Chapter 7: Correlation and Regression — Mark Scheme

7.1 Lines of best fit — Exercise 7A

1. State the type of correlation expected.

   (a) **Negative correlation.** As outdoor temperature increases, less heating is needed, so the heating bill decreases.

   (b) **Negative correlation.** Among adults, memory test scores tend to decrease as age increases.

   (c) **Positive correlation.** Larger fires require more fire engines and also cause more damage, so both tend to increase together (note: this is a classic case where the relationship is driven by a third factor — the size/severity of the fire — rather than fire engines causing damage).

   (d) **No correlation.** Shoe size and mathematics test score are unrelated in adults.

2. Drownings and ice cream sales.

   (a) The strong positive correlation does not mean eating ice cream causes drowning, because correlation does not imply causation. Both variables could be driven by a third factor, so there is no evidence of a direct causal link.

   (b) **Confounding variable: hot weather (temperature/season).** In hot weather more people buy ice cream *and* more people go swimming (leading to more drownings), so warm weather is a plausible confounder that drives both variables up together.

3. Population, x (thousands), and number of secondary schools, y, for 6 towns:

$$x : 15,\ 22,\ 30,\ 45,\ 60,\ 80 \qquad y : 2,\ 3,\ 4,\ 6,\ 8,\ 10$$

   (a) Mean point: $\bar{x} = \dfrac{15+22+30+45+60+80}{6} = \dfrac{252}{6} = 42$, $\quad \bar{y} = \dfrac{2+3+4+6+8+10}{6} = \dfrac{33}{6} = 5.5$. Plot the six points, mark the mean point (42, 5.5), and draw a single straight line by eye through the mean point, following the upward trend of the data (accept any line that reasonably follows the trend and passes through/near the mean point).

   (b) The data shows **strong positive correlation**: as the population of a town increases, the number of secondary schools also increases, and the points lie close to a straight line.

7.2 The regression line — Exercise 7B

1. Regression line $y = -15 + 42x$, where x = hours of sunshine (data range $2 \le x \le 11$), y = daily ice cream sales in £.

   (a) $y = -15 + 42(7) = -15 + 294 = 279$. Predicted sales are **£279**. Since $x = 7$ lies within the data range 2–11, this is interpolation, so the prediction should be reasonably reliable.

   (b) $y = -15 + 42(0) = -15$. This predicts −**£15 of sales**, which is not sensible (sales cannot be negative). This is extrapolation, since $x = 0$ lies outside the data range 2–11, so the model should not be trusted here — the linear relationship need not hold (or even make sense) beyond the range of the data.

   (c) The gradient means that, according to this model, **each extra hour of sunshine per day is associated with an increase of £42 in daily ice cream sales** (on average).

2. Regression line $y = 5 + 3x$, where x = time in hours (data range $0 \le x \le 8$), y = bacteria population in thousands.

   (a) $y = 5 + 3(5) = 5 + 15 = 20$. Predicted population after 5 hours is **20 thousand** (20 000) bacteria.

(b) $x = 500$ is a very long way outside the data range of 0–8 hours, so this would be extreme **extrapolation** — the model gives $y = 5 + 3(500) = 1505$ (i.e. 1 505 000 bacteria), which is unrealistic. Bacteria cannot keep growing at a constant linear rate indefinitely: growth will eventually slow due to limited food, space, and build-up of waste, so the linear trend seen over 0–8 hours cannot be assumed to continue for 500 hours.

3. Regression line $y = 30 + 4.5x$, where x = advertising spend in £1000s (data range $1 \leq x \leq 20$), y = monthly sales in £1000s. Spending $x = 0$ lies outside the data range used to calculate the line (1–20), so predicting at $x = 0$ is **extrapolation**. The line was fitted only to data where the company was already advertising to some extent, so there is no guarantee the same linear trend continues down to zero advertising; the predicted value $y = 30$ (i.e. £30 000) may not reflect what would really happen with no advertising at all (e.g. other factors sustaining baseline sales are not captured by this model), so the prediction should not be trusted.

7.3 Pearson's product moment correlation coefficient — Exercise 7C

1. Classifying values of r (using the convention: 0.7–1 strong, 0.4–0.7 moderate, 0–0.4 weak/little or no correlation):

   (a) $r = 0.89$: **strong positive correlation**.

   (b) $r = -0.31$: **weak negative correlation** (some negative association, but not strong).

   (c) $r = 0.04$: **very weak/no correlation** (r is close to 0, so little to no linear relationship).

   (d) $r = -0.97$: **strong negative correlation**.

2. $r = 0.81$ between hours of weekly training and 100m sprint time improvement.

   (a) Since $r = 0.81$ is close to $+1$, there is **strong positive correlation**: athletes who train more hours per week tend to show a greater improvement in their 100m sprint time.

   (b) This value of r alone does not prove causation. There could be a confounding variable (e.g. natural athletic ability, motivation, or quality of coaching) that causes both more training and greater improvement, or another explanation for the association — correlation does not imply causation.

3. $r = -0.62$ between hours of TV watched per week and exam score, for 50 students.

   (a) Since -0.62 lies between -0.4 and -0.7, this is a **moderate negative correlation**: students who watch more TV per week tend to have somewhat lower exam scores, though the relationship is not very strong.

   (b) The headline wrongly assumes correlation proves causation. There could be a confounding variable (e.g. students with less parental supervision or less interest in academic work might both watch more TV and score lower), or the causation could run the other way (students who are struggling academically may choose to watch more TV). The value of r alone cannot establish that TV-watching *causes* lower exam scores.

4. $r = 0.02$. This tells us that there is **almost no linear correlation** between the two variables – a straight line is a very poor fit to the data. However, it does **not** tell us that there is no relationship at all between the variables: r only measures *linear* association, so the variables could still have a strong **non-linear** relationship (e.g. quadratic) even though r is close to 0.