Chapter 7 Additional Worksheets

Correlation & Regression

These additional worksheets mirror Chapter 7 (Correlation and Regression) of the AQA Use of Mathematics textbook, exercise by exercise: 7.1 / Exercise 7A, 7.2 / Exercise 7B, and 7.3 / Exercise 7C.

7.1 Lines of best fit

When two variables are related, this can be shown on a scatter diagram. If the relationship is roughly linear, a line of best fit can be drawn through the data.

Key point. To draw a line of best fit: use sensible scales for both axes, plot the data, calculate the mean point $(\bar{x}, \bar{y})$, and draw a straight line by eye through the mean point that follows the trend of the data.
Correlation does not imply causation. Two variables can be strongly correlated because a third variable (a **confounder**) affects both of them, rather than one causing the other.

Example (worked, for reference)

A researcher records the number of mobile-phone masts and the number of coffee shops in 6 towns of different sizes.

Masts, x	4	7	9	12	15	20
Coffee shops, y	2	5	6	9	12	15

a) Describe the correlation shown by this data.
b) Suggest a reason for this correlation other than one variable directly causing the other.
Solution.
a) As the number of masts increases, the number of coffee shops also increases – this is positive correlation.
b) Both variables are likely driven by a third variable: the size (population) of the town. Larger towns have both more phone masts and more coffee shops, so the mast count does not directly cause more coffee shops – population size is the confounder.

Exercise 7A

1. For each pair of variables, state whether you would expect positive correlation, negative correlation, or no correlation.
   a) Outdoor temperature and heating bill.
   b) A person's age and their score on a memory test (adults only).
   c) Number of fire engines sent to a fire and amount of fire damage.
   d) Shoe size and mathematics test score, in a sample of adults.
2. A study finds strong positive correlation between the number of people who drown at a beach each month and the number of ice creams sold that month.
   a) Explain why this does not mean that eating ice cream causes drowning.
   b) Suggest a confounding variable that could explain this correlation.
3. The table shows the population (in thousands) and the number of secondary schools in 6 towns.

Population, x	15	22	30	45	60	80
Schools, y	2	3	4	6	8	10

a) Plot a scatter graph and draw a line of best fit by eye.

b) Describe the correlation shown.

7.2 The regression line

The regression line of y on x is the specific line of best fit used to **predict** a value of y from a given value of x. Unlike a line drawn by eye, it is calculated precisely.

Key point. The regression line of y on x has equation $y = a+bx$. Once you have the equation, you can substitute a value of x to predict y.
Interpolation (predicting within the range of the data) is usually reliable. **Extrapolation** (predicting outside the range of the data) is unreliable, since the linear relationship may not continue.

Example (worked, for reference)

A gardener records the amount of fertiliser applied, x grams, and the resulting plant height, y cm, for plants with x ranging from 5 to 40 grams. The regression line of y on x is calculated to be $y = 12.4 + 1.8x$.
a) Predict the height of a plant given 20 g of fertiliser.
b) Predict the height of a plant given 200 g of fertiliser, and comment on the reliability of this prediction.
c) Interpret the value 1.8 in context.
Solution.
a) $y = 12.4 + 1.8 \times 20 = 48.4$ cm. This is interpolation (20 is within 5–40), so should be reasonably reliable.
b) $y = 12.4 + 1.8 \times 200 = 372.4$ cm. This is extrapolation (200 is far outside 5–40) – plants cannot keep growing indefinitely with more fertiliser (and may even be damaged by too much), so this prediction should not be trusted.
c) For every extra gram of fertiliser applied, plant height increases by an estimated 1.8 cm, according to this model.

Exercise 7B

1. A regression line relating hours of sunshine per day, x (ranging from 2 to 11 in the data), to daily ice cream sales in £, y, is $y = -15 + 42x$.

   a) Predict sales on a day with 7 hours of sunshine.

   b) Predict sales on a day with 0 hours of sunshine, and comment on whether this is sensible.

   c) Interpret the gradient, 42, in context.

2. A scientist studying the growth of bacteria finds the regression line of population size y (thousands) on time x (hours, ranging from 0 to 8) to be $y = 5 + 3x$.

   a) Predict the population size after 5 hours.

   b) Explain why using this line to predict the population after 500 hours would be unwise.

3. A company's regression line relating advertising spend x (£1000s, ranging from 1 to 20) to monthly sales y (£1000s) is $y = 30 + 4.5x$. The company spends £0 on advertising one month. Explain why the model's prediction for y in this case may not have a sensible real-world interpretation.

7.3 Pearson's product moment correlation coefficient

The **product moment correlation coefficient**, r, gives a precise numerical measure of the strength and direction of *linear* correlation.

Key point. $-1 \leq r \leq 1$.

r near -1	r near 0	r near $+1$
strong negative linear correlation	little/no linear correlation	strong positive linear correlation

A value of r close to 0 means there is little *linear* correlation – it does not rule out a strong *non-linear* relationship between the variables.

Example (worked, for reference)

A study of 40 cars finds a PMCC of $r = 0.93$ between engine size and top speed, and a PMCC of $r = -0.08$ between the car's colour code (numerical) and top speed.
a) Interpret both values of r.
b) A student claims that because $r = -0.08$ for colour code and speed, colour code has no relationship at all with speed. Comment on this claim.
Solution.
a) $r = 0.93$ is close to $+1$: there is strong positive linear correlation between engine size and top speed – larger engines are associated with higher top speeds. $r = -0.08$ is close to 0: there is little to no *linear* correlation between colour code and top speed.
b) The claim is too strong. $r = -0.08$ only tells us there's little *linear* relationship; it doesn't rule out some other kind of relationship, and in this case colour code is an arbitrary numerical label, so a low r here is expected and doesn't need further interpretation as a real-world effect.

Exercise 7C

1. State whether each value of r indicates strong, moderate, or weak/no correlation, and whether it is positive or negative: **a)** $r = 0.89$ **b)** $r = -0.31$ **c)** $r = 0.04$ **d)** $r = -0.97$.

2. A PE teacher calculates $r = 0.81$ between hours of weekly training and 100m sprint time improvement (in seconds) for a squad of athletes.

   a) Interpret this value of r in context. (Note: a positive r here means more training is associated with a *larger* improvement.)

   b) Explain why this value of r alone does not prove that training *causes* improved sprint times.

3. A dataset of 50 students gives $r = -0.62$ between hours of TV watched per week and exam score.

   a) Describe the strength and direction of this correlation.

   b) A newspaper headline claims "Watching TV lowers exam scores, study proves." Explain what is wrong with this claim.

4. Two variables have a PMCC of $r = 0.02$. Explain what this value does and does not tell you about the relationship between the variables.

Review. After working through this chapter you should:

- be able to draw a line of best fit by eye through the mean point
- understand that correlation does not imply causation, and recognise confounding variables
- be able to use a regression line $y = a + bx$ to make predictions

- be able to distinguish interpolation from extrapolation and comment on reliability
- be able to interpret the PMCC, r, including its limitations