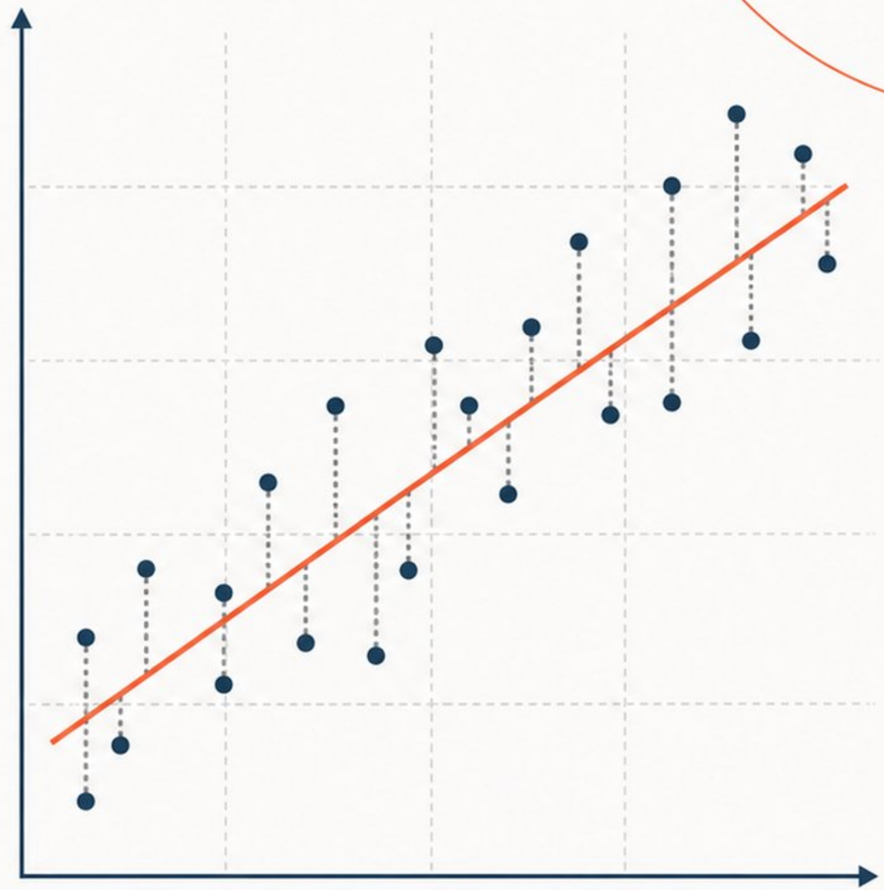
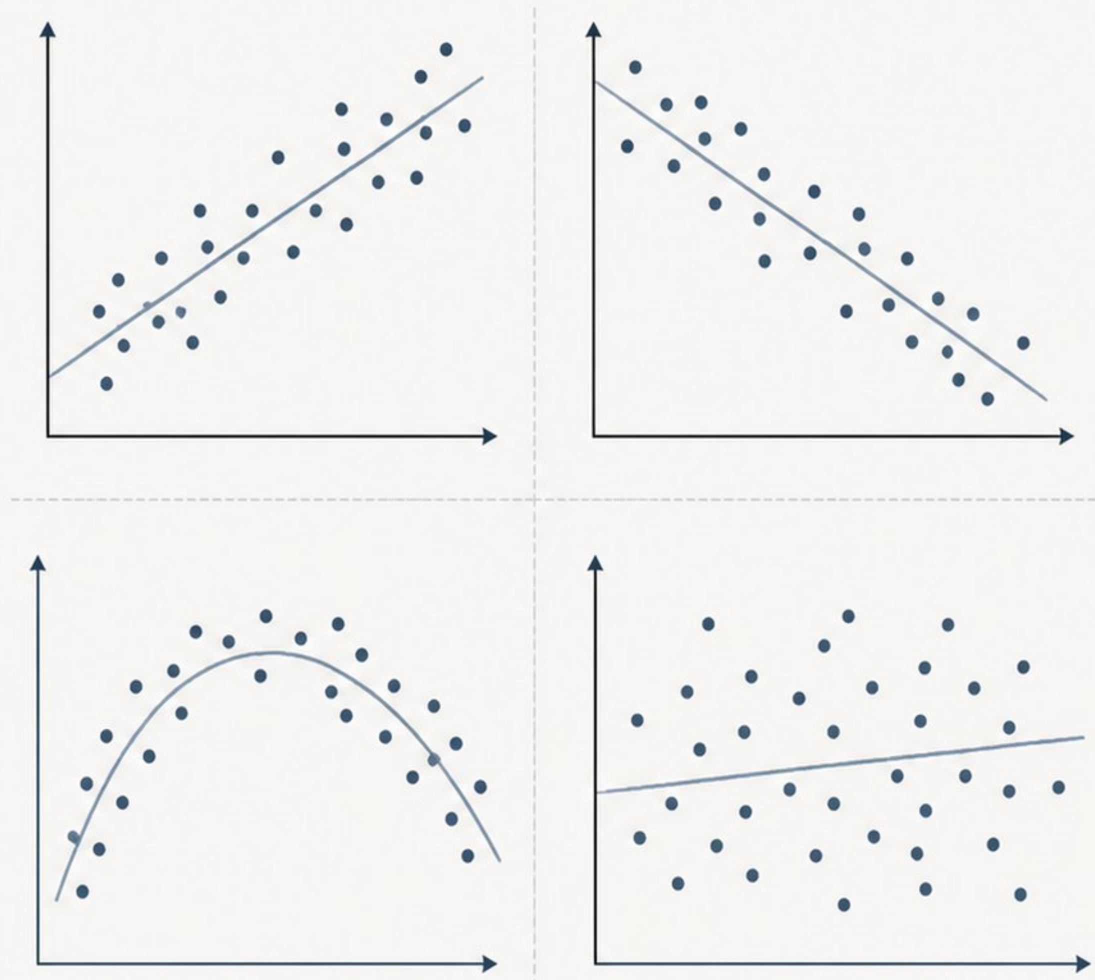
Introduction to Regression: Lines, Residuals, and Patterns

- Course roadmap: scatterplots → fitted lines → residuals → R^2 → responsible limits
- Use fitted lines and residuals to describe two-variable relationships — and know when the model breaks down
- Turn noisy data into actionable insight while staying honest about uncertainty
- Real-world examples: marketing spend vs. sales, height vs. armspan, demand forecasting
- Prerequisites: coordinate graphing, mean, standard deviation, and optionally correlation



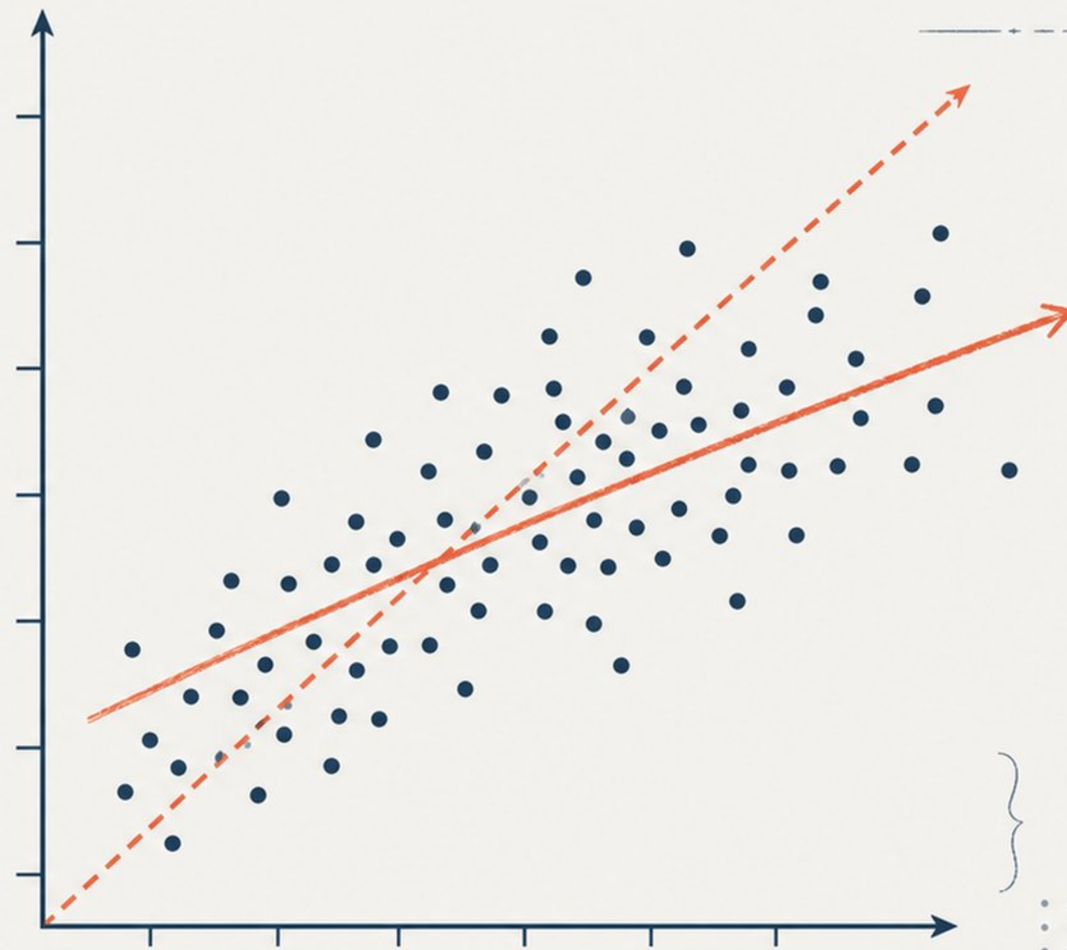
Visualizing Two-Variable Relationships

- Start with a scatterplot: check direction, form, and strength
- Explanatory variable on x-axis; response variable on y-axis
- Drawing a line by eye: “balance” points above and below
- Common trap: judging fit solely by how many points the line touches
- We need an objective method — not just a visual guess



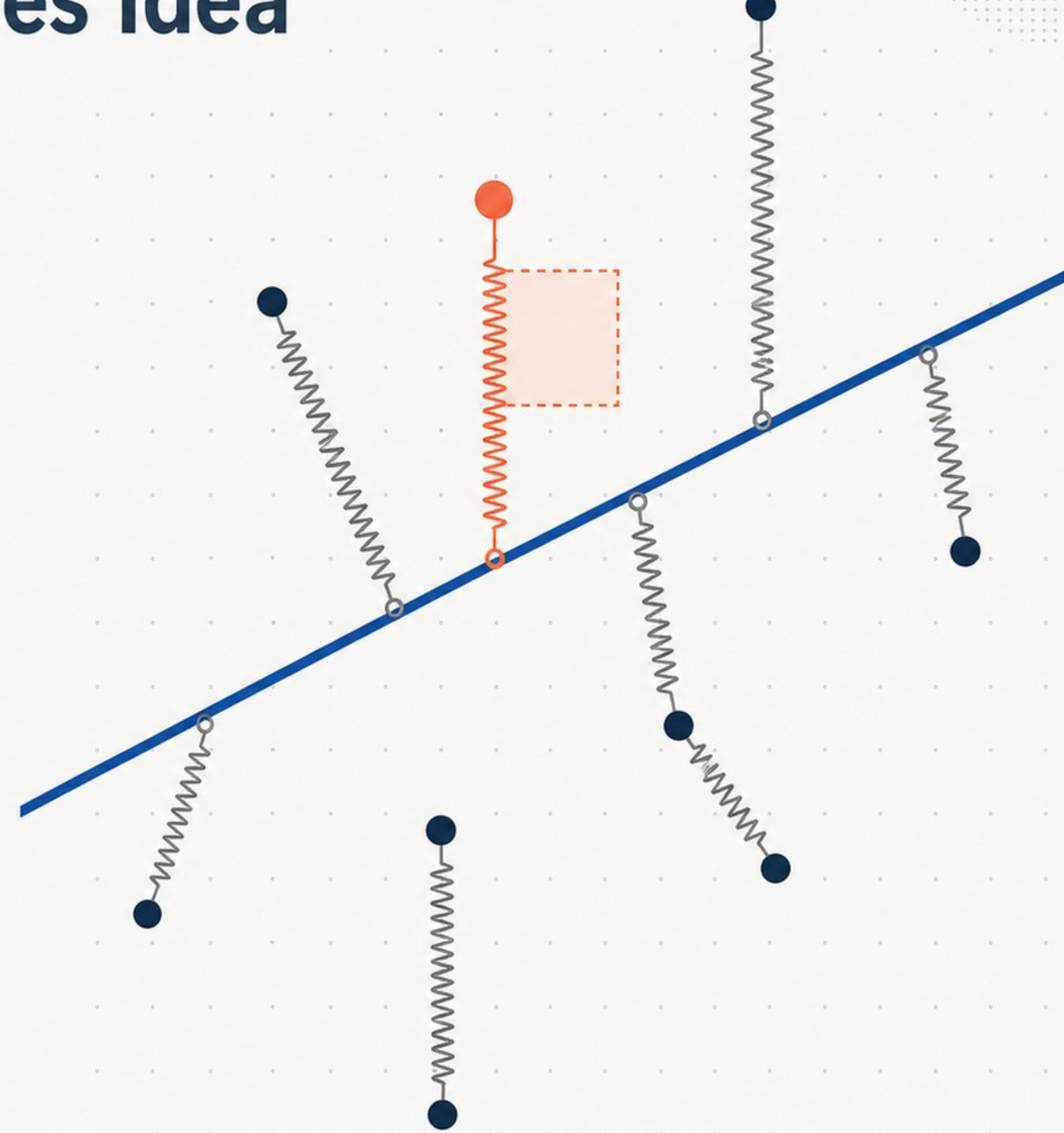
Activity: Build a Scatterplot and Draw a Candidate Line

- Measure height and armspan in pairs, then plot all class data (armspan vs height)
- Draw y=x line: da Vinci predicted armspan equals height
- Draw a line of best fit by eye through the point cloud
- Compare the two lines and discuss why they differ
- Scatterplots reveal patterns summary statistics hide; eye-fitting is only a starting point



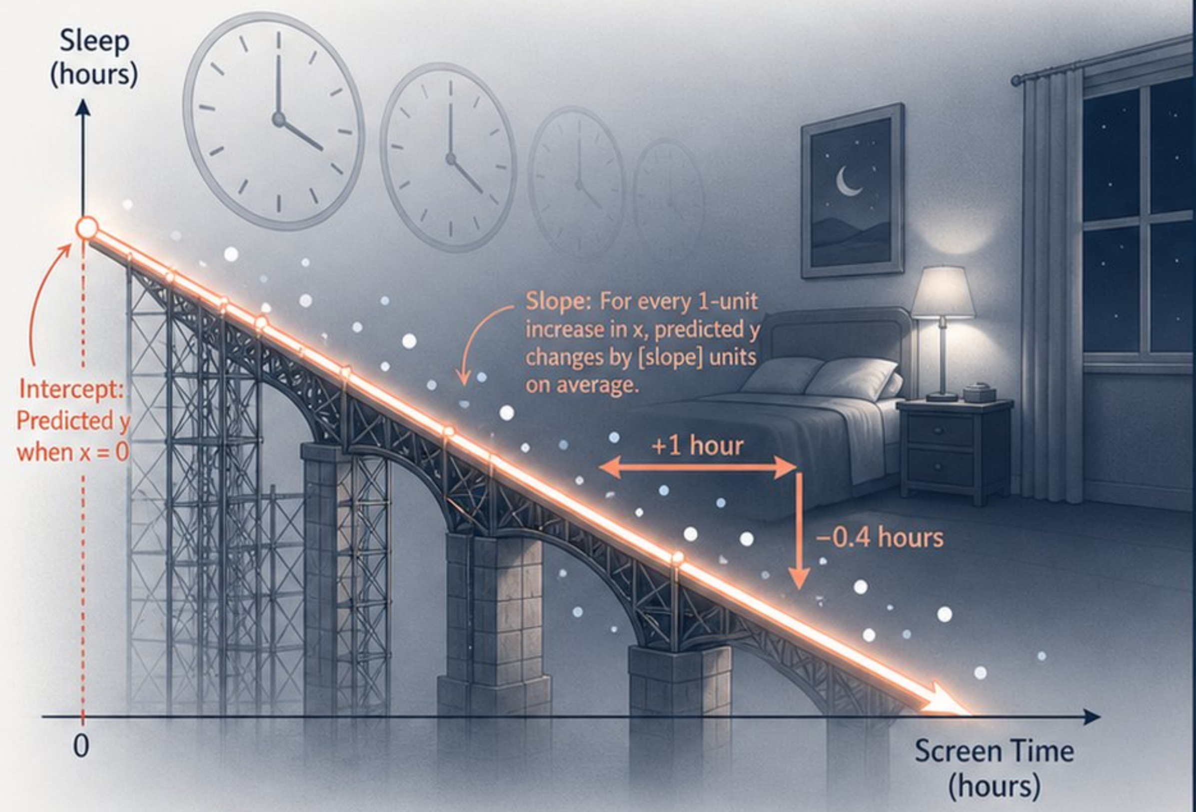
The Least-Squares Idea

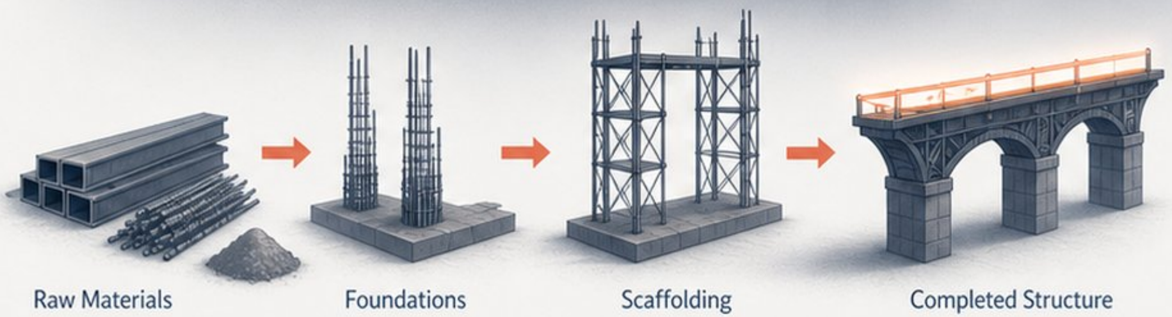
- Residual: vertical distance from data point to the line (observed − predicted)
- Squaring residuals prevents canceling errors and penalizes large misses heavily
- Least-squares chooses the line minimizing the sum of squared residuals
- Physical analogy: springs pulling vertically; equilibrium = minimum total energy
- Mathematically optimal straight line given the data — not just best by eye



Interpreting Slope and Intercept in Context

- **Slope:** For every 1-unit increase in x, predicted y changes by [slope] units on average.
- **Intercept:** Predicted y when x=0. Check if zero is within the observed data.
- **Ex:** Each extra hour of screen time is associated with 0.4 hours less sleep.
- **Danger:** Extrapolating the intercept far outside the data range (e.g., height=0).
- The line describes an average trend — it does not pass through every point.

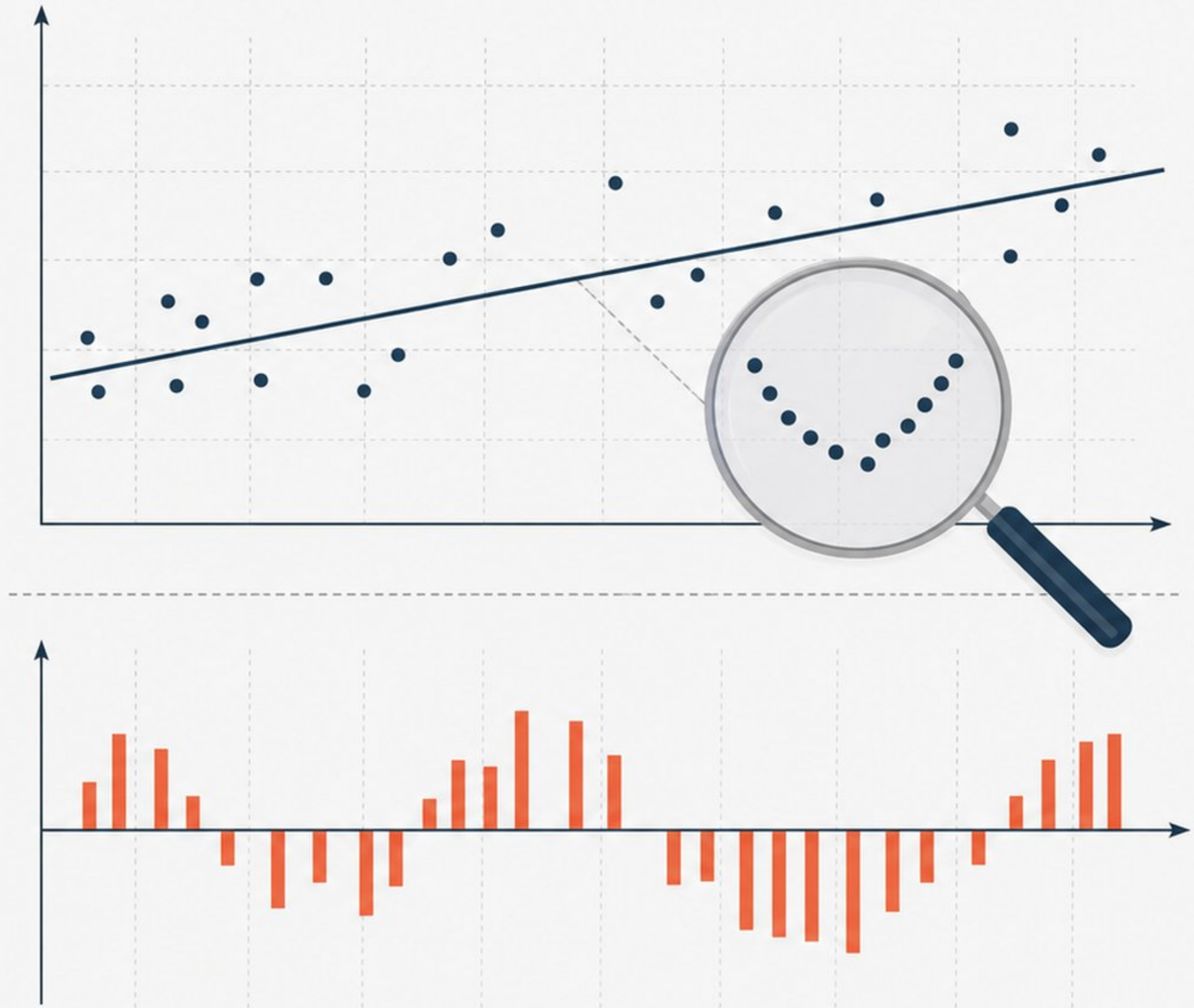




From components to structure — modeling relationships.

Working with Residuals: What the Line Missed

- Residual = observed y − predicted y; positive above the line, negative below
- Residual plot: residuals (y-axis) vs. fitted values or x (x-axis)
- The line captures the average pattern; residuals reveal what was missed
- Ideal: random scatter around the zero line with no obvious shape
- Compute a few residuals by hand before relying on software output

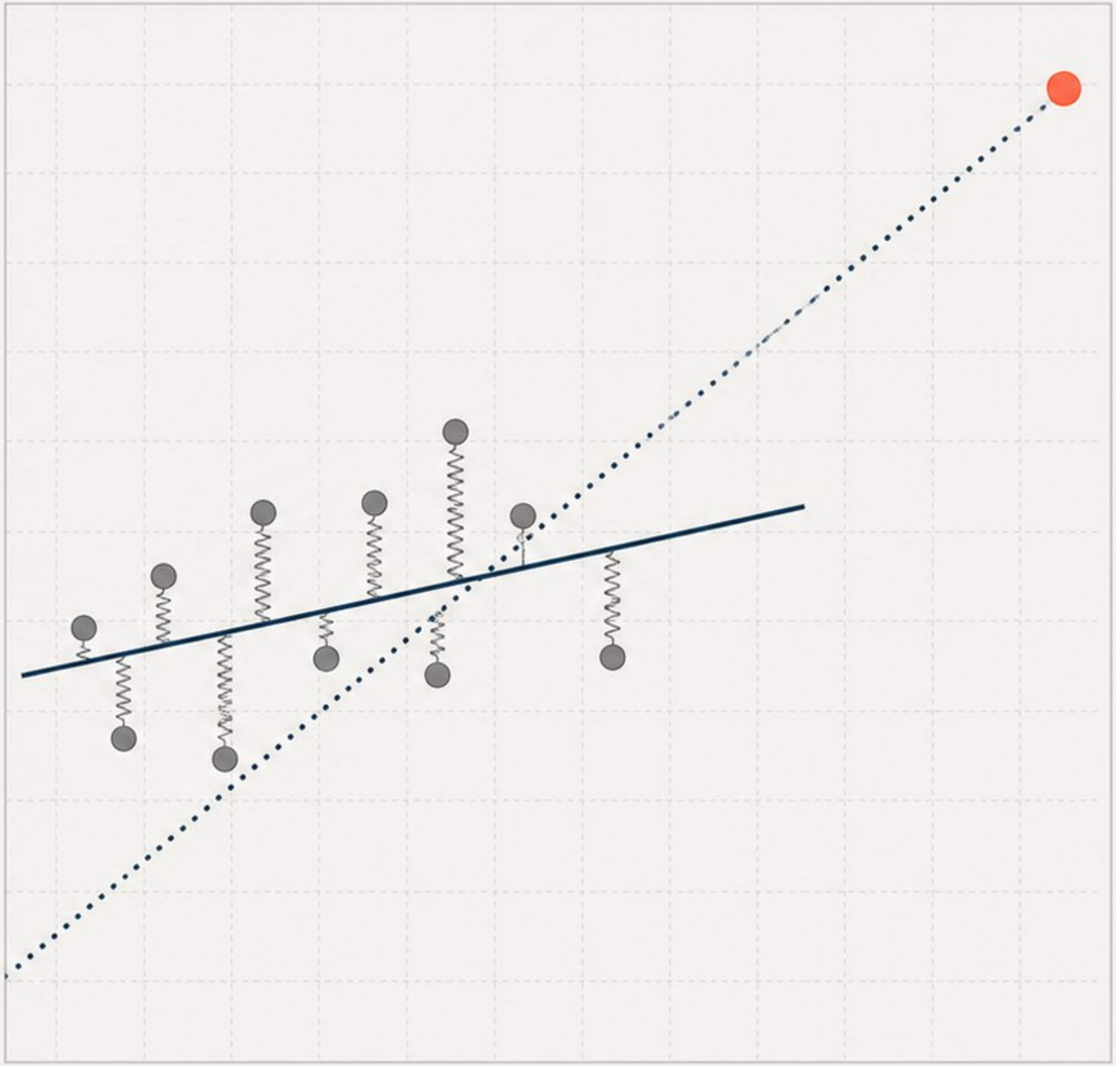


Recognizing Residual Patterns

- **Curved pattern (U-shape/inverted U):** the relationship is not linear
- **Fanning/funneling pattern:** spread changes with predicted values (heteroscedasticity)
- **Outliers:** a single residual far from the random scatter
- Clear pattern means the linear model is inadequate — try a curve or transformation
- **Anscombe's Quartet:** identical stats, different residual patterns — always plot your data

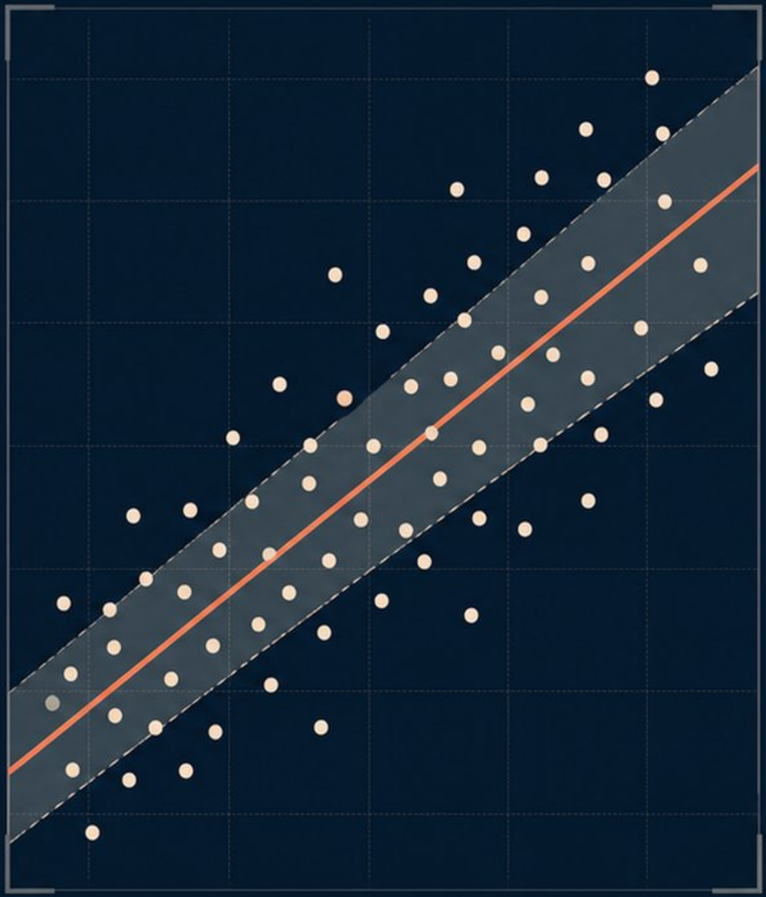
Outliers and Influential Points

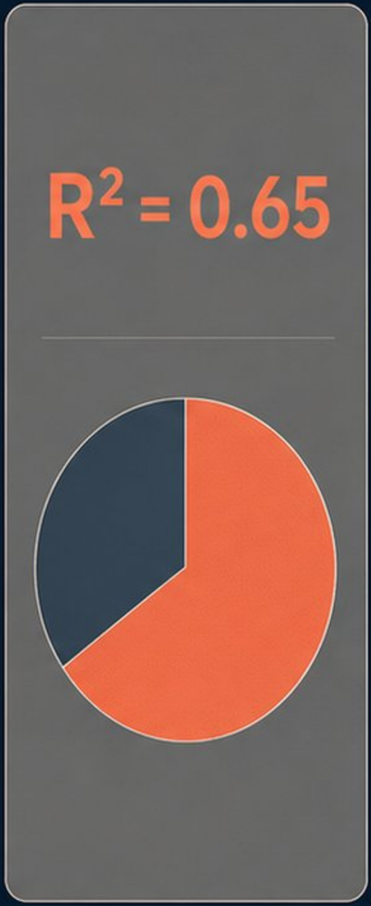
- Outliers in y have large residuals, sitting far above or below the trend.
- High-leverage points have unusual x-values and can pull the regression line.
- Not all outliers are influential; some can dramatically change the slope.
- Always investigate: data error, special case, or genuine relationship?
- In Anscombe's Dataset 4, one high-leverage point determines the entire slope.



R-Squared: How Much Variation Does the Line Capture?

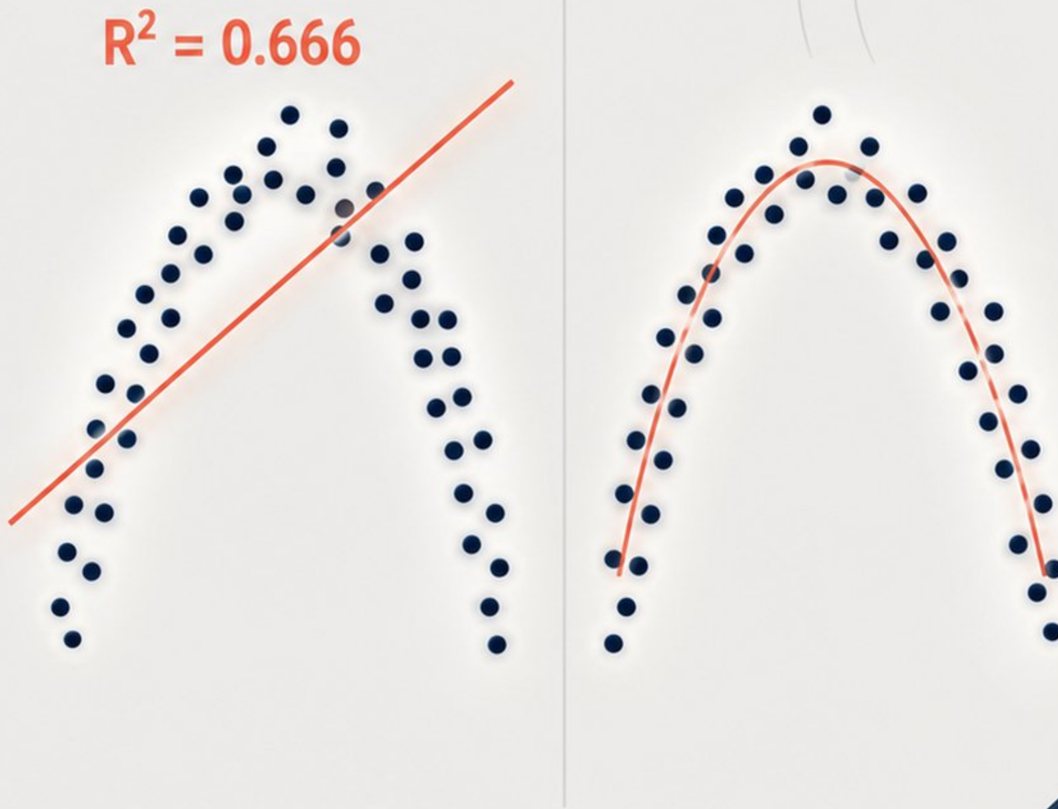
- R^2 measures the proportion of response variation the fitted line accounts for
- Ranges from 0 (none) to 1 (all); e.g., 0.65 means 65% captured
- Physical systems: $R^2 > 0.90$ common; human behavior: 0.30–0.50 can be meaningful
- R^2 describes fit, not causation — never confuse the two
- Adjusted R^2 penalizes extra predictors; use it to compare models





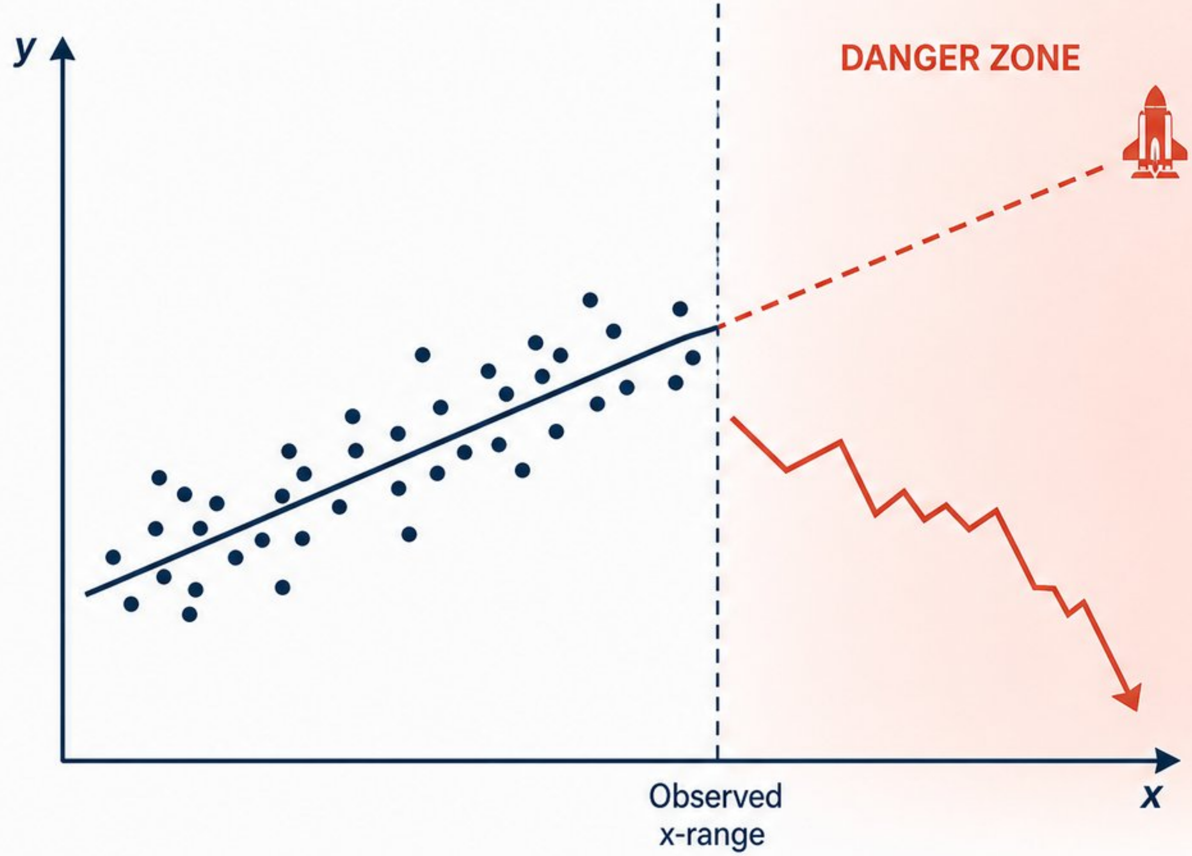
R-Squared Limitations and Traps

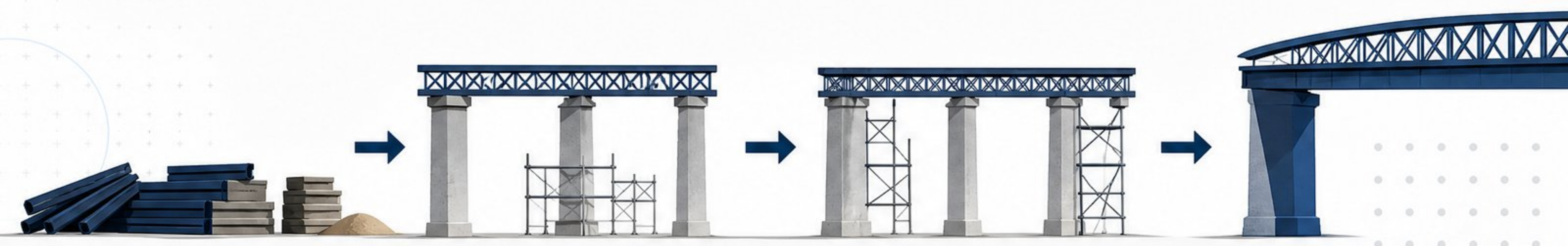
- High R^2 does not guarantee a good model; curved data can still show $R^2 = 0.666$
- R^2 always increases when adding predictors — even random noise inflates apparent fit
- Overfitting memorizes noise: high training R^2, poor performance on new data
- In-sample R^2 is not reliable for causal decisions; collider bias can inflate R^2 while distorting effects
- Always pair R^2 with residual plots and domain knowledge; never use it as the sole quality measure



Describing Relationships Responsibly

- Association is not causation: "x is associated with y" ≠ "x causes y"
- Never extrapolate far beyond observed x-range — the linear trend may not hold
- Challenger: regression at moderate temps couldn't predict O-ring failure at 31°F
- Zillow's $880M loss: models extrapolated past trends, missed a market regime change
- Always check: linear pattern? random residuals? outliers? constant spread?





Communicating Findings to Colleagues

- **Use precise, non-causal language:** “associated with”, not “causes”
- **Always state limitations:** R^2 range, extrapolation risk, model constraints
- **Describe the residual story:** random scatter or curved pattern – what was missed?
- **Follow the checklist:** direction, slope in context, R^2, unusual points, no causal claims
- **Seek specialist help for:** strong residual patterns, influential outliers, causal questions

Guided Practice and Wrap-Up

- Fit a line, compute residuals, then build a residual plot from a small dataset
- Identify curved, fan-shaped, or outlier-driven patterns in the residuals
- Discuss where descriptive regression helps your own work—and where it breaks
- Six essentials: **plot first, minimize** squared errors, inspect residuals, R^2, association ≠ causation, avoid extrapolation
- Keep a quick-reference card: line-fitting steps, diagnostic patterns, communication rules