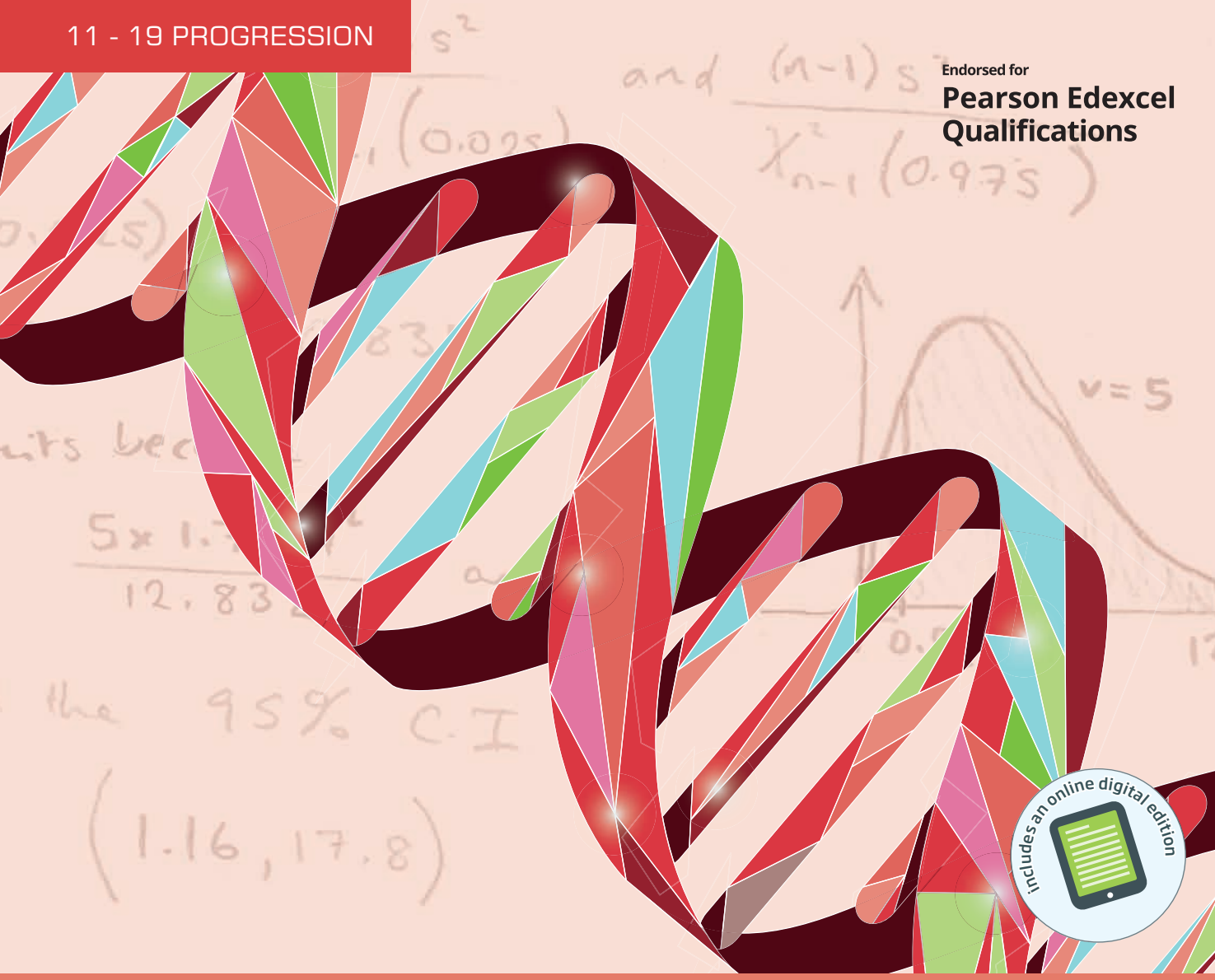
11 - 19 PROGRESSION

Endorsed for
Pearson Edexcel Qualifications



Edexcel AS and A level Further Mathematics

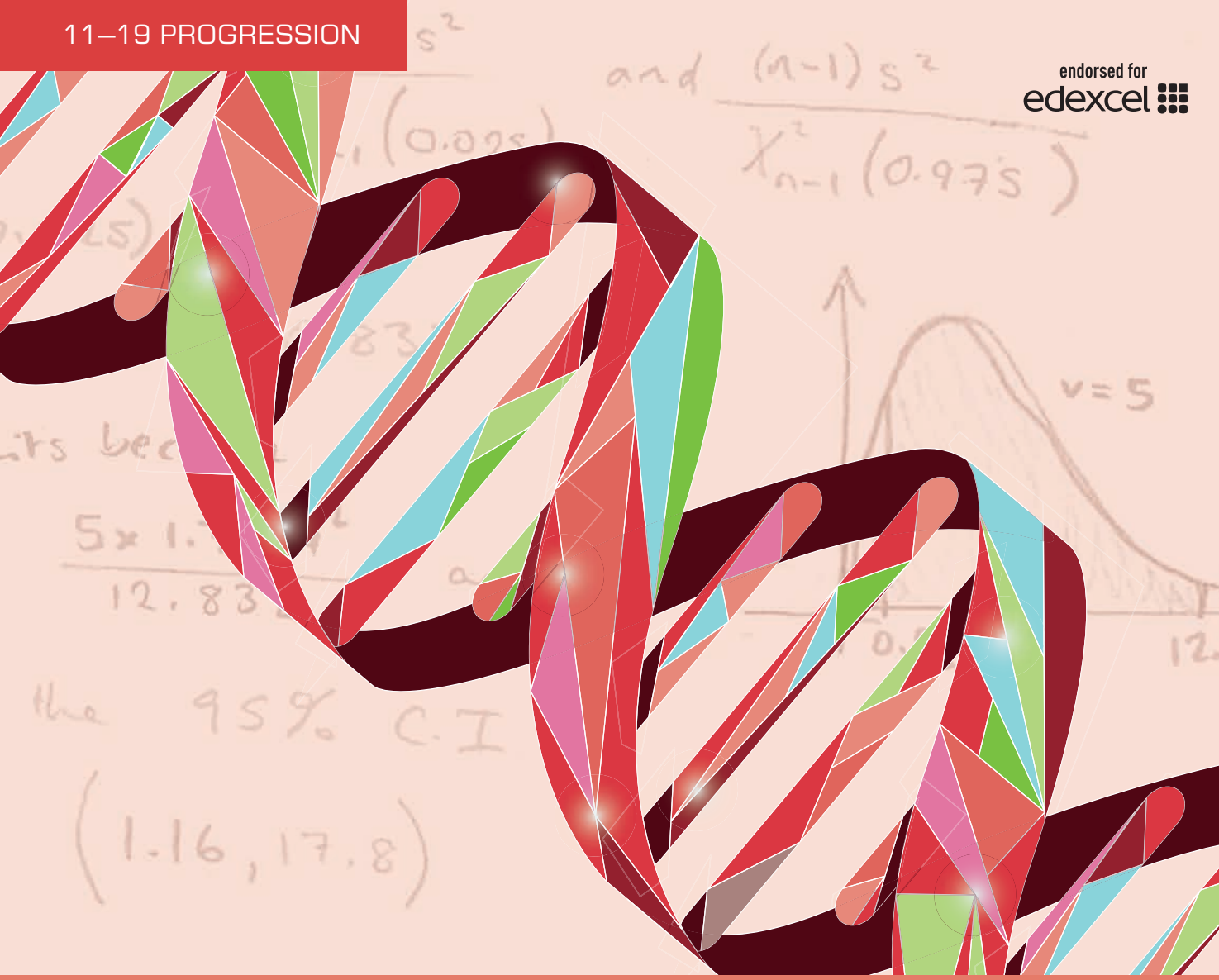
Further Statistics 2

FS2

Pearson

11–19 PROGRESSION

endorsed for edexcel



Edexcel AS and A level Further Mathematics

Further Statistics 2

FS2

Series Editor: Harry Smith

Authors: Greg Attwood, Ian Bettison, Alan Clegg, Gill Dyer, Jane Dyer, John Kinoulty, Keith Pledger, Harry Smith

Pearson

Published by Pearson Education Limited, 80 Strand, London WC2R 0RL.

www.pearsonschoolsandfecolleges.co.uk

Copies of official specifications for all Pearson qualifications may be found on the website: qualifications.pearson.com

First published 2018

21 20 19 18
10 9 8 7 6 5 4 3 2 1

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 978 1 292 18338 1

ISBN 978 1 292 20735 3 (PDF)

Printed in the UK by Bell and Bain Ltd, Glasgow

Acknowledgements
The authors and publisher would like to thank the following for their kind permission to reproduce their photographs:

(Key: b-bottom; c-centre; l-left; r-right; t-top)

123RF: Viparat Kluengsuwanchai 44, 99cr. **Shutterstock:** Gertan 21, 99cl; MPVAN 109, 196b; Acceptphoto 141, 196c. **Alamy Stock Photo:** Vincent Starr Photography/ Cultura Creative (RF) 1, 99l; PCN Photography 92, 99r. **Getty Images:** Ben Queenborough/ Photodisc 163, 196r.

All other images © Pearson Education

A note from the publisher
In order to ensure that this resource offers high-quality support for the associated Pearson qualification, it has been through a review process by the awarding body. This process confirms that this resource fully covers the teaching and learning content of the specification or part of a specification at which it is aimed. It also confirms that it demonstrates an appropriate balance between the development of subject skills, knowledge and understanding, in addition to preparation for assessment.

Endorsement does not cover any guidance on assessment activities or processes (e.g. practice questions or advice on how to answer assessment questions), included in the resource nor does it prescribe any particular approach to the teaching or delivery of a related course.

While the publishers have made every attempt to ensure that advice on the qualification and its assessment is accurate, the official specification and associated assessment guidance materials are the only authoritative source of information and should always be referred to for definitive guidance.

Pearson examiners have not contributed to any sections in this resource relevant to examination papers for which they have responsibility.

Examiners will not use endorsed resources as a source of material for any assessment set by Pearson.

Endorsement of a resource does not mean that the resource is required to achieve this Pearson qualification, nor does it mean that it is the only suitable material available to support the qualification, and any resource lists produced by the awarding body shall include this and other appropriate resources.

Contents

● = A level only

Overarching themes iv
Extra online content vi

1 Linear regression 1
1.1 Least squares linear regression 2
1.2 Residuals 10
Mixed exercise 1 16

2 Correlation 21
2.1 The product moment correlation coefficient 22
2.2 Spearman's rank correlation coefficient 26
2.3 Hypothesis testing for zero correlation 33
Mixed exercise 2 38

3 Continuous distributions 44
3.1 Continuous random variables 45
3.2 The cumulative distribution function 51
3.3 Mean and variance of a continuous distribution 56
3.4 Mode, median, percentiles and skewness 63
3.5 The continuous uniform distribution 71
3.6 Modelling with the continuous uniform distribution 79
Mixed exercise 3 82

● **4 Combinations of random variables** 92
● **4.1** Combinations of random variables 93

Review exercise 1 99

● **5 Estimation, confidence intervals and tests using a normal distribution** 109
● **5.1** Estimators, bias and standard error 110
● **5.2** Confidence intervals 120
● **5.3** Hypothesis testing for the difference between means 127
● **5.4** Use of large sample results for an unknown population 132
Mixed exercise 5 135

● **6 Further hypothesis tests** 141
● **6.1** Variance of a normal distribution 142
● **6.2** Hypothesis testing for the variance of a normal distribution 146
● **6.3** The F-distribution 149
● **6.4** The F-test 155
Mixed exercise 6 159

● **7 Confidence intervals and tests using the t-distribution** 163
● **7.1** Mean of a normal distribution with unknown variance 164
● **7.2** Hypothesis test for the mean of a normal distribution with unknown variance 170
● **7.3** The paired t-test 174
● **7.4** Difference between means of two independent normal distributions 180
● **7.5** Hypothesis test for the difference between means 185
Mixed exercise 7 189

● **Review exercise 2** 196

Exam-style practice paper (AS level) 204

● **Exam-style practice paper (A level)** 206

Appendix 209
Binomial cumulative distribution function 209
Percentage points of the normal distribution 214
Percentage points of the χ^2 distribution 215
Critical values for correlation coefficients 216
Percentage points of Student's t-distribution 217
Percentage points of the F-distribution 218
Poisson cumulative distribution function 219
Formulae 220

Answers 225

Index 256

Overarching themes

The following three overarching themes have been fully integrated throughout the Pearson Edexcel AS and A level Mathematics series, so they can be applied alongside your learning and practice.

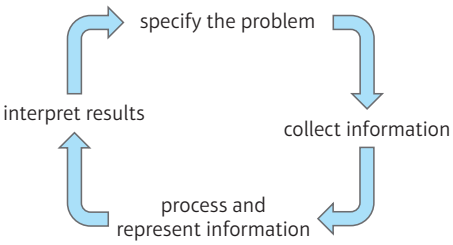
1. Mathematical argument, language and proof

- Rigorous and consistent approach throughout
- Notation boxes explain key mathematical language and symbols
- Dedicated sections on mathematical proof explain key principles and strategies
- Opportunities to critique arguments and justify methods

2. Mathematical problem solving

- Hundreds of problem-solving questions, fully integrated into the main exercises
- Problem-solving boxes provide tips and strategies
- Structured and unstructured questions to build confidence
- Challenge boxes provide extra stretch

The Mathematical Problem-solving cycle



3. Mathematical modelling

- Dedicated modelling sections in relevant topics provide plenty of practice where you need it
- Examples and exercises include qualitative questions that allow you to interpret answers in the context of the model
- Dedicated chapter in Statistics & Mechanics Year 1/AS explains the principles of modelling in mechanics

Finding your way around the book

Access an online digital edition using the code at the front of the book.

Each chapter starts with a list of objectives

The real world applications of the maths you are about to learn are highlighted at the start of the chapter with links to relevant questions in the chapter

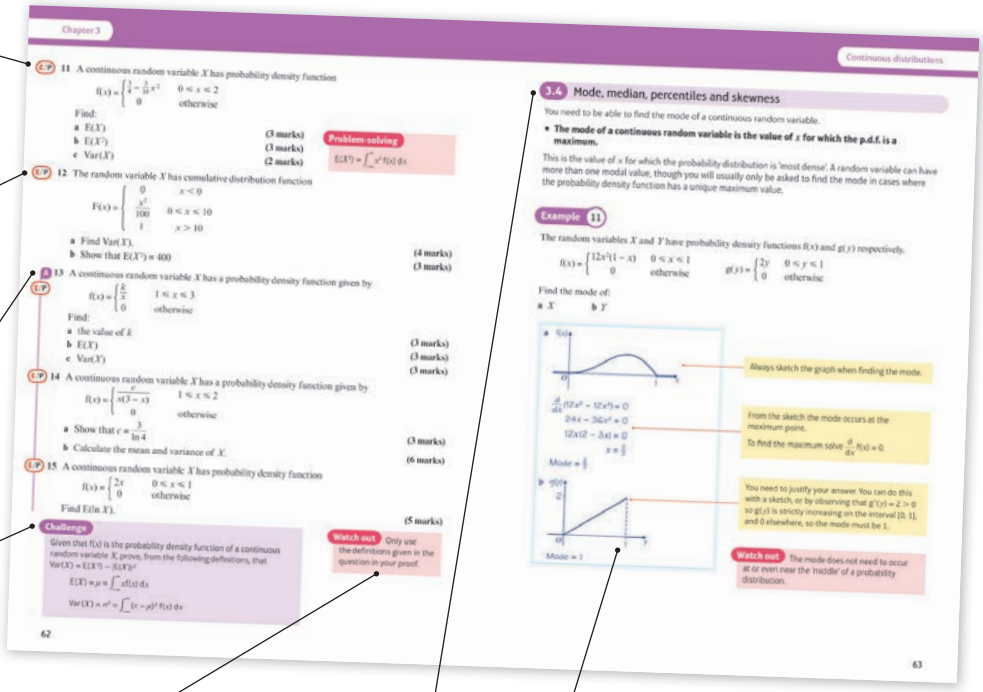
The *Prior knowledge check* helps make sure you are ready to start the chapter

Exercise questions are carefully graded so they increase in difficulty and gradually bring you up to exam standard

Exercises are packed with exam-style questions to ensure you are ready for the exams

A level content is clearly flagged

Challenge boxes give you a chance to tackle some more difficult questions



Exam-style questions are flagged with (E)

Problem-solving questions are flagged with (P)

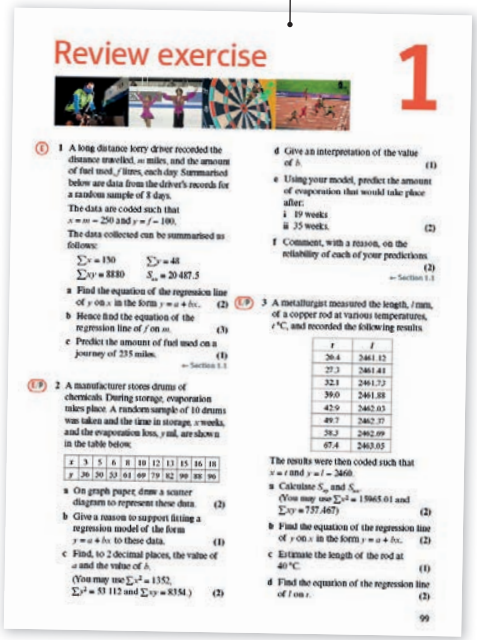
Problem-solving boxes provide hints, tips and strategies, and *Watch out* boxes highlight areas where students often lose marks in their exams

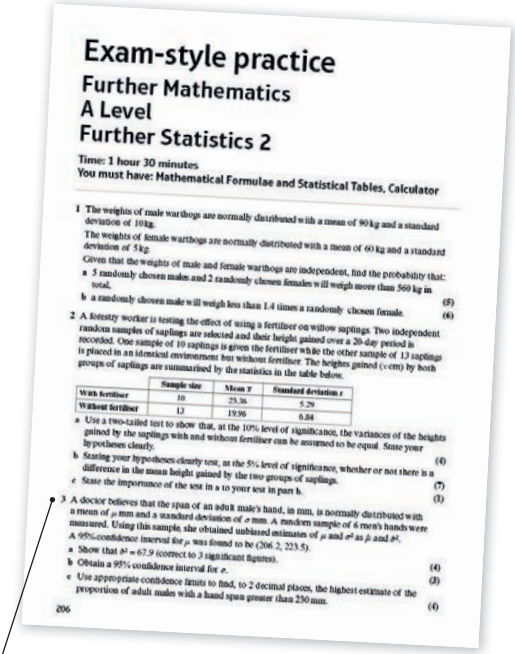
Each section begins with explanation and key learning points

Step-by-step worked examples focus on the key types of questions you'll need to tackle

Each chapter ends with a *Mixed exercise* and a *Summary of key points*

Every few chapters a *Review exercise* helps you consolidate your learning with lots of exam-style questions





AS and A level practice papers at the back of the book help you prepare for the real thing.

Extra online content

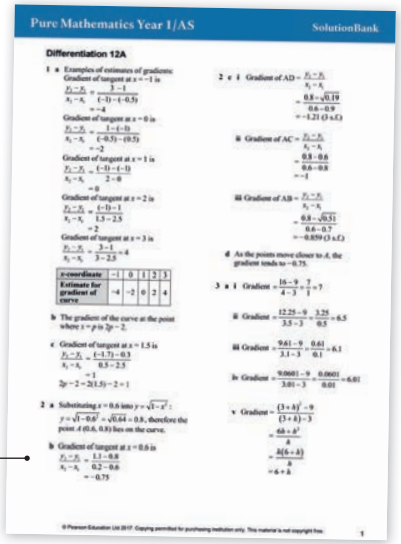
Whenever you see an *Online* box, it means that there is extra online content available to support you.

SolutionBank

SolutionBank provides a full worked solution for every question in the book.

Online Full worked solutions are available in SolutionBank.

Download all the solutions as a PDF or quickly find the solution you need online


Pure Mathematics Year 1/AS SolutionBank

Differentiation 12A

Use of technology

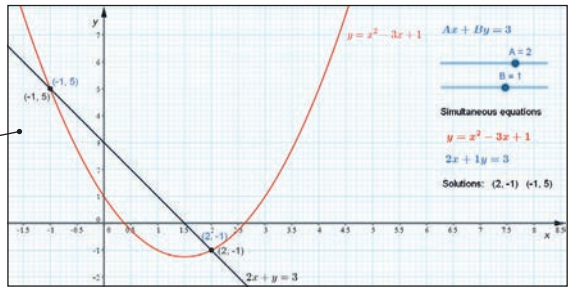
Explore topics in more detail, visualise problems and consolidate your understanding using pre-made GeoGebra activities.

Online Find the point of intersection graphically using technology.

GeoGebra

GeoGebra-powered interactives

Interact with the maths you are learning using GeoGebra's easy-to-use tools



Access all the extra online content for free at:

www.pearsonschools.co.uk/fs2maths

You can also access the extra online content by scanning this QR code:

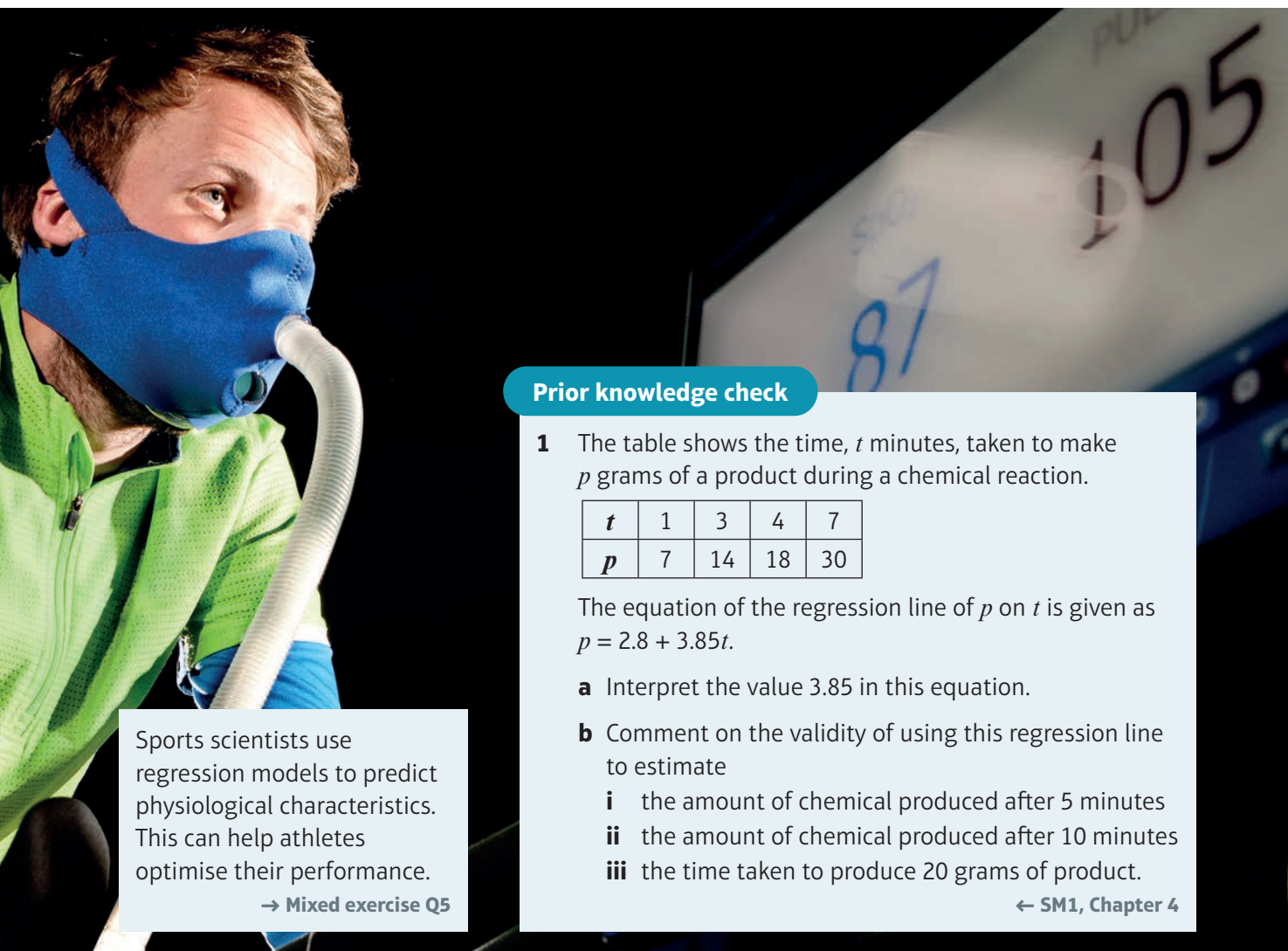
Linear regression

1

Objectives

After completing this chapter, you should be able to:

- Calculate the equation of a regression line using raw data or summary statistics → pages 2-8
- Use coding to find the equation of a regression line → pages 8-10
- Calculate residuals and use them to test for linear fit and identify outliers → pages 10-15
- Calculate the residual sum of squares (RSS) → pages 13-15



Sports scientists use regression models to predict physiological characteristics. This can help athletes optimise their performance.

→ Mixed exercise Q5

Prior knowledge check

1 The table shows the time, t minutes, taken to make p grams of a product during a chemical reaction.

t	1	3	4	7
p	7	14	18	30

The equation of the regression line of p on t is given as $p = 2.8 + 3.85t$.

a Interpret the value 3.85 in this equation.

b Comment on the validity of using this regression line to estimate

i the amount of chemical produced after 5 minutes

ii the amount of chemical produced after 10 minutes

iii the time taken to produce 20 grams of product.

← SM1, Chapter 4

1.1 Least squares linear regression

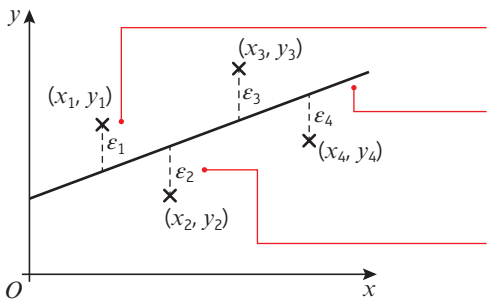
When you are analysing bivariate data, you can use a **least squares regression line** to predict values of the dependent (response) variable for given values of the independent (explanatory) variable. If the response variable is y and the explanatory variable is x, you should use the regression line of $\boldsymbol{y}$ **on** $\boldsymbol{x}$, which can be written in the form $y = a + bx$.

Links 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. This is called **interpolation**. Making predictions for values outside of the range of the given data is called **extrapolation** and produces a less reliable prediction. ← SM1, Section 4.2

The least squares regression line is the line that minimises the **sum of the squares of the residuals** of each data point.

- **The residual of a given data point is the difference between the observed value of the dependent variable and the predicted value of the dependent variable.**

Notation The Greek letter epsilon (ε) is sometimes used to denote a residual.



ε_1 is the residual of the data point (x_1, y_1)

The least squares regression line of y on x is the straight line that minimises the value of $\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_2^4$. In general, if each data point has residual ε_i, the regression line minimises the value of $\sum \varepsilon_i^2$.

The observed value of the dependent variable, y_2, is **less** than the predicted value, so the residual of (x_2, y_2) will be **negative**.

You need to be able to find the equation of a **least squares regression line** using raw data or summary statistics.

- **The equation of the regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ is:**

$$\boldsymbol{y = a + bx}$$

where $\boldsymbol{b = \frac{S_{xy}}{S_{xx}}}$ **and** $\boldsymbol{a = \overline{y} - b\overline{x}}$

Watch out You can calculate a and b directly from raw data using your calculator. However, you might be given summary statistics in the exam so you need to be familiar with these formulae.

S_{xy} and S_{xx} are known as **summary statistics** and you can calculate them using the following formulae:

- $\boldsymbol{S_{xy} = \sum xy - \frac{\sum x \sum y}{n}}$

$$\boldsymbol{S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\boldsymbol{S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}}$$

Example 1

The results from an experiment in which different masses were placed on a spring and the resulting length of the spring measured, are shown below.

Mass, x (kg)	20	40	60	80	100
Length, y (cm)	48	55.1	56.3	61.2	68

a Calculate S_{xx} and S_{xy}.
(You may use $\sum x = 300$ $\sum x^2 = 22\,000$ $\overline{x} = 60$ $\sum xy = 18\,238$ $\sum y^2 = 16\,879.14$ $\sum y = 288.6$ $\overline{y} = 57.72$)

b Calculate the regression line of y on x.

c Use your equation to predict the length of the spring when the applied mass is:
i 58 kg
ii 130 kg

d Comment on the reliability of your predictions.

Online Explore the calculation of a least squares regression line using GeoGebra.

a $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$

$= 22\,000 - \frac{300^2}{5}$

$= 4000$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

$= 18\,238 - \frac{300 \times 288.6}{5}$

$= 922$

Use the standard formulae to calculate S_{xx} and S_{xy}. Write down any formulae you are using before you substitute.

b $b = \frac{S_{xy}}{S_{xx}} = \frac{922}{4000} = 0.2305$

$a = \overline{y} - b\overline{x}$

$= 57.72 - 0.2305 \times 60$

$= 43.89$

Use the formulae to calculate a and b. If you want to check your answer using your calculator, make sure you use the correct mode for linear regression with bivariate data. On some calculators this mode is labelled $\boldsymbol{y = a + bx}$.

$y = 43.89 + 0.2305x$

Remember to write the equation at the end. The numbers should be given to a suitable degree of accuracy.

c i $y = 43.89 + 0.2305 \times 58$

$= 57.3$ cm (3 s.f.)

ii $y = 43.89 + 0.2305 \times 130$

$= 73.9$ cm (3 s.f.)

Substitute the given values into the equation of the regression line.

d Assuming the model is reasonable, the prediction when the mass is 58 kg is reliable since this is within the range of the data.

This is called **interpolation**.

The prediction when the mass is 130 kg is less reliable since this is outside the range of the data.

This is called **extrapolation**.

Example 2

A scientist working in agricultural research believes that there is a linear relationship between the amount of a certain food supplement given to hens and the hardness of the shells of the eggs they lay. As an experiment, controlled quantities of the supplement were added to the hens' normal diet for a period of two weeks and the hardness of the shells of the eggs laid at the end of this period was then measured on a scale from 1 to 10, with the following results:

Food supplement, f (g/day)	2	4	6	8	10	12	14
Hardness of shells, h	3.2	5.2	5.5	6.4	7.2	8.5	9.8

a Find the equation of the regression line of h on f.

(You may use $\sum f = 56$ $\sum h = 45.8$ $\bar{f} = 8$ $\bar{h} = 6.543$ $\sum f^2 = 560$ $\sum fh = 422.6$)

b Interpret what the values of a and b tell you.

a $S_{fh} = \sum fh - \dfrac{\sum f \sum h}{n}$

$= 422.6 - \dfrac{56 \times 45.8}{7} = 56.2$

$S_{ff} = \sum f^2 - \dfrac{(\sum f)^2}{n}$

$= 560 - \dfrac{56^2}{7} = 112$

$b = \dfrac{S_{fh}}{S_{ff}} = \dfrac{56.2}{112}$

$= 0.5017\ldots$ hardness units per g per day

$a = \bar{h} - b\bar{f}$

$= 6.543 - 0.5017\ldots \times 8$

$= 2.5287\ldots$ hardness units

$h = 2.53 + 0.502f$

b a estimates the shell strength when no supplement is given (i.e. when $f = 0$). Zero is only just outside the range of f so it is reasonable to use this value.

b estimates the rate at which the hardness increases with increased food supplement; in this case for every extra one gram of food supplement per day the hardness increases by 0.502 (3 s.f.) hardness units.

Watch out The variables given might not be x and y. Be careful that you use the correct values when you substitute into the formulae. It can sometimes help to write x next to the explanatory variable in the table (f) and y next to the response variable (h).

When dealing with a real problem do not forget to put the units of measurement for the two constants.

Make sure you give your answer in the context of the question. Don't just say that one value increases as the other increases – you need to comment on the **rate** of increase of hardness.

Example 3

A repair workshop finds it is having a problem with a pressure gauge it uses. It decides to have it checked by a specialist firm. The following data were obtained.

Gauge reading, x (bars)	1.0	1.4	1.8	2.2	2.6	3.0	3.4	3.8
Correct reading, y (bars)	0.96	1.33	1.75	2.14	2.58	2.97	3.38	3.75

(You may use $\sum x = 19.2$ $\sum x^2 = 52.8$ $\sum y = 18.86$ $\sum y^2 = 51.30$ $\sum xy = 52.04$)

a Show that $S_{xy} = 6.776$ and find S_{xx}.

It is thought that a linear relationship of the form $y = a + bx$ could be used to describe these data.

b Use linear regression to find the values of a and b giving your answers to 3 significant figures.

c Draw a scatter diagram to represent these data and draw the regression line on your diagram.

d The gauge shows a reading of 2 bars. Using the regression equation, work out what the correct reading should be.

a $S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n}$

$= 52.04 - \dfrac{19.2 \times 18.86}{8} = 6.776$

$S_{xx} = \sum x^2 - \dfrac{(\sum x)^2}{n}$

$= 52.8 - \dfrac{(19.2)^2}{8} = 6.72$

Quote the formula you are going to use before substituting the values.

b $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{6.776}{6.72} = 1.0083\ldots$

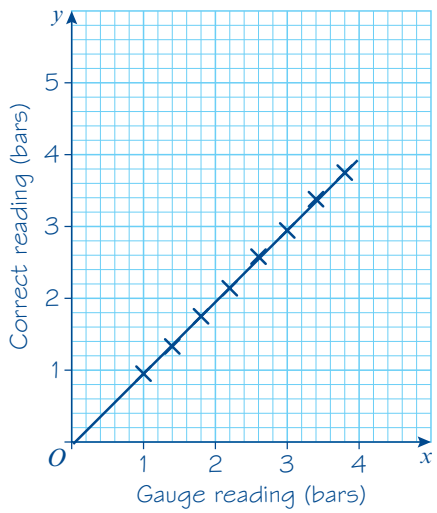
$a = \overline{y} - b\overline{x} = \dfrac{18.86}{8} - 1.0083\ldots \times \dfrac{19.2}{8}$

$= -0.0625$

Regression line is: $y = -0.0625 + 1.008x$

or $y = 1.008x - 0.0625$

c



To draw the regression line either plot the point $(0, a)$ and use the gradient or find two points on the line.

In this case using $x = 1$ gives $y = 0.95$ and using $x = 3$ gives $y = 2.96$.

d $y = (1.008 \times 2) - 0.0625$

$y = 1.95$ bar (3 s.f.)

Exercise 1A

1 The equation of a regression line in the form $y = a + bx$ is to be found. Given that $S_{xx} = 15$, $S_{xy} = 90$, $\overline{x} = 3$ and $\overline{y} = 15$, work out the values of a and b.

2 Given that $S_{xx} = 30$, $S_{xy} = 165$, $\overline{x} = 4$ and $\overline{y} = 8$, find the equation of the regression line of y on x.

3 The equation of a regression line is to be found. The following summary data is given:

$S_{xx} = 40$ $S_{xy} = 80$ $\overline{x} = 6$ $\overline{y} = 12$

Find the equation of the regression line in the form $y = a + bx$.

4 Data is collected and summarised as follows:

$\sum x = 10$ $\sum x^2 = 30$ $\sum y = 48$ $\sum xy = 140$ $n = 4$

a Work out $\overline{x}$, $\overline{y}$, S_{xx} and S_{xy}.

b Find the equation of the regression line of y on x in the form $y = a + bx$.

5 For the data in the table,

x	2	4	5	8	10
y	3	7	8	13	17

Hint Check your answer using the statistical functions on your calculator.

a calculate S_{xx} and S_{xy}

b find the equation of the regression line of y on x in the form $y = a + bx$.

(P) **6** Research was done to see if there is a relationship between finger dexterity and the ability to do work on a production line. The data is shown in the table.

Dexterity score, x	2.5	3	3.5	4	5	5	5.5	6.5	7	8
Productivity, y	80	130	100	220	190	210	270	290	350	400

The equation of the regression line for these data is $y = -59 + 57x$.

a Use the equation to estimate the productivity of someone with a dexterity of 6.

b Give an interpretation of the value of 57 in the equation of the regression line.

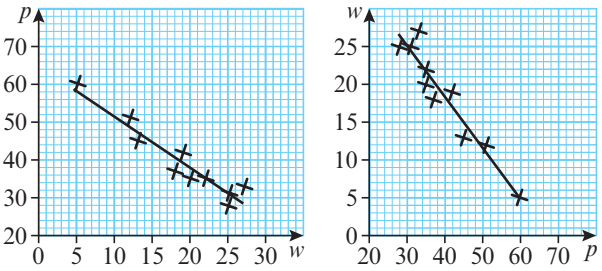
c State, giving in each case a reason, whether or not it would be reasonable to use this equation to work out the productivity of someone with dexterity of:

i 2 **ii** 14

7 A field was divided into 12 plots of equal area. Each plot was fertilised with a different amount of fertilizer (h). The yield of grain (g) was measured for each plot. Find the equation of the regression line of g on h in the form $g = a + bh$ given the following summary data.

$\sum h = 22.09$ $\sum g = 49.7$ $\sum h^2 = 45.04$ $\sum g^2 = 244.83$ $\sum hg = 97.778$ $n = 12$

(P) **8** Research was done to see if there was a relationship between the number of hours in the working week (w) and productivity (p). The data are shown in the two scatter graphs below.



(You may use $\sum p = 397$ $\sum p^2 = 16\,643$ $\sum w = 186$ $\sum w^2 = 3886$ $\sum pw = 6797$)

a Calculate the equation of the regression line of p on w, giving your answer in the form $p = a - bw$.

b Rearrange this equation into the form $w = c + dp$.

The equation of the regression line of w on p is $w = 45.0 - 0.666p$.

c Comment on the fact that your answer to part **b** is different to this equation.

d Which equation should you use to predict:

i the productivity for a 23-hour working week

ii the number of hours in a working week that achieves a productivity score of 40.

(P) **9** In a chemistry experiment, the mass of chemical produced, y and the temperature, x are recorded.

x (°C)	100	110	120	130	140	150	160	170	180	190	200
y (mg)	34	39	41	45	48	47	41	35	26	15	3

Maya thinks that the data can be modelled using a linear regression line.

a Calculate the equation of the regression line of y on x. Give your answer in the form $y = a + bx$.

b Draw a scatter graph for these data.

c Comment on the validity of Maya's model.

(E/P) **10** An accountant monitors the number of items produced per month by a company (n) together with the total production costs (p). The table shows these data.

Number of items, n (1000s)	21	39	48	24	72	75	15	35	62	81	12	56
Production costs, p (£1000s)	40	58	67	45	89	96	37	53	83	102	35	75

(You may use $\sum n = 540$ $\sum n^2 = 30\,786$
$\sum p = 780$ $\sum p^2 = 56\,936$
$\sum np = 41\,444$)

Watch out The numbers of items are given in 1000s. Be careful to choose the correct value to substitute into your regression equation.

a Calculate S_{nn} and S_{np}. **(2 marks)**

b Find the equation of the regression line of p on n in the form $p = a + bn$. **(3 marks)**

c Use your equation to estimate the production costs of 40 000 items. **(2 marks)**

d Comment on the reliability of your estimate. **(1 mark)**

E/P **11** A printing company produces leaflets for different advertisers. The number of leaflets, n, measured in 100s and printing costs £p are recorded for a random sample of 10 advertisers. The table shows these data.

n (100s)	1	3	4	6	8	12	15	18	20	25
p (pounds)	22.5	27.5	30	35	40	50	57.5	65	70	82.5

(You may use $\sum n = 112$ $\sum n^2 = 1844$ $\sum p = 480$ $\sum p^2 = 26\,725$ $\sum np = 6850$)

a Calculate S_{nn} and S_{np}. **(2 marks)**

b Find the equation of the regression line of p on n in the form $p = a + bn$. **(3 marks)**

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

An advertiser is planning to print t hundred leaflets. A rival printing company charges 5p per leaflet.

d Find the range of values of t for which the first printing company is cheaper than the rival. **(2 marks)**

E/P **12** The relationship between the number of coats of paint applied to a boat and the resulting weather resistance was tested in a laboratory. The data collected are shown in the table.

Coats of paint, x	1	2	3	4	5
Protection, y (years)	1.4	2.9	4.1	5.8	7.2

a Use your calculator to find an equation of the regression line of y on x as a model for these results, giving your answer in the form $y = a + bx$. **(2 marks)**

b Interpret the value b in your model. **(1 mark)**

c Explain why this model would not be suitable for predicting the number of coats of paint that had been applied to a boat that had remained weather resistant for 7 years. **(1 mark)**

d Use your answer to part **a** to predict the number of years of protection when 7 coats of paint are applied. **(2 marks)**

In order to improve the reliability of its results, the laboratory made two further observations:

Coats of paint, x	6	8
Protection, y (years)	8.2	9.9

e Using all 7 data points:

i produce a refined model

ii use your new model to predict the number of years of protection when 7 coats of paint are applied

iii give two reasons why your new prediction might be more accurate than your original prediction. **(5 marks)**

Sometimes the original data is coded to make it easier to manage. You can calculate the equation of the original regression line from the coded one by substituting the coding formula into the equation of the coded regression line.

Example 4

Eight samples of carbon steel were produced with a different percentages, $c\%$, of carbon in them. Each sample was heated in a furnace until it melted and the temperature, m in °C, at which it melted was recorded.

The results were coded such that $x = 10c$ and $y = \dfrac{m - 700}{5}$

The coded results are shown in the table.

Percentage of carbon, x	1	2	3	4	5	6	7	8
Melting point, y	35	28	24	16	15	12	8	6

a Calculate S_{xy} and S_{xx}.
(You may use $\sum x^2 = 204$ and $\sum xy = 478$.)

b Find the regression line of y on x.

c Estimate the melting point of carbon steel which contains 0.25% carbon.

a $S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n}$

$= 478 - \dfrac{36 \times 144}{8} = -170$

$S_{xx} = \sum x^2 - \dfrac{(\sum x)^2}{n} = 204 - \dfrac{36^2}{8} = 42$

$\sum x = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36$
$\sum y = 35 + 28 + 24 + 16 + 15 + 12 + 8 + 6 = 144$

b $b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{-170}{42} = -4.047\ldots$

$a = \overline{y} - b\overline{x}$

$= \dfrac{144}{8} + 4.047\ldots \times \dfrac{36}{8} = 36.214\ldots$

$\overline{y} = \dfrac{\sum y}{n}$ and $\overline{x} = \dfrac{\sum x}{n}$

$y = 36.2 - 4.05x$

c Method 1

If $c = 0.25$, then $x = 10 \times 0.25 = 2.5$

$y = 36.214\ldots - 4.047\ldots \times 2.5 = 26.095\ldots$

$y = \dfrac{m - 700}{5}$

$m = 5y + 700$

$= 5 \times 26.095\ldots + 700 = 830$ (3 s.f.)

Watch out y and x are coded values. You can either code the given value of c, then reverse the coding for the resulting value of y (method 1). Or you can convert your regression equation in y and x to an equation in m and c (method 2).

Method 2

$y = 36.214\ldots - (4.047\ldots)x$

$\dfrac{m - 700}{5} = 36.214\ldots - 4.047\ldots \times 10c$

You can find an equation for the regression line of m on c by substituting $y = \dfrac{m - 700}{5}$ and $x = 10c$ into the regression line of y on x, then rearranging into the form $m = p + qc$

$m - 700 = 181.07\ldots - (202.38\ldots)c$

$m = 881.07\ldots - (202.38\ldots)c$

$= 881.07\ldots - (202.38\ldots) \times 0.25$

$= 830$ (3 s.f.)

The estimate for the melting point is 830°C (3 s.f.)

Write a conclusion in the context of the question and give units. If possible, you should check that your answer makes sense. If you substituted $x = 0.25$ into the regression line of y on x you would get a melting point of 35 °C, which is clearly wrong.

Exercise 1B

1 Given that the coding $p = x + 2$ and $q = y - 3$ has been used to get the regression equation $p + q = 5$, find the equation of the regression line of y on x in the form $y = a + bx$.

2 Given the coding $x = p - 10$ and $y = s - 100$ and the regression equation $x = y + 2$, work out the equation of the regression line of s on p.

3 Given that the coding $g = \frac{x}{3}$ and $h = \frac{y}{4} - 2$ has been used to get the regression equation $h = 6 - 4g$, find the equation of the regression line of y on x.

4 The regression line of t on s is found by using the coding $x = s - 5$ and $y = t - 10$.
The regression equation of y on x is $y = 14 + 3x$.
Work out the regression line of t on s.

5 A regression line of c on d is worked out using the coding $x = \frac{c}{2}$ and $y = \frac{d}{10}$

a Given that $S_{xy} = 120$, $S_{xx} = 240$, the mean of $x(\overline{x})$ is 5 and the mean of $y(\overline{y})$ is 6, calculate the regression line of y on x.

b Find the regression line of d on c.

E/P **6** Some data on the coverage area, $a\,m^2$, and cost, £c, of five boxes of flooring were collected.

The results were coded such that $x = \frac{a - 8}{2}$ and $y = \frac{c}{5}$

The coded results are shown in the table.

x	1	5	10	16	17
y	9	12	16	21	23

a Calculate S_{xy} and S_{xx} and use them to find the equation of the regression line of y on x. **(4 marks)**

b Find the equation of the regression line of c on a. **(2 marks)**

c Estimate the cost of a box of flooring which covers an area of $32\,m^2$. **(2 marks)**

E/P **7** A farmer collected data on the annual rainfall, x cm, and the annual yield of potatoes, p tonnes per acre.

The data for annual rainfall was coded using $v = \frac{x - 4}{8}$ and the following statistics were found:

$S_{vv} = 10.21$ $S_{pv} = 15.26$ $S_{pp} = 23.39$ $\overline{p} = 9.88$ $\overline{v} = 4.58$

a Find the equation of the regression line of p on v in the form $p = a + bv$. **(3 marks)**

b Using your regression line, estimate the annual yield of potatoes per acre when the annual rainfall is 42 cm. **(2 marks)**

1.2 Residuals

You can use residuals to check the reasonableness of a linear fit and to find possible outliers.

- **If a set of bivariate data has regression equation $y = a + bx$, then the residual of the data point (x_i, y_i) is given by $y_i - (a + bx_i)$. The sum of the residuals of all data points is 0.**

Consider the following data set:

x	1	2	4	6	7
y	1.2	1.7	3.1	5.2	5.8

The equation of the regression line of y on x is $y = 0.2 + 0.8x$.

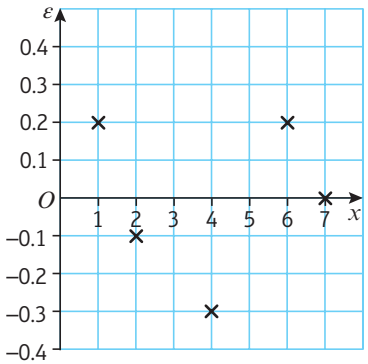
You can calculate the residuals for each data point and record them in a table:

x	y	$y = 0.2 + 0.8x$	ε
1	1.2	1.0	0.2
2	1.7	1.8	−0.1
4	3.1	3.4	−0.3
6	5.2	5.0	0.2
7	5.8	5.8	0

Use ε for the residual column. Remember that if the observed value is **less** than the predicted value then the residual will be negative.

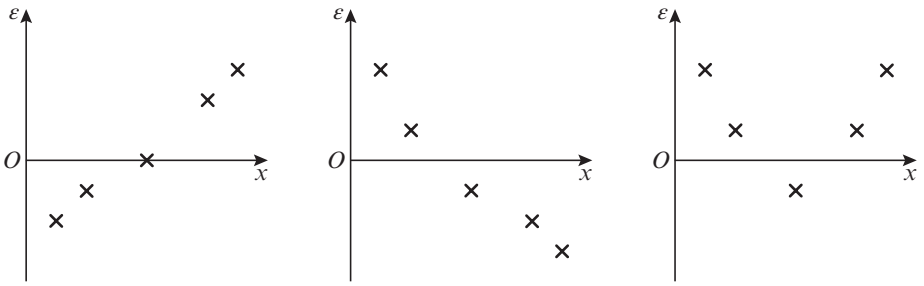
Notice that the sum of the residuals adds up to zero. → **Mixed exercise, Challenge**

The residuals can be plotted on a **residual plot** to show the trend:



The distribution of the residuals around zero is a good indicator of linear fit. You would expect the residuals to be randomly scattered about zero. If you see a trend in the residuals, you would question the appropriateness of the linear model.

Non-random residuals might follow an increasing pattern, a decreasing pattern or an obviously curved pattern. Here are three examples of residual patterns which might indicate that a linear model is not suitable:



Example 5

The table shows the relationship between the temperature, t °C, and the sales of ice cream, s, on five days in June:

Temp, t (°C)	15	16	18	19	21
Sales, s (100s)	12.0	15.0	17.5	p	24.0

Online Explore residuals of data points and reasonableness of fit using GeoGebra.

The equation of the regression line of s on t is given as $s = -17.154 + 1.9693t$.

a Calculate the residuals for the given regression line and hence find the value of p.

b By considering the residuals, comment on whether a linear regression model is suitable for these data.

a

t	s	$s = -17.154 + 1.9693t$	ε
15	12.0	12.3855	−0.3855
16	15.0	14.3548	0.6452
18	17.5	18.2934	−0.7934
19	p	20.2627	$p - 20.2627$
21	24.0	24.2013	−0.2013

$-0.3855 + 0.6452 - 0.7934 + (p - 20.2627) - 0.2013 = 0$

$\Rightarrow p = 21.0$ (3 s.f.)

The residual for $t = 19$ is 0.7373.

b The residuals appear to be randomly distributed around zero therefore it is likely that the linear regression model is suitable.

Calculate the predicted values using the regression line equation.

Write the residual for $t = 19$ in terms of p.

Use the fact that the sum of residuals adds up to zero.

Look at the distribution of the residuals around zero. They alternate signs, and don't follow an obvious pattern.

You can use residuals to identify possible outliers.

Example 6

The table shows the time taken, t minutes, to produce y litres of paint in a factory.

t	2.1	3.7	4.8	6.1	7.2
y	19.2	27.3	26.9	38.5	40.9

The regression line of y on t is given as $y = 9.7603 + 4.3514t$.

One of the y-values was incorrectly recorded.

a Calculate the residuals and write down the outlier.

b Comment on the validity of ignoring this outlier in your analysis.

c Ignoring the outlier, produce a new model.

d Use the new model to estimate the amount of paint that is produced in 4.8 minutes.

a

t	y	$y = 9.7603 + 4.3514t$	ε
2.1	19.2	18.8982	0.3018
3.7	27.3	25.8605	1.4395
4.8	26.9	30.6470	−3.7470
6.1	38.5	36.3038	2.1962
7.2	40.9	41.0904	−0.1904

The incorrect value is 26.9.

b The residuals suggest that this data point does not follow the pattern of the rest of the data, so it is valid to remove it.

c New model: $y = 10.669 + 4.3573t$

d $y = 10.669 + 4.3573 \times 4.8 = 31.6$ litres (3 s.f.)

Look for a data point with a residual that is far larger than the other residuals.

Problem-solving

You could also say that the data point **is** a valid piece of data so it should be used, or that there are only five data points so you should retain them all. You can make any reasonable conclusion as long as you give a reason.

Use your calculator to find the new values of a and b.

Substitute 4.8 into your new equation.

It is often useful to have a numerical value to indicate how closely a given set of data fits a linear regression model. Because the sum of the residuals is 0, you find the **square** of each residual and work out the sum of these values. This is called the **residual sum of squares (RSS)**.

- **You can calculate the residual sum of squares (RSS) for a linear regression model using the formula**

$$\text{RSS} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

The linear regression model is the linear model which minimises the RSS for a given set of data. Unlike the product moment correlation coefficient, which takes values between −1 and 1, the units of the RSS are same as the units of the response variable squared. For this reason, you should only use the RSS to compare goodness of fit for data recorded in the same units.

Note The formula is given in the formulae booklet. You are not expected to be able to derive it.

Links In your A-level course, you used the product moment correlation coefficient, r, to measure the strength and type of linear correlation. ← SM2, Section 1.2

The RSS is also linked to the product moment correlation coefficient, r, by the equation $\text{RSS} = S_{yy}(1 - r^2)$ → Section 2.1

Example 7

The data shows the sales, in 100s, y, of *Slush* at a riverside café and the number of hours of sunshine, x, on five random days during August.

x	8	10	11.5	12	12.2
y	7.1	8.2	8.9	9.2	9.5

Given that $\sum x = 53.7$ $\sum y = 42.9$ $\sum x^2 = 589.09$ $\sum y^2 = 371.75$ $\sum xy = 467.45$

a calculate the residual sum of squares (RSS).

The RSS for five random days in December is 0.0562.

b State, with a reason, which month is more likely to have a linear fit between the number of hours of sunshine and the sales of *Slush*.

a $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 371.75 - \frac{42.9^2}{5}$

$= 3.668$

$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 589.09 - \frac{53.7^2}{5}$

$= 12.352$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

$= 467.45 - \frac{53.7 \times 42.9}{5} = 6.704$

$\text{RSS} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 3.668 - \frac{6.704^2}{12.352}$

$= 0.0294$ (3 s.f.)

b 0.0294 < 0.0562 therefore August is more likely to have a linear fit.

Calculate the summary statistics.

Calculate the RSS using the given formula.

The smaller the value of the RSS, the more likely a linear fit.

Exercise 1C

1 The table shows the relationship between two variables, x and y:

x	1.1	1.3	1.4	1.7	1.9
y	12.2	14.5	16.9	p	23.5

The equation of the regression line of y on x is given as $y = -3.633 + 14.33x$.

Calculate the residuals for the given regression line and hence find the value of p.

E/P **2** The table shows the masses of six baby elephants, m kg, against the number of days premature they were born, x.

x	2	5	8	9	11	15
m	110	105	103	101	96	88

The equation of the regression line of m on x is given as $m = 114.3 - 1.655x$.

a Calculate the residual values. **(2 marks)**

b Draw a residual plot for this data. **(2 marks)**

c With reference to your residual plot, comment on the suitability of a linear model for this data. **(1 mark)**

Hint There is an example of a residual plot on page 11.

E/P **3** Sarah completes a crossword each day. She measures both the time taken, t minutes, and the accuracy of her answers, given as a percentage, p. She records this data for 10 days and the results are shown in the table below:

t	5.1	5.7	6.3	6.4	7.1	7.2	8.0	8.3	8.7	9.1
p	79	81	85	86	89	84	95	96	98	99

The regression line of p on t is given as $p = 51.04 + 5.308t$.

a Calculate the residuals and use your results to identify an outlier. **(3 marks)**

b State, with a reason, whether this outlier should be included in the data. **(1 mark)**

c Ignoring the outlier, produce another model. **(2 marks)**

d Use this model to predict the percentage of correct answers if the crossword takes Sarah 7.8 minutes to complete. **(1 mark)**

E/P **4** The table shows the age, x years, of a particular model of car and the value, y, in £1000s.

x	1.2	1.7	2.4	3.1	3.8	4.2	5.1
y	13.1	12.5	10.9	9.4	7.9	a	5.8

The regression line of y on x is given as $y = 15.7 - 2.02x$.

a Calculate the residuals and hence find the value of a, correct to three significant figures. **(3 marks)**

b By considering the signs of the residuals, explain whether or not the linear regression model is suitable for this data. **(1 mark)**

E/P **5** The table shows the ages of runners, a, against the times taken to complete an obstacle course, t minutes.

a	17	19	22	25	30	38	41	44
t	18	19	21	22	25	28	29	31

$\sum a = 236$ $\sum t = 193$ $\sum a^2 = 7720$ $\sum t^2 = 4821$ $\sum at = 6046$

a State what is measured by the residual sum of squares. **(1 mark)**

b Calculate the residual sum of squares (RSS). **(5 marks)**

The runners then complete a cross-country course. The RSS for the new set of data is 1.154.

c State, with a reason, which data is more likely to have a linear fit. **(1 mark)**

E/P **6** The table shows the amount of rainfall, $d$mm, against the relative humidity, h%, for Stratford-upon-Avon on 7 random days during September.

h	67	69	74	77	79	81	87
d	1.3	1.7	1.9	2.0	2.2	2.4	3.1

Given that $S_{hh} = 289.4$ $S_{dd} = 1.949$ $S_{hd} = 23.13$

a calculate the residual sum of squares (RSS). **(2 marks)**

The RSS for a random sample of 7 days in October is 0.0965.

b State, with a reason, which sample is more likely to have a linear fit. **(1 mark)**

E/P **7** A particular model of car depreciates in value as it gets older. The table below shows the ages, x years, and the values, y £1000s of a random sample of these cars.

x	0.7	1.3	1.8	2.3	2.9	3.8
y	15.4	13.5	12.1	10.1	8.5	5.8

$\sum x = 12.8$ $\sum y = 65.4$ $\sum x^2 = 33.56$ $\sum y^2 = 773.72$ $\sum xy = 120.03$

a Calculate the equation of the regression line of y on x, giving your answer in the form $y = a + bx$. Give the values of a and b correct to 4 significant figures. **(3 marks)**

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

c Use your regression line to estimate the value of a car that is 2 years old. **(1 mark)**

d Calculate the values of the residuals. **(2 marks)**

e Use your answer to part **d** to explain whether a linear model is suitable for these data. **(1 mark)**

f Calculate the residual sum of squares (RSS). **(2 marks)**

A sample for a second model of car has an RSS of 0.2548.

g State, with a reason, which sample is more likely to have a linear fit. **(1 mark)**

Challenge

The table shows the relationship between two variables, x and y.

x	1	5	7
y	9	p	q

Given that the equation of the regression line of y on x is $y = 2 + 4x$, Find the values of p and q.

Mixed exercise 1

E **1** Two variables s and t are thought to be connected by an equation of the form $t = a + bs$, where a and b are constants.

a Use the summary data

$\sum s = 553$ $\sum t = 549$ $\sum st = 31\,185$ $n = 12$ $\bar{s} = 46.0833$

$\bar{t} = 45.75$ $S_{ss} = 6193$

to work out the regression line of t on s. **(3 marks)**

b Find the value of t when s is 50. **(1 mark)**

E **2** A biologist recorded the breadth (x cm) and the length (y cm) of 12 beech leaves. The data collected can be summarised as follows.

$\sum x^2 = 97.73$ $\sum x = 33.1$ $\sum y = 66.8$ $\sum xy = 195.94$

a Calculate S_{xx} and S_{xy}. **(2 marks)**

b Find the equation of the regression line of y on x in the form $y = a + bx$. **(3 marks)**

c Predict the length of a beech leaf that has a breadth of 3.0 cm. **(1 mark)**

E/P **3** 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 are 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

a Calculate S_{xy} and S_{xx}. **(2 marks)**

b Find the equation of the regression line of y on x in the form $y = a + bx$. **(3 marks)**

c Estimate the Gross National Product of a country that has an energy consumption of 100. **(1 mark)**

d Estimate the energy consumption of a country that has a Gross National Product of 3500. **(1 mark)**

e Comment on the reliability of your answer to **d**. **(1 mark)**

E **4** In an environmental survey on the survival of mammals, the tail length t (cm) and body length m (cm) of a random sample of six small mammals of the same species were measured.

These data are coded such that $x = \frac{m}{2}$ and $y = t - 2$.

The data from the coded records are summarised below.

$\sum y = 13.5$ $\sum x = 25.5$ $\sum xy = 84.25$ $S_{xx} = 59.88$

a Find the equation of the regression line of y on x in the form $y = ax + b$. **(3 marks)**

b Hence find the equation of the regression line of t on m. **(2 marks)**

c Predict the tail length of a mammal that has a body length of 10 cm. **(2 marks)**

E/P **5** A sports scientist recorded the number of breaths per minute (r) and the pulse rate per minute (p) for 10 athletes at different levels of physical exertion. The data are shown in the table.

The data are coded such that $x = \frac{r - 10}{2}$ and $y = \frac{p - 50}{2}$

x	3	5	5	7	8	9	9	10	12	13
y	4	9	10	11	17	15	17	19	22	27

(You may use $\sum x = 81$ $\sum x^2 = 747$ $\sum y = 151$ $\sum y^2 = 2695$ $\sum xy = 1413$)

a Calculate S_{xy} and S_{xx}. **(2 marks)**

b Find the equation of the regression line of y on x in the form $y = a + bx$. **(3 marks)**

c Find the equation of the regression line for p on r. **(2 marks)**

d Estimate the number of pulse beats per minute for someone who is taking 22 breaths per minute. **(2 marks)**

e Comment on the reliability of your answer to **d**. **(1 mark)**

E/P **6** A farm food supplier monitors the number of hens kept (x) against the weekly consumption of hen food (y kg) for a sample of 10 small holders. He records the data and works out the regression line for y on x to be $y = 0.16 + 0.79x$.

a Write down a practical interpretation of the figure 0.79. **(1 mark)**

b Estimate the amount of food that is likely to be needed by a small holder who has 30 hens. **(2 marks)**

c If food costs £12 for a 10 kg bag, estimate the weekly cost of feeding 50 hens. **(2 marks)**

E/P **7** Water voles are becoming very rare. A naturalist society decided to record details of the water voles in their area. The members measured the mass (y) to the nearest 10 grams, and the body length (x) to the nearest millimetre, of eight active healthy water voles. The data they collected are in the table.

Body length, x (mm)	140	150	170	180	180	200	220	220
Mass, y (grams)	150	180	190	220	240	290	300	310

a Draw a scatter diagram of these data. **(2 marks)**

b Give a reason to support the calculation of a regression line for these data. **(1 mark)**

c Use the coding $l = \frac{x}{10}$ and $w = \frac{y}{10}$ to work out the regression line of w on l. **(3 marks)**

d Find the equation of the regression line for y on x. **(2 marks)**

e Draw the regression line on the scatter diagram. **(1 mark)**

f Use your regression line to calculate an estimate for the mass of a water vole that has a body length of 210 mm. Write down, with a reason, whether or not this is a reliable estimate. **(2 marks)**

The members of the society remove any water voles that seem unhealthy from the river and take them into care until they are fit to be returned.

They find three water voles on one stretch of river which have the following measurements.

A: Mass 235 g and body length 180 mm

B: Mass 180 g and body length 200 mm

C: Mass 195 g and body length 220 mm

g Write down, with a reason, which of these water voles were removed from the river. **(1 mark)**

E/P **8** A mail order company pays for postage of its goods partly by destination and partly by total weight sent out on a particular day. The number of items sent out and the total weights were recorded over a seven-day period. The data are shown in the table.

Number of items, *n*	10	13	22	15	24	16	19
Weight, *w* (kg)	2800	3600	6000	3600	5200	4400	5200

a Use the coding $x = n - 10$ and $y = \frac{w}{400}$ to work out S_{xy} and S_{xx}. **(4 marks)**

b Work out the equation of the regression line for y on x. **(3 marks)**

c Work out the equation of the regression line for w on n. **(2 marks)**

d Use your regression equation to estimate the weight of 20 items. **(2 marks)**

e State why it would be unwise to use the regression equation to estimate the weight of 100 items. **(1 mark)**

f Use your equation of the regression line found in part **b** to work out the residuals of the coded data points (x, y). **(2 marks)**

g Use your equation of the regression line found in part **c** to work out the residuals of the original data points (n, w). **(2 marks)**

h Explain how your answers to parts **f** and **g** are related to the coding used. **(1 mark)**

E/P **9** The table shows the time, t hours, against the temperature, T°C, of a chemical reaction.

t	2	3	5	6	7	9	10
T	72	68	59	54	50	42	38

Given that the equation of the regression line of T on t is $T = 80.445 - 4.289t$,

a calculate the residual values. **(2 marks)**

b State, with a reason, whether a linear model is suitable in this case. **(1 mark)**

Given that $S_{tt} = 52$, $S_{TT} = 957.43$ and $S_{tT} = -223$,

c calculate the residual sum of squares (RSS). **(2 marks)**

A second chemical reaction has a RSS of 0.8754.

d State, with a reason, which reaction is most likely to have a linear fit. **(1 mark)**

E/P **10** A meteorologist is developing a model to describe the relationship between the number of hours of sunshine, s, and the daily rainfall, f mm, in summer.

A random sample of the number of hours sunshine and the daily rainfall is taken from 8 days and are summarised below:

$\sum s = 53.4$ $\sum s^2 = 395.76$ $\sum f = 29.9$ $\sum f^2 = 131.93$ $\sum sf = 171.66$

a Calculate S_{ss} and S_{sf}. **(2 marks)**

b Find the equation of the regression line of f on s. **(3 marks)**

c Use your equation to estimate the daily rainfall when there is 7.5 hours of sunshine. **(1 mark)**

d Calculate the residual sum of squares (RSS). **(3 marks)**

The table shows the residual for each value of s.

s	3.1	4.2	5.4	6.2	7.1	8.8	9.1	9.5
Residual	−0.177	−0.196	0.256	0.124	x	−0.129	−0.216	−0.032

e Find the value of x. **(2 marks)**

f By considering the signs of the residuals, explain whether or not the linear regression model is suitable for these data. **(1 mark)**

E/P **11** A random sample of 9 baby southern hairy-nosed wombats was taken. The age, x, in days, and the mass, y grams, was recorded. The results were as follows:

x	2	3	4	5	6	7	8	9	10
y	4	5	7	8	9	11	12	11	15

(You may use $S_{xx} = 60$ $S_{yy} = 98.89$ $S_{xy} = 75$)

a Find the equation of the regression line of y on x in the form $y = a + bx$ as a model for these results. Give the values of a and b correct to three significant figures. **(2 marks)**

b Show that the residual sum of squares is 5.14 to three significant figures. **(2 marks)**

c Calculate the residual values. **(2 marks)**

d Write down the outlier. **(1 mark)**

e i Comment on the validity of ignoring this outlier.

ii Ignoring the outlier, produce another model.

iii Use this model to estimate the mass of a baby wombat after 20 days.

iv Comment, giving a reason, on the reliability of your estimate. **(5 marks)**

E/P **12** The annual turnover, £t million of eight randomly selected UK companies, and the number of staff employed in 100s, s, is recorded and the data shown in the table below:

t, £million	1.2	1.5	1.8	2.1	2.5	2.7	2.8	3.1
s, 00s	1.1	1.4	1.7	2.2	2.4	2.6	2.9	3.2

($\sum t = 17.7$ $\sum s = 17.5$ $\sum t^2 = 42.33$ $\sum s^2 = 42.07$ $\sum ts = 42.16$)

a Calculate the equation of the regression line of s on t, giving your answer in the form $s = a + bt$. Give the values of a and b correct to three significant figures. **(3 marks)**

b Use your regression line to predict the number of employees in a UK company with an annual turnover of £2 300 000. **(2 marks)**

The table shows the residuals for each value of t:

t	1.2	1.5	1.8	2.1	2.5	2.7	2.8	3.1
Residual	0.0121	−0.0137	−0.0395	0.1347	−0.0997	p	0.0745	0.0487

c Find the value of p. **(2 marks)**

d By considering the signs of the residuals, or otherwise, comment on the suitability of the linear regression model for these data. **(1 mark)**

e Calculate the residual sum of squares (RSS). **(2 marks)**

A random sample of equivalent companies in France is taken and the residual sum of squares is found to be 0.421.

f State, with a reason, which sample is likely to have the better linear fit. **(1 mark)**

Challenge

A set of bivariate data (x_1, y_1), (x_2, y_2), (x_3, y_3), … (x_n, y_n) is modelled by the linear regression equation $y = a + bx$, where $a = \overline{y} - b\overline{x}$.

a Prove that the sum of the residuals of the data points is 0.

b By means of an example, or otherwise, explain why this condition does not guarantee that the model closely fits the data.

Summary of key points

1 The **residual** of a given data point is the difference between the observed value of the dependent variable and the predicted value of the dependent variable.

2 The equation of the regression line of y on x is:

$$y = a + bx$$

where $b = \dfrac{S_{xy}}{S_{xx}}$ and $a = \overline{y} - b\overline{x}$

3 $S_{xy} = \sum xy - \dfrac{\sum x \sum y}{n}$

$S_{xx} = \sum x^2 - \dfrac{(\sum x)^2}{n}$

$S_{yy} = \sum y^2 - \dfrac{(\sum y)^2}{n}$

4 If a set of bivariate data has regression equation $y = a + bx$, then the residual of the data point (x_i, y_i) is given by $y_i - (a + bx_i)$. The sum of the residuals of all data points is 0.

5 You can calculate the residual sum of squares (RSS) for a linear regression model using the formula

$$\text{RSS} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$