Question 1

Worked Solution

Regression line of pay (y, in £) on points (x): $y = 4.5x - 47$.

Part (a): Describe the correlation between points and pay

The gradient is positive ($4.5 > 0$), so as points increase, pay increases.

Positive correlation.

Part (b): Interpretation of the gradient

The gradient of 4.5 represents the rate of change of pay with respect to points. For every additional point awarded, the predicted pay increases by £4.50 (on average).

For every extra point awarded, pay increases by **£4.50** on average.

Part (c): Why the model might not be appropriate for all jobs

Using the model for low point values gives a predicted pay that is unrealistically low or negative. For example, a job with fewer than $\frac{47}{4.5} \approx 10.4$ points would be predicted to have negative pay, which is not meaningful.

Also, the regression line was fitted to a sample of only 8 employees, so extrapolating to jobs with very different point values may not be valid.

For jobs with fewer than about 11 points, the model predicts negative pay, which is not realistic. [Also accept: the sample of 8 may not represent the full range of jobs in the company.]

Question 2

Worked Solution

Part (a): Range of distances corresponding to recorded value 0 for daily mean visibility

The daily mean visibility is recorded to the nearest 100 m. A recorded value of 0 means the true value was rounded down to 0, so it corresponds to a true value anywhere from 0 m up to (but not including) 50 m.

0 to 50 m (i.e. $0 \leq$ visibility < 50 m). [Accept "0 to 500 m" if treating units as metres with rounding to nearest 100, giving 0 to $<$ 50 dm $=$ 0 to 500 m. Accept 0 to 50dm.]

Part (b): Show the circled point (≈ 5300) is an outlier for visibility

From the table: $Q_1 = 1100$ and IQR $= 1600$ for daily mean visibility.

Upper fence $= Q_3 + 1.5 \times$ IQR.

$Q_3 = Q_1 + \text{IQR} = 1100 + 1600 = 2700$.

$$Q_3 + 1.5 \times \text{IQR} = 2700 + 1.5 \times 1600 = 2700 + 2400 = 5100$$

The circled point has visibility $= 5300 > 5100$.

Upper fence $= Q_1 + \text{IQR} + 1.5 \times \text{IQR} = 1100 + 1600 + 1.5(1600) = 5100$.
Since $5300 > 5100$, the point is an outlier. ✓

Part (c): Interpret the correlation

The scatter diagram (Figure 2) shows a negative correlation between daily mean visibility and daily maximum relative humidity.

As the daily maximum relative humidity increases, the daily mean visibility decreases (negative correlation). Higher humidity is associated with lower visibility.

Part (d): Identify the unlabelled x-axis variable in Figure 3

Figure 3 shows daily mean visibility on the y-axis against a variable with values ranging from approximately 0 to 14 on the x-axis, with a positive association and a regression line. From knowledge of the large data set, the variable that takes small non-negative integer or near-integer values in this range and has a positive relationship with visibility is **daily sunshine hours** (sunshine duration). Cloud cover takes values 0–8 (oktas) and is not a quantitative continuous variable; wind speed (knots) is not integer; daily mean temperature for June would not be near 0 in June; so sunshine is the most appropriate match.

Daily sunshine hours (hours of sunshine). The x-axis values (0 to 14) are consistent with sunshine duration in hours, and more sunshine is associated with better visibility.

Question 3

Worked Solution

Part (a): Describe the correlation in Figure 1 (one month)

Figure 1 shows points clustered with no clear linear pattern — the points are scattered without an obvious upward or downward trend.

No correlation (or weak/little correlation) between daily mean pressure and daily mean air temperature for that single month.

Part (b): What Nadine can infer from Figure 2 (all 2015 data)

Figure 2, using all the 2015 Beijing data, shows a clear negative correlation — the points follow a downward trend and a regression line with negative gradient has been fitted.

There is a **negative correlation**: as daily mean pressure (p) increases, daily mean air temperature (t) tends to decrease. This relationship only becomes apparent when data from across the full year is considered.

Part (c): A value of p suitable for interpolation using Figure 2

From knowledge of the large data set, daily mean pressure in Beijing for 2015 ranges approximately from 990 hPa to 1040 hPa. Any value within this range would be interpolation (predicting within the range of the data).

Any value of p in the range **990 to 1040 hPa** (inclusive). For example, $p = 1010$ hPa.

Part (d): Why it is not meaningful to look for a linear relationship between daily mean wind speed (Beaufort) and daily mean air temperature

In the large data set, daily mean wind speed for Beijing is recorded using the **Beaufort scale**, which is a categorical/qualitative scale expressed in descriptive terms (e.g. "calm", "light breeze"). It is not a continuous quantitative variable, so it is not meaningful to fit a linear regression or calculate a linear correlation coefficient with it.

Daily mean wind speed (Beaufort conversion) is a **qualitative (categorical)** variable in the large data set — it is recorded in descriptive categories (Beaufort scale), not as a continuous numerical measurement. Linear correlation and regression require quantitative variables.

Question 4

Worked Solution

Year 12: 84 students. Year 13: 56 students. Total: 140 students. Stratified sample of 40.

Part (a): How to take the stratified sample

Step 1: Label each year group separately — create a numbered list of all 84 Year 12 students and a numbered list of all 56 Year 13 students.

Step 2: Calculate the number to select from each year group in proportion to their size:

$$\text{Year 12:} \quad \frac{84}{140} \times 40 = 24 \quad \text{students}$$

$$\text{Year 13:} \quad \frac{56}{140} \times 40 = 16 \quad \text{students}$$

Step 3: Use random numbers to select 24 students from the Year 12 list and 16 from the Year 13 list.

Label each year group separately. Use random numbers to select **24 Year 12 students** and **16 Year 13 students**.

Part (b): Effect of an extra 0.5 hours of sleep on test performance

The gradient of the regression line is 5.60. This means for every extra hour of sleep, performance increases by 5.60 marks on average. For an extra 0.5 hours:

$$\Delta p = 5.60 \times 0.5 = 2.8 \text{ marks}$$

An extra 0.5 hours of sleep is associated with an **increase of 2.8 marks** in test performance on average.

Part (c): One limitation of the regression model

[Accept any one valid limitation, for example:]

The model predicts that more sleep always leads to better performance, but this cannot hold indefinitely — sleeping for very many hours is unlikely to continue improving performance (the model would predict unrealistically high scores). Alternatively: the model only accounts for sleep and ignores other factors affecting performance (e.g. prior revision, health). Or: for $s = 0$, the model predicts a score of 26.1 marks, but a student who slept 0 hours might not achieve this.

End of Worked Solutions