General Mathematics Units 3 & 4

Chapter 4 — Data transformation

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Attribution: Classbasics Pty Ltd, vwx.au

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

Theory

Earlier, when we fitted a **least-squares line** to two numerical variables, we assumed the association was **linear** — that the points on the scatterplot were scattered about a straight line. Many real associations are not linear: the points follow a smooth **curve**. Fitting a straight line to curved data gives a poor model, and its **residual plot** gives the warning — the residuals form a clear pattern (often a ∪ or ∩ shape) instead of a random band about zero.

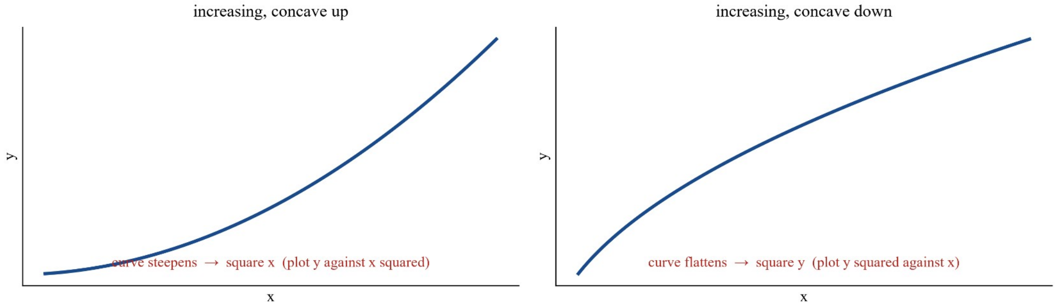
This chapter develops a way to deal with curved data. Rather than fit a curve, we **transform** one of the variables — apply the same function to every value of that variable — so that the **transformed** data lie along a straight line. We may then fit an ordinary least-squares line to the transformed data and use it to make predictions. This section studies the **squared transformation**.

One axis only. A data transformation is always applied to **one variable only** — either every x-value is replaced by x^2, or every y-value is replaced by y^2, but never both at once. The untouched variable keeps its original values. The three transformations in this course are the **square**, the **logarithm** (base 10) and the **reciprocal**; this section is the squared transformation, $x \to x^2$ or $y \to y^2$.

Why squaring straightens a curve. Squaring **stretches the upper end** of a scale and leaves the lower end almost unchanged. The values $1, 2, 3, 4, 5$ become $1, 4, 9, 16, 25$: the gap between the first two grows from 1 to 3, while the gap between the last two grows from 1 to 9. So squaring pulls the large values further apart than the small ones. If a curve bends because its large values are “bunched up”, squaring the right variable spreads those values out and straightens the curve.

Deciding which variable to square. We only transform an association that is **steadily increasing (or steadily decreasing) but curved** — that is, **monotonic** with a definite bend. For an increasing association, read the **concavity** of the curve:

- if the scatter is **concave up** — it bends **upward**, so y increases at an *increasing* rate — square the **explanatory** variable x, and fit
$$y = a + b\,x^2 \text{ (plot } y \text{ against } x^2\text{);}$$
- if the scatter is **concave down** — it bends **downward**, so y increases at a *decreasing* rate and flattens off — square the **response** variable y, and fit
$$y^2 = a + b\,x \text{ (plot } y^2 \text{ against } x\text{).}$$



In both cases the squared transformation stretches the axis on which the values are bunched: a concave-up curve has its high x-values bunched, so we stretch the x-axis (square x); a concave-down curve has its high y-values bunched, so we stretch the y-axis (square y). If you are unsure which will work, **try one and check**: the correct transformation makes the transformed scatterplot look linear and leaves **no pattern** in its residual plot.

Fitting the line to the transformed data. Once the variable to square is chosen, build a new column of transformed values and treat it as an ordinary variable:

- to fit $y = a + b\,x^2$, compute x^2 for every point and fit the least-squares line of y against x^2 (so x^2 plays the role of the explanatory variable);

- to fit $y^2 = a + bx$, compute y^2 for every point and fit the least-squares line of y^2 against x (so y^2 plays the role of the response variable).

In practice the coefficients a and b come from your CAS: enter the original data, create the squared list, and run the regression on the transformed list against the other variable — exactly the two-variable (bivariate) statistics routine used for an ordinary least-squares line.

Making predictions. Substitute into the fitted equation.

- For $y = a + bx^2$ the response y is **not** transformed, so substitute the given x, square it, and read off y directly.
- For $y^2 = a + bx$ the response **is** transformed. Substituting x gives a value of y^2, the *square* of the prediction. You must **back-transform**: take the (positive) **square root** to recover the predicted y.

Forgetting to back-transform is the most common error with a squared response: the number the equation returns is y^2, not y.

Checking the fit with a residual plot. A high correlation for the transformed data is encouraging, but the decisive check is the **residual plot** of the transformed fit. If the residuals are scattered randomly in a band about zero, the straight-line model for the transformed data is appropriate. If a curved pattern remains, the squared transformation has not fully straightened the data and another transformation should be tried.

Using CAS. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (which may be annotated). Your CAS will create a transformed list (an x^2 or y^2 column) with a single formula, fit the least-squares line, and report r and r^2 and the residuals. Learn the by-hand ideas — choosing the variable from the curve's shape, substituting to predict, and back-transforming a squared response — so that you can set the calculation up correctly and interpret what the CAS returns.

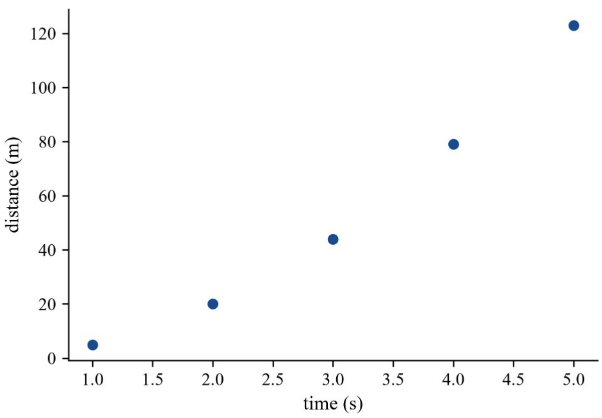
Worked examples

Example 1 — scaffolded

An object is released from rest and the distance d (in metres) it has fallen is recorded at times t (in seconds). The scatterplot of d against t is curved: it increases at an increasing rate (concave up). The data are shown below.

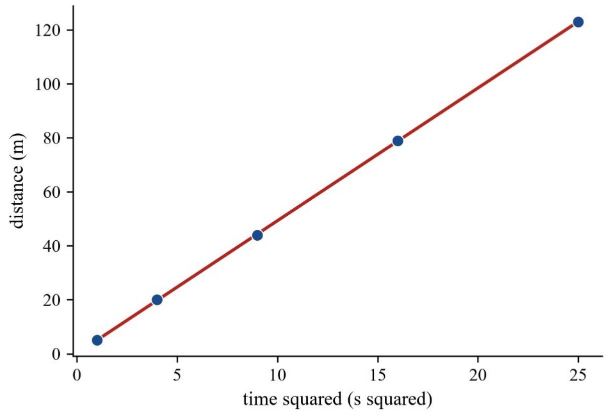
t (s)	1	2	3	4	5
d (m)	5	20	44	79	123

Use a squared transformation to fit a straight-line model, and predict the distance fallen at $t = 6$ s.



Working	Comment
The curve is increasing and **concave up**, so square the explanatory variable: fit $d = a + bt^2$.	Concave up ⇒ square x (here $x = t$); stretch the horizontal axis.

Working	Comment
Form the t^2 column: $1, 4, 9, 16, 25$.	Replace each t by t^2; leave d unchanged.
Plotting d against t^2 gives a straight line. A CAS least-squares fit gives $d = 0.11 + 4.92\,t^2$.	Fit the line of d against t^2; $r \approx 1.000$, so the transformed data are almost perfectly linear.
Predict at $t = 6$: $d = 0.11 + 4.92 \times 6^2 = 0.11 + 4.92 \times 36 = 177.23$ m.	d was not transformed, so substitute and read d directly — no back-transform.



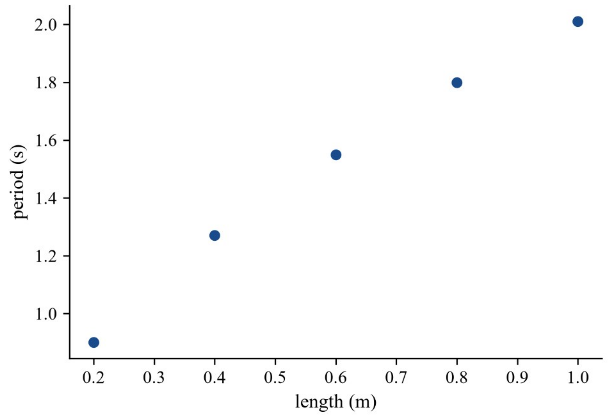
The object is predicted to fall about 177 m in the first 6 seconds.

Example 2 — scaffolded

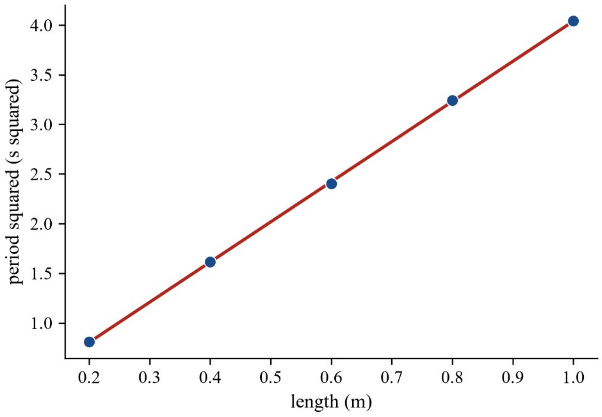
The period T (in seconds) of a pendulum is measured for several lengths L (in metres). The scatterplot of T against L increases but **flattens off** (concave down).

L (m)	0.2	0.4	0.6	0.8	1.0
T (s)	0.90	1.27	1.55	1.80	2.01

Use a squared transformation to fit a straight-line model, and predict the period when $L = 0.5$ m.



Working	Comment
The curve is increasing and **concave down**, so square the response variable: fit $T^2 = a + b\,L$.	Concave down ⇒ square y (here $y = T$); stretch the vertical axis.
Form the T^2 column: $0.81, 1.61, 2.40, 3.24, 4.04$.	Replace each T by T^2; leave L unchanged.
Plotting T^2 against L gives a straight line. A CAS least-squares fit gives $T^2 = -0.01 + 4.04\,L$.	Fit the line of T^2 against L; $r \approx 1.000$.
Predict at $L = 0.5$: $T^2 = -0.01 + 4.04 \times 0.5 = 2.01$.	Substituting gives T^2, the **square** of the period.
Back-transform: $T = \sqrt{2.01} \approx 1.42$ s.	The response was squared, so take the square root to recover T.



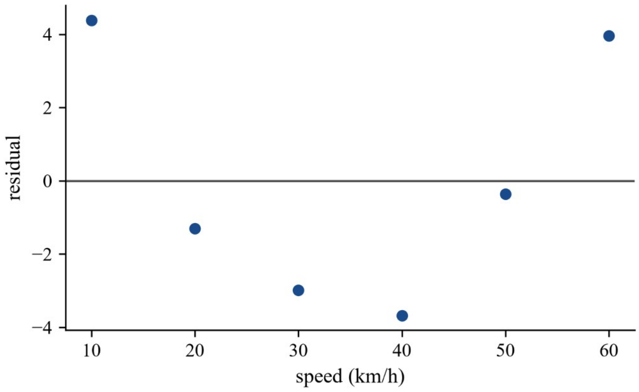
The pendulum of length 0.5 m is predicted to have a period of about 1.42 s.

Example 3 — unscaffolded

The braking distance d (in metres) of a car is recorded at several speeds v (in km/h).

v (km/h)	10	20	30	40	50	60
d (m)	2	5	12	20	32	45

A least-squares line was first fitted to the **raw** data. Its residual plot is shown below. Explain what the residual plot indicates, apply a suitable squared transformation, and use the transformed model to predict the braking distance at $v = 70$ km/h.

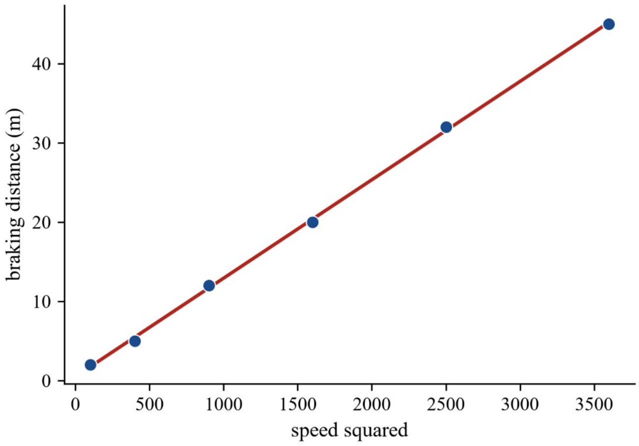


The residuals are positive at both ends and negative in the middle — a clear ∪-shaped pattern, not a random band about zero. This tells us the straight-line model is **not** appropriate: the association is non-linear. The scatter of d against v is increasing and concave up (braking distance rises ever more steeply with speed), so we square the explanatory variable and fit $d = a + b\,v^2$.

Forming the v^2 column $100, 400, 900, 1600, 2500, 3600$ and fitting the least-squares line of d against v^2 gives

$$d = 0.51 + 0.0124\,v^2,$$

with $r \approx 0.9997$. Plotting d against v^2 now gives points lying close to this line, and the residual plot of the transformed fit shows no pattern, confirming that squaring the speed has straightened the data.



Predicting at $v = 70$ km/h, $d = 0.51 + 0.0124 \times 70^2 = 0.51 + 0.0124 \times 4900 \approx 61.3$ m. Because d was not transformed, no back-transform is needed.

∴ the ∪-shaped residual plot signals non-linearity; the model $d = 0.51 + 0.0124\,v^2$ straightens the data and predicts a braking distance of about 61.3 m at 70 km/h.

Example 4 — unscaffolded

The height h (in cm) of a young plant is recorded each week for six weeks. The scatter of h against the week number w increases but flattens off (concave down).

w (week)	1	2	3	4	5	6
h (cm)	8.7	11.3	13.7	15.4	17.2	18.7

Fit a squared transformation and use it to predict the plant's height in week 8, commenting on the reliability of the prediction.

Because the curve is increasing and concave down, we square the **response** variable and fit $h^2 = a + b\,w$. The h^2 column is $75.69, 127.69, 187.69, 237.16, 295.84, 349.69$, and a least-squares fit of h^2 against w gives

$$h^2 = 19.9 + 55.0\,w\,,$$

with $r \approx 1.000$, so the transformed data are close to linear.

Predicting at $w = 8$: $h^2 = 19.9 + 55.0 \times 8 = 459.9$, and back-transforming (the response was squared),

$$h = \sqrt{459.9} \approx 21.4\,\text{cm}\,.$$

Week 8 lies **beyond** the range of the data ($w = 1$ to 6), so this is an **extrapolation**. Even though the model fits the recorded weeks very well, growth may slow further or stop, so a prediction outside the data must be treated with caution.

∴ the model $h^2 = 19.9 + 55.0\,w$ predicts a height of about 21.4 cm in week 8, but as an extrapolation this figure is not reliable.

Practice questions

Scaffolded

Q1. Each scatterplot below shows a curved but steadily **increasing** association. State whether you would square the explanatory variable x or the response variable y to straighten it. (a) The curve bends upward (concave up). (b) The curve rises steeply then flattens off (concave down). (c) y increases at an increasing rate. (d) y increases at a decreasing rate.

Q2. For the data below, y against x is increasing and concave up.

x	2	4	6	8	10
y	9	21	45	75	120

(a) Copy and complete a row of x^2 values.

(b) State which straight-line model, $y=a+b\,x^2$ or $y^2=a+b\,x$, you would fit.

Q3. A squared transformation gives the least-squares line $\hat{y}=3+2\,x^2$. Predict y when (a) $x=5$, (b) $x=8$, (c) $x=10$.

Q4. A response variable was squared, giving the least-squares line $\hat{y}^2=10+6\,x$. For each x, find the predicted y^2 and then back-transform to predict y. (a) $x=4$. (b) $x=9$. (c) $x=15$.

Q5. The braking distance d (m) of a car at speed v (km/h) is modelled by $d=0.5+0.012\,v^2$. (a) Predict the braking distance at $v=50$ km/h. (b) Predict the braking distance at $v=100$ km/h. (c) Using your answers, explain why doubling the speed more than doubles the braking distance.

Q6. For the data below, y against x is increasing and concave up.

x	1	2	3	4	5
y	2	4.5	8.5	14	21

(a) State which variable to square and give the form of the model.

(b) Copy and complete a row of x^2 values.

(c) A CAS fit gives $\hat{y}=1.2+0.8\,x^2$. Predict y when $x=6$.

(d) The residual plot of this transformed fit shows no pattern. What does this tell you about the model?

Unscaffolded

Q7. A squared transformation gives $\hat{y}=4+3\,x^2$. Predict y when $x=6$.

Q8. A response variable was squared, giving $\hat{y}^2=5+4\,x$. Predict y when $x=11$.

Q9. A scatterplot of a town's population against time rises with ever-increasing steepness. State which variable should be squared, and write down the form of the straight-line model that would be fitted.

Q10. The table shows an increasing, concave-up association.

x	1	2	3	4	5	6
y	4	9	17	30	44	63

(a) Using a squared transformation, fit the least-squares line of y against x^2, giving the coefficients correct to two decimal places.

(b) Predict y when $x=7$.

Q11. The table shows an increasing, concave-down association.

x	1	2	3	4	5
y	3.0	4.5	5.6	6.4	7.1

(a) Square the response variable and fit the least-squares line of y^2 against x, giving the coefficients correct to two decimal places.

(b) Predict y when $x=6$ (remember to back-transform).

Q12. A least-squares line fitted to transformed data is $\hat{y}^2=20+5\,x$. A student predicts that when $x=16$, $y=100$. Explain the student's error and give the correct predicted value of y.

Q13. A squared transformation gives $\hat{y}=2+0.5\,x^2$, fitted to data for which x ranged from 1 to 10. Predict y when $x=25$, and comment on the reliability of this prediction.

Q14. A least-squares line is fitted to the raw data of a curved association, and its residual plot shows a clear ∪-shaped pattern. State what this indicates about the straight-line model and what should be done next.

Q15. The period T (s) of a pendulum of length L (m) is modelled by $T^2 = 4.0\,L$. (a) Predict the period when $L = 0.9$ m. (b) Find the length for which the period is 3 s.

Q16. For the data below, apply a squared transformation to the explanatory variable.

x	2	4	6	8	10
y	7	19	40	71	110

(a) Fit the least-squares line of y against x^2 and state the coefficient of determination r^2 (to four decimal places).
(b) Comment on the quality of the fit.

Q17. A squared transformation gives $\hat{y} = 1 + 0.25\,x^2$. Find the value of x for which the model predicts $\hat{y} = 17$ (take the positive value).

Q18. Explain, in terms of how squaring affects large and small numbers, why squaring the x-values can straighten a scatterplot that bends upward (concave up).

Q19. The area A (in square centimetres) of a circular ink blot is measured for several radii r (in cm). The scatter of A against r is increasing and concave up.

r (cm)	1	2	3	4	5
A (cm²)	3.5	12	29	50	79

(a) State which variable to square and fit the least-squares line, giving the coefficients correct to two decimal places.
(b) Predict the area when $r = 6$ cm.

Q20. For a curved association, a least-squares line fitted to y against x has a ∪-shaped residual plot, while a least-squares line fitted to y against x^2 has a residual plot showing only random scatter about zero. State, with a reason, which model should be used.

Answers

Q1. (a) square x; (b) square y; (c) square x; (d) square y. **Q2.** (a) x^2: 4, 16, 36, 64, 100; (b) $y = a + b x^2$ (square x). **Q3.** (a) 53; (b) 131; (c) 203. **Q4.** (a) $y^2 = 34$, $y \approx 5.83$; (b) $y^2 = 64$, $y = 8$; (c) $y^2 = 100$, $y = 10$. **Q5.** (a) 30.5 m; (b) 120.5 m; (c) the distance depends on v^2, so doubling v multiplies the v^2 term by 4 — the distance grows about fourfold, not twofold. **Q6.** (a) square x, $y = a + b x^2$; (b) x^2: 1, 4, 9, 16, 25; (c) 30; (d) no pattern means the straight-line model for the transformed data is appropriate (a good fit). **Q7.** 112. **Q8.** $y^2 = 49$, so $y = 7$. **Q9.** Square the explanatory variable (time); model $P = a + b t^2$. **Q10.** (a) $\hat{y} = 2.27 + 1.69 x^2$; (b) ≈ 85.08. **Q11.** (a) $\hat{y}^2 = -0.66 + 10.35 x$; (b) $y^2 = 61.44$, $y \approx 7.84$. **Q12.** The line predicts y^2, not y: at $x = 16$, $y^2 = 100$, so $y = \sqrt{100} = 10$. The student forgot to back-transform. **Q13.** $\hat{y} = 314.5$; unreliable — $x = 25$ is far beyond the data range (1 to 10), so this is extrapolation. **Q14.** The straight-line model is not appropriate (the pattern shows the association is non-linear); transform the data (e.g. a squared transformation) and refit. **Q15.** (a) $T^2 = 3.6$, $T \approx 1.90$ s; (b) $9 = 4.0 L$, so $L = 2.25$ m. **Q16.** (a) $r^2 = 0.9998$; (b) an excellent fit — the transformed data are very nearly linear. **Q17.** $x = 8$. **Q18.** Squaring stretches large values much more than small ones, so it spreads out the bunched high-x points and pulls the upward-bending curve into a straight line. **Q19.** (a) square r; $\widehat{A} = 0.02 + 3.15 r^2$; (b) ≈ 113.42 cm². **Q20.** Use y against x^2: its residual plot shows random scatter (no pattern), confirming the transformed model is linear, whereas the ∪-shaped residual plot for y against x shows that model is non-linear.

Fully-worked solutions

Q1. For an increasing curved association, concavity decides the variable to square. Concave up (bending upward) ⇒ square the explanatory variable x; concave down (flattening) ⇒ square the response variable y. So (a) square x; (b) square y; (c) "increasing at an increasing rate" is concave up ⇒ square x; (d) "increasing at a decreasing rate" is concave down ⇒ square y.

Q2. (a) Squaring each x-value: $2^2 = 4$, $4^2 = 16$, $6^2 = 36$, $8^2 = 64$, $10^2 = 100$, so the x^2 row is $4, 16, 36, 64, 100$. (b) The association is concave up, so we stretch the x-axis by squaring x and fit $y = a + b x^2$ (plotting y against x^2).

Q3. Substitute into $\hat{y} = 3 + 2 x^2$ (the response is not transformed, so read y directly). (a) $\hat{y} = 3 + 2(5^2) = 3 + 2(25) = 53$. (b) $\hat{y} = 3 + 2(64) = 131$. (c) $\hat{y} = 3 + 2(100) = 203$.

Q4. Substitute into $\hat{y}^2 = 10 + 6 x$ to get y^2, then take the square root. (a) $y^2 = 10 + 6(4) = 34$, so $y = \sqrt{34} \approx 5.83$. (b) $y^2 = 10 + 6(9) = 64$, so $y = \sqrt{64} = 8$. (c) $y^2 = 10 + 6(15) = 100$, so $y = \sqrt{100} = 10$.

Q5. (a) $d = 0.5 + 0.012(50^2) = 0.5 + 0.012(2500) = 0.5 + 30 = 30.5$ m. (b) $d = 0.5 + 0.012(100^2) = 0.5 + 0.012(10000) = 0.5 + 120 = 120.5$ m. (c) The model depends on v^2, not v. Replacing v by $2v$ replaces v^2 by $(2v)^2 = 4v^2$, so the speed-dependent part of the distance is multiplied by 4. Going from 50 to 100 km/h the distance rises from 30.5 to 120.5 m — roughly four times as far — which is why doubling the speed more than doubles the braking distance.

Q6. (a) The scatter is increasing and concave up, so square the explanatory variable and fit $y = a + b x^2$. (b) x^2: $1, 4, 9, 16, 25$. (c) $\hat{y} = 1.2 + 0.8(6^2) = 1.2 + 0.8(36) = 1.2 + 28.8 = 30$. (d) A residual plot with no pattern (random scatter about zero) confirms that the straight-line model for the transformed data is appropriate — the squared transformation has successfully linearised the data.

Q7. $\hat{y} = 4 + 3(6^2) = 4 + 3(36) = 4 + 108 = 112$.

Q8. Substituting into the squared-response line, $\hat{y}^2 = 5 + 4(11) = 49$. Back-transform: $y = \sqrt{49} = 7$.

Q9. "Ever-increasing steepness" means the curve is increasing and concave up, so the large x-values are bunched. Stretch the horizontal axis by squaring the explanatory variable (time t), and fit the straight-line model $P = a + b t^2$.

Q10. (a) Form $x^2 = 1, 4, 9, 16, 25, 36$ and fit the least-squares line of y against x^2. A CAS regression gives $\hat{y} = 2.27 + 1.69\,x^2$ (coefficients to two decimal places). (b) At $x = 7$, $x^2 = 49$, so $\hat{y} = 2.27 + 1.69(49) = 2.27 + 82.81 = 85.08$.

Q11. (a) Form $y^2 = 9.00, 20.25, 31.36, 40.96, 50.41$ and fit the least-squares line of y^2 against x; a CAS regression gives $\hat{y}^2 = -0.66 + 10.35\,x$. (b) At $x = 6$, $\hat{y}^2 = -0.66 + 10.35(6) = -0.66 + 62.10 = 61.44$. Back-transform: $y = \sqrt{61.44} \approx 7.84$.

Q12. Substituting $x = 16$ gives $\hat{y}^2 = 20 + 5(16) = 100$. But this is the predicted value of y^2, the *square* of the response, because a squared transformation was applied to y. The student read this as y. The correct prediction back-transforms by taking the square root: $y = \sqrt{100} = 10$.

Q13. $\hat{y} = 2 + 0.5(25^2) = 2 + 0.5(625) = 2 + 312.5 = 314.5$. However, the model was fitted for x between 1 and 10, and $x = 25$ lies far outside that range. Predicting there is **extrapolation**, and there is no evidence the squared relationship continues so far beyond the data, so the prediction is unreliable.

Q14. A residual plot is the test of whether a straight line is an adequate model: if the model is appropriate the residuals scatter randomly in a band about zero. A clear ∪-shaped pattern is not random scatter, so the straight-line model fitted to the raw data is **not** appropriate — the association is non-linear. (The ∪ shape says the line under-predicts at both ends, where the residuals are positive, and over-predicts in the middle, where they are negative.) What should be done next is to transform one of the variables to straighten the association — for an increasing, concave-up scatter square the explanatory variable and fit $y = a + b\,x^2$; for an increasing, concave-down scatter square the response variable and fit $y^2 = a + b\,x$ — then refit the least-squares line to the transformed data and check its residual plot again, since only a patternless residual plot confirms the transformation has worked.

Q15. (a) $T^2 = 4.0(0.9) = 3.6$, so $T = \sqrt{3.6} \approx 1.90$ s. (b) Set $T = 3$: then $T^2 = 9$, so $9 = 4.0\,L$, giving $L = \dfrac{9}{4.0} = 2.25$ m.

Q16. (a) Form $x^2 = 4, 16, 36, 64, 100$ and fit the least-squares line of y against x^2. The CAS reports a coefficient of determination of $r^2 = 0.9998$ (to four decimal places). (b) Since r^2 is extremely close to 1 (99.98 % of the variation in y explained), the transformed data are very nearly linear — the squared transformation gives an excellent fit.

Q17. Set $\hat{y} = 17$ in $\hat{y} = 1 + 0.25\,x^2$: $17 = 1 + 0.25\,x^2$, so $0.25\,x^2 = 16$ and $x^2 = 64$. Taking the positive root, $x = 8$.

Q18. Squaring is not an even stretch: it barely changes small numbers but greatly increases large ones (for example $1, 2, 3, 4, 5$ become $1, 4, 9, 16, 25$, so the gaps grow from 1 to 9 across the range). A concave-up scatter bends upward because its high x-values are bunched close together while y climbs quickly. Replacing x by x^2 spreads those high x-values further apart, pulling the steep right-hand part of the curve outwards until the points line up straight.

Q19. (a) The area rises ever more steeply with radius (concave up), so square the explanatory variable r. Form $r^2 = 1, 4, 9, 16, 25$ and fit the least-squares line of A against r^2: $\widehat{A} = 0.02 + 3.15\,r^2$ (coefficients to two decimal places; the slope ≈ 3.15 is close to π, as expected for the area of a circle). (b) At $r = 6$, $r^2 = 36$, so $\widehat{A} = 0.02 + 3.15(36) = 0.02 + 113.4 = 113.42$ cm².

Q20. The residual plot is the decisive test of linearity. Fitting y against x leaves a ∪-shaped pattern in the residuals, which shows that model is non-linear. Fitting y against x^2 leaves only random scatter about zero, which shows the transformed model is linear and appropriate. So the model using y against x^2 should be used.

Theory

When we fit a least-squares line to two numerical variables, we are assuming that the association between them is **linear**. A **residual plot** is the check: if it shows a random band of points above and below the line $\text{residual} = 0$, a straight line is a good model; but if it shows a clear **curved pattern**, the association is **non-linear** and the straight-line model is unsuitable. Rather than abandon least squares, we can **transform** the data — replace one of the variables by a new quantity — so that the transformed data *are* linear, fit the line to them, and make predictions from that line. This section develops the **logarithmic (base 10) transformation**, the most useful transformation for data that grow rapidly or span several orders of magnitude.

The base-10 logarithm.

Recall that the **logarithm to base 10** of a positive number y, written $\log_{10} y$, is the power to which 10 must be raised to give y:

$$\log_{10} y = c \text{ means } 10^c = y\,.$$

The two operations undo each other, so $10^{\log_{10} y} = y$. Taking the logarithm of a set of numbers that range over several orders of magnitude **compresses** them: values in the thousands and values in the ones are brought close together, because their logarithms differ only by the number of powers of ten between them. This is exactly the property that lets a logarithm straighten a rapidly rising curve. In practice you read $\log_{10} y$ from the log (base 10) key of your CAS, and recover a value with the antilogarithm 10^c.

Why a logarithm linearises rapid growth.

Suppose a quantity increases by a roughly **constant factor** in each time period — it doubles, or triples, or grows by 20% — rather than by a constant amount. Plotted against time the points curve upward ever more steeply (this is **exponential** growth). Taking $\log_{10}$ of the quantity turns repeated *multiplication* into repeated *addition*: multiplying by the same factor each step adds the same amount to the logarithm each step, so the transformed points fall on a straight line. A logarithm is therefore the natural transformation whenever a variable changes by a constant **percentage** or **factor**.

Transforming one axis only.

A transformation is always applied to **one variable only** — one axis of the scatterplot. There are two choices.

- **Log the response variable ($y \to \log_{10} y$).** Use this when y grows very rapidly as x increases, so the scatterplot bends steeply **upward**. The transformed model is
$$\log_{10} y = a + b\,x\,.$$
- **Log the explanatory variable ($x \to \log_{10} x$).** Use this when the values of x themselves span several orders of magnitude, or when y rises quickly and then **levels off**. The transformed model is
$$y = a + b\log_{10} x\,.$$

(The square and reciprocal transformations, met elsewhere in this chapter, straighten other shapes of curve; each is likewise applied to a single axis.) **Choosing which variable to log:** take the logarithm of the variable whose values stretch over several powers of ten — that is the variable responsible for the curvature. If the response climbs by a constant factor, log the response; if the explanatory values run from, say, tens to hundreds of thousands, log the explanatory variable.

Fitting and checking.

The method has four steps.

1. **Transform** the chosen variable, replacing each value by its $\log_{10}$.

2. **Fit** the least-squares line to the transformed data, using your CAS regression tool exactly as for untransformed data. This gives the coefficients a and b.
3. **Check** with a residual plot of the *transformed* data. If it is now a random band about zero, the transformation has succeeded and the line is a good model.
4. **Predict** by substituting into the fitted line and, if the response was logged, **back-transforming** with a power of ten.

Predicting and back-transforming. For a **log-y** model $\log_{10} y = a + bx$, substitute the value of x to obtain $\log_{10} \hat{y}$, then raise 10 to that power:

$$\hat{y} = 10^{a+bx}.$$

For a **log-x** model $y = a + b\log_{10} x$, the response is not transformed, so substituting x gives $\hat{y}$ directly; to find the x that produces a given y, rearrange to $\log_{10} x = \dfrac{y-a}{b}$ and back-transform with $x = 10^{(y-a)/b}$.

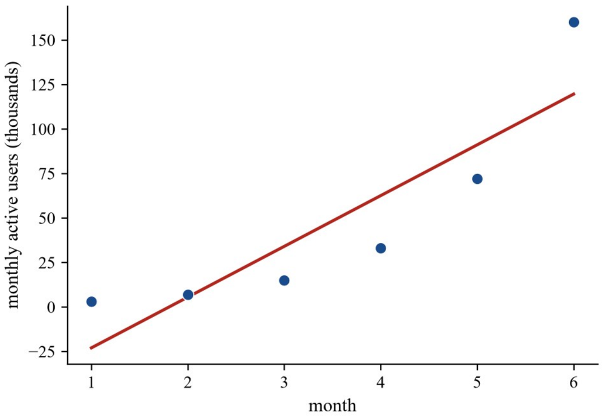
Interpreting the slope of a log-y line. Because a step of 1 in x adds b to $\log_{10} y$, it **multiplies** y by 10^b. So 10^b is the factor by which the response grows for each one-unit increase in the explanatory variable, and 10^a is the predicted response when $x = 0$. A factor greater than 1 describes growth; a factor less than 1 (so b is negative) describes decay.

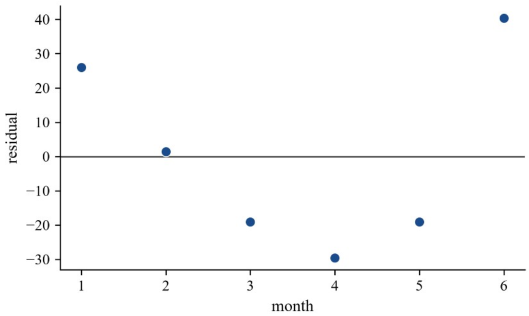
Using CAS. General Mathematics is a technology-active, open-book course. In practice you place the logged values in a new list, run the two-variable regression to obtain a and b, and let the calculator draw the residual plot. Learn the by-hand structure above so that you can decide *which* variable to log, and so that you can interpret and back-transform the fitted equation. As always, a prediction made outside the range of the data is an **extrapolation** and should be treated with caution.

A worked illustration. The monthly active users of a new app (in thousands) over its first six months are shown below.

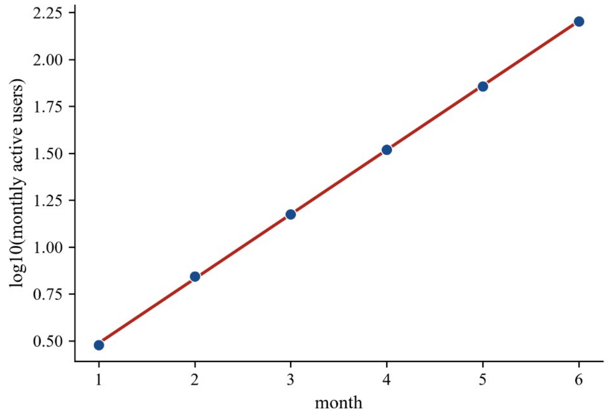
month x	1	2	3	4	5	6
users y (thousands)	3	7	15	33	72	160

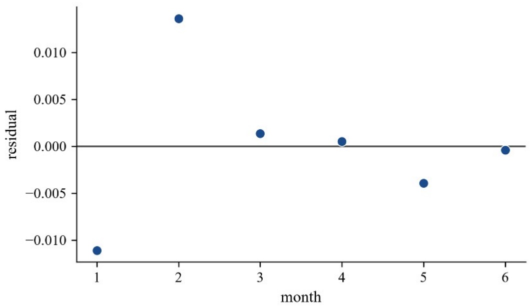
The scatterplot bends steeply upward, and a straight line fits poorly — the residual plot shows a clear curved (∪-shaped) pattern, so the association is non-linear.





The users roughly double each month (a constant factor), so we log the **response**. Replacing each y by $\log_{10} y$ gives $0.48, 0.85, 1.18, 1.52, 1.86, 2.20$, and these fall on a line. The residual plot of the transformed data is now a random band about zero.





The least-squares line of the transformed data (coefficients to three significant figures) is

$$\log_{10} y = 0.145 + 0.343\,x\,, r^2 = 0.9998\,.$$

Since $10^{0.343} \approx 2.20$, the number of users multiplies by about 2.2 each month. To predict the users in month 8, substitute $x = 8$: $\log_{10} \hat{y} = 0.145 + 0.343 \times 8 = 2.889$, so $\hat{y} = 10^{2.889} \approx 770$ thousand users (an extrapolation, so treat it with care).

Worked examples

Example 1 — scaffolded

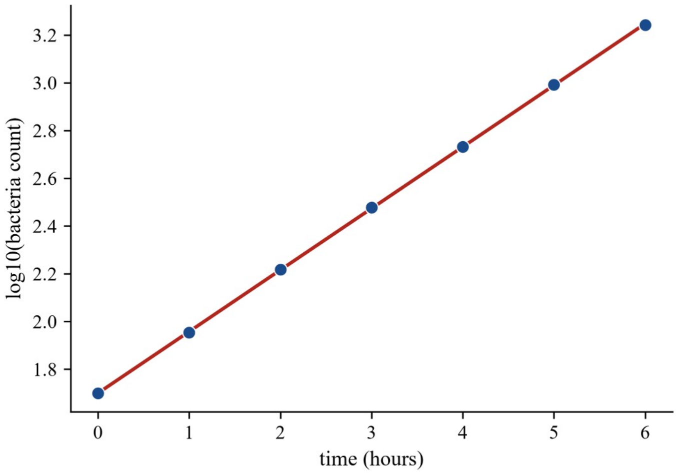
A colony of bacteria is counted every hour. The counts were

time t (hours)	0	1	2	3	4	5	6
count N	50	90	165	300	540	980	1750

The count grows by roughly a constant factor each hour. Transform the data with a logarithmic (base 10) transformation, fit the least-squares line, and interpret the slope.

Working	Comment
The count multiplies each hour (ratios $90/50 = 1.8$, $165/90 \approx 1.8$, …), so log the **response**: replace N by $\log_{10} N$.	Rapid, constant-factor growth ⇒ log the y-axis.
$\log_{10} N: 1.70, 1.95, 2.22, 2.48, 2.73, 2.99, 3.24$.	Read from the CAS log (base 10); values to two decimal places.
Plot $\log_{10} N$ against t: the points lie on a straight line.	The transformation has linearised the data.
Least-squares line: $\log_{10} N = 1.70 + 0.258\,t$, with $r^2 \approx 1.0000$.	Fit the line to $(t, \log_{10} N)$ on the CAS; coefficients to three significant figures.
$10^b = 10^{0.258} \approx 1.81$.	10^b is the factor the count multiplies by each hour.

The count increases by a factor of about 1.81 every hour, and the transformed data are almost perfectly linear.



Example 2 — scaffolded

Using the fitted line $\log_{10} N = 1.70 + 0.258t$ from Example 1, predict the bacteria count after 8 hours.

Working	Comment
Substitute $t = 8$: $\log_{10} \widehat{N} = 1.70 + 0.258 \times 8$.	Predict the *logarithm* of the count first.
$\log_{10} \widehat{N} = 1.70 + 2.064 = 3.764$.	Evaluate the right-hand side.
$\widehat{N} = 10^{3.764} \approx 5810$.	Back-transform: raise 10 to that power.
$t = 8$ lies beyond the data ($0 \le t \le 6$), so this is an extrapolation.	Predictions outside the data range are less reliable.

The model predicts about 5810 bacteria after 8 hours.

Example 3 — unscaffolded

On five islands the area (in km²) and the number of plant species were recorded.

area (km²)	1	10	100	1000	10 000
species	8	17	25	34	42

Fit an appropriate logarithmic model and use it to predict the number of species on an island of area 500 km².

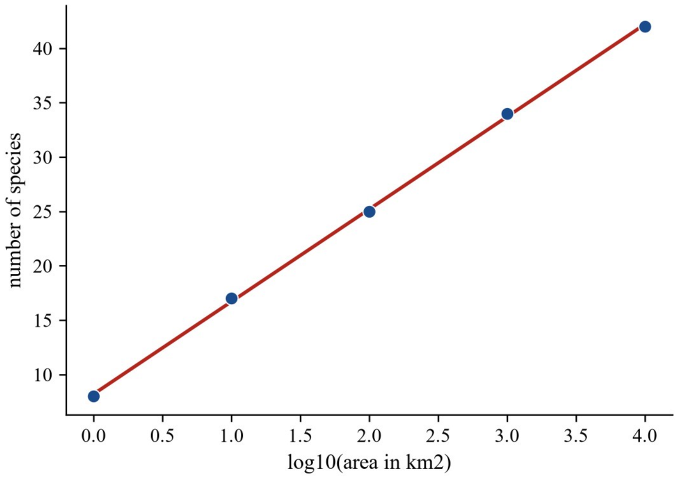
The **areas** span four orders of magnitude (from 1 to 10 000 km²), while the species counts do not, so the curvature comes from the x-axis: we log the **explanatory** variable. Replacing each area by $\log_{10}(\text{area})$ gives $0, 1, 2, 3, 4$, and plotting species against these values produces a straight line. The least-squares line (coefficients to three significant figures) is

$$\text{species} = 8.20 + 8.50 \log_{10}(\text{area}), r^2 = 0.9996,$$

so each ten-fold increase in area adds about 8.5 species. For an area of 500 km², $\log_{10} 500 \approx 2.699$, so

$$\widehat{\text{species}} = 8.20 + 8.50 \times 2.699 \approx 31.1.$$

∴ the model predicts about 31 plant species on an island of area 500 km².



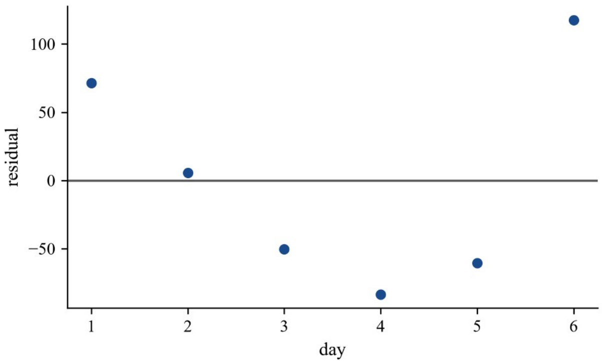
Example 4 — unscaffolded

The number of people in a town who had heard a particular rumour was recorded each day.

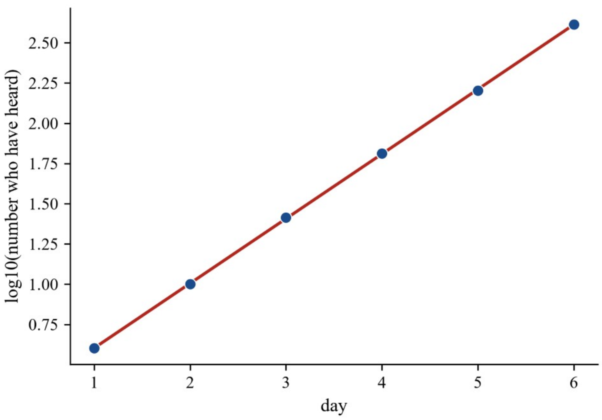
day x	1	2	3	4	5	6
number R	4	10	26	65	160	410

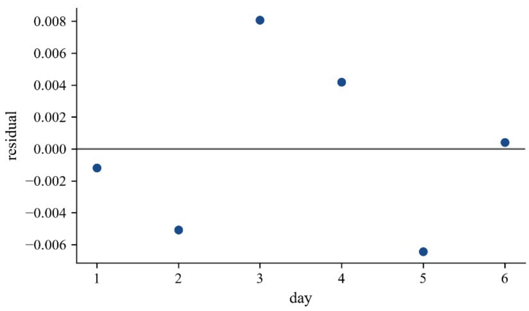
Show that a straight-line model is unsuitable, apply a logarithmic transformation, confirm the fit, and predict how many people will have heard the rumour on day 8.

Fitting a least-squares line to the raw data and plotting the residuals gives a clear curved (∪-shaped) pattern — the residuals are positive at the ends and negative in the middle — so the raw association is non-linear.



The number roughly multiplies each day, so we log the response. Replacing R by $\log_{10} R$ gives $0.60, 1.00, 1.41, 1.81, 2.20, 2.61$, which lie on a straight line, and the residual plot of the transformed data is now a random band about zero — the transformation has succeeded.





The least-squares line of the transformed data is

$$\log_{10} R = 0.201 + 0.402\,x, r^2 = 0.9999.$$

Since $10^{0.402} \approx 2.52$, the number of people who have heard the rumour multiplies by about 2.5 each day. To predict day 8, substitute $x=8$: $\log_{10}\widehat{R} = 0.201 + 0.402 \times 8 = 3.417$, so $\widehat{R} = 10^{3.417} \approx 2610$.

∴ the model predicts that about 2610 people will have heard the rumour on day 8, though this is an extrapolation beyond the six days of data.

Practice questions

Scaffolded

Q1. (a) Convert each value to a logarithm (base 10), correct to two decimal places: 40, 400, 4000, $40\,000$. (b) Find the value represented by each logarithm (the antilogarithm), correct to the nearest whole number: 1.70, 2.30, 3.85.

Q2. The population of a small colony of rabbits was recorded each year.

year x	0	1	2	3	4
population P	20	46	105	240	550

(a) Copy and complete the row of $\log_{10} P$ values, correct to two decimal places.
(b) A CAS fits the least-squares line $\log_{10} P = 1.30 + 0.360\,x$. State the value of r^2 to four decimal places (you should find it is very close to 1).
(c) Predict the population when $x=6$ by finding $\log_{10}\widehat{P}$ and back-transforming.
(d) Find 10^b and interpret it in context.

Q3. A sound's loudness L (in a certain scale) was measured at four sound intensities I.

intensity I	1	10	100	1000
loudness L	2	11	21	29

(a) Copy and complete the row of $\log_{10} I$ values.
(b) The least-squares line is $L = 2.10 + 9.10\log_{10} I$ with $r^2 = 0.9983$. Should the intensity or the loudness be logged, and why?
(c) Predict the loudness when $I = 200$.
(d) Interpret the slope 9.10 in context.

Q4. A different bacteria colony was counted each hour over the first 6 hours. Its transformed data give the least-squares line $\log_{10} N = 1.45 + 0.301\,t$, where N is the count and t is the time in hours. (a) Predict the count when $t=10$, correct to the nearest hundred. (b) Find 10^b and state what it tells you about the colony. (c) Explain why the prediction in (a) should be treated with more caution than a prediction at $t=5$.

Q5. A culture of microbes was counted each hour.

time t (hours)	0	1	2	3	4	5
count C	12	38	120	380	1180	3700

(a) Copy and complete the row of $\log_{10} C$ values, correct to two decimal places.
(b) The least-squares line is $\log_{10} C = 1.08 + 0.498\,t$, with $r^2 \approx 1.0000$. Find 10^b and describe the growth in words.
(c) Predict the count when $t=6$, correct to the nearest ten.
(d) Predict the count when $t=8$, correct to the nearest hundred.

Q6. For each situation, state whether you would apply a logarithmic transformation to the explanatory variable, to the response variable, or to neither, and give a brief reason. (a) A scatterplot of population against year curves steeply upward; the populations run from hundreds to millions. (b) A scatterplot of test score against study time is a straight band of points with a random residual plot. (c) A scatterplot of reaction rate against concentration rises quickly and then flattens; the concentrations range from 0.1 to 1000.

Unscaffolded

Q7. (a) Convert 6, 250 and $90\,000$ to logarithms (base 10), correct to two decimal places. (b) Find the value represented by the logarithm 2.85, correct to the nearest whole number.

Q8. The value V of an investment (in dollars) was recorded every two years.

year x	0	2	4	6	8	10
value V (\$)	5000	7200	10 500	15 100	21 800	31 500

(a) Explain why a logarithmic transformation of V is appropriate, and state the fitted line $\log_{10} V = 3.70 + 0.0800\,x$ (with $r^2 \approx 1.0000$).
(b) Predict the value after 12 years, correct to the nearest hundred dollars.
(c) By how many times does the investment grow each year? Answer to two decimal places.

Q9. In a survey the number of different bird species seen was recorded at five sites of widely differing area.

area sampled (m²)	5	50	500	5000	50 000
species	12	28	45	60	78

(a) State, with a reason, which variable should be logged.
(b) The least-squares line is $\text{species} = 0.337 + 16.4\,\log_{10}(\text{area})$, with $r^2 = 0.9994$. Predict the number of species for an area of 1000 m².
(c) Find the area (to the nearest ten m²) at which the model predicts 40 species.

Q10. For a set of data, a straight-line model gives $r^2 = 0.86$ and a residual plot with a clear curved pattern, while a log-y model gives $r^2 = 0.998$ and a random residual plot. Which model should be reported, and why? What does the residual plot add that r^2 alone does not?

Q11. The spread of the rumour in Example 4 is modelled by $\log_{10} R = 0.201 + 0.402\,x$, where x is the day. On which day does the model first predict that more than 1000 people will have heard the rumour?

Q12. For the bacteria model of Example 1, $\log_{10} N = 1.70 + 0.258\,t$, evaluate 10^a (where a is the intercept) and explain what it represents in the context of the colony.

Q13. The distances of the eight planets from the Sun range from about 58 million km to about 4500 million km, and are to be plotted against planet number (1 for the closest, and so on). A scatterplot of distance against planet number curves steeply upward. Explain which variable should be transformed with a base-10 logarithm and why.

Q14. A model $y = a + b\log_{10} x$ has been fitted to data in which y rises rapidly for small x and then almost levels off. Explain, in terms of the shape of the curve, why logging x (rather than y) straightens this association.

Q15. The value V of a car (in dollars) was recorded each year.

age t (years)	0	1	2	3	4	5
value V (\$)	30 000	24 000	19 200	15 360	12 288	9830

(a) Convert the values to $\log_{10} V$ (two decimal places) and describe what the transformed plot looks like.
(b) The least-squares line is $\log_{10} V = 4.48 - 0.0969\,t$. Predict the value at age 8, correct to the nearest ten dollars.
(c) Find 10^b and interpret it in context.

Q16. For a log-y model $\log_{10} y = a + bx$, explain what it means for the slope b to be **negative**, and give the factor by which y changes for each one-unit increase in x in terms of b.

Q17. After a logarithmic transformation and a least-squares fit, the residual plot of the *transformed* data still shows a clear curved pattern. What does this tell you about the transformation, and what would you try next?

Q18. Two scatterplots are each strongly non-linear. In the first, y increases by roughly a constant factor as x increases by 1. In the second, the x-values range over four orders of magnitude while y increases steadily. For each, state which variable you would transform with a base-10 logarithm.

Q19. A data set has already been transformed by logging the response. The pairs $(x, \log_{10} y)$ are

x	10	20	30	40	50
$\log_{10} y$	0.90	1.35	1.80	2.25	2.70

(a) The least-squares line of these transformed data is $\log_{10} y = 0.45 + 0.045x$. Verify that the point $(30, 1.80)$ lies on this line.

(b) Predict y when $x = 60$, correct to the nearest ten.

Q20. Explain why, on a log-y model of exponential growth, the intercept a and slope b have the interpretations "10^a is the starting value" and "10^b is the growth factor", and why these are more meaningful summaries of the data than the intercept and slope of a straight line fitted to the raw values.

Answers

Q1. (a) $1.60, 2.60, 3.60, 4.60$; (b) 50, 200, 7079. **Q2.** (a) $1.30, 1.66, 2.02, 2.38, 2.74$; (b) $r^2 \approx 1.0000$; (c) $\log_{10}\widehat{P} = 3.46$, $\widehat{P} \approx 2880$; (d) $10^{0.360} \approx 2.29$, the population multiplies by about 2.29 each year. **Q3.** (a) $0, 1, 2, 3$; (b) log the intensity, since its values span several orders of magnitude (1 to 1000); (c) $L = 2.10 + 9.10\log_{10}200 \approx 23$; (d) each ten-fold increase in intensity raises the loudness by about 9.1 units. **Q4.** (a) $\log_{10}\widehat{N} = 4.46$, $\widehat{N} \approx 28\,800$; (b) $10^{0.301} \approx 2.00$, the count doubles each hour; (c) $t = 10$ lies beyond the 6 hours of data (an extrapolation), whereas $t = 5$ lies inside that range (an interpolation). **Q5.** (a) $1.08, 1.58, 2.08, 2.58, 3.07, 3.57$; (b) $10^{0.498} \approx 3.15$, the count roughly triples each hour; (c) $\widehat{C} \approx 11\,690$; (d) $\widehat{C} \approx 115\,900$. **Q6.** (a) log the response (population), rapid constant-factor growth; (b) neither, the data are already linear; (c) log the explanatory variable (concentration), it spans orders of magnitude and the curve levels off. **Q7.** (a) $0.78, 2.40, 4.95$; (b) $10^{2.85} \approx 708$. **Q8.** (a) V grows by a roughly constant factor, so logging V linearises it; (b) $\log_{10}\widehat{V} = 4.66$, $\widehat{V} \approx \$45\,700$; (c) $10^{0.0800} \approx 1.20$, so about 1.20 times (a 20% rise) each year. **Q9.** (a) log the area, since it spans orders of magnitude (5 to $50\,000$); (b) $0.337 + 16.4\log_{10}1000 \approx 50$ species; (c) $\log_{10}\text{area} = 40 - 0.337/16.4 \approx 2.42$, area ≈ 260 m². **Q10.** Report the log-y model: its higher r^2 and random residual plot show a much better fit. The residual plot reveals the *pattern* of the misfit (curvature), which a single r^2 value can hide. **Q11.** $\log_{10}R = 3 \Rightarrow x = 3 - 0.201/0.402 \approx 6.96$, so the model first predicts over 1000 on day 7. **Q12.** $10^{1.70} \approx 50$, the predicted count at $t = 0$, i.e. the starting number of bacteria. **Q13.** Log the distance (the response): it ranges over about two orders of magnitude and drives the steep upward curve, so logging it straightens the plot. **Q14.** Logging x compresses the large x-values, stretching the crowded small-x region and pulling in the flat tail, which straightens a curve that rises then levels off. **Q15.** (a) $4.48, 4.38, 4.28, 4.19, 4.09, 3.99$, a straight line sloping **downward**; (b) $\log_{10}\widehat{V} = 3.705$, $\widehat{V} \approx \$5070$; (c) $10^{-0.0969} \approx 0.80$, the value falls to about 80% (a 20% loss) each year. **Q16.** A negative b means $\log_{10}y$ decreases as x increases, so y **decays**; each unit rise in x multiplies y by 10^b, which is less than 1. **Q17.** The logarithmic transformation has not fully linearised the data, so a log model is not appropriate; try a different transformation (e.g. reciprocal or square), or log the other variable. **Q18.** First: log the response y (constant-factor growth). Second: log the explanatory variable x (its values span orders of magnitude). **Q19.** (a) $0.45 + 0.045 \times 30 = 1.80$, which matches; (b) $\log_{10}\hat{y} = 0.45 + 0.045 \times 60 = 3.15$, $\hat{y} \approx 1410$. **Q20.** Setting $x = 0$ gives $\log_{10}y = a$, so 10^a is the starting value; each unit rise in x adds b to $\log_{10}y$ and so multiplies y by 10^b, the growth factor. These describe multiplicative (percentage) growth, which is what the data actually do, whereas a raw straight line assumes a constant *amount* of change and fits the curve badly.

Fully-worked solutions

Q1. (a) $\log_{10}40 = 1.60$, $\log_{10}400 = 2.60$, $\log_{10}4000 = 3.60$, $\log_{10}40\,000 = 4.60$ (each is one more than the previous, because each value is ten times the one before). (b) The antilogarithm of c is 10^c: $10^{1.70} \approx 50$, $10^{2.30} \approx 200$, $10^{3.85} \approx 7079$.

Q2. (a) Taking $\log_{10}$ of each population: $\log_{10}20 = 1.30$, $\log_{10}46 = 1.66$, $\log_{10}105 = 2.02$, $\log_{10}240 = 2.38$, $\log_{10}550 = 2.74$. (b) Fitting the least-squares line to $(x, \log_{10}P)$ gives $r^2 \approx 1.0000$ — the transformed data are almost exactly linear. (c) At $x = 6$, $\log_{10}\widehat{P} = 1.30 + 0.360 \times 6 = 1.30 + 2.16 = 3.46$, so $\widehat{P} = 10^{3.46} \approx 2880$. (d) $10^b = 10^{0.360} \approx 2.29$, so the rabbit population multiplies by about 2.29 each year.

Q3. (a) $\log_{10}1 = 0$, $\log_{10}10 = 1$, $\log_{10}100 = 2$, $\log_{10}1000 = 3$. (b) The intensities $1, 10, 100, 1000$ span three orders of magnitude while the loudness values do not, so the curvature comes from the x-axis: log the **intensity**. (c) At $I = 200$, $\log_{10}200 \approx 2.301$, so $L = 2.10 + 9.10 \times 2.301 \approx 23$. (d) The slope 9.10 is the increase in loudness for each unit increase in $\log_{10}I$; since one unit of $\log_{10}I$ is a ten-fold increase in intensity, loudness rises by about 9.1 units each time the intensity is multiplied by ten.

Q4. (a) At $t = 10$, $\log_{10}\widehat{N} = 1.45 + 0.301 \times 10 = 1.45 + 3.01 = 4.46$, so $\widehat{N} = 10^{4.46} \approx 28\,800$. (b) $10^b = 10^{0.301} \approx 2.00$, so the count is multiplied by about 2 each hour — the colony doubles hourly. (c) The colony was counted only over the first 6 hours, so $t=10$ lies beyond the observed data (an extrapolation) and the doubling pattern may not continue — food or space may run out — whereas $t=5$ lies inside the observed range (an interpolation) and is more trustworthy.

Q5. (a) $\log_{10}12=1.08$, $\log_{10}38=1.58$, $\log_{10}120=2.08$, $\log_{10}380=2.58$, $\log_{10}1180=3.07$, $\log_{10}3700=3.57$. (b) $10^b=10^{0.498}\approx 3.15$, so the microbe count roughly **triples** each hour. (c) At $t=6$, $\log_{10}\widehat{C}=1.08+0.498\times 6=1.08+2.988=4.068$, so $\widehat{C}=10^{4.068}\approx 11\,690$. (d) At $t=8$, $\log_{10}\widehat{C}=1.08+0.498\times 8=1.08+3.984=5.064$, so $\widehat{C}=10^{5.064}\approx 115\,900$.

Q6. (a) The response (population) grows by a roughly constant factor and bends the plot steeply upward, so log the **response**. (b) The plot is already a straight band with a random residual plot, so **no** transformation is needed. (c) The curve rises and then levels off and the explanatory values (0.1 to 1000) span four orders of magnitude, so log the **explanatory** variable.

Q7. (a) $\log_{10}6\approx 0.78$, $\log_{10}250\approx 2.40$, $\log_{10}90\,000\approx 4.95$. (b) $10^{2.85}\approx 708$.

Q8. (a) The successive values grow by a roughly constant factor (each is about 1.44 times the value two years earlier), so the association is exponential and logging V straightens it; the fitted line is $\log_{10}V=3.70+0.0800\,x$ with $r^2\approx 1.0000$. (b) At $x=12$, $\log_{10}\widehat{V}=3.70+0.0800\times 12=3.70+0.96=4.66$, so $\widehat{V}=10^{4.66}\approx \$\,45\,700$. (c) The factor per year is $10^b=10^{0.0800}\approx 1.20$, a rise of about $20\,\%$ each year.

Q9. (a) The areas run from 5 to $50\,000$ m² — four orders of magnitude — so log the **area**. (b) At an area of 1000 m², $\log_{10}1000=3$, so $\widehat{\text{species}}=0.337+16.4\times 3=0.337+49.2\approx 50$ species. (c) Setting species $=40$: $\log_{10}(\text{area})=\dfrac{40-0.337}{16.4}\approx 2.42$, so area $=10^{2.42}\approx 260$ m².

Q10. The log-y model has $r^2=0.998$ (against 0.86) and a random residual plot, so it fits far better and should be reported. The residual plot adds information that r^2 hides: a curved residual pattern shows *systematic* misfit — the model is the wrong shape — even when r^2 is fairly high, whereas a random band confirms that a straight line (in the transformed variables) is the right shape.

Q11. More than 1000 people means $\log_{10}R>3$. Setting $\log_{10}R=3$ gives $3=0.201+0.402\,x$, so $x=\dfrac{3-0.201}{0.402}\approx 6.96$. Since x counts whole days, the model first predicts over 1000 on **day** 7.

Q12. The intercept is $a=1.70$. Setting $t=0$ gives $\log_{10}N=1.70$, so $N=10^{1.70}\approx 50$. This is the predicted count at the start ($t=0$) — the initial number of bacteria in the colony.

Q13. The distances range from about 58 to 4500 million km — roughly two orders of magnitude — and it is this rapid growth in the response that bends the plot steeply upward. Logging the **distance** compresses the large values, so the transformed plot of $\log_{10}(\text{distance})$ against planet number is straight.

Q14. When y rises quickly for small x and then flattens, the points are crowded together at small x and stretched out along a nearly flat tail at large x. Taking $\log_{10}x$ compresses the large x-values (pulling in the tail) and spreads out the small x-values (opening up the crowded region), which straightens the curve. Logging y instead would not fix a curve of this shape.

Q15. (a) $\log_{10}30\,000=4.48$, $\log_{10}24\,000=4.38$, $\log_{10}19\,200=4.28$, $\log_{10}15\,360=4.19$, $\log_{10}12\,288=4.09$, $\log_{10}9830=3.99$; the transformed plot is a straight line sloping **downward** (the value decays). (b) At $t=8$, $\log_{10}\widehat{V}=4.48-0.0969\times 8=4.48-0.7752=3.7048\approx 3.705$, so $\widehat{V}=10^{3.705}\approx \$\,5070$. (c) $10^b=10^{-0.0969}\approx 0.80$, so each year the value falls to about $80\,\%$ of the previous year's value — a loss of about $20\,\%$ per year.

Q16. If $b<0$ then increasing x *decreases* $\log_{10}y$, so y itself decreases as x grows — the response **decays** rather than grows. Each one-unit increase in x multiplies y by 10^b; because b is negative, $10^b<1$, confirming a shrinking response.

Q17. A curved residual plot for the transformed data means the logarithmic transformation has *not* linearised the association, so a log model is not appropriate here. The next step is to try a different transformation — for example a reciprocal or a square transformation, or logging the other variable — and re-check the residual plot each time until it shows a random band.

Q18. First scatterplot: y grows by a constant factor, exponential growth, so transform the **response** with $\log_{10} y$. Second scatterplot: the x-values span four orders of magnitude, so the curvature is driven by the x-axis; transform the **explanatory** variable with $\log_{10} x$.

Q19. (a) Substituting $x=30$: $0.45+0.045\times 30=0.45+1.35=1.80$, which equals the given $\log_{10} y$ value, so the point lies on the line. (b) At $x=60$, $\log_{10}\hat{y}=0.45+0.045\times 60=0.45+2.70=3.15$, so $\hat{y}=10^{3.15}\approx 1410$.

Q20. In the model $\log_{10} y=a+b\,x$, substituting $x=0$ gives $\log_{10} y=a$, hence $y=10^{a}$ — the value of the response at the start. Increasing x by 1 adds b to $\log_{10} y$, which multiplies y by 10^{b}; so 10^{b} is the constant factor by which the response grows each step. A straight line fitted to the raw data instead assumes y changes by the same *amount* each step, which does not match multiplicative growth, so it fits the curve poorly and its slope and intercept are not meaningful summaries of the data.

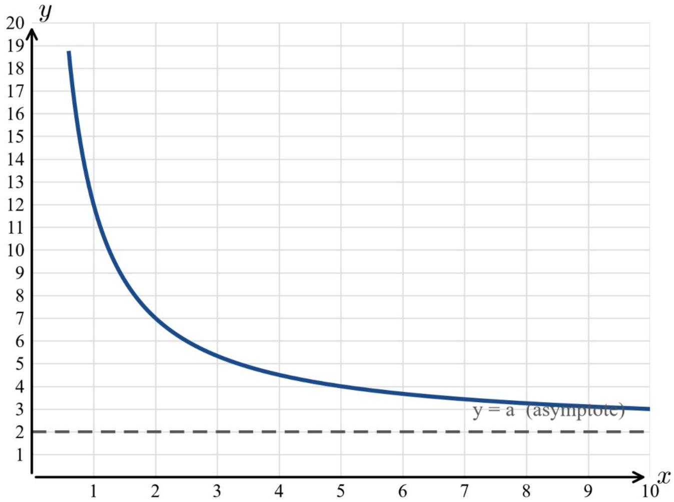
Theory

When a straight line is not enough. In an earlier chapter we fitted a **least-squares line** $y = a + bx$ to two numerical variables and checked the fit with a **residual plot**. If the points of a scatterplot lie in a clearly **curved** band, or if the residual plot shows a **pattern** rather than a random scatter about zero, then a straight line is the wrong model. Rather than abandon the powerful least-squares method, we **transform** one of the variables so that the transformed data become linear, fit the line to the transformed data, and use that line to make predictions. This section studies one such transformation: the **reciprocal transformation**.

The reciprocal function. The reciprocal of a number x is $\frac{1}{x}$. As x increases, its reciprocal $\frac{1}{x}$ **decreases** towards 0; as x decreases towards 0, its reciprocal grows without bound. The graph of $y = \frac{1}{x}$ is therefore a curve that is **steep** for small x and **flattens** as x increases, approaching the horizontal axis but never touching it. Adding a constant and a multiplier gives the general shape

$$y = a + b\frac{1}{x},$$

whose graph falls (for $b > 0$) steeply at first and then levels off towards the horizontal line $y = a$, called a **horizontal asymptote**.



It is exactly this shape — a scatter of points that drops steeply and then flattens out (or rises steeply and then flattens) — that a reciprocal transformation is designed to **straighten**.

The two reciprocal transformations. The transformation is **applied to one axis only** — you reciprocate *either* the explanatory variable *or* the response variable, not both.

- **Reciprocal of x (transform the explanatory variable).** Replace each x by $\frac{1}{x}$ and fit the least-squares line

  $$y = a + b\frac{1}{x}.$$

  A scatterplot of y against $\frac{1}{x}$ is then a **straight line** with slope b and intercept a.

- **Reciprocal of y (transform the response variable).** Replace each y by $\frac{1}{y}$ and fit the least-squares line

  $$\frac{1}{y} = a + bx.$$

A scatterplot of $\frac{1}{y}$ against x is then a **straight line**.

In both cases the transformed data are linear, so the coefficients a and b are found by the ordinary least-squares method — in practice, by entering the transformed values into the CAS two-variable statistics and reading off the line, exactly as for untransformed data.

Choosing the transformation. A reciprocal transformation **compresses the large values** on whichever axis it is applied to and stretches the small ones, which is what pulls a flattening curve back into a straight line. The steep-then-flat shape tells you that a **reciprocal** is worth trying, but on its own it does **not** decide which axis: a curve of this shape can usually be straightened by reciprocating *either* variable, so the shape gives you **candidates**, not a single answer. Two things help you choose:

- **the level the data flatten towards.** The model $y = a + b\frac{1}{x}$ levels off at the horizontal asymptote $y = a$, so it can follow data that settle towards a value clearly **above zero**. Reciprocating the response instead gives $\frac{1}{y} = a + bx$, that is $y = \frac{1}{a+bx}$, which keeps falling towards **zero**, so it suits data that appear to be heading for zero.
- **which candidate actually fits better.** Reciprocating x has the practical advantage of leaving the response in its original units, so no back-transformation is needed and the slope and intercept can be interpreted directly; but if reciprocating y gives the better fit, use it.

You then **confirm** the choice by fitting: the correlation coefficient r of the transformed data should be close to ± 1, and the **residual plot of the transformed fit** should show a random scatter about zero with no leftover pattern. If a pattern remains, fit the other reciprocal instead (or try a different transformation), and keep the model with the patternless residual plot and the larger $|r|$.

Making predictions. Once the line is fitted to the transformed data, a prediction is made by **substituting** into the equation — and, where the *response* was reciprocated, by **back-transforming** the result.

- If you fitted $y = a + b\frac{1}{x}$, substitute the value of x; the equation returns y directly.
- If you fitted $\frac{1}{y} = a + bx$, substitute x to obtain a value for $\frac{1}{y}$, then **take the reciprocal** to recover the prediction
$$y = \frac{1}{a+bx}.$$

As always, a prediction made **within** the range of the data (**interpolation**) is reliable, while one made **beyond** the range (**extrapolation**) assumes the pattern continues and must be treated with caution.

Using CAS. General Mathematics is a technology-active course: an approved CAS calculator or software may be used throughout, including in both end-of-year examinations, where you may also take in one bound reference (which may be annotated). In practice you create a new list or column holding the transformed values — $\frac{1}{x}$ or $\frac{1}{y}$ — and then run the same least-squares regression on the transformed list that you would on any other, reading off a, b and r. Knowing the by-hand method below lets you set the calculator up correctly, try each candidate variable, and judge from r and the residual plot whether the transformation has done its job.

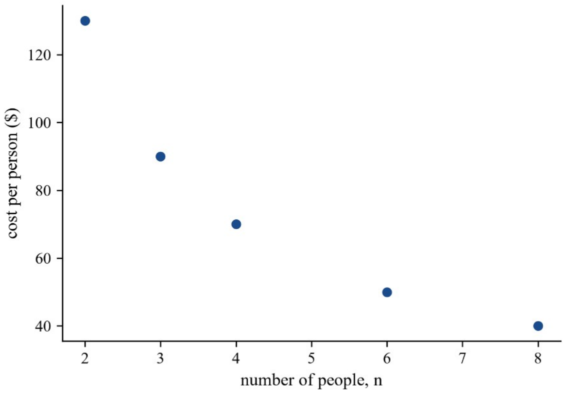
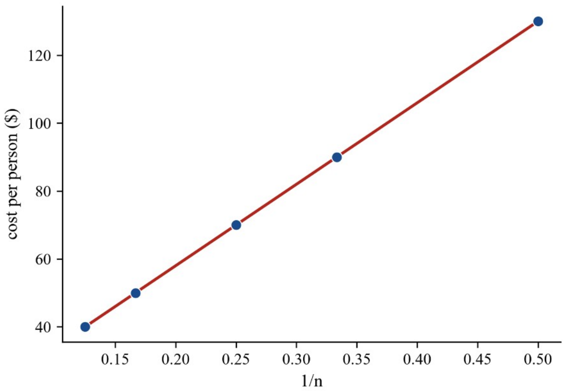
Worked examples

Example 1 — scaffolded

A group of friends hires a minibus for a day trip. The fixed hire charge is shared equally between them, and each person also pays $10 for lunch. The table shows the cost per person, c dollars, when n people go on the trip.

Number of people, n	2	3	4	6	8
Cost per person, c ($)	130	90	70	50	40

Show that a reciprocal transformation straightens the data, and find the least-squares line.

Working	Comment
A scatterplot of c against n falls steeply and then flattens — a curve, not a line.	The reciprocal shape. The cost levels off well above \$0 (each person still pays for lunch), so reciprocate x (here n); the fit will confirm it.
	The raw data are clearly non-linear.
Compute $\frac{1}{n}$: $0.500, 0.333, 0.250, 0.167, 0.125$.	Reciprocal of each n (3 decimal places).
Plot c against $\frac{1}{n}$ — the points now lie on a straight line.	The transformation has linearised the data.
	Straight-line scatter, so least-squares applies.
Least-squares line (from CAS): $c = 10 + 240\frac{1}{n}$, with $r = 1$.	Intercept $a = 10$, slope $b = 240$.

The correlation coefficient of the transformed data is $r = 1$, confirming that reciprocating n has straightened the data perfectly. The model is $c = 10 + 240\frac{1}{n}$.

Example 2 — scaffolded

Use the model $c = 10 + 240\frac{1}{n}$ from Example 1 to answer the following.

Working	Comment
(a) $n = 5$: $c = 10 + 240 \times \frac{1}{5} = 10 + 48 = \$58$.	$n = 5$ lies between observed values (2–8) — interpolation, so reliable.

Working	Comment
(b) $n=12$: $c=10+240\times\frac{1}{12}=10+20=\$30$.	$n=12$ is beyond the data — extrapolation, so treat with caution.
(c) As n grows large, $240\times\frac{1}{n}\to 0$, so $c\to\$10$.	The intercept \$10 is the horizontal asymptote — the lunch each person pays, which does not depend on n.

So the model predicts a cost per person of \$58 for 5 people and \$30 for 12 people, and it can never fall below the \$10 lunch cost.

Example 3 — unscaffolded

The concentration y (mg/L) of a drug in a patient's blood is measured x hours after a dose is given:

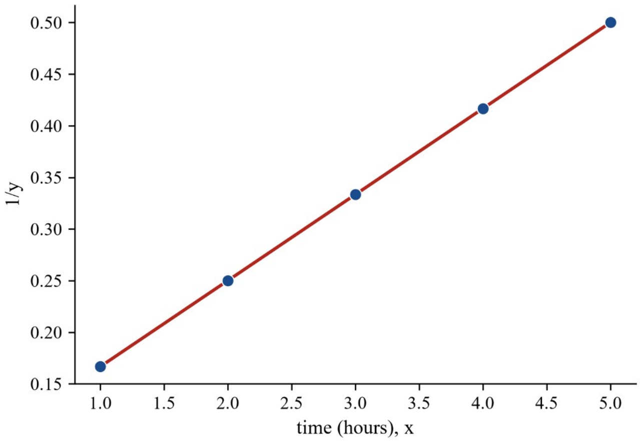
Time, x (hours)	1	2	3	4	5
Concentration, y (mg/L)	6	4	3	2.4	2

A scatterplot of y against x is curved. Show that reciprocating the concentration straightens the data, fit the least-squares line, and predict the concentration after 2.5 hours.

The concentration falls steeply and then flattens as time increases — the reciprocal shape, so a reciprocal transformation is worth trying. Either variable is a candidate; here the concentration is falling away towards zero rather than levelling off at a value above zero, which is what $\frac{1}{y}=a+bx$ describes, so — as the question asks — reciprocate the **response** y. The reciprocals (to 3 decimal places) are

$$\frac{1}{y}=0.167, 0.250, 0.333, 0.417, 0.500.$$

Plotting $\frac{1}{y}$ against x gives points lying on a straight line, so the transformation has worked.



The least-squares line fitted to the transformed data (from CAS) is

$$\frac{1}{y}=0.0833+0.0833x, r=1.$$

To predict the concentration after $x=2.5$ hours, substitute and then **back-transform**:

$$\frac{1}{y}=0.0833+0.0833\times 2.5\approx 0.292, y\approx\frac{1}{0.292}\approx 3.4.$$

Since 2.5 lies within the observed times (1–5 hours), this is interpolation. The prediction is given to 1 decimal place, as the recorded concentrations are.

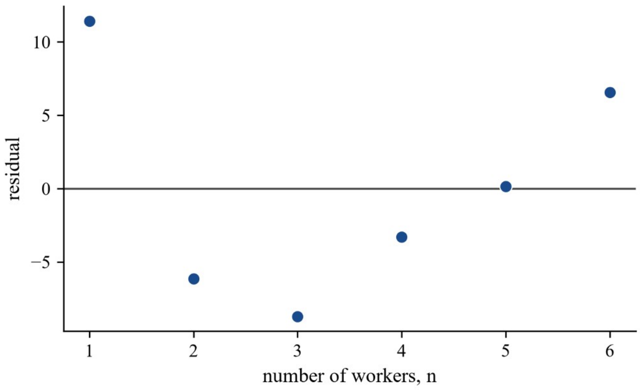
$\therefore$ the model predicts a concentration of about 3.4 mg/L after 2.5 hours.

Example 4 — unscaffolded

A workshop task is carried out by n workers, and the time t (hours) to finish is recorded:

Number of workers, n	1	2	3	4	5	6
Time, t (hours)	50	25	15	13	9	8

A straight line $t = 46 - 7.43n$ was fitted directly to the data. Its residual plot is shown below. Explain what the residual plot indicates, then use a reciprocal transformation to model the data and predict the time for 8 workers.

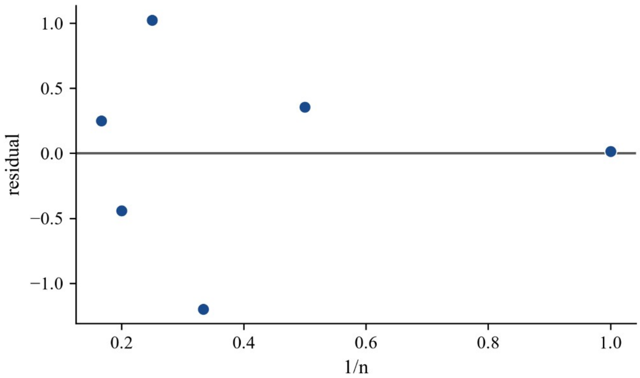


The residual plot of the straight-line fit is not a random scatter about zero: the residuals are positive at the ends and negative in the middle, forming a clear $\cup$-shaped pattern. A pattern in the residual plot means the association is **non-linear**, so a straight line is not an adequate model.

The data fall steeply and then flatten as n increases, which is the reciprocal shape. Either variable is a candidate, but for a fixed amount of work the time should be roughly proportional to $\frac{1}{n}$, and reciprocating n keeps the time in hours, so reciprocate n and check the fit. Fitting the least-squares line to t against $\frac{1}{n}$ (from CAS) gives

$$t = -0.70 + 50.7\frac{1}{n}, r = 0.999.$$

The residual plot of this transformed fit shows a random scatter about zero with no leftover pattern, confirming that the transformed model is a good linear fit.



To predict the time for $n=8$ workers, substitute:

$$t=-0.70+50.7\times\frac{1}{8}=-0.70+6.3375=5.6375\approx 5.6.$$

Because $n=8$ is beyond the observed range (1–6 workers), this is an extrapolation and should be treated with caution.

∴ the transformed model predicts about 5.6 hours for 8 workers, with the caution that this is an extrapolation.

Practice questions

Scaffolded

Q1. The current I (amps) flowing through a component of resistance R (ohms) at a fixed voltage was measured:

Resistance, R (ohms)	2	3	4	6	12
Current, I (amps)	6	4	3	2	1

(a) Copy and complete the row of reciprocals $\frac{1}{R}$ (to 3 decimal places).

(b) State which axis has been transformed.

(c) The least-squares line of I against $\frac{1}{R}$ is $I=a+b\frac{1}{R}$; from CAS, $a=0$ and $b=12$. Write the equation.

(d) Predict the current when $R=8$ ohms.

Q2. A least-squares line fitted to transformed data is $y=3+60\frac{1}{x}$. (a) Predict y when $x=5$. (b) Predict y when $x=20$. (c) State what value y approaches as x becomes very large, and explain why.

Q3. For the data below, a reciprocal transformation of y is used.

x	2	3	6	8	14
y	5	4	2.5	2	1.25

(a) Copy and complete the row of reciprocals $\frac{1}{y}$.

(b) The least-squares line of $\frac{1}{y}$ against x is $\frac{1}{y}=0.1+0.05x$. State the slope and intercept.

(c) Predict y when $x=10$, remembering to back-transform.

Q4. A scatterplot of two numerical variables, with x explanatory and y the response, shows the points dropping steeply for small values of x and then levelling off towards the horizontal line $y=20$ as x increases. (a) Is a straight-line model appropriate for these data? Explain. (b) Name a transformation that could straighten this curve. (c) Both reciprocal transformations are candidates for a curve of this shape. State which variable you would reciprocate, using the level the data flatten towards to justify your choice.

Q5. A straight line was fitted to some curved data and the residual plot was drawn. The residuals were negative for small and large x and positive in the middle, forming a $\cap$-shaped pattern. (a) What does this residual plot tell you about the straight-line model? (b) After a suitable reciprocal transformation, what should the new residual plot look like if the transformation has succeeded?

Q6. For a fixed mass of gas at constant temperature, the pressure P (kPa) was measured at several volumes V (litres):

Volume, V (litres)	1	2	4	5	10
Pressure, P (kPa)	60	30	15	12	6

(a) Complete the reciprocals $\frac{1}{V}$.

(b) The least-squares line of P against $\frac{1}{V}$ is $P=0+60\frac{1}{V}$. Write it as $P=\frac{60}{V}$.

(c) Predict the pressure when $V=3$ litres.

(d) Interpret the slope 60 in terms of the product $P\times V$.

Unscaffolded

Q7. A car travels a fixed route at various average speeds s (km/h); the travel time t (hours) is recorded:

Speed, s (km/h)	30	50	60	75	100
Time, t (hours)	10	6	5	4	3

Show that reciprocating s linearises the data, state the least-squares line $t=a+b\frac{1}{s}$ given $a=0$, $b=300$, and predict the time at a speed of 120 km/h.

Q8. For the data below, a reciprocal transformation of y gives the least-squares line $\frac{1}{y}=0.01+0.01x$.

x	1	3	4	7	9
y	50	25	20	12.5	10

Predict y when $x=6$.

Q9. A reciprocal transformation may be applied to the explanatory variable or to the response variable, but only to one of them. (a) Explain why applying it to both axes at once is not part of this method. (b) A set of data is straightened by reciprocating the response variable. Give the form of the least-squares line that is fitted, and explain how a prediction for y is obtained from it.

Q10. The travel time t (hours) for a fixed journey and the average speed s (km/h) are modelled by the least-squares line $t=320\frac{1}{s}$. (a) Interpret the slope 320 in the context of the journey. (b) Predict the travel time at an average speed of 80 km/h. (c) Explain why the intercept is 0 for this relationship.

Q11. The average response time t (milliseconds) of a website was recorded for different numbers of servers s:

Servers, s	1	2	3	5	6	9
Response time, t (ms)	92	44	31	17	16	9

A reciprocal transformation of s gives the least-squares line $t=-0.74+92.4\frac{1}{s}$ with $r=0.999$. (a) Predict the response time with 4 servers. (b) The straight-line fit of t against s had $r=-0.82$. Explain what the change in r shows.

Q12. For fixed distance travel, the model $t = 320\frac{1}{s}$ from Q10 was found. A second journey gives the model $t = 500\frac{1}{s}$. (a) Which journey is longer, and how can you tell from the equations? (b) Predict the time for the second journey at 100 km/h.

Q13. The value y of a quantity was recorded against x:

x	1	2	4	6	10
y	16	12	8	6	4

A reciprocal transformation of y gives the least-squares line $\frac{1}{y} = 0.0417 + 0.0208x$ with $r = 1$. (a) Predict y when $x = 3$. (b) State whether this prediction is an interpolation or an extrapolation.

Q14. Explain, in your own words, the full modelling process when a scatterplot of two numerical variables is curved: what you check, what you transform, how you confirm the transformation, and how you make a prediction.

Q15. Two residual plots are drawn for the same data. The first, from a straight-line fit, shows a clear curved band of residuals. The second, from a least-squares line fitted after reciprocating the explanatory variable, shows residuals scattered randomly about zero. (a) Which model should be used for prediction, and why? (b) What would you conclude if the second residual plot also showed a curved pattern?

Q16. A construction job requires a fixed amount of work. The time t (days) to finish it with w workers was recorded:

Workers, w	2	3	4	6	8
Time, t (days)	12	8	6	4	3

The least-squares line of t against $\frac{1}{w}$ is $t = 24\frac{1}{w}$. (a) Predict the time to finish the job with 5 workers. (b) Interpret the slope 24 in terms of the total amount of work.

Q17. A straight-line model of some curved data has correlation coefficient $r = -0.87$. After a reciprocal transformation of the explanatory variable, the least-squares line of y against $\frac{1}{x}$ has $r = 0.998$. (a) What does the change in the value of r indicate about the transformation? (b) Which model should be used to make predictions?

Q18. The reciprocal model $I = 12\frac{1}{R}$ relates the current I (amps) to the resistance R (ohms) for resistances between 2 and 12 ohms. (a) Predict the current at $R = 10$ ohms. (b) A student uses the model to predict the current at $R = 0.5$ ohms. Explain why this prediction should be treated with caution.

Q19. The light intensity L (lux) at a distance d (metres) from a lamp along a bench was recorded:

Distance, d (m)	1	2	3	4	6
Intensity, L (lux)	124	64	44	34	24

A reciprocal transformation of d gives the least-squares line $L = 4 + 120\frac{1}{d}$ with $r = 1$. (a) Predict the intensity at a distance of 5 metres. (b) Interpret the intercept 4 in the context of the model.

Q20. After exercise, a person's heart rate above their resting rate, h (beats per minute above rest), was recorded m minutes into recovery:

Time, m (minutes)	1	2	3	4	5	6
Rate above rest, h (bpm)	82	38	28	19	17	13

(a) The shape of the data suggests a reciprocal transformation, and both were tried: the least-squares line of h against $\frac{1}{m}$ has $r = 0.998$, while the line of $\frac{1}{h}$ against m has $r = 0.995$. State which variable you would reciprocate, and give a reason.

(b) The least-squares line of h against $\frac{1}{m}$ is $h = -0.61 + 81.9\frac{1}{m}$ with $r = 0.998$. Predict the rate above rest at $m = 8$ minutes.

(c) State whether the prediction in (b) is reliable, giving a reason.

Answers

Q1. (a) $0.500, 0.333, 0.250, 0.167, 0.083$; (b) the explanatory variable R; (c) $I = 12\frac{1}{R}$; (d) 1.5 amps. **Q2.** (a) 15; (b) 6; (c) $y \to 3$, because $60\frac{1}{x} \to 0$ as x grows, leaving the intercept. **Q3.** (a) $0.2, 0.25, 0.4, 0.5, 0.8$; (b) slope 0.05, intercept 0.1; (c) $\frac{1}{y} = 0.6$, so $y \approx 1.67$. **Q4.** (a) No — the data are curved, not straight; (b) a reciprocal transformation; (c) the explanatory variable — $y = a + b\frac{1}{x}$ levels off at $y = a \approx 20$, while reciprocating y would force the model down towards 0. **Q5.** (a) The pattern shows the association is non-linear, so a straight line is not a good model; (b) a random scatter about zero with no pattern. **Q6.** (a) $1, 0.5, 0.25, 0.2, 0.1$; (b) $P = \frac{60}{V}$; (c) 20 kPa; (d) the slope 60 is the constant product $P \times V$. **Q7.** $t = 300\frac{1}{s}$; at $s = 120$, $t = 2.5$ hours. **Q8.** $\frac{1}{y} = 0.07$, so $y \approx 14.29$. **Q9.** (a) The method straightens the curve by transforming one axis; reciprocating both changes the model entirely and is not this technique; (b) fit $\frac{1}{y} = a + bx$; substitute x, then back-transform: $y = \frac{1}{a + bx}$. **Q10.** (a) The slope 320 is the fixed distance of the journey (320 km); (b) 4 hours; (c) at "infinite" speed the time would be 0, so the line passes through the origin. **Q11.** (a) about 22.4 ms; (b) r moved from -0.82 to 0.999 (nearly 1), so the transformed data are far more linear — the transformation worked. **Q12.** (a) The second journey is longer — its slope (500) is larger, and the slope is the distance; (b) 5 hours. **Q13.** (a) $\frac{1}{y} = 0.104$, so $y \approx 9.6$; (b) interpolation ($x = 3$ is within 1–10). **Q14.** Check the scatter/residual plot for curvature; try a reciprocal transformation — either axis is a candidate, so fit one and, if a pattern remains, the other; confirm with r near ± 1 and a patternless residual plot; predict by substituting and back-transforming if y was reciprocated. **Q15.** (a) The transformed model, because its residual plot has no pattern (the linearity assumption holds); (b) that this reciprocal transformation has not straightened the data — reciprocating the other variable, or a different transformation, would be needed. **Q16.** (a) 4.8 days; (b) the slope 24 is the total work in worker-days. **Q17.** (a) r close to 1 means the transformed data are almost perfectly linear, so the transformation straightened the curve; (b) the transformed (reciprocal) model. **Q18.** (a) 1.2 amps; (b) $R = 0.5$ is outside the data range (2–12 ohms), so it is an extrapolation and the model may not hold there. **Q19.** (a) 28 lux; (b) the intercept 4 is the intensity the model approaches as the distance becomes very large (a background level of about 4 lux). **Q20.** (a) the explanatory variable m — both transformations straighten the data, but h against $\frac{1}{m}$ has the larger r (0.998 against 0.995) and leaves h in bpm, so no back-transformation is needed; (b) about 9.6 bpm above rest; (c) not reliable — $m = 8$ is beyond the data (1–6 minutes), so it is an extrapolation.

Fully-worked solutions

Q1. (a) The reciprocals are $\frac{1}{2} = 0.500$, $\frac{1}{3} = 0.333$, $\frac{1}{4} = 0.250$, $\frac{1}{6} = 0.167$, $\frac{1}{12} = 0.083$. (b) Because we reciprocated R, the **explanatory** variable has been transformed. (c) With $a = 0$ and $b = 12$ the line is $I = 12\frac{1}{R}$. (d) At $R = 8$, $I = 12 \times \frac{1}{8} = 1.5$ amps.

Q2. (a) $y = 3 + 60 \times \frac{1}{5} = 3 + 12 = 15$. (b) $y = 3 + 60 \times \frac{1}{20} = 3 + 3 = 6$. (c) As x becomes very large, $\frac{1}{x} \to 0$, so $60\frac{1}{x} \to 0$ and y approaches the intercept, 3. The line $y = 3$ is the horizontal asymptote.

Q3. (a) The reciprocals are $\frac{1}{5}=0.2$, $\frac{1}{4}=0.25$, $\frac{1}{2.5}=0.4$, $\frac{1}{2}=0.5$, $\frac{1}{1.25}=0.8$. (b) In $\frac{1}{y}=0.1+0.05\,x$ the slope is 0.05 and the intercept is 0.1. (c) At $x=10$, $\frac{1}{y}=0.1+0.05\times 10=0.6$; back-transform: $y=\frac{1}{0.6}\approx 1.67$.

Q4. (a) No. The points lie on a curve, not a straight line, so a linear model would fit poorly and its residual plot would show a pattern. (b) The steep-then-flat shape is the reciprocal shape, so a **reciprocal transformation** could straighten it. (c) The data level off towards $y=20$, well above zero. Fitting $y=a+b\frac{1}{x}$ gives a model with the horizontal asymptote $y=a$, which can sit at 20; reciprocating the response instead gives $y=\frac{1}{a+b\,x}$, which falls away towards 0 and cannot level off at 20. So reciprocate the **explanatory** variable and fit $y=a+b\frac{1}{x}$, confirming the choice with r and the residual plot.

Q5. (a) A residual plot should show a random scatter about zero if the straight-line model is appropriate. A clear ∩-shaped pattern means the association is **non-linear**, so a straight line is not a good model and a transformation is needed. (b) If the reciprocal transformation has succeeded, the residuals of the transformed fit should be **randomly scattered about zero** with no pattern.

Q6. (a) $\frac{1}{1}=1$, $\frac{1}{2}=0.5$, $\frac{1}{4}=0.25$, $\frac{1}{5}=0.2$, $\frac{1}{10}=0.1$. (b) With intercept 0 and slope 60, $P=60\frac{1}{V}=\frac{60}{V}$. (c) At $V=3$, $P=\frac{60}{3}=20$ kPa. (d) Rearranging, $P\times V=60$, so the slope is the constant product of pressure and volume — the pressure and volume are inversely proportional.

Q7. The travel time falls steeply and then flattens as the speed increases — the reciprocal shape. For a fixed route, time × speed = distance is constant, so t should be proportional to $\frac{1}{s}$: reciprocate the speed. The reciprocals $\frac{1}{s}$ are $0.0333, 0.0200, 0.0167, 0.0133, 0.0100$, and a plot of t against $\frac{1}{s}$ is a straight line. With $a=0$ and $b=300$ the model is $t=300\frac{1}{s}=\frac{300}{s}$ (the slope 300 is the distance in km). At $s=120$, $t=\frac{300}{120}=2.5$ hours.

Q8. Substitute $x=6$ into $\frac{1}{y}=0.01+0.01\,x$: $\frac{1}{y}=0.01+0.01\times 6=0.07$. Because the response was reciprocated, back-transform: $y=\frac{1}{0.07}\approx 14.29$. ($x=6$ lies within the observed values 1–9, so this is an interpolation.)

Q9. (a) The reciprocal transformation straightens a curve by reciprocating **one** axis, leaving the other unchanged, so that a line can be fitted to the transformed pair. Reciprocating both axes produces a different model altogether and is not part of this single-axis technique. (b) Reciprocating the response means the straight line is fitted to $\frac{1}{y}$ against x, so the line has the form $\frac{1}{y}=a+b\,x$. To predict, substitute the value of x to obtain a value for $\frac{1}{y}$, then **back-transform** by taking the reciprocal: $y=\frac{1}{a+b\,x}$.

Q10. (a) Rearranging $t=320\frac{1}{s}$ gives $t\times s=320$; for a fixed journey, $\text{time}\times\text{speed}=\text{distance}$, so the slope 320 is the **distance**, 320 km. (b) At $s=80$, $t=320\times\frac{1}{80}=4$ hours. (c) At an extremely high speed the journey would take almost no time, so as $\frac{1}{s}\to 0$ the time $\to 0$; the line therefore passes through the origin and the intercept is 0.

Q11. (a) At $s=4$, $t=-0.74+92.4\times\frac{1}{4}=-0.74+23.1=22.36\approx 22.4$ ms (and $s=4$ lies within the data 1–9, so this is interpolation). (b) The straight-line fit had $r=-0.82$, only moderately strong, while the transformed fit has $r=0.999$, almost exactly 1. The jump towards 1 shows the reciprocal transformation has straightened the data, so the transformed model fits far better.

Q12. (a) Each model has the form $t=\text{distance}\times\frac{1}{s}$, so the slope is the distance. The second journey has the larger slope (500 against 320), so it is the **longer** journey — 500 km against 320 km. (b) At $s=100$, $t=500\times\frac{1}{100}=5$ hours.

Q13. (a) At $x=3$, $\frac{1}{y}=0.0417+0.0208\times 3=0.0417+0.0624=0.1041$; back-transform: $y=\frac{1}{0.1041}\approx 9.6$. (b) $x=3$ lies between the smallest and largest observed x (1 and 10), so it is an **interpolation**.

Q14. First **check for curvature**: draw the scatterplot, or fit a straight line and inspect its residual plot. A curved band or a patterned residual plot shows the association is non-linear. Next **transform**: the steep-then-flat reciprocal shape says a reciprocal transformation is worth trying, but not which axis — reciprocating the explanatory variable (fit $y=a+b\frac{1}{x}$) and reciprocating the response (fit $\frac{1}{y}=a+bx$) are both candidates, so choose one, using the level the data flatten towards as a guide. Then **confirm**: the transformed data should have r close to ± 1 and a residual plot with no pattern; if a pattern remains, fit the other reciprocal and keep the better model. Finally **predict** by substituting into the fitted line, and, if the response was reciprocated, back-transform by taking the reciprocal of the result — using interpolation where possible and treating extrapolation with caution.

Q15. (a) Use the **transformed** model. A residual plot with no pattern means the linearity assumption is satisfied for the transformed data, whereas the curved residual plot of the straight-line fit shows that model is inadequate. (b) A curved pattern remaining in the second residual plot would mean the reciprocal transformation did **not** straighten the data — a different transformation (or a different choice of axis) would be needed.

Q16. (a) At $w=5$, $t=24\times\frac{1}{5}=4.8$ days. (b) Rearranging, $t\times w=24$, so the number of days times the number of workers is constant; the slope 24 is the **total work**, measured in worker-days.

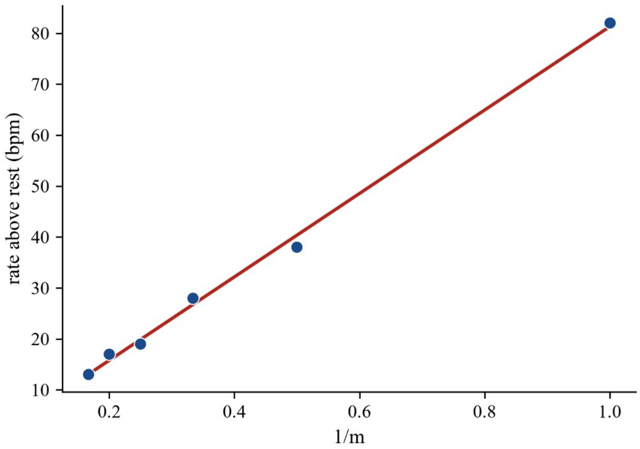
Q17. (a) The correlation coefficient measures how linear the data are. Moving from $r=-0.87$ for the raw straight-line fit to $r=0.998$ for the transformed fit means the transformed points are almost perfectly linear, so the reciprocal transformation has straightened the curve. (b) Predictions should be made from the **transformed** model $y=a+b\frac{1}{x}$, which fits the data well.

Q18. (a) At $R=10$, $I=12\times\frac{1}{10}=1.2$ amps (within the data range 2–12 ohms, so reliable). (b) $R=0.5$ ohms lies **outside** the range of resistances used to build the model (2–12 ohms). Using the model there is an **extrapolation**: the reciprocal pattern is only known to hold within the measured range, and may break down beyond it, so the prediction is unreliable.

Q19. (a) At $d=5$, $L=4+120\times\frac{1}{5}=4+24=28$ lux. (b) As the distance d becomes very large, $120\times\frac{1}{d}\to 0$, so L approaches the intercept, 4. The intercept 4 is therefore the background intensity (about 4 lux) that the light level tends towards far from the lamp.

Q20. (a) The rate above rest falls steeply and then flattens as time increases — the reciprocal shape — so a reciprocal transformation is called for, and both variables are candidates. The fits decide it: h against $\frac{1}{m}$ gives $r=0.998$ against

$r = 0.995$ for $\frac{1}{h}$ against m, and reciprocating the **explanatory** variable m also leaves h in beats per minute, so no back-transformation is needed. Plotting h against $\frac{1}{m}$ gives a straight-line scatter, confirming the choice.

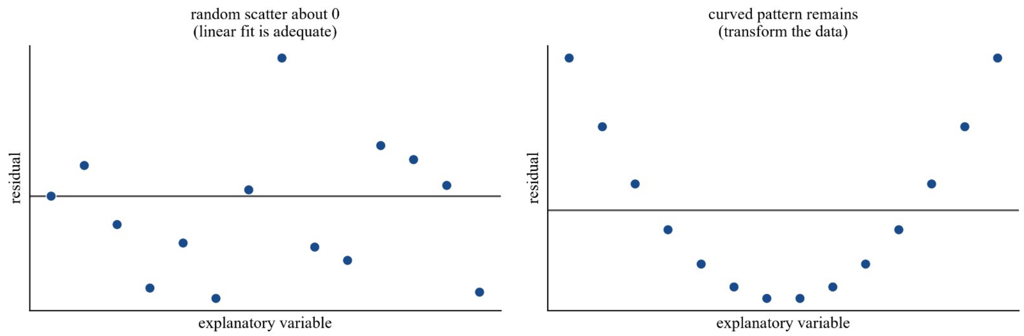


(b) At $m = 8$, $h = -0.61 + 81.9 \times \frac{1}{8} = -0.61 + 10.2375 = 9.6275 \approx 9.6$ bpm above rest. (c) The prediction is **not** reliable: $m = 8$ minutes is beyond the observed range (1–6 minutes), so it is an extrapolation and assumes the pattern continues past the data.

Theory

A least-squares line is only a good model when the association between the two numerical variables is **linear**. When a scatterplot is **curved**, a straight line misses the shape of the data: it fits badly, and its **residual plot** shows a clear pattern instead of a random band. The remedy studied in the previous sections of this chapter is a **data transformation** — we replace one of the variables by its **square**, its **logarithm** (base 10) or its **reciprocal**, so that the transformed data fall along a straight line. This section brings those ideas together and answers the central question: **which transformation should you choose, and how do you carry it through to a model and a prediction?**

The residual-plot signal. After fitting a least-squares line, plot the residuals against the explanatory variable. If the residuals form a **random band about zero**, the linear model is adequate. If they form a **curved pattern** (for example a $\cup$ or a $\cap$ shape), the association is not linear and the data should be transformed.



The six transformations. A transformation is **applied to one axis only** — to the explanatory variable x **or** to the response variable y, never both at once. With three functions and two choices of axis there are six options:

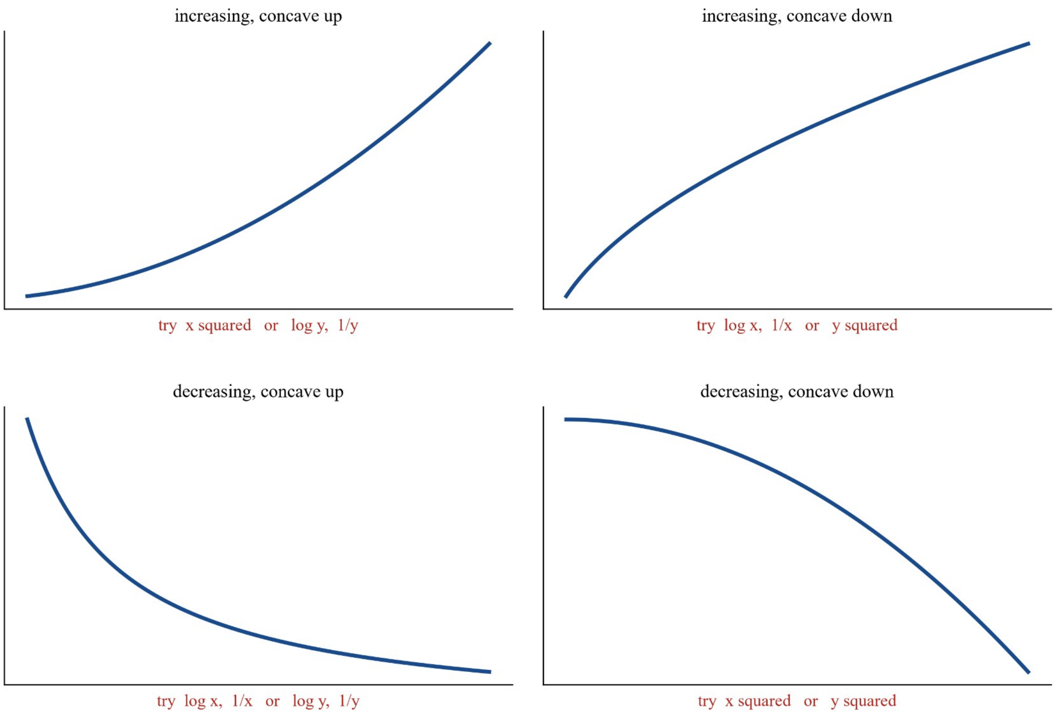
$$x^2, \log_{10} x, \frac{1}{x}, y^2, \log_{10} y, \frac{1}{y}.$$

What matters is how each function changes the **spacing** of the values it is applied to:

Transformation	Effect on the larger values	Strength
square (x^2 or y^2)	**stretches** them further apart	—
logarithm ($\log_{10} x$ or $\log_{10} y$)	**compresses** them closer together	moderate
reciprocal ($\frac{1}{x}$ or $\frac{1}{y}$)	**compresses** them strongly (and reverses their order)	strong

Straightening a curve is a matter of stretching the axis that is too "bunched" or compressing the axis that is too "stretched" until the points line up.

Matching the transformation to the shape. Look at the **direction** (increasing or decreasing) and the **concavity** (does the curve bend upward like a valley, or downward like a dome?). Each of the four curved shapes can be straightened by transforming **either** axis, which is why more than one transformation can work.



Read the chart like this. For an **increasing, concave-up** curve the large x-values are crowded under a steeply rising tail, so we either **stretch the large** x with x^2, **or** pull down the large y-values by compressing y with $\log_{10} y$ or $\dfrac{1}{y}$. For an **increasing, concave-down** curve the reasoning is reversed: **compress the large** x with $\log_{10} x$ or $\dfrac{1}{x}$, **or stretch the** y with y^2. The two decreasing cases follow the same logic. A **decreasing, concave-up** curve drops steeply and then flattens along a long tail, so **compress the** x**-axis** with $\log_{10} x$ or $\dfrac{1}{x}$ to pull that tail in and even out the fall, **or** compress the y-axis with $\log_{10} y$ or $\dfrac{1}{y}$ to pull in the widely spaced large y-values at the steep left-hand end. A **decreasing, concave-down** curve does the reverse — it falls slowly and then plunges — so an axis must be **stretched**: x^2 spreads apart the large x-values under the plunge, **or** y^2 spreads apart the bunched large y-values along the flat start.

Several transformations may help — choose the best fit. The chart gives **candidates**, not a single answer. Because the three functions differ in strength, one usually straightens the data better than the others, and a transformation that is **too strong** will **overshoot** — instead of removing the curvature it over-corrects, leaving the transformed points bent the other way. Since the reciprocal is the strongest of the three, it is the one that overshoots most often. So you **try a candidate, then judge it** by two things:

- the **residual plot** of the transformed fit should now show a **random band** (no leftover pattern);
- the **coefficient of determination** r^2 should be **larger** (closer to 1) than for the untransformed line — and, when comparing candidates, you keep the transformation giving the **largest** r^2.

A method for the whole process.

1. Draw the scatterplot; if in doubt, fit a line and look at its residual plot to confirm non-linearity.
2. Note the direction and concavity, and read the candidate transformations from the chart.
3. For a candidate, compute the transformed variable, fit the least-squares line to the transformed data, and check its residual plot and r^2.
4. If more than one candidate is plausible, compare their r^2 values and keep the best.
5. **Write the model** in terms of the transformed variable.
6. **Predict** by substituting into the model — transforming the given x first, and **back-transforming** at the end if the transformation was applied to y.

Writing the model and predicting. With the least-squares line written as $\hat{y}=a+b\,x$, the six transformations give the model forms and back-transformations below (only the transformed term changes):

Transform	Model (fitted to the transformed data)	To predict $\hat{y}$
x^2	$\hat{y}=a+b\,x^2$	substitute x
$\log_{10} x$	$\hat{y}=a+b\log_{10} x$	substitute x
$\frac{1}{x}$	$\hat{y}=a+b\left(\frac{1}{x}\right)$	substitute x
y^2	$\hat{y}^2=a+b\,x$	$\hat{y}=\sqrt{a+b\,x}$
$\log_{10} y$	$\log_{10}\hat{y}=a+b\,x$	$\hat{y}=10^{a+b\,x}$
$\frac{1}{y}$	$\frac{1}{\hat{y}}=a+b\,x$	$\hat{y}=\frac{1}{a+b\,x}$

When the transformation is on x the response $\hat{y}$ is read straight off; when it is on y you must **undo** the transformation (square root, power of 10, or reciprocal) to return to the original units. As always, a prediction made **outside** the range of the data is an **extrapolation** and should be treated with caution.

Using CAS. General Mathematics is technology-active. In practice you enter the data, create a new column holding the transformed variable (x^2, $\log_{10} x$, $1/x$, and so on), and run the least-squares routine on the transformed column to read off a, b and r^2. Learn the by-hand steps below so that you can choose the transformation sensibly and check that the CAS output is reasonable.

Worked examples

Example 1 — scaffolded

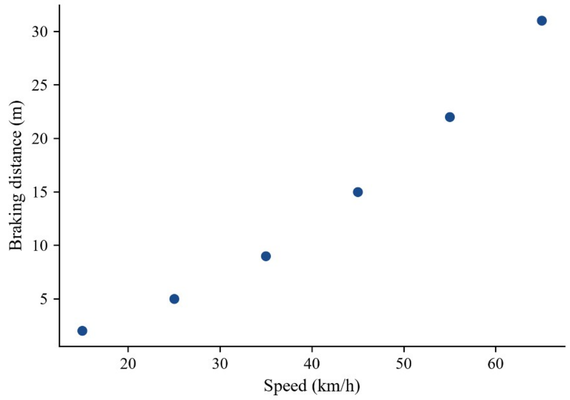
The braking distance of a car was measured at six speeds.

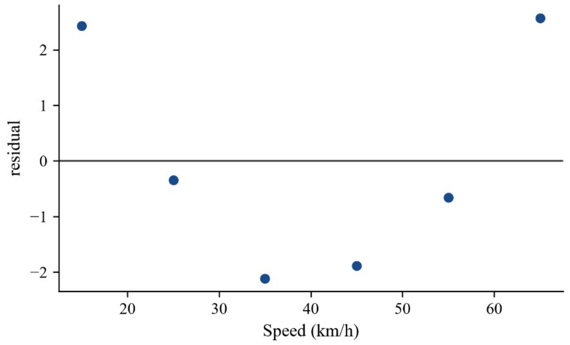
Speed s (km/h)	15	25	35	45	55	65
Braking distance d (m)	2	5	9	15	22	31

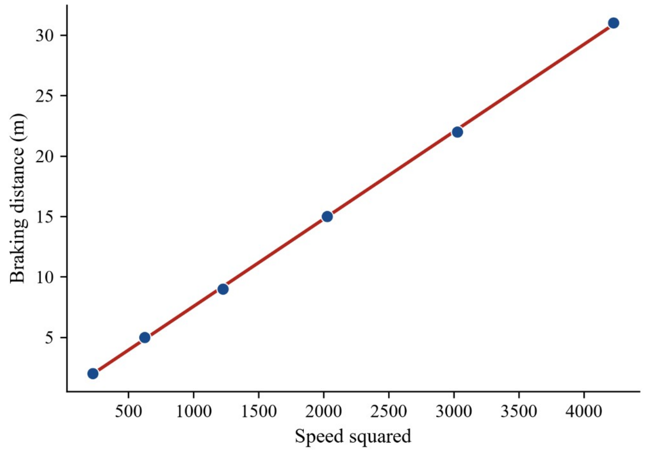
Choose a suitable transformation, fit a model and predict the braking distance at 50 km/h.

Working	Comment
The scatterplot rises and is **concave up** (the gap between successive points grows).	Describe direction and concavity first.
A straight line gives $r^2=0.9651$, but its residual plot is a clear $\cup$ shape.	High r^2 alone is not enough — the residual pattern reveals the curvature.
Concave up $\Rightarrow$ candidates are s^2 (stretch x) or $\log_{10} d\,, 1/d$ (compress y).	Read the candidates from the chart.
Fitting d against s^2 gives $r^2=0.9998$; fitting $\log_{10} d$ against s gives $r^2=0.9664$, and $1/d$ against s gives $r^2=0.7340$.	Compare candidates by r^2; the two y-compressions both over-correct here.
The s^2 transformation wins, and its residual plot is now a random band.	Largest r^2 **and** a patternless residual plot.
Least squares on (s^2, d) gives $\hat{d}=0.336+0.00722\,s^2$.	The model, written with the transformed term s^2.

Working	Comment
At $s = 50$: $\hat{d} = 0.336 + 0.00722 \times 50^2 = 0.336 + 18.05 = 18.4$ m.	Transform the given x ($50^2 = 2500$), then substitute.







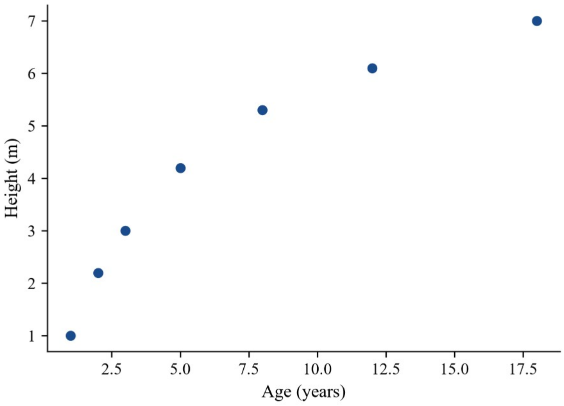
Example 2 — scaffolded

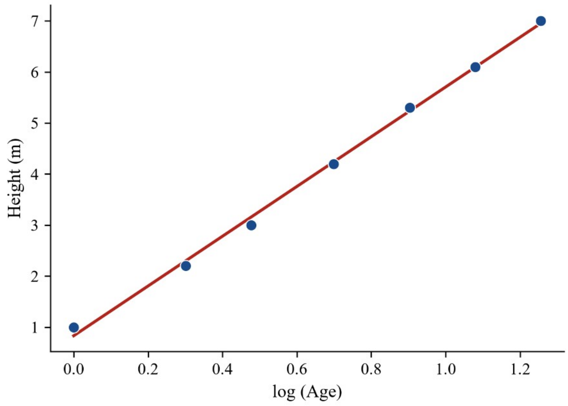
The height of a young tree was recorded at several ages.

Age a (years)	1	2	3	5	8	12	18
Height h (m)	1.0	2.2	3.0	4.2	5.3	6.1	7.0

Choose a suitable transformation, write the model and predict the height at age 6.

Working	Comment
The scatterplot increases but is **concave down** — it rises quickly then levels off.	Direction increasing, concavity downward.
Concave down $\Rightarrow$ candidates are $\log_{10} a$, $1/a$ (compress x) or h^2 (stretch y).	Read the candidates from the chart.
r^2: h against $\log_{10} a$, 0.9975; h^2 against a, 0.9734; h against $1/a$, 0.8204.	Two candidates help; $\log_{10} a$ is best and $1/a$ overshoots.
Least squares on $(\log_{10} a, h)$ gives $\hat{h} = 0.831 + 4.875 \log_{10} a$.	The model, with the transformed term $\log_{10} a$.
At $a = 6$: $\hat{h} = 0.831 + 4.875 \times \log_{10} 6 = 0.831 + 4.875 \times 0.7782 = 4.$ m.	$\log_{10} 6 = 0.7782$; the age 6 lies inside the data, so this is interpolation.





Example 3 — unscaffolded

A bacterial culture was counted every hour.

Time t (hours)	0	1	2	3	4	5
Count N	50	82	128	205	340	560

Choose a transformation, write the model and estimate the count at $t=6$ hours.

The scatterplot increases and is **concave up**, with the count multiplying by roughly the same factor each hour — the signature of exponential growth. A straight line gives only $r^2=0.8870$. For a concave-up curve the candidates are t^2 (stretch x) or a compression of y. Transforming y with the logarithm gives $r^2=0.9996$, better than the t^2 transformation ($r^2=0.9918$), and its residual plot is a random band, so we transform y with $\log_{10}$.

Fitting $\log_{10} N$ against t gives

$$\log_{10}\widehat{N}=1.697+0.209\,t\,.$$

Because the transformation is on y, a prediction must be **back-transformed**. At $t=6$,

$$\log_{10}\widehat{N}=1.697+0.209\times 6=2.951\,,\ \widehat{N}=10^{2.951}\approx 893\,.$$

Note that $t=6$ is just beyond the data, so this is an extrapolation.

$\therefore$ the model is $\log_{10}\widehat{N}=1.697+0.209\,t$, predicting about 890 bacteria at $t=6$ hours.

Example 4 — unscaffolded

The time to build a wall was recorded for different numbers of workers.

Workers w	2	3	4	5	6	8
Time T (days)	38	25	18	15	12	9

Choose a transformation, write the model and predict the time for 10 workers.

The scatterplot **decreases** and is **concave up** — the time falls steeply at first, then flattens towards a floor as more workers are added. A straight line gives only $r^2=0.8255$, and for a decreasing concave-up curve the candidates compress an axis: $\log_{10} w$ or $\dfrac{1}{w}$ on x, or a compression of y. The reciprocal of w gives the best fit, $r^2=0.9992$ (compared with 0.9519 for $\log_{10} w$), with a random residual plot — which also makes physical sense, since more workers should halve the time when their number doubles.

Fitting T against $\frac{1}{w}$ gives

$$\hat{T} = -0.83 + 77.43\left(\frac{1}{w}\right).$$

The transformation is on x, so no back-transformation is needed. At $w = 10$,

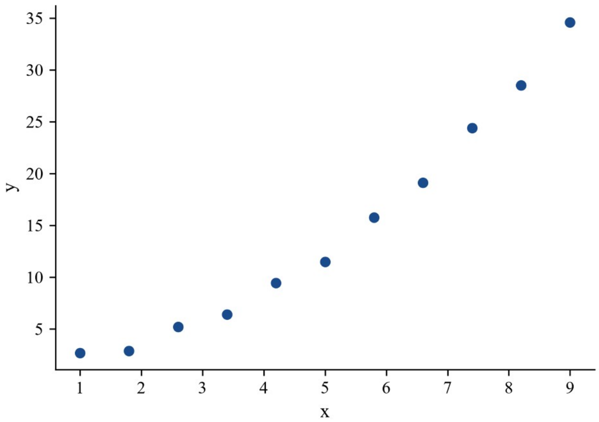
$$\hat{T} = -0.83 + 77.43 \times \frac{1}{10} = -0.83 + 7.743 = 6.9\,\text{days}.$$

$\therefore$ the model $\hat{T} = -0.83 + 77.43\left(\frac{1}{w}\right)$ predicts about 6.9 days for 10 workers.

Practice questions

Scaffolded

Q1. The scatterplot below shows a curved association between two variables.



(a) State the direction of the association (increasing or decreasing).
(b) State the concavity (concave up or concave down).
(c) List the transformations that could be tried to straighten this data.
(d) Explain how you would decide which one to use.

Q2. An object is dropped and the distance it has fallen is recorded each second.

Time t (s)	1	2	3	4	5	6
Distance d (m)	5	19	45	82	124	178

(a) The scatterplot is increasing and concave up. Name a transformation of t that should straighten it.
(b) A line fitted to the raw data gives $r^2 = 0.9620$; the t^2 transformation gives $r^2 = 0.9997$. State which model is better and why.
(c) The least-squares line for d against t^2 is $\hat{d} = 0.32 + 4.96t^2$. Predict the distance fallen after 7 seconds.

Q3. The response to a drug was measured at several doses.

Dose x (mg)	1	2	4	8	16	32
Response y	10	23	33	46	57	70

(a) The data rise quickly then level off (concave down), and the doses span a wide range. Name a suitable transformation of x.

(b) Fitting y against $\log_{10} x$ gives $\hat{y} = 10.19 + 39.39 \log_{10} x$. State the value of r^2 you would expect to be higher: the raw fit or the transformed fit.

(c) Predict the response at a dose of 24 mg ($\log_{10} 24 = 1.3802$).

Q4. A colony of plants is counted each week.

Time t (weeks)	0	1	2	3	4	5
Number P	20	31	52	80	128	205

(a) The growth is concave up and multiplies by a roughly constant factor each week. Name a transformation of P that should straighten it.

(b) The least-squares line for $\log_{10} P$ against t is $\log_{10} \widehat{P} = 1.299 + 0.203\,t$. Write the step needed to turn a value of $\log_{10} \widehat{P}$ back into a number of plants.

(c) Predict the number of plants at $t = 6$ weeks.

Q5. The time to drain a tank depends on how many pumps are used.

Pumps x	1	2	3	4	6
Time T (min)	61	29	21	15	10

(a) The scatterplot decreases and is concave up. Name a transformation of x that should straighten it.

(b) The least-squares line for T against $\frac{1}{x}$ is $\widehat{T} = -0.19 + 60.87\left(\frac{1}{x}\right)$. State whether a prediction from this model needs back-transforming, and why.

(c) Predict the draining time with 5 pumps.

Q6. A student fitted four models to the same curved data — mass m (grams) against length ℓ (cm) — and recorded the coefficient of determination for each.

Model	raw (m vs ℓ)	ℓ^2	$\log_{10} \ell$	$\frac{1}{\ell}$
r^2	0.9525	0.9997	0.7956	0.5822

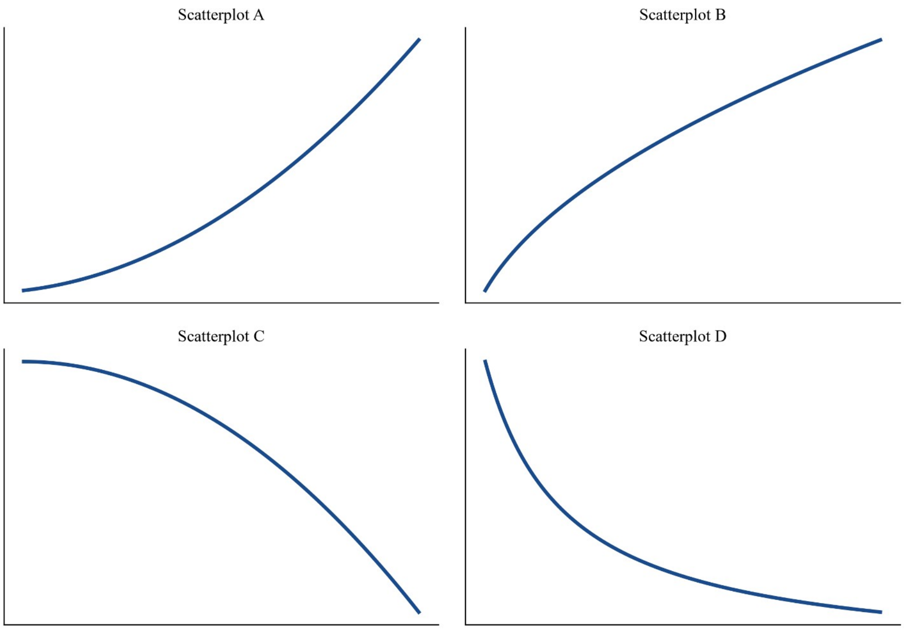
(a) Which transformation gives the best linear fit?

(b) The fitted model is $\widehat{m} = 3.88 + 0.628\,\ell^2$. Predict the mass when $\ell = 14$ cm.

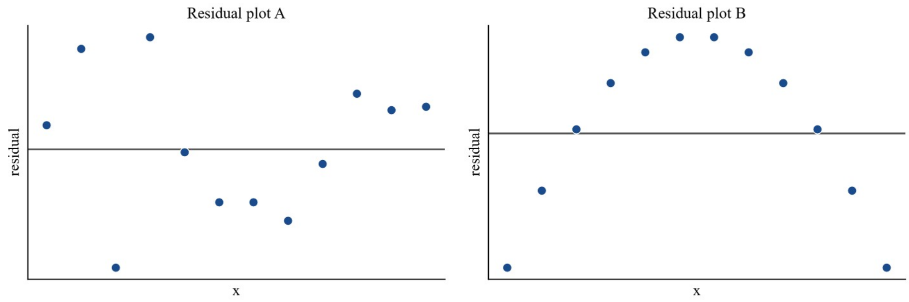
(c) Explain why the reciprocal transformation fits so poorly here.

Unscaffolded

Q7. For each scatterplot below, name a transformation (of x or of y) that could straighten it.



Q8. After fitting a straight line to some data, a student obtains the two residual plots below for two different candidate models.



State which model the student should keep, and justify your choice.

Q9. The power generated by a small wind turbine was measured at several wind speeds.

Speed v (m/s)	2	4	6	8	10
Power P (W)	6	25	55	102	152

The association is increasing and concave up. Fit the model $\hat{P} = 0.559 + 1.533\,v^2$ obtained from the v^2 transformation, and use it to predict the power at a wind speed of 12 m/s.

Q10. A skill score was measured against the number of practice sessions completed.

Sessions x	1	3	5	10	20	40
Score y	12	28	36	48	60	71

(a) Describe the shape of the association and name a suitable transformation.

(b) The transformed model is $\hat{y} = 10.95 + 37.27\log_{10} x$. Predict the score after 15 sessions.

Q11. An investment is valued at the start of each even-numbered year.

Time t (years)	0	2	4	6	8	10
Value V ($)	1000	1450	2250	3200	4850	7000

(a) Explain why a logarithmic transformation of V is appropriate.

(b) The model is $\log_{10}\widehat{V} = 3.000 + 0.0850\,t$. Predict the value after 12 years, and state one reason to treat the prediction with caution.

Q12. The time for a delivery van to complete a fixed route was recorded at several average speeds.

Speed s (km/h)	20	40	50	80	100
Time T (hours)	5.8	3.1	2.4	1.6	1.2

The reciprocal transformation of s gives $\widehat{T} = 0.144 + 113.87\left(\frac{1}{s}\right)$. Use it to predict the travel time at an average speed of 60 km/h.

Q13. Two variables x and y are related by the data below.

x	1	2	3	4	5	6	7
y	2	9	20	34	53	75	101

A student tries four transformations and finds $r^2 = 0.9615$ (raw), 0.9997 (x^2), 0.7990 ($\log_{10} x$) and 0.5652 ($\frac{1}{x}$). (a) State the best transformation and write down what it tells you about the shape of the data. (b) The best model is $\hat{y} = 0.89 + 2.06\,x^2$. Predict y when $x = 8$.

Q14. Explain why, in this course, a transformation is applied to **one axis only** and never to both x and y at the same time.

Q15. A curved scatterplot is increasing and concave up. A student decides to compress the y-axis but cannot decide between $\log_{10} y$ and $\frac{1}{y}$. Explain how the student should choose between these two candidates.

Q16. A model fitted after a $\log_{10} y$ transformation is $\log_{10}\hat{y} = 1.4 + 0.25\,x$. (a) Predict y when $x = 4$. (b) Predict y when $x = 8$.

Q17. For a set of data, a straight line has $r^2 = 0.72$ and a residual plot with a clear $\cap$-shaped pattern. A classmate says "the r^2 is high, so the linear model is fine". Explain why the classmate is wrong.

Q18. A scatterplot **decreases** and is **concave down** (it falls slowly at first, then more steeply). Using the transformation chart, name the transformations that could straighten it.

Q19. Two transformations of the same curved data give models with $r^2 = 0.981$ and $r^2 = 0.994$, and both residual plots look random. Which model should be reported, and why?

Q20. The stopping distance d (metres) of a vehicle is modelled from its speed s (km/h) by the transformed line $\hat{d} = 1.2 + 0.0065\,s^2$. (a) Predict the stopping distance at $s = 40$ km/h. (b) Predict the stopping distance at $s = 80$ km/h. (c) The speed in (b) is double that in (a). By what factor does the predicted stopping distance (approximately) increase, and why does doubling the speed **not** double the distance?

Answers

Q1. (a) Increasing; (b) concave up; (c) x^2, or $\log_{10} y$ or $\frac{1}{y}$; (d) fit each candidate, then keep the one with the most random residual plot and the largest r^2. **Q2.** (a) t^2; (b) the t^2 model, because its r^2 is closer to 1 (and its residual plot is random); (c) 243.4 m. **Q3.** (a) $\log_{10} x$; (b) the transformed ($\log_{10} x$) fit; (c) 64.6. **Q4.** (a) $\log_{10} P$; (b) raise 10 to the power of $\log_{10} \widehat{P}$, i.e. $\widehat{P} = 10^{\log_{10} \widehat{P}}$; (c) 329 plants. **Q5.** (a) $\frac{1}{x}$; (b) no — the transformation is on x, so $\widehat{T}$ is already in minutes; (c) 12.0 min. **Q6.** (a) ℓ^2; (b) 127.0 g; (c) $\frac{1}{\ell}$ **compresses** the explanatory axis, but an increasing concave-up curve needs that axis **stretched** (ℓ^2) or the response axis compressed — compressing ℓ bends the data further from a line, so r^2 drops below even the raw fit. **Q7.** A: increasing concave up — x^2 (or $\log_{10} y / \frac{1}{y}$); B: increasing concave down — $\log_{10} x / \frac{1}{x}$ (or y^2); C: decreasing concave down — x^2 (or y^2); D: decreasing concave up — $\log_{10} x / \frac{1}{x}$ (or $\log_{10} y / \frac{1}{y}$). **Q8.** Model A — its residual plot is a random band about 0, whereas B still shows a curved pattern, so B is not yet linear. **Q9.** 221.3 W. **Q10.** (a) Increasing, concave down (levels off), so $\log_{10} x$; (b) 54.8. **Q11.** (a) the value multiplies by a roughly constant factor (about 1.5) over each two-year step — geometric/exponential growth, which $\log_{10} V$ straightens; (b) about \$10 500; caution: $t = 12$ is an extrapolation beyond the data. **Q12.** 2.04 hours. **Q13.** (a) x^2 (largest r^2); the data are increasing and concave up; (b) 132.7. **Q14.** Transforming both axes changes the model into a form whose slope, intercept and predictions can no longer be interpreted simply in the original variables; the course keeps to single-axis transforms so the fitted line stays usable. **Q15.** Fit both, and keep whichever gives the more random residual plot and the larger r^2; the reciprocal is the stronger compression and may overshoot. **Q16.** (a) $10^{2.4} \approx 251$; (b) $10^{3.4} \approx 2512$. **Q17.** A high r^2 does not prove linearity; the $\cap$-shaped residual pattern shows systematic curvature the line has missed, so the data should be transformed. **Q18.** x^2 (stretch x) or y^2 (stretch y). **Q19.** The $r^2 = 0.994$ model, because both fits are sound (random residuals) so the larger r^2 gives the better linear fit. **Q20.** (a) 11.6 m; (b) 42.8 m; (c) about 4 times; squaring means the distance depends on s^2, and doubling s multiplies s^2 by 4.

Fully-worked solutions

Q1. The points climb from left to right, so the association is **increasing**; each step up is larger than the last, so the curve bends upward — **concave up**. From the chart, an increasing concave-up curve is straightened by stretching the x-axis with x^2, or by compressing the y-axis with $\log_{10} y$ or $\frac{1}{y}$. To choose, transform the data each way, fit a least-squares line, and keep the transformation whose residual plot is the most random and whose r^2 is closest to 1.

Q2. (a) The curve increases and is concave up, so stretching x with t^2 should straighten it. (b) The t^2 model has $r^2 = 0.9997$ compared with $r^2 = 0.9620$ for the raw line; since 0.9997 is closer to 1 (and the t^2 residual plot is a random band), the t^2 **model** is better. (c) Transform the x-value: $7^2 = 49$, then

$$\hat{d} = 0.32 + 4.96 \times 49 = 0.32 + 243.04 = 243.4\,\text{m}.$$

Q3. (a) The data increase, are concave down and the doses range over a wide scale, so a $\log_{10} x$ transformation is suitable (it compresses the large doses). (b) The **transformed** ($\log_{10} x$) fit will have the higher r^2, because it removes the curvature the straight line could not follow. (c) At $x = 24$, $\log_{10} 24 = 1.3802$, so

$$\hat{y} = 10.19 + 39.39 \times 1.3802 = 10.19 + 54.37 = 64.6.$$

Q4. (a) Multiplying by a constant factor each week is exponential growth, straightened by transforming the response with $\log_{10} P$. (b) Since the model gives $\log_{10} \widehat{P}$, raise 10 to that power: $\widehat{P} = 10^{\log_{10} \widehat{P}}$. (c) At $t = 6$,

$$\log_{10}\hat{P}=1.299+0.203\times 6=1.299+1.218=2.517\,,\,\hat{P}=10^{2.517}\approx 329\text{ plants.}$$

Q5. (a) A decreasing, concave-up curve that flattens towards a floor is straightened by $\frac{1}{x}$. (b) **No** back-transformation is needed: the transformation is on the explanatory variable x, so $\hat{T}$ comes out directly in minutes. (c) At $x=5$,

$$\hat{T}=-0.19+60.87\times\frac{1}{5}=-0.19+12.174=12.0\text{ min.}$$

Q6. (a) The largest coefficient of determination is 0.9997, for the ℓ^2 transformation, so it gives the best linear fit. (b) At $\ell=14$, $\ell^2=196$, so

$$\hat{m}=3.88+0.628\times 196=3.88+123.09=127.0\text{ g.}$$

(c) The winning model $\hat{m}=3.88+0.628\,\ell^2$ has a positive ℓ^2 coefficient, so the data are **increasing and concave up**, and for that shape the chart calls for **stretching** the explanatory axis (ℓ^2) or **compressing** the response axis ($\log_{10} m$ or $\frac{1}{m}$). The reciprocal $\frac{1}{\ell}$ does neither — it compresses the **explanatory** axis, the opposite of what this shape needs — so it bends the points further from a straight line instead of straightening them. The table shows this: both compressions of ℓ score below even the raw fit (0.9525), and the reciprocal, being the stronger of the two compressions, drops furthest ($r^2=0.5822$ compared with 0.7956 for $\log_{10}\ell$).

Q7. Reading direction and concavity: **A** rises and is concave up $\Rightarrow$ x^2 (or compress y with $\log_{10} y/\frac{1}{y}$). **B** rises and is concave down $\Rightarrow$ $\log_{10} x/\frac{1}{x}$ (or y^2). **C** falls and is concave down $\Rightarrow$ x^2 (or y^2). **D** falls and is concave up $\Rightarrow$ $\log_{10} x/\frac{1}{x}$ (or $\log_{10} y/\frac{1}{y}$).

Q8. Keep **Model A**. A residual plot tests linearity: A's residuals scatter randomly above and below 0 with no pattern, showing the transformed data are linear, while B's residuals still curve, showing B has not removed the non-linearity. The random residual plot, not merely a high r^2, is what confirms the model.

Q9. The v^2 transformation straightens the increasing, concave-up data. Transforming the given speed, $12^2=144$, so

$$\hat{P}=0.559+1.533\times 144=0.559+220.75=221.3\text{ W.}$$

(The transformation is on x, so no back-transformation is required.)

Q10. (a) The score rises quickly at first and then levels off — **increasing and concave down** — and the sessions span a wide range, so a $\log_{10} x$ transformation is appropriate. (b) At $x=15$, $\log_{10} 15=1.1761$, so

$$\hat{y}=10.95+37.27\times 1.1761=10.95+43.83=54.8\,.$$

Q11. (a) Each recorded step is two years, and over each of them the value grows by roughly the same **factor**: $1450\div 1000=1.45$, $2250\div 1450=1.55$, $3200\div 2250=1.42$, and so on — about 1.5 every two years. That is geometric growth, and a $\log_{10} V$ transformation turns constant-factor growth into a straight line. (b) At $t=12$,

$$\log_{10}\hat{V}=3.000+0.0850\times 12=3.000+1.020=4.020\,,\,\hat{V}=10^{4.020}\approx \$\,10\,500\,.$$

It should be treated with caution because $t=12$ lies **outside** the data range (0 to 10 years) — an extrapolation that assumes the growth pattern continues unchanged.

Q12. The reciprocal of s straightens the decreasing, concave-up data. At $s=60$,

$$\hat{T}=0.144+113.87\times\frac{1}{60}=0.144+1.898=2.04\text{ hours.}$$

Q13. (a) The largest r^2 is 0.9997, for the x^2 transformation; that x^2 (not $\log_{10} x$ or $\frac{1}{x}$) straightens the data tells us the trend is **increasing and concave up**. (b) At $x = 8$, $x^2 = 64$, so

$$\hat{y} = 0.89 + 2.06 \times 64 = 0.89 + 131.84 = 132.7.$$

Q14. A single-axis transformation replaces one variable by a simple function of itself, leaving a model whose slope and intercept can still be read in terms of the original variables and whose predictions back-transform cleanly. Transforming **both** axes produces a relationship between two transformed quantities that no longer corresponds to a straightforward model of y in terms of x, so the course restricts itself to transforming one axis only.

Q15. Both $\log_{10} y$ and $\frac{1}{y}$ compress the large y-values, but by different amounts (the reciprocal is stronger). The student should **fit both**, examine the residual plots, and compute r^2 for each, then keep the transformation whose residual plot is the more random and whose r^2 is larger. If the reciprocal overshoots (its residuals curve the opposite way), the logarithm is the better choice.

Q16. The transformation is on y, so each prediction is back-transformed with a power of 10. (a) At $x = 4$: $\log_{10} \hat{y} = 1.4 + 0.25 \times 4 = 2.4$, so $\hat{y} = 10^{2.4} \approx 251$. (b) At $x = 8$: $\log_{10} \hat{y} = 1.4 + 0.25 \times 8 = 3.4$, so $\hat{y} = 10^{3.4} \approx 2512$.

Q17. The coefficient of determination measures how much variation a **straight line** explains, but it does not test whether a line is the right shape. A residual plot does: the clear $\cap$-shaped pattern here shows the residuals are positive in the middle and negative at the ends — systematic curvature the line has missed. Despite $r^2 = 0.72$, the data are non-linear and should be transformed.

Q18. A decreasing, concave-down curve falls slowly at first and then plunges, so from the chart it is straightened by **stretching** an axis. Stretching with x^2 spreads the large x-values apart, easing the plunge at the right-hand end while squeezing the flat start; y^2 does the same job from the other side, spreading apart the large y-values bunched along that flat start. So the candidates are x^2 (stretch the x-axis) or y^2 (stretch the y-axis).

Q19. Both models are valid, since a random residual plot confirms that each has removed the curvature. The one with $r^2 = 0.994$ explains a greater share of the variation and so is the **better linear fit**; report that model.

Q20. The transformation is on x (s^2), so predictions are read directly. (a) $\hat{d} = 1.2 + 0.0065 \times 40^2 = 1.2 + 0.0065 \times 1600 = 1.2 + 10.4 = 11.6$ m. (b) $\hat{d} = 1.2 + 0.0065 \times 80^2 = 1.2 + 0.0065 \times 6400 = 1.2 + 41.6 = 42.8$ m. (c) $42.8 \div 11.6 \approx 3.7$, i.e. roughly 4 **times** as far. Because the model depends on s^2, doubling the speed multiplies the s^2 term by $2^2 = 4$; the distance grows with the **square** of the speed, not in proportion to it, so it quadruples rather than doubles.