Quantitative Research Methods

Chapter 3: Linear Regression

Prof. Dr. Christoph Weisser

HSBI

Summer Semester 2026

The course at a glance

	Lecture	Topic
	1	Ch. 1 + 2 — Introduction; what is statistical learning
	2	Ch. 2 — Accuracy, bias–variance, Bayes classifier, KNN
▶	**3–4**	**Ch. 3 — Linear regression (estimation, inference, diagnostics)**
	5–6	Ch. 4 — Classification (logistic, LDA/QDA, naive Bayes, ROC)
	7	Ch. 5 — Resampling (validation, CV, bootstrap)
	8	Ch. 6 — Model selection and regularisation (ridge, lasso)
	9	Ch. 7 — Moving beyond linearity (splines, GAMs)
	10	Ch. 8 — Tree-based methods (trees, forests, boosting)
	11	Ch. 10 — Deep learning (MLPs, CNNs, training)
	12	Ch. 13 — Multiple testing (FWER, FDR)

Twelve sessions of 180 minutes. ▶ marks today's chapter. Short exercises every $\sim$20 min; extended exercises every $\sim$45 min; each chapter has a companion lab notebook.

Contents

Optional and advanced material is collected in the **appendix** at the end of the deck.

Notation in this chapter

Symbol	Meaning
$\beta_0,\ \beta_1$	Population intercept and slope; hats ($\hat{\beta}_0, \hat{\beta}_1$) mark least-squares estimates.
$e_i = y_i - \hat{y}_i$	Residual: observed minus fitted value for observation i.
$\mathrm{RSS} = \sum_i e_i^2$	Residual sum of squares — the quantity least squares minimises.
σ, RSE	Error standard deviation; estimated by $\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$.
$\mathrm{SE}(\hat{\beta}_1)$	Standard error: sample-to-sample variability of the slope estimate.
$t = \hat{\beta}_1/\mathrm{SE}(\hat{\beta}_1)$	t-statistic for testing $H_0\colon \beta_1 = 0$.
$R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$	Fraction of variance explained; $\mathrm{TSS} = \sum_i (y_i - \bar{y})^2$.
F	F-statistic: tests all slopes at once, $H_0\colon \beta_1 = \cdots = \beta_p = 0$.
h_{ii}	Leverage of observation i: how unusual its predictor values are.
$\mathrm{VIF}(\hat{\beta}_j)$	Variance inflation factor: collinearity of X_j with the other predictors.

Symbols are listed in order of first appearance; each is defined in full on the slide where it first occurs.

Sources and references

Primary textbook

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023).
An Introduction to Statistical Learning, with Applications in Python. Springer.
`www.statlearning.com`.

About these slides

Compared with the standard chapter deck, every concept is now introduced with explicit *motivation*, *intuition*, *formal definition*, *worked example*, and *practical interpretation*. Slide titles are descriptive; formulas are explained symbol-by-symbol.

Learning objectives for this chapter

By the end of this chapter you will be able to:

- **Estimate** simple and multiple linear regression coefficients by ordinary least squares, and interpret each coefficient correctly — including the "holding others fixed" clause.
- **Test** the relevant hypotheses (t for one coefficient, F for several at once) and **construct** confidence intervals.
- **Distinguish** a confidence interval for the regression line from a prediction interval for an individual outcome.
- **Diagnose** the four classical regression problems — nonlinearity, heteroscedasticity, outliers / leverage, collinearity — and apply the standard remedies.
- **Encode** qualitative predictors and **add** interaction terms following the hierarchy principle.
- **Compare** linear regression with nonparametric alternatives (KNN regression).

Roadmap of this chapter

The path through linear regression

1. Motivation: the Advertising problem.
2. Simple linear regression: model, estimation, inference.
3. Multiple linear regression: confounding, the four key questions.
4. Extensions: dummies, interactions, polynomials.
5. Diagnostics: nonlinearity, heteroscedasticity, leverage, collinearity.
6. Comparison with KNN regression.

Outline

Why is linear regression still in every textbook?

Linear regression was published in 1805. Why are we still teaching it in 2026?

Three reasons

1. It is the *most widely deployed* statistical model in industry — finance, medicine, marketing, public policy.
2. It comes with a *complete inference toolkit* (standard errors, p-values, prediction intervals).
3. It is the *benchmark* against which fancier methods (random forests, neural nets) are routinely compared. Many of those fancier methods only barely beat linear regression on real data.

This chapter gives you everything you need to fit, interpret, diagnose, and defend a linear regression in any setting.

Concrete client problem: the Advertising data

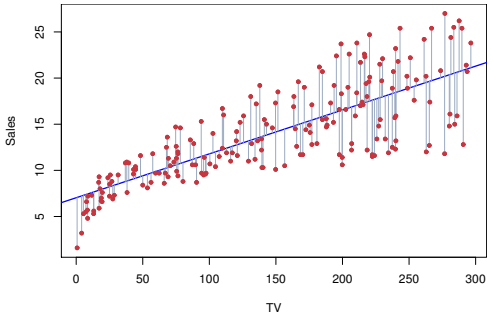
The brief

A marketing client has spent budget on TV, radio, and newspaper advertising in 200 different markets. Sales are recorded per market. The client asks:

"Where should I put next year's budget?"

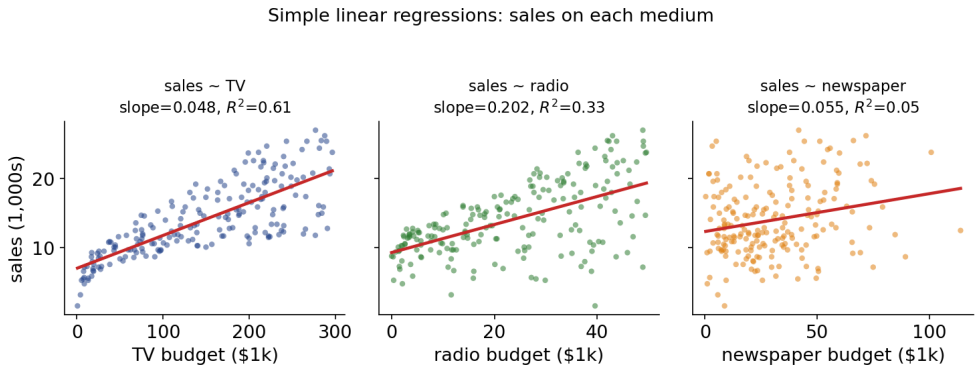
Variables

- Response: `sales` (units, in 1,000s).
- Predictors: `TV`, `radio`, `newspaper` budgets (in \$1,000).
- $n = 200$ markets.



Source: Figure 3.1 — ISLP, James et al. (2023).

The raw data: sales against each medium



First look

Sales rise clearly with TV and radio spend; the newspaper trend is weak. Each panel is a *simple* regression — the multiple regression later will revise these single-predictor slopes.

Seven concrete questions linear regression can answer

Before we look at any formulas, notice that one data set lets us answer many *different* questions:

1. Is there *any* relationship between advertising and sales?
2. How *strong* is that relationship?
3. Which media *contribute* — and which do not?
4. How *precise* are our estimated effects?
5. How well can we *predict* next quarter's sales?
6. Is the relationship between spend and sales *linear*?
7. Is there *synergy* (interaction) among the media?

Each question maps to a specific tool we will develop in this chapter. Keep them in mind as we go.

What linear regression delivers — and what it does not

What you get

- Coefficients you can interpret.
- Standard errors and tests.
- Confidence and prediction intervals.
- Diagnostic plots.

What you do not automatically get

- *Causality.* A coefficient is an association, not a cause.
- *Validity outside the data range.* Extrapolation is always risky.
- *Robustness to gross outliers.* Leverage points can dominate.
- *Capture of nonlinear effects.* You must add them by hand.

Where this chapter is used in industry

Sector	Concrete application	Which decision it drives
Consumer goods, retail	Marketing mix modelling: weekly sales on spend per channel, price, promotion, seasonality	How to split the annual media budget across channels
Real estate, mortgage lending	Hedonic pricing / automated valuation: price on floor area, rooms, age, location	What to list the property at, and how much to lend against it
Asset management	Factor models: excess return on market, size and value factors	How much of the market move to hedge away
Energy, utilities	Load forecasting: demand on temperature, day type, calendar effects	How much power to buy day-ahead, and what to bid
HR, legal	Pay-equity audit: log salary on grade, tenure, location *plus* a group indicator	Which pay adjustments the board signs off
Manufacturing	Yield or quality on process settings: temperature, pressure, feed rate	Which setting to change, and by how much

In industry — The common thread

In all six, the *coefficient* is the deliverable — not the prediction. Someone has to defend that number in a meeting, which is why standard errors, intervals and diagnostics matter as much as the fit.

Industry case in depth: marketing mix modelling (MMM)

In industry — The model an agency actually fits

- Response: weekly log(sales) per brand and region; n is a few hundred brand–weeks.
- Predictors: spend per channel (TV, digital, print, in-store), price index, promotion depth, distribution coverage, month dummies.

What it is used for — and who signs off

Each channel coefficient is read as a contribution and an elasticity, then aggregated into a decomposition of the year's sales — the starting point for next year's budget split, signed off by the CMO with finance and re-fitted quarterly. Its reader is the brand manager whose channel budget it moves.

In industry — Every diagnostic in this chapter shows up here

Channels are flighted together, so spend variables are collinear (VIF); spend has diminishing returns, so log terms are needed; promotional weeks are high-leverage points; error spread grows with the sales level.

The honest caveat

MMM estimates *association*, so turning a coefficient into a budget decision is a judgement — good teams cross-check it against geo experiments or holdout regions.

Outline

Why we need a line through the points

Imagine the marketing client's TV-vs-sales scatter plot. We want one number that summarises the relationship: "how much does sales rise, on average, per \$1,000 of TV spend?"

The simplest possible model

Draw a straight line through the cloud of points so that the line *summarises* the trend. The line's slope is exactly the answer we want.

Two questions follow naturally:

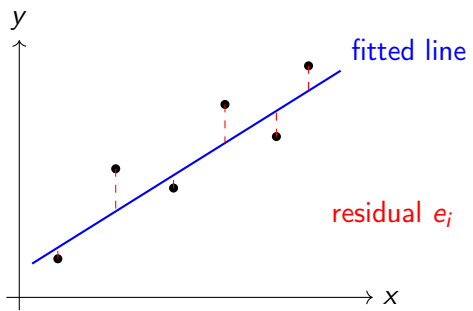
- Among all possible lines, *which one* should we pick?
- How *certain* should we be about that line?

Intuition: minimise the vertical mistakes

Each line makes a mistake on each point. For observation i, the mistake is the vertical distance between the dot y_i and the point on the line above x_i. Call this distance e_i.

We pick the line that makes the total squared mistake as small as possible. Squaring penalises bigger mistakes more.

Why squared and not absolute? Differentiability, tradition (Gauss, 1809), and a clean formula.



The simple linear regression model

We model the response as a straight-line function of the predictor, plus random noise:

Formal model

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{signal (systematic)}} + \underbrace{\varepsilon}_{\text{noise (random)}}$$

with $E[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.

Signal plus noise

The line $\beta_0 + \beta_1 X$ is the part of Y *shared* by every observation at a given X; ε is the part *unique* to each one. Because the noise averages to zero, $Y \approx \beta_0 + \beta_1 X$ describes the *typical* outcome at each X.

Reading the model symbol-by-symbol

How to read this

Symbol	Meaning
Y	Response, the quantity we want to predict (e.g. sales).
X	Predictor / feature, the quantity we measure (e.g. TV spend).
β_0	Intercept: expected Y when $X = 0$.
β_1	Slope: expected change in Y for a 1-unit rise in X.
ε	Error: everything in Y that X does not explain.
σ^2	Variance of the error — the noise floor.

$\beta_0, \beta_1, \sigma^2$ are unknown population parameters; we estimate them from the data and write the estimates with a hat: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$.

Predictions and residuals: definitions

Once we have estimates, two derived quantities matter on every slide for the rest of the chapter.

Fitted values

$$\hat{y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The line's prediction at x_i.

Residuals

$$e_i \;=\; y_i - \hat{y}_i.$$

What was left over — the vertical mistake on observation i.

Takeaway

Residuals are the workhorse of every diagnostic plot you will see. A good model has structureless residuals; structure in residuals is information the model is missing.

The principle: pick the line with smallest squared error

Among all possible (β_0, β_1) we pick the one that minimises:

Residual sum of squares (RSS)

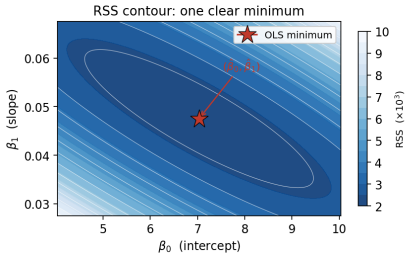
$$\mathrm{RSS}(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} \big(y_i - \beta_0 - \beta_1 x_i\big)^2.$$

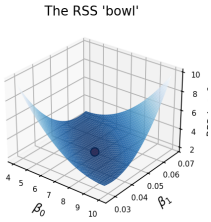
How to read this

Each term in the sum is the vertical mistake for one observation, *squared*. RSS adds them all up. We pick the line that makes this number smallest.

The RSS surface: what least squares is minimising

Least squares minimises $\mathrm{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$ [Advertising: sales on TV]

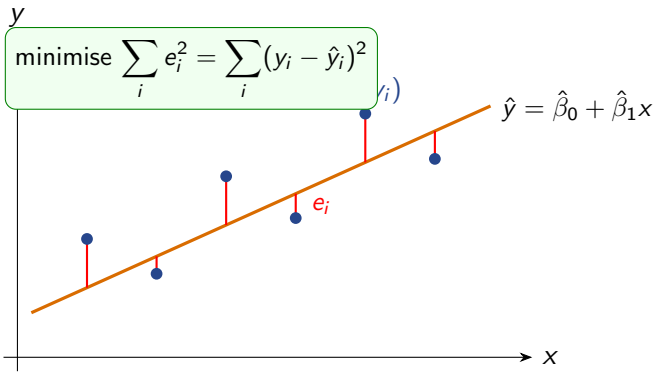




A single lowest point

$\mathrm{RSS}(\beta_0, \beta_1)$ is a convex *bowl* over the plane of candidate coefficients. It has exactly one lowest point — so there is a *unique* best line, and the calculus on the next slides pins it down in closed form, with no trial-and-error search.

Least-squares geometry: minimise the squared residuals



Takeaway

Each red segment is a residual $e_i = y_i - \hat{y}_i$. OLS slides the line until the *sum of squared* segment lengths is as small as possible — larger misses are penalised more heavily.

Deriving the slope: step 1 of 2

Set the partial derivative of RSS with respect to β_1 to zero:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i \right) \stackrel{!}{=} 0.$$

After simplifying (and using the equation for β_0 from the next step) we get the closed form:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{cov}}(X, Y)}{\widehat{\mathrm{var}}(X)}.$$

How to read this

The slope is the sample covariance of X and Y divided by the sample variance of X. Strong covariation of X and Y pushes the slope up; more variation in X (the denominator) pulls it back toward zero.

Deriving the intercept: step 2 of 2

Set the partial derivative of RSS with respect to β_0 to zero:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \stackrel{!}{=} 0.$$

Solve for β_0:

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.}$$

Two consequences for free

- The fitted line always passes through $(\bar{x}, \bar{y})$.
- The residuals sum to zero by construction; they are uncorrelated with X.

A concrete worked example

Suppose we have just five data points (small enough to do by hand):

$$(x_1, y_1), \ldots, (x_5, y_5) = (1, 2),\ (2, 3),\ (3, 5),\ (4, 4),\ (5, 6).$$

Compute means: $\bar{x} = 3$, $\bar{y} = 4$.

Worked example

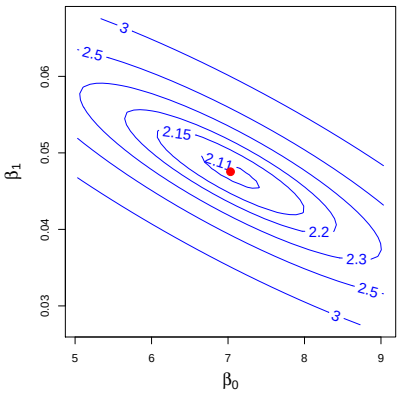
$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 9, \qquad \sum (x_i - \bar{x})^2 = 10.$$

Hence:

$$\hat{\beta}_1 = \tfrac{9}{10} = 0.9, \qquad \hat{\beta}_0 = 4 - 0.9 \cdot 3 = 1.3.$$

The fitted line is $\hat{y} = 1.3 + 0.9x$. Sanity check at $x = 3$: $\hat{y} = 4 = \bar{y}$ ✓.

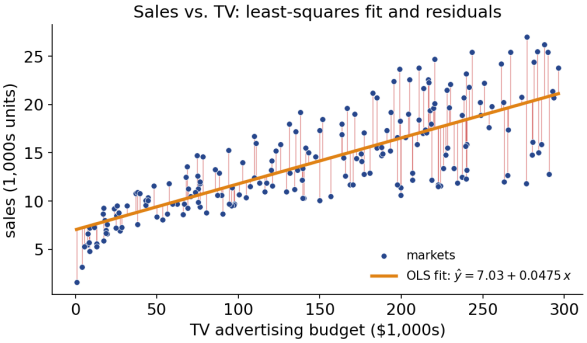
The same bowl, seen from above: RSS contours on real data



Each blue curve joins the (β_0, β_1) pairs with the same RSS on the Advertising data. The red dot at the centre is the least-squares solution $(\hat{\beta}_0, \hat{\beta}_1) = (7.03,\ 0.0475)$ — the bottom of the bowl.

Source: Figure 3.2 — ISLP, James et al. (2023).

Recreating Figure 3.1 on the Advertising data



Takeaway

The orange line is the OLS fit of `sales` on `TV`; each red segment is a residual e_i. Least squares picks the one line that makes the total *squared* residual length smallest.

Exercise 3.1 — Least squares by hand [Math]

Exercise

You observe five data points:

$$(x_i, y_i) = (1,3),\ (2,4),\ (3,2),\ (4,5),\ (5,6).$$

1. Compute the means $\bar{x}$ and $\bar{y}$.
2. Compute $\sum(x_i - \bar{x})(y_i - \bar{y})$ and $\sum(x_i - \bar{x})^2$, then the least-squares estimates $\hat{\beta}_1$ and $\hat{\beta}_0$.
3. Write the fitted line and predict y at $x = 6$.
4. Interpret the slope in words.

Solution — Exercise 3.1 (1/2)

Solution

Method. Every OLS quantity is built from deviations from the means, so we first compute $\bar{x}, \bar{y}$, then the co-movement S_{xy} and the predictor spread S_{xx}.

(1) Means.

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3, \qquad \bar{y} = \frac{3+4+2+5+6}{5} = 4.$$

(2) Cross-product and spread. Tabulate deviations:

$x_i - \bar{x}$	-2	-1	0	1	2
$y_i - \bar{y}$	-1	0	-2	1	2

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 2 + 0 + 0 + 1 + 4 = 7, \qquad \sum (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

Hence

$$\hat{\beta}_1 = \frac{7}{10} = 0.7, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4 - 0.7 \cdot 3 = 1.9.$$

Solution — Exercise 3.1 (2/2)

Solution

(3) Fitted line and prediction.

$$\hat{y} = 1.9 + 0.7\,x, \qquad \hat{y}(6) = 1.9 + 0.7 \cdot 6 = 6.1.$$

Sanity check: at $x = \bar{x} = 3$, $\hat{y} = 1.9 + 2.1 = 4 = \bar{y}$ ✓ (the line passes through $(\bar{x}, \bar{y})$).
(4) Interpretation. A one-unit increase in x is associated with an average increase of 0.7 units in y. The intercept 1.9 is the predicted y when $x = 0$ (only meaningful if $x = 0$ is within a sensible range).

Common mistake

treating $\hat{y}(6) = 6.1$ as just as trustworthy as an in-range prediction. $x = 6$ lies outside the observed range $[1, 5]$, so this is (mild) extrapolation — the further we move from the data, the less the fitted line can be trusted.

Why estimates carry uncertainty

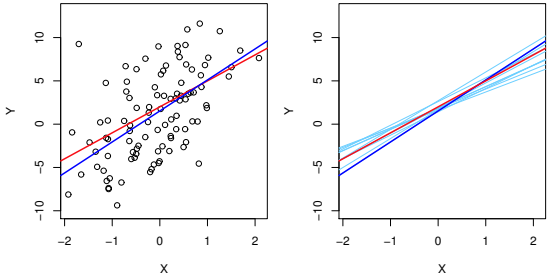
Our $\hat{\beta}_0$ and $\hat{\beta}_1$ were computed from a particular sample. A different sample from the same population would have produced slightly different estimates.

Thought experiment

Imagine drawing 100 independent samples of 200 markets each and fitting OLS each time. We would get 100 different lines, each near but not on the true population line. The *spread* of those 100 lines is the sampling variability we care about.

Standard errors quantify this spread *from a single sample*. Confidence intervals and hypothesis tests are built on top.

Population vs. sample lines (illustration)



Takeaway

The flatter heavy line is the *population* line: fixed, and never observed. Every other line is least squares on one sample; they wobble around it, and a standard error measures that wobble from a *single* sample.

Source: Figure 3.3 — ISLP, James et al. (2023).

Standard errors of the OLS estimates

A *standard error* measures how much an estimate would bounce around from one random sample to the next. Under the classical assumptions of independent errors with constant variance σ^2:

Formulas

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right],$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

In practice σ^2 is replaced by $\hat{\sigma}^2 = \mathrm{RSS}/(n-2)$, the squared *residual standard error*.

Reading the standard-error formulas

How to read this

- The *denominator* of $\mathrm{SE}(\hat{\beta}_1)^2$ is $\sum(x_i - \bar{x})^2$. *More spread-out predictors* $\Rightarrow$ *smaller SE for the slope.* (Long lever arm.)
- Both formulas have σ^2 in the numerator. *Less noise* $\Rightarrow$ *smaller SEs.* Obvious in hindsight.
- $\mathrm{SE}(\hat{\beta}_0)^2$ shrinks like $1/n$ in n, so *more data* $\Rightarrow$ *tighter intercept estimate.*

Practical implication

If you can choose where to measure, *spread the x_i's widely*. A study that measures $x \in \{2, 3, 4\}$ produces a much less precise slope than one that measures $x \in \{1, 5, 10\}$.

Confidence interval: construction

Approximately,

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

is a 95 % confidence interval for β_1. (Strictly, use $t_{0.025,\, n-2}$ instead of 2.)

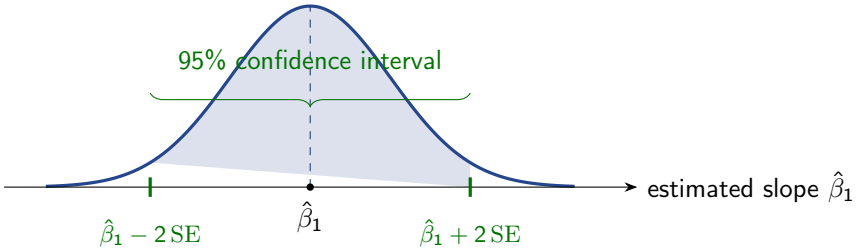
Advertising example

For TV: $\hat{\beta}_1 = 0.0475$, $\mathrm{SE}(\hat{\beta}_1) = 0.0027$.
95 % CI: $[0.0475 - 2 \cdot 0.0027,\ 0.0475 + 2 \cdot 0.0027] = [0.0421,\ 0.0529]$.
Reading: a \$1,000 increase in TV spend is associated with between 42 and 53 extra units of sales, with 95 % confidence.

In industry — Marketing: what the controller actually asks for

Before signing off a budget shift, finance asks for the *interval*, not the point estimate: it shows whether the channel's payback is comfortably positive or could plausibly be zero.

A picture of the 95% confidence interval



Reading the picture

The bell is the sampling distribution of $\hat{\beta}_1$; the shaded band $\hat{\beta}_1 \pm 2\,\mathrm{SE}$ is the 95% confidence interval. The *t*-test asks the yes/no version of the same question — *does* 0 *fall outside this band?* — and if it does, the slope is significant at the 5% level.

Confidence interval: interpretation pitfalls

Common misreading

"There is a 95 % probability that β_1 lies in $[a, b]$."

This is *wrong* under classical (frequentist) inference.

Correct interpretation

The probability statement is about the random *interval*, not the fixed parameter β_1: *If we repeated the experiment many times, 95 % of the constructed intervals would contain the true β_1.*

For a Bayesian probability statement about β_1 given the data, you would need a prior and a posterior.

Exercise 3.2 — Standard error, t and CI [Math]

Exercise

A simple linear regression is fit on $n = 10$ observations. You are given $\hat{\beta}_1 = 2.5$, residual standard error $\mathrm{RSE} = 4.0$, and $\sum (x_i - \bar{x})^2 = 16$.

1. Compute the standard error $\mathrm{SE}(\hat{\beta}_1) = \mathrm{RSE} / \sqrt{\sum (x_i - \bar{x})^2}$.
2. Compute the t-statistic for $H_0 : \beta_1 = 0$.
3. Build a 95% confidence interval for β_1 (use $t_{0.025,\,8} = 2.306$).
4. Is the slope significant at the 5% level? Justify.

Solution — Exercise 3.2

Solution

(1) Standard error.

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\mathrm{RSE}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{4.0}{\sqrt{16}} = \frac{4.0}{4} = 1.0.$$

(2) t-statistic.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} = \frac{2.5}{1.0} = 2.5, \qquad \mathrm{df} = n - 2 = 8.$$

(3) 95% confidence interval.

$$\hat{\beta}_1 \pm t_{0.025,8} \cdot \mathrm{SE}(\hat{\beta}_1) = 2.5 \pm 2.306 \cdot 1.0 = [0.194,\ 4.806].$$

(4) Significance. $|t| = 2.5 > 2.306$, equivalently the CI excludes 0. We reject $H_0 : \beta_1 = 0$ at the 5% level; the two-sided p-value (≈ 0.037) is below 0.05. There is evidence of a real linear association between x and y.

Note: with only $n = 10$ the critical value 2.306 is well above the large-sample value 1.96; small samples demand larger t's to reach significance.

Common mistake: dividing the RSE by $\sqrt{n}$. The slope's SE divides by $\sqrt{\sum (x_i - \bar{x})^2}$ — more spread-out predictor values give a more precise slope, regardless of n.

The first question: is there *any* relationship?

Before we do anything fancy, we test whether the predictor has any effect at all.

Hypotheses

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 \neq 0.$$

In English: "*TV has no effect on sales*" against "*TV has some effect on sales*".

Why start here

If $\beta_1 = 0$ the line is flat and X tells us nothing about Y. Rejecting H_0 is what *licenses* interpreting the slope at all — so this test comes *before* we read off effect sizes or build confidence intervals.

The t-statistic, step by step

Test statistic

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{n-2} \text{ under } H_0.$$

How to read this

Numerator: *how far* from zero is our estimate? Denominator: *how surprising* would such a deviation be by sampling fluctuation alone? A large $|t|$ says: "too far from zero to be a coincidence."

Advertising TV regression

$t = 0.0475/0.0027 \approx 17.6$. The two-sided p-value is essentially zero — we reject the null at any reasonable level.

Reading the regression output

	Coef.	Std. Err.	t	p
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

How to read this

- Each \$1,000 of TV advertising is associated with about 47.5 extra units of sales on average.
- The slope is highly statistically significant ($t = 17.67$).
- The intercept (about 7,032 units of sales when TV = \$0) is also significantly different from zero — there is a non-zero baseline of sales.

Two summary measures of fit

After fitting any regression we want to know how well it fits.

Residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}}.$$

The average size of a residual, in the units of Y. Smaller is better.

Coefficient of determination R^2

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \quad \mathrm{TSS} = \sum (y_i - \bar{y})^2.$$

Proportion of variance in Y that the model explains.

R^2: how to read it

How to read this

R^2 ranges between 0 and 1 in simple linear regression.

- $R^2 = 1$: model explains all variation in Y. The points sit exactly on the line.
- $R^2 = 0$: knowing X tells us nothing about Y.
- In simple linear regression $R^2 = \mathrm{cor}(X, Y)^2$.

Advertising TV regression

$\mathrm{RSE} = 3.26$, $R^2 = 0.61$. Translation: typical prediction is off by 3,260 units of sales; the model explains 61 % of the variability in sales across markets.

Common misinterpretations of R^2

High R^2 does *not* prove

- the model is *correct*,
- the predictors *cause* Y,
- the model will *predict well* on new data.

Low R^2 does *not* prove

- the model is *useless* — in finance an R^2 of 0.05 can be very valuable,
- the predictors are *irrelevant*,
- there is *no signal* to be found.

In industry — Asset management: factor models

Regress a fund's excess return on market, size and value factors: R^2 is the share of its movement explained by known factors — the rest is what the manager charges for.

Exercise 3.3 — R^2, RSE and cor^2 [Concept]

Exercise

For the five points of Exercise 3.1 the fit gave $\mathrm{TSS} = \sum(y_i - \bar{y})^2 = 10$ and $\mathrm{RSS} = \sum(y_i - \hat{y}_i)^2 = 5.1$ with $n = 5$.

1. Compute R^2 and state in words what it means.
2. Compute the residual standard error $\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$ and give its units interpretation.
3. Recall $\mathrm{cor}(x, y) = 0.7$. Verify the identity $R^2 = \mathrm{cor}(x, y)^2$ that holds in simple linear regression, and explain why R^2 alone can be misleading.

Solution — Exercise 3.3 (1/2)

Solution

(1) Coefficient of determination.

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{5.1}{10} = 0.49.$$

So 49% of the variance of y is explained by the linear fit on x; the remaining 51% is unexplained (noise or missing structure).

(2) Residual standard error.

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}} = \sqrt{\frac{5.1}{3}} = \sqrt{1.7} \approx 1.30.$$

RSE is measured in the *units of* y: a typical prediction misses the observed value by about 1.30 units. Unlike R^2 it is not scale-free.

Solution — Exercise 3.3 (2/2)

Solution

(3) Identity. Substituting the given correlation:

$$\mathrm{cor}(x, y)^2 = 0.7^2 = 0.49 = R^2.\checkmark$$

Why it holds: with one predictor the fitted values are a linear function of x ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$), so $\mathrm{cor}(y, \hat{y}) = |\mathrm{cor}(x, y)|$; and in any OLS fit with an intercept, $R^2 = \mathrm{cor}(y, \hat{y})^2$. With several predictors only the latter survives, so the identity is special to *simple* regression.
Why R^2 alone can mislead: it can be high even when the linear form is wrong (curvature masked by a strong trend), and it never decreases when predictors are added — so a large R^2 is not by itself evidence of a good or useful model. Check the residual plot and the RSE as well.

Common mistake

reading $R^2 = 0.49$ as "49% of points are predicted correctly". It is the share of *variance* explained — a statement about spread, not about hit rates.

Extended Exercise 3.L1 — Simple regression from A to Z [Math]

Extended exercise (15 min)

A dealer records the resale value y (in \$1000s) of a used car against its age x (in years) for six cars:

x_i (age)	1	2	3	4	5	6
y_i (value)	19	18	17	15	12	9

Carry out a *complete* simple linear regression by hand:

1. Estimate $\hat{\beta}_0, \hat{\beta}_1$ and write the fitted line.
2. Compute the fitted values, residuals, RSS, RSE and R^2.
3. Compute $\mathrm{SE}(\hat{\beta}_1)$, the t-statistic and the 95% confidence interval for β_1.
4. Predict the value of a 7-year-old car with a 95% *prediction* interval. Interpret every number.

Useful: $t_{0.025,\,4} = 2.776$.

Solution — Extended Exercise 3.L1 (1/3)

Solution

Summaries. $n = 6$, $\bar{x} = 3.5$, $\bar{y} = 15$. With $x_i - \bar{x} = (-2.5, -1.5, -0.5, 0.5, 1.5, 2.5)$ and $y_i - \bar{y} = (4, 3, 2, 0, -3, -6)$,

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = -35, \quad S_{xx} = \sum (x_i - \bar{x})^2 = 17.5, \quad S_{yy} = \sum (y_i - \bar{y})^2 = 74.$$

(1) Slope, intercept, line.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-35}{17.5} = -2, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 15 - (-2)(3.5) = 22.$$

$$\boxed{\hat{y} = 22 - 2\,x}$$ each extra year of age lowers value by \$2000.

Solution — Extended Exercise 3.L1 (2/3)

Solution

(2) Fit quality. Fitted $\hat{y}_i = (20, 18, 16, 14, 12, 10)$; residuals $e_i = (-1, 0, 1, 1, 0, -1)$.

$$\mathrm{RSS} = \sum e_i^2 = 4, \qquad \mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}} = \sqrt{\frac{4}{4}} = 1.0, \qquad R^2 = 1 - \frac{\mathrm{RSS}}{S_{yy}} = 1 - \frac{4}{74} = 0.946.$$

Interpretation: a typical prediction misses the true value by about $\mathrm{RSE} = \$1000$; age alone explains $R^2 = 94.6\%$ of the variation in resale value.

(3) Inference on the slope.

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\mathrm{RSE}}{\sqrt{S_{xx}}} = \frac{1}{\sqrt{17.5}} = 0.239, \qquad t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{-2}{0.239} = -8.37.$$

Since $|t| = 8.37 \gg 2.776$, we reject $H_0 : \beta_1 = 0$: age is a highly significant predictor of value.

Solution — Extended Exercise 3.L1 (3/3)

Solution

95% CI for β_1.

$$\hat{\beta}_1 \pm t_{0.025,4}\,\mathrm{SE}(\hat{\beta}_1) = -2 \pm 2.776(0.239) = [-2.66,\ -1.34].$$

We are 95% confident the annual drop is between \$1340 and \$2660; the interval excludes 0, matching the t-test.

(4) Prediction at $x_0 = 7$. $\hat{y}_0 = 22 - 2(7) = 8$ (\$8000). A *prediction* interval for one new car adds the irreducible error (the "1+"):

$$\hat{y}_0 \pm t\,\mathrm{RSE}\sqrt{1 + \tfrac{1}{n} + \tfrac{(x_0 - \bar{x})^2}{S_{xx}}} = 8 \pm 2.776(1)\sqrt{1 + \tfrac{1}{6} + \tfrac{12.25}{17.5}} = 8 \pm 3.79,$$

i.e. [4.21, 11.79] (\$4210–\$11,790). It is wide because n is small, $x_0 = 7$ extrapolates beyond the data, and it targets a *single* car rather than the average.

Common mistake

dropping the "1+" under the root. That gives the *confidence* interval for the average 7-year-old car — 8 ± 2.58 — far too narrow for one specific car.

Outline

Why one predictor is rarely enough

Real applications rarely involve a single explanatory variable. The client cares about TV, radio, *and* newspaper at once — and they may interact.

Takeaway

Multiple regression lets us study the effect of one predictor *while controlling for the others*. This "ceteris paribus" interpretation is what makes regression so useful for inference.

Why not just fit p separate simple regressions?

Separate one-predictor fits cannot answer "what is the effect of TV *holding radio fixed*?" If predictors move together, each isolated slope quietly absorbs the others' effects. One joint model disentangles them.

The multiple linear regression model

Formal model

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

Each β_j is interpreted as:

The expected change in Y for a one-unit increase in X_j, holding all other predictors fixed.

The clause that matters

The phrase "holding others fixed" is exactly what makes regression so useful for inference — and so easy to misinterpret in the presence of confounders or collinearity.

Exercise 3.4 — Multiple OLS on Advertising [Python]

Exercise

Using `statsmodels`, fit `sales` $\sim$ `TV` + `radio` + `newspaper` on the Advertising data.

1. Write the code to fit the model and print the summary.
2. Given the estimated coefficients (intercept 2.94, TV 0.0458, radio 0.189, newspaper -0.001), interpret the TV and radio slopes.
3. Newspaper has a large p-value. What do you conclude, and how much extra sales does \$1000 of radio buy?

Solution — Exercise 3.4 (1/2)

Solution

(1) Code.

```
import pandas as pd # data handling
import statsmodels.api as sm # OLS with full inference output

ad = pd.read_csv("../ALL CSV FILES - 2nd Edition/Advertising.csv",
                 index_col=0) # load data; column 0 = row labels
X = sm.add_constant(ad[["TV", "radio", "newspaper"]]) # + intercept col
res = sm.OLS(ad["sales"], X).fit() # fit sales on all three media
print(res.summary()) # coef table: SE, t, p, R2, F
# TV ~0.0458, radio ~0.189, newspaper ~-0.001 (p ~0.86)
# R-squared ~0.897
```

Solution — Exercise 3.4 (2/2)

Solution

(2) Interpretation (ceteris paribus). Holding the other media fixed, spending an extra \$1000 on *TV* is associated with about $0.0458 \times 1000 = 45.8$ additional units sold; an extra \$1000 on *radio* is associated with about $0.189 \times 1000 = 189$ additional units. (Budgets are in thousands of dollars and sales in thousands of units.)

(3) Newspaper. Its coefficient is essentially zero with a large p-value, so once TV and radio are in the model, newspaper adds no detectable predictive value. \$1000 of radio buys roughly 189 extra units — radio is far more effective per dollar than the tiny (non-significant) newspaper effect. The model explains about 90% of the variance in sales ($R^2 \approx 0.897$).

Common mistake

reading the near-zero newspaper slope as "newspaper advertising can never work". The model only says that *given* TV and radio, newspaper adds no predictive value in these markets; its positive simple-regression slope arises because newspaper spend travels with radio spend (see Exercise 3.5).

A cautionary tale: the Advertising slopes

Comparing single-predictor and multiple-predictor slopes side by side:

	Univariate slope	Multiple slope
TV	0.0475	0.0458
Radio	0.2025	0.1885
Newspaper	0.0547	−0.0010

What just happened?

Newspaper looks predictive on its own — but its effect essentially *disappears* once TV and radio are in the model. The univariate correlation came from newspaper spend co-moving with radio spend.

Why a univariate t-test can mislead

Confounding in one sentence

Two predictors that move together can each look significant on their own, but only one of them actually drives the response.

Practical rule

Always run *multiple* regression when several predictors might be related to one another. Univariate t-tests are a useful diagnostic but not a substitute for joint analysis.

Exercise 3.5 — Why newspaper "works" alone [Concept]

Exercise

On the Advertising data, a *simple* regression of sales on newspaper alone gives a positive, significant slope. But in the *multiple* regression of Exercise 3.4 the newspaper coefficient is essentially zero and not significant. The correlation between newspaper and radio spending is about 0.35.

1. Explain how newspaper can look helpful on its own yet be useless in the joint model.
2. Which variable is the true driver, and what role is newspaper playing in the simple regression?
3. State the general lesson for interpreting simple correlations.

Solution — Exercise 3.5

Solution

(1) Confounding. Markets that spend more on radio also tend to spend somewhat more on newspaper ($\mathrm{cor} \approx 0.35$). Radio genuinely raises sales. In a simple regression, newspaper acts as a *proxy* for the radio spending that travels with it, so it "inherits" a positive slope. Once radio is included in the multiple regression, radio absorbs that effect and newspaper's own coefficient collapses toward zero.
(2) True driver. Radio (and TV) drive sales. In the simple regression newspaper is a surrogate / confounded stand-in for radio, not a cause of sales.
(3) General lesson. A marginal (simple) association can be entirely due to correlation with an omitted variable. To assess the partial effect of a predictor — its contribution holding others fixed — you must include the other relevant predictors. Simple correlations invite confounding; "correlation is not causation" is the practical warning.

Common mistake

concluding "newspaper advertising cannot work". The model says newspaper adds nothing *given TV and radio, in these data* — an observational statement, not a causal verdict.

Four questions to ask of every multiple regression

After fitting any non-trivial linear model, ask:

1. Is at least one predictor useful? (F-test)
2. Which subset of predictors is useful? (t-tests, Ch. 6)
3. How well does the model fit the data? (RSE, R^2, plots)
4. How accurate is a prediction at a new x_0? (CIs, PIs)

Each question has a tailored statistical answer; we treat them in turn.

Q1. Is *any* predictor useful? — the F-test

Test the joint null that no predictor matters at all:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

F-statistic

$$F \;=\; \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \;\sim\; F_{p,\,n-p-1} \text{ under } H_0.$$

How to read this

Numerator: *average* reduction in RSS per added predictor. Denominator: *average* residual variance per remaining degree of freedom. Large $F \Rightarrow$ predictors are explaining more than chance.

Why not just p separate t-tests?

Multiple-testing inflation

With p predictors and $\alpha = 0.05$ each, even under the *global null* we expect $0.05p$ rejections by chance alone. With $p = 20$ predictors that is one false positive per regression, on average.

Takeaway

The F-test handles this multiple-comparisons issue *globally*. We will return to formal multiple-testing corrections in Chapter 13.

Q2. Which predictors are useful?

This is the variable-selection question. Three approaches in increasing sophistication:

Approach 1: individual t-tests

Read off the regression table; useful when p is small. Beware multiplicity.

Approach 2: subset selection (Ch. 6)

Best subset, forward stepwise, backward stepwise. Use information criteria (C_p, AIC, BIC) or cross-validation to choose.

Approach 3: regularisation (Ch. 6)

The lasso adds an ℓ_1 penalty that shrinks irrelevant coefficients to exactly zero — automatic variable selection.

Q3. How well does the model fit?

Three complementary tools:

- **Residual standard error (RSE):** typical residual magnitude in the units of Y — a concrete number.
- R^2 **(and adjusted** R^2**):** proportion of variance explained, useful for relative comparisons.
- **Residual plots:** structure in residuals signals model misspecification (next section).

Important nuance

Adding any predictor never *decreases* R^2, even if the predictor is pure noise. *Adjusted* R^2 (Ch. 6) penalises this; cross-validation is more reliable still.

Q4. How precise is a prediction?

At a new point x_0 we have two distinct uncertainty quantities:

Confidence interval for $E[Y \mid X = x_0]$

Quantifies the uncertainty in the *regression line* at x_0. Narrows as $n \to \infty$.

Prediction interval for an individual Y_0

Quantifies the uncertainty in the line *plus* the inherent noise in Y_0. Includes σ^2 even at infinite n.

Key fact (often confused)

Prediction intervals are always wider than confidence intervals. Use prediction intervals when communicating uncertainty about an individual outcome.

Exercise 3.6 — The F-test vs. many t-tests [Math]

Exercise

A multiple regression with $p = 3$ predictors is fit on $n = 100$ observations, giving $\mathrm{TSS} = 1000$ and $\mathrm{RSS} = 100$.

1. Compute the overall F-statistic $F = \dfrac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$.
2. Compare with $F_{0.05,\,3,\,96} \approx 2.70$. What do you conclude?
3. Suppose instead $p = 100$ predictors were all truly useless. Roughly how many would appear "significant" at $\alpha = 0.05$ using individual t-tests? Why does this motivate the F-test?

Solution — Exercise 3.6

Solution

(1) F-statistic.

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} = \frac{(1000-100)/3}{100/(100-3-1)} = \frac{900/3}{100/96} = \frac{300}{1.0417} \approx 288.$$

(2) Conclusion. $288 \gg 2.70$, so we overwhelmingly reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. At least one predictor is related to the response.

(3) Multiple testing. If all 100 predictors are useless, each individual t-test still rejects with probability $\alpha = 0.05$ by chance, so we expect about $100 \times 0.05 = 5$ "significant" predictors purely from noise. Scanning many t-tests therefore almost guarantees false positives. The F-test gives a single overall test whose Type I error stays at α regardless of p, which is why we consult it *first* before looking at individual coefficients.

Common mistake

reading the significant F as "all three predictors matter". It says only that *at least one* does; the individual t-tests (and Chapter 6) decide which.

Extended Exercise 3.L3 — Multiple regression on Advertising [Python]

Extended exercise (15 min)

Using `statsmodels` and `Advertising.csv`, fit the multiple regression `sales` $\sim$ `TV` + `radio` + `newspaper`.

1. Fit the model and read the summary: interpret each slope, the overall F-test, and R^2 vs adjusted R^2.
2. In a *simple* regression `sales` $\sim$ `newspaper` the newspaper slope is significant, yet here it is not. Explain the apparent contradiction (confounding).
3. Drop newspaper and refit. Compare coefficients, R^2 and RSE. Which model do you keep, and why?

Solution — Extended Exercise 3.L3 (1/3)

Solution

(1) Fit and summary.

```
import pandas as pd, numpy as np, statsmodels.api as sm # libraries
ad = pd.read_csv("../ALL CSV FILES - 2nd Edition/Advertising.csv",
                 index_col=0) # load data; col 0 = row labels
X = sm.add_constant(ad[["TV", "radio", "newspaper"]]) # design + const
res = sm.OLS(ad["sales"], X).fit() # least-squares fit
print(res.summary()) # full inference table
# const 2.939 | TV 0.0458 (p<.001) | radio 0.189 (p<.001)
# newspaper -0.0010 (p=0.86, 95% CI [-0.013, 0.010])
# R2 = 0.897 | adj-R2 = 0.896 | F = 570.3 (p = 1.6e-96) | RSE = 1.686
```

Solution — Extended Exercise 3.L3 (2/3)

Solution

Reading it. Holding the others fixed, \$1000 more on *TV* adds ≈ 45.8 units of sales and \$1000 more on *radio* adds ≈ 188 units; *newspaper* is ≈ 0 ($p = 0.86$; its CI straddles 0). The F-statistic 570 (tiny p) tests $H_0 : \beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{news}} = 0$ and is decisively rejected — at least one medium matters. Here $R^2 = 0.897$ and adjusted $R^2 = 0.896$ are almost equal because there are only three predictors.

(2) Why newspaper "flips". Alone, `sales` $\sim$ `newspaper` has slope 0.055 ($p = 0.001$). But $\text{cor}(\texttt{radio}, \texttt{newspaper}) = 0.35$: markets with big radio budgets also buy more newspaper. Alone, newspaper acts as a *proxy* for radio and borrows its effect; once radio is in the model, newspaper adds nothing. This is textbook *confounding*.

Solution — Extended Exercise 3.L3 (3/3)

Solution

(3) Drop newspaper and refit.

```
X2 = sm.add_constant(ad[["TV", "radio"]]) # drop newspaper column
res2 = sm.OLS(ad["sales"], X2).fit() # refit the smaller model
# const 2.921 | TV 0.0458 | radio 0.188
# R2 = 0.897 (unchanged) | adj-R2 0.896 -> 0.896 | RSE 1.686 -> 1.681
```

The TV and radio slopes barely move, R^2 is essentially unchanged, and RSE improves slightly. **Keep the two-predictor model**: it is simpler, loses no explanatory power, and both adjusted R^2 and RSE (which penalise useless predictors) favour it. Parsimony wins, and it agrees with the confounding diagnosis in part (2).

Common mistake

citing the unchanged R^2 as the evidence. Plain R^2 can only fall when a predictor is dropped — adjusted R^2 and RSE, which penalise dead weight, are the fair judges.

Outline

What if a predictor is categorical?

Many real predictors are categorical: gender, region, brand, day of the week. We cannot use them directly — regression needs numbers.

Solution: dummy coding

Replace a categorical predictor with one or more 0/1 indicator variables. The math then works exactly as before.

Coding "owns a home?"

`Own` $\in$ {Yes, No} becomes one column: $x = 1$ for owners, $x = 0$ for non-owners. Its coefficient is then the average Y-difference between owners and non-owners.

Two levels: a single 0/1 dummy

For a binary predictor with levels A and B:

$$x_i = \begin{cases} 1 & \text{if observation } i \text{ is in level A,} \\ 0 & \text{if in level B (the baseline).} \end{cases}$$

How to read this

The coefficient β_j is interpreted as the average difference in Y between level A and the baseline B, holding all other predictors fixed. Either level can serve as baseline; only the sign of $\hat{\beta}_j$ flips.

In industry — HR and legal: the adjusted pay gap

Pay-equity audits regress log(salary) on job grade, tenure, performance and location *plus* a protected-group dummy. That single coefficient — the gap remaining after the legitimate factors are held fixed — is the number HR and legal report to the board.

K levels: use $K - 1$ dummies

Use $K - 1$ dummy variables and let one level serve as the baseline.

Why not K dummies?

With K dummies, the dummies sum to a column of ones — collinear with the intercept. The design matrix would not be full rank and OLS would have no unique solution.

Three-region predictor

Suppose region $\in$ {North, South, East}. Set North as baseline. Use two dummies: $x_S = 1$ if South, $x_E = 1$ if East. The coefficients β_S, β_E are the average Y-differences relative to North, holding others fixed.

Exercise 3.7 — Dummy coding factors [Math]

Exercise

1. **Two levels.** Regressing credit-card balance on a student indicator gives $\widehat{\texttt{balance}} = 480.4 + 396.5 \cdot \texttt{student}$, where $\texttt{student} = 1$ for students, 0 otherwise. Give the predicted average balance for students and for non-students, and interpret 396.5.
2. **Three levels.** For region with East as the baseline and dummies for South and West, the fit is

$$\widehat{\texttt{balance}} = 531 - 12.5 \cdot \texttt{South} - 18.7 \cdot \texttt{West}.$$

   Give the predicted mean balance in each region.
3. Why do we use only $k - 1$ dummies for a k-level factor?

Solution — Exercise 3.7

Solution

(1) Two levels.

$$\text{non-student } (\texttt{student} = 0):\ \hat{y} = 480.4, \qquad \text{student } (\texttt{student} = 1):\ \hat{y} = 480.4 + 396.5 = 876.9.$$

The coefficient 396.5 is the *difference* in average balance: students carry, on average, \$396.5 more balance than non-students. The intercept 480.4 is the baseline (non-student) mean.

(2) Three levels. The intercept 531 is the baseline (East) mean; each dummy coefficient is a difference from East.

$$\text{East} = 531, \qquad \text{South} = 531 - 12.5 = 518.5, \qquad \text{West} = 531 - 18.7 = 512.3.$$

(3) Why $k-1$ dummies. A separate dummy for every level plus an intercept would be perfectly collinear (the dummies sum to the all-ones intercept column), so the design matrix is not full rank and the coefficients are not identified — the "dummy variable trap". Dropping one level makes it the baseline that the intercept represents; the remaining $k-1$ coefficients are contrasts against that baseline. Overall differences among levels are still tested jointly with an F-test.

Common mistake: comparing South and West directly by their coefficients as if each were a mean — both are contrasts against *East*. South vs. West is $-12.5 - (-18.7) = +6.2$: South's average balance is \$6.20 *above* West's.

Why a plain linear model can miss the truth

Plain linear regression assumes the effect of X_1 is the *same* at every value of X_2. Often this is wrong.

A simple synergy

Spending \$50k on TV *and* \$30k on radio yields more sales than spending \$50k on either alone. The two channels reinforce each other — they *interact*.

Takeaway

The fix is to add a multiplicative term X_1X_2 to the model.

The interaction model

Two predictors with an interaction

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

How to read this

Differentiate with respect to X_1:

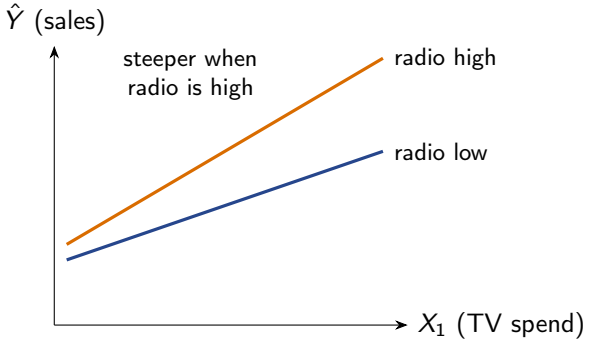
$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2.$$

The slope in X_1 now depends on X_2 — exactly the synergy we wanted to model.

In industry — Retail: price $\times$ promotion

Price elasticity is not one number — shoppers react far more strongly to a cut that is also displayed. Pricing teams model price $\times$ promotion, media teams TV $\times$ digital.

Interaction, drawn: one slope bends with the other predictor



How to read this

The interaction makes the slope of $\hat{Y}$ in X_1 equal to $\hat{\beta}_1 + \hat{\beta}_3 X_2$: different values of X_2 give lines with *different slopes*, so they fan out instead of staying parallel.

Takeaway

Parallel lines = no interaction. A fan = synergy: the predictors amplify each other.

The hierarchy principle

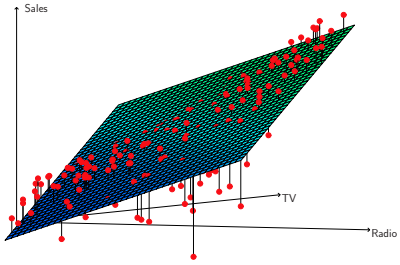
Rule

If an interaction X_1X_2 is in the model, keep the main effects X_1 and X_2 as well, even if their individual t-statistics are small.

Why

Without main effects, the interaction is parameterising deviations from a model in which neither X_1 nor X_2 enters — a strong and usually unintended assumption.

Why Advertising needs an interaction



What the figure shows

The additive model can only fit a *flat plane*: it under-predicts when the budget is split between TV and radio and over-predicts when it is concentrated in one medium. A `TV`×`radio` term lets the surface *bend*, lifting R^2 from about 0.90 to 0.97.

Source: Figure 3.5 — ISLP, James et al. (2023).

Exercise 3.8 — Interactions and hierarchy [Concept]

Exercise

On the Advertising data a model with an interaction gives

$$\widehat{\texttt{sales}} = 6.75 + 0.019\,\texttt{TV} + 0.029\,\texttt{radio} + 0.0011\,(\texttt{TV} \times \texttt{radio}).$$

1. Interpret the interaction coefficient 0.0011 in words.
2. Compute the effective slope of TV when radio $= 0$ and when radio $= 50$.
3. State the *hierarchy principle*. If the TV main effect were not significant but $\texttt{TV} \times \texttt{radio}$ were, should you drop TV?

Solution — Exercise 3.8

Solution

(1) Interaction meaning. The effect of TV on sales is not constant — it *grows* with the radio budget (and vice versa). Concretely, each extra unit of radio raises the per-unit effectiveness of TV by 0.0011. This is a *synergy*: the two media reinforce each other.

(2) Effective TV slope.

$$\frac{\partial \widehat{\texttt{sales}}}{\partial\, \texttt{TV}} = 0.019 + 0.0011 \cdot \texttt{radio}.$$

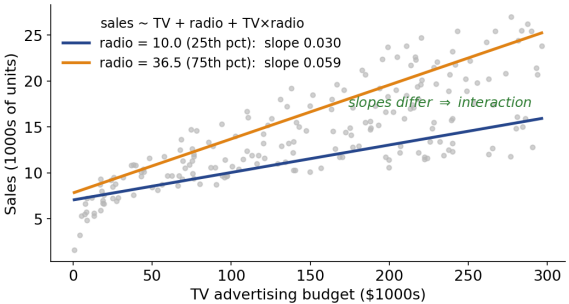
$$\texttt{radio} = 0: \ 0.019, \qquad \texttt{radio} = 50: \ 0.019 + 0.0011 \cdot 50 = 0.019 + 0.055 = 0.074.$$

With a \$50k radio budget, TV is nearly four times as effective per dollar as with no radio.

(3) Hierarchy principle. If an interaction term is included in a model, the corresponding main effects should be kept too, even if their individual p-values are large. So *no* — do not drop TV: the interaction $\texttt{TV} \times \texttt{radio}$ is not interpretable without the TV main effect, and removing it makes the remaining coefficients hard to read and can bias the fit.

Common mistake: quoting 0.019 as "the" TV effect. With an interaction present, that is the slope only at $\texttt{radio} = 0$; the TV effect must be stated at a chosen radio level.

Seeing the interaction: the TV slope depends on radio



Takeaway

On Advertising, an extra \$1k of TV is worth 0.030 thousand units of sales when radio is at its 25th percentile but 0.059 — almost double — at the 75th ($p_{\text{interaction}} \approx 10^{-51}$). Non-parallel fitted lines are the signature of synergy.

Polynomial regression *is* linear regression

Adding X^2, X^3, ... terms keeps the model *linear in the coefficients*, even though the relationship $X \to Y$ is curved.

Quadratic in X, linear in β

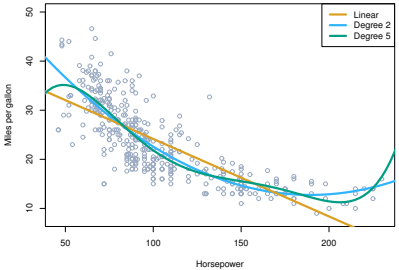
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon.$$

All OLS machinery applies; we simply augment the design matrix.

Takeaway

Polynomial regression is the cheapest way to capture mild curvature. It has known weaknesses (boundary oscillation), but splines (Ch. 7) are the better long-term answer.

Auto: linear vs. quadratic fit



Takeaway

Linear (orange) underfits — a straight line cannot bend. The quadratic (blue) captures the curvature with one extra term. Past degree 4, flexibility adds variance and edge oscillation for little gain.

Source: Figure 3.8 — ISLP, James et al. (2023).

Conceptual summary so far

Where we are in the chapter

We have built the linear model, learnt to estimate it, run inference, handled qualitative predictors and interactions, and explored polynomial extensions.

The next step

A model is only as good as its diagnostics. We now learn to spot the four classical regression problems and what to do about them: nonlinearity, heteroscedasticity, outliers / leverage, collinearity.

Exercise 3.9 — Polynomial fit on Auto [Python]

Exercise

Using the Auto data, model `mpg` against `horsepower`.

1. Fit a linear model and a quadratic (degree-2) model with `statsmodels`.
2. The linear fit gives $R^2 \approx 0.606$; the quadratic gives $R^2 \approx 0.688$ with a highly significant `horsepower`^{2} term. What does this tell you about the relationship?
3. Why is the quadratic model still called a *linear* model?

Solution — Exercise 3.9 (1/2)

Solution

(1) Code.

```
import pandas as pd, numpy as np
import statsmodels.api as sm

auto = pd.read_csv("../ALL CSV FILES - 2nd Edition/Auto.csv",
                   na_values="?").dropna() # "?" -> NaN, then drop
hp = auto["horsepower"].astype(float) # ensure numeric predictor

# Linear model
X1 = sm.add_constant(hp) # intercept + hp
m1 = sm.OLS(auto["mpg"], X1).fit() # fit mpg ~hp

# Quadratic model
X2 = sm.add_constant(np.column_stack([hp, hp**2])) # cols: hp, hp^2
m2 = sm.OLS(auto["mpg"], X2).fit() # fit mpg ~hp + hp^2

print(m1.rsquared, m2.rsquared) # ~0.606 , ~0.688
```

Solution — Exercise 3.9 (2/2)

Solution

(2) Interpretation. The jump from $R^2 \approx 0.606$ to 0.688 and the significant squared term show the relationship is *curved*, not straight: mpg falls steeply as horsepower rises from low values, then flattens. A straight line systematically under- and over-predicts (a U-shaped residual pattern), which the quadratic term corrects.

(3) Still "linear". The model is linear in the *parameters*: $\texttt{mpg} = \beta_0 + \beta_1 \texttt{hp} + \beta_2 \texttt{hp}^2 + \varepsilon$ is linear in $\beta_0, \beta_1, \beta_2$ even though it is nonlinear in the predictor. We simply treat $\texttt{hp}^2$ as an additional predictor, so ordinary least squares still applies.

Common mistake

justifying the quadratic by the R^2 increase alone — R^2 *never* decreases when a term is added. The real evidence is the highly significant t-statistic on $\texttt{hp}^2$ (here $t \approx 10$) together with the disappearance of the U-shape from the residual plot.

Extended Exercise 3.L4 — Dummies and interactions [Integrative]

Extended exercise (15 min)

A store models monthly *spend* y (in \$100s) from a customer's *income* x (in \$1000s) and *region*, a 3-level factor (East, South, West). With East as baseline and dummies D_S, D_W, the model *with interaction* is

$$y = \beta_0 + \beta_1 x + \beta_2 D_S + \beta_3 D_W + \beta_4 (x \cdot D_S) + \beta_5 (x \cdot D_W) + \varepsilon,$$

and the fit gives $(\hat{\beta}_0, \ldots, \hat{\beta}_5) = (30,\ 6,\ -10,\ 20,\ 2,\ -3)$.

1. Write the design-matrix rows for (East, x=40), (South, x=50) and (West, x=40).
2. Give the fitted line *for each region*.
3. Predict spend for an East customer and a West customer, both with x=40.
4. Interpret $\hat{\beta}_4$ and $\hat{\beta}_5$.

Solution — Extended Exercise 3.L4 (1/2)

Solution

(1) Design matrix. Columns $[1,\ x,\ D_S,\ D_W,\ xD_S,\ xD_W]$:

(East, 40)	1	40	0	0	0	0
(South, 50)	1	50	1	0	50	0
(West, 40)	1	40	0	1	0	40

A dummy switches on a level's *intercept* shift; the product column switches on its *slope* shift.
(2) Fitted line per region (substitute the dummies):

$$\text{East: } \hat{y} = 30 + 6x, \qquad \text{South: } \hat{y} = 20 + 8x, \qquad \text{West: } \hat{y} = 50 + 3x.$$

(South intercept $30 - 10$, slope $6 + 2$; West intercept $30 + 20$, slope $6 - 3$.)

Solution — Extended Exercise 3.L4 (2/2)

Solution

(3) Predictions.

East, x=40 : $\hat{y} = 30 + 6(40) = 270$ (\$27,000), West, x=40 : $\hat{y} = 50 + 3(40) = 170$ (\$17,000).

Same income, very different spend — because both the intercept *and* the slope differ by region.
(4) Interpreting the interaction. $\hat{\beta}_4 = 2$: in the South the *effect of income* on spend is 2 units per \$1000 *larger* than in the East (slope 8 vs 6). $\hat{\beta}_5 = -3$: in the West it is 3 units *smaller* (slope 3 vs 6). The interaction lets each region own its *slope*; without it ($\beta_4 = \beta_5 = 0$) all regions would share slope 6 and differ only by a vertical shift. By the *hierarchy principle* we keep the main effects D_S, D_W whenever their interactions are present.

Common mistake

reading $\hat{\beta}_4 = 2$ as the South's slope. It is the *difference* from East: the South's own slope is $6 + 2 = 8$. Dummy interactions, like dummies, are contrasts against the baseline.

Outline

The four classical regression problems

After fitting a regression, the work is not done. Four problems routinely arise and each has a standard diagnostic.

1. **Nonlinearity** — residuals show curvature.
2. **Non-constant variance (heteroscedasticity)** — residual spread depends on the fitted value.
3. **Outliers and high-leverage points** — a few observations dominate the fit.
4. **Collinearity** — two or more predictors carry the same information.

Takeaway

Three of the four announce themselves in a *single* picture — residuals plotted against fitted values — which is exactly why we draw it for every regression. Collinearity is the exception; the VIF catches that one.

Diagnostic 1: nonlinearity

After fitting any linear model, plot the residuals against the fitted values.

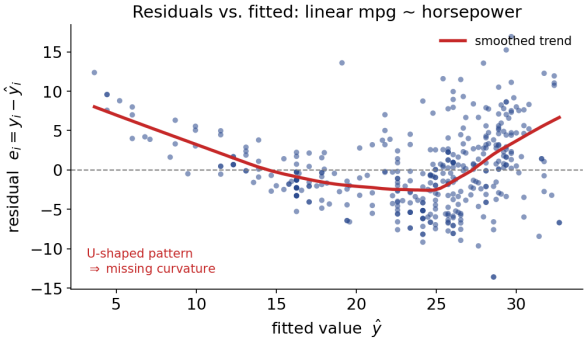
What to look for

A clean linear model has *structureless* residuals, scattered randomly around zero. A clear pattern (curve, U-shape) signals the linear form is wrong.

Reading a U-shape

A U-shape in the residual plot is a classical signature of missing curvature. Remedies: add X^2 (and possibly X^3), use a spline in X, or transform Y.

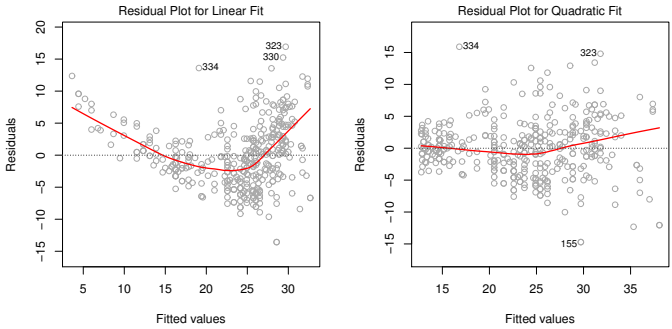
Residuals vs. fitted on the Auto data



Takeaway

A straight-line fit of `mpg` on `horsepower` leaves a clear U in the residuals: the smoothed trend dips then rises. The linear form is missing curvature — add a quadratic term.

Residuals vs. fitted: diagnostic plot



Left: a linear model on `mpg` vs. `horsepower`. Residuals form a clear U. Right: after adding a quadratic term — the pattern is gone.

Source: Figure 3.9 — ISLP, James et al. (2023).

Exercise 3.10 — Reading residual plots [Concept]

Exercise

For each residual-vs-fitted pattern, name the problem and a remedy.

1. Residuals form a clear U-shape (curve) around zero.
2. Residuals fan out into a funnel: small spread at low fitted values, large spread at high fitted values.
3. Residuals are a random band of roughly constant width centred on zero.
4. One point sits far from all others in x-space and pulls the fitted line toward it.

Solution — Exercise 3.10

Solution

(1) Nonlinearity. A systematic curve means the linear functional form is wrong. Remedy: add polynomial terms ($x^2, \ldots$), transform the predictor ($\log x$, $\sqrt{x}$), or use a more flexible model.
(2) Heteroscedasticity. The error variance grows with the mean — non-constant variance. Standard errors and CIs become unreliable. Remedy: transform the response ($\log y$ or $\sqrt{y}$), use weighted least squares, or heteroscedasticity-robust ("sandwich") standard errors.
(3) All good. A structureless horizontal band of constant width is exactly what we want: the linearity, constant-variance and zero-mean assumptions look satisfied. No action needed.
(4) High-leverage point. An extreme x value has unusual influence on the slope. Remedy: verify it is not a data error, report the fit with and without it, and consider a robust regression. (Distinguish leverage — unusual x — from an outlier — unusual y; the dangerous points are high in both.)

Common mistake

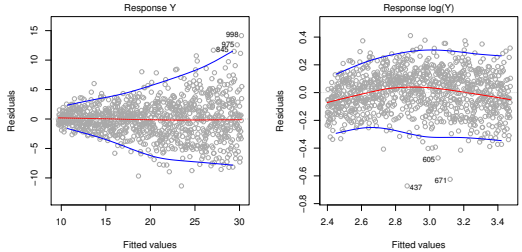
judging these assumptions from the raw y-vs-x scatter or from R^2. A high R^2 can coexist with panels (1), (2) and (4) — only the residual plot shows them cleanly.

Diagnostic 2: non-constant variance

OLS assumes $\mathrm{Var}(\varepsilon \mid X)$ is constant. When it isn't, the standard errors and inference become unreliable.

What to look for

A funnel-shaped residual plot — the spread of residuals grows or shrinks systematically with the fitted value.



Source: Figure 3.11 — ISLP, James et al. (2023).

Three remedies for heteroscedasticity

Fix the model

Transform the response: $\log Y$ for right-skewed positive responses, $\sqrt{Y}$ for count-like data. Fit OLS on the transformed scale.

Reweight the fit

Weighted least squares: give observations with smaller expected variance more weight, recover efficient estimates and valid inference.

Fix the standard errors

Use *heteroscedasticity-consistent* (sandwich / robust) standard errors. The point estimates are unchanged; only the SEs improve.

Outlier vs. high-leverage: not the same thing

These two terms describe *very different* pathologies and need to be diagnosed separately.

Outlier

A point with a large residual — the model fits it badly. May or may not affect the fit much.

High-leverage point

A point with an unusual X value — it can pull the line strongly toward itself.

Worst case

A point that is *both* a high-leverage point *and* an outlier dominates the entire fit. Always flag these.

Detecting outliers and leverage in practice

Two diagnostic statistics, available in every package:

Studentised residuals

Each residual divided by its estimated standard deviation. Studentised residuals greater than 3 in absolute value flag potential outliers.

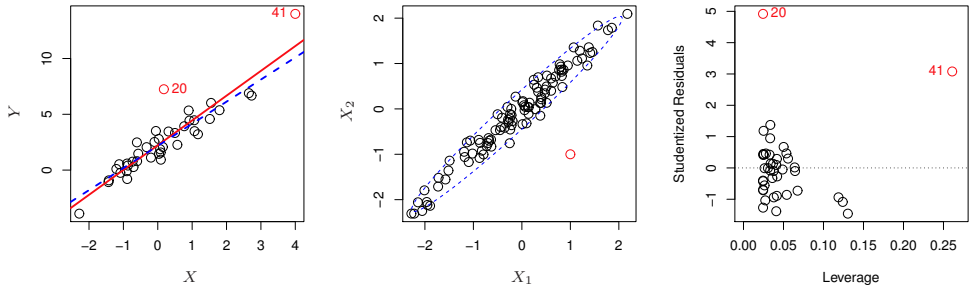
Leverage statistic h_{ii}

The i-th diagonal of the hat matrix, $h_{ii} = (\mathrm{H})_{ii}$. Always $0 \leq h_{ii} \leq 1$ with average $(p+1)/n$. Flag points with $h_{ii} > 2(p+1)/n$ as high-leverage.

Takeaway

Studentised residuals catch unusual y-values; leverage catches unusual x-values. A point that trips *both* rules is the one most likely distorting the fit.

Outliers and leverage: visual examples



Left: point 41 has an extreme x and drags the fit (solid line) away from the fit without it (dashed); point 20 is unusual in y only. Middle: 41 is unremarkable in X_1 or X_2 alone, yet still high-leverage. Right: studentised residual against leverage — 41 stands out on both axes.

Source: Figure 3.13 — ISLP, James et al. (2023).

Exercise 3.11 — Leverage and outliers [Python]

Exercise

Fit `medv` $\sim$ `lstat` on the Boston data.

1. Write code to extract the leverage (hat) values and the studentised residuals.
2. Flag high-leverage points using the rule $h_{ii} > 2(p+1)/n$ and outliers as $|\text{studentised residual}| > 3$.
3. In one sentence each, explain what leverage and a studentised residual measure.

Solution — Exercise 3.11 (1/2)

Solution

(1)–(2) Code.

```
import pandas as pd, numpy as np
import statsmodels.api as sm

boston = pd.read_csv("../ALL CSV FILES - 2nd Edition/Boston.csv")
X = sm.add_constant(boston["lstat"]) # intercept + lstat
res = sm.OLS(boston["medv"], X).fit() # fit medv ~lstat

infl = res.get_influence() # influence diagnostics object
lev = infl.hat_matrix_diag # leverage h_ii
stud = infl.resid_studentized_external # studentized residuals

n, p = X.shape[0], X.shape[1] - 1 # n obs, p predictors
high_lev = np.where(lev > 2*(p+1)/n)[0] # rule: h_ii > 2(p+1)/n
outliers = np.where(np.abs(stud) > 3)[0] # rule: |stud. resid| > 3
print("high leverage:", high_lev) # 34 tracts flagged
print("outliers:", outliers) # 11 tracts, max |stud| = 4.0
```

Solution — Exercise 3.11 (2/2)

Solution

(3) Meaning.

- *Leverage* h_{ii} measures how unusual an observation's *predictor* values are; high-leverage points can strongly influence the fitted line. The average leverage is $(p+1)/n$, hence the $2(p+1)/n$ threshold.
- A *studentised residual* is the residual divided by its estimated standard deviation (with that point left out); it flags observations with unusual *response* values (outliers) on a comparable scale. Points that are both high leverage *and* outliers are the most influential.

What the run shows (Boston). With $n = 506$, $p = 1$ the leverage cut-off is $2(p+1)/n = 4/506 \approx 0.008$, and 34 tracts exceed it — exactly the largest-`lstat` neighbourhoods. Eleven studentised residuals exceed $+3$ (max ≈ 4.0): all are expensive tracts, mostly at the censored top value `medv` $= 50$, which the line under-predicts. The two sets do not overlap here — leverage and outlyingness really are different things.

Common mistake

treating the thresholds as deletion rules. Flagged points call for *inspection* (data error? capped response?) and a with/without sensitivity refit — not automatic removal.

Diagnostic 4: collinearity

Definition

Two or more predictors are highly correlated, so that one is well approximated by a linear combination of the others.

Why it matters

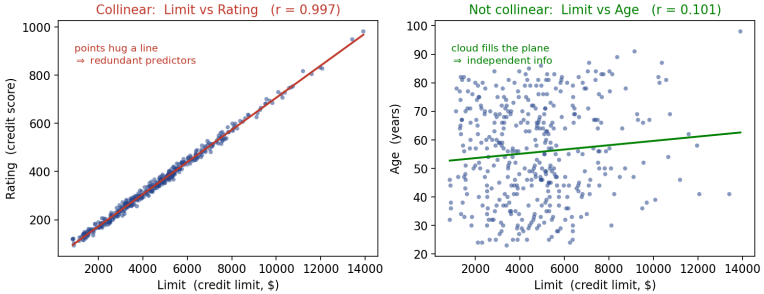
The estimates $\hat{\beta}_j$ become unstable. Inflated standard errors, unstable signs, and counterintuitive coefficients are all classical symptoms.

Curiously, predictions are usually fine; only the inferential interpretation of individual coefficients is corrupted.

In industry — Marketing mix modelling: channels move together

Campaign weeks raise TV, digital and in-store spend at once, so channel coefficients are barely identified: one turns negative, the next refresh flips it back. Check VIFs before anyone reallocates budget on them.

Collinearity, seen: two predictors that move together



Takeaway

When two predictors nearly fall on a line (left, $r = 0.997$), the data cannot separate their individual effects, so each coefficient swings wildly from sample to sample. Uncorrelated predictors (right) carry independent information and cause no such trouble.

Variance inflation factor (VIF)

Quantify collinearity for predictor X_j via:

VIF

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ comes from regressing X_j on all the *other* predictors. If they already explain X_j almost perfectly, $R^2_{X_j|X_{-j}} \to 1$, the denominator $\to 0$, and $\mathrm{VIF} \to \infty$.

How to read this

- $\mathrm{VIF} = 1$: no collinearity.
- $\mathrm{VIF} > 5$ (some authors use 10): problematic.
- Remedies: drop a predictor, combine into an index, or use ridge regression (Ch. 6) which is robust to collinearity.

Exercise 3.12 — Collinearity and VIF [Python]

Exercise

On the Credit data, `limit` and `rating` are almost perfectly correlated.

1. Write code to compute the variance inflation factor (VIF) for the predictors `limit`, `rating` and `age` in a regression of `balance`.
2. You obtain roughly $\mathrm{VIF}(\texttt{limit}) \approx 160$, $\mathrm{VIF}(\texttt{rating}) \approx 161$, $\mathrm{VIF}(\texttt{age}) \approx 1.0$. What does VIF measure, and what is the consequence of these large values?
3. Suggest two fixes.

Solution — Exercise 3.12 (1/2)

Solution

(1) Code.

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence \
    import variance_inflation_factor as vif_

credit = pd.read_csv("../ALL CSV FILES - 2nd Edition/Credit.csv")
cols = ["Limit", "Rating", "Age"] # CSV column names are capitalised
X = sm.add_constant(credit[cols]) # design matrix + intercept column
# VIF_j = 1/(1 - R2 of X_j regressed on the other columns)
vifs = {col: vif_(X.values, i) for i, col in enumerate(X.columns)}
print(vifs) # Limit ~160, Rating ~161, Age ~1.0
```

Solution — Exercise 3.12 (2/2)

Solution

(2) What VIF measures.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is from regressing predictor j on all the others. It is the factor by which the *variance* of $\hat{\beta}_j$ is inflated because of collinearity. $\mathrm{VIF} = 1$ means no collinearity; values above $\sim$ 5–10 are concerning. Here $\mathrm{VIF} \approx 160$ means the standard errors of the `limit` and `rating` coefficients are about $\sqrt{160} \approx 13$ times larger than they would be if the two were uncorrelated. Consequence: hugely inflated standard errors, unstable and possibly wrong-signed coefficients, and low power to detect either effect — even though *jointly* they predict balance well.

(3) Fixes. Drop one of the redundant predictors (keep `limit` *or* `rating`); or combine them into a single index (an average or a principal component); or use a penalised method such as ridge regression, which tolerates collinearity.

Common mistake: concluding the model *predicts* poorly. Collinearity corrupts the attribution to *individual* coefficients; the joint fit and its predictions are typically fine.

Extended Exercise 3.L5 — A diagnostics case study [Integrative]

Extended exercise (15 min)

You regress house `price` on `size`, `bedrooms`, age and `income` ($n = 400$) and inspect the standard diagnostic plots:

- **Residuals vs fitted:** a clear $\cup$-shaped curve, and the spread of residuals *widens* from left to right.
- **Scale–location:** $\sqrt{|\text{standardised resid}|}$ trends steadily upward.
- **Residuals vs leverage:** two points with $|\text{studentised resid}| > 3$; one point on the far right with leverage $h_{ii} = 0.30$ (average $\bar{h} = 5/400 = 0.0125$) and a large Cook's distance.
- **VIF:** size 8.9, bedrooms 7.5, age 1.3, income 9.2.

Diagnose *all four* classical problems and recommend a concrete fix for each.

Solution — Extended Exercise 3.L5 (1/2)

Solution

Nonlinearity. A ∪-shaped residuals-vs-fitted plot is the signature of a *misspecified mean* — a straight model on a curved relationship. *Fix:* add polynomial terms (e.g. $\texttt{size}^2$) or transform a predictor ($\log \texttt{size}$); consider splines. Re-check for a flat, patternless band.

Heteroscedasticity. Widening residuals together with an upward scale–location trend indicate non-constant error variance, so the usual SEs and the t-/F-tests are unreliable. *Fix:* transform the response ($\log$ or $\sqrt{\ }$ of `price`), use weighted least squares, or report heteroscedasticity-robust (White / HC) standard errors.

Solution — Extended Exercise 3.L5 (2/2)

Solution

Outliers. The two points with |studentised resid| > 3 are badly fit in the y-direction; they inflate RSE and can distort tests. *Fix:* check for data-entry errors; report the fit with and without them; consider robust regression — do *not* delete solely because a point is large.
High leverage / influence. The point with $h_{ii} = 0.30$ ($24\times$ the average 0.0125) plus a large Cook's distance is unusual in x *and* actually moves the fit. *Fix:* inspect it and test sensitivity by refitting without it.
Collinearity. VIFs of 8.9, 7.5, 9.2 (all $\gg 5$) show `size`, `bedrooms` and `income` carry overlapping information, inflating their SEs. *Fix:* drop or combine a redundant predictor (e.g. keep size, or use price-per-size), or use ridge / PCA regression; recompute VIFs until all < 5.

Common mistake

conflating the diagnoses — an outlier is unusual in y, a leverage point in x. The dangerous case is both at once, and only Cook's distance measures that combination.

Outline

Linear regression vs. KNN regression

KNN regression is the simplest nonparametric alternative: predict Y at x_0 as the average of the K nearest neighbours' responses.

Linear regression

- Assumes a global functional form.
- Few parameters $\Rightarrow$ low variance.
- Interpretable coefficients.

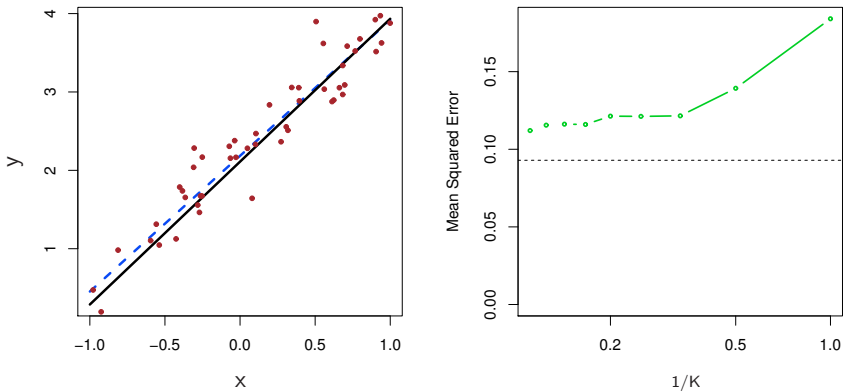
KNN regression

- No functional form.
- Many "effective" parameters $\Rightarrow$ low bias if f is nonlinear.
- Hard to interpret.

Takeaway

Rule of thumb: with few observations, many predictors, or a roughly linear truth, prefer regression. KNN wins only when the truth is strongly nonlinear *and* n is large enough to fill the space.

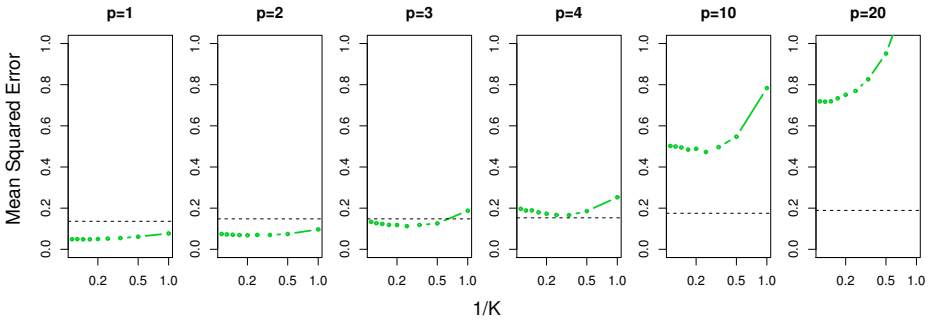
Linear vs. KNN: simulation



The truth here *is* linear (left), so least squares (blue dashed) nails it. Right: KNN's test MSE (green) against $1/K$ stays above the least-squares MSE (dashed).

Source: Figure 3.18 — ISLP, James et al. (2023).

The curse of dimensionality strikes KNN



As p grows, KNN performance deteriorates rapidly: in high dimensions there are no truly "near" neighbours. Linear methods are more robust because they impose structure.

Source: Figure 3.20 — ISLP, James et al. (2023).

Outline

Python lab — run it live in the notebook

Companion notebook: `chapter_03_lab.ipynb`

The code in this section is meant to be **demonstrated live**. Open the companion Jupyter notebook and run it cell by cell:

- §1 *Simple linear regression* and §2 *Multiple linear regression*, with `statsmodels`;
- §3 *Polynomials and interactions via the ISLP* `ModelSpec`;
- §4 *Diagnostics*, then §5 *Comparison with KNN*.

Data loads via the `ISLP` package or the bundled CSV files; the notebook ends with hands-on exercises.

Simple regression with statsmodels

```
import pandas as pd, numpy as np
import statsmodels.api as sm
from ISLP import load_data # textbook data sets

Boston = load_data("Boston") # Boston housing data
X = sm.add_constant(Boston["lstat"]) # predictor + intercept column
y = Boston["medv"] # response: median home value
res = sm.OLS(y, X).fit() # least-squares fit medv ~lstat
print(res.summary()) # coef, SE, t, p, CI, R2, F
```

`statsmodels` produces a comprehensive output table with coefficients, standard errors, t-statistics, p-values, and confidence intervals — everything we need for inference.

Confidence and prediction intervals

```
xnew = pd.DataFrame({"const": 1.0, "lstat": [5, 10, 15]}) # new x
pred = res.get_prediction(xnew) # predictions with uncertainty
pred.summary_frame(alpha=0.05) # 95% intervals of both kinds
# columns: mean, mean_se,
# mean_ci_lower / upper, <- CI
# obs_ci_lower / upper <- prediction interval
```

Both intervals come from the same fit but answer different questions. Always use *prediction* intervals when communicating uncertainty about an individual outcome.

Multiple regression

```
predictors = ["lstat", "age", "rm", "ptratio", "tax"] # five inputs
X = sm.add_constant(Boston[predictors]) # design matrix + intercept
res = sm.OLS(Boston["medv"], X).fit() # fit multiple regression
print(res.summary()) # one t-test per coefficient
```

Adding more predictors typically increases R^2 and shrinks the per-coefficient standard errors — as long as the new variables are not collinear.

Polynomials and interactions via ISLP

```
from ISLP.models import ModelSpec as MS, poly

terms = [poly("lstat", degree=2), "age", # lstat + lstat^2, age
         ("lstat", "age")] # tuple = interaction term
design = MS(terms).fit_transform(Boston) # build the design matrix
res2 = sm.OLS(Boston["medv"], design).fit() # then fit as usual
print(res2.summary())
```

`ModelSpec` produces a `statsmodels`-ready design matrix and handles dummies, interactions, polynomials, and splines cleanly.

Diagnostic plots

```
import matplotlib.pyplot as plt
fig, axes = plt.subplots(1, 2, figsize=(10, 4)) # two panels
axes[0].scatter(res.fittedvalues, res.resid, s=8) # resid vs fitted
axes[0].axhline(0, color="grey", linewidth=0.8) # zero line
axes[0].set_xlabel("Fitted"); axes[0].set_ylabel("Residual")
sm.qqplot(res.resid, line="45", ax=axes[1]) # normality check
plt.tight_layout(); plt.show() # render figure
```

Two diagnostic plots that should be made for every regression: the residual-vs.-fitted plot and the Q-Q plot.

Leverage and VIF

```
from statsmodels.stats.outliers_influence \
    import variance_inflation_factor as vif_

infl = res.get_influence() # influence diagnostics object
lev = infl.hat_matrix_diag # leverage values h_ii
high_lev = np.where(lev > 2*X.shape[1]/X.shape[0])[0] # h > 2(p+1)/n

vifs = [vif_(X.values, j) for j in range(X.shape[1])] # VIF per column
print(dict(zip(X.columns, vifs))) # VIF near 1 fine; > 5 problematic
```

Always inspect leverage and VIF before declaring an analysis complete.

Outline

Chapter 3 in one slide

What we accomplished

Built up linear regression from one predictor to many, with full inferential machinery and diagnostics.

- Estimated β by least squares; derived standard errors, t- and F-tests.
- Saw how multiple regression handles confounding.
- Encoded qualitative predictors and added interaction terms.
- Diagnosed nonlinearity, heteroscedasticity, outliers, leverage, and collinearity.
- Compared linear regression with KNN regression.

Vocabulary checklist

Term	Meaning
SLR / MLR	Simple / multiple linear regression
OLS	Ordinary least squares
RSS / TSS / RSE	Residual / total sum of squares; $\sqrt{\mathrm{RSS}/(n-p-1)}$
R^2 / adj. R^2	Variance explained / penalised version
t-test / F-test	One β_j / all β_j at once
Confidence vs. prediction interval	On the line / on a future Y
Hat matrix H	$\mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top$
Leverage h_{ii}	Diagonal of H; flag if $> 2(p+1)/n$
Studentised residual	Residual divided by its estimated SD
VIF	$1/(1-R^2_{X_j \mid X_{-j}})$
Interaction	Effect of X_1 depends on X_2

Key formulas at a glance

OLS estimates

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Tests

$$t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p-1}, \qquad F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,\, n-p-1}.$$

Fit and intervals (simple regression; $S_{xx} = \sum (x_i - \bar{x})^2$, $\mathrm{TSS} = \sum (y_i - \bar{y})^2$, $t = t_{0.025,\, n-2}$)

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\mathrm{RSE}}{\sqrt{S_{xx}}}, \qquad \mathrm{SE}(\hat{\beta}_0) = \mathrm{RSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \qquad \mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}}, \qquad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\underbrace{\hat{y}_0 \pm t\, \mathrm{RSE}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}_{\text{CI for the mean response } E[Y \mid X = x_0]} \qquad \underbrace{\hat{y}_0 \pm t\, \mathrm{RSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}_{\text{PI for a new observation } Y_0 \text{ — note the “1+”}}$$

Decision rules of thumb

Always

Plot residuals vs. fitted values. Plot residuals vs. each predictor. Compute VIFs. Examine high-leverage observations.

Rules of thumb

- $|t_j| > 2$ or $p < 0.05 \Rightarrow$ "statistically significant" — but always also check effect size.
- $\mathrm{VIF} > 5$ (some say 10) $\Rightarrow$ collinearity worth addressing.
- Funnel-shape residuals $\Rightarrow$ try a log transformation or robust SEs.
- Curved residual mean $\Rightarrow$ add polynomial / spline.

Common pitfalls

Don't

1. Interpret a univariate slope as a causal effect.
2. Trust p-values when residuals are obviously non-normal or heteroscedastic.
3. Add an interaction without keeping the main effects (hierarchy principle).
4. Use stepwise selection with naive p-values — the standard errors are wrong.
5. Report only R^2 on training data without out-of-sample checks.
6. Forget to check leverage — a few points can dominate the entire fit.

Self-check questions

1. Derive the OLS slope $\hat{\beta}_1$ from the normal equations. (pp. 25–26; in full in the appendix, pp. 148–150)
2. Why does adding any predictor never decrease R^2? What is the fix? (p. 69)
3. Suppose two predictors are perfectly correlated. What goes wrong with OLS? With ridge? (pp. 115–117)
4. Interpret an interaction coefficient $\beta_3 X_1 X_2 = 0.4$ when X_1 = TV spend in \$1k and X_2 = radio spend in \$1k. (pp. 84–85)
5. In the Boston data, the simple regression of `medv` on `age` is significantly negative, but in the multiple regression it is not. Explain. (pp. 61–62)
6. State the prediction interval for a new observation at x_0 and contrast with the confidence interval. (p. 70)

Ten things to remember

1. Linear regression: linear in the coefficients, not in X.
2. Least squares minimises the squared residuals.
3. Standard errors $\rightarrow$ confidence intervals and t-tests.
4. F-test answers "is any predictor useful?".
5. Each β_j holds the others fixed (interpretation rule).
6. Univariate slopes can mislead — multiple regression matters.
7. Encode categories as dummies; consider interactions.
8. Diagnose: nonlinearity, heteroscedasticity, outliers, leverage, collinearity.
9. Confidence vs. prediction intervals: line vs. point.
10. Linear regression is hard to beat on small/medium tabular data.

What's next

Chapter 4: *Classification*.

- Logistic regression as the linear model for log-odds.
- Discriminant analysis (LDA, QDA, naive Bayes).
- Confusion matrices, ROC curves, AUC.
- Generalised linear models in one frame.

Appendix — what is in here, and why

Nothing in the appendix is needed to follow the chapter: each slide is a formal derivation, a longer worked exercise, or a side topic. Read it when you want the full story, or when a later chapter sends you back here.

Contents of the appendix

- **Squared versus absolute loss** — the trade-offs of the squared criterion and the robust alternatives. Interesting, but this course always uses squared loss.
- **Deriving least squares (Extended Exercise 3.L2)** — the full calculus derivation of $\hat{\beta}_0$ and $\hat{\beta}_1$. The main deck derives the two normal equations already; this is the same argument done slowly, as homework.
- **Matrix form of multiple regression** — $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. You never compute it by hand; it reappears in Chapter 6 for ridge.
- **Linear vs. polynomial vs. KNN (Extended Exercise 3.L6)** — a full simulation study in Python. Excellent practice, and a natural bridge to Chapter 5 — but it needs a machine, not a lecture slot.

Why squared, and not absolute?

Pros of squaring

- Differentiable everywhere $\Rightarrow$ closed form.
- Larger mistakes punished more $\Rightarrow$ no huge residuals at the optimum.
- Aligns with the Gaussian-error likelihood (Gauss 1809).

Cons of squaring

- Not robust to outliers — one extreme y_i can move the line a lot.
- Robust alternatives (median regression, Huber loss) exist and are useful when outliers dominate.

Takeaway

For a first course we always start with squared loss. Switching to robust losses is a one-line change in modern software.

Extended Exercise 3.L2 — Deriving least squares [Math]

Extended exercise (15 min)

Do not merely *use* the OLS formulas — *derive* them.

1. Write $\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ and set both partial derivatives to zero to obtain the *normal equations*.
2. Solve them for $\hat{\beta}_0$ and $\hat{\beta}_1$.
3. Show the fitted line always passes through $(\bar{x}, \bar{y})$.
4. State precisely what it means that $\hat{\beta}_1$ is *unbiased*.
5. Plug in the summaries $n = 20$, $\bar{x} = 5$, $\bar{y} = 15$, $S_{xx} = 40$, $S_{xy} = 100$ to get the estimates.

Solution — Extended Exercise 3.L2 (1/3)

Solution

Method. $\mathrm{RSS}(\beta_0, \beta_1)$ is a smooth convex function of the two coefficients (a bowl), so its unique minimiser is the point where *both* partial derivatives vanish — two linear equations in two unknowns.

(1) Normal equations. Differentiate the sum of squares and set each derivative to zero:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2\sum(y_i - \beta_0 - \beta_1 x_i) = 0 \;\Rightarrow\; \sum y_i = n\beta_0 + \beta_1 \sum x_i,$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \;\Rightarrow\; \sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2.$$

The factors -2 cancel. Read as statements about residuals: the first says the residuals sum to zero, the second says they are orthogonal to the predictor.

Solution — Extended Exercise 3.L2 (2/3)

Solution

(2) Solve the system. Divide the first normal equation by n:

$$\bar{y} = \beta_0 + \beta_1 \bar{x} \;\Rightarrow\; \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Substitute this into the second equation (note $\sum x_i = n\bar{x}$) and collect the $\hat{\beta}_1$ terms:

$$\sum x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x})\, n\bar{x} + \hat{\beta}_1 \sum x_i^2 \;\Rightarrow\; \sum x_i y_i - n\bar{x}\bar{y} = \hat{\beta}_1 \Big(\sum x_i^2 - n\bar{x}^2 \Big).$$

The centring identities $\sum x_i y_i - n\bar{x}\bar{y} = \sum (x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$ and $\sum x_i^2 - n\bar{x}^2 = \sum (x_i - \bar{x})^2 = S_{xx}$ then give

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

Common mistake: trying to solve the second equation on its own — the uncentred sums collapse to S_{xy}/S_{xx} only *after* substituting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Solution — Extended Exercise 3.L2 (3/3)

Solution

(3) Passes through $(\bar{x}, \bar{y})$. At $x = \bar{x}$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}.$$

This is just the first normal equation restated — the residuals always sum to zero.
(4) Unbiasedness. Treating the x_i as fixed and assuming $\mathbb{E}[\varepsilon_i] = 0$, one shows $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\mathbb{E}[\hat{\beta}_0] = \beta_0$. Meaning: *averaged over many repeated samples* the estimates are centred on the true values — no systematic over- or under-estimation. (Any single estimate still has sampling error, quantified by its SE.)
(5) Plug in.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{100}{40} = 2.5, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 15 - 2.5(5) = 2.5,$$

so $\hat{y} = 2.5 + 2.5\,x$.

Matrix form: a compact view

Stack predictors into a design matrix X that includes a column of ones (for the intercept):

$$\mathsf{y} \;=\; \mathsf{X}\boldsymbol{\beta} \;+\; \boldsymbol{\varepsilon}, \qquad \mathsf{X} \in \mathbb{R}^{n \times (p+1)}.$$

Least squares gives the closed form

$$\boxed{\hat{\boldsymbol{\beta}} \;=\; (\mathsf{X}^{\top}\mathsf{X})^{-1}\mathsf{X}^{\top}\mathsf{y}.}$$

How to read this

- $\mathsf{X}^{\top}\mathsf{X}$ must be invertible: requires $n > p + 1$ *and* no perfect collinearity.
- The hat matrix $\mathsf{H} = \mathsf{X}(\mathsf{X}^{\top}\mathsf{X})^{-1}\mathsf{X}^{\top}$ satisfies $\hat{\mathsf{y}} = \mathsf{H}\mathsf{y}$.
- Diagonal entries h_{ii} are leverages — we will use them in diagnostics.

Extended Exercise 3.L6 — Linear vs polynomial vs KNN [Python]

Extended exercise (15 min)

On `Auto.csv`, model `mpg` $\sim$ `horsepower` (drop the rows where `horsepower` is the string `"?"`).

1. Make a 70/30 split (`random_state=1`) and fit three models: linear, degree-2 polynomial, and KNN regression (standardise x first).
2. Estimate the *test* MSE of each.
3. Which fits best, and why? Answer in terms of the *bias–variance trade-off*, and say what happens to KNN as $K \to 1$.

Solution — Extended Exercise 3.L6 (1/3)

Solution

Load, clean and split.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error as mse # test-error metric

a = pd.read_csv("../ALL CSV FILES - 2nd Edition/Auto.csv") # load Auto data
a = a[a.horsepower != "?"]; a.horsepower = a.horsepower.astype(float) # clean
X, y = a[["horsepower"]].values, a["mpg"].values # one predictor, target mpg
Xtr, Xte, ytr, yte = train_test_split(X, y, test_size=.3, random_state=1) # 70/30
```

Solution — Extended Exercise 3.L6 (2/3)

Solution

Fit the three models and score the test MSE.

```
lin = LinearRegression().fit(Xtr, ytr) # straight-line fit
pol = make_pipeline(PolynomialFeatures(2), LinearRegression()).fit(Xtr, ytr) # adds hp^2
knn = make_pipeline(StandardScaler(), KNeighborsRegressor(10)).fit(Xtr, ytr) # scale, K=10
for name, m in [("linear", lin), ("poly-2", pol), ("KNN10", knn)]:
    print(name, round(mse(yte, m.predict(Xte)), 1)) # test-set MSE
# linear 26.2 | poly-2 19.8 | KNN(K=10) 21.1
```

Solution — Extended Exercise 3.L6 (3/3)

Solution

Test MSE. Linear ≈ 26.2, degree-2 polynomial ≈ 19.8, KNN (K=10) ≈ 21.1: both flexible models clearly beat the straight line.

Why (bias–variance). The true `mpg`–`horsepower` relationship is convex and decreasing, so a straight line *underfits* — high *bias*, largest test MSE. The degree-2 polynomial captures the curvature with one extra parameter; a moderate-K KNN is an equally strong non-parametric competitor.

As $K \to 1$. Each prediction is a single neighbour: the model interpolates the training points (training MSE $\to 0$) but is very wiggly — high *variance*, so test MSE *rises* (here ≈ 38 at K=1); the best K balances the two. *Takeaway:* this one split cannot rank the two flexible fits — 19.8 against 19.3 at K=25 is noise, and the ranking flips with the seed. Scored by 10-fold CV on all 392 cars (see the lab notebook), KNN at $K \approx 25$–50 is marginally *ahead* (18.2–18.6 vs 19.2) — but inside one standard error. Prefer the degree-2 fit for *interpretability*, not for accuracy.

Common mistake

choosing K by training MSE — it falls monotonically to zero at $K = 1$. Only test or CV error reveals the U-shape that picks the right K.