Check your Understanding (Correlation and Linear Regression)

MUTIPLE CHOICE.
Choose the one alternative that best completes the statement or answers the question.

1. In regression analysis, the variable that is being predicted is

A the independent variable

B the dependent variable

C usually denoted by x

D none of the above (B)

2. In the regression equation $y = b_o + b_1x$, bois the

A slope of the line

B independent variable

C y intercept

D none of the above (C)

3. In the regression equation $y = b_o + b_1x$, b_1 is

A the slope of the line

B an independent variable

C the y intercept

D none of the above (A)

4. In regression analysis, the variable that is doing the predicting or explaining is

A the independent variable

B usually denoted by y

C the dependent variable

D none of the above (A)

5. The range of the correlation coefficient is

A 0 to +1

B -1 to 0

C -1 to 0

D -1 to +1 (D)

6. If the slope of the regression equation y = bo + b1xis positive, then

A as x increases y decreases

B as x increases so does y

C Either a or b is correct

D none of the above (B)

Scatter Diagram and Product Moment Correlation Coefficient

1. MI H1 Prelim 2017/Q11

An electric heater was switched on in a cold room and the temperature of the room was noted at five-minute intervals.

Time from switching on electric heater, x (min)	0	5	10	15	20	25	30	35	40
Temperature of room, y (^oC)	0.4	1.5	3.4	5.5	7.7	9.7	11.7	13.5	15.4

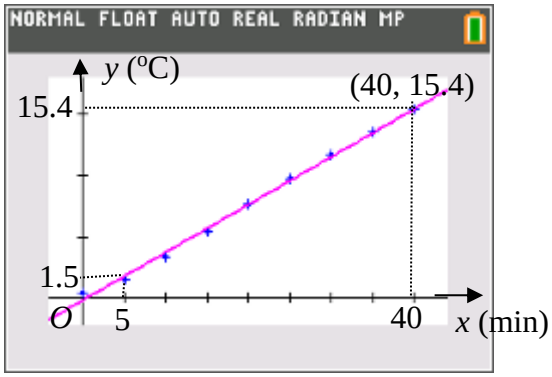
(i) Draw a sketch of the scatter diagram for the data, as shown on your calculator. [2]

(ii) Find the product moment correlation coefficient and comment on its value in the context of the data. [2]

Answers
(ii) 0.999

Solution:

1 (i)



(ii) From graphing calculator,
required coefficient, $r \approx 0.99870 = 0.999$. (3 s.f)

Since $r > 0$ and $|r| = 0.999$ is very close to 1, the value of r suggests that there is a strong positive linear correlation between the temperature of the room and the time from switching on the electric heater.

2. SAJC H1 Prelim 2017/Q7

(a) Eight pairs of values of variables x and y are measured. Draw a sketch of a possible scatter diagram of the data for each of the following cases:

(i) the product moment correlation coefficient is approximately -0.9, [1]
(ii) the product moment correlation coefficient is approximately zero. [1]

(b) A researcher recorded the water temperature T, in °C, and the depth D, in metres, at noon on a certain day at each of the eight locations in a lake. The results are summarized in the table below.

D (m)	10	50	80	120	200	250	340	400
T (°C)	25.0	23.0	22.2	k	16.4	12.4	10.0	4.0

(i) It is known that the regression line of T on D is given by $T = -0.051424D + 25.908$. Show that the value of k is 19.7, correct to 1 decimal place. [1]

(ii) Give a sketch of the scatter diagram for the data. [2]

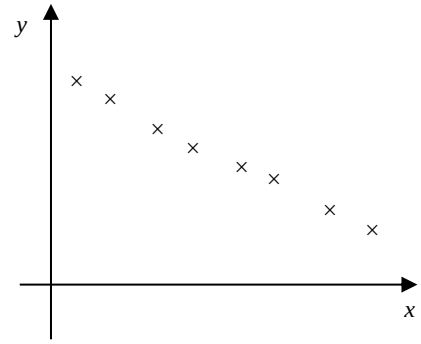
(iii) Calculate the product moment correlation coefficient for the revised data and comment on its value in the context of the question. [2]

(iv) Sketch the regression line T on D on your scatter diagram. [1]

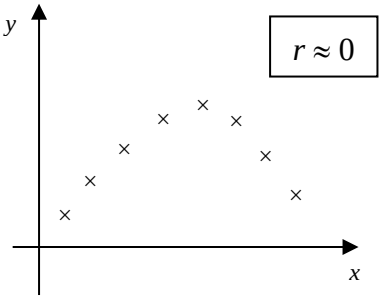
	Answers (iii) $r = -0.992$ (iv) $T = -0.0514D + 25.9$

Solution

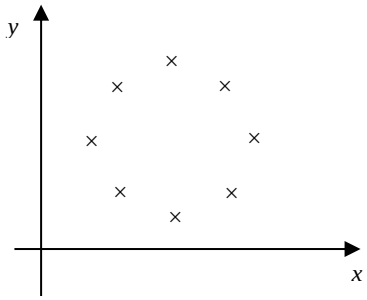
2

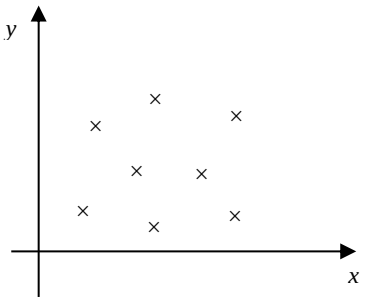


$r \approx -0.9$



$r \approx 0$



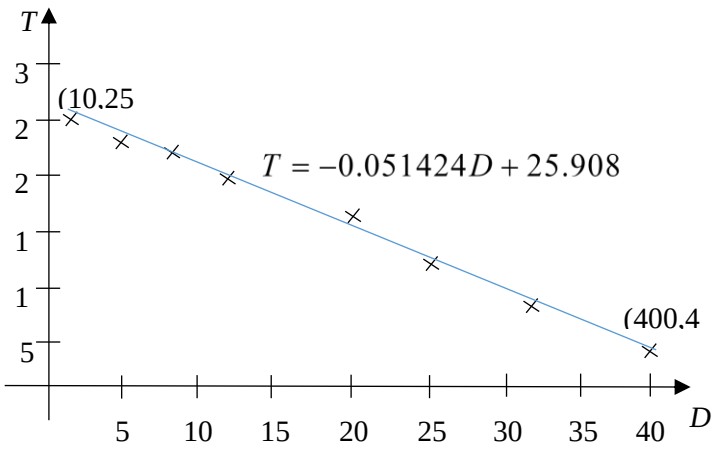


Using $\overline{T} = -0.051424\overline{D} + 25.908$,

substitute $\overline{D} = \dfrac{1450}{8} = 181.25$, $\overline{T} = \dfrac{k+113}{8}$

$\Rightarrow \dfrac{k+113}{8} = -0.051424(181.25) + 25.908$

$\Rightarrow k = 19.6992 = 19.7$ (shown)



$r = -0.992$. It indicates a strong negative linear correlation between the depth of water and the temperature of water. As the depth of water increases, the temperature of water decreases.

3. TJC H1 Prelim 2017/Q8

The number of hours, x, spent daily on revision for mathematics and the marks, y, obtained for the mathematics year-end examination are recorded for 10 randomly selected students. The results are given in the following table.

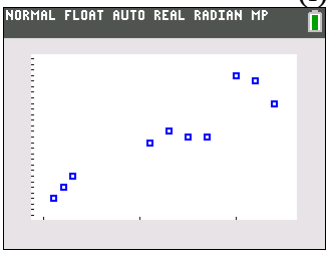
x	1.3	2.1	1.1	2.3	2.7	1.2	3.2	3.4	3.0	2.5
y	68	74	64	76	75	66	85	81	86	75

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

(ii) Find the product moment correlation coefficient and comment on its value in the context of the data. [2]

Answers

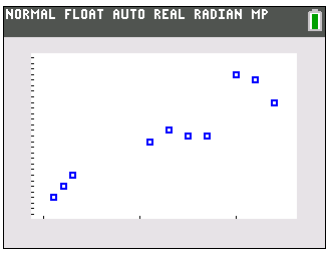
(i)



(ii) $r = 0.938$

$y = 68.5$

Solution

3	(i)  (ii) $r = 0.938$ This indicates a strong positive linear correlation between revision hours and the marks obtained, i.e. as the hours for revision increases, the marks obtained increases linearly.

4. TPJC H1 Prelim 2017/Q9

The year x, and the mean maximum air temperature y, in degrees Celsius, of Singapore, are given in the following table.

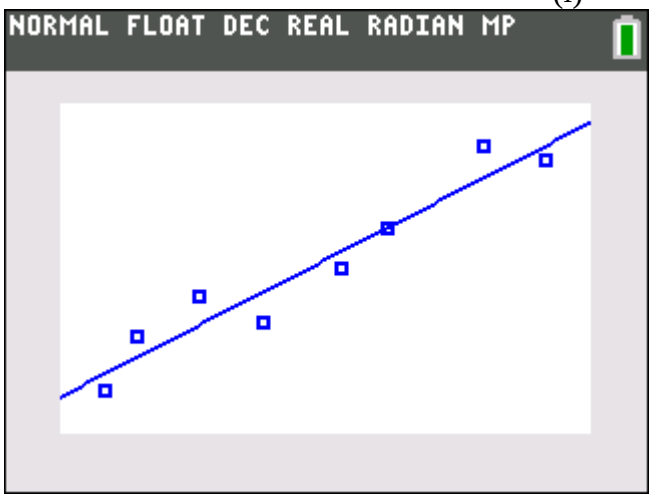
x	1974	1976	1980	1984	1989	1992	1998	2002
y	30.3	30.7	31.0	30.8	31.2	31.5	32.1	32.0

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

(ii) Find the product moment correlation coefficient and comment on its value in the context of the data. [2]

Answers

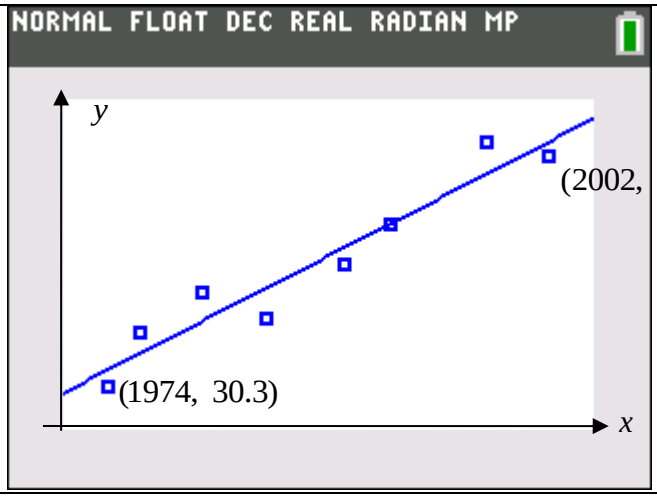
(i)



(ii) $r = 0.960$

Since r is close to 1, there is a strong positive linear relationship between the year and the mean maximum air temperature. In particular, as the year increases, the mean maximum air temperature increases.

Solution

4 (i)	
4 (ii)	Using GC, $r = 0.960$ Since r is close to 1, there is a strong positive linear relationship between the year and the mean maximum air temperature. In particular, as the year increases, the mean maximum air temperature increases.

5. TPJC H1 Prelim 2017/Q9

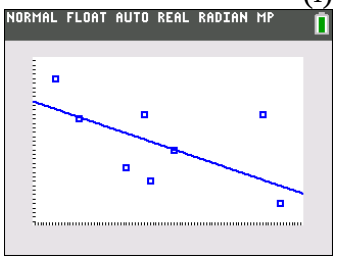
The Mathematics score, x, and the English score, y, of 8 Primary Four students during a year end examination are given in the following table.

Student	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	37	41	49	52	53	57	72	75
y	73	64	53	65	50	57	65	45

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

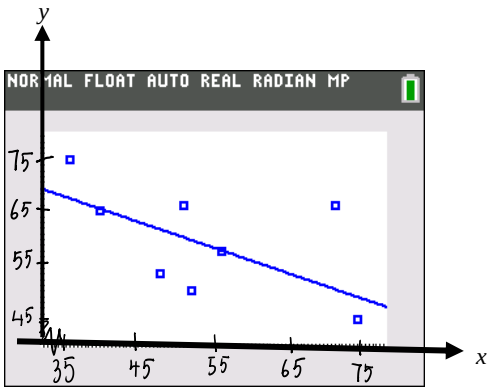
(ii) Find the product moment correlation coefficient. [1]

(iii) The least squares regression line of y on x is used to calculate an estimate of the English score for a student who scored 48 for Mathematics. State, with reasons, whether the estimate will be a reliable one. [2]

	Answers (i)  (ii) $r = -0.53382$ (iii) unreliable

Solution

5)(i)



(ii) $r = -0.53382$

(iii) The estimate is unreliable as from the scatter diagram, the points do not seem to lie close to straight line and r is not close to -1.

Use the appropriate least squares regression line to predict or estimate unknown values

1. MI H1 Prelim 2017/Q11

An electric heater was switched on in a cold room and the temperature of the room was noted at five-minute intervals.

Time from switching on electric heater, x (min)	0	5	10	15	20	25	30	35	40
Temperature of room, y (^oC)	0.4	1.5	3.4	5.5	7.7	9.7	11.7	13.5	15.4

Given that the product moment correlation coefficient = 0.99870.

(i) Find the equation of the regression line of y on x in the form $y = mx + c$, giving the values of m and c correct to 5 decimal places. Draw the line on the scatter diagram in part **(i)** and give an interpretation of m in the context of the question. [3]

(ii) Predict the temperature 2 hours from switching on the electric heater. Give a reason why should this prediction be treated with caution in the context of the question. [2]

(iii) It was later found that the temperature was in fact k ^oC after the electric heater was switched on for 30 minutes, and the equation of the correct regression line of y on x should be $y = 0.4x$. Find the value of k. [2]

Answers
(iii) $y = 0.38933x - 0.14222$
(iv) 46.6 ^oC
(v) $k = 14.9$.

Solution:

1	(i) From graphing calculator, required equation is $y = 0.38933x - 0.14222$. (5 d.p) The temperature of the room is expected to increase by 0.3893^OC for every one minute increase in time from switching on the electric heater. (ii) When $x = 120$, $y = 0.38933(120) - 0.14222 \approx 46.577 = 46.6$. (3 s.f.)

	Required temperature is 46.6 ^{o}C. Possible reasons why prediction should be treated with caution: 1. Extrapolation This prediction of 46.6^{O}C is rather high and should be treated with caution since x = 60 is far outside the range of values of x in the data, i.e. y = 46.6 is an extrapolation. 2. Relationship not linear outside data range. The relationship between the temperature of the room and the time from switching on the electric heater may not be linear any more beyond 40 minutes. 3. Actual temperature too high and can cause a fire. The actual temperature 120 minutes after switching on the electric heater may be high enough to cause a fire which can be a disaster. (iii) For the new data, $\sum x = 180,\ \sum y = 57.1 + k$. $y = 0.4x \Rightarrow \bar{y} = 0.4\bar{x} \Rightarrow \sum y = 0.4\sum x$ $\Rightarrow 57.1 + k = 0.4(180) \Rightarrow k = 14.9.$

2. SAJC H1 Prelim 2017/Q7

	A researcher recorded the water temperature T, in ^{o}C, and the depth D, in metres, at noon on a certain day at each of the eight locations in a lake. The results are summarized in the table below. | D (m)	10	50	80	120	200	250	340	400
---	---	---	---	---	---	---	---	---
T (^{o}C)	25.0	23.0	22.2	19.7	16.4	12.4	10.0	4.0	 **Given that** the Regression line T on D: $T = -0.051424D + 25.908$ (i) Hence, estimate the water temperature when the depth of the water is 550 metres. Comment on the reliability of this estimate. [1] (ii) Given that 1 kilometre = 1000 metre, rewrite your equation from part **(i)** so that it can used to estimate the temperature of the water when the height is given in kilometres. State the value of the regression coefficient. [2]
	Answers (i) $T = -2.38$ (ii) $T = -51.4D + 25.9$

Solution

2	When D = 550,

	$T = -0.051424(550) + 25.908$ $T = -2.38$ Although $r = -0.992$ (3 s.f.) is close to -1, but the estimate may not be reliable as D = 550 falls outside the data range [10, 400]. Hence, the estimate is unreliable. Regression line T on D where height is given in kilometres: $T = -51.424D + 25.908$ $T = -51.4D + 25.9$ (3 s.f.) The value of the regression coefficient is -51.4 .

3. TJC H1 Prelim 2017/Q8

The number of hours, *x*, spent daily on revision for mathematics and the marks, *y*, obtained for the mathematics year-end examination are recorded for 10 randomly selected students. The results are given in the following table.

x	1.3	2.1	1.1	2.3	2.7	1.2	3.2	3.4	3.0	2.5
y	68	74	64	76	75	66	85	81	86	75

(i) Find the equation of the regression line of *y* on *x*, in the form $y = mx + c$, giving the values of *m* and *c* correct to 4 significant figures. Sketch this line on your scatter diagram. [2]

(ii) Use the equation of your regression line to estimate the marks obtained by a student who spends 1.5 hours a day on revision for mathematics. Comment on the reliability of your estimate. [3]

Answers
(i) $y = 8.3802x + 55.893$ (ii) $y = 68.5$

Solution

3	(i) $y = 8.3802x + 55.893$ where $m = 8.380$ (4 s.f.) and $c = 55.89$ (4 s.f.) (ii) When $x = 1.5$, $y = 68.5$ (3 s.f.) The estimate is reliable since *r*-value is close to 1 and $x = 1.5$ is within the data range

4. TPJC H1 Prelim 2017/Q9

The year x, and the mean maximum air temperature y, in degrees Celsius, of Singapore, are given in the following table.

x	1974	1976	1980	1984	1989	1992	1998	2002
y	30.3	30.7	31.0	30.8	31.2	31.5	32.1	32.0

(i) Find the equation of the regression line of y on x, giving your answer in the form $y = mx + c$, where m and c are constants. Sketch this line on your scatter diagram. [2]

(ii) Use the equation of your regression line to estimate the mean maximum air temperature in 2010. Comment on the reliability of your estimate. [2]

(iii) The mean maximum air temperature in 2015 is 31.9 degrees Celsius. Calculate a new estimate of the mean maximum air temperature in 2010. [2]

Answers

(i) $y = 0.0598x - 87.5$

(ii)32.6

Since $x = 2010$ lies outside the given data range, the estimate is not reliable. Extrapolation is not good practice.

(iii) $y = 0.042567x - 53.430$

New estimate is 32.1 degrees Celsius

Solution

4 (i)	Regression line of y on x: $y = 0.059761x - 87.538$ $y = 0.0598x - 87.5$
4 (ii)	When $x = 2010$, $y = 0.059761(2010) - 87.538$ $= 32.6$ Mean maximum air temperature in 2010 is 32.6 degrees Celsius. Since $x = 2010$ lies outside the given data range, the estimate is not reliable. Extrapolation is not good practice.
4 (iii)	New regression line of y on x: $y = 0.042567x - 53.430$ When $x = 2010$, $y = 32.1$ New estimate is 32.1 degrees Celsius

5. TPJC H1 Prelim 2017/Q9

The Mathematics score, x, and the English score, y, of 8 Primary Four students during a year end examination are given in the following table.

Student	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	37	41	49	52	53	57	72	75
y	73	64	53	65	50	57	65	45

(i) Find student *G*'s English score if the correct equation of the regression line of y on x is $y = 88.722 - 0.57976x$. [3]

(ii) Based on the revised data with the correct English score for student *G* as obtained in **(iv)**, calculate an estimate of a student's Mathematics score if his English score is 75. Comment on the reliability of the estimate. [3]

Answers

(i) $p = 50$ (nearest integer) (ii) $x = 34$

Solution

5)

(i) Let the score be p.

$\bar{x} = \frac{436}{8}$, $\bar{y} = \frac{407 + p}{8}$

$\frac{407 + p}{8} = 88.722 - 0.57976\left(\frac{436}{8}\right)$

$\therefore\ p = 50$ (nearest integer)

(ii) From GC, $x = 121.07 - 1.1653y$

When $y = 75$,

$x = 34$ (nearest integer)

The estimate is not reliable as $y = 75$ is outside the given data range $45 \le y \le 73$.