicted change in y for C when x increases by one. [1]

44. [Maximum mark: 6]

Delivery cost y (euros) and parcel mass x (kg) have regression model $y = 1.5x + 2$, with $r = 0.93$. The same masses are then recorded as u grams, so $u = 1000x$.

(a) Express the regression model in terms of u and y. [2]

(b) Use the new equation to estimate the cost for a parcel of mass 2000 g. [2]

(c) State the new value of r and explain why it has not changed. [2]

45. [Maximum mark: 5]

For a plant experiment, fertiliser mass x (g) and plant height y (cm) have model $y = 0.247x + 5.186$. The observed fertiliser masses range from 10 g to 50 g.

(a) Estimate the height when $x = 40$, giving your answer to 3 significant figures. [2]

(b) A student first rounds the equation to $y = 0.2x + 5$. Find the difference between this student's prediction at $x = 40$ and the prediction from the original equation, before final rounding. [2]

(c) State how this error can be avoided. [1]

46. [Maximum mark: 6]

Two plant varieties have height models $h_A = 1.2t + 6$ and $h_B = 0.8t + 13$, where height is in centimetres and age t is in weeks. Both models were fitted using $5 \leq t \leq 15$.

(a) Find the difference between their predicted heights at $t = 10$. [2]

(b) Find the age at which the two models predict equal heights. [2]

(c) Explain why the answer to part (b) requires caution. [2]

47. [Maximum mark: 6]

Ten paired observations have mean point $(6, 20)$ and regression line $y = 3x + 2$. The observation $(15, 38)$ is removed after a recording error is confirmed.

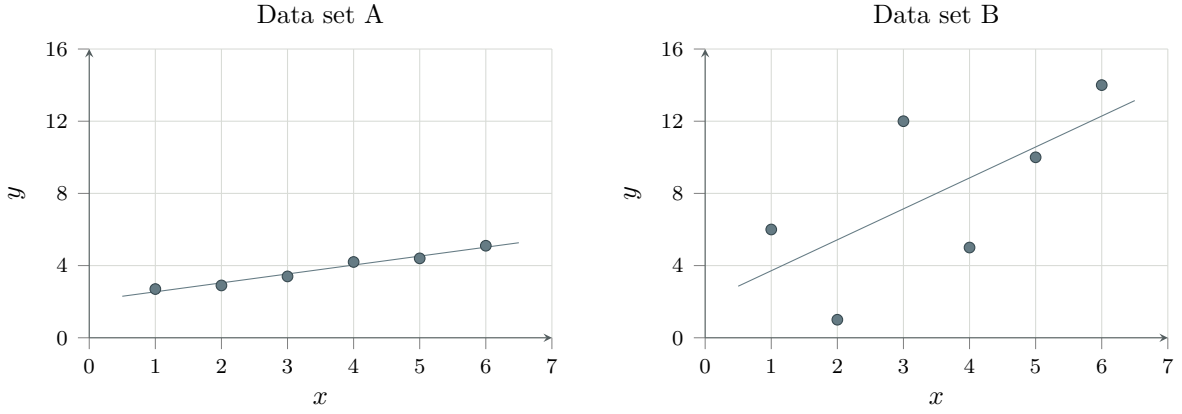
(a) Find the mean point of the remaining nine observations. [3]

(b) Show that $y = 3x + 2$ cannot be the regression line for the remaining observations. [2]

(c) State whether the new mean point alone is enough to find the new regression line. [1]

48. [Maximum mark: 5]

The two diagrams use the same axis scales.



(a) State which data set has the stronger linear correlation, giving a reason. [2]

(b) State which regression line has the larger gradient. [1]

(c) Explain what the diagrams show about the difference between gradient and strength of correlation. [2]

49. [Maximum mark: 6]

Four paired observations have regression line $y = 2x + 5$.

x	1	2	3	4
Observed y	8	7	12	13

(a) Find the four predicted y-values. [2]

(b) For each observation, calculate observed y minus predicted y. [2]

(c) Identify the observation whose prediction has the greatest absolute error, and state whether the model overestimates or underestimates y there. [2]

50. [Maximum mark: 10]

A sample of students records revision time x (hours) and test score y (out of 100). The observed revision times satisfy $2 \leq x \leq 10$. The mean point is $(6, 58)$, and the regression model is $y = 4x + 34$, with $r = 0.86$.

(a) Verify that the mean point lies on the regression line. [1]

(b) Estimate the score when $x = 8$. [2]

(c) Explain why this use of the model is more appropriate than a prediction at $x = 16$. [2]

(d) A report claims: "A student who does no revision will score 34. Also, $r = 0.86$ proves that revision causes 86% of the score." Identify and correct two errors in these claims. [4]

(e) One student revises for eight hours and scores 61. Find observed score minus predicted score. [1]

4. AA HL exclusive challenge Q51–60

For HL students: deeper applications of the same correlation and regression ideas. Use technology where needed.

51. Five paired observations have regression line $y = 2x + 3$. It is also known that $q = p + 2$.

x	1	2	3	4	5
y	5	p	8	q	14

(a) Find the mean point, then use $q = p + 2$ to determine p and q.

(b) Find Pearson's r for the completed data and describe the linear correlation.

(c) Explain why the mean-point condition alone gives insufficient information to determine both p and q.

52. A researcher records temperature T (°C) and energy use E (kWh). For $5 \le T \le 25$, the regression model is $E = 0.32T + 2.8$, with $r = 0.94$. Another report uses temperature F (°F) and energy J (MJ), where $F = 1.8T + 32$ and $J = 3.6E$.

(a) Express the regression model in the form $J = aF + b$.

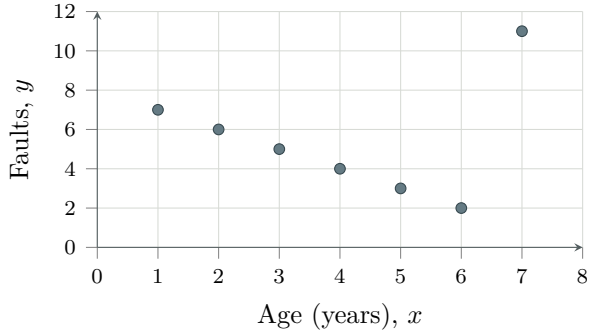
(b) State the new value of r, explaining your answer.

(c) Estimate J when $F = 59$.

(d) Explain why the negative intercept of your new model does not describe a physically possible negative energy use.

53. Seven machines are tested. Their ages x (years) and numbers of faults y during the test are shown.

x	1	2	3	4	5	6	7
y	7	6	5	4	3	2	11



(a) Identify the observation that does not follow the pattern of the other six.

(b) Find r and the regression line of y on x for all seven observations.

(c) Repeat part (b) after omitting the unusual observation.

(d) A technician removes it only because this increases $|r|$. Explain why this is not a valid reason, and state one check that should be made.

54. Consider again the seven observations below.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

(a) Find r and the regression line of y on x.

(b) State a simple curved model that fits all seven observations exactly.

(c) An eighth observation $(0, k)$ is added. The regression line for the eight observations is $y = 5$. Use the mean-point property to find k.

(d) Verify the new value of r using technology. Explain why this value does not tell you whether all eight observations fit the curved model from part (b).

55. Paired observations are collected from two groups. The number of observations in group A is n.

Group	Number	Mean x	Mean y
A	n	3	10
B	4	8	15

For the combined data, the regression line is $y = 1.1x + 6.5$.

(a) Express the combined mean point in terms of n.

(b) Hence determine n.

(c) Find the combined mean point and use the regression line to estimate y when $x = 7$.

(d) Explain why simply averaging the two group mean points would be incorrect here.

56. A data set has regression line $y = 1.8x + 4$, correlation coefficient $r = 0.87$, and $\bar{x} = 6$. New variables are defined by $u = 12 - 3x$ and $v = 5 + 2y$.

(a) Find the original mean point.

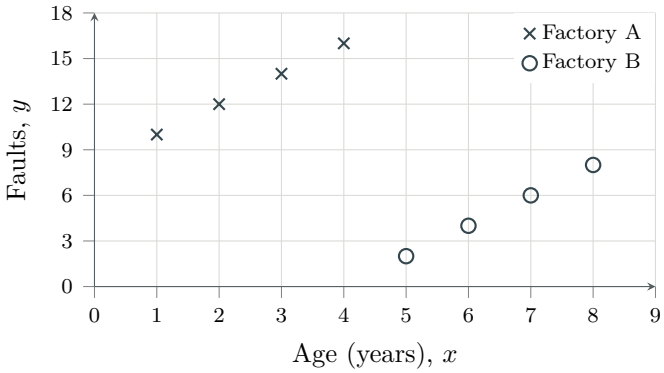
(b) Express the transformed regression line as $v = Au + B$.

(c) Find the new mean point and verify that it lies on your transformed line.

(d) State the correlation coefficient between u and v, explaining its sign and magnitude.

57. Two factories record machine age x (years) and number of faults y in a test.

Factory	Paired observations (x, y)
A	$(1, 10),\ (2, 12),\ (3, 14),\ (4, 16)$
B	$(5, 2),\ (6, 4),\ (7, 6),\ (8, 8)$



(a) Find r for each factory separately and for all eight observations together.

(b) Find the regression line for the combined data.

(c) A report concludes: "Older machines have fewer faults, so keeping a machine longer will reduce faults." Explain two problems with this claim, referring to the grouped data.

58. For cars aged x years, where $1 \le x \le 8$, a fitted sale-price model is $y = -1.8x + 24.6$, where y is in thousands of euros. The correlation coefficient is -0.92.

A seller pays a commission of 5% of the sale price plus €600, and a preparation cost of $(0.4x + 0.8)$ thousand euros.

(a) Show that the model predicts net proceeds of $N = 21.97 - 2.11x$, in thousands of euros.

(b) Find the age at which the model predicts net proceeds of exactly €12 000, to two decimal places. State which ages in the model's range give at least this amount, using the unrounded threshold.

(c) Explain why this threshold is an estimate rather than a guarantee for an individual car.

59. A regression model of y on x is $y = ax + b$. The observed x-values range from 8 to 30. The mean point is $(18, 42)$, the model predicts $y = 50$ at $x = 26$, and $r = 0.56$.

(a) Find a and b.

(b) A student rearranges the equation to predict x when $y = 60$. Find the value obtained.

(c) Explain two reasons for treating that value cautiously.

(d) Use the model to predict y when $x = 20$. Explain why this is a more appropriate use of the model, while still being uncertain.

60. Five paired observations are shown below.

x	1	2	3	4	5
y	4	4	8	8	11

Three proposed lines are

$$L_1:\ y = 1.7x + 1.9, \qquad L_2:\ y = 2x + 1, \qquad L_3:\ y = 1.4x + 2.8.$$

(a) Find the mean point and show that all three lines pass through it.

(b) Find the actual regression line of y on x and Pearson's r.

(c) Explain why passing through the mean point does not identify the regression line uniquely.

(d) The final y-value is accidentally entered as 16. Find the new regression line and r. Explain why checking the raw data remains essential even when $|r|$ is large.

Answer key and solution outlines Q1–60

Numerical regression results use unrounded coefficients before the final answer is rounded. By-eye lines and contextual explanations may vary.

Section 1: Short response — Q1–15

1. Horizontal axis: weekly training time. Vertical axis: 5 km race time.

2. A: strong negative linear correlation. B: strong positive linear correlation. C: no linear correlation.

3. $-0.34,\ 0.81,\ -0.92$, using the sizes of $|r|$.

4. -0.65, 0 and 0.99; Pearson's r must satisfy $-1 \leq r \leq 1$.

5. There is a strong positive linear association: students who practice more tend to score more highly. No; r measures linear association, not a percentage increase.

6. $\bar{x} = 35/5 = 7$, $\bar{y} = 75/5 = 15$; mean point $(7, 15)$.

7. $38 = 2.5(12) + b$, so $b = 8$.

8. $y = -1.5(6) + 26 = 17$.

9. The predicted height increases by 1.8 cm per day. The predicted increase over four days is $1.8(10 - 6) = 7.2$ cm.

10. The model predicts a fare of €4.80 when $d = 0$: a fixed starting charge in this model.

11. $x = 4$: extrapolation. $x = 18$: interpolation. $x = 30$: extrapolation.

12. There is no linear correlation. A non-linear relationship could still be present.

13. All points lie exactly on a straight line with a negative gradient. The gradient need not be -1; r does not give the gradient.

14. An association alone does not establish a cause. Hot, sunny weather can increase both ice-cream sales and time in the sun, which can increase sunburn.

15. The positive gradient means that r is positive. However, r cannot equal 3, since $r \leq 1$; its value cannot be found from this gradient alone.

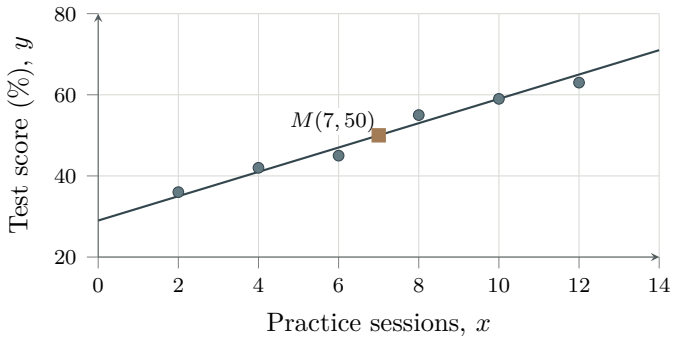
Section 2: Extended response — Q16–30

16.

(a) Plot the six given pairs.

(b) $\bar{x} = 42/6 = 7$, $\bar{y} = 300/6 = 50$, so $M = (7, 50)$.

(c)–(d) Answers vary. For example, $y = 3x + 29$ passes through M and gives $y = 56\%$ at $x = 9$. Accept a sensible line through M and a consistent estimate.



Section 2: Extended response — Q16–30 (continued)

17.

(a) $r = 0.989949\ldots \approx 0.990$; $y = 2.8x + 30.4$.

(b) $y = 2.8(9) + 30.4 = 55.6\%$, which should be close to a sensible by-eye estimate.

(c) Each additional practice session is associated with a predicted increase of 2.8 percentage points in the score.

(d) Other factors, such as prior attainment or motivation, may also affect the scores; the data show association, not proof of cause.

18.

(a) $r = -0.994118\ldots \approx -0.994$: strong negative linear correlation.

(b) $y = -2.542857\ldots x + 112.771428\ldots$ (or $y \approx -2.54x + 113$).

(c) Using unrounded coefficients, $y = 56.828571\ldots$, so about 57 drinks.

(d) 40 °C is outside the observed range 8–28 °C; the same linear pattern may not continue.

19.

(a) $y = -0.797142857\ldots x + 96.428571\ldots$.

(b) $y = 68.528571\ldots \approx 68.5\%$.

(c) $y = -31.114285\ldots \approx -31.1\%$, which is impossible for battery charge. The linear model should not be extrapolated this far.

20.

(a) $r = 0.990664\ldots \approx 0.991$.

(b) $r = 0.404346\ldots \approx 0.404$; including P makes the positive linear association much weaker.

(c) Check for a recording error, an unhealthy plant, or a difference in growing conditions. Do not remove the observation only because it weakens the correlation.

21.

(a) $r = 0$.

(b) The points follow the exact curved relationship $y = x^2$; $r = 0$ tells us only that there is no linear correlation.

(c) The values first decrease and then increase, so a single straight line does not follow the pattern.

22.

(a) $\bar{x} = 6$ and $\bar{y} = 1.65(6) + 5.7 = 15.6$.

(b) $(62 + k)/5 = 15.6$, so $k = 16$.

(c) $y = 1.65(7) + 5.7 = 17.25$; interpolation.

23.

(a) Group B, because $|0.95| > |0.72|$.

(b) Group A: 16 percentage points. Group B: 4 percentage points.

(c) The gradient measures the rate of change of the predicted score. Correlation measures how closely the points follow a straight-line pattern.

24.

(a) $y = 0.8(175) + 34 = 174$ cm.

(b) $x = (180 - 34)/0.8 = 182.5$ cm.

(c) This line was fitted to predict arm span from height. Reversing it does not generally give the regression line fitted to predict height from arm span; a model fitted in that direction is more appropriate.

Section 2: Extended response — Q16–30 (continued)

25.

(a) $r = 0.998812\ldots \approx 0.999$; $y = 1.64x + 10$.

(b) At 12 °C: $29.68 \approx 29.7$ L. At 35 °C: 67.4 L.

(c) The prediction at 12 °C is interpolation within 5–20 °C. The prediction at 35 °C is extrapolation; a high r for only four observed days does not establish that the pattern continues.

26.

(a) $y = 1.02x + 1.6$.

(b) $y = 1.02(25) + 1.6 = 27.1$ °C.

(c) At $x = 20$, the predicted reading is 22 °C. Residual $= 23 - 22 = 1$ °C, so the point is above the line.

27.

(a) $\bar{x} = 3$, $\bar{y} = 12$.

(b) A and B pass through $(3, 12)$; C gives $y = 14.8$ at $x = 3$.

(c) A follows the positive trend and lies close to the observations. B has a negative gradient and does not follow the points. Passing through the mean point alone does not make a line a good fit.

28.

(a) $r = 0.866283\ldots \approx 0.866$.

(b) $r = 0.993056\ldots \approx 0.993$.

(c) Sorting only one list changes which score is paired with which student's hours. The new pairs are not the original observations.

29.

(a) $x = 4.5$, so $y = 4.5(4.5) + 18 = 38.25$ thousand euros: €38 250.

(b) An additional €1000 of weekly advertising is associated with a predicted revenue increase of €4500.

(c) Since $X = 1000x$ and $Y = 1000y$, $Y = 4.5X + 18\,000$.

30.

(a) $y = 0.4(90) + 42 = 78\%$.

(b) $y = 0.4(150) + 42 = 102\%$.

(c) Any three well-explained issues: 150 minutes is outside the observed range; a score above 100% is impossible; a regression prediction is an estimate, not a guarantee; correlation does not prove cause. The model is being used beyond a sensible range.

Section 3: IB exam-style practice — Q31–50

31.

(a) $\bar{x} = 10/4 = 2.5$; $\bar{y} = 20/4 = 5$. Mean point $(2.5, 5)$.

(b) $1.4(2.5) + 1.5 = 3.5 + 1.5 = 5 = \bar{y}$.

(c) $y = 1.4(3.5) + 1.5 = 6.4$.

(d) Interpolation: 3.5 lies within the observed x-range $[1, 4]$.

32.

(a) For each increase of 1°C, the model predicts five more smoothies sold per day, on average.

(b) $s = 5(22) - 30 = 80$ smoothies.

(c) 5°C is outside $[14, 26]$, so this is extrapolation. A negative number of smoothies sold is impossible.

Section 3: IB exam-style practice — Q31–50 (continued)

33.

(a) P: 0.886; Q: -0.886; R: 0.

(b) R has a clear curved pattern. Pearson's r measures linear correlation; it can be zero even when a curved relationship exists.

34.

(a) $\bar{x} = 5$, so $\bar{y} = 1.4(5) + 2 = 9$. Thus $(25 + k)/4 = 9$, giving $k = 11$.

(b) $y = 1.4(7) + 2 = 11.8$.

(c) y, using a known value of x.

35.

(a) $\bar{y} = -0.75(12) + 22 = 13$.

(b) $\sum y = 8(13) = 104$.

(c) The gradient measures predicted change in y per unit of x; r measures direction and strength of linear correlation. They are not generally equal. The negative gradient only tells us that $r < 0$.

36.

(a) Predicted $d = 24 - 2(5) = 14$. Observed $d = 18$ is 4 faulty items above the prediction.

(b) $\Delta d = -2(1.5) = -3$: a predicted decrease of 3 faulty items per week.

(c) The observed association alone does not establish cause. For example, machine age or production volume may also affect faults and maintenance time.

37.

(a) $y = 0.8(180) - 70 = 74$ kg.

(b) $70 = 0.8x - 70 \Rightarrow x = 175$ cm.

(c) The line was fitted to predict mass from height. Rearranging it does not generally give the regression that predicts height from mass; a model fitted in that direction is needed.

38.

(a) $(\bar{x}, \bar{y}) = (64/8, 152/8) = (8, 19)$.

(b) $a = (19 - 7)/(8 - 2) = 2$; $7 = 2(2) + b \Rightarrow b = 3$.

(c) $y = 2(10) + 3 = 23$.

39.

(a) $\bar{x} = 0$, $\bar{y} = 15/5 = 3$. The mean point $(0, 3)$ lies on $y = 3$.

(b) $y = x^2 + 1$.

(c) The points follow a curve: y falls and then rises. A horizontal line misses this pattern. Passing through the mean is not enough to make a linear model appropriate.

40.

(a) $y = 8(4) + 35 = 67$.

(b) $y = 8(10) + 35 = 115$.

(c) Ten sessions is outside the observed range $[0, 5]$. Also, 115 is impossible for a test marked out of 100.

41.

(a) Strong positive linear correlation.

(b) Association does not establish cause; the study has not isolated the effect of the number of lifeguards.

(c) For example, number of beach visitors: busier beaches may employ more lifeguards and also have more people needing rescue. Other well-explained third variables are acceptable.

Section 3: IB exam-style practice — Q31–50 (continued)

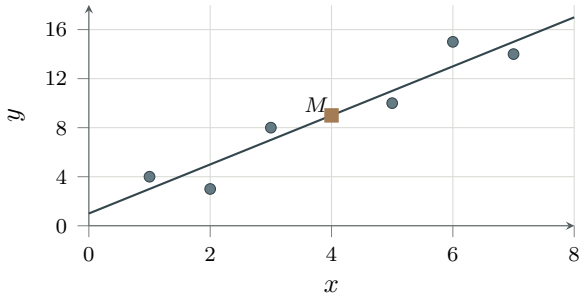
42.

(a) Mark M at $(4, 9)$.

(b) A reasonable rising line passes through M, with roughly as many points above as below.

(c) Answers vary. One suitable line uses $(0, 1)$ and $(8, 17)$: gradient 2, so $y = 2x + 1$.

(d) For this example, $y = 2(4.5) + 1 = 10$. Accept a consistent estimate from a reasonable line.



43.

(a) A positive r requires a positive regression gradient; the reported gradient is negative.

(b) r must lie in $[-1, 1]$; -1.06 is outside this range.

(c) Strong negative linear correlation.

(d) A decrease of 0.4 in y.

44.

(a) $x = u/1000$, so $y = 0.0015u + 2$.

(b) $y = 0.0015(2000) + 2 = 5$ euros.

(c) $r = 0.93$. A positive change of units changes the axis scale but not the direction or strength of the linear correlation.

45.

(a) $y = 0.247(40) + 5.186 = 15.066 \approx 15.1$ cm.

(b) Rounded-model prediction $= 0.2(40) + 5 = 13$ cm. It is $15.066 - 13 = 2.066$ cm lower.

(c) Keep the original coefficients (or full calculator precision) and round only the final answer.

46.

(a) $h_A = 18$ cm, $h_B = 21$ cm; B is predicted to be 3 cm taller.

(b) $1.2t + 6 = 0.8t + 13 \Rightarrow 0.4t = 7 \Rightarrow t = 17.5$ weeks.

(c) 17.5 lies outside $[5, 15]$; the result uses extrapolation, and growth may not continue to follow either line.

47.

(a) Original totals: $\sum x = 60$, $\sum y = 200$. New means: $\bar{x} = (60 - 15)/9 = 5$, $\bar{y} = (200 - 38)/9 = 18$. Mean point $(5, 18)$.

(b) The old line gives $y = 3(5) + 2 = 17 \neq 18$, so it does not pass through the new mean point.

(c) No. Many lines pass through $(5, 18)$; further information is needed.

48.

(a) A: its points lie much closer to a straight line.

(b) B has the larger gradient.

(c) Gradient describes the rate of change, while correlation strength describes how closely the observations follow a straight line. A steeper line need not have a stronger correlation.

Section 3: IB exam-style practice — Q31–50 (continued)

49.

(a) Predicted y-values: $7, 9, 11, 13$.

(b) Observed minus predicted: $1, -2, 1, 0$, respectively.

(c) At $(2, 7)$, the error has greatest magnitude, 2. The model overestimates y by 2.

50.

(a) $4(6) + 34 = 58$.

(b) $y = 4(8) + 34 = 66$.

(c) 8 lies in $[2, 10]$, so this is interpolation. A prediction at 16 is extrapolation; the same trend may not continue.

(d) At zero hours, the model is being extrapolated and cannot guarantee a score of 34. Also, $r = 0.86$ describes strong positive linear correlation; it neither means "86% of the score" nor proves that revision causes the association.

(e) $61 - 66 = -5$.

Section 4: AA HL exclusive challenge — Q51–60

51.

(a) $\bar{x} = 3$; $\bar{y} = 2(3) + 3 = 9$. Thus $5 + p + 8 + q + 14 = 45$, so $p + q = 18$. Combining this with $q - p = 2$ gives $p = 8$, $q = 10$. Mean point $(3, 9)$.

(b) $r = 0.953462\ldots \approx 0.953$: strong positive linear correlation.

(c) The mean-point condition gives only $p + q = 18$, one equation with two unknowns. This condition alone allows many pairs (p, q).

52.

(a) $T = (F - 32)/1.8$. Therefore

$$J = 3.6\left(0.32\frac{F - 32}{1.8} + 2.8\right) = 0.64F - 10.4.$$

(b) $r = 0.94$. Both scale changes are positive; neither the direction nor the strength of the linear correlation changes.

(c) $J = 0.64(59) - 10.4 = 27.36$ MJ (approximately 27.4 MJ).

(d) The observed temperatures become $41 \leq F \leq 77$. The intercept corresponds to $F = 0$, far outside that range, so it is not a meaningful prediction of energy use.

53.

(a) $(7, 11)$.

(b) $r = 0.051571\ldots \approx 0.0516$;

$$y = \frac{1}{14}x + \frac{36}{7} \approx 0.0714x + 5.14.$$

(c) Without $(7, 11)$: $r = -1$ and $y = -x + 8$.

(d) Making $|r|$ larger is not a reason to discard a genuine observation. Check, for example, for a recording error or different testing conditions. Retain valid data or explain a justified separate analysis.

Section 4: AA HL exclusive challenge — Q51–60 (continued)

54.

(a) $r = 0$ and the regression line is $y = 4$.

(b) $y = x^2$.

(c) The new mean point is $(0, (28 + k)/8)$. Since it lies on $y = 5$, $(28 + k)/8 = 5$, so $k = 12$.

(d) $r = 0$ again. Pearson's r measures linear correlation, so it does not test whether a quadratic fits. In fact, $(0, 12)$ does not lie on $y = x^2$, although the original seven points do.

55.

(a) $(\bar{x}, \bar{y}) = \left(\dfrac{3n + 32}{n + 4}, \dfrac{10n + 60}{n + 4}\right)$.

(b) Substitute the mean point into the regression line:

$$10n + 60 = 1.1(3n + 32) + 6.5(n + 4) = 9.8n + 61.2.$$

Thus $0.2n = 1.2$ and $n = 6$.

(c) $\bar{x} = 50/10 = 5$, $\bar{y} = 120/10 = 12$. Mean point $(5, 12)$. At $x = 7$, predicted $y = 1.1(7) + 6.5 = 14.2$.

(d) There are 6 observations in A and 4 in B. Their means require weights 6 and 4; an unweighted average would incorrectly treat the groups as equally large.

56.

(a) $\bar{y} = 1.8(6) + 4 = 14.8$. Original mean point $(6, 14.8)$.

(b) $x = 4 - u/3$, so

$$v = 5 + 2\left[1.8\left(4 - \frac{u}{3}\right) + 4\right] = 27.4 - 1.2u.$$

(c) $\bar{u} = 12 - 3(6) = -6$, $\bar{v} = 5 + 2(14.8) = 34.6$. Verification: $27.4 - 1.2(-6) = 34.6$.

(d) $r_{uv} = -0.87$. Reversing just the x-scale reverses the direction; translations and nonzero scale changes preserve the magnitude of r.

57.

(a) Within factory A: $r = 1$. Within factory B: $r = 1$. Combined:

$$r = -\frac{11}{21} \approx -0.524.$$

(b) $y = -\dfrac{22}{21}x + \dfrac{96}{7} \approx -1.05x + 13.7$.

(c) Within each factory, faults increase with age; the combined negative trend hides the differences between the two factories. Also, observational correlation does not establish that keeping a machine longer causes fewer faults. Factory conditions may affect the association.

Section 4: AA HL exclusive challenge — Q51–60 (continued)

58.

(a) In thousands of euros,

$$\begin{aligned} N &= 0.95y - 0.6 - (0.4x + 0.8) \\ &= 0.95(-1.8x + 24.6) - 0.6 - 0.4x - 0.8 \\ &= 21.97 - 2.11x. \end{aligned}$$

(b) $N = 12 \Rightarrow x = 9.97/2.11 = 997/211 = 4.725118\ldots$, approximately 4.73 years. Within the fitted range, predicted proceeds are at least €12 000 for $1 \leq x \leq 997/211$. Do not round the upper limit upward when applying the inequality.

(c) The regression is a prediction; individual car prices differ from the line, and $r \neq -1$. The assumed costs may also vary. The age threshold is within the observed range, but this does not make it a guarantee.

59.

(a) $42 = 18a + b$, $50 = 26a + b$. Subtraction gives $8 = 8a$, so $a = 1$, $b = 24$.

(b) $x = 60 - 24 = 36$.

(c) This reverses a y-on-x regression, which does not generally produce the appropriate x-on-y model. Also, 36 lies outside the observed x-range $[8, 30]$.

(d) $y = 20 + 24 = 44$. This predicts in the intended direction and uses interpolation. However, $r = 0.56$ is only moderate, so individual observations can differ noticeably from the prediction.

60.

(a) $(\bar{x}, \bar{y}) = (3, 7)$. At $x = 3$, L_1 gives $5.1 + 1.9 = 7$, L_2 gives $6 + 1 = 7$, and L_3 gives $4.2 + 2.8 = 7$.

(b) Actual regression: $y = 1.8x + 1.6$; $r = 0.948683\ldots \approx 0.949$.

(c) Many lines pass through a single point. Passing through the mean is a necessary property, but does not alone determine the regression gradient.

(d) With the incorrect entry: $y = 2.8x - 0.4$ and $r = 0.903696\ldots \approx 0.904$. Both correlations are strong, but one input error changes the model substantially. A large $|r|$ does not verify that the data were entered correctly.