



eClassroom

A-LEVEL MATHEMATICS • STATISTICS 1

Correlation and Regression

Worksheet

Pearson Edexcel International A Level (WST01) & A Level (9MA0) • Calculator and statistical tables allowed: show every step

Free Correlation and Regression worksheet for A-Level and International A-Level Maths, with full worked solutions



Examples
& Your turn
page 2



Fluency
core skills
page 4



Reasoning
explain it
page 6



Problem
solving
page 7



Exam
practice
page 9

Tap a section to jump straight to it

By the end of this worksheet you can: calculate S_{xx} , S_{yy} and S_{xy} • find and interpret the product moment correlation coefficient • find the regression line of y on x • interpret and use the line • *Stretch:* coding, units, rates of change, outliers and the reliability of estimates.

Which questions are for me?

CORE The methods every A-level student must secure. Start here.

STRETCH Summary statistics, coded variables, lines that must be interpreted in context, and estimates you must judge.

Every section has both, clearly labelled. Questions get harder as you go, and every skill in Sections A to D is needed in Section E.

Track your progress

Section	B	C	D	E
Core marks	23	10	13	21
Stretch marks	16	14	17	13
My score				

Full worked solutions

Exam questions by topic

Statistics 1 exam questions

Formula sheet builder

Name

Class

Date



Join students already acing their exams.

Book your FREE lesson at www.eclassroom.org.uk

Page 1 of 11

A

Worked examples & Your turn

Read each example carefully. Cover it up, then answer the question beside it using the same method.

CORE

Secure the method

Worked example A1: S values and r

$n = 6$, $\sum x = 30$, $\sum y = 48$, $\sum x^2 = 170$, $\sum y^2 = 410$, $\sum xy = 260$. Find r

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \text{ and similarly for } S_{yy} \text{ and } S_{xy}$$

$$S_{xx} = 20, S_{yy} = 26, S_{xy} = 260 - 240 = 20$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{20}{\sqrt{520}}$$

$$r = 0.877$$

Your turn A1

$n = 5$, $\sum x = 20$, $\sum y = 35$, $\sum x^2 = 90$, $\sum y^2 = 265$, $\sum xy = 150$. Find r

Worked example A2: the regression line

$S_{xx} = 40$, $S_{xy} = 60$, $\bar{x} = 5$, $\bar{y} = 20$. Find the regression line of y on x

$$b = \frac{S_{xy}}{S_{xx}}, \text{ then } a = \bar{y} - b\bar{x}$$

$$b = 1.5, a = 20 - 7.5$$

$$y = 12.5 + 1.5x$$

Your turn A2

$S_{xx} = 25$, $S_{xy} = -40$, $\bar{x} = 4$, $\bar{y} = 30$. Find the regression line of y on x

Worked example A3: interpreting r

For 20 students, $r = -0.92$ between hours of screen time and hours of sleep. Interpret this value.

Strength, direction, and the context.

strong negative: more screen time, less sleep

Your turn A3

For 15 cities, $r = 0.35$ between population and rainfall. Interpret this value.



STRETCH

Worked example A4: using the line

$y = 12.5 + 1.5x$ was found from data with x from 2 to 9. Estimate y when $x = 6$ and when $x = 20$, and comment.

$x = 6$: 21.5, inside the data (interpolation)

$x = 20$: 42.5, far outside the data (extrapolation)

21.5 reliable; 42.5 unreliable

Your turn A4

$y = 36.4 - 1.6x$ was found from data with x from 1 to 8. Estimate y when $x = 5$ and when $x = 25$, and comment.

Worked example A5: decoding a line

$p = x - 50$ and $q = \frac{y}{10}$. The line of q on p is $q = 2 + 0.3p$. Find the line of y on x

Substitute both codings, then rearrange for y

$$\frac{y}{10} = 2 + 0.3(x - 50), \text{ so } y = 20 + 3x - 150$$

$$y = -130 + 3x$$

Your turn A5

$p = \frac{x - 20}{2}$ and $q = y - 100$. The line of q on p is $q = 5 + 1.2p$. Find the line of y on x

Worked example A6: S values under coding

$w = 4d$. Given $S_{ww} = 400$ and $S_{tw} = 80$, find S_{dd} and S_{td}

$$S_{ww} = 4^2 S_{dd} \text{ and } S_{tw} = 4S_{td}$$

$$S_{dd} = \frac{400}{16}, \quad S_{td} = \frac{80}{4}$$

$$S_{dd} = 25, \quad S_{td} = 20$$

Your turn A6

$w = \frac{d}{2}$. Given $S_{ww} = 36$ and $S_{tw} = -12$, find S_{dd} and S_{td}

Worked example A7: a rate of change

$y = 4.2 + 0.35x$. By how much must x increase for y to increase by 7?

Use the gradient only: $0.35 \times (\text{increase in } x) = 7$

20

Your turn A7

$g = 15 + 0.8h$. By how much must h increase for g to increase by 12?



B**Fluency**

Each question is a designed sequence. From one part to the next, ask yourself: what changed, and what did it do to the answer?

CORE*Secure the method***B1** The data are

x	2	4	5	7	9
y	3	6	7	10	14

(a) Find S_{xx} **(b)** Find S_{yy} **(c)** Find S_{xy} **(d)** Find r **[4 marks]****B2** For the data in B1,**(a)** find b **(b)** find a **(c)** estimate y when $x = 6$ **(d)** interpret the gradient.**[4 marks]****B3** Describe the correlation shown by each value, or explain why it is impossible.**(a)** $r = 0.95$ **(b)** $r = -0.3$ **(c)** $r = 1.2$ **(d)** $r = -0.98$ **[4 marks]****B4** $n = 10$, $\sum x = 60$, $\sum y = 250$, $\sum x^2 = 400$, $\sum y^2 = 6600$ and $\sum xy = 1580$ **(a)** Find S_{xx} , S_{yy} , S_{xy} **(b)** Find r **(c)** Find b **(d)** Find the line of y on x **[4 marks]**

CORE

continued: using lines

- B5** The line $y = 13 + 2x$ was found from data with x from 2 to 10
- (a) Estimate y when $x = 7$
 - (b) Estimate y when $x = 25$
 - (c) State, with a reason, which estimate is more reliable. [3 marks]
- B6** A study records the number of hours, h , that 12 students revised and their test marks, m
- (a) State the explanatory variable.
 - (b) State which regression line should be used to estimate a mark from hours of revision.
 - (c) Explain why this line should not be used to estimate hours from a mark.
 - (d) $r = 0.88$. Explain why this does not prove that revising causes higher marks. [4 marks]

STRETCH

- B7** $p = x - 100$ and $q = \frac{y - 50}{4}$. For the coded data, $r = 0.9$ and the line of q on p is $q = 1.2 + 0.8p$
- (a) Write down r for x and y
 - (b) Find the line of y on x
 - (c) Estimate y when $x = 110$
 - (d) Explain why r did not change. [4 marks]
- B8** $w = 5d$. $S_{tt} = 20$, $S_{tw} = 120$ and $S_{ww} = 900$
- (a) Find S_{dd}
 - (b) Find S_{td}
 - (c) Find r between t and d
 - (d) Find the gradient of the line of d on t [4 marks]
- B9** For plants, the mass y g and height x cm satisfy the regression line $y = -8 + 0.6x$
- (a) By how much does y increase for each 10 cm increase in height?
 - (b) By how much must the height increase for the mass to increase by 15 g?
 - (c) Estimate the mass of a plant of height 1.2 m
 - (d) Explain what the line predicts for a plant 10 cm tall, and what this shows. [4 marks]
- B10** For the data in B4, a new point is added at $(\bar{x}, \bar{y})$
- (a) State $\bar{x}$ and $\bar{y}$
 - (b) State the new S_{xx} , S_{yy} , S_{xy}
 - (c) State the new value of r
 - (d) State the new regression line. [4 marks]



C**Reasoning**

Show your working and explain your thinking. A correct answer without a reason does not score full marks.

CORE*Secure the method*

- C1** For outdoor temperature and the number of cold drinks sold, $r = 0.91$. Ama writes “strong positive correlation”.
Explain why this would not score the interpretation mark, and improve it. [2 marks]
- C2** Weekly sales y are in hundreds of dollars, and $y = 1.5 + 0.8x$. Ben estimates sales of \$5.50 when $x = 5$.
Explain his error, and give the correct estimate. [2 marks]
- C3** $S_{xx} = 50$, $S_{yy} = 72$ and $S_{xy} = 54$. Cal writes $r = \frac{54}{50} = 1.08$.
Explain why his answer must be wrong, and find r . [3 marks]
- C4** $w = 4d$ and $S_{ww} = 320$. Dee writes $S_{dd} = \frac{320}{4} = 80$.
Explain her error, and find S_{dd} . [3 marks]

STRETCH

- C5** Data with x from 2 to 12 give the line $y = 25 - 1.5x$, where y is a waiting time in minutes. Eli says the estimate for $x = 20$ is unreliable “because it is extrapolation”.
Find the estimate, and give a stronger reason. [3 marks]
- C6** $g = 10 + 0.4h$. To find the increase in h that gives a 6 cm increase in g , Fin solves $6 = 10 + 0.4h$.
Explain his error, and find the correct increase. [3 marks]
- C7** The regression line of y on x is $y = 4 + 2x$. Gail uses it to estimate x when $y = 20$, and gets 8.
Explain why this is not appropriate. [3 marks]
- C8** (a) Show that the regression line $y = a + bx$ passes through $(\bar{x}, \bar{y})$. (2)
(b) Explain why the gradient b and the correlation coefficient r always have the same sign. (3)
[5 marks]



D**Problem solving**

You decide the method. Read the whole question first, and show every step.

CORE

Secure the method

- D1** A gardener records the height, h cm, and the leaf length, l cm, of each of 40 plants of the same variety. The following summary statistics are calculated from these data.

$$\sum h = 1240 \quad \sum h^2 = 39400 \quad S_{ll} = 96 \quad S_{hl} = 108$$

- (a) Show that $S_{hh} = 960$

(1)

- (b) Calculate the product moment correlation coefficient between h and l , and interpret your value in context.

(2)

[3 marks]

- D2** A coach records the number of training sessions, x , and the number of goals scored, y , by each of 8 players in a season. The data are summarised as

$$\sum x = 52 \quad \sum y = 140 \quad \sum xy = 950 \quad \sum y^2 = 2530 \quad S_{xx} = 42$$

- (a) Find the exact value of S_{xy} and the exact value of S_{yy}

(2)

- (b) Calculate the product moment correlation coefficient between x and y , and interpret it in context.

(2)

[4 marks]

- D3** A courier records the distance, x miles, and the time taken, t minutes, for 25 deliveries, and finds the regression line of t on x . All of the distances are then converted to kilometres by multiplying by 1.6, and the times are not changed.

State, giving a reason in each case, the effect of this conversion on the product moment correlation coefficient, on the gradient of the regression line and on its intercept.

[3 marks]

- D4** For 20 children, the regression line of height, h cm, on age, t months, is $h = 12.4 + 0.86t$, and the product moment correlation coefficient between h and t is 0.94

The data are coded using $x = t - 10$ and $y = \frac{h - 20}{2}$

Write down the product moment correlation coefficient between x and y , and find the equation of the regression line of y on x

[3 marks]



STRETCH

- D5** A company records the outside temperature, t °C, and the daily heating cost, d dollars, on 15 days. The costs are coded using $w = 8d$, and

$$S_{tt} = 40 \quad S_{tw} = -224 \quad S_{ww} = 1600$$

The regression line of w on t is $w = 50 - 5.6t$

- (a) Find S_{td} and S_{dd} , and hence the product moment correlation coefficient between t and d

(3)

- (b) Write down the equation of the regression line of d on t

(1)

[4 marks]

- D6** For a sample of adults, the regression line of hand span, g cm, on height, l cm, is $g = -30 + 0.65l$

- (a) Use the regression line to estimate the hand span of an adult of height 1.8 m

(2)

- (b) The line predicts that, for each increase of x cm in height, hand span increases by 13 cm. Find the value of x

(2)

[4 marks]

- D7** From the same 10 observations, the regression line of y on x is $y = 2x + 3$ and the regression line of x on y is $x = 0.4y + 0.6$

Both regression lines pass through the point $(\bar{x}, \bar{y})$. Find $\bar{x}$ and $\bar{y}$, and hence find $\sum x$ and $\sum y$. You must show your working clearly.

[4 marks]

- D8** The midday temperature, f °F, is recorded in a city on 12 days, giving $\sum f = 744$ and $\sum f^2 = 46\,548$

- (a) Find the mean and the standard deviation of these temperatures.

(2)

The temperatures are converted to °C using $c = \frac{5}{9}(f - 32)$

- (b) Find the mean and the standard deviation of the temperatures in °C

(2)

- (c) The product moment correlation coefficient between the temperature in °F and the number of visitors to a park is 0.83. State, with a reason, its value when the temperature is in °C

(1)

[5 marks]



E**Exam practice**

Full-size exam questions in the style of WST01. Write your answers in the spaces, and show every step.

CORE

Exam questions for everyone

- E1** Statistical models are often used by scientists and engineers to describe and make predictions about real-world situations.

(a) Give one reason why statistical models are used.

(1)

An engineer wants to model the relationship between the weather and the output of a solar panel. She records the number of hours of sunshine, s , and the electricity generated by the solar panel, e kWh, on each of 14 randomly selected days in summer. To simplify the calculations, the electricity values are coded using $y = 10e$, and the coded data are summarised by

$$\sum s = 98 \quad \sum y = 616 \quad S_{ss} = 52.5 \quad S_{sy} = 315 \quad S_{yy} = 2016$$

(b) Calculate the product moment correlation coefficient between s and y

(2)

(c) Give an interpretation, in context, of your product moment correlation coefficient.

(1)

(d) Show that the regression line of y on s is $y = 2 + 6s$

(3)

(e) Hence find the equation of the regression line of e on s

(1)

(f) Use your answer to part (e) to estimate the electricity generated by the solar panel on a day with 9 hours of sunshine, and give an interpretation, in context, of the gradient of your regression line.

(2)

(Total for Question E1 is 10 marks)



- E2** A forester believes that taller trees tend to have thicker trunks. She measures the trunk diameter, d cm, and the height, h metres, of each of 10 trees chosen at random from a wood. The data are summarised as follows.

$$\sum d = 450 \quad \sum h = 180 \quad \sum dh = 8320 \quad \sum h^2 = 3340 \quad S_{dd} = 1250$$

- (a) Find the exact values of S_{dh} and S_{hh}

(3)

- (b) Calculate the product moment correlation coefficient between d and h

(2)

- (c) Show that the equation of the regression line of h on d can be written as $h = 10.08 + 0.176d$

(3)

- (d) Give an interpretation, in context, of the gradient of this regression line for the trees in the wood.

(1)

- (e) Use the regression line from part (c) to estimate the height of a tree in this wood with a trunk diameter of 0.5 metres, giving your answer to 3 significant figures.

(2)

(Total for Question E2 is 11 marks)



STRETCH

- E3** A teacher records the number of hours, x , that each of 10 students in her class spent revising for a test, and the mark, y , that each student scored in the test. The test was marked out of 100

Student	A	B	C	D	E	F	G	H	I	J
Hours, x	2	4	5	6	7	8	9	10	11	12
Mark, y	30	38	95	44	48	50	55	58	62	66

An outlier is a value greater than $Q_3 + 1.5(Q_3 - Q_1)$ or less than $Q_1 - 1.5(Q_3 - Q_1)$

- (a) Show that the mark of student C is an outlier among the marks.

(3)

Student C is removed. For the remaining 9 students, $\sum x = 69$, $\sum y = 451$, $\sum x^2 = 615$ and $\sum xy = 3761$

- (b) Show that the regression line of y on x for these 9 students is $y = 23.1 + 3.53x$

(3)

- (c) Interpret the gradient of this regression line in context.

(1)

- (d) Use the regression line to estimate the mark of a student who revises for 7.5 hours, and the mark of a student who revises for 30 hours. Comment on the reliability of each of your estimates.

(3)

- (e) State, giving a reason, whether the product moment correlation coefficient for all 10 students would be greater or less than the product moment correlation coefficient for the 9 students.

(1)

- (f) Using the regression line, find the increase in the estimated test mark for each extra 5 hours that a student spends revising.

(2)

(Total for Question E3 is 13 marks)

Want more exam practice? The **Exam Questions by Topic** booklets have 10 more full exam questions on Statistics 1 Chapter 5, Correlation and Regression, with a full mark scheme.

[Open the booklet](#)

Finished? Mark your work with the **full worked solutions**, fill in your scores on the front page, then practise real exam questions on the **Statistics 1 exam questions page**.



Join students already acing their exams.

Book your FREE lesson at www.eclassroom.org.uk