RAFFLES INSTITUTION
H2 Mathematics 9758
2025 Year 6

Additional Practice Questions for Chapter S6: Correlation and Regression (Solution)

1 **8863/2007/01/Q8**

Seven cities in a certain country are linked by rail to the capital city. The table below shows the distance of each city from the capital and the rail fare from the city to the capital.

City	A	B	C	D	E	F	G
Distance, x km	124	44	76	148	16	180	104
Rail fare, \$ y	156	53	99	169	23	177	138

(i) Give a sketch of the scatter diagram for the data as shown on your calculator. [2]

(ii) Calculate the product moment correlation coefficient. [1]

You are given that the regression line of y on x has equation $y = 16.7 + 1.01x$, where the coefficients are given correct to 3 significant figures.

(iii) Calculate the equation of the regression line of x on y, giving your answer in the form $x = a + by$. [1]

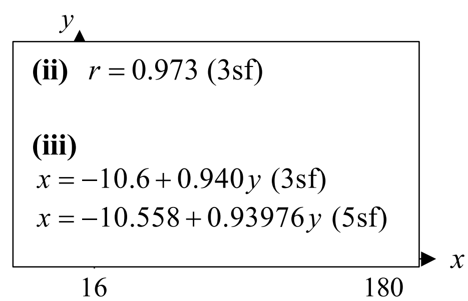
(iv) Use the appropriate regression line to estimate

(a) the rail fare from a city that is 28 km from the capital, [2]

(b) the distance of a city from the capital if the rail fare is \$198. [2]

(v) Comment briefly on the reliability of the estimates in part **(iv)**. [2]

[(ii) $r = 0.973$ (3sf) **(iii)** $x = -10.6 + 0.940y$ (3sf) **(iv)(a)** \$45 **(b)** 176 km **]**

Solution:	
(i)	 **(ii)** $r = 0.973$ (3sf) **(iii)** $x = -10.6 + 0.940y$ (3sf) $x = -10.558 + 0.93976y$ (5sf)
(iv) (a)	Given $x = 28$, use the regression line of y on x to get $y = 44.98 = 45$ (nearest integer). So the estimated rail fare is \$45.
(iv) (b)	Given $y = 198$, use the regression line of x on y, i.e. $x = -10.558 + 0.93976y$ (5s.f) to get $x = 176$ (3s.f.). So the estimated distance is 176 km.
(v)	The value $x = 28$ is within the given range of values of x where $16 \le x \le 180$. Since interpolation was used, and the correlation coefficient 0.973 is close to 1, indicating a strong positive linear correlation between the variables, the estimate obtained in **(a)** is likely to be reliable. The value $y = 198$ is beyond the given range of values of y where $23 \le y \le 177$. Since extrapolation was used, the estimate obtained in **(b)** is unlikely to be reliable.

2 **8863/2008/01/Q11**

An engineering company makes cranes. The numbers, x, sold in each three-month period for two years, together with the profits, y thousand dollars, on the sale of these cranes are given in the following table.

x	15	17	13	21	16	22	14	18
y	290	350	270	430	340	410	300	360

(i) Give a sketch of the scatter diagram for the data as shown in your calculator. [2]

(ii) Find $\bar{x}$ and $\bar{y}$, and mark the point $(\bar{x}, \bar{y})$ on your scatter diagram. [2]

(iii) Calculate the equation of the regression line of y on x, and draw this line on your scatter diagram. [2]

(iv) Calculate the product moment correlation, and comment on its value in relation to your scatter diagram. [2]

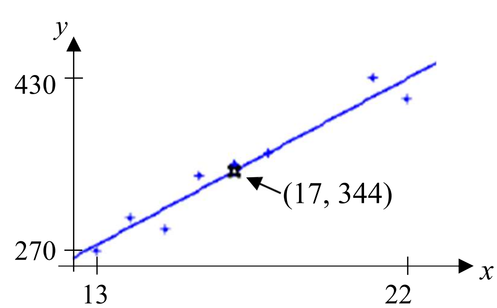
(v) For the next three-month period, the sales target is 20 cranes. Estimate the corresponding profit. [2]

(vi) The company's sales director uses the regression line in **(iii)** to predict the profit if 40 cranes were to be sold in a three-month period. Comment on the validity of this prediction. [2]

[(ii) $\bar{x} = 17, \bar{y} = 344$ **(iii)** $y = 17.1x + 53.3$ **(iv)** 0.969 **(v)** \$395,000**]**

Solution:

(i)

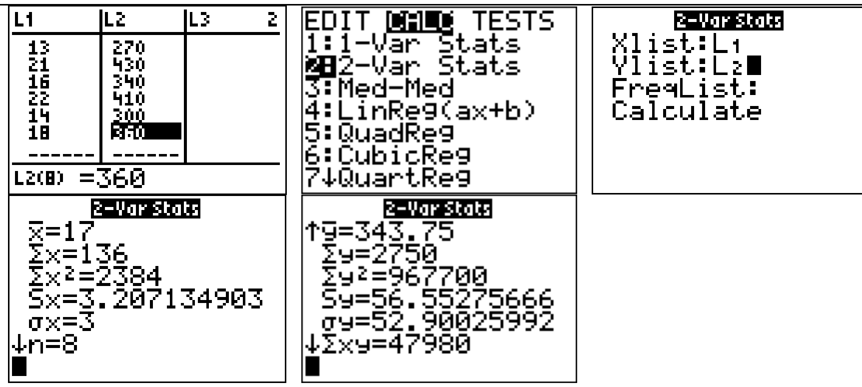


(ii)

$$\bar{x} = \frac{\sum x}{n} = 17$$

$$\bar{y} = \frac{\sum y}{n} = 343.75 = 344 \quad \text{(to nearest integer)}$$

OR: use GC to find $\bar{x}$ and $\bar{y}$:



(iii) **To use GC to draw regression line :**

Linear regression of y on x is $y = 17.1x + 53.3$ (3 s.f)

(iv) $r = 0.969$ (3 s.f)

The product moment correlation coefficient is close to 1, suggesting a strong positive linear correlation between x and y. This is congruent with the scatter diagram which shows most of the points lying close to the regression line with a positive gradient.

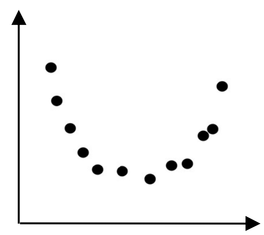
(v) When $x = 20$, $y = 17.083(20) + 53.333 = 394.993$

i.e, the estimated corresponding profit is \$395 000.

(vi) Since $x = 40$ is **outside the given range of values of** $\boldsymbol{x}$, **extrapolation** is used. The director's prediction would be unreliable as the linear model may not be applicable outside the given range of data.

3 Explain why it is advisable to draw a scatter diagram before interpreting a correlation coefficient calculated for a sample drawn from a bivariate distribution. [1]

Sketch a scatter diagram that shows an obvious relation between two variables but which yields a linear product-moment correlation coefficient that is close to zero. [1]

Solution:
A scatter diagram helps to assess if a linear relationship exists between the 2 variables, or a non-linear relationship fits better. It also helps to identify any outliers in the data that do not fit in the general pattern. One possible scatter diagram is 

4 **9740/2015/NJC/02/Q12 (modified)**

A medical officer wishes to investigate a patient's heart-beat rate h beats per minute (bpm) when he walks at different speeds s km/h. The data is shown below:

s (km/h)	1	1.5	2	2.5	3	3.5	4	4.5	5
h (bpm)	60	63	66	75	86	99	150	110	130

(i) Sketch a scatter plot of the above data. [1]

(ii) One of the values of h appears to be incorrect. Indicate the corresponding point on your diagram by labelling it P. [1]

Omit P for the remainder of this question.

(iii) Calculate the product moment correlation coefficient for this set of data. Use the equation of an appropriate regression line to predict the value of s when $h = 100$, justifying your choice of regression line. [4]

It is suggested to use one of the following two models instead:

Model (I): $h = a + bs^2$,

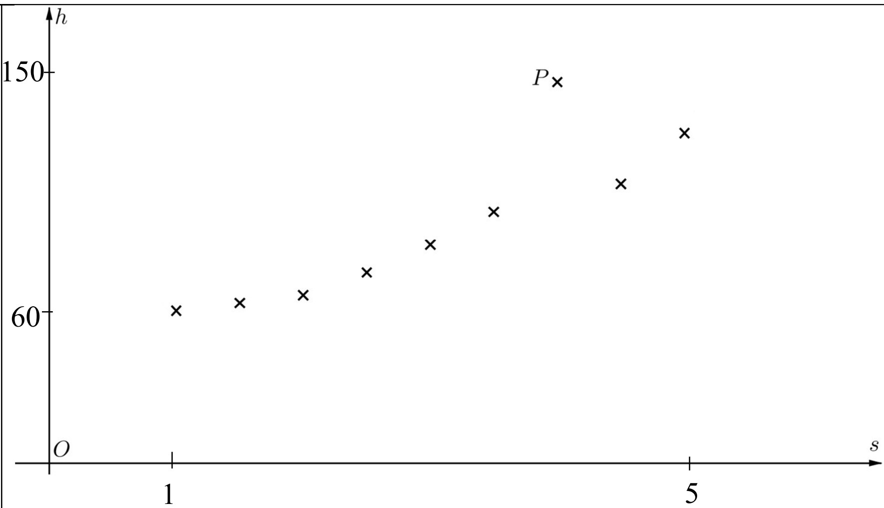
Model (II): $h = a + be^s$,

where a and b are real constants.

(iv) Determine which of the two models is a better choice, giving a reason for your answer. [2]

(v) Suppose a new data pair $\left(\bar{s}, \bar{h}\right)$ is added to the table above, where $\bar{s}$ and $\bar{h}$ are the patient's sample mean walking speed (in km/h) and his sample mean heart-beat rate (in bpm) respectively, based on the data above. Without any calculations, explain whether the equation of the regression line you have obtained in part **(iii)** would change. [2]

[(iii) 0.9812; $s = 3.67$ **(iv)** model (I) **(v)** $(\bar{s}, \bar{h})$ **]**

Solution:	
(i), (ii)	
(iii)	r-value = 0.9812; We use the h on s line since s is the independent variable. h on s line: $h = 35.851 + 17.486s$ When $h = 100$, $s = 3.669 = 3.67$ (3 s.f.)
(iv)	For model (I), r-value = 0.9898 For model (II), r-value = 0.9380 Hence model (I) is more suitable since the correlation coefficient is nearer to 1, which suggests a stronger linear relationship between h and s^2 as compared to that between h and e^s.
(v)	Since the data point $(\bar{s}, \bar{h})$ lies on the regression line, the answer in part **(iii)** will not change as the equation of the regression line is not affected. **OR** The addition of the data point $(\bar{s}, \bar{h})$ does not add further to the sum of squares of the vertical residuals between the current regression line and the data points, hence the vertical residuals remain minimised with the same line.

5 9740/2014/02/Q8

(a) Sketch a scatter diagram that might be expected when x and y are related approximately by $y = px^2 + t$ in each of the cases **(i)** and **(ii)** below. In each case your diagram should include 6 points, approximately equally spaced with respect to x, and with all x-values positive.

(i) p and t are both positive.

(ii) p is negative and t is positive. [2]

(b) The ages in months (m) and prices in dollars (P) of a random sample of ten used cars of a certain model are given in the table.

m	11	20	28	36	40	47	58	62	68	75
P	112800	102600	76500	72000	72000	69000	65800	57000	50600	47600

It is thought that the price after m months can be modelled by one of the formulae

$$P = am + b, \qquad P = c\ln m + d,$$

where a, b, c and d are constants.

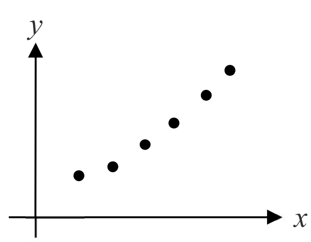
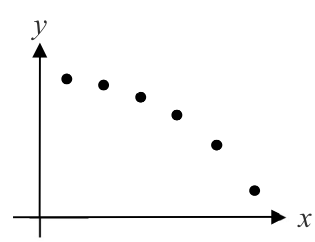
(i) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

(A) m and P,

(B) $\ln m$ and P. [2]

(ii) Explain which of $P = am + b$ and $P = c\ln m + d$ is the better model and find the equation of a suitable regression line for this model. [3]

(iii) Use the equation of your regression line to estimate the price of a car that is 50 months old. [1]

Solution:			
(a)(i)		**(ii)**	
(b)(i)	Product moment correlation coefficient between m and P is –0.9470 (4dp). Product moment correlation coefficient between $\ln m$ and P is –0.9749 (4dp).		
(ii)	–0.9749 is nearer to –1 compared to –0.9470. The scatter diagram of P on m also suggests a non-linear relationship between the 2 variables. Thus $P = c \ln m + d$ is the better model. $P = 195690 - 33660\ln m$ $P = 196000 - 33700\ln m$ (3 s.f.)		
(iii)	From GC, estimated price is \$64000 (3 s.f.)		

6 **9740/2007/02/Q11**

Research is being carried out into how the concentration of a drug in the bloodstream varies with time, measured from when the drug is given. Observations of successive times give the data shown in the following table.

Time (t minutes)	15	30	60	90	120	150	180	240	300
Concentration (x micrograms per litre)	82	65	43	37	22	19	12	6	2

It is given that the value of the product moment correlation coefficient for this data is -0.912, correct to 3 decimal places. Draw a scatter diagram for the data. [1]

Calculate the equation of the regression line of x on t. [2]

Calculate the corresponding estimated value of x when $t = 300$, and comment on the suitability of the linear model. [2]

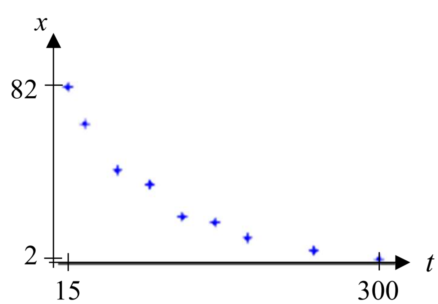
The variable y is defined by $y = \ln x$. For the variables y and t,

(i) calculate the product moment correlation coefficient and comment on its value, [2]

(ii) calculate the equation of the appropriate regression line. [3]

Use a regression line to give the best estimate that you can of the time when the drug concentration is 15 micrograms per litre. [2]

Solution:



Equation of regression line of x on t :

$x = 66.194 - 0.25970t$

$x = 66.2 - 0.260t$ (3 s.f.)

When $t = 300$, $x = 66.194 - 0.25970(300) = -11.7$ (3 s.f.)

Even though $r \approx -0.912$ indicates a high negative linear correlation between the variables, x and t, the linear model may not be suitable as the scatter diagram shows that the points follow a curved nature.

Also, it is not possible to have $x = -11.7$ as concentration of blood cannot be negative.

(i) $r = -0.994$ (3 s.f.)

	The r obtained is closer to -1 than when using variables x and t.
(ii)	$y = 4.6206 - 0.012343t$ $y = 4.62 - 0.0123t$ (3 s.f.)
	When $x = 15$, $y = \ln 15$ $\therefore \ln 15 = 4.6206 - 0.012343t$ $\Rightarrow t = 155$ (3 s.f.)

7 AJC/2010/MidYr/02/Q5

Mr. Lee recorded the length of time, y minutes, taken to travel to work when leaving home x minutes after 6.00 am on seven selected mornings. The results are as follows:

x /min	0	10	20	30	40	50	60
y /min	16	27	28	39	43	48	51

(i) Calculate the equations of the least squares regression lines of y on x and x on y. [2]

(ii) Mr. Lee forgot to add an observation for a particular morning but remembered that the two regression lines remain unchanged on its inclusion. Find this missing observation. [1]

Mr. Lee needs to report to work by 8.00 am. He leaves home x minutes after 6.00 am and he arrives at his workplace z minutes before 8.00 am.

(iii) Find z in terms of x. [1]

(iv) Hence estimate, to the nearest minutes, the latest time Mr. Lee can leave home without arriving late at work. [2]

[**(i)** $y = 0.579x + 18.6$ (3sf); $x = 1.67y - 30$ (3sf) **(ii)** (30,36) **(iii)** $z = 101.35714 - 1.57857x$ **(iv)** 7.04 am]

Solution:	
(i)	The least squares regression line y on x is $y = 0.579x + 18.6$ (3sf). The least squares regression lines x on y is $x = 1.67y - 30$ (3sf).
(ii)	Since the two regression lines pass through $(\bar{x}, \bar{y})$,ie. (30,36), the missing observation is (30,36).
(iii)	$x + y + z = 120 \Rightarrow z = 120 - x - y$ $\therefore\ z = (120 - x) - (0.57857x + 18.643)$ $= 101.35714 - 1.57857x$
(iv)	When $z \geq 0$, $x \leq \dfrac{101.35714}{1.57857} = 64.2$ (3 s.f.) $\therefore$ Latest time is 7.04 am (nearest minute)

8 9740/2016/02/Q8

A website about electric motors gives information about the percentage efficiency of motors depending on their power, measured in horsepower. Xian has copied the following table for a particular type of electric motor, but he has copied one of the efficiency values wrongly.

Power, x	1	1.5	2	3	5	7.5	10	20	30	40	50
Efficiency, y%	72.5	82.5	84.0	87.4	87.5	88.5	89.5	90.2	91.0	91.7	92.4

(i) Plot a scatter diagram on graph paper for these values, labeling the axes, using a scale of 2 cm to represent 10% efficiency on the y-axis and an appropriate scale for the x-axis. On your diagram, circle the point that Xian has copied wrongly. [2]

For parts **(ii)**, **(iii)** and **(iv)** of this question you should **exclude** the point for which Xian has copied the efficiency value wrongly.

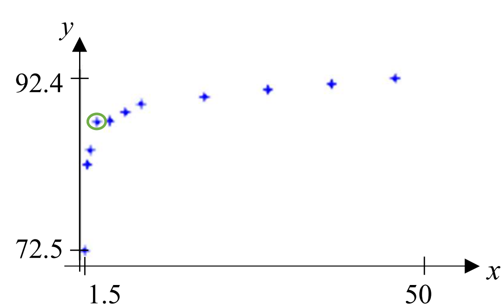
(ii) Explain from your scatter diagram why the relationship between x and y should not be modelled by an equation of the form $y = ax + b$. [1]

(iii) Suppose that the relationship between x and y are modelled by an equation of the form $y = \frac{c}{x} + d$, where c and d are constants. State with a reason whether each of c and d is positive or negative. [2]

(iv) Find the product moment correlation coefficient and the constants c and d for the model in part **(iii)**. [3]

(v) Use the model $y = \frac{c}{x} + d$ with the values of c and d found in part **(iv)**, to estimate the efficiency value (y) that Xian has copied wrongly. Give two reasons why you would expect this estimate to be reliable. [3]

[(i) Circle (3, 87.4) **(iv)** -0.980 (3sf) ; $c = -17.5$ (3sf) and $d = 91.8$ (3sf) **(v)** 85.9 **]**

Solution:	
(i)	 [Remember to use graph paper, and circle the point (3, 87.4) whose y-value is wrong.]
(ii)	Scatter diagram does not display linear relationship, since as x increases, y increases at a decreasing rate. Thus it should not be modelled as an equation of the form $y = ax + b$
(iii)	$y = \frac{c}{x} + d$

	All x and y values are positive, thus $d > 0$ to account for large values of x when $\frac{c}{x} \to 0$. Values of y increases as x increase, so $c < 0$
(iv)	Product moment correlation coefficient is -0.980 (3sf) $c = -17.5$ (3sf) and $d = 91.8$ (3sf)
(v)	Value of y when $x = 3$ is $85.922 = 85.9$ (3sf) $r = -0.97955$ is near -1 which suggest good fit of the model. Besides, the value 3 is within the given range of values of x, so we did not extrapolate the information.

9 For 3 years a factory records the number of units it produces each quarter and the total cost of producing the units. The following table summarises these results.

Units of output (x) in 1000's	15	30	57	75	10	25	49	70	20	34	60	76
Total cost (y) in £1000's	33	49	71	99	34	40	61	82	40	51	84	93

(i) Give a sketch of the scatter diagram for the data as displayed on your graphing calculator. [2]

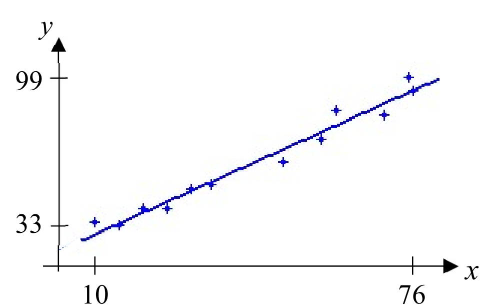
(ii) Calculate the equation of the regression line of y on x, draw this line on your diagram and calculate the product moment correlation coefficient. [3]

The selling price of each unit of output is £1.80.

(iii) Use your graph to estimate the level of output (correct to 2 significant figures) at which the total cost and the total income balance. [2]

(iv) Give a brief interpretation of what this value means. [1]

[(ii) $y = 19.3 + 0.970x$, $r = 0.983$, **(iii)** 23000 units (2 sf) **]**

Solution:	
(i)	
(ii)	$y = 19.3 + 0.970x$, $r = 0.983$ (3 s.f.)
(iii)	The regression line cuts the line $y = 1.8x$ at $x = 23.25$ (4 s.f.) The level of output required is 23000 units (2 s.f.)

(iv)	The factory makes a profit if the output is above this level. If the output is below this level, it suffers a loss.

10 **MJC Prelims 9740/2007/02/Q9**

In an investigation of a shrub, research workers measured the average width y (in cm) and stem density x (number of stems per m^2) at ten sites with the following results:

x	4	5	6	9	11	14	15	19	21	22
y	0.8	2	0.65	0.6	0	0.65	0.55	0.35	0.40	0.38

(i) Calculate the correlation coefficient for the data. **[2]**

(ii) Give a sketch of the scatter diagram for the data, as shown on your calculator, and comment on your answer in **(i)**. **[3]**

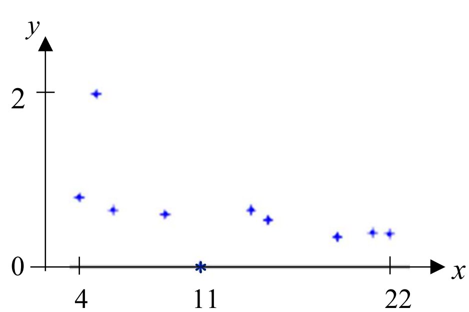
(iii) Identify the two data pairs which should be removed, and calculate the correlation coefficient, and the regression line of y on x, for the revised data. **[3]**

(iv) Use the regression line of y on x in **(iii)** to demonstrate that it is unwise to extrapolate beyond the range of the data. **[2]**

(v) A new data pair (a, b) is now obtained at an eleventh site, where $4 \le a \le 22$. Using the revised data in **(iii)** and the new data pair (a, b), the regression equation of y on x is calculated and found to be identical to that using only the revised data. Find a possible data pair (a, b). **[2]**

[(ii) $r = -0.542$ **(iii)** $r = -0.916$, $y = 0.840 - 0.0213x$ (3s.f) **(v)** $(13.75, 0.5475)$ **]**

Solution:

(i)	$r = -0.542$ (3sf)
(ii)	 The scatter diagram indicates a strong negative linear correlation between x and y. This is not consistent with the value of r in (i) due to the presence of the outliers, (5,2) and (11,0).
(iii)	The data pairs which should be removed are (5,2) and (11,0). For the revised data, $r = -0.916$ (3sf) Regression line of y on x: $y = 0.83992 - 0.021267x$

	$y = 0.840 - 0.0213x$ (3 s.f.)
(iv)	Substitute a large value of x ($x > 39$) and obtain a negative value of y, which is impossible. Note: $0.83992 - 0.021267x < 0 \Rightarrow x > 39.494$ Eg. Substituting $x = 40$ into the equation of line of y on x gives $y = -0.01076 < 0$. But in this context, y represents stem density, which can't be negative. Thus it is unwise to extrapolate beyond the range of the data.
(v)	$(a,b) = (\bar{x}, \bar{y}) = (13.75, 0.5475)$

11 **9740/2015/02/10**

In an experiment the following information was gathered about air pressure P, measured in inches of mercury, at different heights above sea-level h, measured in feet.

h	2000	5000	10000	15000	20000	25000	30000	35000	40000	45000
P	27.8	24.9	20.6	16.9	13.8	11.1	8.89	7.04	5.52	4.28

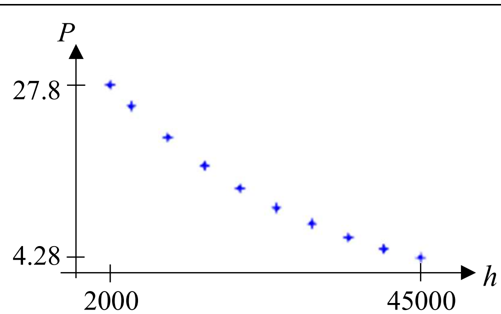
(i) Draw a scatter diagram for these values, labelling the axes. [1]

(ii) Find, correct to 4 decimal places, the product moment correlation coefficient between
(a) h and P,
(b) $\ln h$ and P,
(c) $\sqrt{h}$ and P. [3]

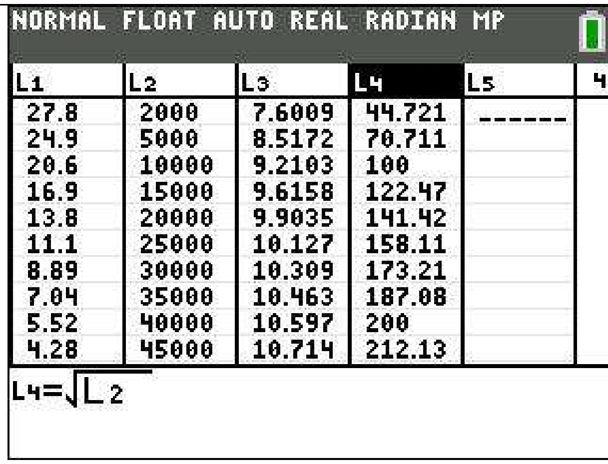
(iii) Using the most appropriate case from part **(ii)**, find the equation which best models air pressure at different heights. [3]

(iv) Given that 1 metre = 3.28 feet, re-write your equation from part **(iii)** so that it can be used to estimate the air pressure when the height is given in metres. [2]

[**(ii)** -0.9807, -0.9748, -0.9986; **(iii)** $P = 34.8 - 0.147\sqrt{h}$; **(iv)** $P = 34.8 - 0.266\sqrt{h}$]

Solution:	
(a)(i)	

(ii)



NORMAL FLOAT AUTO REAL RADIAN MP

L1	L2	L3	L4	L5
27.8	2000	7.6009	44.721	------
24.9	5000	8.5172	70.711	
20.6	10000	9.2103	100	
16.9	15000	9.6158	122.47	
13.8	20000	9.9035	141.42	
11.1	25000	10.127	158.11	
8.89	30000	10.309	173.21	
7.04	35000	10.463	187.08	
5.52	40000	10.597	200	
4.28	45000	10.714	212.13	

L4=√L2

(a)

(b)

(c)

(iii) (c) has product moment correlation coefficient nearest to -1.

Equation of best model:

$P = 34.789 - 0.14687\sqrt{h}$ (5sf)

$P = 34.8 - 0.147\sqrt{h}$ (3sf)

(iv) If height in metres is h, then height in feet is $3.28h$.

$P = 34.789 - 0.14687\sqrt{3.28h}$ (where h is in metres)

$P = 34.8 - 0.266\sqrt{h}$ (3sf)

12 **9740/2013/02/Q10**

(i) Sketch a scatter diagram that might be expected when x and y are related approximately as given in each of the cases (A), (B) and (C) below. In each case your diagram should include 6 points approximately equally spaced with respect to x, and with all x- and y-values positive. The letters a, b, c, d, e and f represent constants.

(A) $y = a + bx^2$, where a is positive and b is negative,

(B) $y = c + d\ln x$, where c is positive and d is negative,

(C) $y = e + \dfrac{f}{x}$, where e is positive and f is negative. [3]

A motoring website gives the following about the distance travelled, y km, by a certain type of car at different speeds, x kmh^{-1}, on a fixed amount of fuel.

Speed, x	88	96	104	112	120	128
Distance, y	148	147	144	138	126	107

(ii) Draw the scatter diagram for these values, labelling the axes. [1]

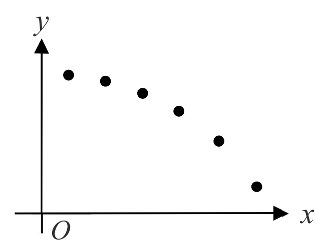
(iii) Explain which of the three cases in part **(i)** is the most appropriate for modelling these values, and calculate the product moment correlation coefficient for this case. [2]

(iv) It is required to estimate the distance travelled at a speed of 110 kmh^{-1}. Use the case that you identified in part **(iii)** to find the equation of a suitable regression line, and use your equation to find the required estimate. [3]

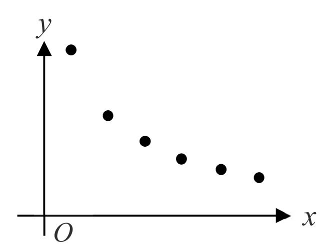
Solution:

(i)

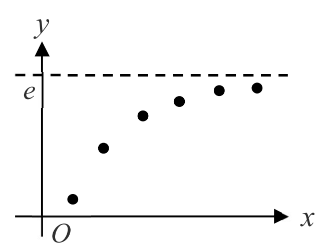
For (A): $y = a + bx^2$, with $a > 0$ and $b < 0$.

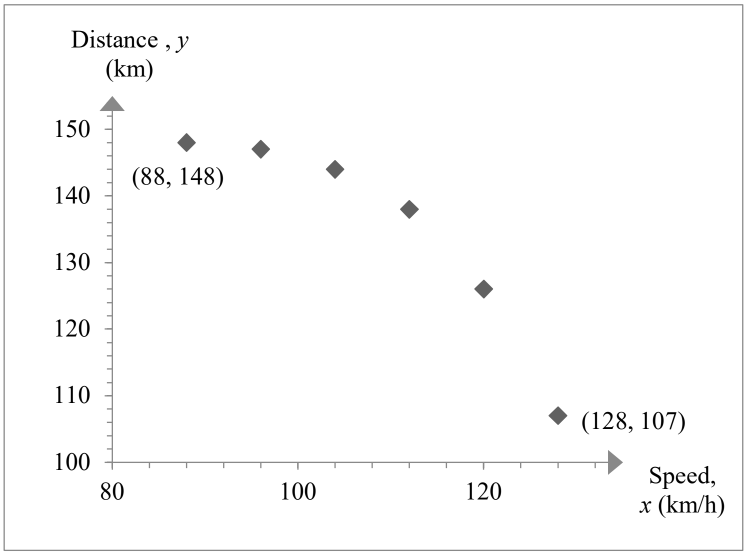


For (B): $y = c + d\ln x$, with $c > 0$ and $d < 0$.



For (C): $y = e + \dfrac{f}{x}$, with $e > 0$ and $f < 0$.



<table>
<tr><td>(ii)</td><td>
</td></tr>
<tr><td>(iii)</td><td>From the scatter diagram, we see that y decreases at an increasing rate as x increases, which is consistent with the scatter plot (A). Thus (A) is the most appropriate model.
Using model (A), the product moment correlation coefficient is -0.939 (3 s.f.).</td></tr>
<tr><td>(iv)</td><td>The required equation is $y = 189.7475 - 0.004619785x^2$, which is $y = 190 - 0.00462x^2$ to 3 s.f.
At a speed of 110 km/h, the distance travelled is estimated to be
$y = 189.7475 - 0.004619785(110)^2 = 134$ km (3 s.f.).</td></tr>
</table>

13 **9758/2018/02/09**

Many electronic devices need a fan to keep them cool. In order to maximise the lifetime of such fans, the speed they run at is reduced when conditions allow. Running a fan at a lower speed reduces the power required. The following table gives details, for a particular type of fan, of the power required (P watts) at different fan speeds (R revolutions per minute).

Fan speed (R)	3600	4500	5400	6300	7200	8100	9000	9900
Power (P)	0.22	0.34	0.52	0.78	1.06	1.48	2.04	2.64

(i) Draw a scatter diagram of these data. Use your diagram to explain whether the relationship between P and R is likely to be well modelled by an equation of the form $P = aR + b$, where a and b are constants. [2]

(ii) By calculating the relevant product moment correlation coefficients, determine whether the relationship between P and R is modelled better by $P = aR + b$ or by $P = aR^2 + b$. Explain how you decide which model is better, and state the equation in this case. [5]

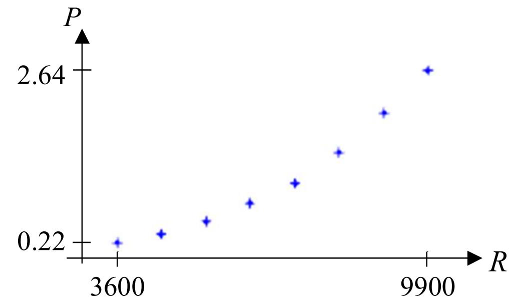
(iii) Use your equation to estimate the speed of the fan when the power is 0.9 watts. Explain whether your estimate is reliable. [2]

(iv) Use your equation to estimate the power used when the speed of the fan is 3300 revolutions per minute. Explain whether your estimate is reliable. [2]

(v) Re-write your equation from part **(ii)** so that it can be used when the speed of the fan, R, is given in revolutions per second. [1]

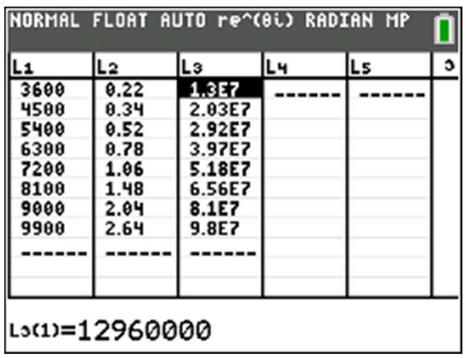
Solution:

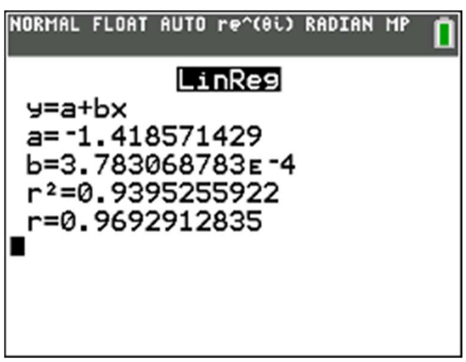
(i)



The scatter plots suggests that the relationship is unlikely to be well-modelled by the linear model $P = aR + b$ because as R increases, P increases at an increasing rate.

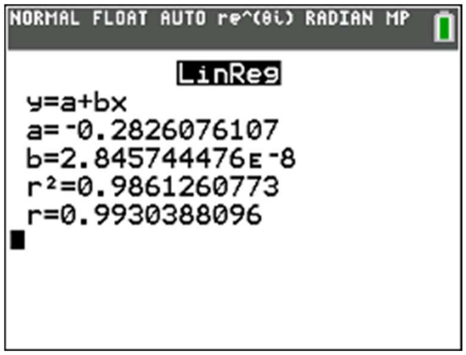
(ii)





For model, $P = aR + b$

$r = 0.96929$

	 For model, $P = aR^2 + b$ $r = 0.99304$ Since the product moment correlation coefficient, r, for model $P = aR^2 + b$ is nearer to 1 compared to model $P = aR + b$, we confirmed the relationship by drawing the scatter diagram. Thus, $P = 0.000000028457R^2 - 0.28261$
(iii)	$0.9 = 2.8457\times10^{-8}R^2 - 0.28261 \quad \Rightarrow \quad R = 6450$ (3sf) $r = 0.993$ is near 1 which suggest the model is a good fit. Also, the value 0.9 is within the given range of values of P, so we did not extrapolate the information and therefore the estimation is reliable.
(iv)	$P = 2.8457\times10^{-8}(3300)^2 - 0.28261 = 0.0273$ (3sf) The value 3300 lies outside the given range of values of R, so we extrapolate the information and therefore the estimation is not reliable.
(v)	If R is given in revolution per second, then the revolutions per minute is $60R$. $P = 2.8457\times10^{-8}\left(60R\right)^2 - 0.28261$ $P = 0.00010245R^2 - 0.28261$

14 9740/2010/02/Q10

A car is placed in a wind tunnel and the drag force F for different wind speeds v, in appropriate units is recorded. The results are shown in the table.

v	0	4	8	12	16	20	24	28	32	36
F	0	2.5	5.1	8.8	11.2	13.6	17.6	22.0	27.8	33.9

(i) Draw the scatter diagram for these values, labelling the axes clearly. [2]

It is thought that the drag force F can be modelled by one of the formulae

$$F = a + bv \quad \text{or} \quad F = c + dv^2$$

where a, b, c and d are constants.

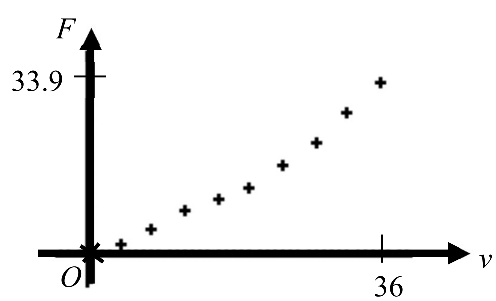
(ii) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

(a) v and F,

(b) v^2 and F. [2]

(iii) Use your answers to parts **(i)** and **(ii)** to explain which of $F = a + bv$ or $F = c + dv^2$ is the better model. [1]

(iv) It is required to estimate the value of v for which $F = 26.0$. Find the equation of a suitable regression line, and use it to find the required estimate. Explain why neither the regression line of v on F nor the regression line of v^2 on F should be used. [4]

Solution:	
(i)	
(ii)(a)	$r = 0.9860$ (4dp)
(ii)(b)	$r = 0.9907$ (4dp)
(iii)	$F = c + dv^2$ is the better model since $\lvert r \rvert$ is closer to 1 as compared to $F = a + bv$. The scatter diagram also shows a curvilinear trend (as v increases, F increases at an increasing rate), indicating that a linear model might not be appropriate.
(iv)	A suitable regression line is F on v^2. Using GC, $F = 3.19565 + 0.024242v^2$ When $F = 26.0, v = 30.7$ units (3 s.f.) Neither the regression line of v on F nor the regression line of v^2 on F should be used, as v is the independent variable. Therefore, the regression line of F on v^2 should be used regardless of estimating F or v. [Note: Cambridge exam report states answer as 30.7 (using F on v^2) or 31.0 (using F on v)]