Correlation and regression

Two measurements on each member of a sample: do they move together? The correlation coefficient r says how closely the points follow a line; the regression line is that line, and it predicts one variable from the other, as long as we stay inside the data.

I SCATTER DIAGRAMS AND CORRELATION

SL 4.4 · bivariate data, Pearson's r, correlation and causation

I.1 Describing a correlation from r

Definition 1 *Pearson's correlation coefficient*

For bivariate data (x, y), **Pearson's correlation coefficient** r (GDC) measures how close the points are to a straight line: $-1 \leq r \leq 1$. Close to 1: strong positive linear correlation; close to -1: strong negative; close to 0: none.

Example 1

Describe the correlation when $r = -0.81$, when $r = 0.35$ and when $r = 0.02$.

WHAT YOU DO	WHAT YOU WRITE
Step 1 The sign gives the direction	negative; positive; positive
Step 2 The size gives the strength	0.81 strong; 0.35 weak; 0.02 none
Step 3 Say it in words	strong negative; weak positive; no linear correlation

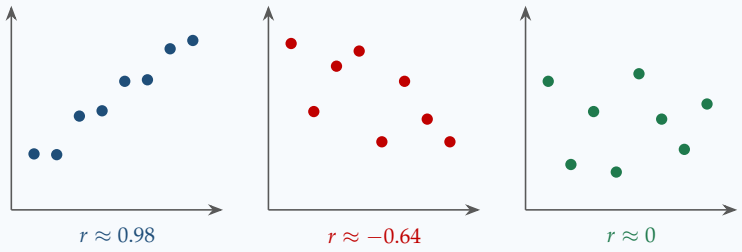


Figure 1 — Three scatter diagrams: $r \approx 0.98$, strong positive; $r \approx -0.64$, moderate negative; $r \approx 0$, none.

APPLY

1 Describe $r = 0.96$.

2 Describe $r = -0.47$.

I.2 r on the GDC

Example 2

Find r for x: 1, 2, 3, 4, 5, 6 and y: 2, 5, 4, 8, 9, 11, and describe the correlation.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Enter x and y in two lists	Statistics: lists $X1$, $Y1$
Step 2 Linear regression gives r	$r = 0.961$
Step 3 Describe it	strong positive linear correlation

APPLY

3 Find r for x: 2, 4, 6, 8 and y: 9, 7, 6, 2.

I.3 Correlation is not causation

Watch out 1

Correlation is not causation: both variables may depend on a third. And r measures a *linear* link only: points on a parabola can have $r \approx 0$.

Example 3

Across many towns, the number of doctors and the number of deaths each year are strongly positively correlated. Do doctors cause deaths?

WHAT YOU DO	WHAT YOU WRITE
Step 1 Look for a third variable	the population of the town
Step 2 Conclude in words	no: larger towns have more of both; the correlation shows no cause

APPLY

4 In children, shoe size and reading score are positively correlated. Why?

II REGRESSION LINES

SL 4.4 · 4.10 · y on x, x on y, interpolation and extrapolation

II.1 The regression line of y on x

Definition 2 *regression line of y on x*

The **regression line of y on x**, $y = ax + b$ from the GDC, makes the sum of the squared vertical distances to the points as small as possible. It passes through the mean point $(\bar{x}, \bar{y})$.

Example 4

For the data of Example 2, find the regression line of y on x and check that it passes through the mean point.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Linear regression on the GDC	$y = 1.74x + 0.4 \;\; (a = 1.742857\ldots)$
Step 2 The mean point	$\bar{x} = 3.5, \;\; \bar{y} = 6.5$
Step 3 Check	$1.742857 \times 3.5 + 0.4 = 6.5$

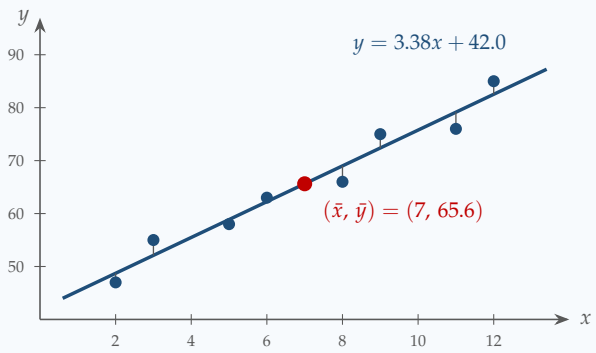


Figure 2 — Exam-style question 3's eight points, their mean point $(7, 65.6)$ and the regression line $y = 3.38x + 42.0$.

APPLY

1 The regression line of y on x for x: 2, 4, 6, 8 and y: 9, 7, 6, 2.

II.2 Predicting: interpolation and extrapolation

Property 1 *predicting*

Estimate y from x with the line of y on x; estimate x from y with the line of x **on** y (SL 4.10). Reliable when $|r|$ is close to 1 and the value lies inside the data (**interpolation**); outside (**extrapolation**) it is not.

Example 5

The line $y = 0.9x + 12$ was found from data with $20 \leq x \leq 60$ and $r = 0.95$. Estimate y when $x = 45$ and when $x = 90$, and comment.

WHAT YOU DO	WHAT YOU WRITE
Step 1 Substitute $x = 45$	$y = 0.9 \times 45 + 12 = 52.5$
Step 2 Inside the data, $\|r\|$ close to 1	reliable (interpolation)
Step 3 $x = 90$ is outside the data	93, but not reliable (extrapolation)

APPLY

2 Example 4's line: estimate y when $x = 4.5$.

3 A line with $r = 0.12$: is an estimate inside the data reliable?

II.3 The mean point

Example 6

The regression line of y on x is $y = 2.4x + 7$ and $\bar{x} = 5$. Find $\bar{y}$.

WHAT YOU DO	WHAT YOU WRITE
Step 1 The line passes through $(\bar{x}, \bar{y})$	substitute $x = \bar{x}$
Step 2 Compute	$\bar{y} = 2.4 \times 5 + 7 = 19$

APPLY

4 $y = ax + 2$, $\bar{x} = 8$, $\bar{y} = 26$: find a.

II.4 The line of x on y

Watch out 2

Do not rearrange $y = ax + b$ to find x: the line of x on y is a different line, unless $r = \pm 1$.

Example 7

For the data of Example 2, estimate x when $y = 7$.

WHAT YOU DO	WHAT YOU WRITE
Step 1 x from y: the line of x on y (GDC)	$x = 0.530y + 0.0522$
Step 2 Substitute $y = 7$	$x = 3.77$
Step 3 Inside the data, $r = 0.961$	a reliable estimate

APPLY

5 For x: 2, 4, 6, 8 and y: 9, 7, 6, 2, estimate x when $y = 5$.

III EXAM-STYLE QUESTIONS

from the shortest to the longest · every one in steps, with its marks

A Reading r and a regression line

EXAM-STYLE QUESTION 1 Paper 1 · no calculator · [Maximum mark: 4]

For 12 cars, the age x years and the value y thousand euros give $r = -0.92$ and the regression line of y on x, $y = -2.5x + 80$. The ages range from 2 to 15 years, and the mean age is 10 years.

(a) Describe the correlation between age and value. [1]

(b) Use the regression line to estimate the value of a car aged 12 years. [1]

(c) Explain why the line should not be used to estimate the value of a car aged 40 years. [1]

(d) Find the mean value of the 12 cars. [1]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Sign and size of r, in words	strong negative (linear) correlation	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Substitute $x = 12$	$-2.5 \times 12 + 80 = 50$: 50 000 euros	A1

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Outside the data: say why	40 is far outside the data: extrapolation *The line also gives a negative value.*	R1

WORKING (d)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 The mean point: substitute $\bar{x}$	$-2.5 \times 10 + 80 = 55$: 55 000 euros *The regression line passes through $(\bar{x}, \bar{y})$.*	A1

EXAM-STYLE QUESTION 2 Paper 1 · no calculator · [Maximum mark: 4]

For a set of bivariate data, the regression line of y on x is $y = 0.8x + 4$ and $\bar{x} = 15$.

(a) Find $\bar{y}$. [2]

(b) Explain why this line should not be used to estimate x when $y = 20$. [2]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 The line passes through $(\bar{x}, \bar{y})$	the line passes through $(\bar{x}, \bar{y})$	(M1)
Step 2 Substitute	$\bar{y} = 0.8 \times 15 + 4 = 16$	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Estimating x, not y	to estimate x from y	R1
Step 2 Name the right line	use the regression line of x on y	R1

Not the line of y on x rearranged.

B Data on the GDC (Paper 2)

EXAM-STYLE QUESTION 3

Paper 2 · calculator · [Maximum mark: 7]

Eight students recorded the hours x they studied for a test and their score y.

x	2	3	5	6	8	9	11	12
y	47	55	58	63	66	75	76	85

(a) Find the value of Pearson's correlation coefficient r. [1]

(b) Find the equation of the regression line of y on x. [2]

(c) Estimate the score of a student who studied for 10 hours. [2]

(d) Estimate the number of hours studied by a student who scored 70. [2]

WORKING (a)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 r on the GDC	$r = 0.979$ $(0.979458\ldots)$	A1

WORKING (b)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Regression line of y on x	$y = 3.38x + 42.0$ $(3.38043\ldots, 41.9619\ldots)$ *A1 for each coefficient.*	A1A1

WORKING (c)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 Substitute $x = 10$	$3.38043 \times 10 + 41.9619$	(M1)
Step 2 Compute	75.8 $(75.7663\ldots)$	A1

WORKING (d)

WHAT YOU DO	WHAT YOU WRITE	
Step 1 The line of x on y	line of x on y: $x = 0.284y - 11.6$ *Not the line of (b) rearranged, which gives 8.29.*	(M1)
Step 2 Substitute $y = 70$	8.24 hours $(8.24158\ldots)$	A1

IV WHICH METHOD? READ THE QUESTION

4.10 · the whole topic on one page

THE QUESTION SAYS	YOU KNOW	YOU WRITE
"describe the correlation"	the sign and the size of r	strong / moderate / weak, positive / negative D1
"find r", "find the regression line"	the GDC	r to 3 s.f.; $y = ax + b$ with a, b to 3 s.f. D2
"estimate y when $x = \ldots$"	the line of y on x	substitute; check it is interpolation P1
"estimate x when $y = \ldots$"	the line of x on y	the other line, from the GDC P1
"is the estimate reliable?"	$\lvert r\rvert$ and the range of the data	yes if $\lvert r\rvert \approx 1$ and inside; no if outside or r weak P1
the mean point, or an unknown mean	$(\bar{x}, \bar{y})$ is on the line	substitute $\bar{x}$ to get $\bar{y}$ D2
"does x cause y?"	correlation only	no conclusion about cause from r W1

In the exam 1 *command terms*

Describe the correlation: strength and direction. **Interpret** the gradient: "for each extra unit of x, y increases by a on average". **Justify**: interpolation or extrapolation, and the value of r.

V ON THE GDC: PAPER 2

SL 4.4 · 4.10 · NumWorks · every calculator move of this topic

1 r and the regression line

Statistics → V1: x, V2: y → Regression → Linear

Enter x and y as one series, choose the linear model. **On paper:** $r = 0.979$ and $y = 3.38x + 42.0$.

STATISTICS · REGRESSION

```
X1: 2,3,5,6,8,9,11,12
Y1: 47,55,58,63,66,75,76,85
y=a*x+b a=3.380435 b=41.96196
r=0.9794583
```

2 The line of x on y

Statistics → swap the columns → Regression → Linear

Swap the roles of the two lists: the GDC then gives the line of x on y. **On paper:** $x = 0.284y - 11.6$.

STATISTICS · REGRESSION

```
X1: 47,55,58,63,66,75,76,85
Y1: 2,3,5,6,8,9,11,12
a=0.2837915 b=-11.62382
```

ANSWERS TO THE APPLY QUESTIONS

I.1 strong positive

I.2 weak (to moderate) negative

I.3 -0.965

I.4 both grow with age

II.1 $y = -1.1x + 11.5$

II.2 8.24

II.3 no: the correlation is very weak

II.4 $a = 3$

II.5 $x = -0.846y + 10.1$: 5.85

Written and verified by a practising IB mathematics teacher, produced with AI assistance; every result recomputed by software, every figure computed from its function. Not affiliated with or endorsed by the International Baccalaureate Organization. © 2026 Hence