Exercises: Regression and Correlation

Exercise 10F

1 Nine students sat a French and a Spanish test. The table gives their results. Find the r-value and describe the correlation between the two sets of scores.

Subject	A	B	C	D	E	F	G	H	I
French	56	56	65	65	50	25	87	44	35
Spanish	87	91	85	91	75	28	92	66	58

2 A social psychologist thinks that there is a correlation between income and education. She found that people with higher income have more years of education. The results of her survey are shown below.

You can also phrase it that people with more years of education have higher income.

Person	A	B	C	D	E	F	G	H	I	J
Income (thousand $)	125	100	40	35	41	29	35	24	50	60
Years of education	19	20	16	16	18	12	14	12	16	17

a Find the r-value.
b What can you say about the strength of the correlation?
c What does the sign of the r-value indicate?

3 Does a car stop more slowly as it gets older? The table below shows the age (in months) of a car and its stopping distance (in metres) from a speed of $40\,\text{km}\,\text{h}^{-1}$.

Age (months)	9	15	24	30	38	46	53	60	64	76
Stopping distance (m)	28.4	29.3	37.6	36.2	36.5	35.3	36.2	44.1	44.8	47.2

a Find the r-value.
b What happens to the stopping distance as the car gets older?
c Describe the strength of the correlation.

4 Kelly has always been told to stop chatting on the computer and focus on her studies. Kelly wants to see if there will be any effect on school grades first, and decides to survey 10 friends. Here are Kelly's results.

GPA	3.1	2.4	2.0	3.8	2.2	3.4	2.9	3.2	3.7	3.5
Chat time (h/week)	14	16	20	7	25	9	15	13	4	14

An A grade is worth 4 points, a B is 3 points, a C is 2 points, a D 1 point, and an F is 0 points. The average grade is called the grade point average, GPA.

a Find the r-value.
b Describe the correlation.
c Based on the survey, would Kelly's grade increase if the chat time decreased?

5 Mo had always been told to stop playing computer games and get on with some work, and so he decided to conduct a survey of 10 friends to see the effect on GPA. The results are in the table below.

GPA	2.7	3.8	1.5	3.6	2.2	3.8	2.0	1.9	2.5	3.0
Game time (h/week)	10	24	25	17	5	26	14	30	22	7

a Find the r-value.
b Describe the correlation.
c Based on the survey, would Mo's grade increase if his game time decreased?

6 Find and interpret the r-value for the leaning tower of Pisa data.

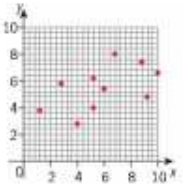
Review exercise

1 Phrases **i**, **ii**, **iii**, **iv**, and **v** represent descriptions of the correlation between two variables:

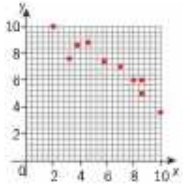
i High positive linear correlation
ii Low positive linear correlation
iii No correlation
iv Low negative linear correlation
v High negative linear correlation

Which phrase best represents the relationship between the two variables shown in each of the scatter diagrams below?

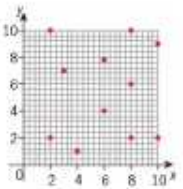
a



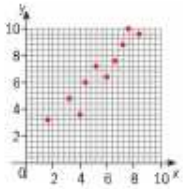
b



c



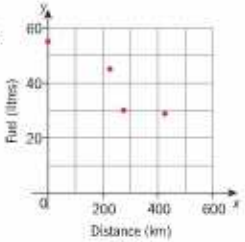
d



EXAM-STYLE QUESTIONS

2 The following table gives the amount of fuel in a car's fuel tank, and the number of kilometres traveled after filling the tank.

Distance traveled (km)	0	220	276	500	680	850
Amount of fuel in tank (litres)	55	43	30	24	10	6



a Copy the scatter diagram and plot the remaining points. The mean distance traveled $\bar{x}$ is 421 km, and the mean amount of fuel in the tank $\bar{y}$ is 28 litres. This point is plotted on the scatter diagram.
b Sketch the line of best fit through the mean point.
c A car traveled 350 km. Use your line of best fit to estimate the amount of fuel left in the tank.

3 This table shows the ages of ten policemen and the time that it took them to run 100 m.

Age	22	23	24	25	32	35	39	45	45	50
Time	10.9	11.1	10.8	12.0	11.2	12.1	12.6	13	12.7	13.6

a Plot the data on a scatter diagram.
b Find the mean age and the mean time.
c Draw a line of best fit through the mean point.
d How long would you expect a 30-year-old policeman to take to run the 100 m?

Review exercise

EXAM-STYLE QUESTIONS

1 The table shows the number of push-ups that David can do each minute for 6 minutes.

Minutes	1	2	3	4	5	6
Push-ups	7	8	5	3	2	2

a Show the points on a scatter diagram along with a line of best fit.
b What happens to the number of push-ups as the time increases?
c Find the equation of the regression line.
d Find the r-value and use it to describe the relationship.

2 The heights and weights of a sample of 11 students are:

Height (m) h	1.36	1.47	1.54	1.56	1.59	1.63	1.66	1.67	1.69	1.74	1.81
Weight (kg) w	52	50	67	62	69	74	59	87	77	73	67

a Write down the regression line of w on h.
b Use the regression line to estimate the weight of someone whose height is 1.6 m.

EXAM-STYLE QUESTIONS

3 A psychologist wants to investigate the relationship between the IQ of a child and the IQ of their mother. She measures the IQ of a sample of 8 children and mothers:

Child's IQ x	87	91	94	98	103	108	111	123
Mother's IQ y	94	96	89	102	98	94	116	117

a Write down the correlation coefficient between x and y.
b Find the regression line of y on x.
c Use the regression line to estimate the IQ of the mother of a child with an IQ of 100.
Using your answer to part **a**, explain how accurate you think this estimate is likely to be.

4 Eight students took two mathematics tests. We would like to know if we could predict the result of Test 2 from Test 1.
The percentage results are given below:

Test 1	54	72	32	68	55	80	45	77
Test 2	31	38	16	34	27	41	22	37

a Plot your results on a scatter diagram.
b Describe the correlation from your diagram.
c Copy and complete the sentence 'Students with a high score on Test 1 tend to have a ..... score on Test 2'.
d Find the equation of the line of best fit.
e If another student achieved a score of 40 in Test 1, what can we predict that this student will get in Test 2?

5 The height of a pot plant was measured for the first 8 weeks from when it was bought:

Week x	0	1	2	3	4	5	6	7	8
Height (cm) y	23.5	25	26.5	27	28.5	31.5	34.5	36	37.5

a Plot these pairs of values on a scatter diagram taking 1cm to represent 1 week on the horizontal axis and 1cm to represent 2cm on the vertical axis.
b Write down the value of the mean point.
c Plot the mean point on the scatter diagram. Name it L.
d **i** Write down the correlation coefficient, r, for these readings.
ii Comment on this result.
e Find the equation of the regression line of y on x.
f Draw the line of regression on the scatter diagram.
g Using your equation estimate the height of the plant after $4\frac{1}{2}$ weeks.
h DJ uses the equation to claim that a plant would be 62.8 cm tall after 30 weeks. Comment on his claim.

EXAM-STYLE QUESTIONS

6 Personality researchers studied the behavior of a group of 10 teenagers. The researchers assessed a personality variable called 'agreeableness'. This is a measure of how nice that person is to be around. The questions ask about how cheerful, stubborn, polite, bossy and cooperative the person is. Each teenager's average of those items is reported.

They also created a measure of behavior problems. The youths reported on various behaviors in the last six months, including cheating, swearing, stealing and fighting. Each teenager's sum of behavior problems is reported.

Participant	Agreeableness factor	Behavior problems
George	4.3	5
Bill	3.0	22
Ronald	3.4	10
Jimmy	3.3	12
Gerald	2.9	23
Laura	4.0	21
Hilary	4.7	2
Nancy	2.4	35
Eleanor	2.9	12
Elizabeth	4.7	4

a Construct a scatter diagram and show the regression line.
b What happens as the agreeableness factor increases?
c Find the r-value.
d Describe the correlation.
e Copy and complete the sentence 'Teenagers who were more agreeable tended to have _____ behavior problems'.
f Write down the equation of the regression line.
g Michelle was absent for the behavior problems questions but scored 4.5 on her agreeableness. Estimate her score for the behavior problems.

7 Each day, a clothing factory recorded the number x of coats it produces and the total production cost y dollars. The results for nine days are shown in the following table.

x	26	44	65	43	50	31	68	46	57
y	400	582	784	625	699	448	870	537	724

a Write down the equation of the regression line of y on x.
Use your regression line as a model to answer the following:
b Interpret the meaning of **i** the gradient **ii** the y-intercept.
c Estimate the cost of producing 70 coats.
d The factory sells the boxes for \$19.99 each. Find the smallest number of coats that the factory should produce in one day in order to make a profit.

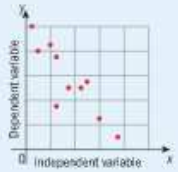
CHAPTER 10 SUMMARY

- Bivariate analysis is concerned with the relationships between pairs of variables (x, y) in a data set.

Scatter diagrams

- **Scatter diagrams** (also called scatter plots) are used to investigate the possible relationship between two variables that both relate to the same 'event'.
- The relationship between two variables is called their **correlation**.
- To draw a scatter diagram plot the (x, y) values from the data table as dots on a graph. The pattern formed by the dots can give us some indication of the correlation.

  The **independent variable** should be on the horizontal axis with the **dependent variable** on the vertical axis.



- A general upward trend in the pattern of dots shows **positive** correlation.
- A general downward trend in the pattern of dots shows **negative** correlation.
- Scattered points with no trend may indicate correlation close to **zero**.
- A correlation between two data sets does not necessarily mean that one causes the other.

The line of best fit

- A **line of best fit** or trend line is drawn on a scatter diagram to find the direction of an association between two variables and to show the trend. This line of best fit can then be used to make predictions.
  - If the line rises from left to right then there is a **positive** correlation.
  - If the line falls from left to right then there is a **negative** correlation.
  - Strong positive and negative correlations have data points very close to the line of best fit.
  - Weak positive and negative correlations have data points that are not clustered near or on the line of best fit.

Continued on next page

- To draw a line of best fit by eye draw a line that will balance the number of points above the line with the number of points below the line. An improvement is to have a reference point: one point for the line to pass through. This is the **mean point** and is calculated by finding the mean of the x-values and the mean of the y-values.

- The equation of the line of best fit, also called the **regression line**, can be used for prediction purposes.

Least squares regression

- A **residual** is the vertical distance between a data point and the graph of a regression equation.
- The least squares regression line is the one that has the smallest possible value for the sum of the squares of the residuals.
- The formula for finding the gradient, or slope (m) of a regression line is

$m = \frac{S_{xy}}{(S_x)^2}$, where

$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$ and $(S_x)^2 = \sum x^2 - \frac{(\sum x)^2}{n}$

- You should use your GDC to find the equation of the regression line in examinations.

Continued on next page

Measuring correlation

- The **Pearson product–moment correlation coefficient** (denoted by r) is a measure of the correlation between two variables X and Y, giving a value between $+1$ and -1 inclusive. It is widely used in the sciences as a measure of the strength of **linear** dependence between two variables.
- The formula for finding Pearson's correlation coefficient is $r = \frac{S_{xy}}{S_x S_y}$

where

$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$, $S_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$ and $S_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$.

- A quick way to interpret the r-value is:

r-value	Correlation
$0 < \lvert r \rvert \le 0.25$	Very weak
$0.25 < \lvert r \rvert \le 0.5$	Weak
$0.5 < \lvert r \rvert \le 0.75$	Moderate
$0.75 < \lvert r \rvert \le 1$	Strong