# Mathematics

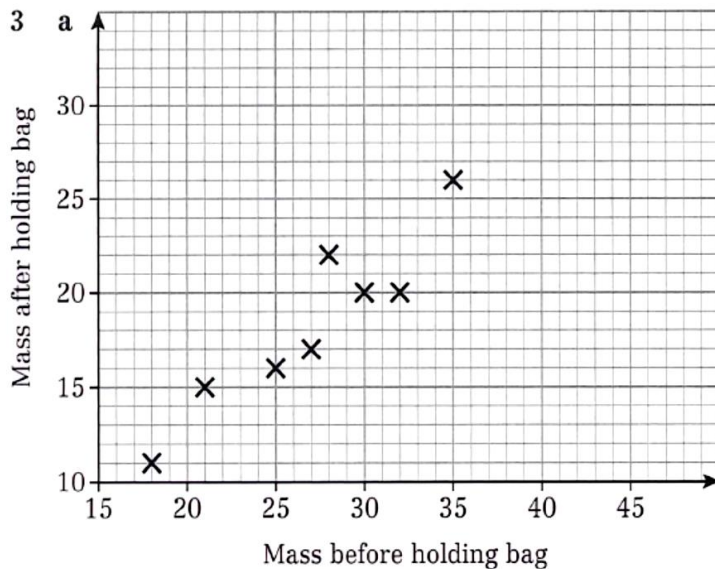## Edexcel IAL

# S1

## Worksheet Answers

### Correlation and Regression

### Eng. Nagy Elraheb

## Correlation and Regression:

## Exercise 1:

1   a   Positive correlation.
    b   The longer the treatment, the more hair growth observed.
2   a   No correlation.
    b   The scatter diagram does not support the statement that hotter cities have less rainfall.

3   a



Mass before holding bag

    b   There is positive correlation. If a student guessed a greater mass before touching the bag, they were more likely to guess a greater mass after touching it.

4   a



Time taken (s)
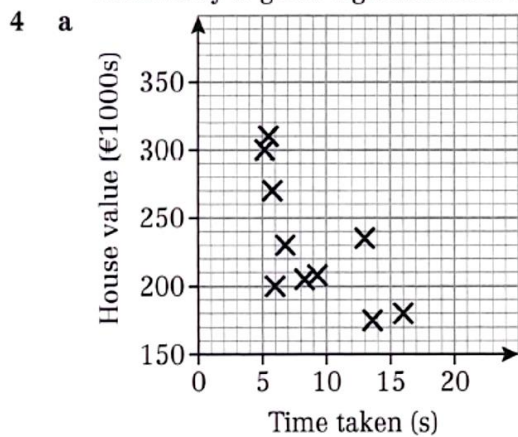
    b   Weak negative correlation.
    c   For example, there may be a third variable that influences both house value and internet connection, such as distance from built-up areas.

**5 b i** There is no reason to believe that the data collected is incorrect.

**ii** 22.3 is an outlier so may not be representative of the typical rainfall.

**c**
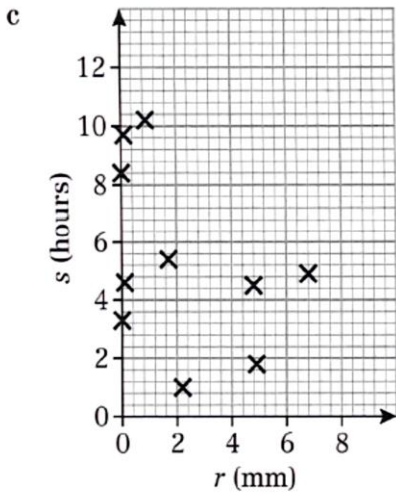


**d** No correlation

**e** For example, there could be a causal relationship as days with more rainfall will have more clouds, and therefore less sunshine.

# Exercise 2:

**1 a, b**
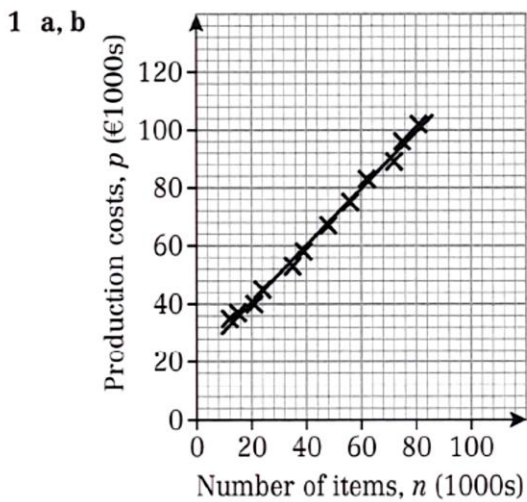


**c** If the number of items produced per month is zero, the production costs will be approximately €21,000. If the number of items per month increases by 1000 items, the production costs increase by approximately €980.

**d** The prediction for 74 000 is within the range of the data (interpolation) so is more likely to be accurate. The prediction for 95 000 is outside the range of the data (extrapolation) so is less likely to be accurate.

**2 a**
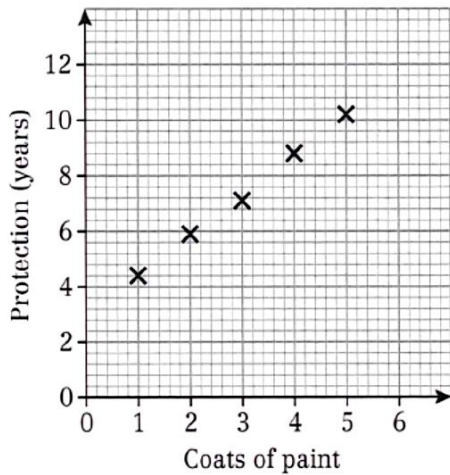


Coats of paint

**b** A gradient of 1.45 means that for every extra coat of paint, the protection will increase by 1.45 years, therefore if 10 coats of paint are applied, the protection will be 14.5 years longer than if no coats of paint were applied. After 10 coats of paint, the protection will last 2.93 + 14.5 = 17.43 years.

**3 a**



Age, $a$ (months)
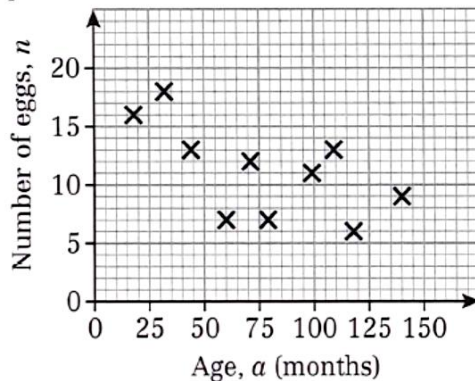
**b** The scatter diagram shows negative correlation, therefore the gradient in the regression equation should be negative.

**4** This is not sensible as there are unlikely to be any houses with no bedrooms.

**5 a** Each visitor spends €740.

  **b** €1161 million

  **c** As this involves interpolation the value of money spent is reliable.

## Exercise 3:

**1** $a = -3, b = 6$

**2** $y = -14 + 5.5x$

**3** $y = 2x$

**4 a** $\bar{x} = 2.5, \bar{y} = 12, S_{xx} = 5, S_{xy} = 20$

  **b** $y = 2 + 4x$

**5 a** $S_{xx} = 40.8, S_{xy} = 69.6$     **b** $y = -0.294 + 1.71x$

6 a $y = -59 + 57(6) = 283$
  b For each dexterity point, productivity increases by 57.
  c i No, because this is extrapolation as it is outside the range of data.
    ii No, because this is extrapolation as it is outside the range of data.
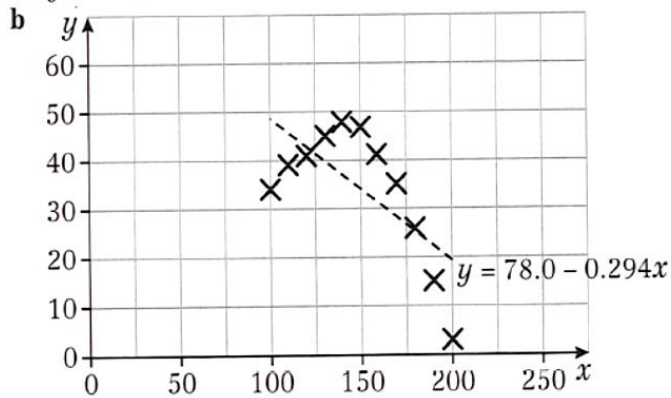7 $g = 1.50 + 1.44h$
8 a $p = 65.4 - 1.38w$
  b $w = 47.4 - 0.72p$
  c The gradient of the second regression line is calculated using different summary statistics rather than just the reciprocal of the summary statistics used for the first regression line.
  d i The first one.   ii The second one.
9 a $y = 78.0 - 0.294x$
  b



  c Model is not valid since data does not follow a linear pattern.
10 a $S_{nn} = 6486$, $S_{np} = 6344$
   b $p = 21.0 + 0.978n$
   c €60,100 (3 s.f.)
   d Reliable, as 40 000 items lies inside the range of the data.
11 a $S_{nn} = 589.6$, $S_{np} = 1474$
   b $p = 20 + 2.5n$
   c The increase in cost, in dollars, for every 100 leaflets printed.
   d $t > 8$
12 a $y = -0.07 + 1.45x$
   b Number of years protection per coat of paint.
   c Unreliable, as 7 coats lies outside the range of the data.
   d 10.08 years
   e i $0.4779 + 1.247x$
     ii 9.2 years (2 s.f.)
     iii The answer now uses interpolation not extrapolation and the number of data points has increased, which increases accuracy in prediction.

## Exercise 4:

1 $y = 6 - x$

2 $s = 88 + p$

3 $y = 32 - 5.33x$

4 $t = 9 + 3s$

5 a $y = 3.5 + 0.5x$        b $d = 35 + 2.5c$

6 a $S_{xy} = 162.2$, $S_{xx} = 190.8$; $y = 7.87 + 0.850x$ (3 s.f.)

  b $c = 22.3 + 2.13a$ (3 s.f.)

  c $90.46 or $90.56

7 a $p = 3.03 + 1.49v$ (3 s.f.)   b 10.1 tonnes (3 s.f.)


## Exercise 5:

1 0.985 (3 s.f.)

2 0.202 (3 s.f.)

3 a 9.71 (3 s.f.)

  b 0.968 (3 s.f.)

  c There is positive correlation. The greater the age, the taller the person.

4 a $S_{LL} = 30.3$, $S_{TT} = 25.1$, $S_{LT} = 25.35$

  b 0.919 (3 s.f.)

  c The value of the correlation coefficient is close to 1 and the points lie on an approximate straight line, therefore a linear regression model is suitable.

5 a 0.866 (3 s.f.)

  b There is positive correlation. The higher the IQ, the higher the mark in the general knowledge test.

6 0.973

7 a

| $p$ | 0 | 5 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| $q$ | 0 | 17 | 12 | 10 | 6 |

  b 0.974 (3 s.f.)       c 0.974 (3 s.f.)

8 a $S_{pp} = 10$, $S_{tt} = 5.2$, $S_{pt} = 7$

  b 0.971 (3 s.f.)

  c 0.971 (3 s.f.)

9 a $S_{xx} = 1601$, $S_{yy} = 1282$, $S_{xy} = -899$

  b -0.627 (3 s.f.)

  c The shopkeeper is wrong. There is negative correlation. Sweet sales actually decrease as newspaper sales increase.

10 a $S_{ff} = \Sigma f^2 - \dfrac{(\Sigma f)^2}{n} = \Sigma(10x)^2 - \dfrac{(\Sigma 10x)^2}{n}$

$= 100\Sigma x^2 - \dfrac{100(\Sigma x)^2}{n} = 100\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)$

$= 100 S_{xx} = 100 \times 111.48 = 11\,148$

  b 0.934 (3 s.f.)

  c The PMCC suggests strong linear correlation but the scatter diagram suggests non-linear fit so a linear regression model is not suitable.

**11 a** $S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n} = 22.02 - \dfrac{12^2}{7} = 1.448\ldots$

$S_{xy} = \Sigma xy - \dfrac{\Sigma x \Sigma y}{n} = 180.37 - \dfrac{12 \times 97.7}{7}$

$\qquad\qquad\qquad = 12.884\ldots$

$S_{yy} = \Sigma y^2 - \dfrac{(\Sigma y)^2}{n} = 1491.69 - \dfrac{97.7^2}{7}$

$\qquad\qquad\qquad = 128.077\ldots$

$r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{12.884\ldots}{\sqrt{1.448\ldots \times 128.077\ldots}} = 0.946 \ (3 \text{ s.f.})$

**b** −2.29345, 0.22765, 0.8382, 1.16985, 1.39095, 0.61205, −1.94575

**c** Residuals are not randomly scattered about zero (they 'rise and fall') so this indicates that a linear model is not a good model for this data.