General Mathematics Units 3 & 4

Chapter 3 — Investigating and modelling linear associations

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Chapter 3 Investigating and modelling linear associations — 3A Fitting a least-squares line

Theory

When a scatterplot of two numerical variables shows a linear association, we can summarise that association with a single straight line — a **linear model**. This section is about **fitting** that line: finding its equation from the data. Interpreting the line and using it to make predictions come later in this chapter.

Explanatory and response variables. Before fitting a line we must decide which variable is which. The **explanatory variable** (also called the independent variable) is the one we use to explain or predict the other; the **response variable** (the dependent variable) is the one whose value we are trying to explain or predict. In modelling we always place the **explanatory variable on the horizontal (x) axis** and the **response variable on the vertical (y) axis**. For example, if we use the number of hours studied to predict a test mark, then hours studied is explanatory (x) and the mark is the response (y). The choice matters: the line that best predicts y from x is **not** the same as the line that best predicts x from y, so swapping the roles generally changes the model.

The line of best fit. A straight line fitted to model the association between two numerical variables is called a **line of best fit**. Following the convention used throughout this course we write its equation as

$$y = a + b\,x\,,$$

where a is the **intercept** (the value of y where the line crosses the vertical axis, i.e. when $x = 0$) and b is the **slope** (the amount by which y changes for each increase of one unit in x). Writing the intercept first, $a + b\,x$, is the standard form in General Mathematics.

Residuals and the least-squares idea. For a given line, the value it predicts for a data point with explanatory value x is $\hat{y} = a + b\,x$. The **residual** of a data point is the *vertical* gap between the actual value y and the predicted value $\hat{y}$:

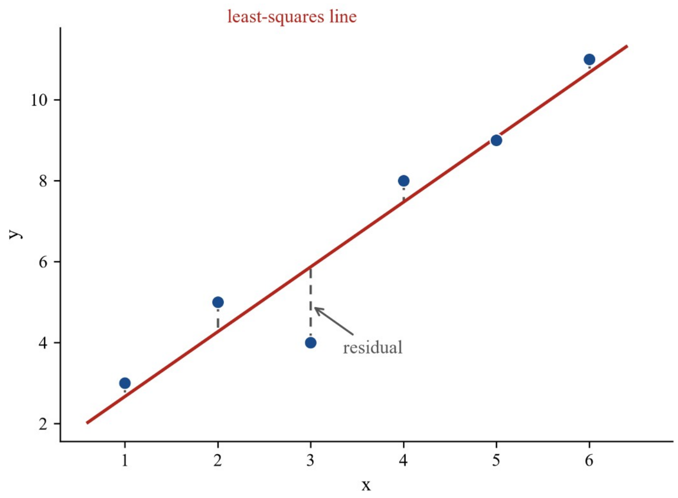
$$\text{residual} = y - \hat{y}\,.$$

A point above the line has a positive residual, a point below it has a negative residual, and a point exactly on the line has a residual of zero. A good line makes the residuals small overall.

Many different lines could be drawn through a cloud of points, and different people would draw different lines “by eye”. To obtain **one** line that everyone agrees on, we use the **least-squares** rule: the **least-squares line of best fit** (usually shortened to the **least-squares line**) is the line that makes the **sum of the squared residuals** as small as possible,

$$\text{minimise} \sum_{i=1}^{n} (\,y_i - \hat{y}_i)^2\,.$$

Squaring the residuals treats gaps above and below the line equally and penalises large gaps heavily, so the least-squares line balances the points as tightly as possible. It is the line used throughout this course, and for any set of data it is **unique**.



Finding the slope and intercept. The coefficients can be found from the summary statistics of the data. If the two variables have means $\bar{x}$ and $\bar{y}$, standard deviations s_x and s_y, and correlation coefficient r, then the slope and intercept of the least-squares line are

$$b = r\frac{s_y}{s_x}, a = \bar{y} - b\bar{x}.$$

Find b first, then substitute it into the formula for a. Two useful facts follow from these formulas: the slope b has the **same sign as** r (an upward line for a positive association, a downward line for a negative one), and, because $a = \bar{y} - b\bar{x}$ rearranges to $\bar{y} = a + b\bar{x}$, the least-squares line **always passes through the mean point** $(\bar{x}, \bar{y})$.

Rounding. When you compute b and a by hand, keep the full-precision value of b in your calculator and only round at the **end** — rounding b too early will throw out the value of a. Quote a and b to the number of decimal places or significant figures the question specifies (three significant figures is common).

Using CAS. General Mathematics is a technology-active course, so in practice you fit the line with your calculator: enter the paired data into the statistics editor, choose the explanatory variable as x and the response as y, and run the **linear regression** (least squares) command. The calculator reports the intercept a, the slope b and the correlation r directly. Read a and b from the output and write the equation as $y = a + bx$, taking care which number is the intercept and which is the slope. Learning the by-hand formulas lets you check that the calculator's line is reasonable.

Worked examples

Example 1 — scaffolded

A CAS one-variable and correlation summary for the number of hours x that several students spent on homework in a week and their score y (out of 30) on a class test gives $\bar{x} = 5$, $s_x = 2$, $\bar{y} = 20$, $s_y = 6$ and $r = 0.9$. Find the equation of the least-squares line.

Working	Comment
Hours studied is explanatory (x); test score is the response (y).	Set up the roles before fitting.
$b = r\frac{s_y}{s_x} = 0.9 \times \frac{6}{2}$.	Apply the slope formula.
$b = 0.9 \times 3 = 2.7$.	Find the slope first.
$a = \bar{y} - b\bar{x} = 20 - 2.7 \times 5$.	Substitute b into the intercept formula.
$a = 20 - 13.5 = 6.5$.	Find the intercept.
$y = 6.5 + 2.7x$.	Write the equation in the form $y = a + bx$.

The least-squares line is $y = 6.5 + 2.7x$.

Example 2 — scaffolded

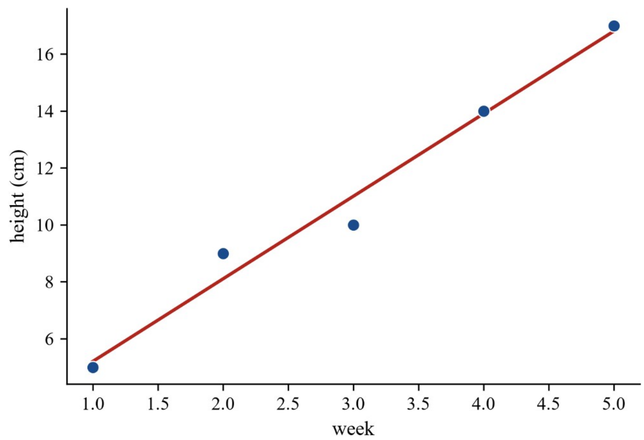
A seedling's height y (cm) was measured at the end of each of its first five weeks, giving the data below. Entering the data into CAS produces the two-variable statistics shown. Fit a least-squares line to model height in terms of the week number.

Week (x)	1	2	3	4	5
Height (y)	5	9	10	14	17

Statistic	Value
$\bar{x}$	3
s_x	1.581
$\bar{y}$	11
s_y	4.637
r	0.9889

Working	Comment
Week number is explanatory (x); height is the response (y).	The week is used to predict height.
$b = r\dfrac{s_y}{s_x} = 0.9889 \times \dfrac{4.637}{1.581} \approx 2.9.$	Slope formula, keeping full precision.
$a = \bar{y} - b\bar{x} = 11 - 2.9 \times 3 = 11 - 8.7 = 2.3.$	Intercept formula.
$y = 2.3 + 2.9x.$	Equation in the form $y = a + bx$.

The least-squares line is $y = 2.3 + 2.9x$. The line is drawn through the data below; notice it passes through the mean point $(3, 11)$.



Example 3 — unscaffolded

The age x (years) and value y (thousands of dollars) of six used cars of the same model were recorded, and a linear regression was run on CAS. The calculator screen reports the values below. Write down the equation of the least-squares line.

Output	Value
a	19.8
b	−1.8

Output	Value
r	−0.9930

The calculator lists the intercept as $a=19.8$ and the slope as $b=-1.8$. Substituting these into the form $y=a+bx$ gives $y=19.8+(-1.8)x$, which is written $y=19.8-1.8x$. The slope is negative, agreeing with the negative correlation ($r=-0.9930$): the fitted value falls as age increases.

∴ the least-squares line is $y=19.8-1.8x$.

Example 4 — unscaffolded

The age x (years) and trunk diameter y (cm) of six trees of the same species are recorded below. Fit a least-squares line, giving the coefficients correct to three significant figures, and confirm that the line passes through the mean point.

Age (x)	2	3	5	6	8	10
Diameter (y)	8	11	14	18	22	27

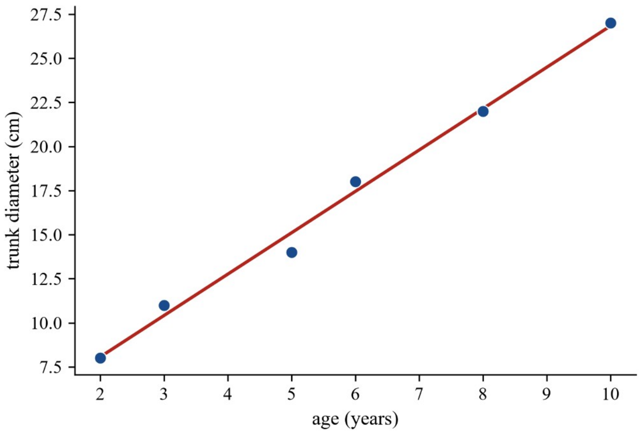
Age is the explanatory variable (x) and trunk diameter is the response (y), since we are modelling diameter in terms of age. Entering the six pairs into CAS and running the linear regression gives $\bar{x}=5.667$, $\bar{y}=16.67$ and the least-squares coefficients

$$b=2.35, a=3.38(\text{to three significant figures}),$$

so the least-squares line is $y=3.38+2.35x$.

To check the mean-point property, substitute $\bar{x}=5.667$ into the equation:
$a+b\bar{x}=3.375+2.3456\times5.667=16.67=\bar{y}$ (using unrounded coefficients). The line passes through $(\bar{x},\bar{y})=(5.667,16.67)$, as expected.

∴ the least-squares line is $y=3.38+2.35x$, and it passes through the mean point $(5.667,16.67)$.



Practice questions

Scaffolded

Q1. A CAS summary for two numerical variables gives $\bar{x}=8$, $s_x=3$, $\bar{y}=50$, $s_y=12$ and $r=0.8$. (a) Find the slope $b=r\frac{s_y}{s_x}$. (b) Find the intercept $a=\bar{y}-b\bar{x}$. (c) Write the equation of the least-squares line in the form $y=a+bx$.

Q2. The table gives the number of members x (hundreds) of a gym and its monthly profit y (thousands of dollars) for five months, together with the CAS two-variable statistics.

x	10	20	30	40	50
y	15	25	28	40	42

Statistic	Value
$\bar{x}$	30
s_x	15.811
$\bar{y}$	30
s_y	11.158
r	0.9778

(a) State the explanatory and response variables.
(b) Find the slope b, correct to two decimal places.
(c) Find the intercept a, correct to two decimal places.
(d) Write the equation of the least-squares line.

Q3. A CAS linear-regression screen reports the values below for the association between a quantity x and a quantity y.

Output	Value
a	4.3
b	1.45
r	0.9889

(a) State the value of the intercept.
(b) State the value of the slope.
(c) Write the equation of the least-squares line.

Q4. For each study, state which variable is the explanatory variable and which is the response variable. (a) Using a person's height to predict their arm span. (b) Using the maximum daily temperature to predict the number of cold drinks a shop sells. (c) Using the number of hours a machine has run to predict the amount of wear on a part. (d) Using a student's hours of sleep to predict their reaction time.

Q5. The table gives the number of coaching sessions attended x and the number of goals y scored in a season by six players, with the CAS statistics shown.

x	1	3	4	6	7	9
y	6	9	11	14	16	19

Statistic	Value
$\bar{x}$	5
s_x	2.898
$\bar{y}$	12.5
s_y	4.764
r	0.9994

(a) Find the slope b, correct to two decimal places.
(b) Find the intercept a, correct to two decimal places.
(c) Write the equation of the least-squares line.

Q6. A least-squares line has slope $b = 2.4$, and the data have means $\bar{x} = 10$ and $\bar{y} = 35$. (a) Use $a = \bar{y} - b\bar{x}$ to find the intercept a. (b) Write the equation of the least-squares line. (c) Verify that the line passes through the mean point $(10, 35)$.

Unscaffolded

Q7. For two numerical variables a CAS summary gives $\bar{x} = 4$, $s_x = 1.5$, $\bar{y} = 30$, $s_y = 9$ and $r = -0.7$. Find the equation of the least-squares line.

Q8. The table gives the amount of rainfall x (mm) in a week and the number of caterpillars y found on a plant. Use CAS to find the equation of the least-squares line, giving a and b correct to two decimal places.

x	5	10	15	20	25	30
y	42	38	35	28	25	19

Q9. A CAS linear-regression screen reports $a = 116.3612$, $b = -3.0617$ and $r = -0.9983$ for the association between a person's age x (years) and the number of minutes y they take to recover after exercise. Write the equation of the least-squares line, with a and b correct to three significant figures.

Q10. The table gives the number of years x a car has been driven and its fuel efficiency y (km/L). Find the equation of the least-squares line, with a and b correct to two decimal places.

x	2	4	6	8	10	12
y	38	34	31	25	22	17

Q11. A researcher wants to model how the number of pages in a printed report depends on the number of words it contains. (a) State the explanatory variable and the response variable. (b) State which variable is placed on the horizontal axis. (c) Write the general form of the equation that would be fitted.

Q12. A least-squares line has slope $b = 1.6$, and the data have means $\bar{x} = 15$ and $\bar{y} = 42$. Find the intercept a and write the equation of the line.

Q13. For two numerical variables the summary statistics are $\bar{x} = 12$, $\bar{y} = 60$, $s_x = 4$, $s_y = 20$ and $r = 0.85$. Find the equation of the least-squares line.

Q14. The table gives the number of weeks x a plant food has been used and the height y (cm) of a plant. Find the equation of the least-squares line (coefficients to two decimal places) and confirm that it passes through the mean point.

x	1	2	4	5	7	8
y	3	6	8	11	14	17

Q15. The table gives five data points that lie exactly on a straight line.

x	1	2	3	4	5
y	7	10	13	16	19

(a) Find the correlation coefficient r.
(b) Find the equation of the least-squares line.
(c) Explain what the value of r and the residuals tell you about the fit.

Q16. The table gives the floor area x (m²) and the sale price y (hundreds of thousands of dollars) of six apartments. Use CAS to find the equation of the least-squares line, giving the coefficients correct to three significant figures.

x	20	30	40	50	60	70
y	12	18	21	27	30	36

Q17. A CAS linear-regression screen reports $a = 64.8$, $b = -3.1$ and $r = -0.9979$ for two numerical variables x and y. (a) State the slope and the intercept. (b) Write the equation of the least-squares line.

Q18. In your own words, answer the following. (a) Explain what is meant by a residual, and what it means for a residual to be negative. (b) Explain the rule that defines the least-squares line.

Q19. The table gives the number of hours x a group of trainees studied and their score y on a practical test. Find the equation of the least-squares line, with a and b correct to two decimal places.

x	3	4	6	7	9	10
y	20	24	29	33	40	44

Q20. A gardener wants to model the mass y (g) of a pumpkin from the number of days x since it began to grow. Measurements on five pumpkins give the CAS statistics $\bar{x} = 140$, $s_x = 31.623$, $\bar{y} = 75.6$, $s_y = 10.807$ and $r = 0.9948$. (a) State the explanatory and response variables. (b) Find the equation of the least-squares line, with a and b correct to two decimal places.

Answers

Q1. (a) $b=3.2$; (b) $a=24.4$; (c) $y=24.4+3.2x$. **Q2.** (a) explanatory x (members), response y (profit); (b) $b=0.69$; (c) $a=9.30$; (d) $y=9.30+0.69x$. **Q3.** (a) intercept 4.3; (b) slope 1.45; (c) $y=4.3+1.45x$. **Q4.** (a) explanatory height, response arm span; (b) explanatory temperature, response cold drinks sold; (c) explanatory running hours, response wear; (d) explanatory hours of sleep, response reaction time. **Q5.** (a) $b=1.64$; (b) $a=4.29$; (c) $y=4.29+1.64x$. **Q6.** (a) $a=35-2.4\times10=11$; (b) $y=11+2.4x$; (c) at $x=10$, $y=11+2.4\times10=35=\overline{y}$. **Q7.** $b=-4.2$, $a=46.8$; $y=46.8-4.2x$. **Q8.** $b=-0.92$, $a=47.27$; $y=47.27-0.92x$. **Q9.** $a=116$, $b=-3.06$; $y=116-3.06x$. **Q10.** $b=-2.10$, $a=42.53$; $y=42.53-2.10x$. **Q11.** (a) explanatory = number of words, response = number of pages; (b) number of words; (c) $y=a+bx$ (pages $=a+b\times$ words). **Q12.** $a=42-1.6\times15=18$; $y=18+1.6x$. **Q13.** $b=4.25$, $a=9$; $y=9+4.25x$. **Q14.** $b=1.88$, $a=1.37$; $y=1.37+1.88x$; passes through $(4.5,9.83)$. **Q15.** (a) $r=1$; (b) $y=4+3x$; (c) $r=1$ means a perfect positive linear association, and every residual is 0, so all five points lie exactly on the line. **Q16.** $b=0.463$, $a=3.17$; $y=3.17+0.463x$. **Q17.** (a) slope -3.1, intercept 64.8; (b) $y=64.8-3.1x$. **Q18.** (a) a residual is the vertical gap $y-\hat{y}$ between a data point and the line; a negative residual means the point lies below the line (actual value less than predicted); (b) the least-squares line is the unique line that makes the sum of the squared residuals as small as possible. **Q19.** $b=3.36$, $a=9.83$; $y=9.83+3.36x$. **Q20.** (a) explanatory x (days), response y (mass); (b) $b=0.34$, $a=28.00$; $y=28.00+0.34x$.

Fully-worked solutions

Q1. The slope is $b=r\dfrac{s_y}{s_x}=0.8\times\dfrac{12}{3}=0.8\times4=3.2$. The intercept is $a=\overline{y}-b\overline{x}=50-3.2\times8=50-25.6=24.4$. Hence the least-squares line is $y=24.4+3.2x$.

Q2. Members (x) is the explanatory variable and profit (y) is the response, since we model profit in terms of membership. The slope is $b=r\dfrac{s_y}{s_x}=0.9778\times\dfrac{11.158}{15.811}\approx0.69$. The intercept is $a=\overline{y}-b\overline{x}=30-0.69\times30\approx9.30$ (keeping the full value of b before rounding). The least-squares line is $y=9.30+0.69x$.

Q3. A CAS regression screen lists the intercept as a and the slope as b. Here $a=4.3$ and $b=1.45$, so, substituting into $y=a+bx$, the least-squares line is $y=4.3+1.45x$.

Q4. The explanatory variable is the one used to predict the other, and the response is the one being predicted. (a) Height is used to predict arm span, so height is explanatory and arm span is the response. (b) Temperature is used to predict drink sales, so temperature is explanatory and cold drinks sold is the response. (c) Running hours is used to predict wear, so running hours is explanatory and wear is the response. (d) Hours of sleep is used to predict reaction time, so hours of sleep is explanatory and reaction time is the response.

Q5. The slope is $b=r\dfrac{s_y}{s_x}=0.9994\times\dfrac{4.764}{2.898}\approx1.64$. The intercept is $a=\overline{y}-b\overline{x}=12.5-1.6429\times5\approx4.29$ (using the unrounded slope 1.6429). The least-squares line is $y=4.29+1.64x$.

Q6. (a) The intercept is $a=\overline{y}-b\overline{x}=35-2.4\times10=35-24=11$. (b) The least-squares line is therefore $y=11+2.4x$. (c) Substituting $x=10$ gives $y=11+2.4\times10=11+24=35$, which equals $\overline{y}$, so the line passes through $(\overline{x},\overline{y})=(10,35)$ — as every least-squares line must.

Q7. The slope is $b=r\dfrac{s_y}{s_x}=-0.7\times\dfrac{9}{1.5}=-0.7\times6=-4.2$, which is negative because $r<0$. The intercept is $a=\overline{y}-b\overline{x}=30-(-4.2)\times4=30+16.8=46.8$. The least-squares line is $y=46.8-4.2x$.

Q8. Entering the six pairs into CAS with rainfall as x and caterpillars as y and running the linear regression gives $b=-0.92$ and $a=47.27$ (to two decimal places). The slope is negative, matching the downward trend in the data. The least-squares line is $y=47.27-0.92x$.

Q9. The calculator lists the intercept as $a = 116.3612$ and the slope as $b = -3.0617$. Rounded to three significant figures these are $a = 116$ and $b = -3.06$. Substituting into $y = a + bx$ gives $y = 116 - 3.06x$.

Q10. Entering the six pairs into CAS with years driven as x and fuel efficiency as y gives $b = -2.10$ and $a = 42.53$ (to two decimal places). The least-squares line is $y = 42.53 - 2.10x$.

Q11. (a) The number of words is used to predict the number of pages, so the number of words is the explanatory variable and the number of pages is the response. (b) The explanatory variable is always placed on the horizontal axis, so the number of words goes on the horizontal axis. (c) The fitted equation has the form $y = a + bx$; that is, pages $= a + b \times$ words.

Q12. The intercept is $a = \overline{y} - b\overline{x} = 42 - 1.6 \times 15 = 42 - 24 = 18$. The least-squares line is $y = 18 + 1.6x$.

Q13. The slope is $b = r\dfrac{s_y}{s_x} = 0.85 \times \dfrac{20}{4} = 0.85 \times 5 = 4.25$. The intercept is $a = \overline{y} - b\overline{x} = 60 - 4.25 \times 12 = 60 - 51 = 9$. The least-squares line is $y = 9 + 4.25x$.

Q14. Entering the six pairs into CAS gives means $\overline{x} = 4.5$ and $\overline{y} = 9.833$, and least-squares coefficients $b = 1.88$ and $a = 1.37$ (to two decimal places), so the line is $y = 1.37 + 1.88x$. Checking the mean-point property with the unrounded coefficients, $a + b\overline{x} = 1.3733 + 1.88 \times 4.5 = 9.833 = \overline{y}$, so the line passes through $(4.5, 9.83)$.

Q15. (a) The five points lie exactly on a straight line, so the linear association is perfect and CAS returns $r = 1$. (b) Running the regression (or noticing that y increases by 3 for each increase of 1 in x, starting from $y = 7$ at $x = 1$) gives $b = 3$ and $a = 4$, so the line is $y = 4 + 3x$. (c) Because $r = 1$ the association is a perfect positive line; every predicted value $\hat{y} = 4 + 3x$ equals the actual value, so every residual is 0 and all five points lie exactly on the least-squares line.

Q16. Entering the six pairs into CAS with floor area as x and price as y gives the least-squares coefficients $b = 0.463$ and $a = 3.17$ (to three significant figures). The least-squares line is $y = 3.17 + 0.463x$.

Q17. (a) On the CAS screen b is the slope and a is the intercept, so the slope is -3.1 and the intercept is 64.8. (b) Substituting into $y = a + bx$ gives the least-squares line $y = 64.8 - 3.1x$.

Q18. (a) A residual is the vertical distance between an actual data point and the fitted line, $\text{residual} = y - \hat{y}$, where $\hat{y} = a + bx$ is the value the line predicts. A negative residual means $y < \hat{y}$, so the data point lies **below** the line — the line over-predicts at that point. (b) The least-squares line is the unique straight line for which the **sum of the squared residuals** $\sum (y_i - \hat{y}_i)^2$ is as small as possible; squaring makes gaps above and below the line count equally and penalises large gaps, giving one agreed line of best fit.

Q19. Entering the six pairs into CAS with hours studied as x and score as y gives $b = 3.36$ and $a = 9.83$ (to two decimal places). The least-squares line is $y = 9.83 + 3.36x$.

Q20. (a) The number of days (x) is used to predict the pumpkin's mass (y), so days is the explanatory variable and mass is the response. (b) The slope is $b = r\dfrac{s_y}{s_x} = 0.9948 \times \dfrac{10.807}{31.623} \approx 0.34$, and the intercept is $a = \overline{y} - b\overline{x} = 75.6 - 0.34 \times 140 = 75.6 - 47.6 = 28.00$. The least-squares line is $y = 28.00 + 0.34x$.

Theory

Recall the least squares line. When two numerical variables have a linear association, we model it with the **least squares line** (or **least squares regression line**)

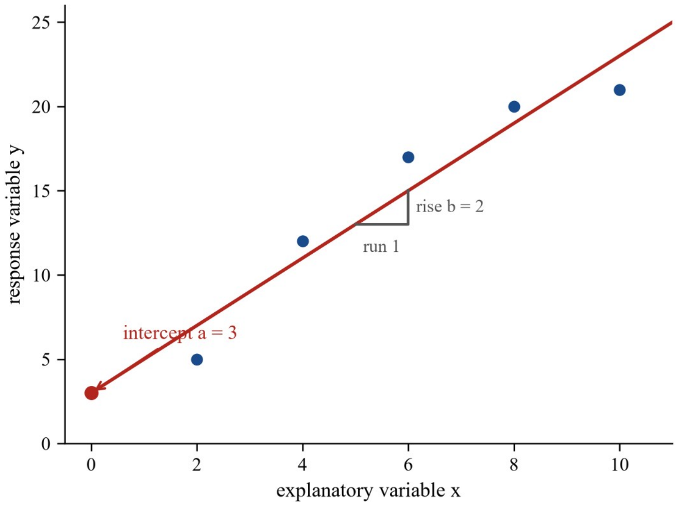
$$y = a + bx,$$

where x is the **explanatory** variable, y is the **response** variable, b is the **slope** and a is the y**-intercept**. Having fitted the line to a set of data — by reading the coefficients from a CAS linear-regression screen, or from the formulas

$$b = r\frac{s_y}{s_x}, a = \bar{y} - b\bar{x},$$

which we met when we first fitted the line — this section is about what the line *tells us*: how to read the slope and intercept in context, how to use the line to make predictions, and how far those predictions can be trusted.

The scatterplot below shows five data points and their least squares line $y = 3 + 2x$. Read off the two numbers that describe the line: the **intercept** $a = 3$ is the height at which the line crosses the vertical axis (where $x = 0$), and the **slope** $b = 2$ is the amount the line rises for each increase of one unit in x.



Interpreting the slope. The slope b is the **average change in the response variable for each one-unit increase in the explanatory variable**. Its sign gives the direction of the association:

- if $b > 0$, the response **increases** by b units (on average) for every one-unit increase in x;
- if $b < 0$, the response **decreases** by $|b|$ units (on average) for every one-unit increase in x.

A slope interpretation must always be written **in context**, naming the two variables and their units, and must include the words “on average” — the line describes the overall trend, not what happens to any single individual. For example, for the fitted line

$$\text{mark} = 42 + 3.5 \times (\text{hours}),$$

the slope 3.5 means that, on average, a student’s mark **increases by** 3.5 **marks for each additional hour** of study.

Interpreting the intercept. The intercept a is the value the line predicts for the response when the explanatory variable is **zero** ($x = 0$). Substituting $x = 0$ into $y = a + bx$ gives $y = a$, so a is simply the predicted response at $x = 0$.

The intercept is only **meaningful** when

- $x = 0$ lies within, or close to, the range of the observed x-values, **and**

- $x = 0$ makes sense in the context.

If **either** condition fails, the intercept is still needed to draw the line, but it should **not** be interpreted as a real quantity — it is then a prediction beyond what the data support, or one with no sensible meaning in the context, and may even be impossible (a negative price, a negative count, and so on).

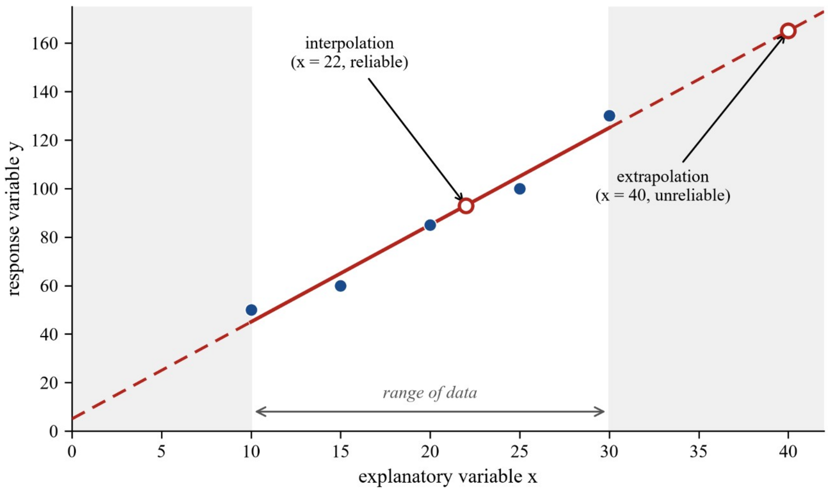
Making predictions. To predict the response for a chosen value of the explanatory variable, **substitute that value of** x **into the equation** $y = a + bx$. The result is a **predicted value**, which we write $\hat{y}$ (read "y-hat") to keep it distinct from an actual observed value y. For the mark example above, a student who studies 6 hours has a predicted mark of

$$\hat{y} = 42 + 3.5 \times 6 = 63.$$

Interpolation and extrapolation. The data were collected over a limited range of x-values, from the smallest observed value to the largest. A prediction made:

- **inside** that range is called **interpolation**;
- **outside** that range is called **extrapolation**.

Interpolation is generally reliable, because we have evidence — the data themselves — that the linear relationship holds across that range. **Extrapolation is unreliable**: outside the observed range we have no evidence that the linear pattern continues, and the relationship may bend, level off, or break down entirely. The further beyond the data we go, the less the prediction can be trusted, and a model can produce values that are clearly impossible once we push it too far.

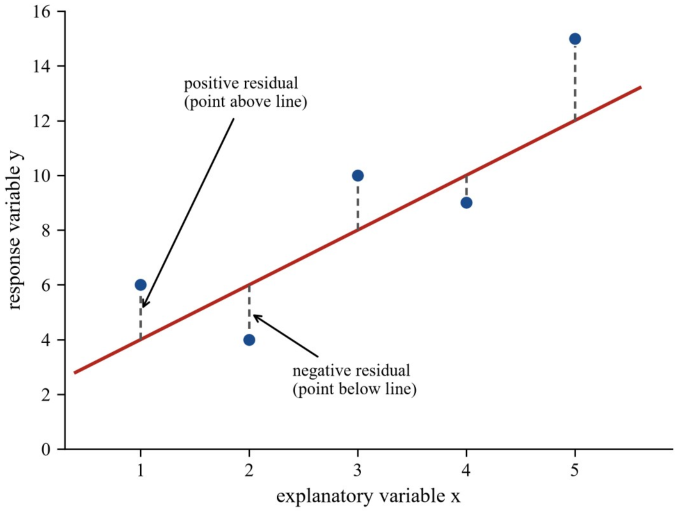


Residuals. No line passes exactly through every point. For a data point with explanatory value x and **observed** response y, the line gives the predicted response $\hat{y} = a + bx$, and the **residual** is the difference

$$\text{residual} = y - \hat{y} = \text{observed} - \text{predicted}.$$

- A point **above** the line has a **positive** residual (the line **under**-predicts it).
- A point **below** the line has a **negative** residual (the line **over**-predicts it).
- A point exactly on the line has a residual of 0.

A residual measures the vertical gap between a point and the line, so small residuals mean the line fits that point well. The vertical dashed segments in the diagram below are the residuals. Examining all of the residuals together to judge whether a straight line was the right model is the subject of a later section in this chapter; here we simply calculate the residual for a single point.



Using CAS. In practice you enter the two lists into your CAS, run the linear-regression command, and read off a and b directly; the calculator will also report a predicted value for any x you supply. Learn the by-hand substitution below so that you understand exactly what the calculator is reporting and can interpret it in context.

Worked examples

Example 1 — scaffolded

A gym fitted the least squares line $y=42+3.5x$ relating a member's weekly fitness score y to the number of training sessions per week x. The members in the study trained between 1 and 8 times per week. Interpret the slope and intercept, and use the line to make two predictions.

Working	Comment
Slope $b=3.5$: on average the fitness score increases by 3.5 points for each extra training session per week.	The slope is the average change in the response per one-unit increase in x; state it in context with "on average".
Intercept $a=42$: the predicted fitness score for a member who does 0 sessions per week.	The intercept is the predicted response at $x=0$.
$x=0$ is below the smallest observed value (1), so this is an extrapolation; treat it with caution.	$x=0$ lies outside the data range, so the intercept is not a reliable figure.
Predict for $x=5$: $\hat{y}=42+3.5\times5=59.5$.	Substitute $x=5$ into $y=a+bx$.
5 lies in 1–8, so this is **interpolation** — reliable.	A prediction inside the data range.
Predict for $x=12$: $\hat{y}=42+3.5\times12=84$.	Substitute $x=12$.
12 is beyond the largest observed value (8), so this is **extrapolation** — unreliable.	No data support the line this far out.

Example 2 — scaffolded

The least squares line relating the price y (in \$1000s) of a used car to its age x (in years) is

$$y=20-1.5x.$$

The sample contained cars aged 1 to 6 years. Interpret the slope, predict the price of a 4-year-old car, and find the residual for a particular 4-year-old car that sold for \$15,500.

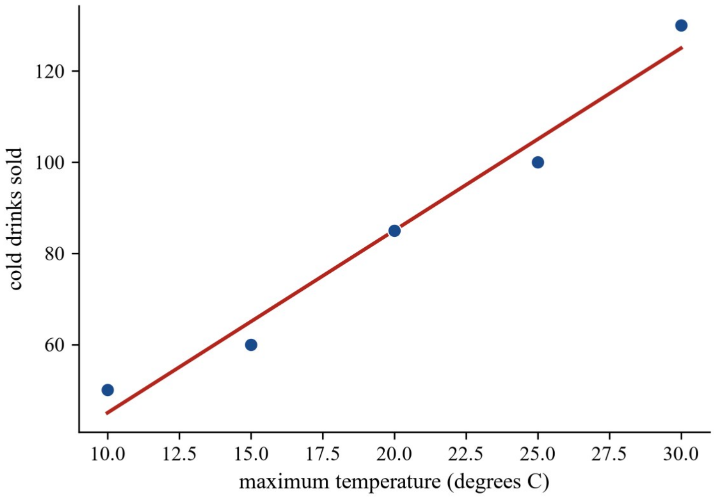
Working	Comment
Slope $b=-1.5$: on average the price **falls** by \$1500 for each additional year of age.	b is negative, so the response decreases; 1.5 in \$1000s is \$1500.
Predict for $x=4$: $\hat{y}=20-1.5\times4=14$, i.e. \$14,000.	Substitute $x=4$ into the line (4 is within 1–6, so interpolation).

Working	Comment
Observed price = $\$15500$, so observed $y = 15.5$ (in $1000s).	Convert the actual sale price to the same units.
Residual = $y - \hat{y} = 15.5 - 14 = 1.5$, i.e. $+\$1500$.	residual = observed − predicted.
The residual is positive, so this car sold for $1500 **more** than predicted; it lies **above** the line.	A positive residual means the line under-predicted this point.

Example 3 — unscaffolded

On five days a kiosk recorded the maximum temperature x (°C) and the number of cold drinks sold y. A CAS gives the least squares line $y = 5 + 4x$ from the data below. Interpret the slope and the intercept, predict the sales on a 22°C day, and find the residual for the 30°C day.

temperature (°C)	10	15	20	25	30
drinks sold	50	60	85	100	130



The slope is $b = 4$, so on average the kiosk sells 4 more cold drinks for each 1°C increase in the maximum temperature. The intercept is $a = 5$, the number of drinks the line predicts at 0°C. Because the coldest day recorded was 10°C, the value $x = 0$ is well outside the data range, so this is an extrapolation; the intercept is **not a meaningful** figure here.

Predicting the sales on a 22°C day (which lies inside the observed range 10–30°C, so this is interpolation):

$$\hat{y} = 5 + 4 \times 22 = 93 \text{ drinks.}$$

For the 30°C day the line predicts $\hat{y} = 5 + 4 \times 30 = 125$ drinks, while the kiosk actually sold 130. The residual is

$$\text{residual} = 130 - 125 = 5.$$

∴ the slope is 4 drinks per °C, the intercept (5) is not meaningful, the predicted sales at 22°C are 93 drinks, and the 30°C day has a residual of $+5$ (it sold 5 more drinks than predicted).

Example 4 — unscaffolded

For a suburb, the least squares line relating a house's price y (in $1000s) to its distance x (in km) from the city centre is $y = 620 - 18x$. The houses in the study were between 2 km and 20 km from the centre. Interpret the slope and intercept, predict the price of a house 8 km out, find the residual for an 8 km house that sold for $500,000, and explain what goes wrong if the line is used to predict the price 40 km from the centre.

The slope is $b = -18$, so on average a house's price **falls by $18,000 for each additional kilometre** from the city centre. The intercept is $a = 620$, the predicted price of a house right in the city centre ($x = 0$). Since the closest house in the study was 2 km out, $x = 0$ is only just beyond the data and is a sensible location, so \$620,000 is a reasonable interpretation of the intercept.

Predicting for $x = 8$ (within the range 2–20, so interpolation):

$$\hat{y} = 620 - 18 \times 8 = 476 \text{, i.e. } \$476\,000 .$$

An 8 km house that sold for \$500,000 has observed $y = 500$, so its residual is

$$\text{residual} = 500 - 476 = 24 \text{, i.e. } + \$24\,000 ,$$

meaning it sold for \$24,000 more than the line predicts (it lies above the line).

Predicting for $x = 40$ (well beyond the largest observed distance, 20 km — an extrapolation) gives

$$\hat{y} = 620 - 18 \times 40 = -100 ,$$

a predicted price of $-\$100\,000$. A negative price is impossible, which shows vividly that the linear model cannot be trusted outside the range of the data.

∴ the price falls \$18,000 per km, the intercept (\$620,000) is meaningful, the predicted price at 8 km is \$476,000, the residual is $+\$24\,000$, and extrapolating to 40 km gives an impossible negative price — a warning against extrapolation.

Practice questions

Scaffolded

Q1. A least squares line $y = 8 + 2.5x$ relates a student's quiz mark y (out of 30) to the number of hours studied x. Students studied between 1 and 10 hours. (a) State the slope. (b) Interpret the slope in context. (c) State the intercept. (d) Interpret the intercept in context. (e) Predict the quiz mark for a student who studies 6 hours.

Q2. The least squares line $y = 80 - 4x$ relates the number of hot chocolates sold y to the maximum daily temperature x (°C). Temperatures ranged from 3°C to 20°C. (a) Interpret the slope in context. (b) Interpret the intercept, and state whether it is meaningful. (c) Predict the sales on a 10°C day, and state whether this is interpolation or extrapolation. (d) Predict the sales on a 25°C day. Comment on the reliability of this prediction.

Q3. A least squares line is $y = 3 + 2x$. A data point has $x = 5$ and an observed value $y = 17$. (a) Find the predicted value $\hat{y}$ at $x = 5$. (b) Find the residual for this point. (c) State whether the point lies above or below the line.

Q4. Data were collected for explanatory values x ranging from 4 to 15. For each prediction below, state whether it is interpolation or extrapolation. (a) $x = 6$ (b) $x = 10$ (c) $x = 2$ (d) $x = 20$

Q5. A least squares line $y = 12 + 0.5x$ relates a plant's height y (cm) to its age x (months). The plants were between 5 and 30 months old. (a) Interpret the slope in context. (b) Predict the height of a 10-month-old plant. (c) Predict the height of a 40-month-old plant, and explain why this prediction is unreliable.

Q6. A least squares line is $y = 5 + 3x$. Complete the table of predicted values and residuals.

x	observed y	predicted $\hat{y}$	residual
2	13		
5	19		
8	29		

(a) Fill in the predicted values.
(b) Fill in the residuals.
(c) Which point does the line fit exactly, and which point does it fit worst?

Unscaffolded

Q7. The least squares line $y = -60 + 0.15x$ relates the weekly amount saved y ($) to the weekly income x ($) of a group of workers. Interpret the slope in context.

Q8. For the same line $y = -60 + 0.15x$, interpret the intercept and state, with a reason, whether it is meaningful.

Q9. The least squares line $y = 38 + 2.5x$ relates annual salary y (in $1000s) to years of experience x. Experience in the study ranged from 1 to 25 years. Predict the salary of a worker with (a) 10 years and (b) 40 years of experience, stating in each case whether the prediction is reliable.

Q10. A least squares line is $y = 100 - 2.5x$. A data point has $x = 20$ and observed value $y = 55$. Find the predicted value and the residual, and state whether the point lies above or below the line.

Q11. The least squares line $y = 500 - 30x$ relates a house's price y (in $1000s) to its distance x (km) from a train station. Use the line to predict the price at $x = 20$ km, and explain what your answer shows about extrapolation.

Q12. The number of items sold y was recorded for $x = 1, 2, 3, 4, 5$ television advertisements, and a CAS gives the least squares line $y = 4 + 3x$.

x	1	2	3	4	5
observed y	8	10	11	16	20

Find the predicted value and the residual for the $x = 3$ data point.

Q13. The least squares line $y = 95 - 1.2x$ relates a phone battery's capacity y (%) to the number of days x since it was fully charged and left unused. Interpret the slope in context.

Q14. A least squares line is $y = 25 + 4x$, fitted to data with x between 1 and 15. For what value of x does the line predict $y = 65$? Is this prediction reliable?

Q15. A least squares line is $y = 7 + 2x$. A data point at $x = 6$ has a residual of -3. Find the observed value y at that point.

Q16. The least squares line $y = 1.5 + 0.08x$ relates the fuel used y (litres) to the distance driven x (km) on a trip. (a) Interpret the slope in context. (b) Interpret the intercept in context. (c) Predict the fuel used on a 200 km trip.

Q17. (a) Explain the difference between interpolation and extrapolation. (b) Explain why extrapolation can give unreliable predictions.

Q18. A least squares line is $y = 2 + 1.5x$. Four data points are given below.

x	2	4	6	8
observed y	6	8	10	15

(a) Find the residual for each point.
(b) Which point does the line fit exactly?
(c) For which point does the line over-predict the response?

Q19. The least squares line $y = 12 + 0.75x$ relates the height y (m) of a gum tree to its age x (years). The trees measured were between 2 and 40 years old. (a) Interpret the slope in context. (b) Interpret the intercept, and state whether it is meaningful. (c) Predict the height of a 20-year-old tree. (d) A 20-year-old tree was measured at 30 m. Find its residual.

Q20. The least squares line $y = 2 + 1.5x$ relates a tomato plant's yield y (kg) to the number of hours of direct sunlight x it receives each day. The plants received between 3 and 10 hours. (a) Interpret the slope in context. (b) Predict the yield of a plant receiving 8 hours of sunlight. (c) Predict the yield of a plant receiving 15 hours, and comment on the reliability. (d) A plant receiving 8 hours yielded 13 kg. Find its residual, and state whether the line over- or under-predicted its yield.

Answers

Q1. (a) 2.5; (b) on average the mark increases by 2.5 for each extra hour studied; (c) 8; (d) the predicted mark for 0 hours of study is 8 (an extrapolation, since $x=0$ is below the range); (e) $\hat{y}=23$. **Q2.** (a) on average sales fall by 4 hot chocolates for each 1°C rise in temperature; (b) 80 is the predicted sales at 0°C, an extrapolation (0°C is below the range 3–20), so not reliably meaningful; (c) $\hat{y}=40$, interpolation; (d) $\hat{y}=-20$, an impossible negative value — extrapolation, unreliable. **Q3.** (a) $\hat{y}=13$; (b) residual $=4$; (c) above the line. **Q4.** (a) interpolation; (b) interpolation; (c) extrapolation; (d) extrapolation. **Q5.** (a) on average the plant grows 0.5 cm per month; (b) 17 cm; (c) $\hat{y}=32$ cm, unreliable because 40 months is beyond the observed range (5–30), i.e. extrapolation. **Q6.** (a) predicted $11, 20, 29$; (b) residuals $+2, -1, 0$; (c) fits the $x=8$ point exactly (residual 0) and fits the $x=2$ point worst (residual $+2$). **Q7.** On average, savings increase by $0.15 (15 cents) for each extra $1 of weekly income. **Q8.** The intercept -60 is the predicted saving for someone with $0 income, i.e. saving $-\$60$. It is not meaningful: $x=0$ is outside the data and a negative saving is not sensible here. **Q9.** (a) $\hat{y}=63$, i.e. $63,000, interpolation — reliable; (b) $\hat{y}=138$, i.e. $138,000, extrapolation ($40>25$) — unreliable. **Q10.** $\hat{y}=50$; residual $=5$; above the line. **Q11.** $\hat{y}=500-30\times 20=-100$, i.e. $-\$100\,000$. A negative price is impossible, showing the line cannot be extrapolated this far. **Q12.** $\hat{y}=13$; residual $=11-13=-2$. **Q13.** On average the battery's capacity falls by 1.2 percentage points for each additional day. **Q14.** $x=10$; reliable, since 10 lies within the range 1–15 (interpolation). **Q15.** $y=16$. **Q16.** (a) on average 0.08 litres of fuel are used per km driven; (b) 1.5 litres is the predicted fuel used at 0 km (before driving); (c) $\hat{y}=17.5$ litres. **Q17.** (a) interpolation predicts inside the range of the observed x-values; extrapolation predicts outside it. (b) Outside the data there is no evidence the linear pattern continues, so the relationship may change and predictions can be badly wrong or even impossible. **Q18.** (a) residuals $+1, 0, -1, +1$; (b) the $x=4$ point (residual 0); (c) the $x=6$ point (residual -1, so the line over-predicts). **Q19.** (a) on average the tree grows 0.75 m per year; (b) 12 m is the predicted height at age 0, not meaningful (a newly planted tree is not 12 m tall, so $x=0$ makes no sense in this context); (c) $\hat{y}=27$ m; (d) residual $=3$ m. **Q20.** (a) on average the yield increases by 1.5 kg per extra hour of daily sunlight; (b) $\hat{y}=14$ kg; (c) $\hat{y}=24.5$ kg, unreliable — 15 hours is extrapolation (>10); (d) residual $=-1$ kg, the line over-predicted the yield.

Fully-worked solutions

Q1. The line is $y=8+2.5x$. (a) The slope is the coefficient of x, $b=2.5$. (b) On average the quiz mark increases by 2.5 marks for each additional hour of study. (c) The intercept is the constant term, $a=8$. (d) It is the mark the line predicts for $x=0$ hours of study; since the smallest observed value is 1 hour, $x=0$ is just outside the data (extrapolation), so it is the "no study" mark but should be read with caution. (e) $\hat{y}=8+2.5\times 6=8+15=23$.

Q2. The line is $y=80-4x$. (a) The slope -4 means that, on average, 4 fewer hot chocolates are sold for each 1°C rise in the maximum temperature. (b) The intercept 80 is the predicted number of sales at 0°C; the coldest day in the data was 3°C, so $x=0$ is an extrapolation and the figure is not reliably meaningful. (c) $\hat{y}=80-4\times 10=40$; since 10 lies in 3–20 this is interpolation. (d) $\hat{y}=80-4\times 25=-20$. A negative number of sales is impossible, and 25°C is beyond the observed range, so this extrapolation is unreliable.

Q3. With $x=5$, the predicted value is $\hat{y}=3+2\times 5=13$. The residual is observed $-$ predicted $=17-13=4$. Because the residual is positive, the observed point lies **above** the line.

Q4. The data cover x from 4 to 15. A value inside $[4, 15]$ gives interpolation, a value outside gives extrapolation. (a) 6 is inside — interpolation. (b) 10 is inside — interpolation. (c) $2<4$ — extrapolation. (d) $20>15$ — extrapolation.

Q5. The line is $y=12+0.5x$. (a) The slope 0.5 means the plant grows, on average, 0.5 cm for each additional month of age. (b) $\hat{y}=12+0.5\times 10=17$ cm. (c) $\hat{y}=12+0.5\times 40=32$ cm, but 40 months is beyond the largest observed age (30 months), so this is extrapolation and there is no evidence the linear growth continues that far — the prediction is unreliable.

Q6. Substitute each x into $\hat{y}=5+3x$: at $x=2$, $\hat{y}=11$; at $x=5$, $\hat{y}=20$; at $x=8$, $\hat{y}=29$. The residuals are observed $-$ predicted: $13-11=+2$, $19-20=-1$, and $29-29=0$. The $x=8$ point has residual 0, so the line passes exactly through it; the $x=2$ point has the largest residual in size ($+2$), so it is fitted worst.

Q7. The slope is 0.15. For each extra \$1 of weekly income, the model predicts, on average, an extra \$0.15 (15 cents) saved per week. Equivalently, the group saves about 15% of each additional dollar of income.

Q8. The intercept is -60: it is the saving the line predicts for a worker whose weekly income is \$0, namely $-\$60$. This is not meaningful, for two reasons — an income of \$0 lies outside the range of incomes in the study (extrapolation), and a saving of $-\$60$ has no sensible interpretation as a genuine amount saved.

Q9. The line is $y=38+2.5x$. (a) $\hat{y}=38+2.5\times10=63$, i.e. \$63,000. Since 10 years lies inside the observed range 1–25, this is interpolation and is reliable. (b) $\hat{y}=38+2.5\times40=138$, i.e. \$138,000. But 40 years is well beyond the largest observed value (25 years), so this is extrapolation and cannot be relied upon.

Q10. At $x=20$, the predicted value is $\hat{y}=100-2.5\times20=100-50=50$. The residual is $55-50=5$. Since the residual is positive, the observed point lies **above** the line.

Q11. At $x=20$, $\hat{y}=500-30\times20=500-600=-100$, i.e. a predicted price of $-\$100\,000$. A house cannot have a negative price, so the line has been pushed far past where it can be trusted: this illustrates why extrapolation — predicting well outside the range of the data — is unreliable.

Q12. At the $x=3$ point the line predicts $\hat{y}=4+3\times3=13$. The observed value is 11, so the residual is $11-13=-2$. The negative residual means the actual sales were 2 below the line, i.e. the line over-predicted this point.

Q13. The slope is -1.2. On average, the battery's capacity drops by 1.2 percentage points for each additional day it is left unused after being fully charged.

Q14. Set the predicted value equal to 65: $65=25+4x$, so $4x=40$ and $x=10$. Because $x=10$ lies inside the range of the data (1 to 15), this is interpolation and the prediction is reliable.

Q15. At $x=6$ the predicted value is $\hat{y}=7+2\times6=19$. Since residual = observed − predicted, the observed value is $y=\hat{y}+\text{residual}=19+(-3)=16$.

Q16. The line is $y=1.5+0.08x$. (a) The slope 0.08 means that, on average, 0.08 litres of fuel are used for each additional kilometre driven. (b) The intercept 1.5 is the predicted fuel used at $x=0$ km — the litres accounted for before any distance is driven (for example, starting and idling). (c) $\hat{y}=1.5+0.08\times200=1.5+16=17.5$ litres.

Q17. (a) A prediction is **interpolation** when the explanatory value lies inside the range of x-values in the data, and **extrapolation** when it lies outside that range. (b) Inside the data we have evidence that the linear relationship holds, so interpolation is generally reliable. Outside the data we have no such evidence: the true relationship may curve, level off, or change entirely, so an extrapolated prediction may be far from correct — and pushed far enough, the line can predict impossible values.

Q18. Substitute each x into $\hat{y}=2+1.5x$: the predicted values are $5, 8, 11, 14$. (a) The residuals are $6-5=+1$, $8-8=0$, $10-11=-1$ and $15-14=+1$. (b) The $x=4$ point has residual 0, so the line fits it exactly. (c) The line over-predicts where the residual is negative — the $x=6$ point (predicted 11, observed 10).

Q19. The line is $y=12+0.75x$. (a) The slope 0.75 means the tree grows, on average, 0.75 m per year. (b) The intercept 12 is the height the line predicts at age 0. Although age 0 is only just below the smallest observed age (2 years), a newly planted tree is certainly not 12 m tall, so $x=0$ makes no sense in this context — and failing that one condition is enough for the intercept not to be meaningful. (c) $\hat{y}=12+0.75\times20=12+15=27$ m. (d) The residual is $30-27=3$ m (the tree is 3 m taller than predicted).

Q20. The line is $y=2+1.5x$. (a) The slope 1.5 means the yield increases, on average, by 1.5 kg for each additional hour of daily sunlight. (b) $\hat{y}=2+1.5\times8=14$ kg; 8 lies in 3–10, so this is interpolation. (c) $\hat{y}=2+1.5\times15=24.5$ kg, but 15 hours is beyond the largest observed value (10), so this extrapolation is unreliable. (d) The observed yield is 13 kg, so the residual is $13-14=-1$ kg. The negative residual means the line **over-predicted** this plant's yield.

Theory

Earlier in this chapter we learned to **fit** a least-squares line to two numerical variables and to **interpret** its slope and intercept. This section draws those skills together into a single, repeatable procedure — a **regression analysis** — that takes you from a set of paired data all the way to a justified conclusion about the association. Just as importantly, it adds the step that decides whether a straight line was the right model at all: the **residual analysis**.

The explanatory and response variables. A regression analysis models one numerical variable from another. The **explanatory variable** (EV) is the one used to explain or predict, and is placed on the horizontal (x) axis; the **response variable** (RV) is the one being predicted, and is placed on the vertical (y) axis. Every step below depends on getting these roles the right way round, so identify them first.

The six steps of a regression analysis.

1. **Construct the scatterplot** of the response variable against the explanatory variable, and describe the association by its **direction** (positive or negative), **form** (linear or non-linear) and **strength** (strong, moderate or weak). Only if the form looks linear does a straight-line model make sense.
2. **Quantify the association** with the **correlation coefficient** r (its sign gives the direction, its size the strength) and the **coefficient of determination** r^2, which as a percentage gives the variation in the response variable explained by the explanatory variable.
3. **Fit the least-squares line** $\hat{y} = a + bx$.
4. **Interpret the slope and the intercept** in the context of the two variables, including their units, and use the line to make predictions where appropriate.
5. **Carry out a residual analysis** — construct the **residual plot** and use it to check the assumption that the association is linear.
6. **Write a conclusion** in the context of the data, stating the model, how much of the variation it explains, and whether a linear model is appropriate.

Step 1 — the scatterplot. With the explanatory variable on the horizontal axis, plot each pair as a point. A band of points trending upward is a **positive** association, a band trending downward is **negative**; points hugging a straight path are **linear** in form, while a bending band is **non-linear**; and a tight band is a **strong** association, a loose cloud a **weak** one.

Step 2 — r and r^2. The correlation coefficient r lies between -1 and 1; its sign is the direction and its size the strength of the **linear** association. Squaring gives the coefficient of determination r^2 (between 0 and 1), and $r^2 \times 100\,\%$ is the percentage of the variation in the response variable explained by the variation in the explanatory variable. In practice the CAS **linear-regression** command reports both r and r^2 at once. Where that output is available, read r^2 from it rather than squaring a quoted r: a quoted r has itself already been rounded, so squaring it can change the fourth decimal place of r^2.

Step 3 — the least-squares line. The least-squares line is written

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

where $\hat{y}$ (read "y-hat") is the **predicted** value of the response variable, b is the slope and a is the y-intercept. From the summary statistics the coefficients are

$$b = r\frac{s_y}{s_x}, a = \overline{y} - b\overline{x},$$

where $\overline{x}, \overline{y}$ are the means and s_x, s_y the sample standard deviations of the two variables. Because s_x and s_y are both positive, **the slope b always has the same sign as r**. In practice you read a and b straight from the CAS regression output; the formulas let you find the line from summary statistics and check that the technology is reasonable. Unless told otherwise we give a and b correct to **two decimal places** (and quote r, r^2 to four).

Step 4 — interpreting the coefficients. The interpretations must always name the variables and their units.

- The **slope** b is the change in the **response variable** for each **one-unit increase** in the explanatory variable. A positive slope means the response increases; a negative slope means it decreases.
- The **intercept** a is the predicted value of the response variable when the explanatory variable is **zero**. This is only meaningful when $x = 0$ is sensible and lies within (or close to) the range of the data; otherwise it is a mathematical starting value only.

The line can then **predict** a response for a given value of the explanatory variable by substitution. Predicting *within* the range of the data is **interpolation** (reliable); predicting *well beyond* the data is **extrapolation** and is unreliable, because there is no evidence the linear pattern continues.

Step 5 — residual analysis: the check of linearity. Fitting a straight line does not, by itself, prove the association *is* straight — a line can be fitted to any data. The **residual** of a data point measures how far the actual value sits above or below the line:

$$\text{residual} = y - \hat{y},$$

the observed value minus the value the line predicts. A point **above** the line has a **positive** residual; a point **below** it has a **negative** residual. (The least-squares line is built so that the residuals always add to zero.)

A **residual plot** graphs each residual against the corresponding value of the explanatory variable, with a horizontal line at 0. Its shape is the diagnostic:

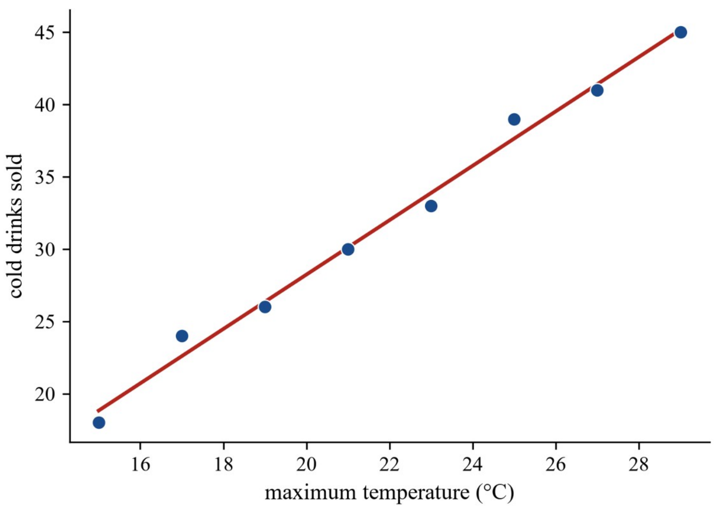
- if the residuals are scattered **randomly** in a band about 0 with **no clear pattern**, the straight line has captured the trend and a **linear model is appropriate**;
- if the residuals show a **clear pattern** — most often a **curve** (for example negative in the middle and positive at both ends, or the reverse) — then the association is really **non-linear** and a **linear model is *not* appropriate**, even if r^2 is high.

This last point is the reason the residual plot matters: a large r^2 measures how close the points are to the *line*, but it cannot reveal a gentle bend. The residual plot magnifies that bend. When the residual plot is clearly curved, the next step is to **transform** one of the variables to straighten the data and then repeat the regression — the subject of the next chapter.

Residual plot appearance	What it means	Action
random band about 0, no pattern	the association is linear	the linear model is appropriate; keep it
clear curved pattern	the association is non-linear	a linear model is not appropriate; transform the data

Step 6 — the conclusion. A good conclusion states the fitted line, interprets r^2 in context, and — crucially — reports the verdict of the residual plot: whether the linear model is an appropriate fit or whether the data should be transformed.

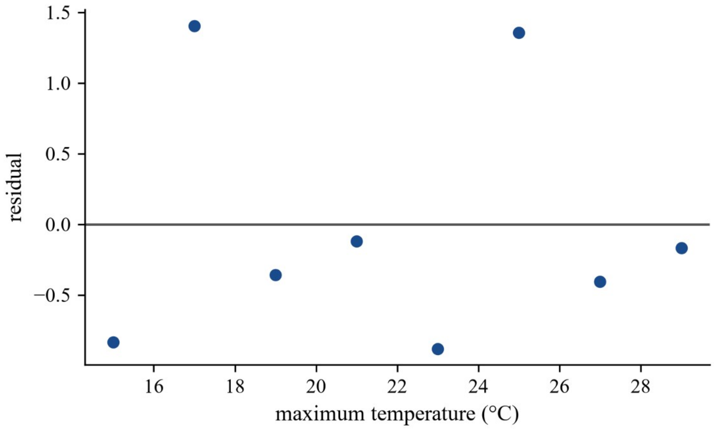
A worked illustration. The scatterplot below plots the number of cold drinks a kiosk sold against the daily maximum temperature (°C), the explanatory variable, for eight days, with the least-squares line drawn in.



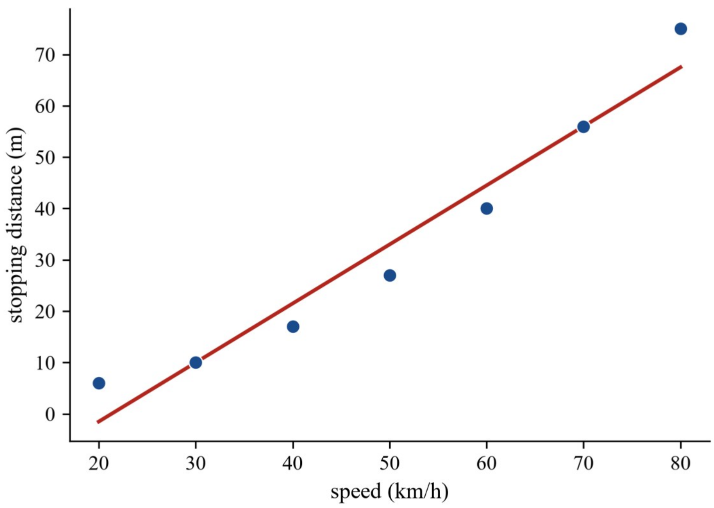
The association is strong, positive and linear ($r = 0.9953$, so $r^2 = 0.9906$, meaning about 99 % of the variation in drinks sold is explained by temperature), and the least-squares line is

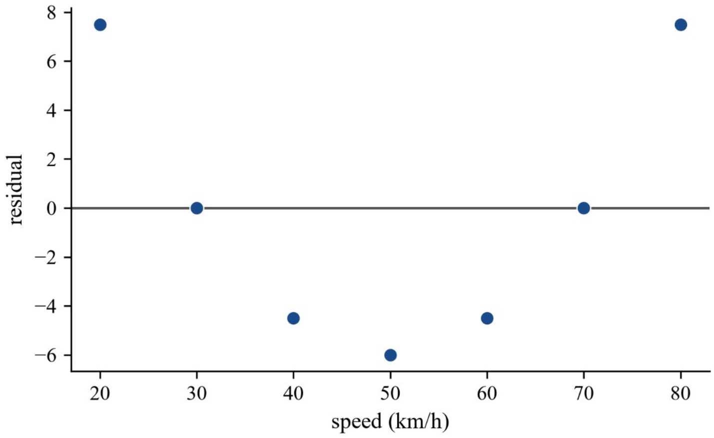
$$\hat{y} = -9.38 + 1.88\,x\,.$$

The slope says the kiosk sells about 1.88 more drinks for each extra degree; the intercept -9.38 is not meaningful, as 0 °C lies far below the observed temperatures. The residual plot is a **random band about** 0, confirming a linear model is appropriate.



Contrast this with a study of a car's stopping distance against its speed. Its scatterplot also looks almost straight and gives a high $r^2 = 0.9514$, yet its residual plot is a clear ∪**-shaped curve** — the residuals are positive at the low and high speeds and negative in between.





That curved pattern is the warning: despite the high r^2, stopping distance is **not** a linear function of speed, so a linear model is not appropriate here and the data would need to be transformed. The residual plot has caught what r^2 alone could not.

Using CAS. General Mathematics is a technology-active, open-book course. In practice you enter the two lists and run the **linear-regression** command: the CAS reports r, r^2 and the coefficients a and b, draws the scatterplot with the fitted line, and produces the residual plot directly. (The two-variable statistics screen is a separate command; it gives the means and standard deviations, not r or r^2.) Learn the by-hand formulas and the reading of a residual plot so that you can set the technology up correctly, check its output, and — above all — judge whether a straight line was the right model.

Worked examples

Example 1 — scaffolded

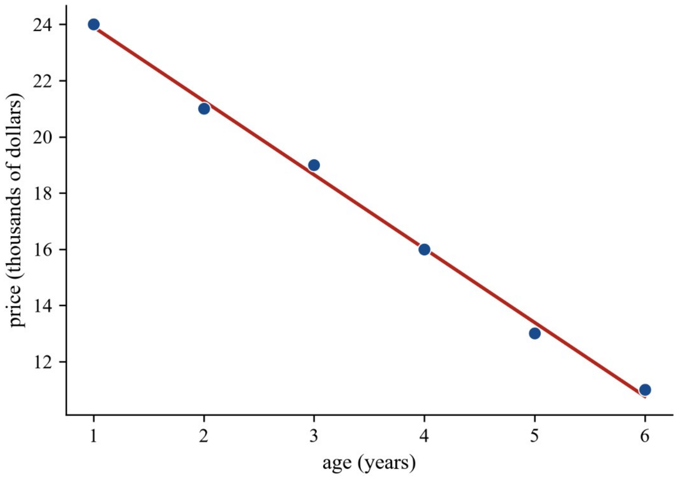
The age (in years) and price (in thousands of dollars) of six second-hand cars of the same model are recorded below. Carry out a full regression analysis: describe the association, find r^2 and the least-squares line, interpret its slope and intercept, check the residual plot, and conclude.

Age x (years)	1	2	3	4	5	6
Price y ($000)	24	21	19	16	13	11

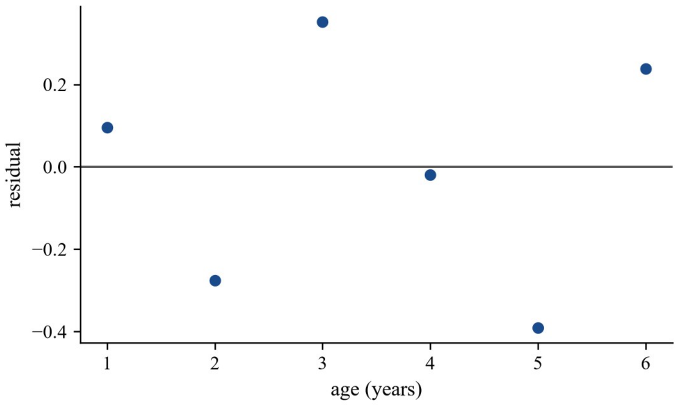
The CAS two-variable statistics give $\overline{x}=3.5$, $\overline{y}=17.33$, $s_x=1.87$ and $s_y=4.93$, and the CAS linear-regression command gives $r=-0.9983$.

Working	Comment
Explanatory variable = **age**; response variable = **price**.	Price is being predicted from age.
The scatterplot shows a strong, negative, linear association.	Points fall in a tight straight band.
The same regression screen reports $r^2=0.9965$.	Read r^2 off the CAS, not by squaring the rounded r; about $99.65\,\%$ of the variation in price is explained by age.
$b=r\dfrac{s_y}{s_x}=-0.9983\times\dfrac{4.93}{1.87}=-2.63.$	Slope has the same sign as r.
$a=\overline{y}-b\,\overline{x}=17.33-(-2.6286)(3.5)=26.53.$	Keep the full-precision $b=-2.6286$ here, not the rounded -2.63; the CAS value is 26.53.
Least-squares line: $\hat{y}=26.53-2.63\,x$.	Coefficients rounded to two decimal places.

Working	Comment
Slope: the price **falls by $2630** for each additional year of age.	$b=-2.63$ thousand dollars per year.
Intercept: a car of age 0 (brand new) is predicted to cost **$26 530**.	$x=0$ is just outside the data, so a mild extrapolation.



The residuals (observed − predicted) are $0.10, -0.27, 0.36, -0.01, -0.38, 0.25$, and the residual plot below scatters randomly about 0 with no pattern.



The random residual plot confirms the association is linear, so the model $\hat{y}=26.53-2.63x$ is an appropriate fit; it explains about 99.65% of the variation in price, and the price of this model of car falls by about $2630 per year of age.

Example 2 — scaffolded

A regression of test mark (response) on the number of hours studied (explanatory) for a class gave the summary statistics $\bar{x}=5$, $\bar{y}=60$, $s_x=2$, $s_y=15$ and $r=0.90$. Find the least-squares line, interpret its slope and intercept, and find the residual for a student who studied for 4 hours and scored 55.

Working	Comment
$b=r\dfrac{s_y}{s_x}=0.90\times\dfrac{15}{2}=6.75.$	Slope from the formula.
$a=\bar{y}-b\bar{x}=60-6.75\times5=26.25.$	Intercept from the means.
Least-squares line: $\hat{y}=26.25+6.75x$.	$\hat{y}$ is the predicted mark.
Slope: the mark **rises by** 6.75 for each extra hour studied.	Change in response per one-unit increase in x.
Intercept: a student who studies 0 hours is predicted to score 26.25.	$x=0$ interpretation (an extrapolation below the data).

Working	Comment
For $x = 4$: $\hat{y} = 26.25 + 6.75 \times 4 = 53.25$.	Predicted mark for that student.
Residual = $y - \hat{y} = 55 - 53.25 = 1.75$.	Positive, so the student scored **above** the line.

A positive residual means the actual mark (55) lies above the value the model predicts (53.25).

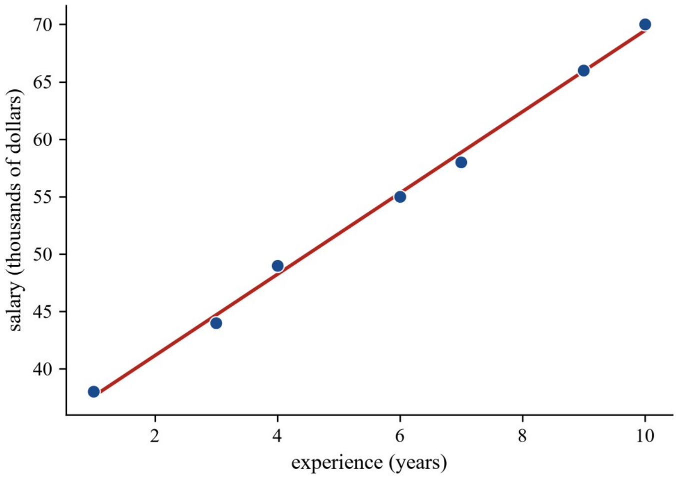
Example 3 — unscaffolded

For seven employees, the number of years of experience and the annual salary (in thousands of dollars) were recorded.

Experience x (years)	1	3	4	6	7	9	10
Salary y ($000)	38	44	49	55	58	66	70

Conduct a regression analysis and state, with reasons, whether a linear model is appropriate.

The explanatory variable is experience and the response variable is salary. The scatterplot shows a strong, positive, linear association.

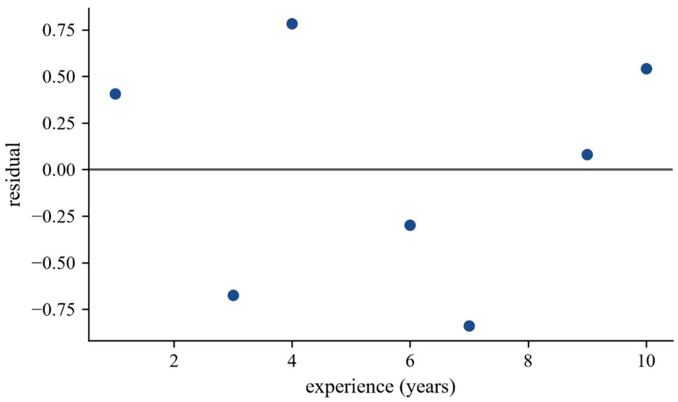


A CAS regression gives $r = 0.9985$ and $r^2 = 0.9971$: about $99.71\,\%$ of the variation in salary is explained by the variation in experience. The least-squares line is

$$\hat{y} = 34.05 + 3.54\,x.$$

The slope 3.54 means salary rises by about $3540 for each additional year of experience, and the intercept 34.05 predicts a starting salary (at 0 years) of about $34 050 — a slight extrapolation, as the data begin at 1 year.

The residual plot scatters randomly about 0 with no clear pattern.



Because the residual plot shows no pattern, the assumption of linearity is supported.

∴ a linear model is appropriate; $\hat{y}=34.05+3.54\,x$ explains about 99.71 % of the variation in salary, with salary increasing by about \$3540 per year of experience.

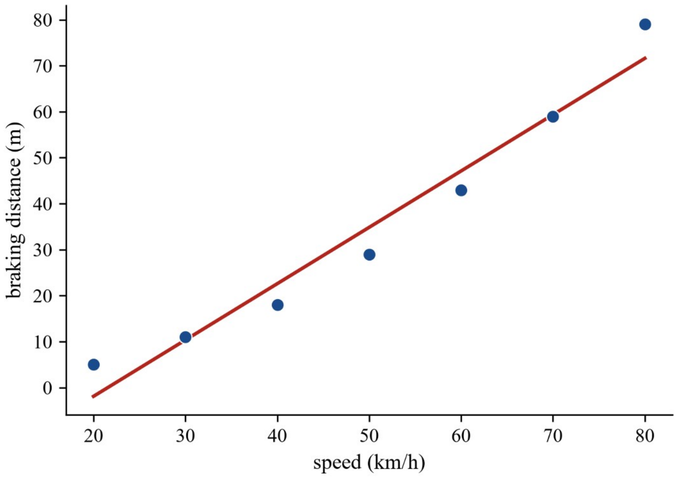
Example 4 — unscaffolded

The speed of a car (in km/h) and its braking distance (in metres) were measured on seven trials.

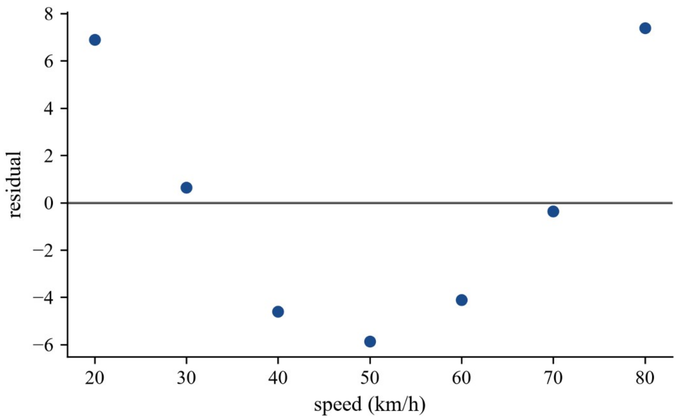
Speed x (km/h)	20	30	40	50	60	70	80
Braking distance y (m)	5	11	18	29	43	59	79

A least-squares line is fitted and gives $r^2=0.9600$ and $\hat{y}=-26.39+1.23\,x$. Using a residual analysis, decide whether a linear model is appropriate.

At first sight the fit looks excellent: $r^2=0.9600$ means about 96 % of the variation in braking distance is "explained" by the line, and the scatterplot below looks almost straight.



But the residual plot tells a different story. The residuals are **positive** at the lowest and highest speeds and **negative** for the middle speeds, tracing a clear **∪-shaped curve** rather than a random band.



This systematic pattern shows that the straight line over-predicts in the middle of the range and under-predicts at the ends — exactly what happens when a curved relationship is forced onto a line. The high r^2 measured only how close the points were to the line; it could not reveal the bend that the residual plot exposes.

∴ despite $r^2=0.9600$, the curved residual plot shows the association is **non-linear**, so a linear model is **not appropriate**; the data should be transformed to achieve linearity before modelling (the method of the next chapter).

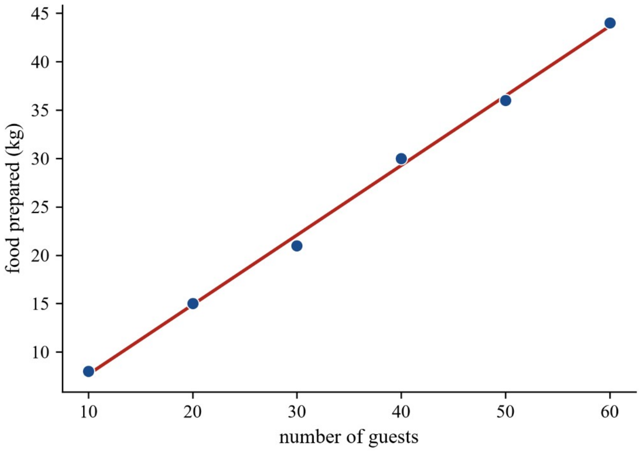
Practice questions

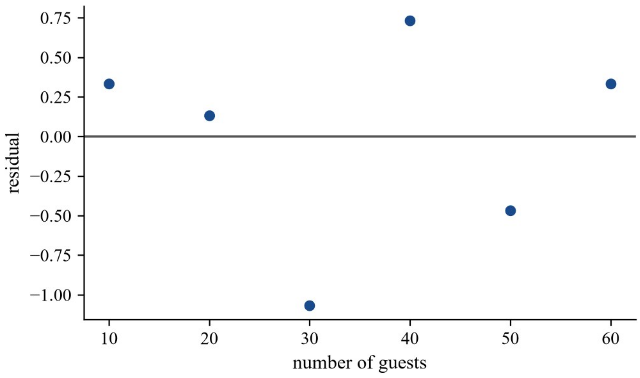
Scaffolded

Q1. The number of guests at a function and the mass of food (kg) prepared were recorded for six functions.

Guests x	10	20	30	40	50	60
Food y (kg)	8	15	21	30	36	44

A CAS regression gives $r = 0.9988$ and the least-squares line $\hat{y} = 0.47 + 0.72x$. (a) State the explanatory and response variables. (b) The scatterplot is shown below. Describe the association (direction, form, strength). (c) Find r^2 and interpret it as a percentage in context. (d) Interpret the slope of the least-squares line in context. (e) Interpret the intercept, and state whether it is meaningful here. (f) The residual plot is shown below. State whether a linear model is appropriate, with a reason.





Q2. A regression of a plant's height (response, cm) on the number of weeks since planting (explanatory) gave $\bar{x} = 5$, $\bar{y} = 40$, $s_x = 2$, $s_y = 8$ and $r = 0.9$. (a) Find the slope $b = r\dfrac{s_y}{s_x}$. (b) Find the intercept $a = \bar{y} - b\bar{x}$. (c) Write down the least-squares line. (d) Interpret the slope in context.

Q3. For six data pairs the least-squares line is $\hat{y} = 2.67 + 2.71x$.

x	1	2	3	4	5	6
y	5	9	10	14	16	19

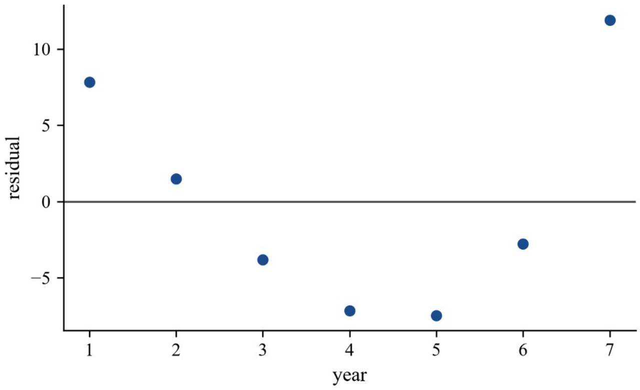
(a) Complete the table of predicted values $\hat{y}$ (to two decimal places).
(b) Find the residual $y - \hat{y}$ for each point.
(c) State whether the residuals show a clear pattern, and hence whether a linear model is appropriate.

Q4. For seven students, the hours studied and the test mark (out of 70) were recorded. A CAS regression gives $r = 0.9967$ and $\hat{y} = 17.96 + 6.04x$. (a) State the explanatory and response variables. (b) Find r^2 and interpret it in context. (c) Interpret the slope in context. (d) Interpret the intercept in context, and comment on whether it is meaningful. (e) Use the line to predict the mark of a student who studies for 5.5 hours. (f) The residual plot shows a random band about 0. State whether a linear model is appropriate.

Q5. The number of members of a new club at the end of each of its first seven years was recorded. A least-squares line $\hat{y} = -13.14 + 7.32x$ is fitted, with $r^2 = 0.8180$.

Year x	1	2	3	4	5	6	7
Members y	2	3	5	9	16	28	50

(a) Find the residual for each year (to two decimal places).
(b) Describe the pattern in the residual plot below.
(c) State, with a reason, whether a linear model is appropriate, and what should be done next.



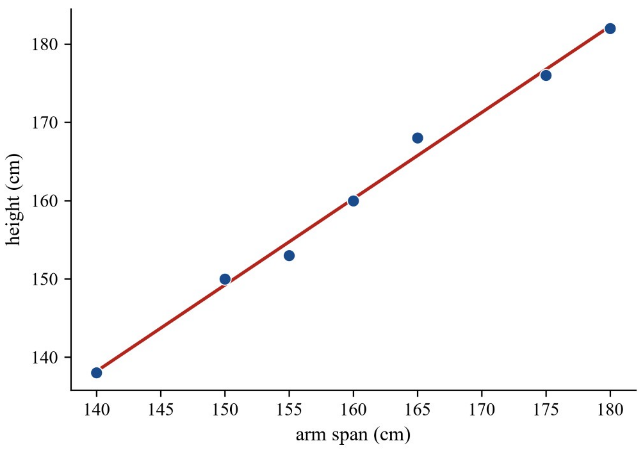
Q6. A firm models its weekly sales revenue y (in thousands of dollars) against its weekly advertising spend x (in thousands of dollars) by the least-squares line $\hat{y} = 12 + 3.5x$, with $r^2 = 0.72$. (a) Interpret the slope in context. (b) Interpret the intercept in context. (c) Interpret r^2 as a percentage in context. (d) Predict the revenue when \$8000 is spent on advertising. (e) Explain why using the line to predict the revenue for a spend of \$50 000 would be unreliable.

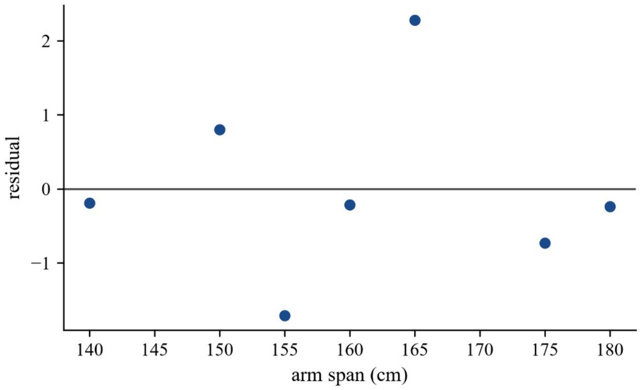
Unscaffolded

Q7. For seven people, the arm span and the height (both in cm) were measured.

Arm span x (cm)	140	150	155	160	165	175	180
Height y (cm)	138	150	153	160	168	176	182

A CAS regression gives $r = 0.9967$ and $\hat{y} = -15.98 + 1.10x$. Conduct a full regression analysis (describe the association, find and interpret r^2, interpret the slope and intercept, use the residual plot below to check linearity, and conclude).





Q8. The least-squares line relating a car's engine size x (litres) to its fuel consumption y (litres per 100 km) is $\hat{y}=4.2+3.1x$, with $r^2=0.81$. (a) Interpret the slope in context. (b) Interpret the intercept, and state whether it is physically meaningful. (c) Interpret r^2 as a percentage in context. (d) Predict the fuel consumption of a car with a 2.0-litre engine.

Q9. The daily maximum temperature (°C) and the number of hot coffees a café sold were recorded on seven days.

Temperature x (°C)	5	8	12	16	20	24	28
Hot coffees y	52	47	44	38	33	29	24

A CAS regression gives $r=-0.9981$. Find the least-squares line (to two decimal places), interpret its slope, and interpret its intercept.

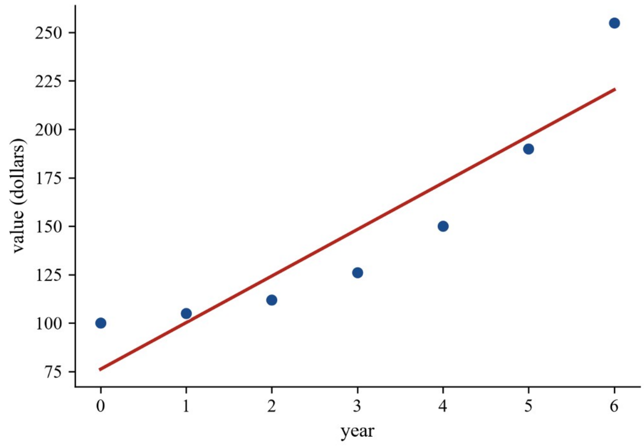
Q10. The least-squares line relating a house's floor area x (m²) to its price y (in thousands of dollars) is $\hat{y}=45+2.8x$, with $r^2=0.66$. (a) Interpret r^2 as a percentage in context. (b) Interpret the slope in context. (c) Predict the price of a house with a floor area of 150 m². (d) Explain why predicting the price of a house with a floor area of 1000 m² would be unreliable.

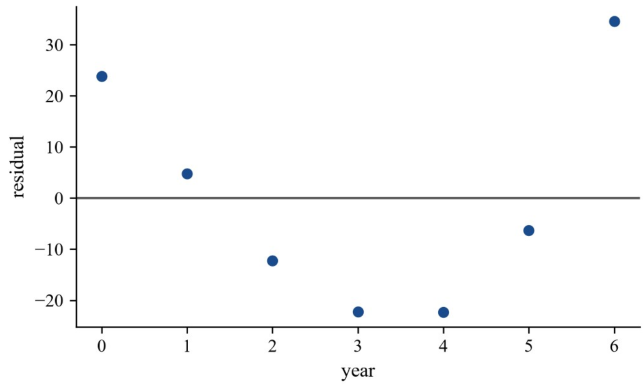
Q11. In your own words, explain (a) what a residual is, (b) what a residual plot shows, and (c) how the appearance of the residual plot is used to decide whether a linear model is appropriate.

Q12. The value of an investment (in dollars) at the end of each of seven years was recorded. A least-squares line $\hat{y}=76.18+24.04x$ is fitted, with $r^2=0.8447$.

Year x	0	1	2	3	4	5	6
Value y ($)	100	105	112	126	150	190	255

(a) State what $r^2=0.8447$ suggests about the fit, taken on its own.
(b) The scatterplot and residual plot are shown below. Describe the residual plot.
(c) State, with a reason, whether a linear model is appropriate, and what should be done next.





Q13. The number of hours per day a student watched television and their test mark were recorded, and the least-squares line $\hat{y} = 82.86 - 5.94\,x$ was fitted.

Hours x	0	1	2	3	4	5
Mark y	82	78	70	66	60	52

(a) Find the residual for the student who watched 2 hours.

(b) Find the residual for the student who watched 4 hours.

(c) State what the sign of a residual tells you about the position of a point relative to the line.

Q14. A regression of a machine's output y (items per hour) on its age x (years) gave $\bar{x} = 20$, $\bar{y} = 150$, $s_x = 5$, $s_y = 12$ and $r = -0.8$. (a) Find the slope and intercept of the least-squares line. (b) Write down the least-squares line. (c) Predict the output of a machine aged 25 years.

Q15. The amount of fertiliser applied (grams) and the yield of a crop (kg) were recorded for six plots.

Fertiliser x (g)	0	20	40	60	80	100
Yield y (kg)	24	32	38	47	53	60

A CAS regression gives $r = 0.9988$ and $\hat{y} = 24.33 + 0.36\,x$. (a) Find r^2 and interpret it in context. (b) Interpret the slope in context. (c) Interpret the intercept, and explain why here it *is* meaningful. (d) Predict the yield when 50 g of fertiliser is applied.

Q16. State whether each statement is **true** or **false**. Correct each false statement. (a) A residual is the vertical distance between a data point and the least-squares line. (b) If the residual plot shows a clear curved pattern, the linear model is a good fit. (c) r^2 gives the percentage of the variation in the response variable explained by the explanatory variable. (d) The slope b and the correlation coefficient r always have the same sign. (e) It is always safe to use the least-squares line to predict far beyond the range of the data.

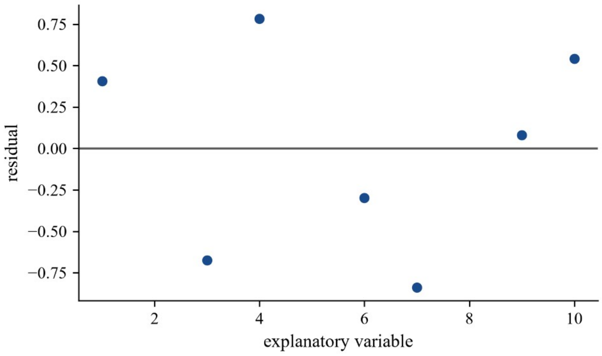
Q17. The least-squares line relating a person's height x (cm) to their weight y (kg) is $\hat{y} = -45 + 0.6x$. (a) Interpret the slope in context. (b) State the value the intercept gives, and explain why the intercept is not meaningful here.

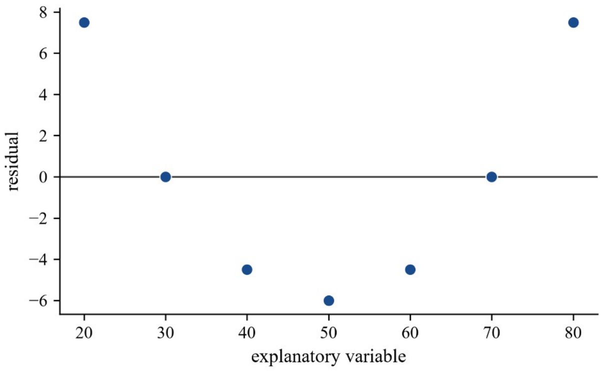
Q18. The age (years) and price ($000) of seven second-hand motorbikes were recorded.

Age x (years)	2	3	4	5	6	7	8
Price y ($000)	28	25	22	18	16	12	9

A CAS regression gives $r = -0.9986$ and $\hat{y} = 34.46 - 3.18x$. (a) Find and interpret r^2 in context. (b) Interpret the slope in context. (c) Predict the price of a 10-year-old motorbike, and state why this prediction should be treated with caution.

Q19. Two students each fit a least-squares line to a different data set. Their residual plots are shown below: plot A scatters randomly about 0, while plot B traces a clear curve.



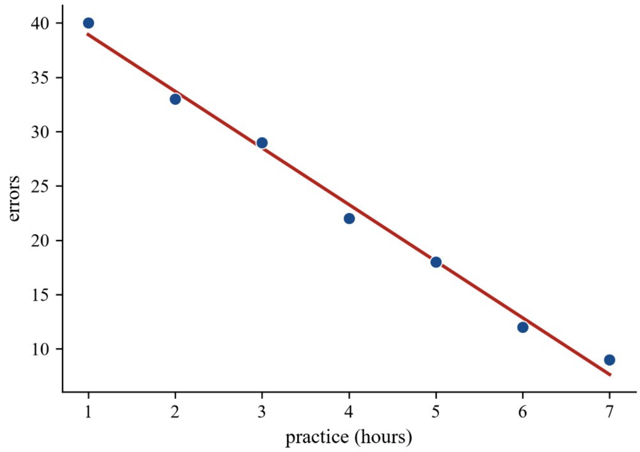


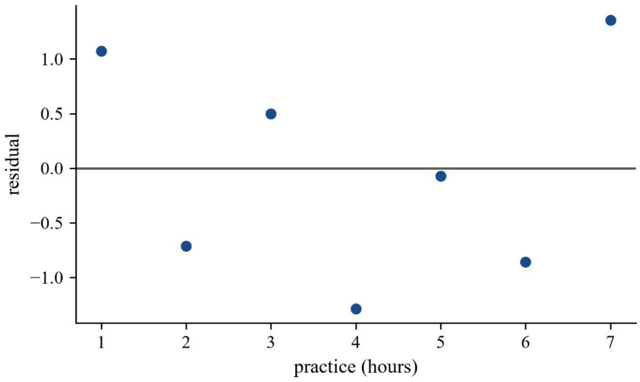
(a) For which data set is a linear model appropriate?
(b) State what plot B tells you about that data set's association.
(c) State what the second student should do next.

Q20. In a keyboard course, the number of hours each of seven students practised and the number of errors they made in a typing test were recorded.

Practice x (hours)	1	2	3	4	5	6	7
Errors y	40	33	29	22	18	12	9

A CAS regression gives $r = -0.9960$ and $\hat{y} = 44.14 - 5.21x$. Write a full regression report: describe the association, find and interpret r^2, interpret the slope and intercept, predict the errors for 5 hours of practice, use the residual plot below to check linearity, and conclude.


errors
40
35
30
25
20
15
10
1
2
3
4
5
6
7
practice (hours)


residual
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
7
practice (hours)

Answers

Q1. (a) EV guests, RV food; (b) strong, positive, linear; (c) $r^2 = 0.9977$, about $99.77\,\%$ of the variation in the mass of food is explained by the number of guests; (d) about 0.72 kg more food per additional guest; (e) 0.47 kg predicted for 0 guests — not really meaningful (no food needed for no guests, and $x = 0$ is below the data); (f) random band about 0, so a linear model is appropriate. **Q2.** (a) $b = 3.6$; (b) $a = 22$; (c) $\hat{y} = 22 + 3.6\,x$; (d) the height increases by 3.6 cm per extra week since planting. **Q3.** (a) $5.38, 8.09, 10.80, 13.51, 16.22, 18.93$; (b) $-0.38, 0.91, -0.80, 0.49, -0.22, 0.07$; (c) no clear pattern, so a linear model is appropriate. **Q4.** (a) EV hours studied, RV test mark; (b) $r^2 = 0.9934$, about $99.34\,\%$ of the variation in the mark is explained by hours studied; (c) the mark rises by about 6.04 per extra hour studied; (d) 17.96 predicted for 0 hours (an extrapolation below the data, but plausible as a base mark); (e) $\hat{y} = 17.96 + 6.04 \times 5.5 \approx 51$ marks; (f) yes — the random band supports linearity. **Q5.** (a) $7.82, 1.50, -3.82, -7.14, -7.46, -2.78, 11.90$; (b) a clear ∪-shaped curve (positive at the ends, negative in the middle); (c) not appropriate — the curved pattern shows the association is non-linear; the data should be transformed to achieve linearity. **Q6.** (a) revenue rises by \$3500 for each extra \$1000 of advertising; (b) with no advertising ($x = 0$), predicted revenue is \$12 000; (c) $72\,\%$ of the variation in revenue is explained by advertising spend; (d) $\hat{y} = 12 + 3.5 \times 8 = 40$, i.e. \$40 000; (e) \$50 000 is far beyond the data, so it is extrapolation and unreliable. **Q7.** Strong, positive, linear; $r^2 = 0.9934$ (about $99.34\,\%$ of the variation in height explained by arm span); $\hat{y} = -15.98 + 1.10\,x$; slope: height rises 1.10 cm per cm of arm span; intercept -15.98 cm is not meaningful (an arm span of 0 is impossible and far below the data); residual plot random, so a linear model is appropriate. **Q8.** (a) fuel consumption rises by 3.1 L/100 km per extra litre of engine size; (b) 4.2 L/100 km for a 0-litre engine — not physically meaningful (no engine has size 0, and $x = 0$ is outside the data); (c) $81\,\%$ of the variation in fuel consumption is explained by engine size; (d) $\hat{y} = 4.2 + 3.1 \times 2.0 = 10.4$ L/100 km. **Q9.** $\hat{y} = 57.53 - 1.20\,x$; slope: about 1.20 fewer hot coffees sold per extra degree; intercept: 57.53 coffees predicted at 0 °C (an extrapolation below the data). **Q10.** (a) $66\,\%$ of the variation in price is explained by floor area; (b) price rises by \$2800 per extra m²; (c) $\hat{y} = 45 + 2.8 \times 150 = 465$, i.e. \$465 000; (d) 1000 m² is far beyond the data, so it is extrapolation and unreliable. **Q11.** (a) a residual is the observed value minus the predicted value, $y - \hat{y}$, i.e. the vertical gap from a point to the line; (b) it plots each residual against x, with a line at 0; (c) a random band about 0 means the linear model is appropriate, while a clear (usually curved) pattern means the association is non-linear and a linear model is not appropriate. **Q12.** (a) taken alone, $r^2 = 0.8447$ suggests the line explains about $84\,\%$ of the variation — a seemingly good fit; (b) a clear ∪-shaped curve (positive at the ends, negative in the middle); (c) not appropriate — the curved residual plot shows the association is non-linear; the data should be transformed and the regression repeated. **Q13.** (a) $\hat{y} = 82.86 - 5.94 \times 2 = 70.98$, residual $70 - 70.98 = -0.98$; (b) $\hat{y} = 82.86 - 5.94 \times 4 = 59.10$, residual $60 - 59.10 = 0.90$; (c) a positive residual means the point lies above the line (the model under-predicts); a negative residual means it lies below the line (the model over-predicts). **Q14.** (a) $b = -1.92$, $a = 188.4$; (b) $\hat{y} = 188.4 - 1.92\,x$; (c) $\hat{y} = 188.4 - 1.92 \times 25 = 140.4$ items per hour. **Q15.** (a) $r^2 = 0.9977$, about $99.77\,\%$ of the variation in yield is explained by the amount of fertiliser; (b) yield rises by 0.36 kg per extra gram of fertiliser; (c) 24.33 kg is the predicted yield with no fertiliser ($x = 0$), which is meaningful because 0 g is an actual value in the data; (d) $\hat{y} = 24.33 + 0.36 \times 50 = 42.33$ kg. **Q16.** (a) True; (b) False — a curved pattern means a linear model is *not* appropriate; (c) True; (d) True; (e) False — predicting far beyond the data is extrapolation and is unreliable. **Q17.** (a) weight rises by 0.6 kg per extra cm of height; (b) the intercept is -45 kg (the predicted weight at height 0 cm), which is not meaningful — a height of 0 cm is impossible and lies far outside the data, and a negative weight is impossible. **Q18.** (a) $r^2 = 0.9971$, about $99.71\,\%$ of the variation in price is explained by age; (b) the price falls by \$3180 per year of age; (c) $\hat{y} = 34.46 - 3.18 \times 10 = 2.66$, i.e. \$2660, but 10 years is beyond the data (which stop at 8), so this is extrapolation and should be treated with caution. **Q19.** (a) the data set for plot A; (b) plot B's clear curve shows that association is non-linear, so a linear model is not appropriate; (c) transform the data to achieve linearity, then repeat the regression on the transformed data. **Q20.** Strong, negative, linear; $r^2 = 0.9920$ (about $99.20\,\%$ of the variation in errors explained by hours of practice); $\hat{y} = 44.14 - 5.21\,x$; slope: about 5.21 fewer errors per extra hour of practice; intercept: 44.14 errors predicted at 0 hours (an extrapolation just below the data); prediction $\hat{y} = 44.14 - 5.21 \times 5 \approx 18$ errors; residual plot random, so a linear model is appropriate.

Fully-worked solutions

Q1. (a) The mass of food is predicted from the number of guests, so **guests** is the explanatory variable and **food** the response. (b) The points rise in a tight, straight band, so the association is **strong, positive and linear**. (c) The regression screen reports $r^2 = 0.9977$ (read it from the CAS rather than squaring the rounded r), so about $99.77\,\%$ **of the variation in the mass of food is explained by the number of guests**. (d) The slope 0.72 means about 0.72 **kg more food is needed for each additional guest**. (e) The intercept 0.47 is the predicted food for 0 guests; it is not really meaningful (no guests should need no food, and $x = 0$ sits below the data). (f) The residual plot is a **random band about** 0 with no pattern, so a **linear model is appropriate**.

Q2. (a) $b = r\dfrac{s_y}{s_x} = 0.9 \times \dfrac{8}{2} = 0.9 \times 4 = 3.6$. (b) $a = \overline{y} - b\,\overline{x} = 40 - 3.6 \times 5 = 40 - 18 = 22$. (c) The least-squares line is $\hat{y} = 22 + 3.6\,x$. (d) The slope 3.6 means the plant's height increases by about 3.6 **cm for each extra week** since planting.

Q3. (a) Substituting each x into $\hat{y} = 2.67 + 2.71\,x$ gives predicted values $5.38, 8.09, 10.80, 13.51, 16.22, 18.93$. (b) The residuals $y - \hat{y}$ are $5 - 5.38 = -0.38$, $9 - 8.09 = 0.91$, $10 - 10.80 = -0.80$, $14 - 13.51 = 0.49$, $16 - 16.22 = -0.22$ and $19 - 18.93 = 0.07$. (c) These residuals alternate in sign and show **no clear pattern**, so the residual plot is a random band about 0 and a **linear model is appropriate**.

Q4. (a) The mark is predicted from the hours studied, so **hours studied** is explanatory and the **test mark** is the response. (b) $r^2 = 0.9967^2 = 0.9934$, so about $99.34\,\%$ **of the variation in the mark is explained by the hours studied**. (c) The slope 6.04 means the mark rises by about 6.04 **marks per extra hour** studied. (d) The intercept 17.96 is the predicted mark for 0 hours of study; since the data do not include 0 hours this is a mild extrapolation, though it is a plausible "base" mark. (e) For $x = 5.5$, $\hat{y} = 17.96 + 6.04 \times 5.5 = 17.96 + 33.22 = 51.18 \approx 51$ marks. (f) The residual plot is a **random band about** 0, which supports the assumption of linearity, so a **linear model is appropriate**.

Q5. (a) Substituting each year into $\hat{y} = -13.14 + 7.32\,x$ and subtracting from y gives residuals $7.82, 1.50, -3.82, -7.14, -7.46, -2.78, 11.90$. (b) These are **positive at the two ends and negative through the middle**, so the residual plot traces a clear **$\cup$-shaped curve**. (c) A curved residual plot shows the association is **non-linear**, so a **linear model is not appropriate**; the data should be **transformed to achieve linearity** and the regression repeated.

Q6. (a) The slope 3.5 (in thousands of dollars per thousand dollars) means revenue rises by **$3500 for each extra $1000 spent on advertising**. (b) The intercept 12 predicts a revenue of **$12 000 when nothing is spent** on advertising ($x = 0$). (c) $r^2 = 0.72$, so $72\,\%$ **of the variation in revenue is explained by advertising spend**. (d) With $8000 spent, $x = 8$ and $\hat{y} = 12 + 3.5 \times 8 = 12 + 28 = 40$, i.e. **$40 000**. (e) A spend of $50 000 ($x = 50$) is far outside the range of the data, so predicting there is **extrapolation** and the linear pattern cannot be assumed to continue.

Q7. Arm span is the explanatory variable and height the response. The scatterplot shows a **strong, positive, linear** association. Squaring the correlation, $r^2 = 0.9967^2 = 0.9934$, so about $99.34\,\%$ **of the variation in height is explained by arm span**. The least-squares line is $\hat{y} = -15.98 + 1.10\,x$: the slope means height increases by about 1.10 **cm per cm of arm span**, and the intercept -15.98 cm is **not meaningful** (an arm span of 0 cm is impossible and lies far below the data). The residual plot is a **random band about** 0, so the assumption of linearity holds. $\therefore$ a **linear model is appropriate**, explaining about $99.34\,\%$ of the variation in height.

Q8. (a) The slope 3.1 means fuel consumption rises by 3.1 **L/**100 **km for each extra litre of engine size**. (b) The intercept 4.2 is the predicted consumption for a 0-litre engine; it is **not physically meaningful**, since no car has a 0-litre engine and $x = 0$ is outside the data. (c) $r^2 = 0.81$, so $81\,\%$ **of the variation in fuel consumption is explained by engine size**. (d) For a 2.0-litre engine, $\hat{y} = 4.2 + 3.1 \times 2.0 = 4.2 + 6.2 = 10.4$ **L/**100 **km**.

Q9. The slope has the sign of r, so it is negative. Using the given line, $\hat{y} = 57.53 - 1.20\,x$. The slope -1.20 means the café sells about 1.20 **fewer hot coffees for each extra degree** of maximum temperature. The intercept 57.53 predicts about 58 **hot coffees at** 0 **°C**; since the coldest day in the data is 5 °C, this is a (mild) extrapolation.

Q10. (a) $r^2 = 0.66$, so $66\,\%$ **of the variation in price is explained by floor area**. (b) The slope 2.8 (thousands of dollars per m²) means price rises by **$2800 per extra m²**. (c) For 150 m², $\hat{y} = 45 + 2.8 \times 150 = 45 + 420 = 465$, i.e. **$465 000**. (d) A floor area of 1000 m² is far larger than any house in the data, so predicting its price is **extrapolation** and unreliable — there is no evidence the straight-line trend continues that far.

Q11. (a) A **residual** is the difference between an observed value and the value the line predicts, $\text{residual} = y - \hat{y}$; graphically it is the vertical distance from the data point to the least-squares line (positive above the line, negative below). (b) A **residual plot** graphs each residual against the corresponding x-value, with a horizontal line at 0. (c) If the residuals scatter **randomly** in a band about 0 with **no pattern**, the straight line has captured the trend and a **linear model is appropriate**; if they form a **clear pattern** — typically a curve — the association is **non-linear** and a linear model is **not appropriate**.

Q12. (a) Taken on its own, $r^2 = 0.8447$ suggests the line explains about $84\,\%$ of the variation, which looks like a good fit. (b) Substituting into $\hat{y} = 76.18 + 24.04\,x$ and subtracting from y gives residuals that are **positive at the ends and negative in the middle** — a clear **∪-shaped curve**. (c) Because the residual plot is clearly **curved**, the association is **non-linear** and a **linear model is not appropriate**, despite the high r^2; the data should be **transformed to achieve linearity** and the regression repeated. (This is the situation where r^2 alone would mislead — the residual plot reveals the bend.)

Q13. (a) For $x = 2$, $\hat{y} = 82.86 - 5.94 \times 2 = 82.86 - 11.88 = 70.98$, so the residual is $70 - 70.98 = -0.98$. (b) For $x = 4$, $\hat{y} = 82.86 - 5.94 \times 4 = 82.86 - 23.76 = 59.10$, so the residual is $60 - 59.10 = 0.90$. (c) A **positive** residual means the point lies **above** the line (the model under-predicts); a **negative** residual means it lies **below** the line (the model over-predicts).

Q14. (a) $b = r\dfrac{s_y}{s_x} = -0.8 \times \dfrac{12}{5} = -0.8 \times 2.4 = -1.92$, and $a = \bar{y} - b\,\bar{x} = 150 - (-1.92)(20) = 150 + 38.4 = 188.4$. (b) The least-squares line is $\hat{y} = 188.4 - 1.92\,x$. (c) For $x = 25$, $\hat{y} = 188.4 - 1.92 \times 25 = 188.4 - 48 = 140.4$ **items per hour**.

Q15. (a) The regression screen reports $r^2 = 0.9977$ (read it from the CAS rather than squaring the rounded r), so about $99.77\,\%$ **of the variation in yield is explained by the amount of fertiliser**. (b) The slope 0.36 means the yield rises by about 0.36 **kg per extra gram of fertiliser**. (c) The intercept 24.33 is the predicted yield when **no fertiliser** is applied; here it **is meaningful**, because $x = 0$ g is an actual value in the data (the first plot), so it is interpolation rather than extrapolation. (d) For 50 g, $\hat{y} = 24.33 + 0.36 \times 50 = 24.33 + 18 = 42.33$ **kg**.

Q16. (a) **True** — the residual $y - \hat{y}$ is exactly the vertical distance from a point to the line. (b) **False** — a clear curved pattern in the residual plot means the association is non-linear, so a linear model is **not** a good fit. (c) **True** — r^2, as a percentage, is the variation in the response explained by the explanatory variable. (d) **True** — since $b = r\,s_y / s_x$ and $s_x, s_y > 0$, the slope and r share the same sign. (e) **False** — predicting far beyond the data is extrapolation, which is unreliable.

Q17. (a) The slope 0.6 means weight rises by about 0.6 **kg for each extra cm of height**. (b) The intercept is -45 kg, the predicted weight at a height of 0 cm. It is **not meaningful**: a height of 0 cm is impossible and lies far outside the range of real heights, and it gives an impossible negative weight — the line simply should not be extended to $x = 0$.

Q18. (a) The regression screen reports $r^2 = 0.9971$ (read it from the CAS rather than squaring the rounded r), so about $99.71\,\%$ **of the variation in price is explained by age**. (b) The slope -3.18 (thousands of dollars per year) means the price **falls by about $3180 for each additional year** of age. (c) For a 10-year-old motorbike, $\hat{y} = 34.46 - 3.18 \times 10 = 34.46 - 31.8 = 2.66$, i.e. **$2660**; however 10 years is beyond the oldest motorbike in the data (8 years), so this is **extrapolation** and the prediction should be treated with caution.

Q19. (a) A linear model is appropriate for the data set giving **plot A**, whose residuals scatter randomly about 0 with no pattern. (b) Plot B's residuals form a **clear curve**, which shows that data set's association is **non-linear**, so a straight-line model does not fit it well. (c) The second student should **transform the data** (for example by squaring, taking logarithms, or using reciprocals of one variable) to straighten the association, then **repeat the regression** on the transformed data.

Q20. The number of errors is the response and the hours of practice the explanatory variable. The scatterplot shows a **strong, negative, linear** association. Squaring the correlation, $r^2 = (-0.9960)^2 = 0.9920$, so about $99.20\,\%$ **of the variation in the number of errors is explained by the hours of practice**. The least-squares line is $\hat{y} = 44.14 - 5.21\,x$: the slope -5.21 means students make about 5.21 **fewer errors for each extra hour of practice**, and the intercept 44.14 predicts about 44 **errors with no practice** ($x = 0$, a slight extrapolation below the data). Predicting for 5 hours, $\hat{y} = 44.14 - 5.21 \times 5 = 44.14 - 26.05 = 18.09 \approx 18$ **errors**. The residual plot is a **random band about** 0 with no pattern, so the assumption of linearity is supported. $\therefore$ a **linear model is appropriate**: it explains about $99.20\,\%$ of the variation in errors, with errors falling by about 5.21 per hour of practice.