A Level H2 Math

Correlation and Linear Regression Test 4

Q1

In a medical study, researchers investigated the effect of varying amounts of calcium intake on the bone density of Singaporean women of age 50 years. A random sample of eighty 50-year-old Singaporean women was taken.

(i) Explain, in the context of this question, the meaning of the phrase 'random sample'. [1]

The daily calcium intake (x mg) of the women was varied and the average percentage increase in bone density (y%) was measured. The data is as shown in the table below.

x (in mg)	700	800	900	1000	1050	1100
y (%)	0.13	0.78	1.38	1.88	2.07	2.10

(ii) Calculate the product moment correlation coefficient and suggest why its value does not necessarily mean that the best model for the relationship between x and y is $y = a + bx$. [2]

(iii) Draw a scatter diagram representing the data above. [2]

The researchers suggest that the change in bone density can instead be modelled by the equation $\ln(P - y) = a + bx$.

The product moment correlation coefficient between x and $\ln(P - y)$ is denoted by r. The following table gives values of r for some possible values of P.

P	3	5	10
r		–0.993803	–0.991142

(iv) Calculate the value of r for $P = 3$, giving your answer correct to 6 decimal places. Use the table and your answer to suggest with reason, which of 3, 5 or 10 is the most appropriate value of P. [2]

The Healthy Society wants to recommend a daily calcium intake that would ensure an average of 1.8% increase in bone density.

(v) Using the value of P found in part **(iv)**, calculate the values of a and b and use your answer to estimate the daily calcium intake that the Health Society should recommend. Comment on the reliability of the estimate obtained. [4]

(vi) Give an interpretation, in the context of the question, of the meaning of the value of P. [1]

Q2

The following table shows the mass (m) of a foetus, in grams, taken at various weeks (t).

t	12	16	20	24	28	32	36
m	14	100	300	600	1005	1702	2622

(i) Draw a scatter diagram to illustrate the data, labelling the axes clearly. [1]

(ii) Calculate the product moment correlation coefficient between t and m, giving your answer correct to 5 decimal places. Explain why this value does not necessarily mean that the linear model is the best model for the relationship between t and m. [2]

It is proposed that the mass of the foetus at week t can be modelled by

$$m = at^b,$$

where a and b are positive constants.

(iii) By using logarithm to transform $m = at^b$ into a linear equation, calculate the value of the product moment correlation coefficient and give two reasons why this model may be a better model. [4]

(iv) Calculate the values of a and b. [2]

(v) Using the equation of a suitable regression line, estimate the mass of the foetus at 26 weeks, giving your answer to the nearest grams. Comment on the reliability of the estimate. [2]

Q3

(i) 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 6 points, approximately equally spaced with respect to x, and with all x- and y-values positive. The letters a, b, c and d 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. [2]

The following table shows the Gross Domestic Product (GDP) per capita, \$$x$, and infant mortality rate, y, for a sample of 9 countries.

x (\$)	1375	2502	10569	2966	11539	2036	4260	1433	7427
y	115	69	18	65	17	83	44	112	27

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

(iii) Calculate the product moment correlation coefficient, and explain why its value does not necessarily mean that a linear model is the best model for the relationship between x and y. [2]

(iv) State which of the two cases in part **(i)** is more appropriate for modelling the relationship between x and y. Calculate the product moment correlation coefficient and the equation of the appropriate regression line for this case. [3]

(v) Use the regression line in part **(iv)** to find an estimate of the infant mortality rate for a country with GDP per capita of \$723. Comment on the reliability of your estimate. [3]

Answers

Correlation and Linear Regression Test 4

Q1

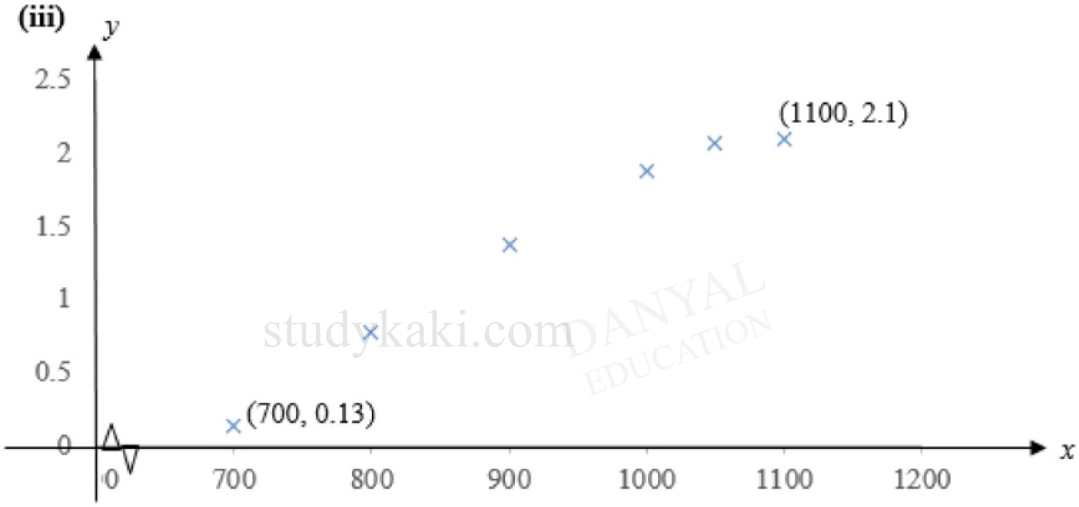
(i) The phrase 'random sample' means that every 50-year-old Singaporean woman has an **equal probability of being included in the sample**.

(ii) $r = 0.988$ (to 3 s.f.)

Although the r-value = 0.988 is close to 1, the value is not 1 so there may be another model with $|r|$ closer to 1.

Hence a linear model may not be the best model for the relationship between x and y.

(iii)



(iv) Using the GC, when $P = 3$, $r = -0.995337$ (to 6 d.p.)

When $P = 3$, $|r|$ is closest to 1 and thus, $P = 3$ is the most appropriate value.

(v) When $P = 3$, using the GC, $a = 3.2446 = 3.24$ (to 3 s.f.)

$b = -0.0030988 = -0.00310$ (to 3 s.f.)

When $y = 1.8$, and $P = 3$,

$$\ln(3-1.8) = 3.2446 - 0.0030988x$$

$$x = 988$$

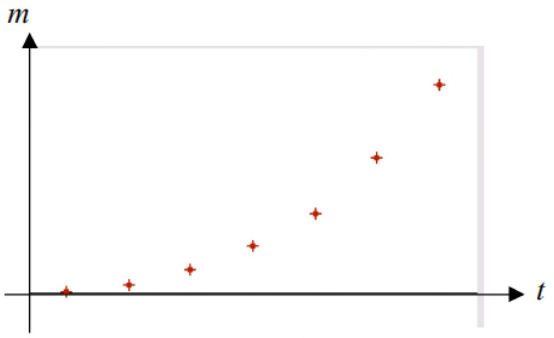
Thus, the recommended daily calcium intake is 988 mg.

Since the r value is -0.995 is close to -1, there is a strong negative linear correlation between $\ln(P-y)$ and x. Also since the value of $y = 1.8$ is within the data range, thus, the estimate obtained is reliable.

(vi) The value of P is the maximum percentage increase in bone density achievable as the daily calcium intake increases.

Q2

(i)



(ii)

The product moment correlation coefficient between t and m is $r = 0.94597$ (5 d.p.).

A value of 0.94597 for r suggests that there is a strong positive linear correlation between t and m. However, the points on the scatter diagram **show a curvilinear relationship**. Therefore this value of r does not necessarily mean that the linear model is best model for the relationship between t and m.

(iii)

$$m = at^b$$

$$\ln m = \ln\left(at^b\right)$$

$$\ln m = b\ln t + \ln a$$

The product moment correlation coefficient between $\ln t$ and $\ln m$ is $r = 0.98967 = 0.990$ (3 sig fig)

Reason 1: From the scatter diagram, as t increases, the **weight of the foetus increases at an increasing rate**.

Reason 2: The value of r between $\ln t$ and $\ln m$ is 0.98967, which is closer to 1 as compared to that between t and m, hence indicating a **stronger positive linear correlation** between $\ln t$ and $\ln m$.

Hence $m = at^b$ is a better model.

(iv)

From GC,

$$\ln m = -8.3764 + 4.5938\ \ln t \quad \text{(5 sig fig)}$$

$$\ln a = -8.3764 \quad \text{and} \quad b = 4.59$$

$$a = 2.30\times10^{-4}$$

(v)

When $t = 26$, $\ln m = -8.3764 + 4.5938 \ln 26$

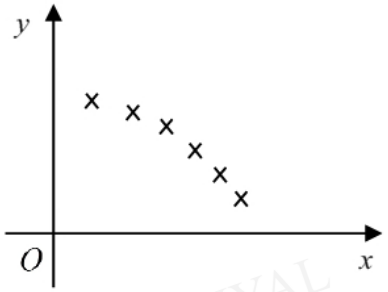
$$m = 728 \text{ (nearest grams)}$$

Since the value of 26 is within the range of values of t and the value of r is close to 1, this estimate is reliable.

Q3

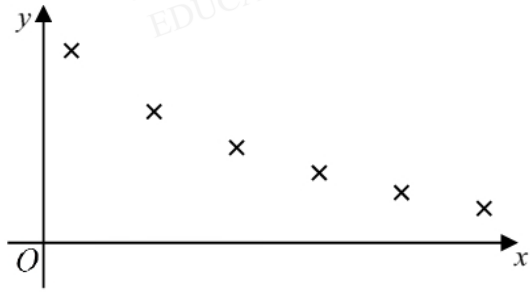
(i)

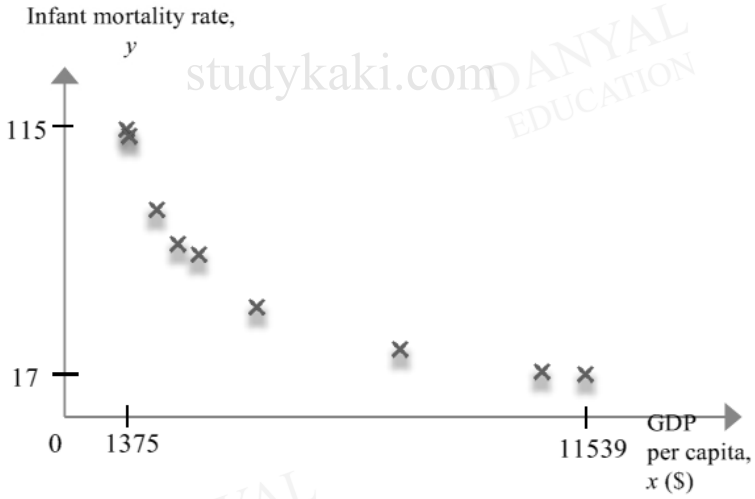
For (A): $y = a + bx^2$, with $a > 0$ and $b < 0$.



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

(ii)





(iii)

From GC, the product moment correlation coefficient is −0.898 (3s.f.). Since −0.898 is close to −1, it suggests a strong negative linear correlation between X and Y. However, it can be observed from the scatter plot that the values of y are decreasing at a decreasing rate with increasing values of x, which will not be the case if the data follows a linear model (the decrease in y should be approximately constant for a linear model).

(iv)

(B) is the appropriate model.

Using model (B), the product moment correlation coefficient is -0.978 (3s.f.).

From GC, $y = 430.30 - 45.010(\ln x)$.

$\therefore y = 430 - 45.0\ln x$ (to 3s.f.)

(v)

At a GDP per capita of \$723, the infant mortality rate is estimated to be

$y = 430.30 - 45.010(\ln 723) = 133.98$ (5s.f.) $= 134$ (3s.f.)

Since $x = 723$ is outside the range of the data values, the estimation is not reliable.