AI Research Labs

Section 3 of 6 · 9 lessons

Linear Regression

How regression models are built, how their coefficients are calculated and interpreted, and how reliable their predictions are.

3.1 Regression Fundamentals

3.2 Simple and Multiple Linear Regression

3.3 Ordinary Least Squares (OLS)

3.4 Calculating Regression Coefficients by Hand

3.5 Interpreting Coefficients

3.6 Residuals and Regression Assumptions

3.7 R^2, Adjusted R^2, RMSE and Other Metrics

3.8 Standard Errors, Confidence Intervals and p-values

3.9 Multicollinearity and VIF



Free course with quizzes, progress tracking and Scoopy, an AI tutor:
airesearchlabs.si/learn/ml

Machine Learning Foundations · Edition November 2026 · © 2026 AI Research Labs. Free for personal study.

Where this section fits

Every lesson is a step in building a model. These are the steps this section teaches; the Big picture PDF and the online map show how they connect to the rest.

R1

SUPERVISED LEARNING · REGRESSION LANE · STEP R1

Choose the model

What shape of relationship am I assuming?

Lessons here: 3.1, 3.2

R2

SUPERVISED LEARNING · REGRESSION LANE · STEP R2

Fit it

Which coefficients fit the data best?

Lessons here: 3.3, 3.4 · feeds into Control complexity

R3

SUPERVISED LEARNING · REGRESSION LANE · STEP R3

Read and check it

What do the coefficients mean, and can I trust them?

Lessons here: 3.5, 3.6, 3.8, 3.9 · feeds into Control complexity, Choose the model

R4

SUPERVISED LEARNING · REGRESSION LANE · STEP R4

Score it

How far off are the predictions, in units people understand?

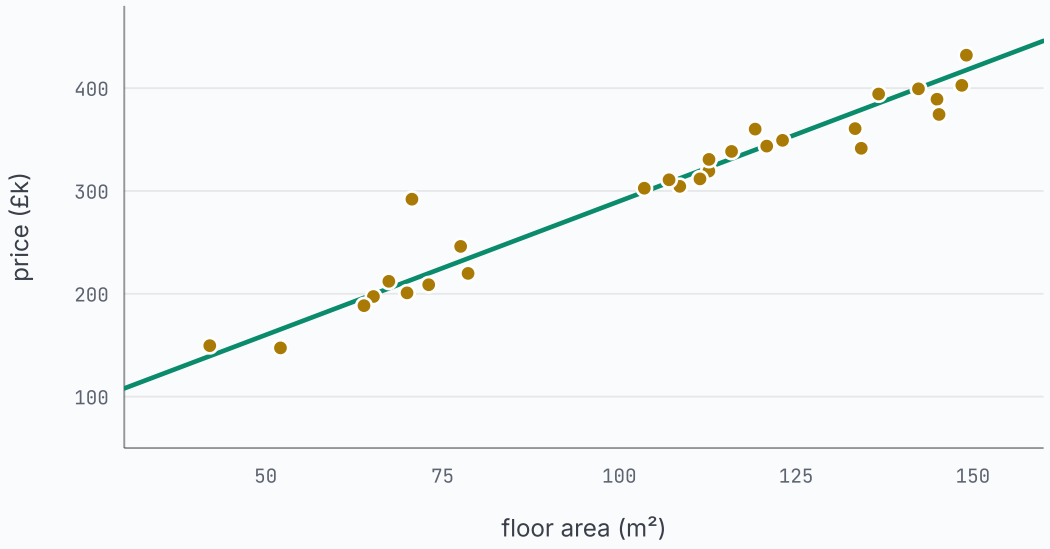
Lessons here: 3.7 · feeds into Evaluate honestly

Regression Fundamentals

Regression predicts a number from inputs. The model, the noise term, what "fitting" means, and how predictions are made, using house prices as a running example.

The question regression answers

Regression predicts a quantitative response Y (a number) from one or more predictors $X_1, \ldots, X_p$. How much will this house sell for? What will CO concentration be at noon? How many patients will arrive next week?



Each dot is a house. The line summarises the trend: on average, more floor area means a higher price. The scatter around it is the noise the model can't explain.

Signal plus noise

Regression models assume the response is a systematic part plus random noise:

$$Y = f(X) + \varepsilon$$

$f(X)$ is the **signal**: the part of Y that depends on the predictors. ε is **noise**: everything else (unmeasured factors, measurement error, pure chance), assumed to average out to zero. Two houses with identical area and bedrooms still sell for different prices; that difference is ε.

Linear regression assumes f is linear: $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Fitting and predicting

- **Fitting** means estimating the unknown coefficients from training data, giving $\hat{\beta}_0, \hat{\beta}_1, \ldots$ (lesson 3.3).
- **Fitted values** $\hat{y}_i$ are the model's predictions for the training points.

- **Predicting** a new case means plugging its predictor values x^* into the fitted equation: $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*$.
- **Residuals** $r_i = y_i - \hat{y}_i$ are the observed minus fitted values: how far each training point sits from the line.

Why start with linear regression?

It is the simplest useful ML model: fast, interpretable, with a closed-form solution and a full toolkit for uncertainty (standard errors, p-values, confidence intervals). Many ideas in the rest of the course (loss, overfitting, regularisation) are clearest here first.

Cheatsheet

Signal + noise

$Y = f(X) + \varepsilon$, $\mathbb{E}[\varepsilon] = 0$. Linear regression: $f(X) = \beta_0 + \sum_j \beta_j X_j$.

Residual

$r_i = y_i - \hat{y}_i$ (observed – fitted). Positive → model under-predicted.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/regression-fundamentals

Simple and Multiple Linear Regression

One predictor or many, the model is a weighted sum of inputs. Why a model with X² or interaction terms is still "linear" regression: it only has to be linear in the coefficients.

Simple linear regression

One predictor:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 is the **intercept** (the predicted Y when $X = 0$) and β_1 the **slope** (the change in predicted Y per one-unit increase in X). The fitted model is a straight line.

Multiple linear regression

Several predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

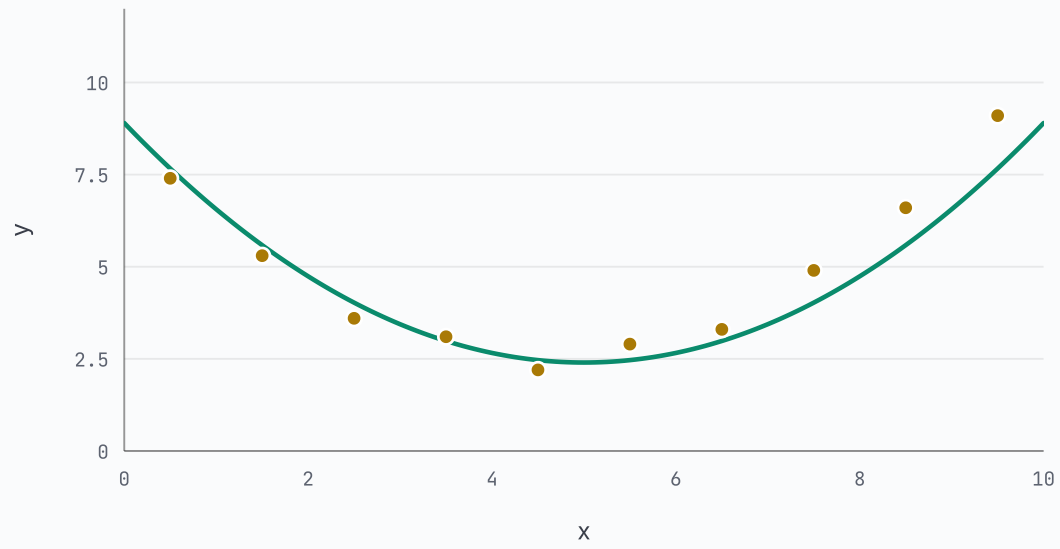
With two predictors the fitted model is a **plane** in 3-D (X_1, X_2, Y) space; with more it is a **hyperplane**. The algebra is identical, it just can't be drawn. It's a **parametric** model: a finite list of $p + 1$ coefficients.

"Linear" means linear in the coefficients

This surprises almost everyone. The model below is still **linear regression**:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

The curve is a parabola, but the *unknowns* $\beta_0, \beta_1, \beta_2$ each appear on their own, multiplied by a known number and added up. Create a new column $X_2 = X^2$ and it becomes an ordinary multiple regression with two predictors. The same fitting method (OLS) works unchanged.



ŷ = 8.9 − 2.6x + 0.26x² (linear in β)

A curved fit from linear regression: the predictors are x and x², and the model is a weighted sum of them.

Interactions are linear too

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$X_1 X_2$ is just another column computed from the data. The interaction lets the effect of X_1 depend on X_2: the slope for X_1 is $\beta_1 + \beta_3 X_2$.

MODEL	LINEAR REGRESSION?	WHY
$Y = \beta_0 + \beta_1 X + \beta_2 X^2$	Yes	Each β multiplies a known column
$Y = \beta_0 + \beta_1 \log X$	Yes	$\log X$ is a known column
$Y = \beta_0 + \beta_1 X_1 X_2$	Yes	$X_1 X_2$ is a known column
$Y = \beta_0 e^{\beta_1 X}$	No	β_1 sits inside an exponential
$Y = \beta_0 + X^{\beta_1}$	No	β_1 is an exponent

TEST

Treat the data as fixed numbers. If the model is a sum of (coefficient × something computed from the data), it is linear regression, however curved the picture.

Common doubts

FIRST PASS **Why is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ still called linear regression when X is squared?**

Because "linear" refers to the **coefficients**, not the inputs. X^2 is computed from the data, so it's a known number for every row. The model is $\beta_0 \cdot 1 + \beta_1 \cdot (X) + \beta_2 \cdot (X^2)$: each unknown multiplies a known value, and the terms are added. Treat X^2 as a second predictor column and it's an ordinary multiple regression.

FIRST PASS **Why is a model with an interaction term $X_1 X_2$ still linear regression?**

Same reason: $X_1 X_2$ is a column you can compute from the data before fitting. The model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ is still a weighted sum with unknown weights $\beta_0, \ldots, \beta_3$, so OLS fits it exactly as before.

FIRST PASS **What does "linear in the coefficients" mean?**

The unknown parameters appear only to the first power, each multiplied by something known, and the pieces are added. The inputs themselves can be transformed in any way (squared, logged, multiplied together). A model like $\beta_0 e^{\beta_1 X}$ is **not** linear in its coefficients, because β_1 is inside the exponential.

Cheatsheet

Multiple linear regression

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$; prediction $\hat{y}^* = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_j^*$.

"Linear" regression

Linear in the **coefficients**, not the inputs. X^2, $\log X$, $X_1 X_2$ are allowed.

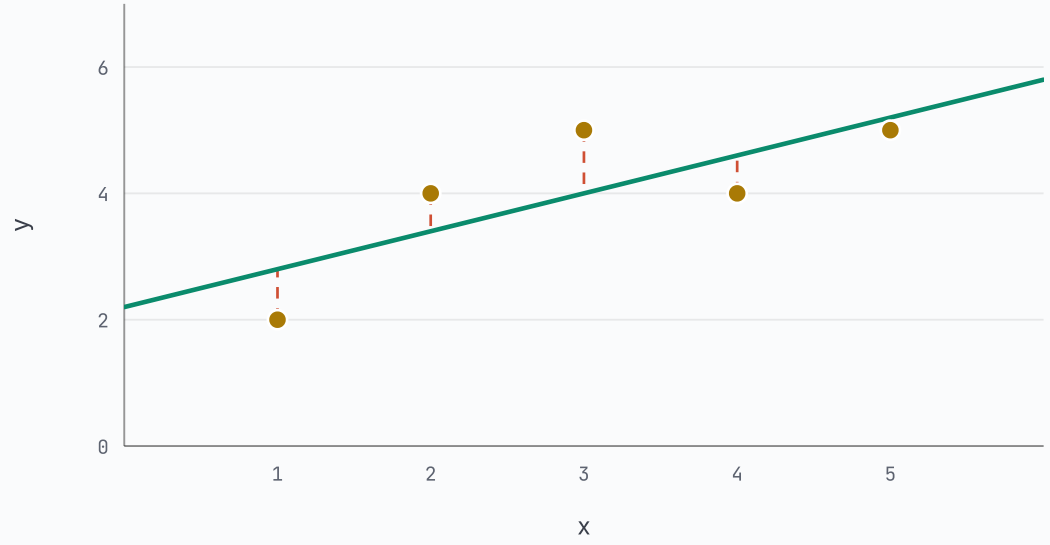
Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/simple-and-multiple-linear-regression

Ordinary Least Squares (OLS)

OLS picks the line that minimises the residual sum of squares. Residuals and RSS, the design matrix and its dimensions, the normal equations, the closed-form solution, and when it breaks.

The problem we're solving

Many lines could pass through a cloud of points. Which is "best"? OLS's answer: the one with the smallest total **squared vertical distance** to the points.



Dataset B: five points. The dashed lines are residuals. OLS makes the sum of their squares as small as possible.

Residuals and RSS

For observation i, the residual is observed minus fitted: $r_i = y_i - \hat{y}_i$. OLS chooses coefficients to minimise the **residual sum of squares**:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2 = \sum_i r_i^2$$

This is exactly ERM with squared loss (lesson 2.2).

WHY SQUARES?

Raw residuals would cancel out (positives and negatives). Absolute values are awkward to minimise. Squares are smooth, so calculus gives a closed-form answer, and they punish big misses more.

The design matrix

Stack the data into matrices. With n observations and p predictors, the **design matrix** X has n rows and $p+1$ columns: a first column of ones (for the intercept), then one column per predictor.

$$X = \underset{n\times(p+1)}{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}, \quad \beta = \underset{(p+1)\times 1}{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}, \quad y = \underset{n\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}$$

Then all the fitted values at once are $\hat{y} = X\beta$, and $\mathrm{RSS}(\beta) = (y - X\beta)^\top (y - X\beta)$.

EXAMPLE: AREA, BEDROOMS, PRICE

HOUSE	AREA (M²)	BEDROOMS	PRICE (£K)
1	50	1	150
2	70	2	200
3	80	2	230
4	100	3	280

Price is the response, so it goes in y, not X. Two predictors means $p = 2$:

$$X = \underset{4\times 3}{\begin{pmatrix} 1 & 50 & 1 \\ 1 & 70 & 2 \\ 1 & 80 & 2 \\ 1 & 100 & 3 \end{pmatrix}}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad y = \begin{pmatrix} 150 \\ 200 \\ 230 \\ 280 \end{pmatrix}$$

Three coefficients: intercept, area, bedrooms. One coefficient per column of X.

The normal equations

Setting the derivative of RSS to zero gives the **normal equations**, and if $X^\top X$ is invertible, the closed-form solution:

OLS SOLUTION

$$X^\top X\,\hat{\beta} = X^\top y \qquad \Longrightarrow \qquad \hat{\beta} = (X^\top X)^{-1} X^\top y$$

Shapes: $X^\top X$ is $(p+1)\times(p+1)$; $X^\top y$ is $(p+1)\times 1$; so $\hat{\beta}$ is $(p+1)\times 1$.

▶ Show the derivation (optional)

WORKED EXAMPLE VIA MATRICES

Data $(x, y) = (1, 3), (2, 5), (3, 7)$. Design matrix rows $(1, x_i)$.

$$X^\top X = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \quad X^\top y = \begin{pmatrix} 15 \\ 34 \end{pmatrix}$$

For a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the inverse is $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. Here $\det = 42 - 36 = 6$:

$$\hat{\beta} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} 15 \\ 34 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 210 - 204 \\ -90 + 102 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

So $\hat{y} = 1 + 2x$, and every residual is zero (the points lie exactly on the line).

When it breaks

$(X^\top X)^{-1}$ doesn't exist when $X^\top X$ is **singular**:

- **Perfect multicollinearity:** one predictor is an exact linear combination of others (e.g. height in cm and height in inches).
- **More coefficients than observations:** $p + 1 > n$.

If $X^\top X$ is *nearly* singular, the solution exists but is unstable (lesson 3.9). Ridge regression (5.4) fixes both cases.

Common doubts

FIRST PASS **The data have Area, Bedrooms and Price. Why does the coefficient vector have three entries, not four?**

Count columns of the **design matrix**, not columns of the spreadsheet. Price is the **response**, so it goes into y, not into X. The predictors are Area and Bedrooms ($p = 2$), plus the column of ones for the intercept: $p + 1 = 3$ columns, so $\beta = (\beta_0, \beta_1, \beta_2)$. There's one coefficient per column of X, and no β_3.

FIRST PASS **I find the OLS matrix derivation slow. What do I actually need to understand?**

Keep these, which carry most of the meaning:

- $Y = X\beta + \varepsilon$ and what each piece is (with its dimensions).
- Residual vector $y - X\beta$, and RSS = sum of squared residuals.
- OLS chooses β to minimise RSS.
- The result: normal equations $X^\top X\hat{\beta} = X^\top y$, solution $\hat{\beta} = (X^\top X)^{-1}X^\top y$, and that it fails when $X^\top X$ is singular.

The step-by-step calculus (expanding the quadratic and differentiating) can be deferred until you need it. It is collapsed above under "Show the derivation".

Cheatsheet

RSS

$\mathrm{RSS} = \sum_i (y_i - \hat{y}_i)^2 = (y - X\beta)^\top (y - X\beta)$.

OLS normal equations

$X^\top X \hat{\beta} = X^\top y \Rightarrow \hat{\beta} = (X^\top X)^{-1} X^\top y$ (if invertible).

Design matrix shape

X: $n \times (p+1)$, first column ones. β: $(p+1) \times 1$.
$X^\top X$: $(p+1) \times (p+1)$.

2×2 inverse

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/ordinary-least-squares

Calculating Regression Coefficients by Hand

The slope and intercept formulas for simple regression, a full worked example with a table of deviations, and why a predictor with little variation gives an unreliable slope.

The formulas

SIMPLE REGRESSION $Y = B_0 + B_1 X + E$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The **numerator** S_{xy} measures how x and y move together: positive when above-average x goes with above-average y.
- The **denominator** S_{xx} measures how spread out x is.
- The fitted line always passes through the point of means $(\bar{x}, \bar{y})$.

Worked example: dataset B

$x = 1, 2, 3, 4, 5$ and $y = 2, 4, 5, 4, 5$. So $n = 5$, $\bar{x} = 3$, $\bar{y} = 4$.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
1	2	-2	-2	4	4
2	4	-1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
Sum				$S_{xy} = 6$	$S_{xx} = 10$

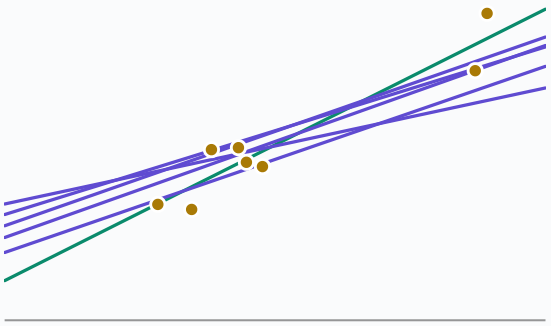
$\hat{\beta}_1 = 6/10 = 0.6$ and $\hat{\beta}_0 = 4 - 0.6 \times 3 = 2.2$. The fitted line is $\hat{y} = 2.2 + 0.6x$.

Fitted values: $(2.8, 3.4, 4.0, 4.6, 5.2)$. Residuals: $(-0.8, 0.6, 1.0, -0.6, -0.2)$. They sum to zero, which always happens when the model has an intercept.

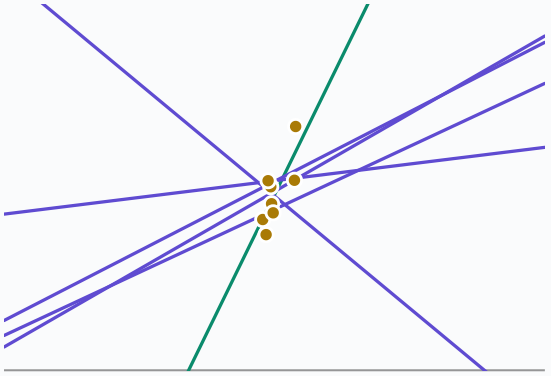
RSS $= 0.64 + 0.36 + 1.00 + 0.36 + 0.04 = 2.4$. We'll reuse this in 3.7 and 3.8.

Why low variance in X is a problem

Look at the denominator S_{xx}. If the x values barely vary, S_{xx} is tiny, and small changes in a few y values swing the slope a lot. If x is constant, $S_{xx} = 0$ and the slope can't be computed at all.



Spread-out x: six resampled datasets give six similar slopes.



Clustered x (4.5–5.5): same truth, same noise, wildly different slopes.

DESIGN LESSON

To estimate how Y changes with X, you need to observe X over a decent range. This is limited information about the slope, and it shows up later as a large standard error (3.8).

Common doubts

FIRST PASS **Do I need to do the full manual calculation (means, deviations, products, squares)?**

For a closed-book exam with "calculate by hand" questions, yes: practise the table above until you can do it in a few minutes, because it earns method marks even if the arithmetic slips. Conceptually, what matters is: the slope is "how x and y co-vary" divided by "how much x varies", and the line passes through $(\bar{x}, \bar{y})$. If you're short on time, understand that first and return to the drill before the exam.

FIRST PASS **Why does low variance in the predictor make the slope unreliable?**

The slope's denominator is $S_{xx} = \sum(x_i - \bar{x})^2$. When the x values are clustered, S_{xx} is small, so tiny changes in the y values (noise) produce big changes in $\hat{\beta}_1$. Intuitively, if you only see houses between 99 and 101 m², you have almost no information about how price changes with size. In 3.8 this appears as $\mathrm{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 / S_{xx}}$ getting large.

Cheatsheet

Simple regression by hand

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

Slope by hand

1) Means $\bar{x}, \bar{y}$. 2) Deviations. 3) Products & squares. 4) Sum → S_{xy}, S_{xx}. 5) Divide; then intercept. 6) Check residuals sum to 0.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/calculating-coefficients-by-hand

Interpreting Coefficients

What a regression coefficient means, the "holding other predictors fixed" clause, units and standardisation, why coefficients can change sign, and why association is not causation.

THE INTERPRETATION SENTENCE

$\hat{\beta}_j$ is the predicted change in Y for a one-unit increase in X_j, **holding all other included predictors fixed**.

WORKED EXAMPLE

Fitted model: $\hat{y} = 12 + 3x_1 - 2x_2$.

- Holding x_2 fixed, +1 in x_1 raises the prediction by 3.
- Holding x_1 fixed, +1 in x_2 lowers it by 2.
- The intercept 12 is the prediction when $x_1 = x_2 = 0$.
- For $(x_1, x_2) = (4, 5)$: $\hat{y} = 12 + 12 - 10 = 14$.

Four classic misreadings

1. **A coefficient is not a simple correlation.** It's the effect *given the other predictors*, so it can change size or even sign when you add or remove predictors.
2. **Size depends on units.** Measuring a predictor in grams instead of kilograms multiplies its coefficient by 1/1000. To compare importance fairly, **standardise** the predictors first; then each coefficient is the change in Y per one-SD change in X_j.
3. **Association is not causation.** In observational data, a positive coefficient doesn't mean changing X_j will change Y. A confounder may drive both.
4. **The intercept may be meaningless.** It's the prediction at $X = 0$, which may be impossible (a house with zero floor area).

The negative bedrooms puzzle

A house-price model with floor area, bedrooms and distance to centre gives bedrooms a **negative** coefficient. Does adding a bedroom make a house cheaper? Not necessarily. *Holding floor area fixed*, more bedrooms means each room is smaller, which buyers may value less. Correlation between area and bedrooms also makes individual coefficients unstable (3.9). The coefficient answers a narrower question than it first appears to.

HOW IT'S ASKED

"Interpret -4 in $\hat{y} = 5 + 2x_1 - 4x_2$." Answer: holding x_1 fixed, each one-unit increase in x_2 is associated with a 4-unit decrease in predicted y. Mention "holding fixed" and "associated with", not "causes".

Cheatsheet

Coefficient sentence

"Holding the other predictors fixed, a one-unit increase in X_j is associated with a $\hat{\beta}_j$ change in predicted Y."

Comparing coefficients

Raw sizes depend on units. Standardise predictors first. And association ≠ causation.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/interpreting-coefficients

Residuals and Regression Assumptions

The assumptions behind OLS inference, and how residual plots reveal curvature, non-constant variance, autocorrelation and outliers. How to describe a residual plot in an exam.

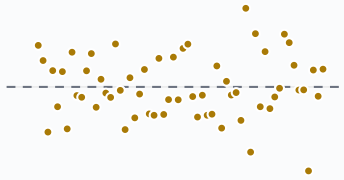
The assumptions

OLS *fitting* needs no assumptions: it just minimises RSS. But standard errors, p-values and confidence intervals (3.8) rely on assumptions about the noise ε:

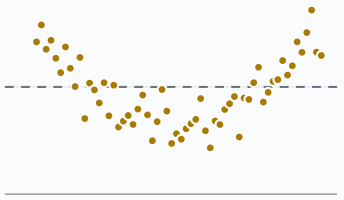
- **Mean zero** at every X (the linear form is right, nothing systematic is missing).
- **Constant variance** (homoskedasticity).
- **Independence** between observations.
- **Roughly normal**, for exact t-tests in small samples.

We can't see ε, but residuals $r_i = y_i - \hat{y}_i$ are our estimates of it. The main tool is a plot of **residuals against fitted values**.

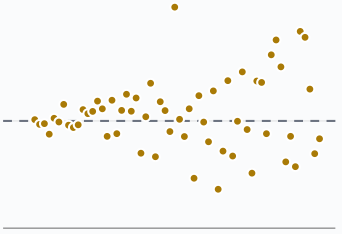
Reading residual plots



Random cloud, constant width: assumptions look reasonable.



U-shape: missing curvature. Try adding X^2.



Fan: spread grows with the fitted value (heteroskedasticity).

WHAT YOU SEE	WHAT IT SUGGESTS	WHAT YOU MIGHT DO
Random cloud around 0, constant width	Assumptions look reasonable	Proceed (this isn't proof)
Curve or systematic pattern	Nonlinearity or missing features	Add X^2, interactions, new predictors
Fan: spread grows with fitted value	Heteroskedasticity (non-constant variance)	Transform Y (e.g. log), weighted least squares. SEs and p-values are unreliable.
Waves or runs of same-sign residuals over time	Autocorrelation (time dependence)	Time-series methods
Heavy tails or extreme points in a Q–Q plot	Non-normal errors, outliers	Investigate outliers; robust loss

A **Q–Q plot** compares the residuals' quantiles with those of a normal distribution. Points along the diagonal mean "approximately normal"; curling away at the ends means heavy tails.

> WORDING THAT EARNS MARKS
>
> Describe the pattern, name the assumption it challenges, and suggest a possible cause or action. For example: "The residuals fan out as fitted values increase, suggesting non-constant variance; standard errors may be unreliable; consider modelling log(Y)." A plot never *proves* anything.

Cheatsheet

Residual plot patterns

Random cloud → OK. Curve → missing nonlinearity. Fan → heteroskedasticity. Waves in time → autocorrelation. Heavy tails → outliers / non-normal.

OLS noise assumptions

Mean zero · constant variance · independent · (roughly normal for t-tests).

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/residuals-and-regression-assumptions

R^2, Adjusted R^2, RMSE and Other Metrics

MAE, MSE, RMSE, MAPE and average error; R^2 as improvement over predicting the mean; out-of-sample R^2 that can go negative; and adjusted R^2, which penalises useless predictors. With worked calculations.

Error metrics

Let $e_i = y_i - \hat{y}_i$ on n evaluation observations, ideally held-out data.

MEASURE	FORMULA	NOTES
MAE	$\frac{1}{n}\sum \lvert e_i \rvert$	Average miss in units of Y. Robust to outliers.
Average error	$\frac{1}{n}\sum e_i$	Signed bias: positives and negatives cancel.
MSE	$\frac{1}{n}\sum e_i^2$	Squared units.
RMSE	$\sqrt{\frac{1}{n}\sum e_i^2}$	Units of Y. Penalises large errors more than MAE.
RSS / SSE	$\sum e_i^2$	Total squared error; grows with n.
MAPE	$\frac{100\%}{n}\sum \left\lvert \frac{e_i}{y_i} \right\rvert$	Percentage error; breaks if $y_i \approx 0$.

WORKED EXAMPLE

Test residuals $2, -1, 3$. MAE $= (2+1+3)/3 = 2$. MSE $= (4+1+9)/3 = 4.67$. RMSE $= \sqrt{4.67} = 2.16$. Average error $= (2-1+3)/3 = 1.33$ (the model tends to under-predict).

Always compare against a **baseline**: the "average predictor" that predicts $\bar{y}_{\text{train}}$ for everyone. A model is only useful if it beats that.

R^2

COEFFICIENT OF DETERMINATION

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \qquad \text{TSS} = \sum_i (y_i - \bar{y})^2$$

TSS is the squared error of the baseline "always predict the mean". So R^2 is the fraction of the baseline's squared error the model removes. In-sample with an intercept, $0 \le R^2 \le 1$.

DATASET B

RSS $= 2.4$ (lesson 3.4). TSS $= 4 + 0 + 1 + 0 + 1 = 6$. $R^2 = 1 - 2.4/6 = 0.6$: the line removes 60% of the mean-only model's squared error.

R^2 CAUTIONS

- $R^2 = 0.75$ does **not** mean 75% of predictions are correct.
- A high in-sample R^2 doesn't guarantee good out-of-sample prediction.
- A low R^2 doesn't mean the predictors are useless: the signal may be nonlinear, or Y very noisy.
- R^2 says nothing about correct specification: check residuals.

Out-of-sample R^2

$$R^2_{\text{OS}} = 1 - \frac{\sum_{\text{test}}(y_i - \hat{y}_i)^2}{\sum_{\text{test}}(y_i - \bar{y}_{\text{train}})^2}$$

It uses held-out data and the **training** mean as the baseline, so it can be **negative**: the model does worse than just predicting the training average. You want it high and close to the in-sample value.

Adjusted R^2

Ordinary R^2 **never decreases** when you add a predictor, even pure noise, because OLS can always set its coefficient to zero and usually finds a tiny bit of in-sample improvement. Adjusted R^2 penalises model size:

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

It goes down when an unhelpful predictor is added and up when a predictor improves fit enough to justify itself.

WORKED EXAMPLE

RSS $= 40$, TSS $= 100$, $n = 20$, $p = 2$. $R^2 = 0.60$. $R^2_{\text{adj}} = 1 - \frac{40/17}{100/19} = 1 - \frac{2.353}{5.263} = 0.553$.

Dataset B ($n = 5$, $p = 1$): $R^2_{\text{adj}} = 1 - \frac{2.4/3}{6/4} = 1 - 0.533 = 0.467$. The penalty is large when n is tiny.

CHOOSING A METRIC	USE
Typical error in plain units, robust to outliers	MAE
Large errors are especially bad	RMSE
Relative error, y never near 0	MAPE
Improvement over the mean baseline	R^2 (out-of-sample)
Comparing in-sample fits with different numbers of predictors	Adjusted R^2 (or better, CV)

Cheatsheet

MAE / RMSE

MAE $= \frac{1}{n}\sum|e_i|$ (robust). RMSE $= \sqrt{\frac{1}{n}\sum e_i^2}$ (punishes big errors). Both in units of Y.

R^2

$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, $\text{TSS} = \sum(y_i - \bar{y})^2$. Share of the mean-baseline's error removed.

Adjusted R^2

$1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$. Falls when a useless predictor is added.

Out-of-sample R^2

Baseline is $\bar{y}_{\text{train}}$ on test data. Can be negative.

R^2 misreadings

Not "% correct". High in-sample ≠ good prediction. Low ≠ useless predictors.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/r-squared-adjusted-r-squared-and-rmse

Standard Errors, Confidence Intervals and p-values

How uncertain is each coefficient? Standard errors, t-statistics, p-values and confidence intervals, with a full worked example, a real regression output to read, and what a p-value is not.

Coefficients are estimates

Collect a new sample and you'd get different $\hat{\beta}$s. Their sampling uncertainty, under uncorrelated errors with equal variance σ^2, is

COEFFICIENT UNCERTAINTY

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}, \qquad \hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1}, \qquad \mathrm{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \left[(X^\top X)^{-1}\right]_{jj}}$$

$n-p-1$ are the **residual degrees of freedom**: n observations minus $p+1$ estimated coefficients.

For simple regression this reduces to $\mathrm{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 / S_{xx}}$. The SE is smaller with **less noise, more data, more spread in X, and less collinearity**.

The t-test

Is the true β_j really non-zero? Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ with

$$t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p-1} \text{ under } H_0$$

The **p-value** is the probability, *assuming H_0 is true* and the model assumptions hold, of a $|t|$ at least as large as the one observed. If $p < 0.05$ we call the coefficient statistically significant at 5%.

Confidence intervals

$$\hat{\beta}_j \pm t_{n-p-1,\,1-\alpha/2} \times \mathrm{SE}(\hat{\beta}_j)$$

With $\alpha = 0.05$: about 95% of intervals built this way (over repeated samples) contain the true β_j. If the interval excludes 0, the two-sided test rejects H_0 at 5%. For large n, $t_{0.975} \approx 1.96$, so "estimate ± 2 SE" is a good shortcut.

WORKED EXAMPLE: DATASET B ($Y = 2.2 + 0.6X$)

RSS $= 2.4$, df $= 5 - 1 - 1 = 3$, so $\hat{\sigma}^2 = 0.8$.

$\mathrm{SE}(\hat{\beta}_1) = \sqrt{0.8/10} = \sqrt{0.08} = 0.283$.

$t = 0.6/0.283 = 2.12$. The critical value $t_{3,\,0.975} = 3.182$. Since $2.12 < 3.182$ we **fail to reject** H_0 ($p \approx 0.12$).

95% CI: $0.6 \pm 3.182 \times 0.283 = 0.6 \pm 0.90 = (-0.30,\ 1.50)$. It includes 0, consistent with the test.

Moral: a positive slope from 5 points is not reliable evidence of a positive true slope.

Size is not significance

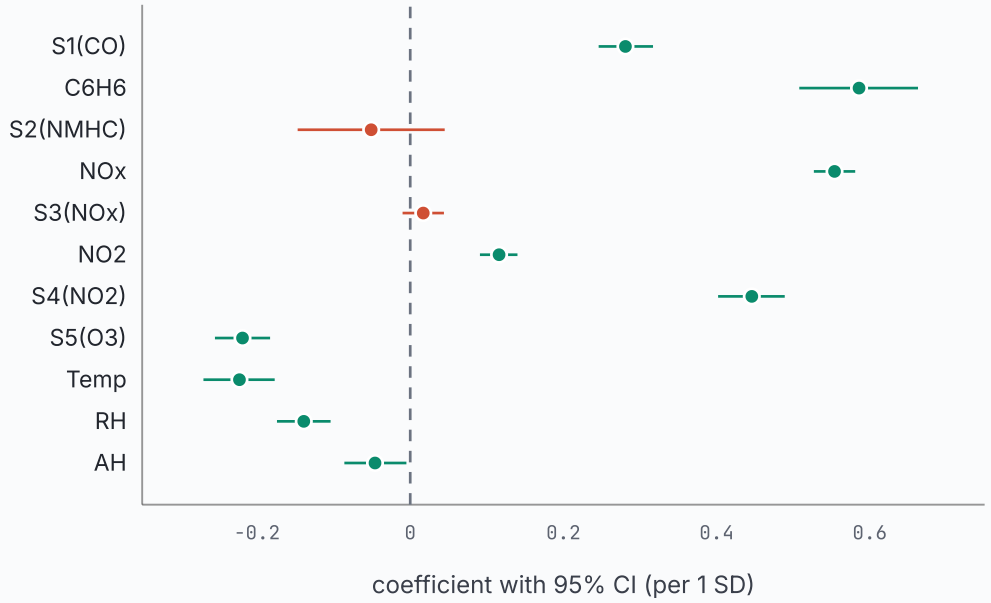
Predictor 1: $\hat{\beta} = 6$, SE $= 3$, so $t = 2$. Predictor 2: $\hat{\beta} = 2$, SE $= 0.5$, so $t = 4$. Predictor 2 has the smaller coefficient but stronger evidence of being non-zero, because it's estimated more precisely. A large coefficient can come with large uncertainty.

WHAT A P-VALUE IS NOT

- Not the probability that H_0 is true.
- Not a measure of effect size: with huge n, trivial effects become "significant".
- A large p-value isn't proof of no effect: small n or collinear predictors can hide a real one.

Reading a real output

Predicting carbon monoxide (mg/m³) from 11 standardised sensor and weather readings in a public air-quality dataset ($n \approx 7{,}700$, 80/20 split). Coefficients are the change in CO per one-SD change in each predictor.



Two intervals (coral) cross zero: those coefficients are not significant at 5%. Training R^2 = 0.888; test R^2 = 0.885.

PREDICTOR	COEF	SE	T	P	VIF
C6H6 (benzene)	0.586	0.040	14.8	0.000	39.9
S2(NMHC) sensor	−0.051	0.049	−1.04	0.298	61.1
NOx	0.554	0.014	39.7	0.000	5.0
S4(NO2) sensor	0.446	0.022	20.0	0.000	12.7
Temperature	−0.223	0.024	−9.5	0.000	14.2

1. **Overall fit:** $R^2 \approx 0.89$ in and out of sample, so little overfitting (n is large relative to $p = 11$).
2. **Significance:** nine of eleven predictors have tiny p-values; S2(NMHC) and S3(NOx) don't.
3. **Why is S2(NMHC) insignificant?** Its VIF is 61 and it correlates 0.98 with benzene. It's nearly a duplicate, so its SE is inflated. That's not evidence the sensor is physically irrelevant: given benzene is in the model, it adds little.
4. **Association ≠ causation:** the negative temperature coefficient doesn't mean cooling the air raises CO; it reflects correlated seasonal and traffic patterns.

Cheatsheet

Standard errors

$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$, $\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(X^\top X)^{-1}]_{jj}}$. Simple: $\sqrt{\hat{\sigma}^2 / S_{xx}}$.

t-test & CI

$t = \hat{\beta}_j / \text{SE}$, df $= n - p - 1$. CI: $\hat{\beta}_j \pm t_{1-\alpha/2}\,\text{SE}$ (≈ ±2 SE).

p-value sentence

"If the true coefficient were zero (and assumptions hold), a t-statistic at least this extreme would occur with probability p."

p-value misreadings

Not P(H_0 true). Not effect size. Large p ≠ no effect.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/standard-errors-confidence-intervals-and-p-values

Multicollinearity and VIF

When predictors carry nearly the same information, coefficients become unstable and standard errors explode, even though predictions can stay fine. The variance inflation factor, how to compute it, and what to do.

The problem

Recall $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$. If $X^\top X$ is **nearly singular**, its inverse has huge entries, so coefficient variances explode. Two causes:

- **Multicollinearity:** a predictor is (almost) a linear combination of others, e.g. two sensors measuring nearly the same pollutant, or height in cm and in inches.
- **Over-parameterisation:** the number of predictors p is close to, equal to or larger than n.

TWO PEOPLE CARRYING ONE SOFA

If two predictors carry nearly the same information, the model can't tell which deserves the credit. Their combined effect is well determined, so predictions are fine, but the split between them is arbitrary and flips from sample to sample.

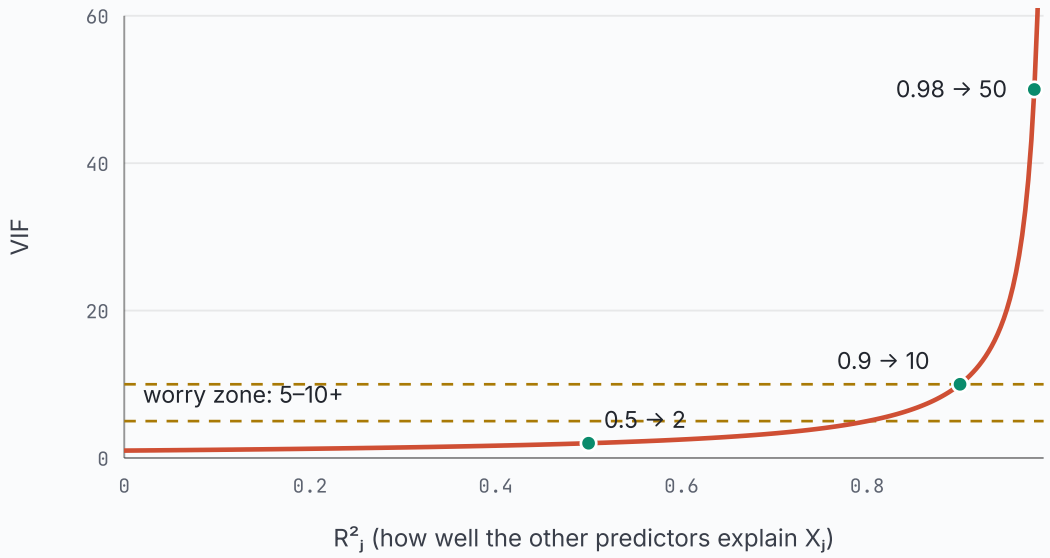
In a simulation with true $\beta_1 = 2$, uncorrelated predictors give estimates clustered around 2; predictors correlated at 0.97 give estimates scattered widely, even though the signal never changed.

Variance inflation factor

VIF

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

R_j^2 is the R^2 from regressing predictor X_j on all the *other* predictors (not on Y).



VIF explodes as a predictor becomes predictable from the others. The variance of its coefficient is inflated by this factor.

Rule of thumb: worry above 5–10. A related diagnostic is the **condition number** of $X^\top X$: huge values mean near-singularity. VIF is a diagnostic, not a decision rule.

WORKED EXAMPLE

A predictor regressed on the others has $R_j^2 = 0.95$. $\mathrm{VIF} = 1/(1-0.95) = 20$: its coefficient's variance is 20 times what it would be with uncorrelated predictors, so its SE is $\sqrt{20} \approx 4.5$ times larger.

What it does and doesn't do

MULTICOLLINEARITY DOES	IT DOES NOT NECESSARILY
Inflate standard errors (predictors look insignificant)	Harm predictions on similar data
Make coefficients unstable, sometimes wrongly signed	Reduce overall R^2
Make interpretation unreliable	Mean the predictors are irrelevant

Remedies: drop or combine redundant predictors; collect more spread-out data; or regularise. Ridge regression (5.4) shares the credit between correlated predictors instead of giving them huge opposite-sign coefficients.

HOW IT'S ASKED

"Two sensors correlate at 0.98. One has a large p-value although it relates to the response on its own. Why?" Given the other sensor, it adds little unique information, so its coefficient is poorly determined (inflated SE, high VIF). Insignificance here isn't irrelevance.

Cheatsheet

VIF

$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$, R_j^2 from X_j on other predictors. Worry > 5–10.

Multicollinearity effects

Inflated SEs, unstable/wrong-sign coefficients. Predictions often fine. Fix: drop/combine, more data, ridge.

Online version with the interactive quiz: airesearchlabs.si/learn/ml/linear-regression/multicollinearity-and-vif