Stats1 Chapter 4: Correlation

Chapter Practice

Key Points

1 **Bivariate data** is data which has pairs of values for two variables.

2 **Correlation** describes the nature of the linear relationship between two variables.

3 When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.

4 The **regression line** of y on x is written in the form $y = a + bx$.

5 The coefficient b tells you the change in y for each unit change in x.

- If the data is positively correlated, b will be positive.
- If the data is negatively correlated, b will be negative.

6 You should only use the regression line to make predictions for values of the dependent variable that are within the range of the given data.

Chapter Exercises

1 A survey of British towns recorded the number of serious road accidents in a week (x) in each town, together with the number of fast food restaurants (y). The data showed a strong positive correlation. Katie states that this shows that building more fast food restaurants in her town will cause more serious road accidents. Explain whether the data supports Katie's statement.

2 The following table shows the mean CO_2 concentration in the atmosphere, c (ppm), and the increase in average temperature compared to the 30-year period 1951–1980, t (°C).

Year	2015	2013	2011	2009	2007	2005	2003	2001	1999	1997	1995	1994
c (ppm)	401	397	392	387	384	381	376	371	368	363	361	357
t (°C)	0.86	0.65	0.59	0.64	0.65	0.68	0.61	0.54	0.41	0.47	0.45	0.24

Source: Earth System Research Laboratory (CO_2 data); GISS Surface Temperature Analysis, NASA (temperature data)

a Draw a scatter diagram to represent this data.

b Describe the correlation between c and t.

c Interpret your answer to part **b**.

3 The table below shows the packing times for a particular employee for a random sample of orders in a mail order company.

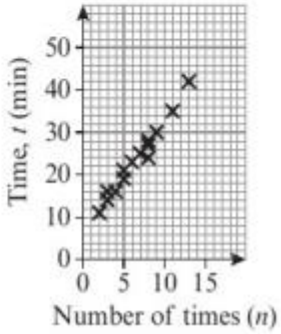
Number of items (n)	2	3	3	4	5	5	6	7	8	8	8	9	11	13
Time (t min)	11	14	16	16	19	21	23	25	24	27	28	30	35	42

A scatter diagram was drawn to represent the data.

a Describe the correlation between number of items packed and time taken. **(1 mark)**

The equation of the regression line of t on n is $t = 6.3 + 2.64n$.

b Give an interpretation of the value 2.64. **(1 mark)**



Chapter Exercises

4 Energy consumption is claimed to be a good predictor of Gross National Product.
An economist recorded the energy consumption (x) and the Gross National Product (y) for eight countries. The data is shown in the table.

Energy consumption (x)	3.4	7.7	12.0	75	58	67	113	131
Gross National Product (y)	55	240	390	1100	1390	1330	1400	1900

The equation of the regression line of y on x is $y = 225 + 12.9x$.

The economist uses this regression equation to estimate the energy consumption of a country with a Gross National Product of 3500.

Give two reasons why this may not be a valid estimate. **(2 marks)**

5 The table shows average monthly temperature, t (°C), and the number of pairs of gloves, g, a shop sells each month.

t (°C)	6	6	50	10	13	16	18	19	16	12	9	7
g	81	58	50	42	19	21	4	2	20	33	58	65

The following statistics were calculated for the data on temperature:
mean = 15.2, standard deviation = 11.4

An outlier is an observation which lies ±2 standard deviations from the mean.

a Show that $t = 50$ is an outlier. **(1 mark)**

b Give a reason whether or not this outlier should be omitted from the data. **(1 mark)**

The equation of the regression line of t on g for the remaining data is $t = 18.4 - 0.18g$.

c Give an interpretation of the value −0.18 in this regression equation. **(1 mark)**

Chapter Exercises

6 James placed different masses (m) on a spring and measured the resulting length of the spring (s) in centimetres. The smallest mass was 20 g and the largest mass was 100 g.

He found the equation of the regression line of s on m to be $s = 44 + 0.2m$.

a Interpret the values 44 and 0.2 in this context. **(2 marks)**

b Explain why it would not be sensible to use the regression equation to work out:

i the value of s when $m = 150$ **ii** the value of m when $s = 60$. **(2 marks)**

7 A student is investigating the relationship between the price (y pence) of 100 g of chocolate and the percentage (x%) of cocoa solids in the chocolate.

The data obtained is shown in the table.

Chocolate brand	x (% cocoa)	y (pence)
A	10	35
B	20	55
C	30	40
D	35	100
E	40	60
F	50	90
G	60	110
H	70	130

a Draw a scatter diagram to represent this data. **(2 marks)**

The equation of the regression line of y on x is $y = 17.0 + 1.54x$.

b Draw the regression line on your diagram. **(2 marks)**

The student believes that one brand of chocolate is overpriced and uses the regression line to suggest a fair price for this brand.

c Suggest, with a reason, which brand is overpriced. **(1 mark)**

d Comment on the validity of the student's method for suggesting a fair price. **(1 mark)**

Chapter Exercises

You will need access to the large data set and spreadsheet software to answer these questions.

1 Investigate the relationship between daily mean windspeed, w, and daily maximum gust, g, in Leeming in 2015.

- **a** Draw a scatter diagram of w against g for the entire data set for Leeming in 2015.
- **b** Describe the correlation shown.
- **c** Comment on whether there is likely to be a causal relationship between mean windspeed and maximum gust.

The equation of the regression line of g on w is given by $g = 4.97 + 2.15w$.

- **d** Use the equation of the regression line to predict the maximum gust on a day when the mean windspeed is:
  i 0.5 knots **ii** 5 knots **iii** 12 knots **iv** 40 knots.
- **e** Comment on the accuracy of each prediction in part **d**.
- **f** Calculate the equation of the regression line of w on g, and use it to predict the mean windspeed on a day when the maximum gust was 30 knots.

2 Use a similar approach to investigate the daily total sunshine and daily mean total cloud cover in Heathrow in 1987.

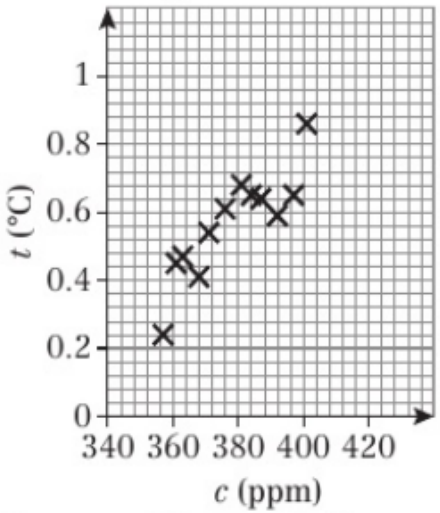
- **a** Use a regression model to suggest values for the missing total sunshine data in the first half of May.
- **b** Do you think there is a causal relationship between these two variables? Give a reason for your answer.

Hint You can use the SLOPE and INTERCEPT functions in some spreadsheets to find the values of a and b in a regression equation.

Chapter Answers

1 The data shows that the number of serious road accidents in a week strongly correlates with the number of fast food restaurants. However, it does not show whether the relationship is causal. Both variables could correlate with a third variable, e.g. the number of roads coming into a town.

2 a



b Strong positive correlation.

c As mean CO_2 concentration in the atmosphere increases, mean global temperatures also increase.

3 a Strong positive correlation.

b If the number of items increases by 1, the time taken increases by approximately 2.64 minutes.

4 (1) 3500 is outside the range of the data (extrapolation).
(2) The regression equation should only be used to predict a value of GNP (y) given energy consumption (x).

5 a Mean + 2SD = 15.2 + 2 × 11.4 = 38; 50 > 38

b The outlier should be omitted as it is very unlikely that the average temperature was 50 °C.

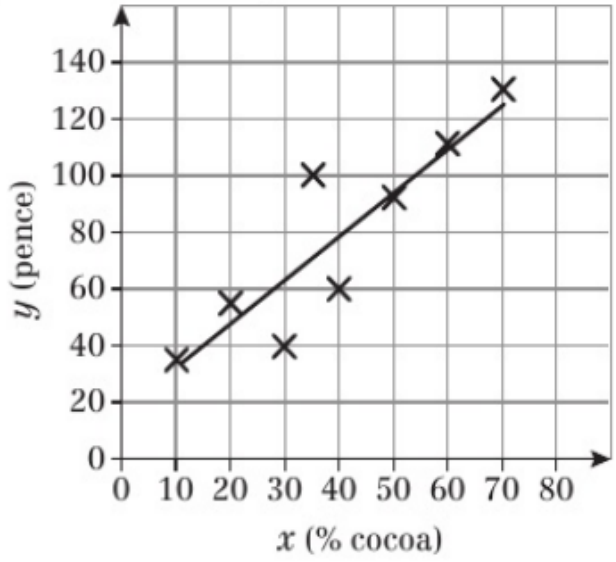
c If the temperature increases by approximately 1 °C, the number of pairs of gloves sold each month decreases by 1.8.

6 a 44 is the length in centimetres of the spring with no mass attached. If a mass of 1 g is attached, the spring would increase in length by approximately 0.2 cm.

b i Outside the range of the data (extrapolation)

ii The regression equation should only be used to predict a value of s given m

7 a & b



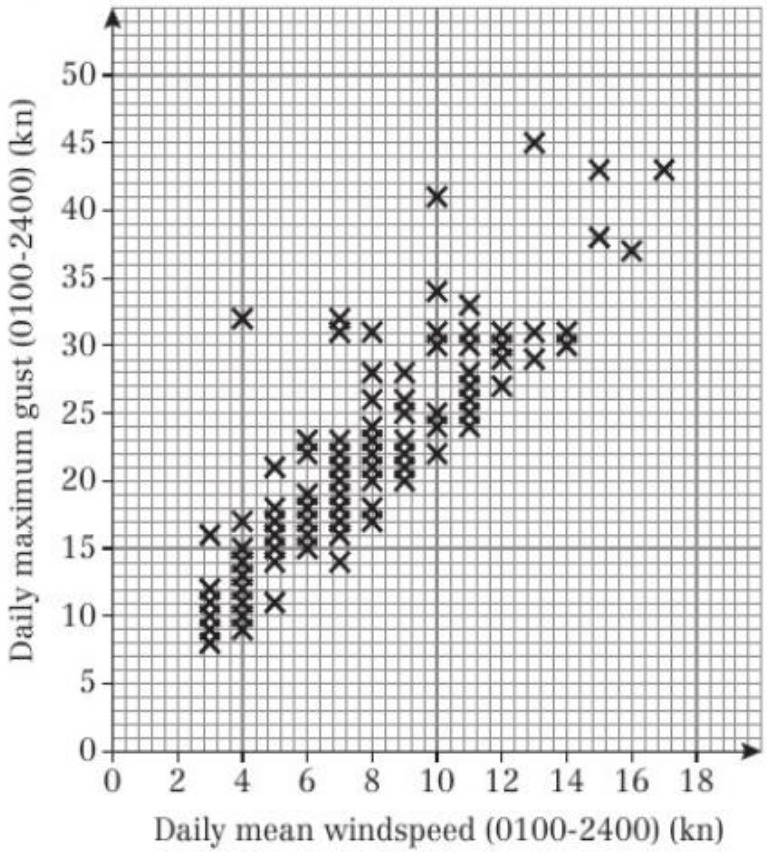
c Brand D is overpriced, since it is a long way above the line.

d The regression equation should be used to predict a value for y given x so the student's method is valid.

Chapter Answers

Large data set

1 a



b Moderate positive correlation.

c The relationship is causal as the maximum gust is related to the mean windspeed.

d **i** 6.05 **ii** 15.7 **iii** 30.8 **iv** 91.0

e Parts **ii** and **iii** are within the range of the data (interpolation) so are more likely to be accurate. Parts **i** and **iv** are outside the range of the data (extrapolation) so are less likely to be accurate.

f $w = 0.053 + 0.35g$; 10.6 knots

2 a Regression equation: $s = 15.1 - 1.8c$
Estimated missing values: 1.8, 8.5, 6.3, 11.6, 3.0, 7.1, 12.2, 14.4, 9.5, 6.3, 2.0, 3.9, 7.2, 3.1, 3.7, 3.9, 0.9

b The relationship is causal because daily sunshine is related to daily mean cloud cover.