

Chapter S8

CORRELATION & REGRESSION

At the end of this chapter, students should be able to

- use a scatter diagram to determine if there is a plausible linear relationship between two variables
- calculate and interpret the product moment correlation coefficient as a measure of the fit of a linear model to the scatter diagram
- find the equation of the least-squares regression line
- use the appropriate regression line to make prediction or estimate a value in practical situations (interpolation and extrapolation), and explain how well the situation is modelled by the linear regression model
- use a square, reciprocal or logarithmic transformation to achieve linearity

8.1 Terminology

8.1.1 Dependent and Independent Variables

In statistics, **bivariate data** is data that has two variables. The following table shows 4 sets of bivariate data x and y . Is there any relationship between each pair of variables?

x	y
Student's Mathematics mark	Student's Physics mark
Age of a tree	Trunk circumference
Length of leg of a student	Distance jumped by a student
Monthly advertising expenditure	Monthly sales

In a bivariate pair, an **independent variable** is a variable whose variation does not depend on that of another. In an experiment, the independent variable would be the one which you have "control" over, thus allowing you to vary its value to determine the value of the corresponding **dependent variable**.

It may be possible to identify *from context* the independent variable and the dependent variable. For example, if we are investigating how the length of a pendulum will affect its period of swing, then the length of the pendulum will be the **independent** variable and its period will be the **dependent** variable.

The following table shows another set of bivariate data.

x	10	20	30	40	50	60	70	80
y	20	21	23	24	23	25	28	29

The values of x varies at a *fixed increment*. This is another indication that x is likely to be the **independent** variable (controlled) and y the **dependent** variable. If context is provided, the fact that x increases at fixed intervals would give further evidence that it is the independent variable.

Example 1

- (a) A botanist wishes to investigate the effects of artificial light on the growth of a certain type of plant. A random sample of 7 seeds of the plant species is planted in 7 different pots and the amount of artificial light per day given to each pot is varied. The following results are recorded after a period of 14 days:

Amount of artificial light per day (l hours)	4	8	10	13	14	17	18
Height of plant (h cm)	6	8	10	11	14	14	15

State the independent variable and the dependent variable (if any), justifying your choice.

Solution:

The variable l is the independent variable and the variable h is the dependent variable since the investigation is on how the height of the plant, h cm depended on the amount of artificial light per day, l cm it received.

- (b) Eight newly-born babies were randomly selected. Their head circumference, x cm, and body length, y cm were measured by the pediatrician and tabulated.

x	31	32	33.5	34	35.5	36	36.5	37.5
y	45	49	49	47	50	53	51	51

State the independent variable and the dependent variable (if any), justifying your choice.

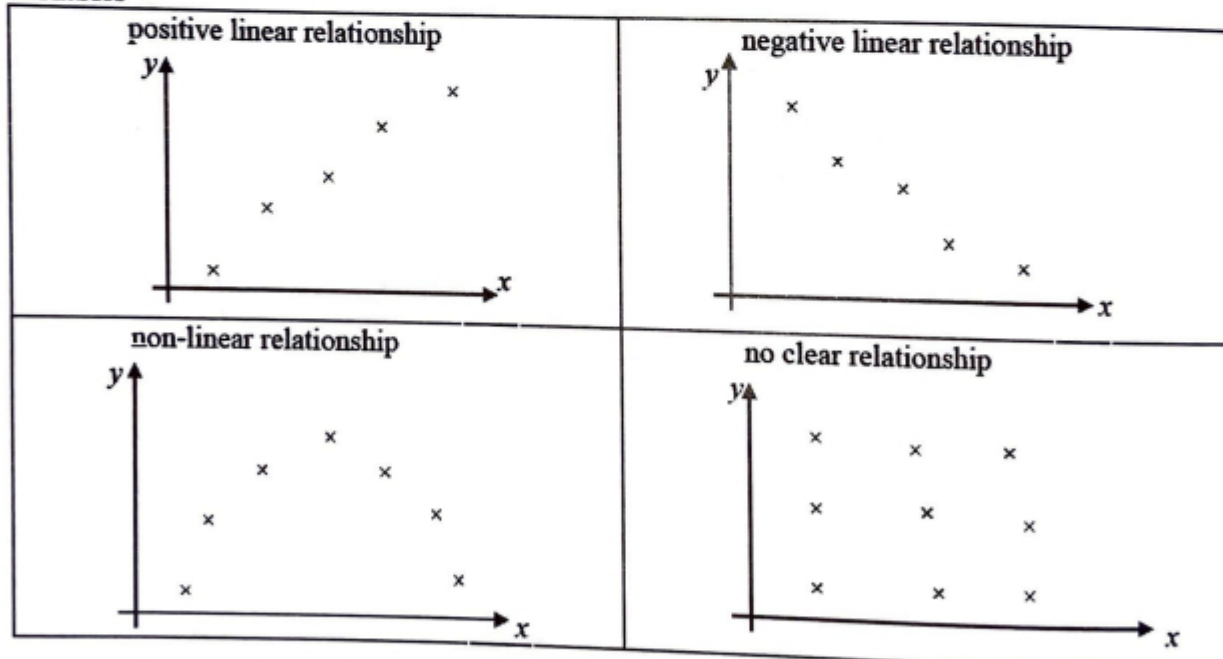
Solution:

There is no independent and dependent variable as it is not clear from context if head circumference is dependent on body length or vice versa.

8.1.2 Scatter Diagram

A **scatter diagram**, in which pairs of data (x_i, y_i) are plotted, can be used to represent bivariate data graphically. A scatter diagram shows the general pattern and relationship between two variables.

Figure 1: Scatter diagrams showing linear, non-linear, and no clear relationships between 2 variables

**Remarks:**

- A scatter diagram gives a pictorial representation of how two variables x and y may be related.
- The independent variable if applicable is plotted on the horizontal axis.
- A scatter diagram is often used to determine if there is a **linear** relationship between two variables.





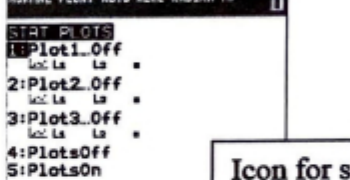









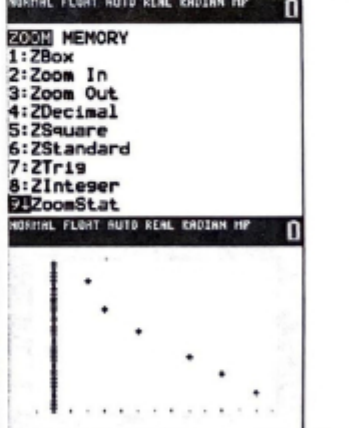

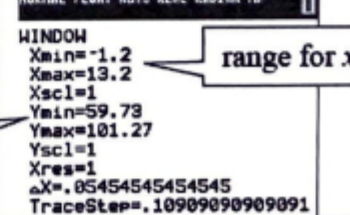
Example 2

Seven students from a class were selected. The number of days of absence and their corresponding final grades were recorded as follows:

Student	A	B	C	D	E	F	G
No. of days of absence, x	10	12	2	0	8	5	3
Final grade, y %	70	65	96	94	75	82	88

Sketch a scatter diagram for the data.

Solution:

<p>Step 1: Press  and select 1:Edit. Key in the values of number of days of absence, x, in list L_1 and final grade, y %, in L_2.</p>		
<p>Step 2: Press   for [STAT PLOTS]. Select 1:Plot1</p>		<p>Icon for scatter plot</p>
<p>Step 3: Notice that Plot1 is highlighted and we can turn it on by positioning the cursor over ON and press . Scroll down  and select the icon for scatter plot at Type: and press .</p> <p>Data for x-axis</p> <p>Data for y-axis</p>		
<p>Step 4: Press  and select 9:ZoomStat. (note: to read the coordinates of each point in the scatter diagram, press   and use the cursor control keys   to move from point to point.)</p>		
<p>Step 5: Press  to see the range of x and y used by the calculator.</p> <p>range for x-axis</p> <p>range for y-axis</p>		

(iv) When $y = 80$, $x = 1.2356(80) + 3.4317 = 102.2797$

$= 102$ (nearest whole number)

Thus, the estimated time taken for a delegate to travel 80 km for the journey is 102 minutes correct to one decimal place.

Remark:

In Examples 6 and 8, the same data set were used and the regression line of y on x and the regression line of x on y are drawn on the same x - y plane below with x on the horizontal axis and y on the vertical axis. Observe that the two lines are **different** and they **intersect at (\bar{x}, \bar{y})** .

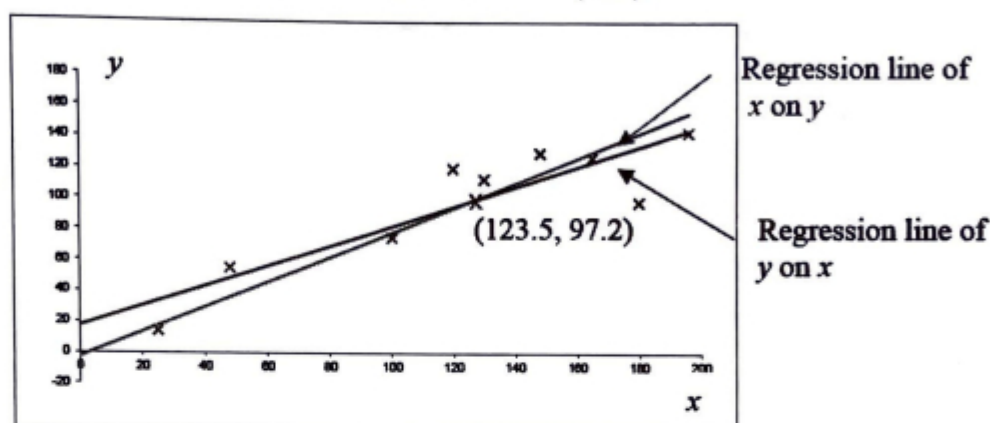


Figure 4: Regression line for y on x and regression line for x on y

<p>Press STATPLOT F1 Y= and key in $\frac{x-3.43}{1.24}$ at $Y_2 =$</p>	<p>NORMAL FLOAT AUTO REAL RADIAN HP</p> <p>Plot1 Plot2 Plot3</p> <p>Y1=0.647X+17.3</p> <p>Y2=$\frac{x-3.43}{1.24}$</p> <p>Y3=</p> <p>Y4=</p> <p>Y5=</p> <p>Y6=</p> <p>Y7=</p> <p>Y8=</p>
<p>Press FORMAT F3 ZOOM and select 9:ZoomStat.</p> <p>The two regression lines can be plotted on the same diagram as shown.</p>	<p>NORMAL FLOAT AUTO REAL RADIAN HP</p>

To find the point of intersection of these two lines:

<p>Press 2ND CALC F4 TRACE for [CALC] and select 5: intersection</p>	<p>NORMAL FLOAT AUTO REAL RADIAN HP</p> <p>CALC INTERSECT</p> <p>Y2DEF=3.43/1.24</p> <p>Intersection</p> <p>X=123.51463 Y=97.21425</p>
---	--

8.3.2 Interpolation Vs Extrapolation

Interpolation is an estimation of a value within two known values in a known sequence of values. **Extrapolation** is an estimation of a value based on extending a known sequence of values. When we are estimating the values of y using values of x that are not in the data range (or vice versa), this is known as extrapolation. In general, we should NOT do extrapolation as there is no way to check if the relationship between x and y continues to hold for values that are outside the given data range.

For the A-Level syllabus, the student may be asked to estimate the value of one variable given the other and to comment on the reliability of the estimate.

(For interpolation), we give two reasons to justify that the estimate is *reliable*: r is close to 1 (or -1) which suggests that there is a strong positive (or negative) linear relationship between the two variables, and the given value lies within the data range.

(For extrapolation) we say that the estimate is *unreliable* as the given value lies outside the data range, thus the estimate may not be reliable due to extrapolation.

Example 9 [N2000/2/10 (part FM)-modified]

An experiment with certain swimming animals was carried out in order to investigate how the speed at which they swam depended on the angle through which their feet moved. The angle θ° through which the hind feet moved was measured, together with the swimming speed $v \text{ ms}^{-1}$. The results are given in the table.

θ	87	92	96	97	98	101	110	114	115	115	116	123	133
v	0.35	0.30	0.50	0.40	0.35	0.45	0.60	0.55	0.55	0.65	0.50	0.70	0.75

- State, giving a reason, which of the least-squares regression lines, θ on v or v on θ , should be used to express a possible linear relation between v and θ .
- Calculate the equation of the line chosen in part (i), give the values of the coefficients to a suitable accuracy.
- Calculate the product moment correlation coefficient.
- Estimate
 - the swimming speed of the animal when the angle through which its hind feet moved is 70° ,
 - the angle through which the animal's feet moved when its swimming speed is 0.32 ms^{-1} .
- Comment briefly on the reliability of the estimates in part (iv).

Solution:

- The regression line v on θ should be used since the experiment investigated how the speed v at which the animals swam depended on the angle θ through which their feet moved, i.e. θ is the independent variable.
- The regression line is $v = 0.0095084\theta - 0.51025$ (5 s.f.)
 $v = 0.01\theta - 0.51$ (2 d.p.)
- $r = 0.910$ (3 s.f.)
- When $\theta = 70$, $v = 0.0095084(70) - 0.51025 = 0.15534$ (5 s.f.)
 $= 0.16$ (2 d.p.)

$$\text{When } v = 0.32, \theta = \frac{0.32 + 0.51025}{0.0095084} = 87.318 \text{ (5 s.f.)}$$

$$= 87 \text{ (nearest whole number)}$$

(The regression line in (ii) is still used because θ is the independent variable)

- The estimate in (a) is unreliable as $\theta = 70$ is outside the data range of θ .
 The estimate in (b) is reliable as $v = 0.32$ is within the data range of v , and r is close to 1.

Example 10

A study is done to investigate whether children who spend more time on reading tend to have higher English Literacy scores than children who do not read sufficiently. The table below shows the number of hours spent on reading per week of a sample of 8 children taken when they were 7 years together with the English Literacy scores determined through a test at age 9 years.

Hours spent on reading per week (x)	12	15	8	10	13	9	6	7
English Literacy score (y)	85	95	73	79	70	76	68	70

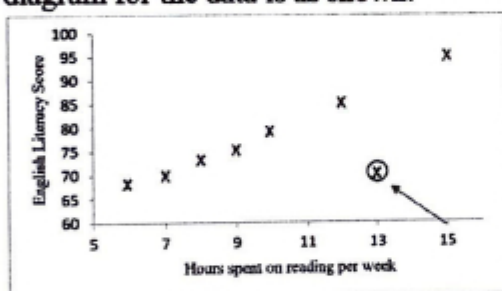
- Draw a scatter diagram for the data.
- It is observed that the English Literacy score for one of the child is suspiciously different. Identify and exclude this point P . Hence find the product moment correlation coefficient and the equation of the estimated regression line of y on x .

With point P removed,

- estimate the English Literacy score of a child who spent 11 hours reading per week.
- estimate the hours spent on reading per week at age 7 years given an English literacy score of 78 at age 9 years to 2 decimal places using the regression line of
 - x on y ,
 - y on x obtained in (ii).
 Comment on the values obtained in (a) and (b).

Solution:

- The scatter diagram for the data is as shown:



- The suspicious point (outlier) is (13, 70). Removing this point and using GC, $r = 0.999$ and the regression line of y on x is
 $y = 3.0322x + 48.978$ (5 s.f.)
 $y = 3.03x + 49.0$ (3 s.f.)
- The estimated English Literacy Score for the child who spent 11 hours reading per week is $3.0322(11) + 48.978 \approx 82.3$ (3 s.f.)
 $= 82$ (whole number)
- By GC, regression line of x on y is $x = 0.32895y - 16.086$
 When $y = 78$, $x = 9.5721$ (5 s.f.)
 $= 10$ (whole number)
 - Substitute $y = 78$ into $y = 3.0322x + 48.978$ from (ii)
 $\Rightarrow x \approx 9.5713$ (5 s.f.)
 $= 10$ (whole number)

The estimated number of hours of reading time corresponding to an English literacy score of 78 in (a) and (b) are 9.5721 and 9.5713 hours respectively. The two values are close.

The reason is since $r = 0.999$ is very close to 1, the regression lines of y on x and x on y almost coincide (i.e. almost identical). Thus it makes little difference which line is used in this case.

Q: Can you identify the independent variable?

Q: Can you suggest a possible reason for the outlier?

Since $r \approx 1$, both regression lines can be used to estimate x given y .

8.4 Transformation of Data Points from Non-Linear to Linear

Two variables may be governed by a relation that is non-linear. For example, the equation $y = a + bx^2$ indicates that y and x are related by a quadratic relationship. However, we can view this equation as y with respect to x^2 . If we let $u = x^2$, then $y = a + bu$ and y and u are linearly related. We see that in making a transformation, $u = x^2$, we have changed two variables that are non-linearly related to two variables that are linearly related.

The table below illustrates some suitable examples of transformation on the variables to obtain a linear relationship.

Equation	Independent Variable	Dependent Variable
$y = a + \frac{b}{x}$	$\frac{1}{x}$	y
$y = a + bx^2$	x^2	y
$y^2 = a + bx$	x	y^2
$y = a + b \ln x$	$\ln x$	y
$\ln y = a + bx$	x	$\ln y$
$y = ae^{bx}$ $\Rightarrow \ln y = \ln a + bx$	x	$\ln y$

Note:

- The above list is not exhaustive. The transformation will usually be given in the question.
- You may be given a scatter diagram and asked to compare two or more proposed models and determine which model is a better fit. Simply state which equation fits the shape of the scatter plot. If there is more than one possibility, compute the product moment correlation coefficient for each model and break the tie by choosing the model with $|r|$ closest to 1,

Example 11

A research worker gave each of eight children a list of words of varying difficulty and asked them to define the meaning of each word. The following table shows the age in years and the number of correctly defined words for each child.

Child	1	2	3	4	5	6	7	8
Age, x (years)	2.5	3.1	4.3	5.0	5.9	7.1	8.1	9.49
Number of correct words, y	9	13	18	25	35	53	81	132

- Using a scatter diagram, comment on the suitability of finding the linear regression line of y on x .
- Find the product moment correlation coefficient between $\ln y$ and $\ln x$ and sketch the scatter diagram of $\ln y$ against $\ln x$. Comment on the suitability of the model $\ln y = a + b \ln x$.
- Find the least-squares regression line of $\ln y$ on $\ln x$ and sketch this line on the scatter diagram in (ii).

Solution:

Solution:

- (i) Given that $x = x_0 e^{-kt}$, taking natural logarithm on both sides,

$$\ln x = \ln(x_0 e^{-kt}) = \ln x_0 + \ln e^{-kt} = \ln x_0 - kt.$$

By GC, $\ln x = -2.8811t + 1.6543$ so that $-k = -2.8811 \Rightarrow k = 2.88$ (3 s.f.)

and $\ln x_0 = 1.6543 \Rightarrow x_0 = e^{1.6543} = 5.23$ (3 s.f.)

See ANNEX for the GC keystrokes

- (ii) When t increases by 1, y is expected to decrease by approximately 2.88.

- (iii) When $x = 1.5$, $y = \ln(1.5)$.

Using the equation of the regression line of y on t , the estimated value of t is 0.433.

Since $r = -0.984$, which is very close to -1 and

$x = 1.5$ is within the data range,

we can use the regression line of y on t to obtain an estimate of t . Thus, the estimate is likely to be reliable.

Recall that since $r \approx -1$, the regression lines of y on t and t on y almost coincide.

Example 13

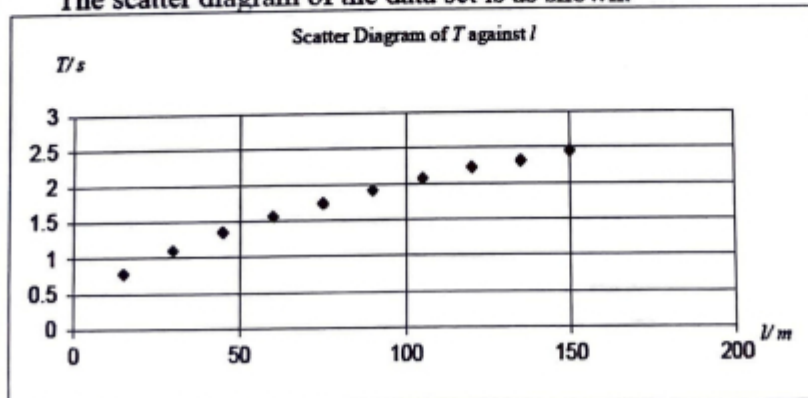
A student wishes to determine the relationship between the length of a pendulum, l , and the corresponding period, T . After conducting the experiment, he obtained the following set of data:

l (cm)	150	135	120	105	90	75	60	45	30	15
T (s)	2.45	2.31	2.22	2.07	1.91	1.74	1.56	1.35	1.10	0.779

- (i) Obtain a scatter diagram of this set of data.
- (ii) The student proposes the following two models:
- $$A: T = a + b \ln(l)$$
- $$B: T^2 = a + bl.$$
- (a) Calculate the product moment correlation coefficient for both models, giving your answers to 4 decimal places.
- (b) Determine which model is appropriate for this set of data.
- (iii) Using the model determined in (ii) part (b), estimate the value of l when $T = 3$ to 1 decimal place. Comment on the suitability of this method.
- (iv) Find a value of l and its corresponding value of T such that the equation of the regression line for the chosen model will remain the same after the addition of this pair of values.

Solution:

- (i) The scatter diagram of the data set is as shown.



<p>(ii)(a) By GC, the product moment correlation coefficient for model A is $r = 0.9871$ and the product moment correlation coefficient for model B is $r = 0.9996$.</p> <p>(b) Since $0.9996 > 0.9871$, model B indicates a stronger positive linear correlation between I and T^2 than the linear correlation between $\ln(I)$ and T of model A. Hence, model B is an appropriate model.</p> <p>(iii) (Using model B) The regression line of T^2 on I is $T^2 = 0.025136 + 0.040060I$. Substituting $T = 3$ into the above equation, we have $3^2 = 0.025136 + 0.040060I$. Thus, $I = 224.0$. Using the regression line may not be appropriate as $T = 3$ is not within the data range even though r is close 1.</p> <p>(iv) Note that $(\bar{I}, \overline{T^2})$ is the <i>only</i> point that we can safely add to the set of data points and not change the regression line. By GC, we find that $\bar{I} = 82.5$ and $\overline{T^2} = 3.3300541$ Thus, $I = 82.5$ and $T = \sqrt{3.3300541} \approx 1.82$.</p>	<p>Note that both equations can apply to this scatter diagram. Thus we compare the value of r as instructed in the question to determine the better model.</p> <p>Q: Are we justified in using the regression line of T^2 on I to estimate I given a value of T?</p> <p>Important: Note that $\overline{T^2} \neq \bar{T}^2$.</p> <p>See ANNEX for the GC keystrokes to find the values of \bar{I} and $\overline{T^2}$.</p>
--	--

Note: For any set of data, adding the point (\bar{x}, \bar{y}) will not alter the equation of the regression line of y on x and the equation of the regression line of x on y .

SUMMARY: For Regression,

a.	We first determine using a scatter diagram that a linear relationship between the variables x and y exist.
b.	For a regression line y on x , x is taken as the independent variable and y as the dependent variable. For a regression line x on y , y is taken as the independent variable and x as the dependent variable.
c.	When it is <u>not obvious</u> which is the <u>independent or dependent variable</u> , then, <ul style="list-style-type: none"> • If you are given x and asked to <i>estimate</i> y, use the regression line y on x. • If you are given y and asked to <i>estimate</i> x, use the regression line x on y
d.	If it is clear from the question that x is the independent variable (e.g. controlled variable), while y is the dependent variable, then there is only one regression line y on x . (Note: you should not find regression line x on y in this case)
e.	Uses of the regression line of y on x : <ul style="list-style-type: none"> • Should only be used to estimate values of y within the range of values of x used in calculating the regression line. [Extrapolation is making a prediction outside the range of values in the sample used to generate the model] [Interpolation is making a prediction within the range of values in the sample] <div data-bbox="513 1077 991 1406" data-label="Figure"> </div> <p>Extrapolation should not be used as we have no way of knowing the relationship between x and y that are not in the data range.</p>
f.	Both regression lines of y on x and x on y will pass through the point (\bar{x}, \bar{y}) .
g.	If the scatter diagram reveals a non-linear relationship between two variables x and y , say of the form, $y = a + bf(x)$, then it is possible to introduce a new variable $u = f(x)$ so that the equation becomes $y = a + bu$ which is linear in u and y .
h.	Given a few non-linear models, we first use the scatter diagram to determine which is the appropriate model. If two or more models are applicable, then the model such that $ r $ is the largest after transformation to a linear equation is in general the best model.

ANNEX

GC Keystrokes for Example 11

To use the GC to transform the data and obtain the regression line of $\ln y$ on $\ln x$.

Press [STAT] and highlight '1:Edit...' and press [ENTER].
Key in the values of x and y into L_1 and L_2 respectively,

In cell L_3 , key in " $\ln(L_1)$ " and press [ENTER].

The command " is to ensure that the GC stores the formula. " can be found on the '+' button.

NORMAL FLOAT AUTO REAL RADIAN HP					
L_1	L_2	L_3	L_4	L_5	L_6
2.5	9				
3.1	13				
4.3	18				
5	25				
5.9	35				
7.1	53				
8.1	81				
9.99	132				
L3="ln(L1)"					

Notice that there is a black icon in the header of L_3 to indicate that a formula is stored. When the cursor is moved to L_3 , the formula " $\ln(L_1)$ " will appear below the screen to indicate that $L_3 = \ln(L_1)$. If " is omitted when keying in " $\ln(L_1)$ " in L_3 , the formula will not stored.

NORMAL FLOAT AUTO REAL RADIAN HP					
L_1	L_2	L_3	L_4	L_5	L_6
2.5	9	.91629			
3.1	13	1.1314			
4.3	18	1.4586			
5	25	1.6094			
5.9	35	1.775			
7.1	53	1.9601			
8.1	81	2.0919			
9.99	132	2.2592			
L3="ln(L1)"					

In cell L_4 , key in " $\ln(L_2)$ " and press [ENTER].

NORMAL FLOAT AUTO REAL RADIAN HP					
L_1	L_2	L_3	L_4	L_5	L_6
2.5	9	.91629			
3.1	13	1.1314			
4.3	18	1.4586			
5	25	1.6094			
5.9	35	1.775			
7.1	53	1.9601			
8.1	81	2.0919			
9.99	132	2.2592			
L4="ln(L2)"					

Press [STAT], highlight 'CALC' and select '4:LinReg(ax+b)' and press [ENTER].

To obtain the product moment correlation coefficient between $\ln y$ and $\ln x$, we key in ' L_3 ' in 'Xlist:' and ' L_4 ' in 'Ylist'. We then scroll down to 'Calculate' and press [ENTER].

NORMAL FLOAT AUTO REAL RADIAN HP					
LinReg(ax+b)					
Xlist:L3					
Ylist:L4					
FreqList:					
Store RegEQ:					
Calculate					

By GC, $r = 0.982$.

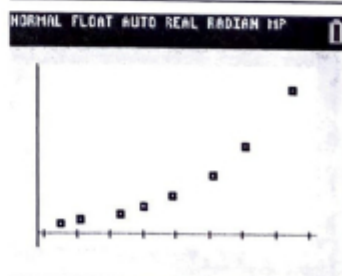
$$\ln y = 1.948 \ln x + 0.247$$

NORMAL FLOAT AUTO REAL RADIAN HP					
LinReg					
Y=Ax+b					
a=1.947690133					
b=.2473300002					
r^2=.9652810252					
r=.9824871629					

Since $r = 0.982$ is very close to 1 and also based on the scatter diagram, we can say that the model $\ln y = a + b \ln x$ is suitable.

Note:

You will need to plot L_4 against L_3 in your GC to see that $\ln x$ and $\ln y$ are linearly related.



GC Keystrokes for Example 12

To use the GC to transform the data and obtain the regression line of $\ln x$ on t .

Press [STAT] and highlight '1:Edit...' and press [ENTER].
Key in the values of t and x into L_1 and L_2 respectively,

In cell L_3 , key in ' $\ln(L_2)$ ' and press [ENTER].

NORMAL FLOAT AUTO REAL RADIAN MP					
L_1	L_2	L_3	θ	L_4	L_5
2	3.22	1.1694			
4	1.63	.4858			
6	.87	-.1165			
8	.41	-.6916			
1	.36	-1.022			

$L_3 = \ln(L_2)$					

Press [STAT], highlight 'CALC' and select '4:LinReg(ax+b)' and press [ENTER].

To obtain the values of x_0 and $-k$, we key in ' L_1 ' in 'Xlist:' and ' L_3 ' in 'Ylist'. We then scroll down to 'Calculate' and press [ENTER].

NORMAL FLOAT AUTO REAL RADIAN MP					
LinReg(ax+b)					
Xlist:L1					
Ylist:L3					
FrcList:					
Store RegEQ:					
Calculate					

Subsequently, we get the results:
From the GC, the regression line of y on t is
 $y = -2.8811t + 1.6543$ or
 $\ln x = -2.8811t + 1.6543$ since $y = \ln x$.
Thus, $k \approx 2.88$ and $x_0 = e^{1.6543} \approx 5.23$.

NORMAL FLOAT AUTO REAL RADIAN MP					
LinReg					
$y = ax + b$					
$a = -2.881121674$					
$b = 1.654308643$					
$r^2 = .9678015913$					
$r = -.9837690742$					

GC Keystrokes for Example 13

To use the GC to calculate \bar{l} and $\overline{T^2}$

To obtain the values of \bar{l} and $\overline{T^2}$, we key in ' L_1 ' in 'Xlist:' and ' L_4 ' in 'Ylist' respectively.

Press [STAT], highlight 'CALC' and select '2:2-Var Stats' and press [ENTER].

NORMAL FLOAT AUTO REAL RADIAN MP					
L_1	L_2	L_3	θ	L_4	L_5
150	2.45	5.0106	6.0025		
135	2.31	4.9852	5.3361		
120	2.22	4.7875	4.9284		
105	2.07	4.654	4.2849		
90	1.91	4.4598	3.6481		
75	1.74	4.3175	3.8276		
60	1.56	4.0943	2.4336		
45	1.35	3.8867	1.8225		
30	1.1	3.4012	1.21		
15	.779	2.7681	.60684		

$L_4 = L_2^2$					

NORMAL FLOAT AUTO REAL RADIAN MP					
2-Var Stats					
Xlist:L1					
Ylist:L4					
FrcList:					
Calculate					

Scrolling down with the navigating buttons:

NORMAL FLOAT AUTO REAL RADIAN MP					
2-Var Stats					
$\bar{x} = 82.5$					
$\Sigma x = 825$					
$\Sigma x^2 = 86625$					
$Sx = 45.41475531$					
$\sigma x = 43.08421985$					
$n = 10$					
$\bar{y} = 3.3300541$					
$\downarrow \Sigma y = 33.300541$					

NORMAL FLOAT AUTO REAL RADIAN MP					
2-Var Stats					
$\uparrow n = 10$					
$\bar{y} = 3.3300541$					
$\Sigma y = 33.300541$					
$\Sigma y^2 = 140.704731$					
$Sy = 1.820016102$					
$\sigma y = 1.726618879$					
$\Sigma xy = 3490.90112$					
$\downarrow \min X = 15$					



NANYANG JUNIOR COLLEGE

DEPARTMENT OF MATHEMATICS

Year Two (2020)

H2 MATHEMATICS Tutorial S8

Correlation and Regression

- 1 A hot metal ball is left to cool down to room temperature. An experiment is conducted to observe the temperature of the metal ball as time varies. The change in successive times is shown in the following table.

Time (t minutes)	1	2	3	4	5	6	7	8	10	12
Temperature (θ °C)	99	86	64	58	43	37	29	25	24	23

Draw a scatter diagram to illustrate the data, labelling the axes clearly. Comment on whether a linear model would be appropriate.

- 2 Sketch a scatter diagram that might be expected when x and y are related approximately as given in each of the cases (A) and (B) below. In each case your diagram should include 5 points, approximately equally spaced with respect to x , and with all x - and y -values positive. The letters a and b represent constants.

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

(B) $y = a + \frac{b}{x}$ where a is positive and b is negative.

- 3 [AJC/2014/II/11]

(a) The linear product correlation coefficient between 2 variables X and Y is denoted by r . A set of 6 bivariate data yields $r = -0.9$ and a second set of 6 different bivariate data also yields $r = -0.9$. Explain, with the aid of a diagram, whether this implies that r is also negative for the combined set of 12 bivariate data.

(b) It is observed in a one-year study that the linear correlation coefficient between the weight gain and the number of sleeping hours per day is close to -1 . Comment briefly upon this statement: "Since the linear correlation coefficient is close to -1 , we can therefore conclude that the weight gain is caused by insufficient amount of sleep per day."

- 4 [MJC/2014/Prelim/II/Q7]

Research is being carried out on how the height of a tree varies with its age. The data collected is given in the following table.

Age of tree (x years)	1	2	3	4	5	6	7	8
Height (y metres)	1.8	2.9	4.0	4.6	5.0	5.3	5.6	5.8

- (i) Draw the scatter diagram for these values, labelling the axes clearly.
- (ii) Calculate the product moment correlation coefficient between x and y , and explain why its value does not necessarily mean that a linear model would be appropriate.
- (iii) Explain why in this context a quadratic model would probably not be appropriate for long-term predictions.

[(ii) $r = 0.952$ (to 3 s.f.)]

5 [PJC/2014/II/10]

- (i) Explain why it is advisable to plot a scatter diagram before interpreting a linear correlation coefficient calculated for a sample drawn from a bivariate distribution.

The table gives the values of six observations of bivariate data, x and y .

x	25	30	35	40	45	50
y	0.067	0.125	1.131	1.000	3.330	7.627

- (ii) Calculate the product moment correlation coefficient between x and y , and explain whether your answer suggests that a linear model is appropriate.
- (iii) Draw a scatter diagram for the data.

One of the values of y appears to be incorrect.

- (iv) Indicate the corresponding point on your diagram by labelling it P , and explain why the scatter diagram for the remaining points may be consistent with a model of the form $\ln y = a + bx$.
- (v) Omitting P , calculate the least squares estimate of a and b for the model $\ln y = a + bx$.
- (vi) Assume that the value of x at P is correct. Estimate the value of y for this value of x .
- (vii) In fact, the variable y represents the percentage of pregnant women of age x years giving birth to a child with Down's Syndrome. It is required to estimate the age of a woman for which $y = 1.131$. Explain why neither the regression line of x on y nor the regression line of x on $\ln y$ should be used.

$$[(ii) r = 0.870, (v) \ln y = -7.7838 + 0.19669x, (vi) y = 0.407]$$

6 Eight pairs of observations on the variables x and y are given in the following table:

x	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0
y	1.84	1.92	2.04	2.10	2.31	2.41	2.56	2.75

- (i) Draw a scatter diagram for the data, explaining your choice of the independent and dependent variables. Based on your observation, what conclusion can you make about the relationship between the variables x and y ?
- (ii) Obtain the product moment correlation coefficient for the data to 4 decimal places.
- (iii) Find the equation of the regression line of y on x giving the coefficient of x and the constant term to 4 decimal places.
- (iv) Use your equation found in (iii) to estimate the value of y to 2 decimal places when $x = 6.5$, giving two reasons why your estimation is valid.
- (v) Find the equation of the regression line of x on y and use it to find the value of y when $x = 6.5$.
- (vi) Compare the two values of y found in (iv) and (v). What do you notice? Why is it so?
- (vii) Plot both the regression lines of y on x and x on y on your scatter plot in part (i).

$$[(ii) 0.9913 (iii) y = 0.1296x + 1.6579 (iv) 2.50 (v) x = 7.5805y - 12.4899, y = 2.51]$$

- 7 Amy travels regularly from home to the tuition centre on Saturdays. She leaves home x minutes after 8 am and takes y minutes to travel to the tuition centre. Ten pairs of data are recorded in the table below.

x	0	5	10	15	20	25	30	35	40	45
y	31	42	33	48	47	53	68	65	71	70

- Draw a scatter diagram to illustrate the data, labelling the axes clearly. Explain your choice of the independent and dependent variables.
- Calculate the equation of the regression line of y on x and the value of the product moment correlation coefficient for the data to 4 decimal places.
- Interpret the coefficient of regression of y on x in the context of the question. Suggest a possible reason for this phenomenon.
- Draw the regression line of y on x on the scatter diagram.
- Amy is supposed to report to the tuition centre at 8.55 am on a particular Saturday. If she leaves home at 8.12 am, explain whether she is likely to report to the tuition centre on time. Comment on the reliability of your answer.
- Use a suitable regression line to estimate the time Amy has to leave her home so that the travelling time to the tuition centre is 45 minutes.

[Note that we still use the regression line of y on x since x is the independent and y is the dependent variable]

[(ii) $y = 0.9455x + 31.5273$, $r = 0.9489$; (vi) 8.14 am]

- 8 One end A of an elastic string was attached to a horizontal bar and a mass, m grams, was attached to the other end B . The mass was suspended freely and allowed to settle vertically below A . The length AB , l mm, was recorded for various masses as follows:

m	100	200	300	400	500	600
l	228	236	256	k	285	301

The equation of the regression line of l on m is $l = 0.15257m + 210.6$.

- A student gave the following calculation to find the unknown k :
 "Since the regression line of l on m is $l = 0.15257m + 210.6$ and k is the value of l when $m = 400$, therefore substituting into the equation gives $k = 271.628 \approx 272$ ". State the mistake in the student's calculation and provide a correct working to show that $k = 278$ correct to the nearest whole number.
- Calculate the product moment correlation coefficient of l and m .
- Give, in context, interpretations for the gradient and vertical intercept of the regression line of l on m .
- State a possible physical limitation in using your regression line of l on m to estimate the length of the string when a mass of 1.2 kilograms is attached at B .

[(ii) 0.991]

9 [2013/HCI/II/7]

The environment agency of a large city is tracking the number of dengue cases reported across several months in year 2013. The data recorded by its researcher is as shown in the table below.

Month, t	1 (Jan)	2 (Feb)	3 (Mar)	4 (Apr)	5 (May)	6 (Jun)	7 (Jul)
Number of dengue cases, n	42	45	55	62	10	80	109

- (i) Draw a scatter diagram for the data. Identify one outlier and indicate it as P on your diagram.

For the rest of the question, exclude P and use only the remaining six months' data.

- (ii) Explain whether a linear model is appropriate.
 (iii) The following models are suggested for the data.

$$(A) n = a - bt^2, \quad (B) n = ae^{bt}, \quad (C) n = a + \frac{b}{t},$$

where a and b are constants, and $b > 0$.

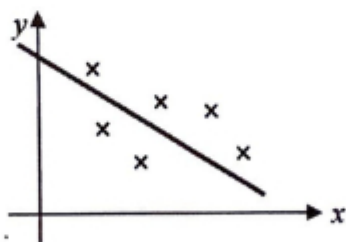
Without calculating the product moment correlation coefficient, state with a reason which model is the most appropriate. Calculate the least squares estimates for a and b using your selected model.

- (iv) It was discovered that there was an error in the transmission of the data. The actual number of dengue cases n' in each month was double that of the data given above. By considering the relationship between n and n' , write down an appropriate regression model of n' in terms of a , b and t .

[(iii) Model B, $a = 34.3$, $b = 0.154$; (iv) $n' = 2n = 2ae^{bt}$]

10(a) [2011 TJC/II/10]

The scatter diagram shows a sample of size 6 of bivariate data, together with the regression line of y on x .



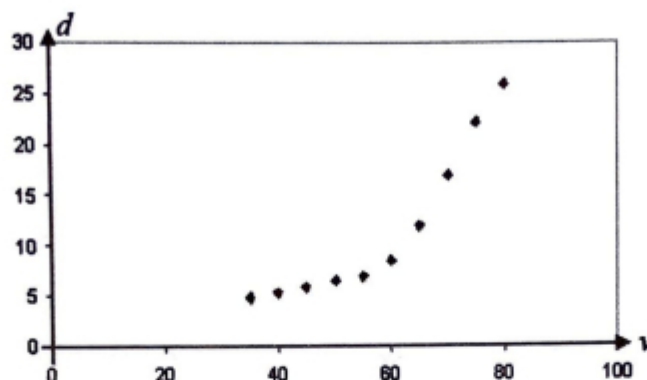
State and giving a reason, whether, for the data shown, the regression line of y on x is the same as the regression line of x on y .

Given that $y = -8x + 86$ is the regression line of y on x for the 6 pairs of observations with $\sum x = 60$.

- (i) Find the value of the mean of y .
 (ii) State, with a reason, the regression line of y on x after a new pair of observation (10, 6) is added.
- (b) A car is travelling along a stretch of road with speed v (km/h) when the brakes are applied. The car comes to rest after travelling a further distance of d m. The values of d for 10 different values of v are given in the table, correct to 2 decimal places.

v	35	40	45	50	55	60	65	70	75	80
d	4.90	5.30	5.93	6.45	7.00	8.50	12.00	17.00	22.10	25.90

It is given that the value of the linear product moment correlation coefficient for this data is 0.919, correct to 3 decimal places. The scatter diagram for the data is shown below:



- (i) Calculate the product moment correlation coefficient between v and \sqrt{d} . What does this indicate about the scatter diagram of the points (v, \sqrt{d}) ?

- (ii) Which regression line is more appropriate: \sqrt{d} on v or d on v ? State the reason and find the equation of the appropriate regression line.

Estimate the distance a car travels after its brake is applied when the car's speed is 100km/h. Comment on the reliability of your answer.

[a(i) 6 (ii) same since $(\bar{x}, \bar{y}) = (10, 6)$ b(i) 0.947 (ii) \sqrt{d} on v , $\sqrt{d} = -0.532 + 0.0656v$, 36.4 m]

11 [9758/2019/II/10]

Abi and Bhani find the fuel consumption for a car driven at different constant speeds. The table shows the fuel consumption, y kilometres per litre, for different constant speeds, x kilometres per hour.

x	40	45	50	55	60
y	22	20	18	17	16

- (i) Abi decides to model the data using the line $y = 35 - \frac{1}{3}x$.
- (a) On a grid paper
- Draw a scatter diagram of the data,
 - Draw the line $y = 35 - \frac{1}{3}x$. [2]
- (b) For a line of best fit $y = f(x)$, the residual for a point (a, b) plotted on the scatter diagram is the vertical distance between $(a, f(a))$ and (a, b) . Mark the residual for each point on your diagram. [1]
- (c) Calculate the sum of the squares of the residuals for Abi's line. [1]
- (d) Explain why, in general, the sum of the squares of the residuals rather than the sum of the residuals is used. [1]

Bhani models the same data using a straight line passing through the points $(40, 22)$ and $(55, 17)$. The sum of the squares of the residuals for Bhani's line is 1.

- (ii) ■State, with a reason, which of the two models, Abi's or Bhani's, gives a better fit. [1]
- (iii) State the coordinates of the point that the least squares regression line must pass through. [1]
- (iv) Use your calculator to find the equation of the least squares regression line of y on x . State the value of the product moment correlation coefficient. [3]
- (v) Use the equation of the regression line to estimate the fuel consumption when the speed is 30 kilometres per hour. Explain whether you would expect this value to be reliable. [2]
- (vi) Cerie performs a similar experiment on a different car. She finds that the sum of the squares of the residuals for her line is 0. What can you deduce about the data points in Cerie's experiment? [1]

Assignment questions

1 [JJC/2014//II/9]

A random sample of nine pairs of values of x and y are given in the table.

x	2.5	2.0	3.0	3.5	5.0	4.0	5.3	7.5	6.0
y	3.20	3.40	3.00	2.86	2.61	2.75	2.57	k	2.55

- (i) The equation of the regression line of y on x is $y = -0.175x + 3.57$. Show that $k = 2.40$.
- (ii) Draw a scatter diagram for this set of data and obtain the product moment correlation coefficient. Comment on the suitability of the linear model.
- (iii) State, with a reason, which of the following would be an appropriate model to represent the above data. 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 negative and d is positive,
 - C. $y = e + \frac{f}{x}$, where e is positive and f is positive.
- (iv) It is required to estimate the value of y for which $x = 8.0$. Find the equation of a suitable regression line, and use it to find the required estimate. Comment on the reliability of your estimation.

2 [YJC/2014//II/6]

A mother monitored the growth of her baby and recorded the height h cm and weight y kg at various stages in the baby development. The results were as follows.

h	50	58	63	68	82	88	96
y	3.93	4.38	5.81	6.68	10.13	13.10	17.45

The mother thought that a model of the form $y = p + qh$, where p and q are constants, might be suitable to describe the relationship between y and h .

- (i) Draw a scatter diagram to illustrate the data.
- (ii) Calculate the value of the product moment correlation coefficient. Explain why its value does not necessarily mean that the best model for the relationship between y and h is $y = p + qh$.
- (iii) Explain which of the following would be the best model to represent the above data.
 - (A) $y = a + be^h$
 - (B) $y = c + dh^3$
- (iv) It is required to estimate the weight of the baby when his height is 75 cm. Find the equation of a suitable regression line, and use it to find the required estimate. Comment on the reliability of the estimation.

Extra Practice Questions

1 ACJC/2014/Prelim/II/Q9

The table below shows the selling price of a particular smartphone T6 over a period of 6 months after its launch.

Month, x	1	2	3	4	5	6
Selling price, y dollars	180	169	149	121	86	52

- (i) Calculate the value of the product moment correlation coefficient, and explain why its value does not necessarily mean that the best model for the relationship between x and y is $y = a + bx$. [2]
- (ii) Draw a scatter diagram to illustrate the data, labelling the axes. [1]
- (iii) Explain how to use the values obtained by calculating the product moment correlation coefficients to decide, for this data, whether $y = c + dx^2$ or $y = a + bx$ is the better model. [1]
- (iv) Estimate the time (in months) for the price of the smartphone to be \$20 using a suitable regression line. Comment on the reliability of your prediction. [2]

[(i) $r = -0.984$, (iii) $r = -0.98446$ $r = -0.99874$ (iv) $x = 6.61$, not reliable]

2 CJC/2014/Prelim/II/Q6

A hot metal ball is left to cool down to room temperature. An experiment is conducted to observe the temperature of the metal ball as time varies. The change in successive times is shown in the following table.

Time (t minutes)	1	2	3	4	5	6	7	8	10	12
Temperature (θ °C)	99	86	64	58	43	37	29	25	24	23

- (i) Draw a scatter diagram to illustrate the data, labelling the axes clearly. Comment on whether a linear model would be appropriate. [3]
- (ii) Referring to the context of the question, state, with a reason, which of the following models is appropriate.
 (A): $\theta = a + b \ln t$, where a and b are constants,
 (B): $\theta = c + \frac{d}{t}$, where c and d are constants. [1]
- (iii) For the appropriate model, calculate the equation of the regression line and the product moment correlation coefficient. [2]
- (iv) Use your equation in part (iii) to predict the temperature at $t = 15$. Comment on the reliability of your prediction. [2]

[(iii) $y = 23.0 + \frac{89.1}{x}$; $r = 0.917$, (iv) $\theta = 28.9^\circ\text{C}$, not reliable]

3 RI/2014/Prelim/II/Q11

A hair stylist tabulated the number of bottles of shampoo sold in his salon in each month. The discounts that was given, x %, and the number of bottles of shampoo sold, y , in each month are as follows.

x	5	10	15	20	25	30	40
y	25	35	55	84	118	151	300

- (i) Draw a scatter diagram for these values, labelling the axes. [1]
- (ii) Calculate the equation of the regression line of y on x . Hence calculate the corresponding estimated value of y when $x = 7.5$. [2]
- (iii) Comment on the suitability of the linear model for this data set. [2]

It is suggested that x and y are related by the equation $y = a + bx^2$.

- (iv) Calculate the product moment correlation coefficient between y and x^2 and comment on its value. [2]
- (v) Estimate the value of a and b . [2]

The cost price of each bottle of shampoo is \$10, and the selling price before discount is \$20. Using the suggested model above, estimate the discount that will maximize the profit for the hair stylist in a month, giving your answer to the nearest whole number. [3]

$$[(ii) y = -46.1 + 7.52x; y = -8.43, (iv) r = 0.995, (v) a = 14.9, b = 0.171; 32\%]$$

4 NJC/2016/Prelim/II/Q8

The table below shows the ages of teak trees, x years, with trunk diameters, y inches. It can be assumed that the diameters of teak trees depend on their ages.

Age (x years)	11	15	28	45	52	57	75	81	88	97
Diameter y (inches)	7.5	11.5	16	19	20.5	21	21.5	21.9	22.2	22.22

- (i) Draw a scatter diagram for these values, labelling the axes. [2]
- (ii) It is desired to predict the diameters of very old trees (of over hundred years old). Explain why, in this context, neither a linear nor a quadratic model is likely to be appropriate. [2]
- (iii) Fit a model of the form $y = a - \frac{b}{x}$ to the data, and calculate the least squares estimates of a and b . Find the product moment correlation coefficient for this model. Use the equation that you have obtained to estimate the diameter of a 40 year-old teak tree, and comment on the reliability of your answer. [4]

$$[(iii) y = 23.9 - \frac{185}{x} \text{ (3sf), } r = -0.994, y = 19.3]$$

5 DHS/2014/Prelim/II/Q10 (modified)

- (i) Sketch a scatter diagram that might be expected for the case when x and y are related approximately by $y = a + b \ln x$, where a is positive and b is negative. Your diagram should include 6 points, approximately equally spaced with respect to x , and with all x - and y -values positive. [1]

About a year ago, Amy decided to go on a healthy eating lifestyle and an exercise regime in order to lose weight. She monitored her weight, y kg, x months after she started and the data is provided below.

No. of months, x	1	3	5	7	9	11
Weight, y	61.4	60.1	59.5	58.9	58.5	58.2

- (ii) Draw the scatter diagram for these values, labelling the axes. [1]
 (iii) Explain which model, $y = a + b \ln x$ or $y = c + dx$, is better for modelling these values. [1]
 For parts (iv) to (vi), use the better model that you have identified in part (iii),

- (iv) Give a contextual interpretation of the value of a or c and calculate the product moment correlation coefficient. [2]
 (v) Use a suitable regression line to estimate the month in which she will reach a weight of 55 kg. You may assume that the model is suitable for short term predictions. [3]
 (vi) Explain, in context, why the model is not suitable for long term predictions. [1]

[(iv) $r = -0.996$, (v) $y = 61.489 - 1.3336 \ln x$; 130th month]

6 DHS/2015/Prelim/II/Q10

In an experiment, different quantities of fertilizer, x ml, were given to seven radish plants of the same height. The heights, y mm, of the plants were measured after ten days. The results are given in the table.

x	10	15	20	25	30	35	40
y	407	412	420	434	450	465	490

- (i) Draw a scatter diagram to illustrate the data. [1]

It is suggested that the height y can be modelled by one of the formulae

$$y = a + bx \quad \text{or} \quad y = c + dx^2 \quad \text{where } a, b, c \text{ and } d \text{ are constants.}$$

- (ii) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between
 (a) x and y , (b) x^2 and y . [2]
 (iii) Use your answers to parts (i) and (ii) to explain which of $y = a + bx$ or $y = c + dx^2$ is the better model. [1]
 (iv) It is desired to estimate the value of x for which $y = 440$. Explain why neither the regression line of x on y nor the regression line of x^2 on y should be used. [1]
 (v) By finding the equation of a suitable regression line, find the required estimate in part (iv). Comment on the reliability of your estimate. [3]

[(ii) (a) $r = 0.9781$ (4 d.p); (b) $y = 399.60 + 0.055326x^2$, $x = 27.0$]

7 VJC/2014/Prelim/II/Q9

A foreign exchange trader studied the daily movement of the rates of Australian Dollar (A\$) and Singapore Dollar (S\$) with respect to one US Dollar (US\$) over 8 trading days. The closing rates of these two currencies for the 8 days are given below.

Day	1	2	3	4	5	6	7	8
S\$ to 1 US\$ (x)	1.2496	1.2499	1.2500	1.2501	1.2502	1.2499	1.2497	1.2503
A\$ to 1 US\$ (y)	1.0572	1.0577	1.0579	1.0581	1.0583	1.0576	1.0572	1.0587

- Draw a scatter diagram for the data. [1]
- Calculate the value of the product moment correlation coefficient between x and y , giving your answer correct to 4 decimal places. Suggest a reason why a linear model may not be the best model for the relationship between x and y . [2]
- The trader used a different model given by the equation $y = ae^{bx} + 1$ to estimate the rate of S\$ to 1 US\$ when the rate of A\$ to 1 US\$ is 1.0582. Find the equation of a suitable regression line, and use it to find the required estimate, giving your answer to 4 decimal places. [4]
- Explain how the trader can conclude that $y = ae^{bx} + 1$ is a better model compared to the linear model $y = c + dx$. [2]
- Comment on the use of the model in part (iii) in predicting the rate of S\$ to 1 US\$ when the rate of A\$ to 1 US\$ is less than 1. [1]

$$[(ii) r = 0.9846, (iii) x = 1.32 + 0.0260 \ln(y - 1); x = 1.2501]$$

8 NJC/2015/Prelim/II/Q12

A medical officer wishes to investigate a patient's walking speed s km/h and his heart-beat rate h beats per minute (bpm). 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

- Sketch a scatter plot of the above data. [1]
 - 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.

- 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): $h = a + be^s$

- Determine which of the two models is a better choice, giving a reason for your answer. [2]
 - Suppose a new data pair (\bar{s}, \bar{h}) 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]
- [(ii) r -value = 0.9812, $s = 3.67$ (iii) (I) r -value = 0.9898 (II) r -value = 0.9380 (v) No]

