

End of Topic Test – Chapter 7: Correlation and Regression

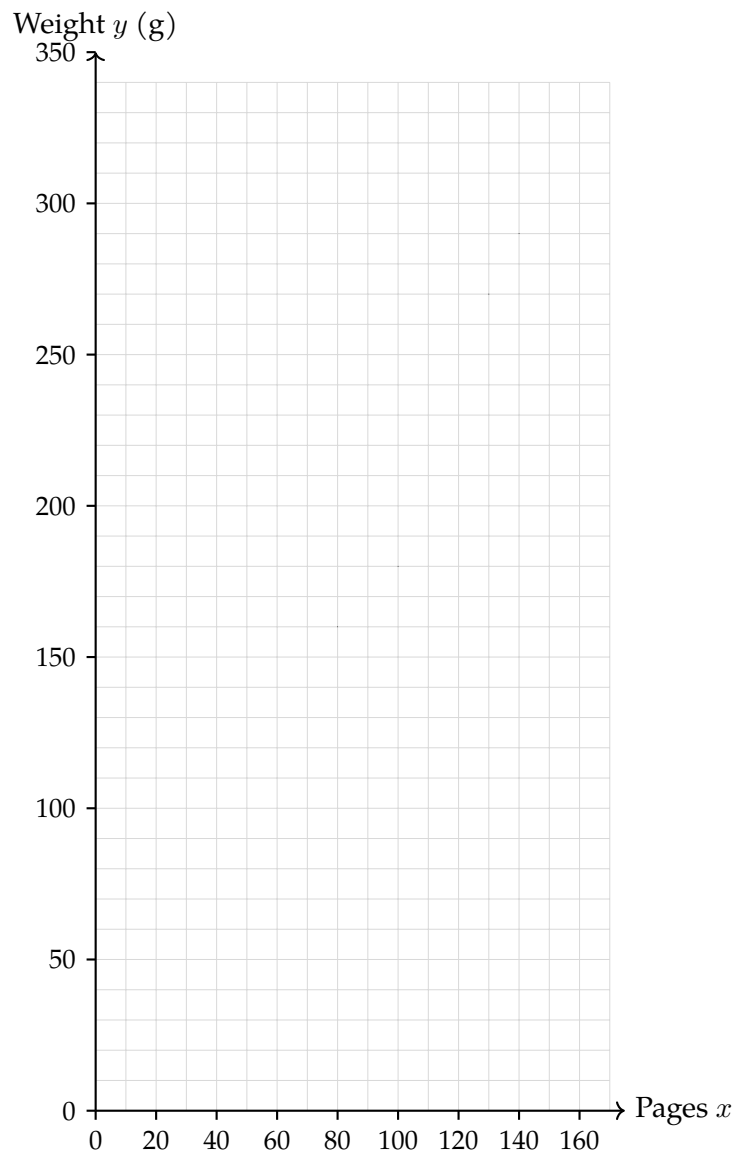
Answer all questions. Use a calculator where appropriate. Show all working and give clear interpretations in context.

1. Number of pages and book weight

The table shows the number of pages x and the weight y (in grams) for a sample of 10 books.

Pages (x)	80	130	100	140	115	90	160	140	105	150
Weight (y in g)	160	270	180	290	230	180	320	270	210	300

- (a) (3 points) On the axes provided, plot the 10 points (x, y) . Do not draw a regression line yet.



- (b) (3 points) Describe the direction and strength of the relationship between number of pages and weight. Give a brief contextual interpretation.

- (c) (6 points) Calculate the product-moment correlation coefficient r for these data. Give your answer to three decimal places and interpret its value in context. Show key steps or calculator output.
- (d) (3 points) Use your regression equation to predict the weight of a book with $x = 125$ pages. State whether this is interpolation or extrapolation and justify your answer.
- (e) (3 points) Give two reasons why the regression line might not predict perfectly for individual books.

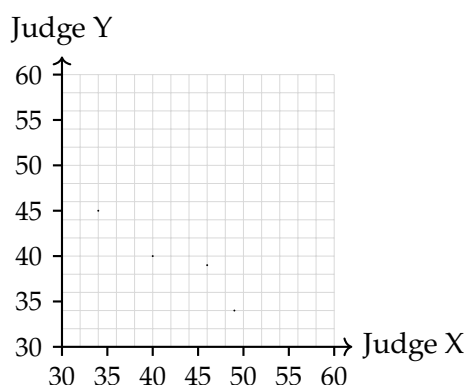
2. Judging consistency in an art competition

Three judges score the same ten paintings. Their marks are shown in the table.

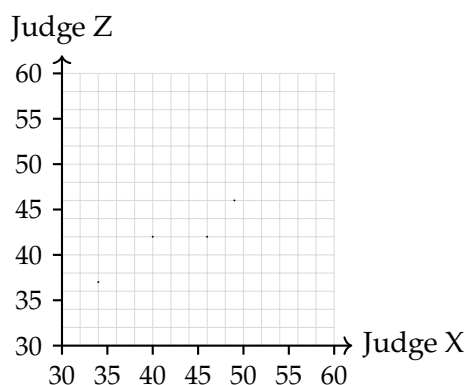
Painting	A	B	C	D	E	F	G	H	I	J
Judge X	40	46	34	49	49	37	41	36	43	38
Judge Y	40	39	45	34	35	38	39	43	40	42
Judge Z	42	42	37	46	45	39	41	37	44	41

After the competition, some artists complain that one judge's marks are not consistent with the others.

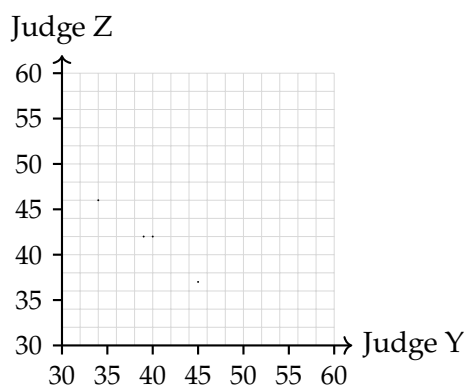
- (a) (3 points) Plot a scatter graph of **Judge X's marks against Judge Y's marks** on the axes provided.



- (b) (3 points) Plot a scatter graph of **Judge X's marks against Judge Z's marks** on the axes provided.



- (c) (3 points) Plot a scatter graph of **Judge Y's marks against Judge Z's marks** on the axes provided.



- (d) (4 points) Calculate the product-moment correlation coefficient between:
- Judge X and Judge Y

- Judge X and Judge Z

Give both answers to three decimal places and comment on the strength of each relationship.

(e) (6 points) Using your results, decide which judge (if any) appears to be marking in a way that is inconsistent with the others. Give a clear explanation using statistical reasoning.

(f) (3 points) State one limitation of using correlation alone to judge fairness in scoring.

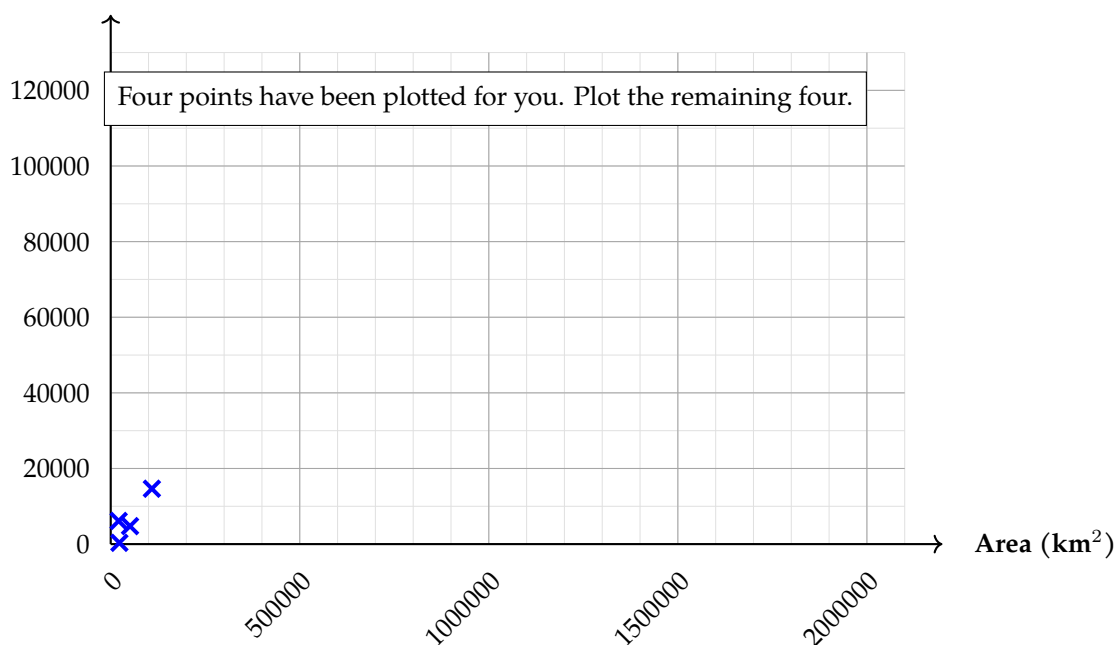
3. Area and population of Central American countries

The table shows the area and population of eight countries in Central America.

Country	Area (km ²)	Population (thousands)
Belize	22966	340.84
Costa Rica	51100	4755.23
El Salvador	21041	6125.51
Guatemala	108889	14647.08
Honduras	112090	8598.56
Mexico	1964375	120286.66
Nicaragua	130370	5848.64
Panama	75420	3608.43

- (a) (4 points) Plot a scatter graph of **area (km²)** on the horizontal axis and **population (thousands)** on the vertical axis. Four points have been plotted for you.

Population (thousands)



- (b) (4 points) Calculate the product-moment correlation coefficient r for the full dataset. Give your answer to three decimal places and comment on the strength and direction of the relationship.
- (c) (5 points) Use your calculator to find the least-squares regression line of population y on area x in the form

$$\hat{y} = a + bx.$$

Give a and b to two decimal places.

- (d) (4 points) Mexico is much larger in both area and population than the other countries. Explain how this point is likely to affect:
- the value of $PMCC$
 - the slope of the regression line
- (e) (3 points) Using your regression line, estimate the population of a country with area 100,000 km². Comment on whether this prediction is likely to be reliable.
- (f) (3 points) Give one limitation of using a simple linear model to predict population from area in this context.

4. Comparing classes: interpretation and residuals

Two teachers fit regression lines of final exam score y on hours revision x :

$$\text{Class A: } \hat{y} = 48 + 2.4x, \quad \text{Class B: } \hat{y} = 52 + 1.9x.$$

- (a) (3 points) Interpret the intercept and slope for each class in context.

- (b) (3 points) Which class shows a greater expected gain in score per extra hour of revision? Explain.

- (c) (4 points) A student in Class A revised for 6 hours and scored 65. Calculate the residual for this student using Class A's regression line. Repeat for Class B (use the same observed score and hours). For which class did the student perform better than expected?

- (d) (5 points) Give two plausible reasons why the two regression lines might differ (consider cohort ability, teaching methods, selection effects, or range of x). For each reason, explain how it could affect the intercept and/or slope.

5. Salary and experience: prediction and model assessment

A careers advisor fits a regression of annual salary y (thousands £) on years' experience x for a sample of employees and obtains

$$\hat{y} = 22.5 + 1.8x, \quad r = 0.76.$$

- (a) (3 points) Interpret the slope and intercept in context.
- (b) (4 points) Use the model to predict the salary for a candidate with 25 years' experience. Comment on whether this prediction is sensible (consider extrapolation and typical industry ranges).
- (c) (3 points) Suggest two additional variables that, if collected and included in a multiple regression model, might improve predictions of salary. Briefly justify each choice.
- (d) (4 points) State two limitations of using a simple linear regression model (one predictor) for making individual salary predictions and suggest how each limitation could be addressed in practice.