Part VI — Regression: Finding the Invisible String

The string nobody can see, but everyone can feel

A cricket coach has years of data on his bowlers: hours of practice per week, and average bowling speed. Plotted as a scatter plot, the dots trend upward — more practice, generally, though not perfectly, goes with higher speed. He wants more than "they seem related." He wants to hand a new trainee a number of practice hours and get back a genuine prediction of what speed to expect.

Plotted, the dots don't sit on a perfect line — no real relationship ever does — but they cluster into a rough diagonal band, as though a string runs through the cloud, taut between the bottom-left and the top-right, with every dot tugging gently at it and settling near, but rarely exactly on, wherever it lands. The question this chapter answers precisely: among every possible straight line you could draw through this cloud, which one is actually the best one? "Looks about right" is not a method. We need something a machine could check without ever looking at the picture.

Here's the trick, and it's a familiar one applied at a new scale. Recall how we measured "how much chaos surrounds a number" back when we built the variance: for every point, measure its distance from a candidate center, square that distance so far misses count more than near misses and so negatives don't cancel positives, then add everything up. A regression line does the identical thing, except the "candidate center" is no longer a single point. It's an entire line, and every dot gets measured by how far it sits from that line, straight up or down.

For any candidate line, and for each individual data point, there's a gap between what the line predicts for that bowler's practice hours, and what that bowler's speed actually was. This gap is called a residual — positive if the real speed beat the line's prediction, negative if it fell short. Square every residual, add them all up, and you get a single number that measures, for that one candidate line, how badly it's failing the whole cloud of dots at once. Now imagine trying every conceivable line, every possible slope, every possible starting height, and asking, for each one, "what's your total squared-residual score?" Somewhere in that infinite haystack of candidate lines is exactly one with the lowest possible score. That line is the least squares line, the line of best fit — and here is the small miracle worth pausing on: it is not a matter of opinion. Given the same dots, everyone who runs this exact procedure lands on the exact same line, every time. That's the whole reason regression is trustworthy rather than merely persuasive. See Figure 6.1 for the least-squares line and its residuals.

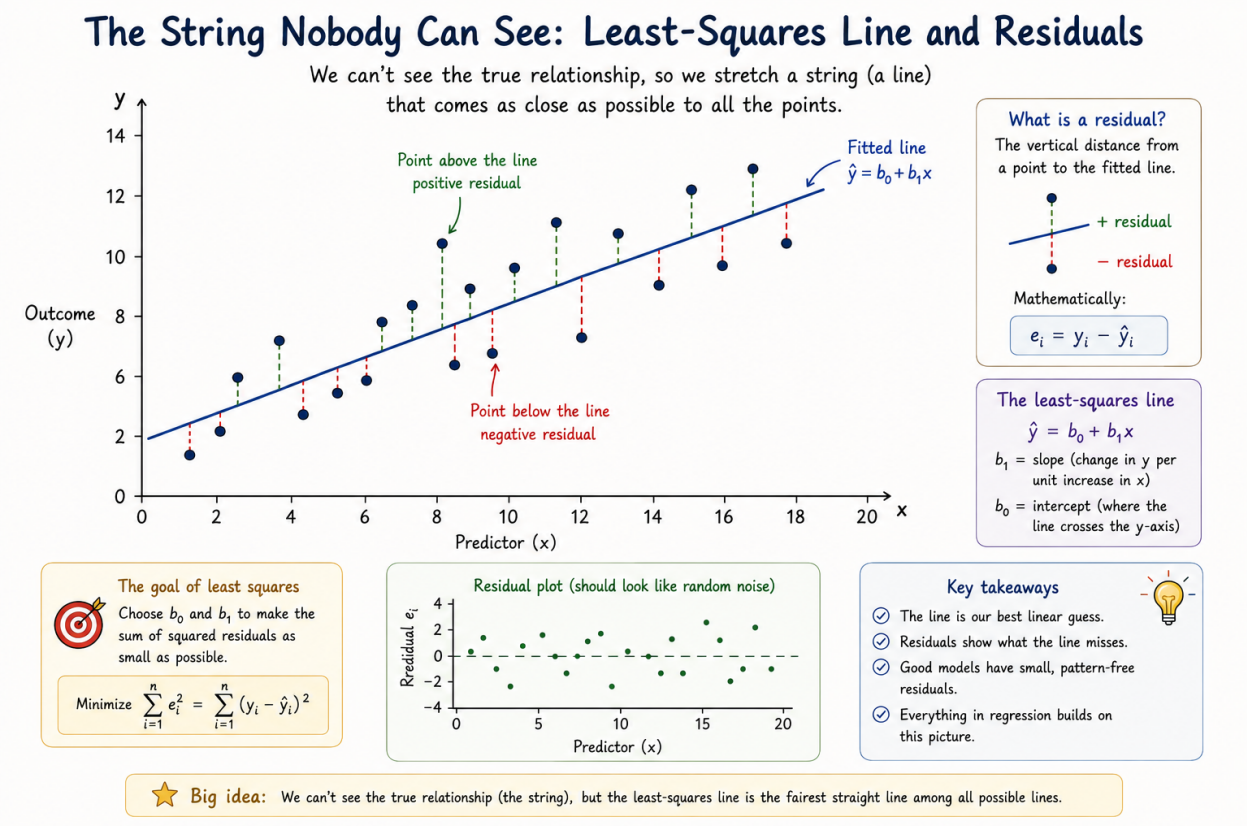

Figure 6.1 — The least-squares line and residuals: the string nobody can see.

Every straight line is fully described by just two numbers. The slope tells you how steeply it rises: for the coach's data, something like speed rising by 0.8 for every added hour of practice, plus a starting value of 118. The 0.8 means each additional hour of weekly practice nudges predicted speed up by 0.8 km/h, on average, across this dataset. The intercept, 118, is the line's starting height, the predicted speed at zero practice hours, a useful mathematical anchor even though nobody actually becomes a competitive bowler with zero practice. Two numbers, and the entire relationship between hours and speed collapses into something you could text a colleague.

THE MATH

Why minimizing squared residuals produces one uniquely best line: squaring guarantees a single smooth bowl-shaped error surface with exactly one lowest point, which is why everyone who runs this procedure lands on the same line.

$$\hat{y} = b_0 + b_1 x, \qquad b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad b_0 = \bar{y} - b_1 \bar{x}$$

> *Worked example:* With speed = $118 + 0.8 \times$ hours: a bowler practicing 10 hours per week is predicted at $118 + 0.8(10) = 126$ km/h; a bowler practicing 20 hours is predicted at $118 + 0.8(20) = 134$ km/h.
>
> **Meaning:** The slope, 0.8, is the whole predictive claim in one number: each added hour of weekly practice nudges predicted speed up by 0.8 km/h on average. The intercept, 118, only anchors the line.

▸ *Interactive version of this math box*

The warning label everyone skips

Here is a warning that has been building since the very first story, and it earns its full weight now that we finally have a concrete tool sharp enough to cut ourselves on. A regression line describes association: how two variables move together in the data actually collected. It does not, by itself, tell you that one causes the other.

Think about it like this: imagine you're a child again, and you notice that every time the ice cream truck's jingle plays, more people are wearing shorts. Does the jingle cause shorts? Obviously not — it's summer, doing both things at once, invisible in the data unless you go looking for it. The regression line connecting "jingle frequency" and "shorts-wearing" would look every bit as clean and confident as the coach's practice-hours line.

The mathematics cannot tell the difference between a real engine of cause and effect and two dots being pulled by the same hidden hand. Perhaps naturally gifted bowlers, who'd throw fast for reasons that have nothing to do with practice, are also simply more motivated, and so it's talent driving both practice and speed, with the arrow of causation reversed from what the line seems to suggest. Perhaps access to quality coaching drives both variables independently.

A strong regression relationship tells you two variables move together reliably. It never, by itself, tells you why. Establishing genuine causation needs the experimental machinery already built (randomized assignment, control groups) or the sharper causal-inference tools waiting in Part XIII. Confusing a strong fit with proof of causation is one of the most common, most consequential errors made across business, medicine, economics, and public policy — not because people are careless, but because a clean line is so visually persuasive that it feels like it must mean more than it does.

THE MATH

Why R-squared is fraction of wrongness erased: it's built from two error tallies, total squared error using just the average as your guess, versus total squared error using the fitted line. R-squared is how much of the first the second manages to remove.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Worked example: If guessing the average speed for every bowler produces total squared error of 1,000, and the fitted line's residuals total only 300, $R^2 = 1 - 300/1{,}000 = 0.7$.

Meaning: R-squared of 0.7 doesn't mean 70 percent accurate; it means the line erased 70 percent of the guessing-error that using no information at all would have produced, leaving 30 percent still unexplained.

▸ *Interactive version of this math box*

How good is "good," really?

A least squares line always exists for any scatter of points — even points with no real relationship at all will produce some "best" line, since the method always finds whichever line minimizes squared error, however poor that minimum still is. So we need a second number, separate from the line itself, that tells us how much to trust it.

This is R-squared, and here's the plainest way to feel what it means: imagine trying to predict each bowler's speed knowing nothing about their practice hours. Your best guess for everyone would just be the overall average speed, and you'd be wrong by some amount for almost every bowler. Now imagine predicting each bowler's speed using the regression line. R-squared is the fraction of your original wrongness that the line manages to erase. An R-squared of 0.7 means the line explains roughly 70% of why bowling speeds differ from one bowler to the next; the remaining 30% is everything else — talent, technique, conditioning — that this simple two-variable model never saw. See Figure 6.2 for the r-to-R-squared bridge.

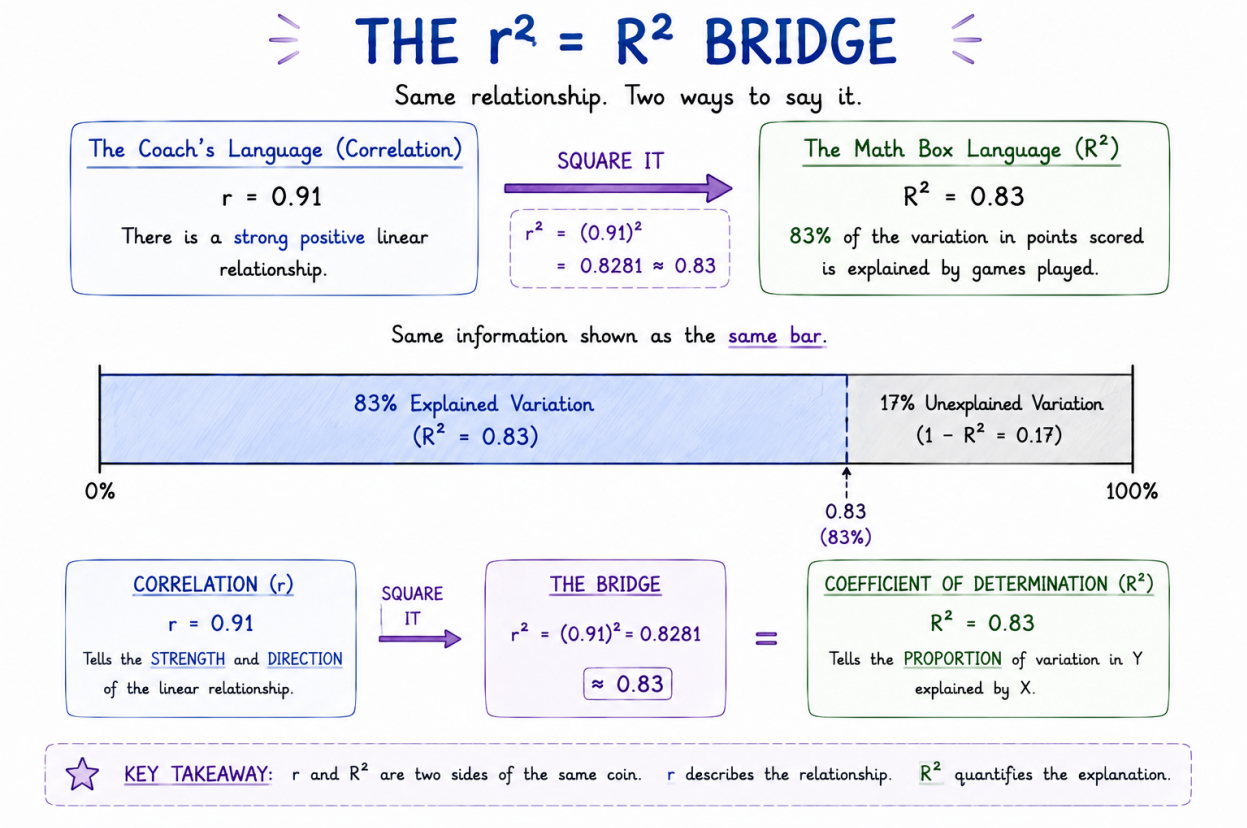

Figure 6.2 — The $r^2 = R^2$ bridge: correlation and coefficient of determination are the same story.

This exact tool underlies an enormous swath of applied work: economists studying education against income, real-estate models turning square footage into price estimates, dose-response curves in medicine, sports analytics predicting performance from measurable inputs, always, if the practitioner is being careful, with the causation warning held firmly in mind.

THE MATH

Why log-odds specifically: odds stretch a bounded 0-to-1 probability into an unbounded 0-to-infinity scale, but taking the log finishes the job, stretching all the way to negative infinity too, the unrestricted range an ordinary straight line needs.

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \;\Rightarrow\; p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

Worked example: With an odds ratio of 1.15 for age, if a 50-year-old's odds are 0.20 (p about 0.167), a 51-year-old's odds become $0.20 \times 1.15 = 0.23$, converting back to $p = 0.23/1.23 \approx 0.187$, a smaller probability jump than the odds multiplier alone might suggest.

> **Meaning:** Fit the easy linear thing on the log-odds scale where it's mathematically safe, then walk the answer back through the doorway to land on a genuine, always-valid probability.

▸ *Interactive version of this math box*

Where the word "regression" actually came from

Here's something genuinely strange worth a moment's detour: why is this entire technique called "regression" at all? The word, in ordinary English, means moving backward or reverting to an earlier state, an odd name for a tool whose entire purpose is fitting a forward-looking predictive line through a scatter of dots. The answer is a specific, curious discovery made by Francis Galton — the same Galton whose bean machine built the normal distribution back in Part III — while he was studying something that, at first glance, seems to have nothing to do with lines through scatter plots at all: the heights of parents and their adult children.

Galton collected height data on many families and noticed something that struck him, initially, as almost paradoxical. Exceptionally tall parents tended to have children who were, on average, taller than the general population — entirely expected, since height runs in families. But those children tended to be somewhat shorter than their exceptionally tall parents, drifting back toward the population's average height rather than continuing to climb higher, generation after generation. Symmetrically, exceptionally short parents tended to have children who were shorter than average, but somewhat taller than their own exceptionally short parents, again drifting back toward the middle rather than away from it.

Galton called this pattern "regression toward mediocrity," later softened to the more familiar "regression toward the mean." It is, quite literally, where this entire statistical technique gets its name — well over a century before anyone was fitting lines to scatter plots of ice cream sales or medical outcomes.

Be precise about what this pattern is, and, just as important, what it is not. Regression to the mean is one of the most widely misunderstood phenomena in applied statistics, mistaken for something mysterious or even causal far more often than its actual, purely statistical nature deserves. It is not that height is somehow being actively pulled back toward the population average by some biological corrective force, punishing extremity. It's a direct, almost mechanical consequence of the fact that no predictor, however strong, correlates perfectly

with the outcome it predicts. Whenever two variables are imperfectly correlated, an extreme value of one will tend, on average, to be paired with a somewhat less extreme value of the other — purely as a mathematical property of imperfect correlation itself, without needing any additional biological or causal explanation whatsoever.

A parent's height is a strong, but imperfect, predictor of a child's height: genetics matters enormously, but so does an enormous tangle of other, partly random influences — nutrition, the other parent's height, sheer developmental chance. Because the correlation is strong but not perfect, an extremely tall parent's other influences on their child's height are not guaranteed to be extreme in the same direction. On average, they pull the child's height back somewhat toward the population mean.

This exact same phenomenon, stripped of the height-specific story, shows up constantly in modern, high-stakes settings, often mistaken for a real effect where none exists. A struggling sports team that fires its coach after an unusually poor losing streak, and then improves under the new coach, may simply be exhibiting ordinary regression to the mean: an unusually poor performance was, in part, due to bad luck that was never going to persist regardless of who was coaching, and a return closer to the team's true average performance level was likely no matter what changed. A patient who is prescribed a new treatment specifically because their symptom levels happened to be at an unusually severe peak may improve afterward partly, or even entirely, because unusually severe measurements tend to be followed by less severe ones on their own — regression to the mean masquerading as a treatment effect.

This is precisely why randomized, controlled experimental designs matter so much. A properly randomized control group experiences the exact same regression-to-the-mean pull the treatment group does, since both groups were equally likely to have been measured at an unusually extreme moment. Only a genuine difference beyond this shared, expected pullback can be honestly credited to the treatment itself.

Galton's own height data, in fact, is precisely what led him to the mathematical technique this Part has built from scratch. Fitting the single line that best summarizes how one variable predicts another was his own tool for measuring exactly how much regression-to-the-mean pull was happening in his height data — how far his children's heights drifted back toward the average relative to their parents'.

The technique kept its inventor's name for the specific phenomenon that led him to invent it, even though the technique itself long ago outgrew that one original height-and-family application, and now fits everything from ice cream sales to insurance claims to disease risk. The name is a historical fossil, in the best sense: a small, permanent reminder, embedded right in the vocabulary every regression model uses today, of the very first pattern that made someone build the tool in the first place.

When the line breaks its own promise

A medical researcher wants to predict whether a patient will develop a certain condition, based on age. She has hundreds of patients, each recorded as either developing the condition (call it 1) or not (call it 0), and tries fitting a straight line directly to this 0/1 outcome, the same way the coach fit speed to hours. For a 95-year-old patient, the line predicts a "probability" of 1.15.

This is not a computational slip. The least squares machinery worked exactly as designed. The problem runs deeper. A straight line, by its very nature, keeps climbing forever; it doesn't know how to stop. But a probability is a fundamentally different kind of number. It lives entirely between 0 and 1, by definition, no exceptions. A tool built for outcomes that can be almost anything breaks the moment you point it at an outcome that's trapped inside a box.

So picture, instead, the shape we need: something that can rise or fall smoothly, but that gently flattens and hugs zero as the predictor shrinks, and gently flattens and hugs one as the predictor grows, never quite touching either boundary, no matter how extreme the input gets. Draw this in your head: flat and low on the left, a steep climb through the middle, flat and high on the right. An S, lying on its side, stretched out. See Figure 6.3 for the logistic S-curve versus the broken linear model.

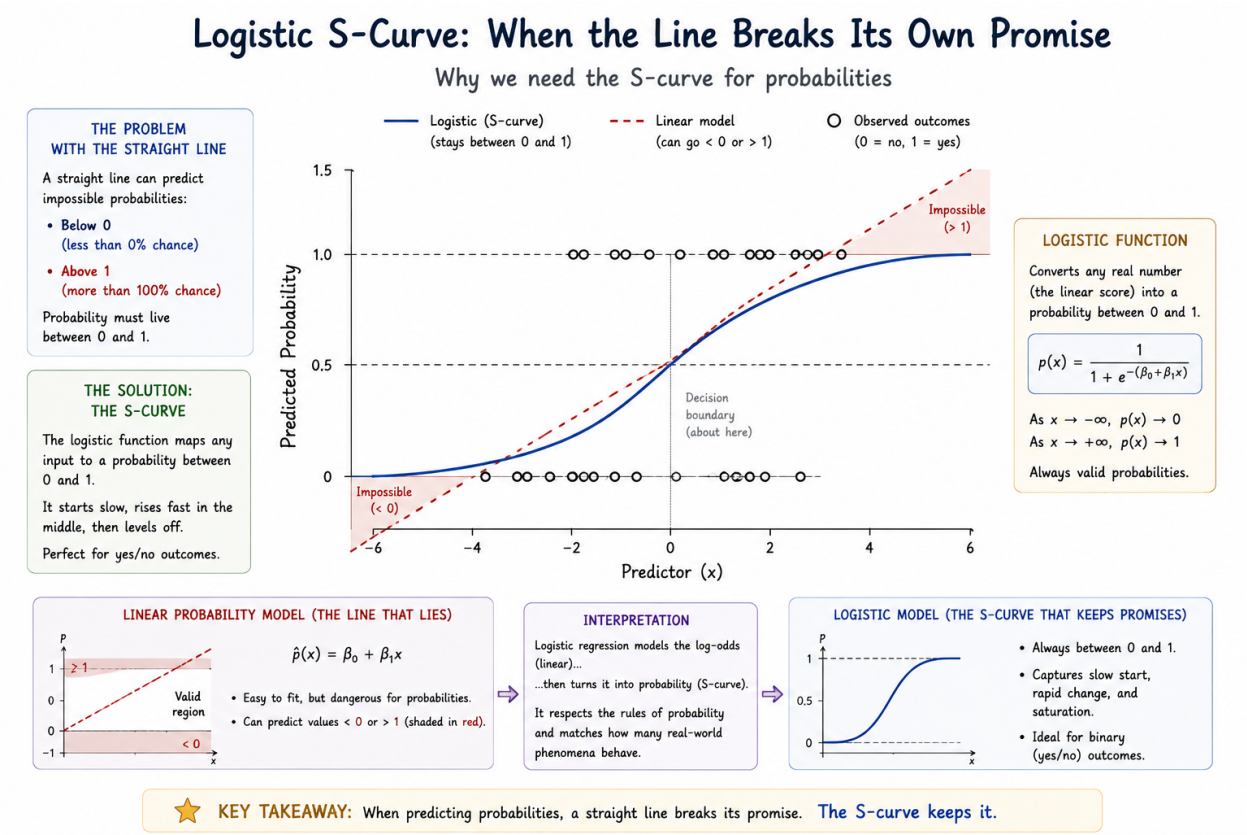

Figure 6.3 — Logistic S-curve: why probabilities need a curve, not a line.

This is the logistic function, or sigmoid curve, and it has a lovely piece of built-in common sense: it's steepest where uncertainty is highest, around the middle, where the outcome is a real toss-up, and flattest where the outcome is already nearly settled, near the extremes.

Think about what one more year of age should intuitively do to a person's risk. For a healthy 20-year-old already at very low risk, one more year barely moves the needle; there's no room to fall further, and the risk factor hasn't yet become dominant. For someone around 60, sitting near an actual 50-50 split, one more year could meaningfully swing things either way — this is where the curve is steepest, most sensitive to small nudges. And for someone already at very high risk, one more year again has diminishing effect, because there's little room left to climb. The S-curve isn't an arbitrary mathematical convenience bolted onto the problem. It's shaped like intuition itself.

Here's the elegant trick for building this curve without throwing away everything just learned about straight lines. Instead of directly predicting the probability, which must obey the 0-to-1 boundary, predict something called the log-odds: a transformed version of probability that can wander freely across all real numbers, with no boundary at all. The odds of an event are just the probability of it happening divided by the probability of it not happening — an event with probability 0.75 has odds of 3-to-1. Take the logarithm of those odds, and the tightly bounded 0-to-1 probability scale stretches out into an unbounded scale running from negative infinity to positive infinity. And on this stretched-out scale, an ordinary straight line makes perfect sense again: fit a regression line predicting log-odds from age, with no boundary problems whatsoever. Then, as a final step, transform the prediction back from log-odds to actual probability using the logistic function, and this transformation step is exactly what guarantees the final answer lands safely between 0 and 1, no matter how extreme the straight-line log-odds prediction underneath it gets. The whole trick, in one sentence: do the easy linear thing on a scale where it's safe to do it, then translate the answer back into the real world using a doorway specifically built to never let an impossible number through.

Because the curve's steepness itself changes depending on where you are along it, a logistic regression's coefficient isn't read the same simple way as a straight line's slope. Instead, it's read as an odds ratio: how much the odds of the outcome multiply for each one-unit increase in the predictor. An odds ratio of 1.15 for age means each additional year multiplies the odds of developing the condition by 1.15, a constant 15% multiplicative nudge on the odds scale, even though its effect on the raw probability itself varies depending on where you start. And the causation warning returns here with full force, arguably more urgency: a logistic regression showing older age strongly associated with a higher probability of a condition doesn't, by itself, prove age causes it. Age might be a genuine biological driver, or a stand-in for decades of accumulated exposure to something else entirely. This matters enormously in

logistic regression's most common real-world home, medical diagnosis and risk prediction, where the stakes of mistaking correlation for causation are measured in real treatment decisions for real people. This is also the exact machinery underneath credit scoring (will this applicant default?), marketing (will this customer churn?), and much of applied political science (will this voter support this candidate?), anywhere the outcome is fundamentally a yes-or-no question rather than a number that can take any value.

The number the line was hiding the whole time

R-squared answers how much wrongness did the line erase. There's an older, more direct question sitting one layer beneath it: before fitting any line at all, how strongly do two variables move together, and in which direction? This is Pearson's correlation coefficient, r, worth building explicitly since it's been referenced by name several times already without ever being written down. See Figure 6.4 for a gallery of Pearson r values from strong positive to strong negative.

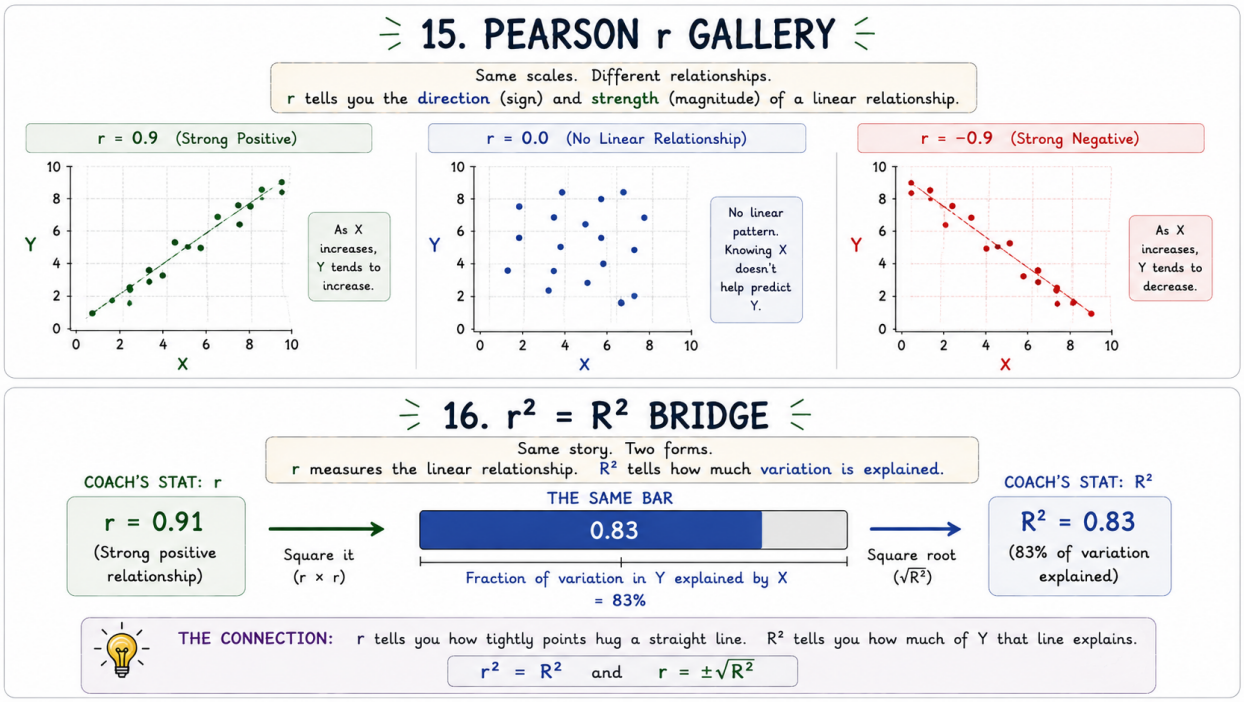

Figure 6.4 — Pearson r gallery: strong positive, none, and strong negative correlation.

THE MATH

Why standardize both variables before multiplying: practice hours and speed are measured in completely different units, so their raw covariance is stuck in a unit

nobody can interpret. Dividing by both standard deviations rescales the result into a pure, unitless number between -1 and 1.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Worked example: If hours and speed produce a numerator of 340, with $\sqrt{\sum(x_i - \bar{x})^2} = 22$ and $\sqrt{\sum(y_i - \bar{y})^2} = 17$, then $r = 340/(22 \times 17) = 340/374 \approx 0.91$, a strong positive relationship. For a simple one-predictor regression, $r^2 = R^2$ exactly, so $0.91^2 \approx 0.83$; the chapter's own R-squared of 0.7 corresponds back to an r of about 0.84.

Meaning: r's sign tells you direction, its magnitude tells you strength, with 0 meaning no linear relationship and plus-or-minus 1 meaning a perfectly straight line. This is also why r is a purely linear measure, and why Anscombe's Quartet ahead can produce nearly identical r values for a curved dataset, an outlier-dominated dataset, and a genuinely linear one: r only asks how well a straight line fits, never whether a straight line makes sense here at all.

▸ *Interactive version of this math box*

When one string isn't the whole story

The coach's line used a single predictor, practice hours, to predict speed. Real prediction problems are rarely this generous. A medical researcher predicting blood pressure plausibly needs age, weight, sodium intake, and family history all at once, not one at a time, since treating them one at a time throws away exactly the information that matters most: how these predictors relate to each other, not just to the outcome.

THE MATH

Why adding more predictors changes the meaning of each coefficient: in simple regression, the slope for hours captured everything hours was associated with. With several predictors fit simultaneously, each coefficient instead captures that one predictor's own contribution holding every other included predictor fixed, a genuinely different, usually more honest quantity.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Worked example: A simple regression of blood pressure on age alone might find a coefficient of 0.9. Add weight to the same model, and age's coefficient might shrink to 0.5, not because age stopped mattering, but because some of what looked like age's effect in the simple model was actually weight's effect, riding along with age since older patients also tend to weigh more.

Adding more predictors always improves R-squared on training data, whether or not those predictors genuinely help, which is why researchers report adjusted R-squared, a version that specifically penalizes each additional predictor, rising only when a new variable improves the fit by more than chance alone would predict.

A second danger: multicollinearity, when two or more predictors are themselves strongly correlated, say weight and body-mass index. The model struggles to tell which overlapping predictor deserves credit, producing unstable coefficients that swing if one is dropped. A variance inflation factor (VIF) is the standard diagnostic for this.

Logistic regression extends the identical way: log-odds = b0 + b1(age) + b2(smoking) + b3(cholesterol), each coefficient again read as its own odds ratio holding every other predictor fixed, exactly the language a clinical paper uses when it reports an adjusted odds ratio.

Meaning: This is precisely the confounding problem Part XIII builds in full, showing up here in miniature: a predictor's simple-regression coefficient can be inflated or deflated by anything else it correlates with, and including that variable in the same model is one direct way of adjusting for it.

▸ *Interactive version of this math box*

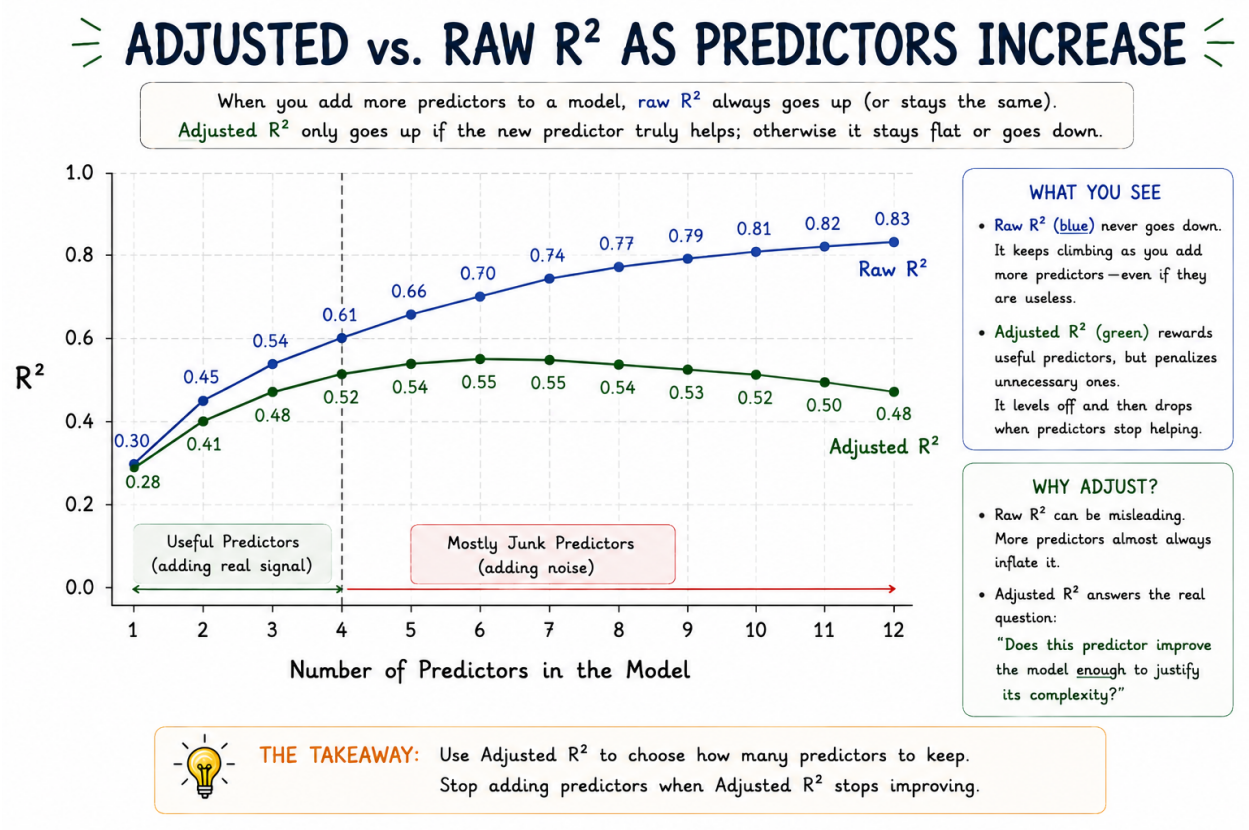

Figure 6.5 — Adjusted versus raw R-squared as predictors increase: useful signal versus noise.

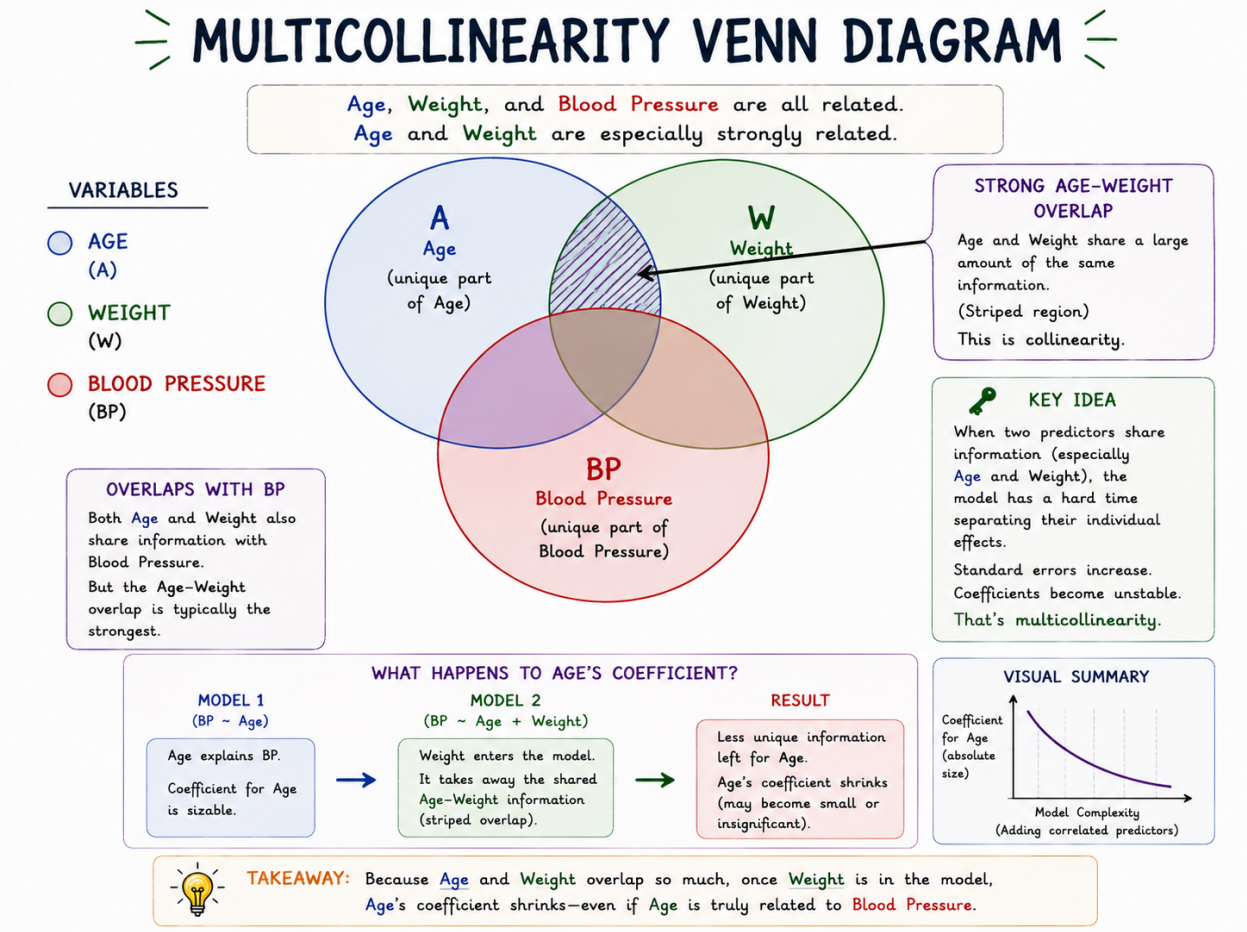

Figure 6.6 — Multicollinearity: why a correlated predictor's coefficient shrinks when another enters the model.

One idea, wearing three different outfits

An insurance analyst wants to predict how many claims a policyholder will file in a year, based on age and driving history. She tries a straight line directly, and for one combination of risk factors, the model predicts −0.4 claims. Nobody can file a negative number of claims.

Notice something: this is the exact same story as the medical researcher's impossible 1.15 probability, showing up in a new setting. A straight line, unconstrained, keeps climbing or falling forever, and once again it's been pointed at an outcome that lives inside a box, this time a box with a floor (zero) but no ceiling, since claim counts can be 0, 1, 2, 3, climbing indefinitely, but never negative and never fractional.

This repetition is worth a second look, because it's about to pay off with something satisfying: instead of inventing a brand-new fix from scratch every time a new kind of outcome breaks the straight line, step back and ask what the fix itself actually consists of, structurally, so the same fix can be reused. Look back at what solved the probability problem: three ingredients working together. A random component: the fact that the outcome follows some specific,

known kind of randomness (a yes/no outcome follows the binomial pattern built earlier). A systematic component: an ordinary, boring straight-line combination of predictors, exactly the part that never needed fixing. And a link function: the doorway (the logistic transformation, in that case) connecting the two, translating an unbounded straight-line prediction into something that respects the outcome's real-world boundaries.

This three-part recipe, random component, systematic component, link function, is a generalized linear model, or GLM, and it turns out to be genuinely broad, not a one-off trick invented for probabilities alone. A claims count follows a different kind of randomness than a yes/no outcome, the counting pattern built earlier for rare events, so it needs a different doorway. Since a count can be any non-negative number with no upper limit at all, the natural fix is the log link: predict the logarithm of the expected count using an ordinary straight line, then exponentiate the result to get back an actual predicted count. Exponentiating any real number, positive, negative, huge, tiny, always produces something strictly positive. The negative-claims problem simply cannot happen anymore, by mathematical construction, using exactly the same conceptual move that fixed the probability problem: do the easy linear thing on a transformed scale, then walk back through a doorway that only lets valid answers out the other side.

Here's the part that deserves the full weight of attention: the very first regression line built in this Part, the coach's straight line, was never a separate technique from these two. It's the simplest possible GLM, the special case where the outcome is already unbounded and continuous, so the doorway is simply left open, unlocked, doing nothing at all. Three seemingly different tools turn out to be one single idea, dressed for three different occasions. Once this pattern clicks, extending regression to a brand-new kind of outcome — proportions, ordered categories, survival times waiting in Part XII — stops being "learn an entirely new technique" and becomes "ask two questions": what kind of randomness does this outcome naturally follow, and what doorway connects an ordinary straight line to that outcome's valid range? This exact three-part recipe is one of the load-bearing ideas underneath modern machine learning. A model's output layer is very often nothing more than a chosen link function, translating an internal, unbounded computation into a valid final answer — a probability, a count, a class. And the whole apparatus of "loss functions" that trains these models is, at its root, the same maximum-likelihood machinery built earlier, just running at a much larger scale.

THE MATH

Why the log link specifically fixes the negative-claims problem: exponentiating any real number, however negative, always produces something strictly positive, so wrapping a straight-line prediction in an exponential guarantees the final answer can never fall below zero, mirroring how the logistic function guaranteed probabilities never left the 0-to-1 box.

$$\ln(\hat{\mu}) = b_0 + b_1 x \;\Rightarrow\; \hat{\mu} = e^{b_0 + b_1 x}$$

Worked example: If a straight-line fit on the log scale predicts $\ln(\hat{\mu}) = -0.5$ for a policyholder, the predicted claims count is $e^{-0.5} \approx 0.61$ claims, a small positive number, never negative, however the underlying linear combination lands.

Meaning: This is the exact same three-part recipe as logistic regression, random component, systematic component, link function, just swapping the logistic doorway for the exponential one.

▸ *Interactive version of this math box*

Four datasets, one perfect disguise

In 1973, statistician Francis Anscombe built four different datasets, deliberately engineered so that all four would produce nearly identical summary statistics: the same mean, the same variance, the same regression line, the same R-squared. By every single number covered so far, these four datasets are indistinguishable. Plotted, they look nothing alike.

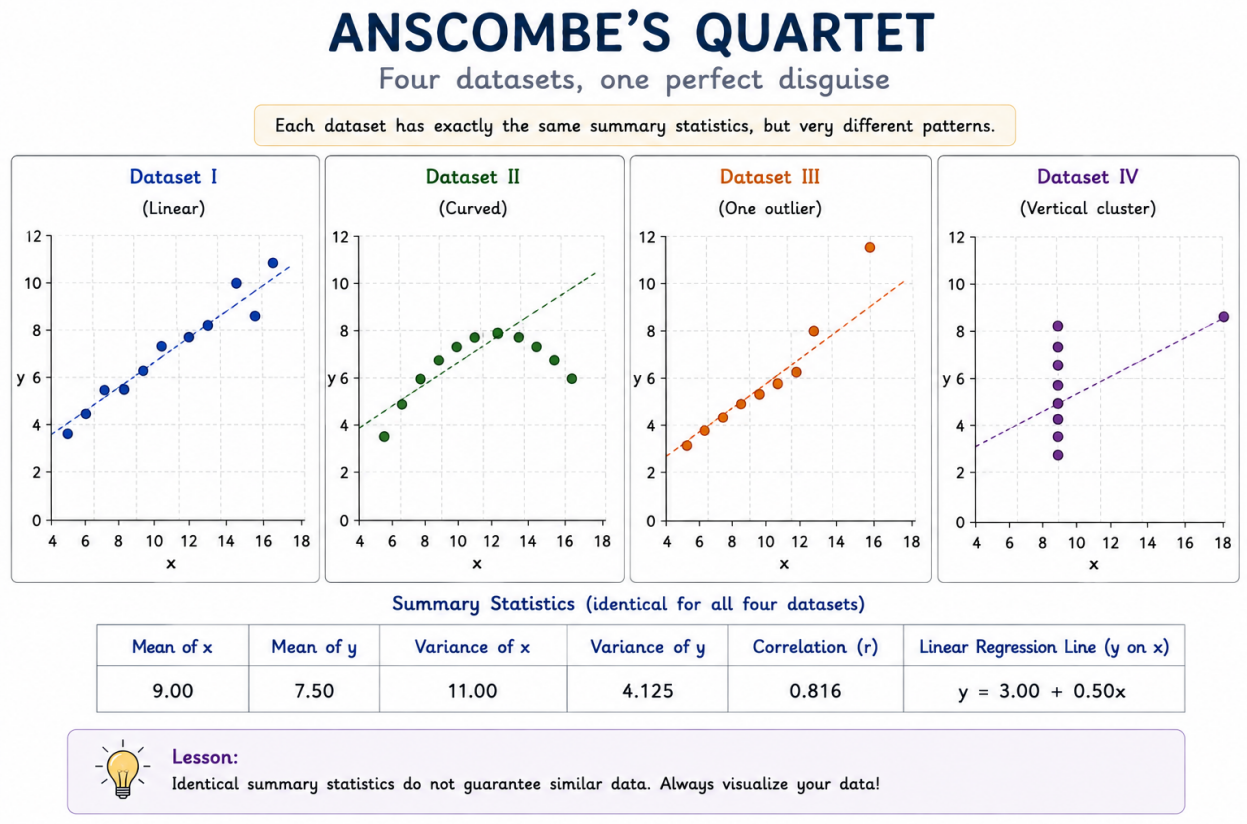

Figure 6.7 — Anscombe's Quartet: four datasets, identical summary statistics, one perfect disguise.

This deserves more than a curiosity's worth of attention, because it's the single most important caution in this entire Part. Dataset I looks exactly like what regression assumes: points scattered fairly evenly and randomly around the fitted line, with no pattern to how far above or below it any point sits. This is the dataset regression was actually built for, and here the line honestly earns its keep. Dataset II is not linear at all; it's a smooth, clearly curved relationship, engineered so that fitting a straight line through it produces the exact same slope, intercept, and R-squared as Dataset I. A straight line is fundamentally the wrong shape here, and no amount of staring at the R-squared number alone would ever reveal that. Dataset III has one single outlier sitting far from an otherwise nearly perfect line among the remaining points, dramatically dragging the fit away from what would otherwise be an almost exact match. Dataset IV is the strangest of all: nearly every point shares the same x-value, clustered in a single vertical line, except one point far to the right that, alone, determines the entire slope. Remove that one point and there's no meaningful relationship left to fit at all.

Four datasets. Identical mean, variance, regression line, R-squared. Four completely different underlying stories. If you only ever looked at the summary numbers, you would conclude all four tell the identical tale, and you would be wrong about three of them.

The fix is refreshingly simple and costs almost nothing: the residual plot, a scatter plot of the model's residuals against the predictor variable. For a well-fitting model, this should look like a formless, random spray of points above and below zero: no shape, no story, just noise. Any visible pattern is a red flag the raw R-squared number was hiding.

Two further checks are worth building into habit alongside the residual plot. Non-constant spread, or heteroscedasticity, shows up in a residual plot as a cone or fan shape rather than a uniform band: predictions might be quite precise for low predictor values and wildly imprecise for high ones. This violates an assumption baked silently into classical confidence intervals for regression coefficients, and often calls for the resampling-based approaches — bootstrapping, permutation testing — built earlier in Part V.

Checking for influential points more broadly, not just Dataset IV's dramatic case but subtler versions of it in real data, extends the outlier framework from Part II directly into the regression context: a responsible analyst checks not just whether outliers exist, but how much any single point is steering the model's conclusions without announcing itself.

It would be easy to walk away from Anscombe's Quartet feeling that regression can't be trusted, that any impressive R-squared might secretly be hiding a curve, an outlier, or a single point pulling all the strings. That's not quite the right lesson. The right lesson is that a fitted line and a diagnostic check are not separate, optional steps. They're two halves of one complete practice, and this closes the loop all the way back to the very first idea from Snow's cholera map: a summary number is a compression of a full shape, and compressions can hide exactly the detail that matters most. Look at the shape. Always look at the shape — this time, the shape of what the model got wrong, not just the shape of the raw data. Rigorous peer review now expects residual plots alongside reported regression results, precisely because Anscombe-style failures have contributed to published findings that later failed to replicate. Financial models are especially vulnerable to Dataset IV's failure mode, where a handful of extreme market events can single-handedly steer a fitted relationship. And policy-relevant models — studying how an intervention affects economic or social outcomes — carry real consequences when this checking is skipped, since an undetected influential point or hidden curve can lead a whole policy recommendation astray on a foundation nobody actually looked at.

The question everyone actually wants answered

Regression can now find the invisible string tying two variables together, in three different flavors, and it knows how to check whether the string it found is even the right shape for the cloud it's threaded through. Genuinely useful. Also, if we're honest, not the question you actually came here for.

"Correlation isn't causation" has been muttered like a superstition throughout this entire Part — a warning label, not a solution. Saying it every few pages doesn't get you any closer to knowing whether X actually causes Y, any more than reading a cigarette pack's warning label tells you your own cholesterol number. Building an actual method for answering that — not just flagging the danger, solving it — is where the next Parts are headed. There are just a few more group comparisons to make first.

Try it yourself: Part VI

▶ PRACTICE MODULE

Open the interactive practice module for this Part

Visualizations for Part VI:

- Anscombe Quartet
- Least Squares Drag
- Logistic Curve
- Pearson r
- Multicollinearity

Quick check *(answers at the back of the book, VI.1–VI.6)*

VI.1. A cricket coach fits speed = 118 + 0.8 × hours. What does the coefficient 0.8 actually mean in plain language?

VI.2. The coach's model has an R-squared of 0.7. What does this number actually tell you, and what does it not tell you?

VI.3. Galton found that children of exceptionally tall parents were taller than average, but shorter than their own parents. Why does this happen, without invoking any biological "correction" force?

VI.4. A sports team fires its coach after an unusually bad losing streak, then improves. Why might this improvement have nothing to do with the new coach?

VI.5. Two bowlers have identical R-squared of 0.7 for the practice-hours-vs-speed relationship, but one has Pearson r = 0.84 and the other has r = -0.84. What's actually different between these two bowlers' data, given that R-squared can't tell them apart?

VI.6. A model predicting blood pressure from age alone finds a coefficient of 0.9. Adding weight to the same model shrinks age's coefficient to 0.5. Has age stopped mattering? What's actually happening?

Explore further *(open-ended — hand these to an AI chat assistant)*

VI.7. Take the exact cricket coach numbers — speed = 118 + 0.8 × hours, R-squared = 0.7. Ask an AI chat assistant to generate a small fictional dataset matching these statistics, then construct a second dataset with the same regression line and R-squared but an obviously curved relationship, the way Anscombe did.

VI.8. Using Galton's regression-to-the-mean idea, ask an AI chat assistant to design a fictional scenario and explain, step by step, how regression to the mean alone could produce an apparent "improvement" with no real intervention at all.

VI.9. *(harder)* Ask an AI chat assistant to walk through why logistic regression uses log-odds instead of predicting probability directly, using the medical researcher's 95-year-old patient example, and to show the actual log-odds-to-probability conversion with real numbers.

VI.10. Using the age/weight blood pressure example, ask an AI chat assistant to simulate a dataset where age and weight are correlated at 0.6, fit both a simple and multivariable regression, and show you directly how much age's coefficient shrinks — then repeat at correlation 0.9 and compare.