Correlation and regression

S2.2 · Statistics · revision notes

Interpreting scatter diagrams, correlation and regression lines for bivariate data, including why correlation does not imply causation. You also use logarithms to change $y = ax^n$ or $y = kb^x$ into a straight-line form and find the constants from a given regression line.

Key facts

- $-1 \leq r \leq 1$; r measures linear correlation only
- b = change in y for each increase of 1 in x
- $y = ax^n \Rightarrow \log y = \log a + n \log x$ (plot $\log y$ against $\log x$)
- Correlation does not imply causation
- Regression line of y on x: $y = a + bx$ (x explanatory, y response)
- Interpolation: reasonably reliable. Extrapolation: unreliable
- $y = kb^x \Rightarrow \log y = \log k + x \log b$ (plot $\log y$ against x)

Notes

Correlation

- Bivariate data are pairs of values (x, y), shown on a scatter diagram.
- Correlation describes how close the points are to a straight line: **positive** (y tends to increase as x increases), **negative** or **none**; strong or weak.

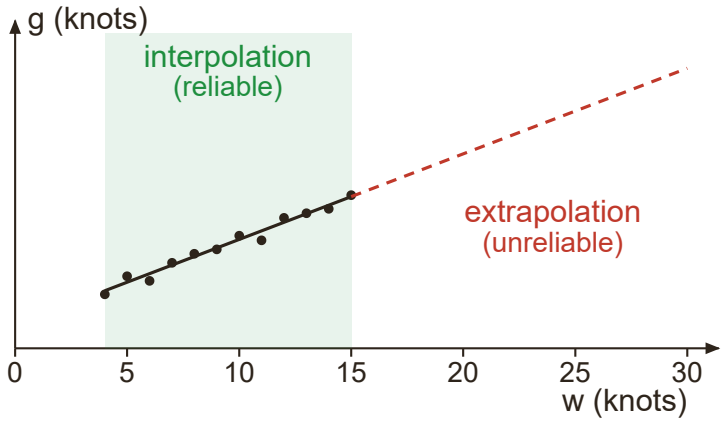


Positive, negative and no correlation

- The product moment correlation coefficient (PMCC), r, measures the strength of *linear* correlation: $-1 \leq r \leq 1$. $r = \pm 1$ means all the points lie on a straight line; r near 0 means little or no linear correlation (there may still be a non-linear relationship). Find r in your calculator's statistics mode.
- **Correlation does not imply causation.** Both variables may be affected by a third variable, or the link may be coincidence.
- If the points on a scatter diagram form distinct groups (e.g. two breeds of dog), one r or one regression line for all the points can mislead; analyse each group separately.

Regression lines

- The regression line of y on x is $y = a + bx$, where x is the **explanatory** (independent) variable and y is the **response** (dependent) variable. The equation is given to you; you are not asked to calculate it.
- b is the change in y for each increase of 1 in x. a is the value of y when $x = 0$, which only makes sense if $x = 0$ is realistic.
- **Interpolation** (estimating within the range of the data) is reasonably reliable when the correlation is strong. **Extrapolation** (outside the range) is unreliable, as the relationship may not continue.

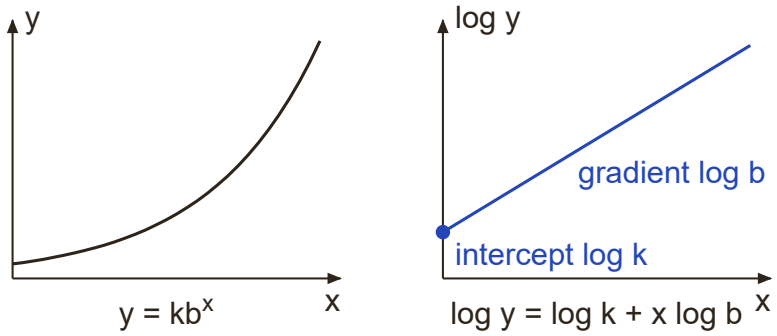


Inside the range of the data: interpolation. Outside it: extrapolation

- Use the regression line of y on x to estimate y from x, not x from y

Changing the variables with logarithms

- $y = ax^n$ gives $\log y = \log a + n \log x$: the graph of $\log y$ against $\log x$ is a straight line with gradient n and intercept $\log a$
- $y = kb^x$ gives $\log y = \log k + x \log b$: the graph of $\log y$ against x is a straight line with gradient $\log b$ and intercept $\log k$



Taking logs turns the curve $y = kb^x$ into a straight line

- Undo the logs at the end: if $\log_{10} k = c$, then $k = 10^c$
- The closer the PMCC of the transformed data is to 1 or -1, the better that model fits.

How to answer each type of question

M = method mark, A = answer mark, B = a mark on its own.

Interpret correlation and comment on a claim of causation

1 to 3 marks | **Grade D**

1. Describe the correlation: direction (positive or negative) and strength, in context.
2. For a causation claim, say that correlation does not imply causation.
3. Suggest a third variable that could affect both.

Example. Over 30 summer days, a beach kiosk records its ice cream sales and the number of people treated for sunburn at the beach first-aid post. The PMCC is 0.87.
(a) Interpret the value 0.87 in context.
(b) A reporter claims that eating ice cream causes sunburn. Comment on this claim.

Model answer.

(a) Strong positive correlation: days with higher ice cream sales tend to have more people treated for sunburn (B1)

(b) Correlation does not imply causation (B1); both are likely to increase on hot, sunny days, so the weather is a third variable affecting both (B1)

Interpret a regression line and comment on reliability

3 to 4 marks | **Grade C**

1. Interpret the gradient in context with units: ‘for each extra 1 …, … increases by …’.
2. Substitute into the equation to estimate.
3. Say whether the value is inside the range of the data (interpolation) or outside it (extrapolation).

Example. Using a sample of 12 days from one UK station, a student finds that the regression line of daily maximum gust, g knots, on daily mean windspeed, w knots, is $g = 5.2 + 1.9w$. The values of w in the sample were between 4 and 15.
(a) Interpret the value 1.9 in context.
(b) Estimate the daily maximum gust on a day with daily mean windspeed 10 knots.
(c) Comment on the reliability of using the line to estimate g when $w = 25$

Model answer.

(a) For each extra 1 knot of daily mean windspeed, the daily maximum gust increases by about 1.9 knots (B1)

(b) $g = 5.2 + 1.9 \times 10$ (M1) $= 24.2$ knots (A1)

(c) 25 knots is outside the range of the data (extrapolation), so the estimate is unreliable (B1)

Use logarithms to find the constants of a model 4 to 6 marks **Grade A**

1. Take logs of the model to get a linear equation.
2. Match its gradient and intercept with the given regression line.
3. Undo the logs (e.g. 10^c) and give the constants to the accuracy asked.
4. For estimates, use the log equation directly to avoid rounding errors.

Example. The number of bacteria, N, in a culture t hours after it is set up is modelled by $N = kb^t$. The regression line of $\log_{10} N$ on t is $\log_{10} N = 2.35 + 0.18t$.
(a) Find the values of k and b, giving your answers to 3 significant figures.
(b) Interpret the value of b in context.
(c) Use the model to estimate the number of bacteria after 5 hours.

Model answer.

(a) $\log_{10} N = \log_{10} k + t \log_{10} b$ (M1)

$k = 10^{2.35} = 224$ (A1), $b = 10^{0.18} = 1.51$ (A1)

(b) The number of bacteria is multiplied by about 1.51 each hour (it grows by about 51% per hour) (B1)

(c) $\log_{10} N = 2.35 + 0.18 \times 5$
$= 3.25$ (M1), so $N = 10^{3.25} \approx 1780$ (A1)

Don't lose marks

- Interpreting the gradient without context or units, or interpreting the intercept when $x = 0$ is not realistic.
- Saying $r = 0$ means 'no relationship': it means no *linear* correlation.
- Claiming that one variable causes the other just because r is close to 1.
- For $y = kb^x$, taking the gradient of the log graph as b instead of $\log b$
- Using the regression line of y on x to estimate x from a value of y

More tips

Memory tricks

- Gradient template: ‘For every increase of 1 [unit] in [x], [y] increases (or decreases) by [b] [units].’
- Power model $y = ax^n$: take logs of both variables. Exponential model $y = kb^x$: take logs of y only.
- If $\log_{10} y = c + mx$, then $y = 10^c \times (10^m)^x$
- Sign check: a negative gradient must go with negative correlation.

Exam technique

- Name both variables and their units in every interpretation.
- For reliability, say whether the value is within the range of the data, and mention the strength of the correlation if you know it.
- Keep the unrounded values of the constants in your calculator for later parts.

What each grade needs

What you need to be able to do, from the first marks up to the top grade.

Grade E — **Describe the correlation shown in a scatter diagram**
State positive, negative or no correlation, and whether it is strong or weak.

Grade E — **Identify the explanatory and response variables**
The explanatory (independent) variable is set or controlled, or affects the other; the response (dependent) variable responds to it.

Grade D — **Interpret the gradient of a regression line**
Say how much the response variable changes, with units, for each increase of 1 in the explanatory variable.

Grade C — **Judge reliability: interpolation or extrapolation**
Estimates within the range of the data are reasonably reliable; estimates outside it are not.

Grade C — **Interpret r, knowing correlation is not causation**
Relate r to how close the points are to a straight line, and recognise that a third variable may explain the link.

Grade B — **Recognise distinct groups in a scatter diagram**
Spot separate clusters (e.g. two species) and explain why one line or one r for all the points may mislead.

Grade A — **Linearise $y = ax^n$ or $y = kb^x$ using logarithms**
Take logs of both sides: $\log y$ is linear in $\log x$ for a power model, and in x for an exponential model.

Grade A* — **Find model constants from a logarithmic regression line**
Match the line to $\log y = \log a + n \log x$ or $\log y = \log k + x \log b$ and undo the logs.

Free notes, Full packs, practice papers and mock exams at chhetriacademy.com. Written for this site; not taken from past papers. Stuck? 1-to-1 tutoring with a free 30-minute trial lesson.