Exam Papers Practice

4.2 Correlation & Regression

Mark Schemes

Course	DP IB Maths
Section	4. Statistics & Probability
Topic	4.2 Correlation & Regression
Difficulty	Medium

To be used by all students preparing for DP IB Maths AI SL
Students of other boards may also find this useful

Question 1

a) (Fairly strong) positive correlation. The better a student performs on the maths test the better they tend to perform on the physics test.

b) (Strong) negative correlation. The more trees a client hugged the lower their reported level of anxiety.

Question 2

a) Input data into your GDC and perform a linear regression $(ax + b)$.

x list: T

y list: C

i) $a = -1.756...$
$= -1.76$ (3sf)

$b = 43.195...$
$= 43.2$ (3sf)

$$C = -1.76T + 43.2$$

ii) $r = -0.9425...$

$$r = -0.942 \text{ (3sf)}$$

b) Sub $T = 11$ into C

$C = -1.76(11) + 43.2$

$= 23.8780... \approx 24$

24 cups of tea

N.B calculator values for a and b used.

c) The estimate from part (b) is made by interpolation and the correlation is strong (r is close to -1).

∴ Very confident that the estimate is accurate.

Question 3

a) Input data into your GDC and perform a linear regression ($ax + b$).

x list: age

y list: height

i) $a = 5.8757...$
$= 5.88$ (3sf)

$b = 78.7259...$
$= 78.7$ (3sf)

$y = 5.88x + 78.7$

ii) $r = 0.9843...$

$r = 0.984$ (3sf)

b) Sub $x = 9$ into y.

$y = 5.88(9) + 78.7$

$y = 131.6079...$

$y = 132$ cm

N.B calculator values for a and b used.

c) The regression line y on x should only be used to find y when given a value x.

Question 4

a) Input data into your GDC and perform a linear regression $(ax + b)$.

x list: distance

y list: calories

i) $a = 62.2075...$
$= 62.2$ (3sf)

$b = 18.7681...$
$= 18.8$ (3sf)

$y = 62.2x + 18.8$

ii) $r = 0.9907...$

$r = 0.991$ (3sf)

b) Rebecca will burn an extra 62.2 calories for every extra 1 km ran.

c) Sub $x = 8$ into y

$y = 62.2(8) + 18.8$

$y = 516.4285...$

$y = 516$ calories (3sf)

N.B calculator values for a and b used.

d) The answer from part (c) is valid and reliable as it was drawn by interpolation and r is very strong (close to 1).

Question 5

a) Input data into your GDC and perform a linear regression $(ax + b)$.

x list: age

y list: percentage of willing people

i) $a = 0.6742...$
$= 0.674$ (3sf)

$b = 38.3809...$
$= 38.4$ (3sf)

$V = 0.674A + 38.4$

ii) $r = 0.9437...$

$r = 0.944$ (3sf)

b) As a person's age increases by 1 year, their age groups approval of the vaccine increases by 0.674%.

c) Sub $A = 95$ into V.

$V = 0.674(95) + 38.4$

$V = 102.4380...$

$V = 102\%$ (3sf)

N.B calculator values for a and b used.

d) The answer in part (c) was drawn via extrapolation, hence it is unreliable. Additionally the percentage is over 100% which is not possible.

Question 6

a) Input data into your GDC and perform a linear regression $(ax + b)$.

x list: distance

y list: price

i) $a = 0.06289...$
$= 0.0629$ (3sf)

$b = 29.0623...$
$= 29.1$ (3sf)

$P = 0.0629d + 29.1$

ii) $r = 0.9634...$

$r = 0.963$ (3sf)

b) Sub $d = 2635$ into P.

$P = 0.0629(2635) + 29.1$

$P = 194.7836...$

$P = 195$ US dollars (3sf)

N.B calculator values for a and b used.

c) This is significantly more than the answer in part (b). The mathematical reason for this is that the answer in part (b) was drawn via extrapolation (2635 km > 1930 km). An additional reason is the other locations are all in Europe where Cairo is in Africa.

Question 7

Complete the information in the following table.

Flavours	A	B	C	D	E	F	G	H
Idris' rank	1	7	3	6	8	2	5	4
Jameel's rank	4	3	1	8	5	7	2	6
Kevin's rank	8	3	6.5	2	1	6.5	4	5

a) Rank the scores in ascending order.

ie. Score = 1 $\therefore$ Rank = 1

Score = 10 $\therefore$ Rank = 8

b) Input the ranks for each part into your GDC and perform a linear regression.

i) x list: Idris y list: Jameel

$r_s = 0.04761...$ $r_s = 0.0476$ (3sf)

ii) x list: Idris y list: Kevin

$r_s = -0.9707...$ $r_s = -0.971$ (3sf)

iii) x list: Jameel y list: Kevin

$r_s = -0.2395...$ $r_s = -0.240$ (3sf)

c) i) The correlation between Idris and Jameel is almost zero, so there is no way of guessing what flavours one likes based on what the other likes.

ii) Idris and Kevin have a strong negative correlation (close to -1), meaning Idris hates what Kevin likes and vice versa.

iii) Jameel and Kevin have a weak negative correlation, meaning there is a slight tendency for one to like what the other does not and vice versa.

d) The rank of flavour A does not change, therefore this will not change any of the answers in part (b).

Question 8

a) Input data into your GDC and perform a linear regression.

i) With student J.

x list: maths y list: physics

$r = 0.7695...$ $r = 0.770$ (3sf)

ii) Input data into your GDC and perform a linear regression.

Without student J.

x list: maths y list: physics

$r = 0.06696...$ $r = 0.0870$ (3sf)

Student	A	B	C	D	E	F	G	H	I	J
Maths rank	6	3	5	8.5	7	2	1	4	8.5	10
Physics rank	9	8	6	7	4	3	5	1	2	10

b) Order the scores from 1-10 and fill in the table.

c) Input the ranks for each part into your GDC and perform a linear regression.

i) With student J.

x list: maths y list: physics

$r_s = 0.3039...$ $r_s = 0.304$ (3sf)

ii) Without student J.

x list: maths y list: physics

$r_s = 0.04184...$ $r_s = 0.0418$ (3sf)

d) Student J is an outlier.

In part (a) we used PMCC (r) and in part (c) we used Spearman's rank c.c (r_s).

Without student J both versions of the c.c show virtually no correlation (close to 0).

With student J included both versions of the c.c are affected, however r_s is less affected.

This is because r_s is less affected by outliers.