1. Gary compared the total attendance, x, at home matches and the total number of goals, y, scored at home during a season for each of 12 football teams playing in a league. He correctly calculated:

$$S_{xx} = 1022500 \qquad S_{yy} = 130.9 \qquad S_{xy} = 8825$$

(a) Calculate the product moment correlation coefficient for these data.

(2)

(b) Interpret the value of the correlation coefficient.

(1)

Helen was given the same data to analyse. In view of the large numbers involved she decided to divide the attendance figures by 100. She then calculated the product moment correlation coefficient between $\frac{x}{100}$ and y.

(c) Write down the value Helen should have obtained.

(1)
(Total 4 marks)

2. The blood pressures, p mmHg, and the ages, t years, of 7 hospital patients are shown in the table below.

Patient	A	B	C	D	E	F	G
t	42	74	48	35	56	26	60
p	98	130	120	88	182	80	135

$$\left[\sum t = 341,\ \sum p = 833,\ \sum t^2 = 18181,\ \sum p^2 = 106397,\ \sum tp = 42948\right]$$

(a) Find S_{pp}, S_{tp} and S_{tt} for these data.

(4)

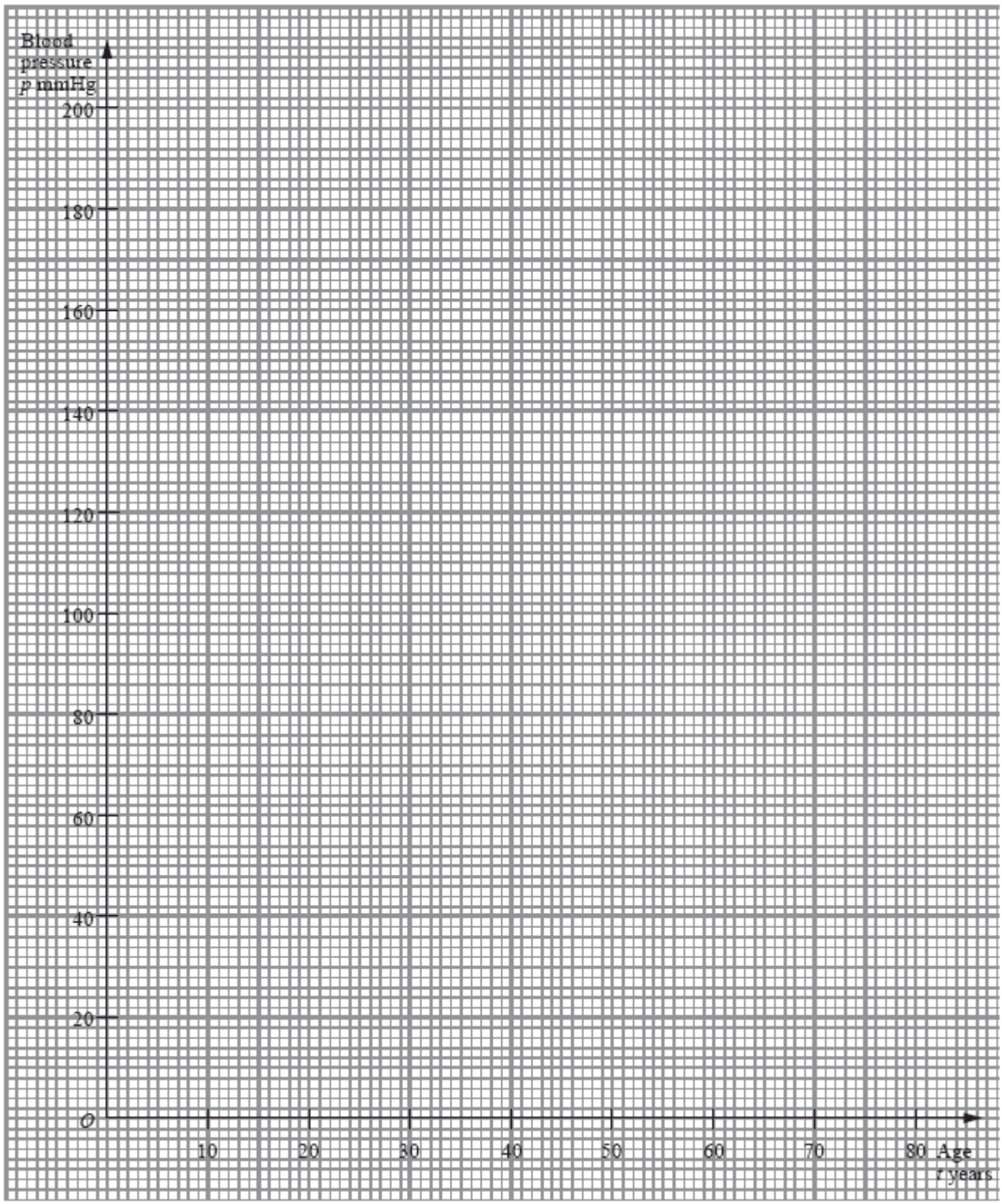
(b) Calculate the product moment correlation coefficient for these data.

(3)

(c) Interpret the correlation coefficient.

(1)

(d) On the graph paper below, draw the scatter diagram of blood pressure against age for these 7 patients.



(2)

(e) Find the equation of the regression line of p on t.

(4)

(f) Plot your regression line on your scatter diagram.

(2)

(g) Use your regression line to estimate the blood pressure of a 40 year old patient.

(2)

(Total 18 marks)

3. The volume of a sample of gas is kept constant. The gas is heated and the pressure, p, is measured at 10 different temperatures, t. The results are summarised below.

$\Sigma p = 445$ $\Sigma p^2 = 38\ 125$ $\Sigma t = 240$ $\Sigma t^2 = 27\ 520$ $\Sigma pt = 26\ 830$

(a) Find S_{pp} and S_{pt}.

(3)

Given that $S_{tt} = 21\ 760$,

(b) calculate the product moment correlation coefficient.

(2)

(c) Give an interpretation of your answer to part (b).

(1)

(Total 6 marks)

4. In a study of how students use their mobile telephones, the phone usage of a random sample of 11 students was examined for a particular week.

The total length of calls, y minutes, for the 11 students were

17, 23, 35, 36, 51, 53, 54, 55, 60, 77, 110

(a) Find the median and quartiles for these data.

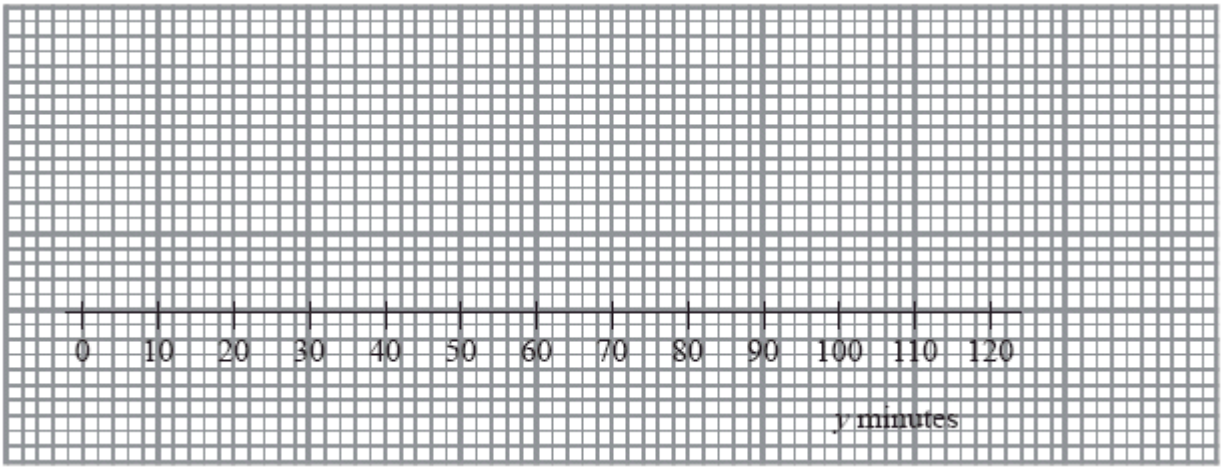
(3)

A value that is greater than $Q_3 + 1.5 \times (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ is defined as an outlier.

(b) Show that 110 is the only outlier.

(2)

(c) Using the graph below draw a box plot for these data indicating clearly the position of the outlier.



(3)

The value of 110 is omitted.

(d) Show that S_{yy} for the remaining 10 students is 2966.9

(3)

These 10 students were each asked how many text messages, x, they sent in the same week.

The values of S_{xx} and S_{xy} for these 10 students are $S_{xx} = 3463.6$ and $S_{xy} = -18.3$.

(e) Calculate the product moment correlation coefficient between the number of text messages sent and the total length of calls for these 10 students.

(2)

A parent believes that a student who sends a large number of text messages will spend fewer minutes on calls.

(f) Comment on this belief in the light of your calculation in part (e).

(1)
(Total 14 marks)

5. As part of a statistics project, Gill collected data relating to the length of time, to the nearest minute, spent by shoppers in a supermarket and the amount of money they spent.
Her data for a random sample of 10 shoppers are summarised in the table below, where t represents time and £m the amount spent over £20.

t (minutes)	£m
15	–3
23	17
5	–19
16	4
30	12
6	–9
32	27
23	6
35	20
27	6

(a) Write down the actual amount spent by the shopper who was in the supermarket for 15 minutes.

(1)

(b) Calculate S_{tt}, S_{mm} and S_{tm}.

(You may use $\Sigma t^2 = 5478$ $\Sigma m^2 = 2101$ $\Sigma tm = 2485$)

(6)

(c) Calculate the value of the product moment correlation coefficient between t and m.

(3)

(d) Write down the value of the product moment correlation coefficient between t and the actual amount spent. Give a reason to justify your value.

(2)

On another day Gill collected similar data. For these data the product moment correlation coefficient was 0.178

(e) Give an interpretation to both of these coefficients.

(2)

(f) Suggest a practical reason why these two values are so different.

(1)
(Total 15 marks)

6. Students in Mr Brawn's exercise class have to do press-ups and sit-ups. The number of press-ups x and the number of sit-ups y done by a random sample of 8 students are summarised below.

$\Sigma x = 272$, $\Sigma x^2 = 10\ 164$, $\Sigma xy = 11\ 222$,

$\Sigma y = 320$, $\Sigma y^2 = 13\ 464$.

(a) Evaluate S_{xx}, S_{yy} and S_{xy}.

(4)

(b) Calculate, to 3 decimal places, the product moment correlation coefficient between x and y.

(3)

(c) Give an interpretation of your coefficient.

(2)

(d) Calculate the mean and the standard deviation of the number of press-ups done by these students.

(4)

Mr Brawn assumes that the number of press-ups that can be done by any student can be modelled by a normal distribution with mean μ and standard deviation σ. Assuming that μ and σ take the same values as those calculated in part (d),

(e) find the value of a such that $P(\mu - a < X < \mu + a) = 0.95$.

(3)

(f) Comment on Mr Brawn's assumption of normality.

(2)
(Total 18 marks)

7. A researcher thinks there is a link between a person's height and level of confidence. She measured the height h, to the nearest cm, of a random sample of 9 people. She also devised a test to measure the level of confidence c of each person. The data are shown in the table below.

h	179	169	187	166	162	193	161	177	168
c	569	561	579	561	540	598	542	565	573

[You may use $\sum h^2 = 272\,094$, $\sum c^2 = 2\,878\,966$, $\sum hc = 884\,484$]

(a) Draw a scatter diagram to illustrate these data.

(4)

(b) Find exact values of S_{hc} S_{hh} and S_{cc}.

(4)

(c) Calculate the value of the product moment correlation coefficient for these data.

(3)

(d) Give an interpretation of your correlation coefficient.

(1)

(e) Calculate the equation of the regression line of c on h in the form $c = a + bh$.

(3)

(f) Estimate the level of confidence of a person of height 180 cm.

(2)

(g) State the range of values of h for which estimates of c are reliable.

(1)
(Total 18 marks)

8. A company owns two petrol stations P and Q along a main road. Total daily sales in the same week for P (£p) and for Q (£q) are summarised in the table below.

	p	q
Monday	4760	5380
Tuesday	5395	4460
Wednesday	5840	4640
Thursday	4650	5450
Friday	5365	4340
Saturday	4990	5550
Sunday	4365	5840

When these data are coded using $x = \dfrac{p - 4365}{100}$ and $y = \dfrac{q - 4340}{100}$,

$$\Sigma x = 48.1, \quad \Sigma y = 52.8, \quad \Sigma x^2 = 486.44, \quad \Sigma y^2 = 613.22 \text{ and } \Sigma xy = 204.95.$$

(a) Calculate S_{xy}, S_{xx} and S_{yy}.

(4)

(b) Calculate, to 3 significant figures, the value of the product moment correlation coefficient between x and y.

(3)

(c) (i) Write down the value of the product moment correlation coefficient between p and q.

(ii) Give an interpretation of this value.

(2)
(Total 9 marks)

1. (a) $r = \dfrac{8825}{\sqrt{1022500 \times 130.9}}$, = awrt **0.763** M1 A1 2

Note

M1 for a correct expression, square root required
Correct answer award 2/2

(b) Teams with high attendance scored more goals
(oe, statement in context) B1 1

Note

Context required (attendance and goals). Condone causality.
B0 for 'strong positive correlation between attendance and goals' on its own oe

(c) 0.76(3) B1ft 1

Note

Value required.
Must be a correlation coefficient between –1 and +1 inclusive.
B1ft for 0.76 or better or same answer as their value from part (a) to at least 2 d.p.

[4]

2. (a) $S_{pp} = 106397 - \dfrac{833^2}{7} = 7270$ M1 A1

$S_{pp} = 42948 - \dfrac{341 \times 833}{7} = 2369,$

$S_{tt} = 18181 - \dfrac{341^2}{7} = 1569.42857....$ or $\dfrac{10986}{7}$ A1 A1 4

Note

M1 for at least one correct expression

1st A1 for S_{pp} = 7270, 2nd A1 for S_{tp} = 2369 or 2370,
3rd A1 for S_{tt} = awrt 1570

(b) $r = \dfrac{2369}{\sqrt{7270 \times 1569.42857...}}$ M1 A1ft

= 0.7013375 **awrt (0.701)** A1 3

Note

M1 for attempt at correct formula and at least one correct value (or correct ft) M0 for $\dfrac{42948}{\sqrt{106397 \times 18181}}$

A1ft All values correct or correct ft. Allow for an answer of 0.7 or 0.70 Answer only: awrt 0.701 is 3/3, answer of 0.7 or 0.70 is 2/3

(c) (Pmcc shows positive correlation.)

Older patients have higher blood pressure B1 1

Note

B1 for comment in context that interprets the fact that correlation is positive, as in scheme.

Must mention age and blood pressure in words, not just "*t*" and "*p*".

(d) Points plotted correctly on graph: –1 each error or omission

(within one square of correct position) B2 2

Note

Record 1 point incorrect as B1B0 on epen. [NB overlay for (60, 135) is slightly wrong]

(e) $b = \dfrac{2369}{1569.42857...} = 1.509466...$ M1 A1

$a = \dfrac{833}{7} - b \times \dfrac{341}{7} = 45.467413...$ M1

$P = 45.5+1.51t$ A1 4

Note

1st M1 for use of the correct formula for b, ft their values from (a)

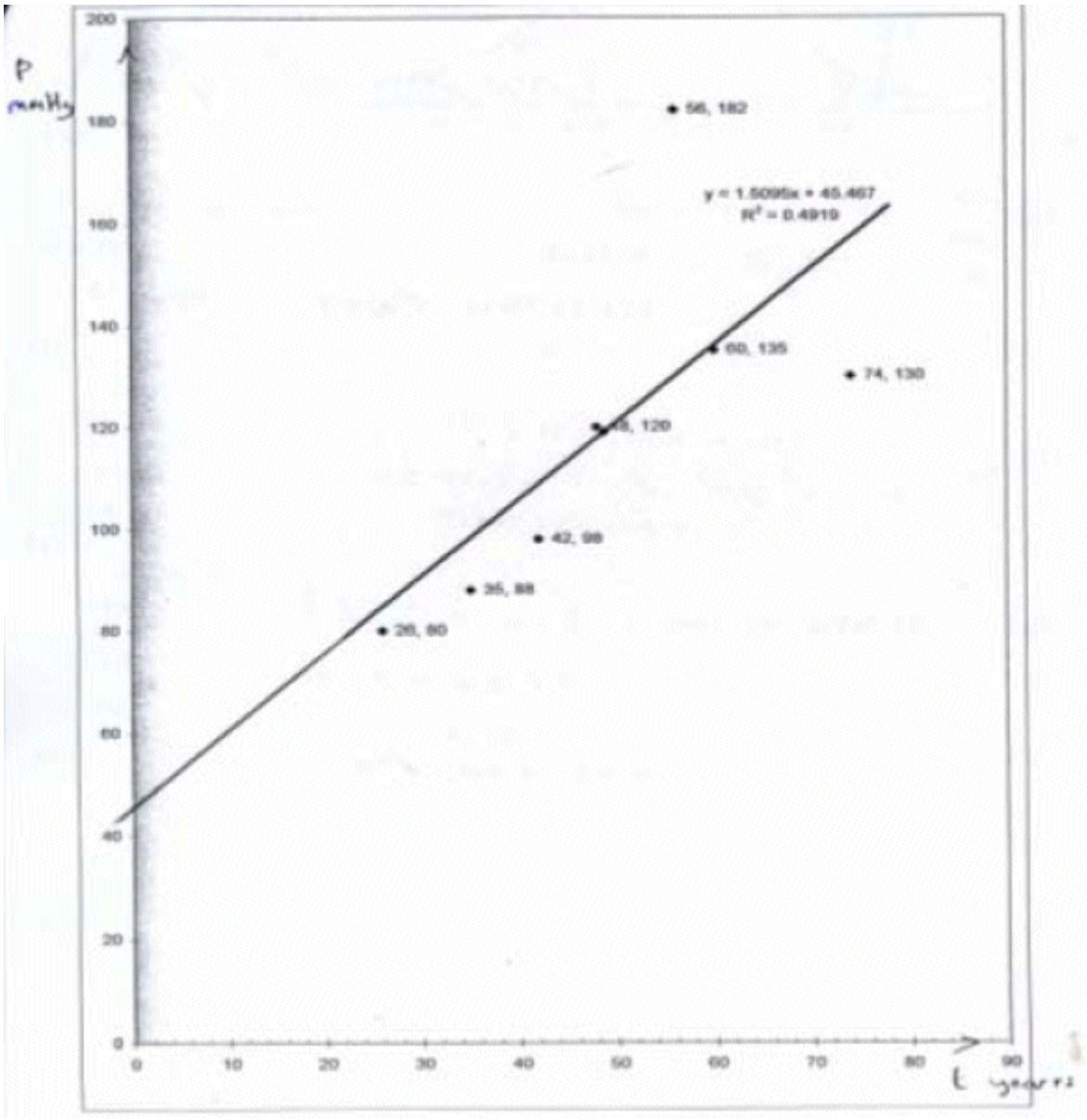
1st A1 allow 1.5 or better

2nd M1 for use of $\bar{y} - b\bar{x}$ with their values

2nd A1 for full equation with a = awrt 45.5 and b = awrt 1.51. Must be p in terms of t, not x and y.

(f) Line drawn with correct intercept, and gradient B1ft B1 2

Diagram for (d) + (f)



Note

1st B1ft	ft their intercept (within one square). You may have to extend their line.
2nd B1	for correct gradient i.e. parallel to given line (Allow 1 square out when $t = 80$)

(g) $t = 40$, $p = 105.84...$ from equation or graph. **awrt 106** M1 A1 2

Note

M1 for clear use of their equation with $t = 40$ or correct value from their graph.

A1 for awrt 106. Correct answer only (2/2) otherwise look for evidence on graph to award M1

[18]

3. (a) $\left(S_{pp} =\right) 38125 - \frac{445^2}{10}$ M1

$= 18322.5$ awrt 18300 A1

$\left(S_{pp} =\right) 26830 - \frac{445 \times 240}{10}$

$= 16150$ awrt 16200 A1 3

Note

M1 for seeing a correct expression

$38125 - \frac{445^2}{10}$ or $26830 - \frac{445 \times 240}{10}$

If no working seen, at least one answer must be exact to score M1 by implication.

(b) $r = \frac{"16150"}{\sqrt{"18322.5" \times 21760}}$ Using their values for method M1

$= 0.8088...$ awrt 0.809 A1 2

Note

Square root and their values with 21760 all in the right places required for method.
Anything which rounds to (awrt) 0.809 for A1.

(c) As the temperature increases
the pressure increases. B1 1

Note

Require a correct statement in **context** using temperature/heat and pressure for B1.

Don't allow "as t increases p increases".

Don't allow proportionality.

Positive correlation only is B0 since there is no interpretation.

[6]

4. (a) $Q_2 = 53$, $Q_1 = 35$, $Q_3 = 60$ B1, B1, B1 3

Note

1st B1 for median
2nd B1 for lower quartile
3rd B1 for upper quartile

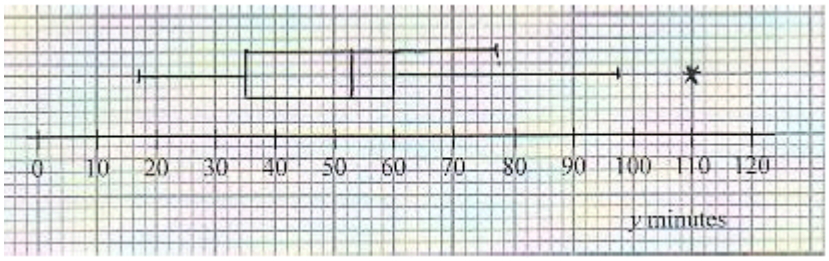
(b) $Q_3 - Q_1 = 25 \Rightarrow Q_1 - 1.5 \times 25 = -2.5$ (no outlier) M1

$Q_3 + 1.5 \times 25 = 97.5$ (so 110 is an outlier) A1 2

Note

M1 for attempt to find one limit
A1 for both limits found and correct. No explicit comment about outliers needed.

(c)



M1
A1ft
A1ft 3

Note

M1 for a box and two whiskers
1st A1ft for correct position of box, median and quartiles. Follow through their values.
2nd A1ft for 17 and 77 or "their" 97.5 and * . If 110 is not an outlier then score A0 here. Penalise no gap between end of whisker and outlier. Must label outlier, needn't be with * .

Accuracy should be within the correct square so 97 or 98 will do for 97.5

(d) $\sum y = 461, \sum y^2 = 24\,219 \therefore S_{yy} = 24219 - \frac{461^2}{10}, = 2966.9(*)$ B1, B1, B1cso 3

Note

1st B1 for $\sum y$ N.B. $\left(\sum y\right)^2$ = 212521 and can imply this mark

2nd B1 for $\sum y^2$ or at least three correct terms of $\sum (y - \bar{y})^2$ seen.

3rd B1 for complete correct expression seen leading to 2966.9. So all 10 terms of $\sum (y - \bar{y})^2$

(e) $r = \frac{-18.3}{\sqrt{3463.6 \times 2966.9}}$ or $\frac{-18.3}{3205.64...} = -0.0057$

AWRT – 0.006 or -6×10^{-3} M1 A1 2

Note

M1 for attempt at correct expression for *r*. Can ft their S_{yy} for M1.

(f) *r* suggests correlation is close to zero so parent's claim is not justified B1 1

Note

B1 for comment rejecting parent's claim on basis of weak or zero correlation
Typical error is "negative correlation so comment is true" which scores B0
Weak negative or weak positive correlation is OK as the basis for their rejection.

[14]

5. (a) (£) 17 Just **17** B1 1

(b) $\Sigma t = 212$ and $\Sigma m = 61$ (Accept as totals under each column in qu.) B1, B1

$S_{tm} = 2485 - \frac{61 \times 212}{10}$, = 1191.8 awrt **1190** or 119 (3sf) M1, A1

$S_{tt} = 983.6$ (awrt 984) and $S_{mm} = 1728.9$ (awrt 1730)(or 98.4 and 173) A1, A1 6

M1 for one correct formula seen, ft. their Σt, Σm
[Use 1st A1 for 1 correct, 2nd A1 for 2 etc]

(c) $r = \frac{1191.8}{\sqrt{983.6 \times 1728.9}}$ M1, A1ft

= 0.913922... awrt **0.914** A1 3

M1 for attempt at correct formula, $\frac{2485}{\sqrt{2101 \times 5478}}$

scores M1A0A0

A1ft ft. their values for S_{tt} etc from (b) but don't give for S_{tt} = 5478 etc (see above)

Answer only (awrt 0.914) scores 3/3, 0.913 (i.e. truncation) can score M1A1ft by implication.

(d) 0.914 (Must be the same as (c) or awrt 0.914) B1ft ($|r|<1$)
e.g. linear transformation, coding does not affect coefficient (or recalculate) dB1 2

2^{nd} B1 dependent on 1^{st} B1 Accept $\sum m = 261$, $\sum m^2 = 8541$, $\sum tm = 6725 \rightarrow 0.914$

(e) 0.914 suggests longer spent shopping the more spent.
(Idea more time, more spent) B1
0.178 different amounts spent for same time. B1 2

One mark for a sensible comment relating to each coefficient

For 0.178 allow "little or no link between time and amount spent". Must be in context.

Just saying 0.914 is strong +ve correlation between amount spent and time shopping and 0.178 is weak correlation … scores B0B0.

(f) e.g. might spend short time buying 1 expensive item OR might spend a long time checking for bargains, talking, buying lots of cheap items. B1g 1

B1g for a sensible, practical suggestion showing that other factors might affect the amount spent.

E.g. different day (weekend vs weekday) or time of day (time spent queuing if busy)

[15]

6. (a) $S_{xx} = 10164 - \frac{272^2}{8} = 916$ M1,A1

Any one method, cao

$S_{yy} = 13464 - \frac{320^2}{8} = 664$ A1

cao

$S_{xy} = 11222 - \frac{272 \times 320}{8} = 342$ A1 4

cao

(**Or** 114.5,83 & 42.75)

(b) $r = \dfrac{342}{\sqrt{916 \times 664}} = 0.43852$ M1A1ftA1 3

formula, all correct ($\sqrt{608224}$), 0.439

(c) Slight / weak evidence, B1
students perform similarly in pressups and situps B1 2

context for +ve

(d) $\bar{x} = \dfrac{272}{8} = 34$ M1A1

$s = \sqrt{\dfrac{10164}{8} - 34^2} = \sqrt{114.5} = 10.700$ M1A1 4

method includes $\sqrt{\ }$, awrt 10.7

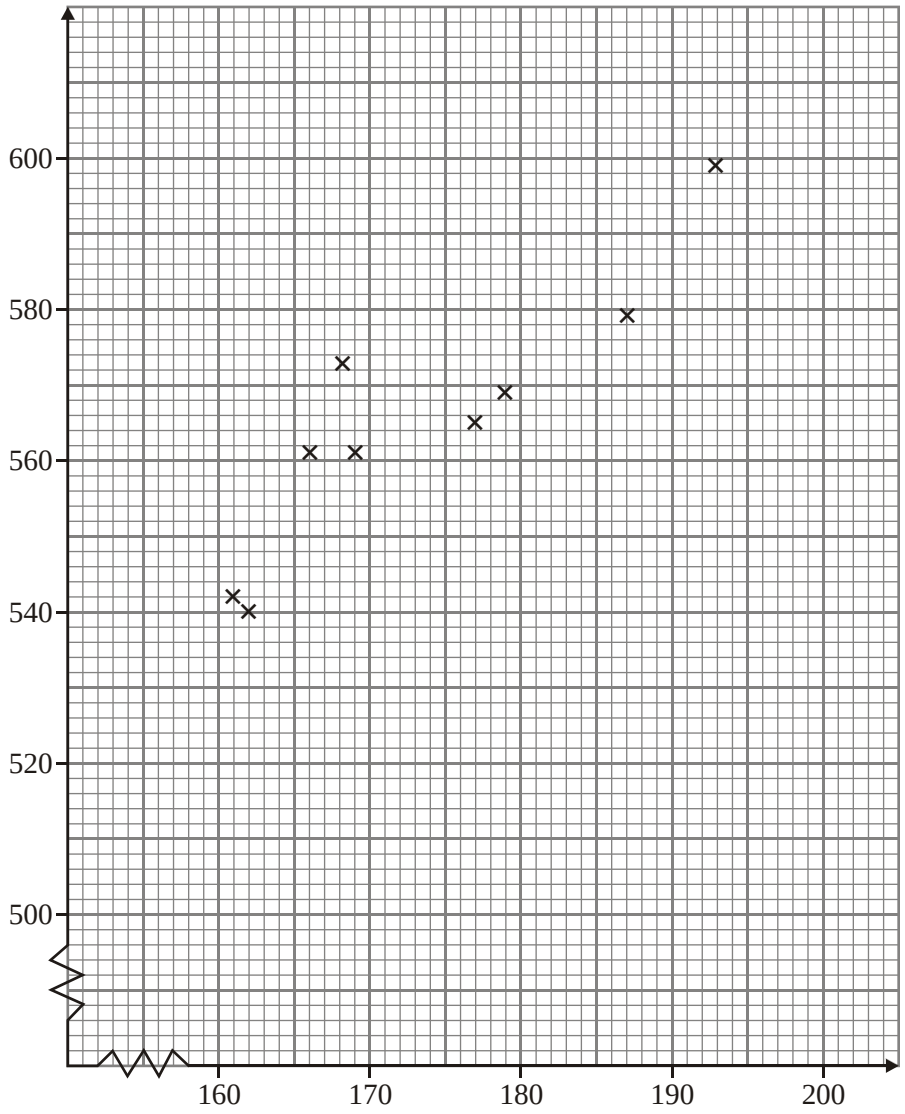
***OR** divisor (n – 1) awrt 11.4*

(e) $a = 1.96 \times 10.700\ldots = 20.9729$ (or 22.4 divisor (n – 1)) 1.96B1
$1.96 \times s$, 21.0 or 22.4 M1A1 3

(f) Pressups discrete, Normal continuous B1
Not a very good assumption B1 dep 2

[18]

7. (a)



Labels (not *x*, *y*)	B1	
Sensible scales allow axis interchange	B1	
Points	B2	4

(–1 ee)

(b) $S_{hc} = 884484 - \dfrac{1562 \times 5088}{9} = \underline{1433\frac{1}{3}}$ M1

correct use of S

$1433\frac{1}{3}$; $1433.\dot{3}$ A1

$S_{hh} = 1000\frac{2}{9}$; $S_{cc} = 2550$ A1; A1 4

$1000\frac{2}{9}$, $1000.\dot{2}$; 2550

(NB: accept :- 9; i.e.:- $159\frac{7}{27}$; $111\frac{11}{81}$; $283\frac{1}{3}$)

(c) $r = \dfrac{1433\frac{1}{3}}{\sqrt{1000\frac{2}{9} \times 2550}}$ M1

substitution in correct formula

$= \underline{0.897488\ldots.}$ A1 ft A1 3

AWRT 0.897(accept 0.8975)

(d) Taller people tend to be more confident B1 1

context

(e) $b = \dfrac{1433.\dot{3}}{1000.\dot{2}} = \underline{1.433014\ldots..}$ M1

$a = \dfrac{5088}{9} - \dfrac{1433.\dot{3}}{1000.\dot{2}} \times \dfrac{1562}{9} = \underline{316.6256\ldots}$ M1

allow use of their b

$\therefore \underline{c = 317 + 1.43h}$ (3sf) A1 3

(f) $h = 180 \Rightarrow c = 574.4$ or $574.5683\ldots.$ M1

subt. of 180

574 – 575 A1 2

(g) $161 \leq h \leq 193$ B1 1

[18]

NB (a) No graph paper ⇒ 0/4

8. (a) $S_{xy} = 204.95 - \dfrac{48.1 \times 52.8}{7} = -157.86142$

$S_{xx} = 155.92428$

$S_{yy} = 214.95714$

correct method M1

AWRT – 158/–22.6 A1

AWRT 156/22.3 A1

AWRT 215/30.7 A1 4

(b) $r = \dfrac{-157.86142}{\sqrt{155.92428... \times 214.95714...}}$ M1 A1 ft

$= \underline{-0.862269}\ldots$ (awrt -0.862) A1 3

SR: No working
r = – 0.862
B1 only

(c) (i) -0.862 B1 ft

(ii) As sales at on petrol station increases, the other decreases; limited pool of customers; close one garage B1 2

[9]

1. Typically candidates successfully used the correct formula in order to calculate the product moment correlation coefficient in part (a). However, a number of candidates lost the accuracy mark by only giving a rounded answer to two decimal places. Providing an interpretation of their value of the correlation coefficient was less straightforward. Most frequently candidates made general remarks and described the correlation as positive without relating this to the context of the question. Of those who did attempt to provide an interpretation, many failed to appreciate that it was the attendance at the matches being compared to the total number of goals scored and not the number of home matches that were played.

Part (c) was answered well overall and correct answers were often justified by accompanying statements which indicated that linear coding does not affect the product moment correlation coefficient. Some candidates, however, seemed unaware of this fact and a common mistake was to divide their original product moment correlation coefficient by 100. In addition many candidates failed to recognise the significance of them being asked to write down their answer and chose to perform a full calculation in order to obtain the product moment correlation coefficient, which sometimes led to processing errors.

2. This was a high scoring question for most candidates. The calculations in parts (a) and (b) were answered very well with very few failing to use the formulae correctly. Part (c) received a good number of correct responses but many still failed to interpret their value and simply described the correlation as strongly positive. The scatter diagram was usually plotted correctly and most knew how to calculate the equation of the regression line although some used S_{pp} instead of S_{tt} and some gave their final equation in terms of y and x instead of p and t . Plotting the line in part (f) proved quite challenging for many candidates and a number with the correct equation did not have the gradient correct. Part (g) was usually well done but some chose to use their graph rather than their equation of the line and lost the final accuracy mark.

3. The vast majority scored full marks in part (a). The most common reason for losing marks for the correlation coefficient was for rounding to less than 3 significant figures without having stated the more accurate answer first. A large proportion of candidates still believe that stating 'it's a high level of correlation' will be enough to gain the mark for interpretation. A fully contextual comment is required here, using the named variables of pressure and temperature and not just the letters p and t.

4. This question was usually answered well. In part (b) some did not realise that they needed to check the lower limit as well in order to be sure that 110 was the only outlier. Part (c) was answered very well although some lost the last mark because there was no gap between the end of their whisker and the outlier. Part (d) was answered very well and most gave the correct values for $\sum y$ and $\sum y^2$ in the appropriate formula. A few tried to use the $\sum (y - \bar{y})^2$ approach but this requires all 10 terms to be seen for a complete "show that" and this was rare.

Part (e) was answered well although some gave the answer as –5.7 having forgotten the10^{-3}, or failed to interpret their calculator correctly. Many candidates gave comments about the correlation being small or negative in part (f) but they did not give a clear reason for rejecting the parent's belief. Once again the interpretation of a calculated statistic caused difficulties.

5. Most candidates knew how to carry out the required calculations in parts (b) and (c) and these were usually completed accurately and with suitable working shown. Although the majority gave an answer of £17 in part (a) £60 and £–3 were sometimes seen. The coding on the variable *m* also caused some confusion with candidates using a value of 261 for $\hat{m}$ and then trying to combine this with the sums of squares given in the question.

In part (d) most knew that the correlation coefficient remain unchanged but some thought the value should be increased by 20 and a few candidates found new values of $\sum m, \sum m^2$ and $\sum tm$ and then seemed surprised when their correlation coefficient was unchanged. In part (e), the commonest response was to simply state that 0.914 represented strong positive correlation whilst 0.178 was weak correlation rather than attempting to interpret the values in terms of time spent shopping and amount of money spent as required. There were a number of sensible practical suggestions offered in response to part (f).

6. Parts (a) and (b) were extremely well answered by candidates; the value of 664 for S_{yy} was occasionally miscopied as 646 from part (a) to part (b). Candidates found it surprisingly difficult to obtain both marks in part (c), with a contextual relationship frequently being omitted. In part (d) the calculation of the mean was straightforward for nearly all candidates. Those candidates who were able to provide a correct formula also accurately found the standard deviation; however, too many candidates at this level were quoting an incorrect formula. Part (e) proved a good discriminator, with relatively few concise solutions; some candidates managed to obtain the correct value of *a* after a page or so of working. Only a handful of candidates were able to see that the number of press-ups is a discrete variable, whereas normal distributions are continuous.

7. This question was familiar to most candidates and many of them answered it very well. This being said, too many used scales that were not sensible for the scatter diagram and far too many ignored the instruction to 'find the exact value'. The interpretation of the correlation coefficient was rarely given in terms of the context of the question and many candidates did not give the values of *a* and *b* to 3 significant figures in spite of previous advice.

8. Parts (a) and (b) were generally well answered with many candidates gaining full marks. This being said, it was not unusual to see ridiculous values for the correlation coefficient and for candidates to follow this through into part (c). Many candidates realised that the value of the correlation coefficient would be the same in (c)(i) and those that attempted (c)(ii) often did so without reference to the context of the question.