Statistics 02.02 *[71 marks]*

The following table shows the systolic blood pressures, p mmHg, and the ages, t years, of 6 male patients at a medical clinic.

Patient	P1	P2	P3	P4	P5	P6
t (years)	40	72	35	47	21	61
p (mmHg)	105	145	100	130	95	132

1a. Determine the value of Pearson's product-moment correlation coefficient, r, for these data. *[2 marks]*

Markscheme

* This sample question was produced by experienced DP mathematics senior examiners to aid teachers in preparing for external assessment in the new MAA course. There may be minor differences in formatting compared to formal exam papers.

$r = 0.946$ **A2**

[2 marks]

1b. Interpret, in context, the value of r found in part (a) (i). *[1 mark]*

Markscheme

the value of r shows a (very) strong positive correlation between age and (systolic) blood pressure **A1**

[1 mark]

The relationship between t and p can be modelled by the regression line of p on t with equation $p = at + b$.

1c. Find the equation of the regression line of p on t. *[2 marks]*

Markscheme

$p = 1.05t + 69.3$ **A1A1**

Note: Only award marks for an equation. Award **A1** for $a = 1.05$ and **A1** for $b = 69.3$. Award **A1A0** for $y = 1.05x + 69.3$.

[2 marks]

A 50-year-old male patient enters the medical clinic for his appointment.

1d. Use the regression equation from part (b) to predict this patient's systolic blood pressure. *[2 marks]*

Markscheme

122 (mmHg) **(M1)A1**

[2 marks]

1e. A 16-year-old male patient enters the medical clinic for his appointment. *[1 mark]*

Explain why the regression equation from part (b) should not be used to predict this patient's systolic blood pressure.

Markscheme

the regression equation should not be used because it involves extrapolation **A1**

[1 mark]

The following table below shows the marks scored by seven students on two different mathematics tests.

Test 1 (x)	15	23	25	30	34	34	40
Test 2 (y)	20	26	27	32	35	37	35

Let L_1 be the regression line of x on y. The equation of the line L_1 can be written in the form $x = ay + b$.

2a. Find the value of a and the value of b. *[2 marks]*

Markscheme

$a = 1.29$ and $b = -10.4$ **A1A1**

[2 marks]

2b. Let L_2 be the regression line of y on x. The lines L_1 and L_2 pass through the same point with coordinates $(p\ ,\ q)$. *[3 marks]*

Find the value of p and the value of q.

Markscheme

recognising both lines pass through the mean point **(M1)**

$p = 28.7$, $q = 30.3$ **A2**

[3 marks]

3a. Find the value of a and the value of b. *[2 marks]*

Markscheme

$a = 1.29$ and $b = -10.4$ **A1A1**

[2 marks]

Let L_2 be the regression line of y on x. The lines L_1 and L_2 pass through the same point with coordinates $(p\ ,\ q)$.

3b. Find the value of p and the value of q. *[3 marks]*

Markscheme

recognising both lines pass through the mean point **(M1)**

$p = 28.7$, $q = 30.3$ **A2**

[3 marks]

3c. Jennifer was absent for the first test but scored 29 marks on the second test. Use an appropriate regression equation to estimate Jennifer's mark on the first test. *[2 marks]*

Markscheme

substitution into **their** x on y equation ***(M1)***

$x = 1.29082(29) - 10.3793$

$x = 27.1$ ***A1***

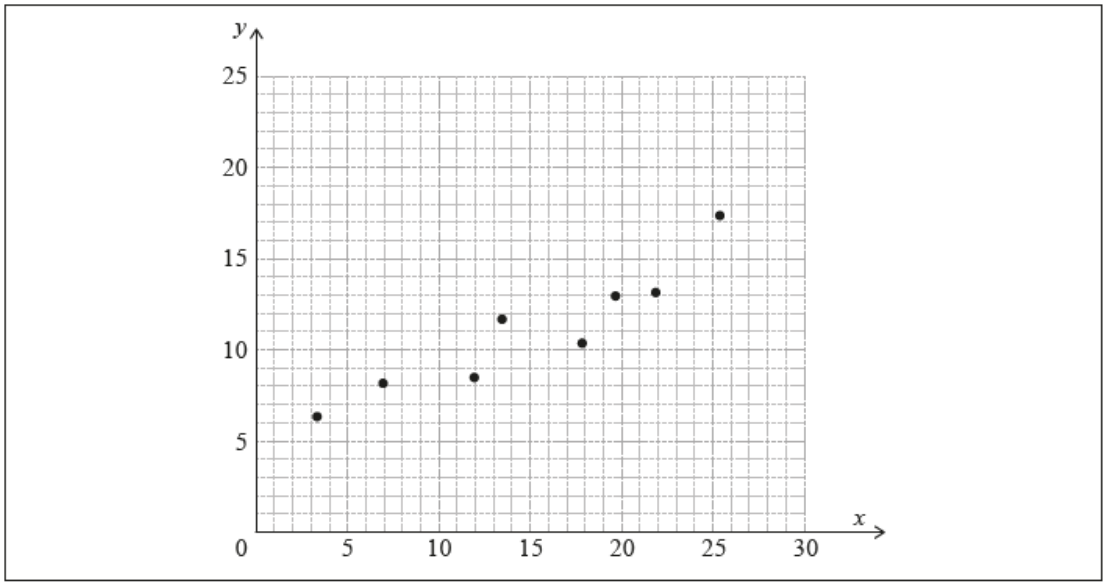
Note: Accept 27.

[2 marks]

The following table shows the data collected from an experiment.

x	3.3	6.9	11.9	13.4	17.8	19.6	21.8	25.3
y	6.3	8.1	8.4	11.6	10.3	12.9	13.1	17.3

The data is also represented on the following scatter diagram.



The relationship between x and y can be modelled by the regression line of y on x with equation $y = ax + b$, where $a, b \in \mathbb{R}$.

4a. Write down the value of a and the value of b. *[2 marks]*

Markscheme

$a = 0.433156\ldots, b = 4.50265\ldots$

$a = 0.433, b = 4.50$ ***A1A1***

[2 marks]

4b. Use this model to predict the value of y when $x = 18$. *[2 marks]*

Markscheme

attempt to substitute $x = 18$ into their equation ***(M1)***

$y = 0.433 \times 18 + 4.50$

$= 12.2994\ldots$

$= 12.3$ ***A1***

[2 marks]

4c. Write down the value of $\bar{x}$ and the value of $\bar{y}$. *[1 mark]*

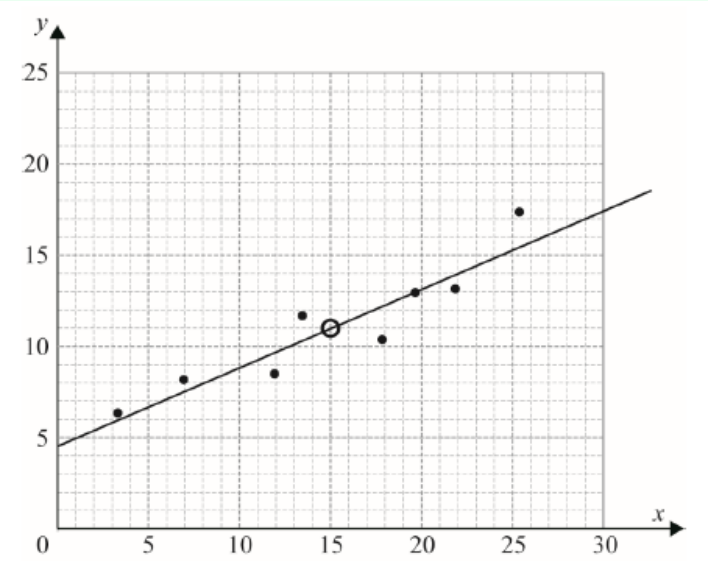
Markscheme

$\bar{x} = 15, \bar{y} = 11$ ***A1***

[1 mark]

4d. Draw the line of best fit on the scatter diagram. *[2 marks]*

Markscheme



A1A1

Note: Award marks as follows:

A1 for a straight line going through $(15,\ 11)$

A1 for intercepting the y-axis between their $b \pm 1.5$ (when their line is extended), which includes all the data for $3.3 \leq x \leq 25.3$.

If the candidate does not use a ruler, award ***A0A1*** where appropriate.

[2 marks]

At a café, the waiting time between ordering and receiving a cup of coffee is dependent upon the number of customers who have already ordered their coffee and are waiting to receive it.

Sarah, a regular customer, visited the café on five consecutive days. The following table shows the number of customers, x, ahead of Sarah who have already ordered and are waiting to receive their coffee and Sarah's waiting time, y minutes.

Number of customers (x)	3	9	11	10	5
Sarah's waiting time (y)	6	10	12	11	6

The relationship between x and y can be modelled by the regression line of y on x with equation $y = ax + b$.

5a. Find the value of a and the value of b. *[2 marks]*

Markscheme

$a = 0.805084\ldots$ and $b = 2.88135\ldots$

$a = 0.805$ and $b = 2.88$ ***A1A1***

[2 marks]

5b. Write down the value of Pearson's product-moment correlation coefficient, r. *[1 mark]*

Markscheme

$r = 0.97777\ldots$

$r = 0.978$ ***A1***

[1 mark]

5c. Interpret, in context, the value of a found in part (a)(i). *[1 mark]*

Markscheme

a represents the (average) increase in waiting time (0.805 mins) per additional customer (waiting to receive their coffee) ***R1***

[1 mark]

5d. On another day, Sarah visits the café to order a coffee. Seven customers have already ordered their coffee and are waiting to receive it. *[2 marks]*

Use the result from part (a)(i) to estimate Sarah's waiting time to receive her coffee.

Markscheme

attempt to substitute $x = 7$ into their equation ***(M1)***

$8.51693\ldots$

8.52 (mins) ***A1***

[2 marks]

Lucy sells hot chocolate drinks at her snack bar and has noticed that she sells more hot chocolates on cooler days. On six different days, she records the maximum daily temperature, T, measured in degrees centigrade, and the number of hot chocolates sold, H. The results are shown in the following table.

Maximum temperature (T)	14	8	4	18	13	11
Number of hot chocolates (H)	79	143	191	58	84	105

The relationship between H and T can be modelled by the regression line with equation $H = aT + b$.

6a. Find the value of a and of b. *[3 marks]*

Markscheme

valid approach ***(M1)***

eg correct value for a or b (or for r or $r^2 = 0.962839$ seen in (ii))

$a = -9.84636,\ b = 221.592$

$a = -9.85,\ b = 222$ ***A1A1 N3***

[3 marks]

6b. Write down the correlation coefficient. *[1 mark]*

Markscheme

-0.981244

$r = -0.981$ ***A1 N1***

[1 mark]

6c. Using the regression equation, estimate the number of hot chocolates that Lucy will sell on a day when the maximum temperature is $12\,^{\circ}\mathrm{C}$. *[2 marks]*

Markscheme

correct substitution into their equation ***(A1)***

eg $-9.85 \times 12 + 222$

103.435 (103.8 from 3 sf)

103 (hot chocolates) ***A1 N2***

[2 marks]

The number of messages, M, that six randomly selected teenagers sent during the month of October is shown in the following table. The table also shows the time, T, that they spent talking on their phone during the same month.

Time spent talking on their phone (T minutes)	50	55	105	128	155	200
Number of messages (M)	358	340	740	731	800	992

The relationship between the variables can be modelled by the regression equation $M = aT + b$.

7a. Write down the value of a and of b. *[3 marks]*

Markscheme

evidence of set up ***(M1)***

eg correct value for a or b (accept $r = 0.966856$)

$4.30161, 163.330$

$a = 4.30, b = 163$ (accept $y = 4.30x + 163$) ***A1A1 N3***

[3 marks]

7b. Use your regression equation to predict the number of messages sent by a teenager that spent 154 minutes talking on their phone in October. *[3 marks]*

Markscheme

valid approach ***(M1)***

eg $4.30\,(154) + 163$

eg 825.778 (825.2 from 3 sf values) ***(A1)***

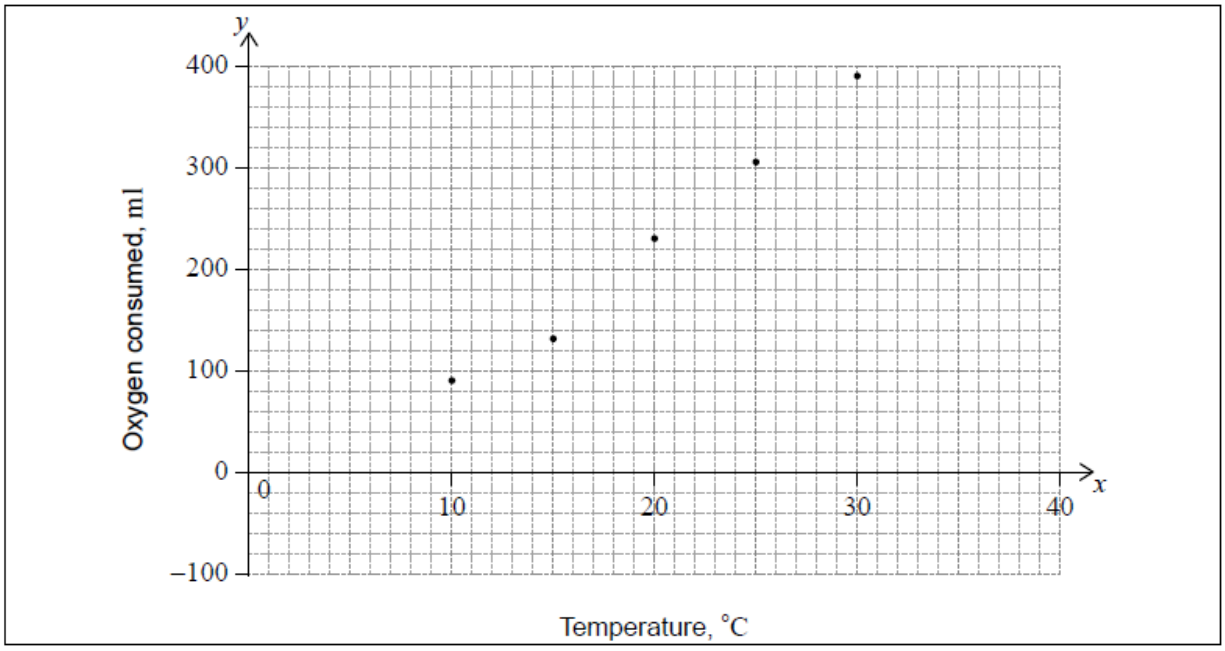
number of messages $= 826$ (must be an integer) ***A1 N3***

[3 marks]

Colorado beetles are a pest, which can cause major damage to potato crops. For a certain Colorado beetle the amount of oxygen, in millilitres (ml), consumed each day increases with temperature as shown in the following table.

Temperature, °C (x)	10	15	20	25	30
Oxygen consumed, ml (y)	90	133	230	306	391

This information has been used to plot a scatter diagram.



8a. Find the equation of the regression line of y on x. *[2 marks]*

Markscheme

* This question is from an exam for a previous syllabus, and may contain minor differences in marking or structure.

$y = 15.5x - 80$ ***(A1)(A1) (C2)***

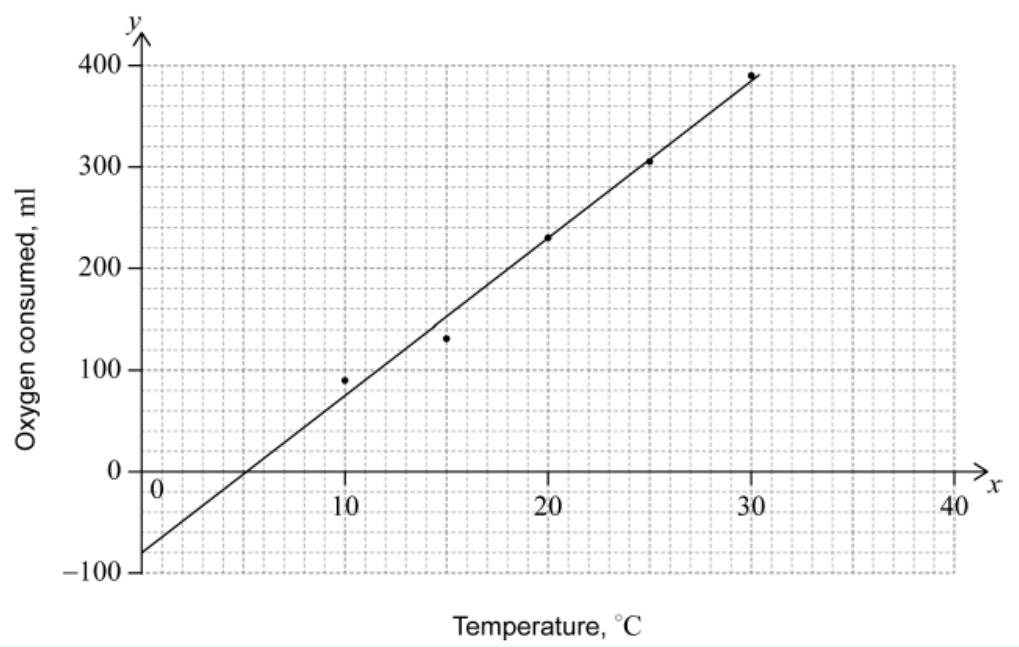
Note: Award ***(A1)*** for $15.5x$; ***(A1)*** for -80. Award at most ***(A1)(A0)*** if answer is not an equation. Award ***(A0)(A1)*(ft)** for $y = -80x + 15.5$.

[2 marks]

The mean point has coordinates (20, 230).

8b. Draw the regression line of y on x on the scatter diagram. *[2 marks]*

Markscheme



(A1)(A1)

(C2)

Note: Award ***(A1)*** for a straight line using a ruler passing through (20, 230); ***(A1)*** for correct y-intercept. If a ruler has not been used, award at most ***(A0)(A1)***.

[2 marks]

8c. In order to estimate the amount of oxygen consumed, this regression line is considered to be reliable for a temperature x such that $a \leq x \leq b$. *[2 marks]*

Write down the value of a and of b.

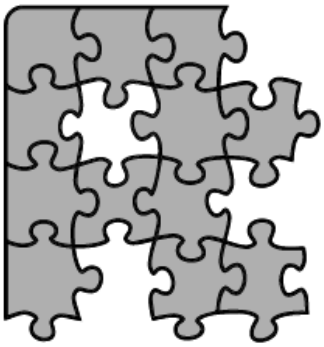
Markscheme

$a = 10$ **AND** $b = 30$ ***(A1)(A1) (C2)***

Note: Accept [10, 30] or $10 \leq x \leq 30$.

[2 marks]

A jigsaw puzzle consists of many differently shaped pieces that fit together to form a picture.



Jill is doing a 1000-piece jigsaw puzzle. She started by sorting the edge pieces from the interior pieces. Six times she stopped and counted how many of each type she had found. The following table indicates this information.

Edge pieces (x)	16	31	39	55	84	115
Interior pieces (y)	89	239	297	402	580	802

Jill models the relationship between these variables using the regression equation $y = ax + b$.

9a. Write down the value of a and of b. *[3 marks]*

Markscheme

* This question is from an exam for a previous syllabus, and may contain minor differences in marking or structure.

valid approach ***(M1)***

eg correct value for a or b (ignore incorrect labels)

$a = 6.92986$, $b = 8.80769$

$a = 6.93$, $b = 8.81$ (accept $y = 6.93x + 8.81$) ***A1A1 N3***

[3 marks]

9b. Use the model to predict how many edge pieces she had found when she had sorted a **total** of 750 pieces. *[3 marks]*

Markscheme

valid approach ***(M1)***

eg 750 = $x + y$, edge + interior = 750

correct working ***(A1)***

eg 750 − x = 6.9298x + 8.807 , 93.4684

93 (pieces) (accept 94) ***A1 N3***

[3 marks]

The marks obtained by nine Mathematical Studies SL students in their projects (x) and their final IB examination scores (y) were recorded. These data were used to determine whether the project mark is a good predictor of the examination score. The results are shown in the table.

Project mark (x)	10	18	14	18	10	18	6	18	14
Examination score (y)	47	61	61	47	54	68	47	54	47

10a. Use your graphic display calculator to write down $\bar{x}$, the mean project mark. *[1 mark]*

Markscheme

14 ***(G1)***

[1 mark]

10b. Use your graphic display calculator to write down $\bar{y}$, the mean examination score. *[1 mark]*

Markscheme

54 ***(G1)***

[1 mark]

10c. Use your graphic display calculator to write down r, Pearson's product–moment correlation coefficient. *[2 marks]*

Markscheme

0.5 ***(G2)***

[2 marks]

The equation of the regression line y on x is $y = mx + c$.

10d. Find the exact value of m and of c for these data. *[2 marks]*

Markscheme

$m = 0.875$, $c = 41.75$ $\left(m = \frac{7}{8}, c = \frac{167}{4}\right)$ ***(A1)(A1)***

Note: Award ***(A1)*** for 0.875 seen. Award ***(A1)*** for 41.75 seen. If 41.75 is rounded to 41.8 do not award ***(A1)***.

[2 marks]

10e. Show that the point M $(\bar{x}, \bar{y})$ lies on the regression line y on x. *[2 marks]*

Markscheme

$y = 0.875(14) + 41.75$ ***(M1)***

Note: Award ***(M1)*** for their correct substitution into their regression line. Follow through from parts (a)(i) and (b)(i).

$= 54$

and so the mean point lies on the regression line ***(A1)***

(accept 54 is $\bar{y}$, the mean value of the y data)

Note: Do not award ***(A1)*** unless the conclusion is explicitly stated and the 54 seen. The ***(A1)*** can be awarded only if their conclusion is consistent with their equation and it lies on the line.

The use of 41.8 as their c value precludes awarding ***(A1)***.

OR

$54 = 0.875(14) + 41.75$ ***(M1)***

$54 = 54$

Note: Award ***(M1)*** for their correct substitution into their regression line. Follow through from parts (a)(i) and (b)(i).

and so the mean point lies on the regression line ***(A1)***

Note: Do not award ***(A1)*** unless the conclusion is explicitly stated. Follow through from part (a).

The use of 41.8 as their c value precludes the awarding of ***(A1)***.

[2 marks]

A tenth student, Jerome, obtained a project mark of 17.

10f. Use the regression line y on x to estimate Jerome's examination score. *[2 marks]*

Markscheme

$y = 0.875(17) + 41.75$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into their regression line.

$= 56.6$ (56.625) ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (b)(i).

[2 marks]

10g. Justify whether it is valid to use the regression line y on x to estimate Jerome's examination score. *[2 marks]*

Markscheme

the estimate is valid ***(A1)***

since this is interpolation **and** the correlation coefficient is large enough ***(R1)***

OR

the estimate is not valid ***(A1)***

since the correlation coefficient is not large enough ***(R1)***

Note: Do not award ***(A1)(R0)***. The ***(R1)*** may be awarded for reasoning based on strength of correlation, but do not accept "correlation coefficient is not strong enough" or "correlation is not large enough".

Award ***(A0)(R0)*** for this method if no numerical answer to part (a)(iii) is seen.

[2 marks]

10h. In his final IB examination Jerome scored 65. *[2 marks]*

Calculate the percentage error in Jerome's estimated examination score.

Markscheme

$\left|\frac{56.6-65}{65}\right| \times 100$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into percentage error formula. Follow through from part (c)(i).

= 12.9 (%)(12.9230...) ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (c)(i). Condone use of percentage symbol. Award ***(G0)*** for an answer of −12.9 with no working.

[2 marks]