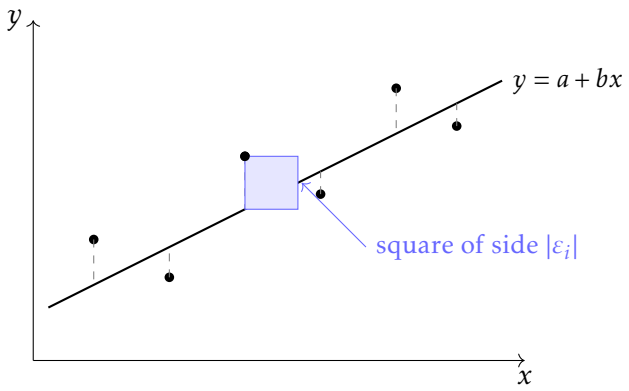
The Method of Least Squares

A scatter diagram with strong linear correlation begs for a line to be drawn through it. But *which* line? Everyone's eyeballed 'line of best fit' is slightly different; we want a principled, reproducible choice.

Definition. Suppose we fit the line $y = a + bx$ to data $(x_1, y_1), \ldots, (x_n, y_n)$. The **residual** (or error) of the ith point is the *vertical* distance from the point to the line:

$$\varepsilon_i = y_i - (a + bx_i).$$



Definition. The **regression line of y on x** is the line $y = a + bx$ that minimises the **sum of the squared residuals**

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - a - bx_i\big)^2.$$

This is the **method of least squares**: geometrically, we are minimising the total area of the squares in the diagram above.

Remark. Why *squared* residuals? Squaring makes all errors positive (so they cannot cancel), penalises large errors heavily, and — crucially — produces a minimisation problem we can solve exactly with algebra. Minimising $\sum |\varepsilon_i|$ is also possible ('least absolute deviations') but has no closed-form answer.

Theorem (The least squares regression line)

The regression line of y on x is $y = a + bx$, where

$$b = \frac{S_{xy}}{S_{xx}}, \qquad a = \bar{y} - b\bar{x},$$

with $S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$ and $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ as in the correlation chapter.

Fact — Since $a = \bar{y} - b\bar{x}$, the regression line **always passes through the mean point** $(\bar{x}, \bar{y})$.

Tip

The least squares line is often a little *flatter* than the line you would draw by eye. Your intuition tries to balance perpendicular distances; least squares only cares about vertical ones.

Calculating Regression Lines

Example

Six students record their revision time x (hours) and test score y:

x	1	2	3	4	5	6
y	52	55	60	61	68	70

(a) Calculate the equation of the regression line of y on x.

(b) Interpret the gradient and the intercept in context.

(a) $\sum x = 21$, $\sum x^2 = 91$, $\sum y = 366$, $\sum xy = 1346$, $n = 6$.

$$S_{xx} = 91 - \frac{21^2}{6} = 17.5, \qquad S_{xy} = 1346 - \frac{21 \times 366}{6} = 65$$

$$b = \frac{65}{17.5} = \frac{26}{7} = 3.71 \text{ (3 s.f.)}$$

$$a = \bar{y} - b\bar{x} = 61 - \frac{26}{7} \times 3.5 = 61 - 13 = 48$$

The regression line is $y = 48 + 3.71x$.

(b) The gradient: each additional hour of revision is associated with an increase of about 3.7 marks, on average. The intercept: a student who does no revision is predicted to score about 48 marks. (The intercept is only meaningful when $x = 0$ lies within or near the range of the data, as it does here.)

Example

A sample of $n = 12$ pairs has summary statistics

$$\sum x = 66, \quad \sum y = 144, \quad \sum x^2 = 406, \quad \sum xy = 825.$$

Find the equation of the regression line of y on x, and use it to estimate y when $x = 6$.

$$S_{xx} = 406 - \frac{66^2}{12} = 43, \qquad S_{xy} = 825 - \frac{66 \times 144}{12} = 33$$

$$b = \frac{33}{43} = 0.767 \text{ (3 s.f.)}, \qquad a = \frac{144}{12} - \frac{33}{43} \times \frac{66}{12} = 12 - 4.22 = 7.78 \text{ (3 s.f.)}$$

The regression line is $y = 7.78 + 0.767x$. *When* $x = 6$:

$$y = 7.78 + 0.767 \times 6 = 12.4 \text{ (3 s.f.)}$$

Since $\bar{x} = 5.5$, *the value* $x = 6$ *is comfortably within the data, so this is interpolation and the estimate is reliable (given reasonably strong correlation).*

Tip

Your calculator's regression mode gives a and b directly from raw data — use it as a check. From *summary statistics* you must use the formulae; both skills are examined.

Now that we have used the formulae, let us see where they come from. Write $Q(a,b) = \sum(y_i - a - bx_i)^2$ for the sum of squared residuals, and treat a and b as the variables to be chosen.

Differentiating with respect to each variable in turn (holding the other fixed) and setting both derivatives to zero:

$$\frac{\partial Q}{\partial a} = -2\sum(y_i - a - bx_i) = 0 \quad \Longrightarrow \quad \sum y = na + b\sum x \quad \Longrightarrow \quad \bar{y} = a + b\bar{x},$$
$$\frac{\partial Q}{\partial b} = -2\sum x_i(y_i - a - bx_i) = 0 \quad \Longrightarrow \quad \sum xy = a\sum x + b\sum x^2.$$

The first equation says the line passes through $(\bar{x}, \bar{y})$. Substituting $a = \bar{y} - b\bar{x}$ into the second and tidying up gives $b\,S_{xx} = S_{xy}$. Since Q is a sum of squares it has a minimum, not a maximum, at this stationary point.

Example

For the revision data above, calculate the residual for the student who revised for 5 hours, and interpret its sign.

The line predicts $y = 48 + \frac{26}{7} \times 5 = 48 + 18.57 = 66.57$. The observed score was 68, so

$$\varepsilon = 68 - 66.57 = 1.43 \text{ (3 s.f.)}$$

The residual is positive: this student scored about 1.4 marks higher than the regression line predicts — the point lies above the line.

Example (Exam style)

An agricultural researcher records the amount of fertiliser x (g/m^2) applied to ten equal plots and the resulting crop yield y (kg). The fertiliser amounts are chosen by the researcher. Summary statistics:

$$n = 10, \quad \sum x = 110, \quad \sum x^2 = 1540, \quad \sum y = 240, \quad \sum y^2 = 5984, \quad \sum xy = 2860.$$

(a) State, with a reason, which is the independent variable, and whether it is controlled.

(b) Calculate the product-moment correlation coefficient.

(c) Find the equation of the regression line of y on x.

(d) Estimate the yield of a plot treated with 12 g/m^2 of fertiliser, and comment on the reliability of your estimate.

(a) Fertiliser amount x is the independent variable — it is chosen by the researcher and is expected to explain the yield. Since its values are set in advance by the experimenter, it is a controlled variable.

(b)

$$S_{xx} = 1540 - \frac{110^2}{10} = 330, \qquad S_{yy} = 5984 - \frac{240^2}{10} = 224,$$

$$S_{xy} = 2860 - \frac{110 \times 240}{10} = 220$$

$$r = \frac{220}{\sqrt{330 \times 224}} = \frac{220}{\sqrt{73920}} = 0.809 \text{ (3 s.f.)}$$

(c) $b = \frac{220}{330} = \frac{2}{3}$, *and* $a = \bar{y} - b\bar{x} = 24 - \frac{2}{3} \times 11 = \frac{50}{3} = 16.7$ *(3 s.f.). The regression line is* $y = 16.7 + 0.667x$.

(d) $y = \frac{50}{3} + \frac{2}{3} \times 12 = \frac{74}{3} = 24.7$ *kg (3 s.f.). With* $\bar{x} = 11$, *the value* $x = 12$ *lies within the range of the data (interpolation), and the correlation is reasonably strong, so the estimate is fairly reliable — though with* $r = 0.81$ *there is still noticeable scatter about the line.*

Example (OCR Further Stats AS, June 2024 (parts))

The ages, x years, and the reaction times, t seconds, in an experiment carried out on a sample of 15 volunteers are summarised as follows.

$$n = 15, \quad \sum x = 762, \quad \sum t = 8.7, \quad \sum x^2 = 44204, \quad \sum t^2 = 5.65, \quad \sum xt = 490.1.$$

(a) Calculate the value of the product moment correlation coefficient between x and t.

(b) Calculate the equation of the line of regression of t on x, giving your answer in the form $t = a + bx$.

(c) Explain the relevance of the quantity $\sum(t - a - bx)^2$ to your answer to part (b).

(d) Estimate the reaction time, in seconds, for a volunteer aged 42.

(a)

$$S_{xx} = 44204 - \frac{762^2}{15} = 5494.4, \qquad S_{tt} = 5.65 - \frac{8.7^2}{15} = 0.604,$$

$$S_{xt} = 490.1 - \frac{762 \times 8.7}{15} = 48.14$$

$$r = \frac{48.14}{\sqrt{5494.4 \times 0.604}} = 0.836 \text{ (3 s.f.)}$$

(b) $b = \frac{48.14}{5494.4} = 0.00876$ *(3 s.f.), and* $a = \bar{t} - b\bar{x} = 0.58 - 0.0087616 \times 50.8 = 0.135$ *(3 s.f.). The regression line is* $t = 0.135 + 0.00876x$.

(c) $\sum(t - a - bx)^2$ *is the sum of the squared residuals: the values of a and b found in part (b) are precisely those which* minimise *this quantity.*

(d) $t = 0.13491 + 0.0087616 \times 42 = 0.503$ *seconds (3 s.f.).*

Example (In class)

The lengths ℓ (cm) and masses m (g) of nine fish of the same species give

$$\sum \ell = 180, \quad \sum \ell^2 = 3654, \quad \sum m = 450, \quad \sum \ell m = 9117.$$

Find the equation of the regression line of m on ℓ, and estimate the mass of a fish of length 22 cm.

Textbook Exercises: [CUP.S] Ch 5 §3; [S1] Ch 10

Which Variable on Which?

The phrase "regression of y on x" means: x is treated as the independent (explanatory or controlled) variable, y as the dependent (response) variable, and we minimise *vertical* errors — errors in predicting y.

Fact —

- If one variable is independent (or controlled), it must play the role of x. We regress the response on the explanatory variable.
- The regression line of x on y reverses the roles: it minimises *horizontal* errors, and is a genuinely **different line** (it is *not* obtained by rearranging the y-on-x equation).
- In a given situation, neither variable may be independent (e.g. marks in two exams) — then either regression could be meaningful, depending on which variable you want to predict.

Remark (Which way round?)**.** You will only ever be asked to calculate the regression line of y on x, with x the independent variable. But understanding the distinction is valuable.

Example

An economist models wages w as a function of years of education E: $w = a + bE$. Why would the model $E = c + dw$ be inappropriate?

The proposed causal direction is from education to wages: education is the explanatory variable and wages the response. Regressing E on w would model years of education as being determined by later wages, which makes no causal sense — your salary at 40 does not change how long you spent at school. If E is treated as controlled, only the regression of w on E is meaningful; to estimate E from a known w we would rearrange the w-on-E line rather than fit a new one.

By contrast, for Latin and Maths marks either regression is sensible: choose y on x to predict y, and x on y to predict x.

Remark (The two regression lines)**.** The y-on-x line has gradient $b_{yx} = \dfrac{S_{xy}}{S_{xx}}$; the x-on-y line (written $x = a' + b_{xy}\,y$) has gradient $b_{xy} = \dfrac{S_{xy}}{S_{yy}}$. Multiplying,

$$b_{yx}\,b_{xy} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2.$$

Both lines pass through $(\bar{x}, \bar{y})$, and they coincide if and only if $r = \pm 1$. For $|r| < 1$ they form a 'scissors' opening about the mean point, and the intuitive line of best fit lies between them. Note also that

$$b_{yx} = r\,\frac{s_y}{s_x},$$

where s_x, s_y are the standard deviations: the gradient is the correlation coefficient rescaled by the spreads. For standardised data the gradient of the regression line is simply r.

Example (OCR S1, June 2015)

The table shows the load a lorry was carrying, x tonnes, and the fuel economy, y km per litre, for 8 different journeys. You should assume that neither variable is controlled.

Load (x tonnes)	5.1	5.8	6.5	7.1	7.6	8.4	9.5	10.5
Fuel economy (y km per litre)	6.2	6.1	5.9	5.6	5.3	5.4	5.3	5.1

$$n = 8, \quad \sum x = 60.5, \quad \sum y = 44.9, \quad \sum x^2 = 481.13, \quad \sum y^2 = 253.17, \quad \sum xy = 334.65.$$

(i) Calculate the equation of the regression line of y on x.

(ii) Estimate the fuel economy for a load of 9.2 tonnes.

(iii) An analyst calculated the equation of the regression line of x on y. Without calculating this equation, state the coordinates of the point where the two regression lines intersect.

(iv) Describe briefly the method required to estimate the load when the fuel economy is 5.8 km per litre.

(i) $S_{xx} = 481.13 - \dfrac{60.5^2}{8} = 23.59875$ *and* $S_{xy} = 334.65 - \dfrac{60.5 \times 44.9}{8} = -4.90625$, *so*

$$b = \frac{-4.90625}{23.59875} = -0.2079 \text{ (4 s.f.)}, \qquad a = \frac{44.9}{8} - (-0.2079) \times \frac{60.5}{8} = 7.18 \text{ (3 s.f.)}$$

The regression line is $y = 7.18 - 0.208x$.

(ii) $y = 7.1848 - 0.2079 \times 9.2 = 5.27$ *km per litre (3 s.f.).*

(iii) Both regression lines pass through the mean point $(\bar{x}, \bar{y}) = (7.56, 5.61)$ *(3 s.f.).*

(iv) Since neither variable is controlled, to predict x from a known y we should use the regression line of x on y: calculate it and substitute $y = 5.8$.

Using the Regression Line

Definition. Using the regression line to estimate y for an x-value *within* the range of the data is **interpolation**; for an x-value *outside* the range of the data it is **extrapolation**.

Fact (Reliability of estimates) — An estimate from a regression line is reliable when:

- it is interpolation, not extrapolation — outside the data range there is no evidence the linear pattern continues;
- the correlation is strong ($|r|$ close to 1);
- we are predicting the response from the explanatory variable (using y on x to predict y).

Beware also of a small section of *curved* data looking convincingly linear: a line fitted to ages 5–10 of a child's height data would predict 3-metre-tall adults.

Example

The regression line $y = 48 + 3.71x$ relates test score to hours of revision, based on data with $1 \le x \le 6$. Comment on the reliability of using it to predict the score of a student who revises for (a) 4.5 hours, (b) 40 hours.

(a) $x = 4.5$ is within the data range: interpolation. With the very strong correlation found earlier ($r = 0.987$), the estimate $y = 48 + 3.71 \times 4.5 \approx 65$ is reliable.

(b) $x = 40$ is far outside the data range: extrapolation. The prediction $y = 48 + 3.71 \times 40 \approx 196$ is meaningless — not least because the test is presumably out of 100. There is no evidence the linear relationship persists beyond 6 hours.

Linearising data with logarithms

Many genuine relationships are not linear, but can be *made* linear by transforming the variables — then all our regression machinery applies.

Fact —

- Power law $y = kx^n$: taking logs, $\log y = \log k + n \log x$, so plotting $\log y$ against $\log x$ gives a straight line with gradient n and intercept $\log k$.
- Exponential growth $y = k\,c^x$: taking logs, $\log y = \log k + x \log c$, so plotting $\log y$ against x gives a straight line with gradient $\log c$ and intercept $\log k$.

Example

For a set of data, the regression line of $Y = \log_{10} y$ on $X = \log_{10} x$ is found to be $Y = 0.301 + 2X$. Express y in terms of x.

$\log_{10} y = 0.301 + 2\log_{10} x = \log_{10} 10^{0.301} + \log_{10} x^2 = \log_{10}\left(10^{0.301} x^2\right)$.
Since $10^{0.301} \approx 2.00$, we obtain the power model $y = 2x^2$.

Remark (Anscombe's quartet)**.** In 1973 Francis Anscombe constructed four data sets of 11 points which share (to 2 d.p.) the same $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy}, correlation coefficient $r \approx 0.816$ and regression line $y = 3 + 0.5x$ — yet their scatter diagrams are wildly different: one is genuinely linear, one is a perfect parabola, one is a perfect line with a single outlier, and one is a vertical stack of points with one influential point. Moral: summary statistics, correlation and regression can be **very** misleading. *Always look at the scatter diagram first.*

Textbook Exercises: [CUP.S] Ch 5 §3; [S1] Ch 10

Effect of Coding on the Regression Line

Fact — Suppose the data are coded by $u = \dfrac{x-h}{p}$ and $v = \dfrac{y-k}{q}$ (with $p, q > 0$), and the regression line of v on u is found to be

$$v = c + d\,u.$$

Then the regression line of y on x is

$$y = \underbrace{k + qc - \frac{qd\,h}{p}}_{\text{intercept}} + \underbrace{\frac{qd}{p}}_{\text{gradient}}\,x.$$

In particular the gradients are related by a factor of q/p — exactly as units demand, since the gradient has units of y per unit of x.

The derivation is nothing more than substituting the codings and rearranging.

$$\frac{y-k}{q} = c + d\,\frac{x-h}{p} \quad\Longrightarrow\quad y = k + qc + \frac{qd}{p}(x-h) = k + qc - \frac{qd\,h}{p} + \frac{qd}{p}\,x.$$

Example

A shop's annual sales are recorded each year. With $u = t - 2000$ (years after 2000) and $v = \dfrac{y}{1000}$ (sales in thousands of pounds), the regression line of v on u is

$$v = 3.2 + 0.15u.$$

Find the regression line of y (sales in pounds) on t (calendar year), and interpret the gradient.

Substituting:

$$\frac{y}{1000} = 3.2 + 0.15(t - 2000)$$

$$y = 3200 + 150(t - 2000) = 150t - 296\,800.$$

The gradient is 150*: sales increased by about £150 per year on average over the period of the data.*

Tip

Do not memorise the boxed formula — just substitute the codings into the coded regression line and rearrange. The arithmetic is friendlier than the algebra.

The following question draws together the whole chapter: the least squares idea, the language of regression, reliability of estimates, and coding.

Example (OCR Further Stats, June 2024)

The coordinates of a set of 10 points are denoted by (x_i, y_i) for $i = 1, 2, \ldots, 10$. For a particular set of values of (x_i, y_i) and any constants a and b it can be shown that

$$\sum (y_i - a - bx_i)^2 = 10(11 - a - 6b)^2 + 126\left(b - \tfrac{83}{42}\right)^2 + \tfrac{139}{14}.$$

(a) (i) Explain why $\sum (y_i - a - bx_i)^2$ is minimised by taking $b = \frac{83}{42}$ and $a = 11 - 6b$.

(ii) Hence explain why the equation of the regression line of y on x for these points is given by the corresponding values of a and b (so that the equation is $y = \frac{83}{42}x - \frac{6}{7}$).

(b) State which of the following terms cannot apply to the variable X if the regression line of y on x can be used for estimating values of Y:

Dependent Independent Controlled Response

(c) Use the regression line to estimate the value of y corresponding to $x = 8$.

(d) State what must be true of the value $x = 8$ if the estimate in part (c) is to be reliable.

(e) Variables u and v are related to x and y by $u = 2 + 4x$ and $v = 8 - 2y$. Show that the gradient of the regression line of v on u is very close to -1.

(a) (i) The last term is constant and both squared brackets are non-negative, so the expression is smallest when both brackets are zero: $b = \frac{83}{42}$ and $11 - a - 6b = 0$, i.e. $a = 11 - 6b = -\frac{6}{7}$.

(ii) $\sum (y_i - a - bx_i)^2$ is the sum of the squared residuals, and the regression line of y on x is by definition the line minimising it — which is exactly what these values of a and b do. Hence the regression line is $y = \frac{83}{42}x - \frac{6}{7}$.

(b) Dependent and Response cannot apply: to use the y-on-x line for predicting Y, the variable X must be the independent (possibly controlled) variable.

(c) $y = \frac{83}{42} \times 8 - \frac{6}{7} = \frac{314}{21} = 15.0$ (3 s.f.).

(d) $x = 8$ must lie within the range of the x-data — the estimate must be interpolation, not extrapolation.

(e) Substituting $v = 8 - 2y$ and $u = 2 + 4x$ into $v = c + d\,u$, the gradients are related by $d = b \times \frac{-2}{4}$, so

$$d = \frac{83}{42} \times \left(-\frac{1}{2}\right) = -\frac{83}{84} = -0.988 \text{ (3 s.f.)},$$

which is very close to -1.

The Population Regression Line

Remark (What regression is really doing)**.** This section explains what regression is really doing — and connects to the estimation theory later in the course.

Just as r estimates a population parameter ρ, our regression line estimates a **population regression line**. The standard model is

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where α, β are fixed unknown population parameters and the errors ε_i are random. The usual assumptions are that the ε_i are independent, with $\mathbb{E}[\varepsilon_i] = 0$ and common variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$ (often also $\varepsilon_i \sim N(0, \sigma^2)$).

The least squares estimates a and b computed from a sample are then *random variables* — different samples give different lines. One can show they are unbiased, with

$$\mathbb{E}[b] = \beta, \qquad \mathrm{Var}[b] = \frac{\sigma^2}{S_{xx}},$$

and under the normality assumption $b \sim N\left(\beta, \dfrac{\sigma^2}{S_{xx}}\right)$. This is what underlies confidence intervals and hypothesis tests for the slope (e.g. testing $\beta = 0$), which you may meet at university.

Be careful to keep two different sets of assumptions apart: the *bivariate normal* assumption used in pmcc hypothesis tests (a statement about the joint distribution of X and Y), and the *normal errors* assumption here (a statement about Y at each fixed x). They are related but not the same.

Textbook Exercises: [Toller] Ch 6