



# Regression, Correlation, and Data Modeling

Four focused lessons, worked examples, a mistake clinic, and 60 practice problems with full explanations.

## In this booster

**35.1** Linear Regression in Depth

**35.2** Residual Analysis and Influential Points

**35.3** Nonlinear Models and Transformations

**35.4** Correlation vs Causation and Data Ethics

Every section teaches the idea, walks through the most common problem types, flags the traps students fall into, and ends with fifteen practice problems. Full answers and explanations are at the back.

## 35.1 Linear Regression in Depth

Learn to build the least-squares line, interpret its slope and intercept in context, make predictions, and recognize when a prediction goes beyond what the data can support.

### Why This Matters

Whenever two numerical variables seem to move together, we want a rule that turns one into a prediction of the other. A **least-squares regression line** (LSRL) is the standard rule: it is the single straight line that fits a scatterplot best, in a precise sense we will define. Students get stuck in three places. They mix up which variable is the explanatory variable  $x$  and which is the response  $y$ . They recite the slope as a number without saying what it means in context. And they trust the line far outside the data, which is the classic extrapolation error.

Regression is everywhere: estimating how much a plant grows per week, how cooking time changes with the weight of a roast, how fuel use rises with speed, how test scores relate to study time, how a company's costs grow with output. It is also the doorway to nearly every statistical model in college courses in science, economics and engineering. If you can write the equation, interpret every number in it, and say when it should not be used, you have the core skill.

### Core Ideas

Start with a **scatterplot** of paired data  $(x, y)$ . The **explanatory variable**  $x$  is the one we use to predict; the **response variable**  $y$  is the one we want to predict. A straight-line model for the data is written with a hat on  $y$  to remind us that it gives a **predicted** value, not an observed one:

#### The least-squares regression line

$$\hat{y} = a + bx$$

The slope is  $b = r \frac{s_y}{s_x}$  and the intercept is  $a = \bar{y} - b\bar{x}$ , where  $\bar{x}, \bar{y}$  are the means,  $s_x, s_y$  the sample standard deviations and  $r$  the correlation.

The LSRL is the line that makes the sum of the squared residuals,  $\sum(y - \hat{y})^2$ , as small as possible. It always passes through the point  $(\bar{x}, \bar{y})$ .

A **residual** is the miss of one prediction:  $\text{residual} = y - \hat{y}$  (observed minus predicted). A positive residual means the point lies above the line. Squaring the residuals keeps misses above and below the line from cancelling, and it punishes large misses more heavily than small ones.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

## Interpreting slope and intercept in context

**Slope  $b$ :** for each 1-unit increase in  $x$ , the predicted  $y$  increases (or decreases, if  $b < 0$ ) by  $|b|$  units, on average.

**Intercept  $a$ :** the predicted value of  $y$  when  $x = 0$ . It has a practical meaning only if  $x = 0$  is a sensible value that is near the data; otherwise it is just the number that positions the line.

Always name the variables, the units, and use the words predicted and on average.

- ◆ The sign of  $b$  always matches the sign of  $r$ , because  $s_x$  and  $s_y$  are positive.
- ◆ If  $r = 0$  the slope is 0 and the line is flat at  $\bar{y}$ .
- ◆ Changing the units of  $x$  or  $y$  changes  $a$  and  $b$  but not  $r$ .
- ◆ Regressing  $y$  on  $x$  and regressing  $x$  on  $y$  give different lines unless  $|r| = 1$ . Do not solve the equation for  $x$  to predict backward.

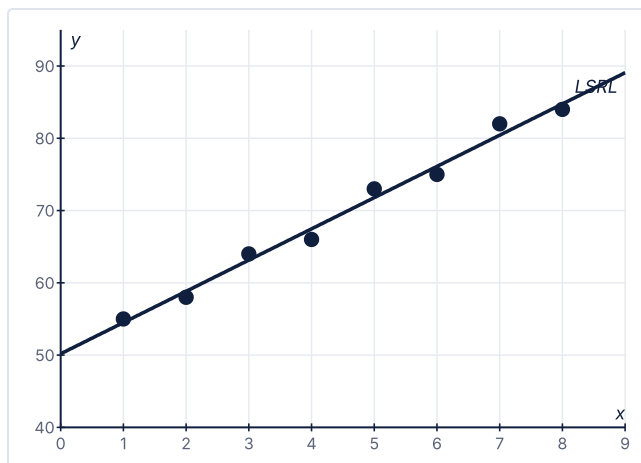
## Extrapolation

**Interpolation** predicts for  $x$ -values inside the range of the data. **Extrapolation** predicts for  $x$ -values outside that range. Extrapolated predictions are unreliable, because we have no evidence that the straight-line pattern continues.

Here is a data set we will reuse. Eight students recorded hours studied per week ( $x$ ) and their score on a unit quiz ( $y$ ):

| Hours $x$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  |
|-----------|----|----|----|----|----|----|----|----|
| Score $y$ | 55 | 58 | 64 | 66 | 73 | 75 | 82 | 84 |

The summary statistics are  $\bar{x} = 4.5$ ,  $s_x \approx 2.449$ ,  $\bar{y} \approx 69.625$ ,  $s_y \approx 10.649$  and  $r \approx 0.994$ . Then  $b = 0.994 \cdot \frac{10.649}{2.449} \approx 4.32$  and  $a = 69.625 - 4.32(4.5) \approx 50.18$ , so  $\hat{y} \approx 50.18 + 4.32x$ .



Quiz score against hours studied with the least-squares line. The line passes through (4.5, 69.6).

Notice how the points sit close to the line: that is what a correlation near 1 looks like. The slope says each extra hour of study goes with about 4.3 more quiz points, on average. The intercept 50.18 says a student who studies 0 hours is predicted to score about 50, which is plausible here, so it has a real meaning.

## Worked Examples

### EXAMPLE 1: INTERPRET SLOPE AND INTERCEPT

A gardener measures a bean plant. The model is  $\hat{y} = 2.5 + 1.8x$ , where  $x$  is days since planting and  $y$  is height in centimeters. Interpret the slope and the intercept.

#### SOLUTION

**Slope:** for each additional day since planting, the predicted height increases by 1.8 cm, on average.

**Intercept:** on day 0 (planting day) the predicted height is 2.5 cm. This is meaningful because  $x = 0$  is a sensible value, though the true height at planting is probably smaller.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

**EXAMPLE 2: PREDICT AND FIND A RESIDUAL**

Use  $\hat{y} = 2.5 + 1.8x$  from Example 1. (a) Predict the height on day 10. (b) On day 10 the plant was actually 22.4 cm tall. Find and interpret the residual.

**SOLUTION**

(a)  $\hat{y} = 2.5 + 1.8(10) = 20.5$  cm.

(b) Residual =  $y - \hat{y} = 22.4 - 20.5 = 1.9$  cm.

The plant was 1.9 cm taller than the model predicted, so the point lies above the line.

**EXAMPLE 3: BUILD THE LINE FROM SUMMARY STATISTICS**

Practice hours  $x$  and score  $y$  have  $\bar{x} = 5$ ,  $s_x = 2$ ,  $\bar{y} = 70$ ,  $s_y = 8$  and  $r = 0.9$ . Find the LSRL and predict  $y$  for  $x = 7$ .

**SOLUTION**

Slope:  $b = r \frac{s_y}{s_x} = 0.9 \cdot \frac{8}{2} = 3.6$ .

Intercept:  $a = \bar{y} - b\bar{x} = 70 - 3.6(5) = 52$ . So  $\hat{y} = 52 + 3.6x$ .

Prediction:  $\hat{y} = 52 + 3.6(7) = 77.2$ .

**EXAMPLE 4: LEAST SQUARES BY HAND**

Find the LSRL for the points (1, 3), (2, 5), (3, 4), (4, 7), (5, 8).

**SOLUTION**

Means:  $\bar{x} = 3$ ,  $\bar{y} = 5.4$ .

Deviations:  $x - \bar{x} = -2, -1, 0, 1, 2$  and  $y - \bar{y} = -2.4, -0.4, -1.4, 1.6, 2.6$ .

$\sum(x - \bar{x})(y - \bar{y}) = 4.8 + 0.4 + 0 + 1.6 + 5.2 = 12$  and  $\sum(x - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$ .

Slope  $b = \frac{12}{10} = 1.2$  and intercept  $a = 5.4 - 1.2(3) = 1.8$ . So  $\hat{y} = 1.8 + 1.2x$ .

**EXAMPLE 5: SPOT BAD EXTRAPOLATION**

A model fitted to children aged 5 to 12 is  $\hat{y} = 80 + 6x$ , with  $x$  the age in years and  $y$  the height in cm. Predict the height of a 40-year-old and comment.

**SOLUTION**

$\hat{y} = 80 + 6(40) = 320$  cm, which is over 10 feet tall.

The age 40 lies far outside the data range of 5 to 12. People stop growing, so the straight-line pattern does not continue. This is extrapolation and the prediction is meaningless.

**EXAMPLE 6: WORK BACKWARD FROM THE LINE**

A regression gives  $\hat{y} = 12 + 2.5x$ . The data have  $\bar{x} = 8$ ,  $s_x = 3$  and  $s_y = 9$ . Find  $\bar{y}$  and  $r$ .

**SOLUTION**

The line passes through  $(\bar{x}, \bar{y})$ , so  $\bar{y} = 12 + 2.5(8) = 32$ .

From  $b = r \frac{s_y}{s_x}$ :  $2.5 = r \cdot \frac{9}{3}$ , so  $r = \frac{2.5}{3} \approx 0.833$ .

The positive slope matches the positive correlation, which is a good check.

**Coach's Playbook**

- 1 Label the variables first.** Decide which is explanatory ( $x$ ) and which is the response ( $y$ ) before any arithmetic.
- 2 Use the two anchor formulas.** Slope from  $r s_y / s_x$ , then intercept from  $\bar{y} - b\bar{x}$ . Do the slope first, always.
- 3 Check the point of means.** Plug  $\bar{x}$  into your line. If you do not get  $\bar{y}$ , you have an arithmetic error.
- 4 Check the sign.** The sign of the slope must match the direction of the scatterplot and the sign of  $r$ .
- 5 Write interpretations in a sentence.** Name the variables with units, say predicted and on average, and compare to the data range.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

- ◆ A mental model: the slope is a rate (change in  $y$  per 1 unit of  $x$ ), the intercept is a starting value.
- ◆ Estimate before you compute. If  $r$  is near 1 and  $s_y$  is about 4 times  $s_x$ , expect a slope near 4.
- ◆ Before predicting, ask: is this  $x$ -value inside the data range? If not, say so.

### Common traps

1) Writing the slope as  $r s_x/s_y$  (upside down). 2) Saying the intercept is always meaningful, even when  $x = 0$  is far outside the data. 3) Interpreting the slope as causing the change, or as exact for every individual. 4) Predicting  $x$  from  $y$  by rearranging the line instead of fitting a new regression.

## Mistake Clinic

### Spot the error 1

A student reads  $\hat{y} = 52 + 3.6x$  (score against hours) and writes: 'Every hour of studying raises your score by exactly 3.6 points.' What is wrong? The statement treats the slope as an exact rule for each person. The slope describes the **predicted change on average**, and the line is only a model with scatter around it. Correct: 'For each additional hour studied, the predicted score increases by 3.6 points, on average.'

### Spot the error 2

Given  $r = 0.9$ ,  $s_x = 2$ ,  $s_y = 8$ ,  $\bar{x} = 5$ ,  $\bar{y} = 70$ , a student computes  $b = 0.9 \cdot \frac{2}{8} = 0.225$ . The ratio is upside down. The slope must be  $r \frac{s_y}{s_x}$  because  $y$  is the response:  $b = 0.9 \cdot \frac{8}{2} = 3.6$ . A quick sense check:  $y$  varies much more than  $x$ , so a 1-unit change in  $x$  should move  $y$  by a lot more than 1 unit.

### Spot the error 3

With  $\hat{y} = 52 + 3.6x$ , a student predicts the hours needed for a score of 80 by solving  $80 = 52 + 3.6x$ , getting  $x \approx 7.78$ . The line predicts score from hours, not hours from score. To predict  $x$  from  $y$  fit a new regression with  $y$  explanatory:  $b' = r \frac{s_x}{s_y} = 0.9 \cdot \frac{2}{8} = 0.225$  and  $a' = 5 - 0.225(70) = -10.75$ . Then  $\hat{x} = -10.75 + 0.225(80) = 7.25$  hours. The two answers differ because  $|r| < 1$ .

### Real-World Connection

A cafe tracks the outdoor temperature  $x$  (degrees Celsius) and cups of iced drink sold  $y$  and finds  $\hat{y} = -35 + 9x$ . Each extra degree goes with about 9 more cups sold, on average. The intercept  $-35$  is not meaningful (negative sales at 0 degrees), which tells the owner not to use the model for cold days far below the data.

### Challenge Corner

A data set has  $\bar{x} = 6$  hours,  $\bar{y} = 40$  points,  $s_x = 2$ ,  $s_y = 10$  and  $r = 0.8$ . (a) Find the LSRL. (b) Rewrite it if  $x$  is measured in minutes instead of hours and every score is raised by 5 bonus points. Solution: (a)  $b = 0.8 \cdot \frac{10}{2} = 4$ ,  $a = 40 - 4(6) = 16$ , so  $\hat{y} = 16 + 4x$ . (b) With minutes  $x' = 60x$  the slope shrinks by 60:  $\frac{4}{60} = \frac{1}{15}$  per minute. Adding 5 to every score adds 5 to the intercept: 21. So  $\hat{y}' = 21 + \frac{1}{15}x'$ . Check:  $x = 6$  hours is  $x' = 360$ , giving  $21 + 24 = 45 = 40 + 5$ , as it must.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

## Practice

### PRACTICE Linear Regression in Depth

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- A model for a basketball player is  $\hat{y} = 2.1 + 0.4x$ , where  $x$  is minutes played and  $y$  is points scored. Which statement interprets the slope?
  - Each additional minute played is associated with **0.4** more predicted points, on average.
  - Each additional point requires **0.4** more minutes, exactly.
  - A player who plays **0** minutes scores **0.4** points.
  - Playing **0.4** minutes causes the player to score **1** point.
- The LSRL for roast weight  $x$  (kg) and cooking time  $y$  (minutes) is  $\hat{y} = 30 + 55x$ . What is the predicted time for a 2.4 kg roast?
  - 85** minutes
  - 132** minutes
  - 192** minutes
  - 162** minutes
- For a data set  $r = -0.6$ ,  $s_x = 5$  and  $s_y = 10$ . What is the slope of the LSRL?
  - 0.3**
  - 1.2**
  - 1.2**
  - 0.6**
- A data set has  $\bar{x} = 7$ ,  $\bar{y} = 40$  and slope  $b = 2$ . What is the intercept  $a$ ?
  - 14**
  - 54**
  - 26**
  - 26**
- A student scored 58 but the model predicted 64.5. What is the residual?
  - 6.5**
  - 6.5**
  - 64.5**
  - 0.9**
- A regression uses data with  $x$  between 10 and 30. For which prediction would extrapolation be a concern?
  - $x = 45$**
  - $x = 12$**
  - $x = 25$**
  - $x = 18$**
- The LSRL is  $\hat{y} = 3 + 0.5x$  and the mean of  $x$  is 10. What must be the mean of  $y$ ?
  - 5**
  - 3.5**
  - 8**
  - 13**
- The LSRL is  $\hat{y} = 20 + 6x$  with  $x$  in hours. Which is the LSRL if  $x$  is measured in minutes?
  - $\hat{y} = 20 + 360x$**
  - $\hat{y} = 1200 + 6x$**
  - $\hat{y} = 0.33 + 6x$**
  - $\hat{y} = 20 + 0.1x$**
- Which description best states what the least-squares criterion does?
  - It chooses the line through the most data points.
  - It chooses the line that minimizes the sum of the squared residuals.
  - It makes the sum of the residuals as large as possible.
  - It chooses the line that minimizes the largest single residual.
- A line has slope 1.5 and the data have  $s_x = 4$ ,  $s_y = 8$ . Find  $r$ .
- Find the LSRL for (2, 5), (4, 9), (6, 10), (8, 16).
- Given  $\bar{x} = 14$ ,  $s_x = 3.5$ ,  $\bar{y} = 82$ ,  $s_y = 7$ ,  $r = -0.8$ , predict  $y$  when  $x = 10$ .
- A data set has  $r = 0.6$ ,  $s_x = 5$ ,  $s_y = 10$ ,  $\bar{x} = 20$ ,  $\bar{y} = 50$ . Using a regression that treats  $y$  as explanatory, predict  $x$  when  $y = 65$ .
- The LSRL is  $\hat{y} = 8 + 0.4x$  with  $x$  in minutes. Write the LSRL with  $x$  in hours.
- A line  $\hat{y} = 40 + 2x$  has  $\bar{x} = 15$  and  $s_y = 4s_x$ . Find  $\bar{y}$  and  $r$ .

### 35.2 Residual Analysis and Influential Points

Learn to read residual plots, judge fit with  $r$  squared and the standard deviation of residuals, and tell outliers, high-leverage points and influential points apart.

#### Why This Matters

Fitting a line is easy. Knowing whether the line deserves your trust is the harder skill. Two data sets can give almost the same  $r$  and the same equation while one is a healthy linear pattern and the other is a curve that a line misrepresents. Students get stuck because they read a high  $r$  as proof that a linear model is right, they confuse a point that is unusual with a point that is **influential**, and they say ' $r^2$  is the percent of points on the line,' which is not what it means.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

Residual analysis is how scientists, engineers and analysts check their models before using them to make decisions: calibrating an instrument, forecasting demand, estimating costs. The tools in this section, residual plots, the standard deviation of residuals  $s$ ,  $r^2$ , and the ideas of outlier and leverage, are exactly the checks used in every statistics course that follows.

Core Ideas

**Residuals and the residual plot**

A **residual** is  $e = y - \hat{y}$ . A **residual plot** graphs the residuals (vertical axis) against  $x$  or against  $\hat{y}$  (horizontal axis).

A linear model is appropriate when the residual plot shows **no clear pattern**: a random-looking cloud centered on the horizontal line at 0 with roughly constant spread. A curve, a U-shape, or a fan shape (spread growing or shrinking) signals a problem.

For an LSRL with an intercept, the residuals always add to 0.

- ◆ **Curved pattern** in the residuals: a line is the wrong shape; try a transformation or a nonlinear model (next section).
- ◆ **Fan shape**: the spread of the response changes with  $x$ , so predictions are more precise in some places than others.
- ◆ **Random scatter**: no evidence against linearity. This does not prove the relationship is linear; it simply shows no problem.

**Measuring fit:  $r^2$  and  $s$**

The **coefficient of determination**  $r^2$  is the fraction of the variation in  $y$  that is accounted for by the linear relationship with  $x$ . Equivalently,  $r^2 = 1 - \frac{SSE}{SST}$ , where  $SSE = \sum (y - \hat{y})^2$  and  $SST = \sum (y - \bar{y})^2$ .

The **standard deviation of the residuals** is  $s = \sqrt{\frac{SSE}{n - 2}}$ . It is the typical size of a prediction error, in the units of  $y$ .

The word to remember in interpretation of  $r^2$  is **variation**. If  $r^2 = 0.84$ , then about 84% of the variation in  $y$  is accounted for by the linear relationship with  $x$ , and about 16% is left over as scatter. The sign of  $r$  is lost when you square it, so find the sign from the slope.

**Outliers, leverage and influence**

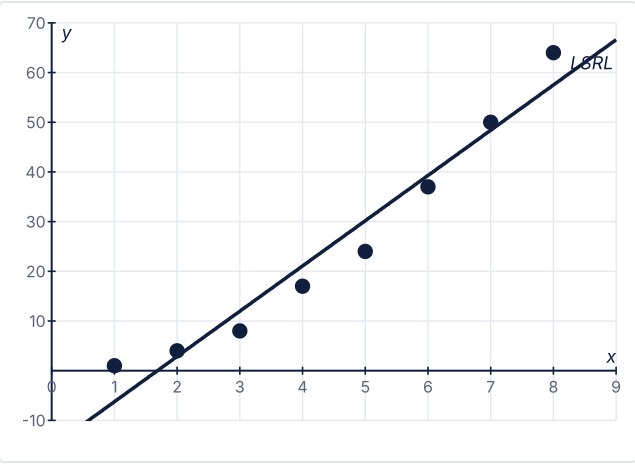
An **outlier** has a large residual (it is far above or below the pattern). A point has **high leverage** if its  $x$ -value is far from  $\bar{x}$ . A point is **influential** if removing it would noticeably change the slope, the intercept or  $r$ .

High-leverage points that break the pattern are usually influential. An outlier in the middle of the  $x$ -range changes the intercept and weakens  $r$  but barely moves the slope.

Here are two data sets whose fits look strong. The first has a nearly linear pattern; the second,  $y$  against  $x = 1, \dots, 8$ , grows like a square:

| $x$         | 1 | 2 | 3 | 4  | 5  | 6  | 7  | 8  |
|-------------|---|---|---|----|----|----|----|----|
| $y$ (set 2) | 1 | 4 | 8 | 17 | 24 | 37 | 50 | 64 |

For set 2 the LSRL is  $\hat{y} \approx -15.36 + 9.11x$  with  $r \approx 0.976$  and  $r^2 \approx 0.952$ . That sounds excellent, yet the residuals are 7.25, 1.14, -3.96, -4.07, -6.18, -2.29, 1.61, 6.50: positive, then negative, then positive. That U-shape is the fingerprint of a curved relationship.

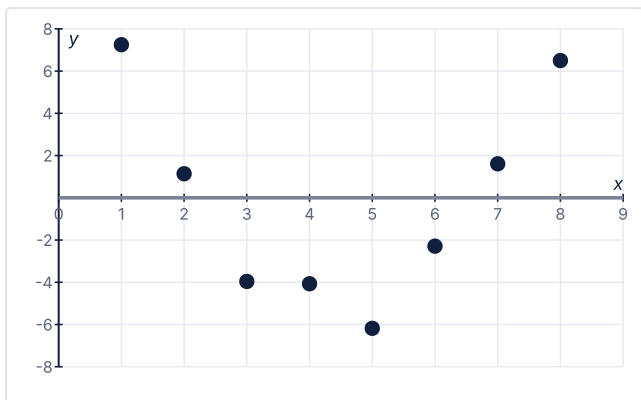


Set 2 with its least-squares line. The points curve above the line at both ends.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →



Residual plot for set 2. The U-shape says a straight line is the wrong model even though  $r$  is large.

## Worked Examples

### EXAMPLE 1: COMPUTE AND CHECK RESIDUALS

For the points  $(1, 3)$ ,  $(2, 5)$ ,  $(3, 4)$ ,  $(4, 7)$ ,  $(5, 8)$  the LSRL is  $\hat{y} = 1.8 + 1.2x$ . Find all residuals and their sum.

#### SOLUTION

Predictions:  $\hat{y} = 3.0, 4.2, 5.4, 6.6, 7.8$ .

Residuals  $y - \hat{y}$ :  $0, 0.8, -1.4, 0.4, 0.2$ .

Sum:  $0 + 0.8 - 1.4 + 0.4 + 0.2 = 0$ , as it must be for an LSRL.

### EXAMPLE 2: READ A RESIDUAL PLOT

A residual plot for a line fitted to set 2 above shows residuals that are positive at the smallest and largest  $x$  and negative in the middle. Is a linear model appropriate?

#### SOLUTION

The pattern is a U-shape, not random scatter.

This means the data curve away from the line. Even with  $r^2 \approx 0.95$ , the linear model is not appropriate; a curved model (such as a quadratic, or a transformation) is better.

### EXAMPLE 3: INTERPRET $R^2$

For a regression of fuel used (liters) on distance (km),  $r = -0.80$ . Find and interpret  $r^2$ .

#### SOLUTION

$$r^2 = (-0.80)^2 = 0.64.$$

About 64% of the variation in fuel used is accounted for by the linear relationship with distance. The remaining 36% is due to other factors and random scatter.

### EXAMPLE 4: STANDARD DEVIATION OF RESIDUALS

Use the residuals from Example 1 to find  $s$  and  $r^2$ . Here  $SST = 17.2$ .

#### SOLUTION

$$SSE = 0^2 + 0.8^2 + (-1.4)^2 + 0.4^2 + 0.2^2 = 0.64 + 1.96 + 0.16 + 0.04 = 2.8.$$

$$s = \sqrt{\frac{2.8}{5-2}} \approx 0.966. \text{ A typical prediction is off by about 0.97 units of } y.$$

$$r^2 = 1 - \frac{2.8}{17.2} \approx 0.837.$$

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

**EXAMPLE 5: INFLUENTIAL POINT**

Add the point (12, 7) to the data of Example 1. The new LSRL is  $\hat{y} \approx 4.27 + 0.310x$  and  $r \approx 0.62$  (before: slope 1.2,  $r \approx 0.915$ ). Is the new point influential?

**SOLUTION**

Its  $x = 12$  is far from  $\bar{x} = 3$ , so it has high leverage.

The old line predicts  $1.8 + 1.2(12) = 16.2$  at  $x = 12$ , yet the point is at 7, a residual of  $-9.2$ .

Adding it cut the slope from 1.2 to 0.31 and  $r$  from 0.915 to 0.62. It is an influential point.

**EXAMPLE 6: RECOVER  $R$  FROM SUMS OF SQUARES**

A regression with a negative slope has  $SST = 250$  and  $SSE = 40$ . Find  $r$ .

**SOLUTION**

$$r^2 = 1 - \frac{40}{250} = 0.84.$$

$|r| = \sqrt{0.84} \approx 0.917$ . The slope is negative, so  $r \approx -0.917$ .

**Coach's Playbook**

- 1 Plot first, compute second.** Always look at the scatterplot and the residual plot before trusting  $r$  or  $r^2$ .
  - 2 Describe the residual plot in words.** Say pattern or no pattern, and say what the pattern means (curve, fan).
  - 3 Separate three ideas.** Outlier = large residual. Leverage = far-away  $x$ . Influential = changes the fit when removed.
  - 4 Test influence by thought experiment.** Imagine removing the point. Would the line tilt a lot? If yes, it is influential.
  - 5 Interpret with the word variation.** ' $r^2$  percent of the variation in  $y$  is accounted for by the linear relationship with  $x$ .'
- ◆ Sum of residuals = 0 is a free arithmetic check for your residual column.
  - ◆ Compare  $s$  to the size of  $y$ : an  $s$  of 2 is small when  $y$  is in the hundreds and large when  $y$  is near 10.
  - ◆ An influential point should be investigated (recording error? real extreme case?), not automatically deleted.

**Common traps**

1) Saying a high  $r$  or  $r^2$  proves the relationship is linear. 2) Saying  $r^2$  is the percent of points on the line, or the percent of predictions that are correct. 3) Calling every outlier influential: an outlier near  $\bar{x}$  barely changes the slope. 4) Forgetting that  $r$  takes the sign of the slope and  $r^2$  does not.

**Mistake Clinic****Spot the error 1**

A student sees  $r = 0.976$  for set 2 and writes: 'The correlation is very high, so a linear model is appropriate.' The conclusion is not justified. The residual plot shows a U-shape, which means the relationship is curved. Correlation measures the strength of the **linear** association only; a clear curve can still have a high  $r$ . Correct: 'The residual plot has a pattern, so a linear model is not appropriate despite  $r = 0.976$ .'

**Spot the error 2**

A student says: 'Since  $r^2 = 0.64$ , the model predicts 64% of the data points exactly.' Wrong meaning. Predictions from a regression are almost never exact. Correct: '64% of the variation in the response is accounted for by the linear relationship with the explanatory variable.' The other 36% is variation the line does not explain.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

**Spot the error 3**

Adding (3, 12) to the points (1, 3), (2, 5), (3, 4), (4, 7), (5, 8), a student says: 'It is an outlier with a huge residual, so it must have changed the slope a lot.' Check: the new LSRL is  $\hat{y} = 2.9 + 1.2x$ , the slope is still 1.2. The point has  $x = 3 = \bar{x}$ , so it has no leverage on the slope. It only raises the intercept (by moving the line up) and lowers  $r$  from 0.915 to 0.519. An outlier is not the same as an influential point.

**Real-World Connection**

A lab calibrates a thermometer by comparing its readings with a trusted standard at 8 temperatures. The residual plot shows the residuals fanning out as the temperature rises. Engineers conclude the thermometer is less precise at high temperatures, which a single  $r^2$  value would hide. Checking residuals protects decisions that depend on predictions.

**Challenge Corner**

The points (1, 3), (2, 5), (3, 4), (4, 7), (5, 8) have LSRL  $\hat{y} = 1.8 + 1.2x$  with  $\bar{x} = 3$  and  $\bar{y} = 5.4$ . Without redoing every sum, find the new LSRL after adding (3, 12). Solution: the new point has  $x = \bar{x}$ , so the new  $\bar{x}$  is still 3. Its deviation in  $x$  is 0, so it adds nothing to  $\sum(x - \bar{x})(y - \bar{y}) = 12$  or to  $\sum(x - \bar{x})^2 = 10$ . The slope stays 1.2. The new  $\bar{y} = \frac{27+12}{6} = 6.5$ . The line passes through (3, 6.5), so  $a = 6.5 - 1.2(3) = 2.9$  and  $\hat{y} = 2.9 + 1.2x$ .

**Practice****PRACTICE Residual Analysis and Influential Points**

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- A residual plot shows points scattered randomly above and below 0 with no pattern and roughly constant spread. What does this suggest?
  - A linear model is appropriate for the data.
  - The correlation must be exactly 0.
  - The data contain no outliers.
  - The relationship is proven to be linear.
- The residual plot for a regression shows a clear U-shape. Which conclusion is best?
  - The line is perfect because residuals are symmetric.
  - The correlation must be negative.
  - A curved model would fit better than a line.
  - The slope must be 0.
- For a regression,  $r = -0.9$ . Which statement is correct?
  - About 90% of the variation in  $y$  is accounted for by  $x$ .
  - About 81% of the variation in  $y$  is accounted for by the linear relationship with  $x$ .
  - About 81% of the points lie on the line.
  - The model predicts  $y$  correctly 81% of the time.
- The LSRL is  $\hat{y} = 5 + 2x$ . What is the residual for the point (4, 15)?
  - 2
  - 13
  - 15
  - 2
- A point lies far to the right of all the other points in  $x$ , but it falls almost exactly on the line suggested by the others. How should it be described?
  - High leverage, but probably not influential.
  - An outlier with a large residual.
  - Influential because it has a large residual.
  - A point with low leverage.
- A point is added in the middle of the  $x$ -range, far above the pattern. Which effect is most likely?
  - The slope changes a lot and  $r$  increases.
  - The slope changes little, the intercept increases, and  $r$  decreases.
  - Nothing changes.
  - The slope becomes negative.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

7. A regression has  $SSE = 36$  and  $n = 11$ . What is  $s$ , the standard deviation of the residuals?  
 (A) 6 (B)  $\sqrt{3.6} \approx 1.9$  (C) 2 (D) 4
8. A regression has  $SST = 200$  and  $SSE = 50$ . What is  $r^2$ ?  
 (A) 0.25 (B) 0.5 (C) 4 (D) 0.75
9. Data: (1, 3), (2, 5), (3, 4), (4, 7), (5, 8). Which single added point would change the slope of the LSRL the most?  
 (A) (3, 6) (B) (12, 7) (C) (3, 12) (D) (12, 16)
10. The LSRL is  $\hat{y} = 4 + 1.5x$ . Find the residual for the point (6, 16).
11. For a regression,  $r = 0.6$ . What percent of the variation in  $y$  is NOT accounted for by the linear relationship with  $x$ ?
12. A line fitted to 5 points leaves residuals 2, -1, 3, -2, -2. Find  $s$  to three decimals.
13. A data set has  $n = 12$ ,  $s_y = 10$  and  $r = -0.8$ . Find  $SSE$ .
14. A regression has  $r^2 = 0.81$  and a negative slope. Find  $r$ .
15. Find the slope of the LSRL for (1, 3), (2, 5), (3, 4), (4, 7), (5, 8), (12, 7) to three decimals.

### 35.3 Nonlinear Models and Transformations

Use logarithms to straighten exponential and power patterns, fit and interpret the transformed line, convert back to the original variables, and choose between competing models.

#### Why This Matters

Many real relationships are not straight lines. Populations and savings grow by a percent each period, so they follow an exponential pattern. The period of a swinging pendulum, the strength of a signal at a distance, and the area of a shape as its size changes follow power patterns. A student who knows only the straight-line tools sees a poor fit and gives up. The key idea of this section is that a **logarithm can turn a curve into a line**, so all the regression tools you already know can be used.

Students get stuck in three places: deciding which variable to take logs of, converting the fitted line back to the original variables (the step where they forget to exponentiate), and choosing between models. Scientists use these transformations to find growth rates, to test scaling laws, and to make forecasts from experimental data, so this skill pays off in biology, physics, economics and computer science.

#### Core Ideas

##### Exponential model

If  $y$  grows or shrinks by a constant **factor** each unit of  $x$ , then  $y = A b^x$ . Taking logarithms gives a straight line:

$$\ln y = \ln A + (\ln b) x.$$

Plot  $\ln y$  against  $x$ . If that plot is linear, fit an LSRL  $\widehat{\ln y} = a + cx$  and convert back with  $\hat{y} = e^a (e^c)^x$ . The growth factor is  $e^c$  per unit of  $x$ .

##### Power model

If  $y$  is proportional to a power of  $x$ , then  $y = A x^p$ . Taking logarithms of both variables gives

$$\ln y = \ln A + p \ln x.$$

Plot  $\ln y$  against  $\ln x$ . If that plot is linear, the slope is the exponent  $p$  and the intercept is  $\ln A$ . Convert back with  $\hat{y} = e^a x^p$ .

Either natural logarithms or base-10 logarithms work, as long as you use the same base to go back: with base 10 write  $\hat{y} = 10^a \cdot (10^c)^x$  or  $\hat{y} = 10^a x^p$ .

| Pattern in the data                               | Transform        | Model after converting back |
|---------------------------------------------------|------------------|-----------------------------|
| Constant ratio of $y$ for equal steps in $x$      | $(x, \ln y)$     | $\hat{y} = A b^x$           |
| $y$ scales with a power of $x$                    | $(\ln x, \ln y)$ | $\hat{y} = A x^p$           |
| Constant difference of $y$ for equal steps in $x$ | none needed      | $\hat{y} = a + bx$          |

Here is an example. A culture of cells is counted each hour:

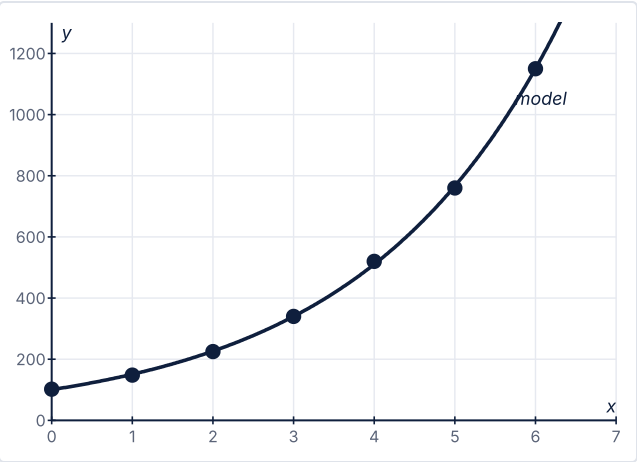
Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

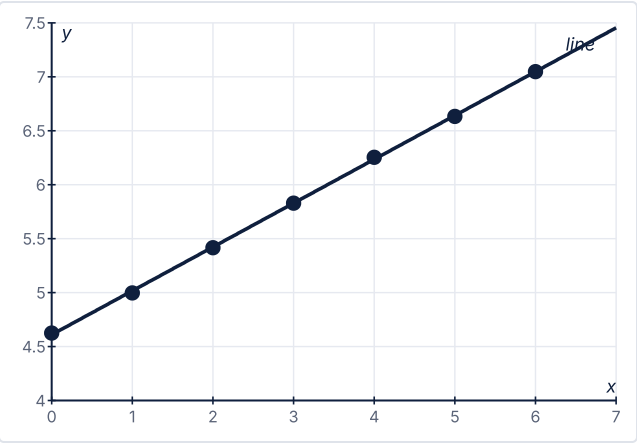
**BOOK NOW →**

| Hour $t$  | 0     | 1     | 2     | 3     | 4     | 5     | 6     |
|-----------|-------|-------|-------|-------|-------|-------|-------|
| Count $y$ | 102   | 148   | 225   | 340   | 520   | 760   | 1150  |
| $\ln y$   | 4.625 | 4.997 | 5.416 | 5.829 | 6.254 | 6.633 | 7.048 |

The counts rise faster and faster, but the values of  $\ln y$  go up by roughly 0.4 each hour: a line. The LSRL of  $\ln y$  on  $t$  is  $\widehat{\ln y} \approx 4.610 + 0.4063t$ , with  $r^2 \approx 0.9998$ . Converting back,  $\hat{y} \approx e^{4.610}(e^{0.4063})^t \approx 100.5(1.501)^t$ : the culture grows by about 50% each hour.



Cell counts and the exponential model fitted through the log transformation.



After taking  $\ln$  of the counts, the points fall on a line.

Choosing a model

Fit the candidate model, then judge it with (1) the **residual plot** (should show no pattern), (2)  $r^2$  (high), and (3) context (does the model make sense?). A high  $r^2$  alone is never enough, and  $r^2$  computed on a transformed scale describes variation in  $\ln y$ , not in  $y$ .

Worked Examples

EXAMPLE 1: CONVERT FROM THE LOG SCALE

A log-transformed fit is  $\widehat{\ln y} = 2.30 + 0.405x$ . Write the model for  $y$  and interpret the growth.

SOLUTION

$\hat{y} = e^{2.30}(e^{0.405})^x \approx 9.97(1.499)^x$ .

Each increase of 1 in  $x$  multiplies the predicted  $y$  by about 1.5, a 50% increase.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

**EXAMPLE 2: PREDICT FROM A POWER MODEL**

A fit with base-10 logs is  $\widehat{\log y} = 0.3 + 1.5 \log x$ . Predict  $y$  when  $x = 4$ .

**SOLUTION**

$$\hat{y} = 10^{0.3} x^{1.5} \approx 1.995 x^{1.5}.$$

For  $x = 4$ ,  $x^{1.5} = 4^{1.5} = 8$ , so  $\hat{y} \approx 1.995(8) \approx 15.96$ .

**EXAMPLE 3: RECOGNIZE EXPONENTIAL DATA**

Is the data  $x = 1, 2, 3, 4, 5$  and  $y = 3, 6, 12, 24, 48$  linear, exponential or neither? Find the model.

**SOLUTION**

Differences are 3, 6, 12, 24: not constant, so not linear.

Ratios are  $6/3 = 12/6 = 24/12 = 48/24 = 2$ : constant, so exponential with  $b = 2$ .

$y = A \cdot 2^x$ ; at  $x = 1$ :  $3 = 2A$ , so  $A = 1.5$  and  $y = 1.5(2)^x$ . Equivalent:  $\ln y$  rises by  $\ln 2$  per step.

**EXAMPLE 4: CHOOSE BETWEEN A LINE AND A TRANSFORMATION**

For the cell data, the linear fit is  $\hat{y} \approx -36.0 + 166.5t$  with  $r^2 \approx 0.898$  and residuals 138, 17.5, -72, -123.6, -110.1, -36.6, 186.8. Compare with the log model ( $r^2 \approx 0.9998$ , residuals within  $\pm 0.02$  of 0 with no pattern). Which model should we use?

**SOLUTION**

The linear residuals are positive, then negative, then positive: a clear U-shape, so the line is the wrong shape even though  $r^2 \approx 0.90$  is fairly high.

The log model has tiny, patternless residuals and a much higher  $r^2$ .

Use the exponential model.

**EXAMPLE 5: PREDICT AND JUDGE**

Use  $\hat{y} = 100.47 e^{0.40634t}$  to predict the count at  $t = 8$  hours. Is this safe?

**SOLUTION**

$$\hat{y} = 100.47 e^{0.40634(8)} = 100.47 e^{3.2507} \approx 2593.$$

The data go only to  $t = 6$ , so  $t = 8$  is mild extrapolation. In real cultures food runs out and growth slows, so treat the number as a rough estimate.

**EXAMPLE 6: FIND A POWER MODEL FROM TWO POINTS**

A relationship  $y = Ax^p$  passes through  $(2, 12)$  and  $(8, 96)$ . Find  $A$  and  $p$ .

**SOLUTION**

Divide:  $\frac{96}{12} = \left(\frac{8}{2}\right)^p$ , so  $8 = 4^p$ , which gives  $p = \frac{\ln 8}{\ln 4} = \frac{3}{2}$ .

Then  $12 = A \cdot 2^{1.5} = A \cdot 2\sqrt{2}$ , so  $A = \frac{12}{2\sqrt{2}} = 3\sqrt{2} \approx 4.243$ .

Check at  $x = 8$ :  $4.243 \cdot 8^{1.5} = 4.243 \cdot 22.627 = 96.0$ .

**Coach's Playbook**

- 1 Name the pattern.** Constant ratio for equal steps in  $x$  means exponential. Straight line only after logging both axes means power.
- 2 Transform and plot.** Make the new scatterplot and fit the LSRL on the transformed variables.
- 3 Check the residuals.** Look at the residual plot of the transformed fit, not the original one.
- 4 Convert back carefully.** Exponentiate the entire prediction:  $\hat{y} = e^{\widehat{\ln y}}$ .
- 5 Interpret in original units.** Say 'multiplies by about 1.5 each hour' or 'y scales like x to the 1.5', not 'ln y goes up 0.4'.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

- ◆ Exponential growth factor  $b = e^c$ ; percent change  $= (b - 1) \cdot 100\%$ .
- ◆ If doubling  $x$  doubles  $y$  the power is 1; if doubling  $x$  quadruples  $y$  the power is 2 (since  $2^2$  is the factor).
- ◆ A calculator check:  $e^{\ln v} = v$ , so a quick test of your conversion is to plug in one data point.

### Common traps

1) Forgetting to exponentiate: a fitted value of  $\ln \hat{y} = 2.7$  means  $\hat{y} = e^{2.7} \approx 14.9$ , not 2.7. 2) Taking logs of only  $y$  when the pattern is a power (needs both). 3) Choosing a model by  $r^2$  alone while ignoring a patterned residual plot. 4) Using the wrong base in the conversion, such as  $e^a$  after fitting with base-10 logs.

## Mistake Clinic

### Spot the error 1

A fitted model is  $\widehat{\ln y} = 1.2 + 0.3x$ . A student predicts  $y$  at  $x = 5$  by computing  $1.2 + 0.3(5) = 2.7$  and answering  $\hat{y} = 2.7$ . The value 2.7 is a prediction of  $\ln y$ , not of  $y$ . Exponentiate:  $\hat{y} = e^{2.7} \approx 14.88$ .

### Spot the error 2

For the cell data a student fits the line  $\hat{y} = -36.0 + 166.5t$ , sees  $r^2 \approx 0.898$ , and predicts  $\hat{y}(8) \approx 1296$ . The residual plot is U-shaped, so the line is the wrong model, and the exponential model gives about 2593 at  $t = 8$ . The line also predicts a negative count of about  $-36$  at  $t = 0$  when the actual count is 102. A high  $r^2$  never replaces a look at the residual plot.

### Spot the error 3

The data satisfy  $y = 2$  at  $x = 1$  and  $y = 8$  at  $x = 2$ . A student says the power is  $p = \frac{8}{2} = 4$  because  $y$  became 4 times larger. The factor 4 equals  $2^2$  because  $x$  doubled, so  $2^p = 4$  and  $p = 2$ . Check with  $y = 2x^2$ : at  $x = 2$ ,  $y = 2 \cdot 4 = 8$ . Correct.

### Real-World Connection

A student times a pendulum with six string lengths  $L$  (meters) and records the swing time  $T$  (seconds). Taking logs of both variables gives the line  $\widehat{\ln T} \approx 0.697 + 0.5015 \ln L$ , so  $\hat{T} \approx 2.007 L^{0.5015}$ : the time grows like the square root of the length, matching the theory. For  $L = 2.5$  m the model predicts  $T \approx 3.18$  seconds. The slope of a log-log plot reveals the exponent in a physical law.

### Challenge Corner

Using  $\hat{y} = 100.47 e^{0.40634t}$  for the cell culture, (a) find the doubling time, and (b) find when the model reaches 5000 cells.

Solution: (a) Doubling means  $e^{0.40634T} = 2$ , so  $T = \frac{\ln 2}{0.40634} \approx 1.706$  hours, independent of the starting count. (b)

$100.47 e^{0.40634t} = 5000$  gives  $e^{0.40634t} = 49.77$ , so  $t = \frac{\ln 49.77}{0.40634} \approx 9.62$  hours, which is extrapolation and a rough estimate.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

## Practice

**PRACTICE** Nonlinear Models and Transformations

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- Data show a constant ratio between successive  $y$ -values when  $x$  increases by 1. Which plot should look linear?
 

(A)  $\ln y$  against  $x$                       (B)  $y$  against  $x$                       (C)  $\ln y$  against  $\ln x$   
 (D)  $y$  against  $\ln x$
- Which transformation linearizes a power relationship  $y = Ax^p$ ?
 

(A) Plot  $\ln y$  against  $x$  .                      (B) Plot  $\ln y$  against  $\ln x$  .                      (C) Plot  $y$  against  $\ln x$  .  
 (D) Plot  $y^2$  against  $x$  .
- A fit is  $\widehat{\ln y} = 1.2 + 0.3x$ . Which is the model for  $y$ ?
 

(A)  $\hat{y} = 1.2 + 0.3x$                       (B)  $\hat{y} \approx 1.2(0.3)^x$                       (C)  $\hat{y} \approx 3.32(1.350)^x$   
 (D)  $\hat{y} \approx 3.32 + 1.350x$
- A fit is  $\widehat{\ln y} = 2 + 0.5x$ . What is the predicted  $y$  when  $x = 2$ ?
 

(A) 3                      (B)  $e^2 \approx 7.39$                       (C)  $e^4 \approx 54.60$                       (D)  $e^3 \approx 20.09$
- A line fit to curved data leaves a U-shaped residual plot. After using  $\ln y$ , the residual plot looks random and  $r^2$  rises. What should the analyst do?
 

(A) Use the transformed (exponential) model.  
 (B) Keep the line because its residuals are easier to compute.  
 (C) Use the line because a U-shape is normal.  
 (D) Discard both models.
- In a power relationship, doubling  $x$  always makes  $y$  four times as large. What is the power  $p$ ?
 

(A) 4                      (B) 2                      (C) 1                      (D)  $\frac{1}{2}$
- Which table shows exponential growth?
 

(A)  $y : 2, 4, 6, 8$                       (B)  $y : 2, 5, 10, 17$                       (C)  $y : 2, 6, 18, 54$   
 (D)  $y : 2, 3, 5, 8$
- With base-10 logs, a fit is  $\widehat{\log y} = 0.5 + 2 \log x$ . Which is the model for  $y$ ?
 

(A)  $\hat{y} = 0.5 + 2x$                       (B)  $\hat{y} \approx 2x^{0.5}$                       (C)  $\hat{y} \approx 0.5x^2$   
 (D)  $\hat{y} \approx 3.162x^2$
- For a regression of  $\ln y$  on  $x$ ,  $r^2 = 0.97$ . Which interpretation is correct?
 

(A) About 97% of the variation in  $\ln y$  is accounted for by the linear relationship with  $x$  .  
 (B) About 97% of the variation in  $y$  is accounted for by  $x$  in any model.  
 (C) The exponential model predicts 97% of the counts exactly.  
 (D) The growth factor is 0.97 .
- A fit is  $\widehat{\ln y} = 3.2 + 0.25x$ . Find  $\hat{y}$  for  $x = 4$  to two decimals.
- For the model in the previous problem, by what percent does the predicted  $y$  increase for each increase of 1 in  $x$ ? Give one decimal.
- A fit is  $\widehat{\log y} = 0.2 + 1.5 \log x$  (base 10). Predict  $y$  when  $x = 100$  to the nearest whole number.
- The data (1, 5), (2, 10), (3, 20), (4, 40) follow an exponential model  $y = Ab^x$ . Find the model.
- A power model  $y = Ax^p$  passes through (1, 4) and (3, 36). Find the model.
- A log-log plot of  $\ln y$  against  $\ln x$  is a line through (1, 2) and (3, 8). Write the power model for  $y$  and predict  $y$  at  $x = 2$  to two decimals.

**35.4** Correlation vs Causation and Data Ethics

Learn why correlation is not causation, how lurking variables, Simpson's paradox and the ecological fallacy mislead, and how to read studies critically and report results honestly.

## Why This Matters

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

A strong correlation is a fact about a pattern in data. A cause-and-effect claim is a much bigger statement about how the world works, and data alone rarely earn it. Students get stuck here because the leap from 'moves together with' to 'makes happen' feels natural. Three recurring mistakes are treating association as causation, forgetting about a hidden third variable, and stretching a result about groups into a claim about individuals.

Headlines, advertisements, health advice and policy debates are full of statistical claims. The ability to ask 'how was this measured, who was included, and what else could explain it?' is one of the most valuable skills this course gives you. It is also an AP Statistics essential: free-response questions regularly ask you to justify a conclusion in context, state what cannot be concluded, and identify a possible lurking variable.

Core Ideas

Association is not causation

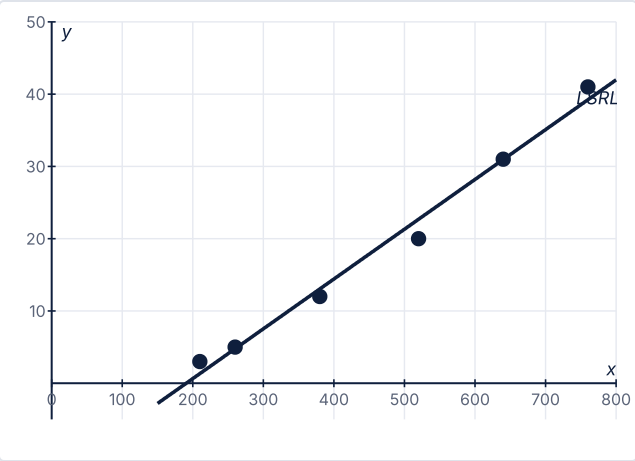
Two variables are **associated** when knowing one helps predict the other. Correlation  $r$  measures only the strength and direction of a **linear** association. It never tells you that a change in  $x$  **causes** a change in  $y$ .

A **lurking variable** is a variable that is not among those studied but affects both. When its effect cannot be separated from the effect of the explanatory variable, it is a **confounding variable**.

Consider a town's monthly data. In six months, ice cream sales (cones) and sunburn cases are:

| Avg. temperature (C) | 8   | 12  | 17  | 22  | 27  | 31  |
|----------------------|-----|-----|-----|-----|-----|-----|
| Cones sold           | 210 | 260 | 380 | 520 | 640 | 760 |
| Sunburn cases        | 3   | 5   | 12  | 20  | 31  | 41  |

The correlation between cones sold and sunburn cases is  $r \approx 0.994$ , extremely strong. But eating ice cream does not burn skin. Temperature is the lurking variable: hot sunny months raise both. In fact the correlation of temperature with cones is 0.996 and with sunburns 0.983.



Sunburn cases against cones sold:  $r = 0.994$ . A hidden variable, temperature, drives both.

When can we claim causation?

A well-designed **randomized comparative experiment** can support cause and effect: subjects are **randomly assigned** to treatment groups, so lurking variables are balanced across groups on average, and there is a **control** or comparison group. Blinding keeps expectations from influencing results.

An **observational study** (no treatments imposed) can show association but cannot by itself establish causation.

- ◆ **Generalizing** requires a random sample from the population. **Causation** requires random assignment. A study can have one, both or neither.
- ◆ Volunteer or convenience samples are likely **biased**, so their results should not be generalized.
- ◆ Strong evidence for causation outside experiments needs many studies, a plausible mechanism, a consistent direction and a dose-response pattern.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

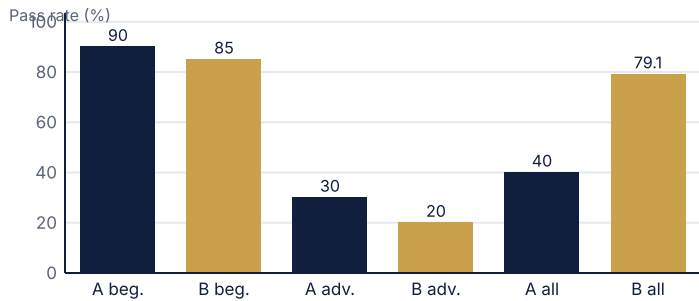
Two traps with groups

**Ecological fallacy:** using a correlation computed from group averages (classes, states, cities) to make claims about individuals. Averages smooth out individual variation, so correlations between averages are usually much stronger than those between individuals.

**Simpson's paradox:** a comparison that holds within every subgroup can reverse when the groups are combined, because the subgroups have very different sizes.

Here is a Simpson's paradox. Two tutoring programs are compared. Program A mostly serves advanced students (a hard group) and Program B mostly beginners. The pass rates are:

| Group     | Program A       | Program B         |
|-----------|-----------------|-------------------|
| Beginners | 18 of 20 (90%)  | 85 of 100 (85%)   |
| Advanced  | 30 of 100 (30%) | 2 of 10 (20%)     |
| Overall   | 48 of 120 (40%) | 87 of 110 (79.1%) |



Program A wins in each group but loses overall: the groups differ in size and difficulty.

Worked Examples

EXAMPLE 1: FIND THE LURKING VARIABLE

A study of a city finds that neighborhoods with more fire trucks sent to fires have larger fire damage. Does sending more trucks cause more damage? Name a lurking variable.

SOLUTION

The correlation is real, but the direction of cause is not as it looks.

Fire **size** is lurking: bigger fires cause both more damage and more trucks sent. No causal claim about trucks follows.

EXAMPLE 2: OBSERVATIONAL OR EXPERIMENT?

(a) A researcher surveys 200 students and finds that those who play an instrument have higher grades. (b) A school randomly assigns 120 volunteer students to 10 minutes of music practice or 10 minutes of quiet reading each morning and compares test gains. Which can support a cause-and-effect conclusion?

SOLUTION

(a) Observational: students chose whether to play. Lurking variables (family resources, motivation) may explain the pattern. Association only.

(b) Randomized comparative experiment: random assignment balances lurking variables, so a causal conclusion is supported, but only for volunteers like these.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

BOOK NOW →

**EXAMPLE 3: RESOLVE A SIMPSON'S PARADOX**

Use the table above. Which program is better?

**SOLUTION**

Within beginners A is better (90% vs 85%) and within advanced A is better (30% vs 20%).

The overall comparison (40% vs 79.1%) reverses because A serves mostly the advanced group, which has a low pass rate for both programs.

Compare within groups. The aggregate rate hides the group sizes; A is the better program for each kind of student.

**EXAMPLE 4: ECOLOGICAL FALLACY**

Four classes have mean weekly reading hours of 4, 7, 10, 13 and mean scores of 72.8, 75.2, 79.2, 83.2, giving  $r \approx 0.994$  for the class means. For the 20 individual students  $r \approx 0.557$ . A headline says: '99% of the variation in student scores is explained by reading time.' Evaluate.

**SOLUTION**

For the class means  $r^2 \approx 0.988$ , so about 99% of the variation in class **means** is accounted for.

For individual students  $r^2 \approx 0.310$ : only about 31% of the variation in student scores.

The headline applies a group-level result to individuals. That is the ecological fallacy.

**EXAMPLE 5: PREDICTION WITHOUT CAUSATION**

For the town data the LSRL of sunburn cases on cones sold is  $\hat{y} \approx -13.14 + 0.0689x$ . Predict sunburn cases when 600 cones are sold and explain what the slope does and does not say.

**SOLUTION**

$\hat{y} = -13.14 + 0.0689(600) \approx 28.2$  cases.

The slope: each additional cone sold goes with about 0.069 more predicted sunburn cases, on average. This is a statement about association; the cause of both is weather. Cutting cone sales would not cut sunburns.

**EXAMPLE 6: WRITE A CAREFUL CONCLUSION**

A survey of 500 adults finds  $r = 0.40$  between cups of tea per day and a health score. Write a conclusion for a report.

**SOLUTION**

$r^2 = 0.16$ , so about 16% of the variation in health scores is accounted for by the linear relationship with tea consumption.

Conclusion: 'In this sample of adults, those who drank more tea tended to have somewhat higher health scores ( $r = 0.40$ ). Because this is an observational survey, we cannot conclude tea improves health; lurking variables such as income, diet and exercise may explain the pattern. The result applies to the population only if the sample was random.'

## Coach's Playbook

- 1 Ask how the data were produced.** Experiment with random assignment, or observational? That alone decides what you may claim.
  - 2 Hunt for the third variable.** What could affect both  $x$  and  $y$ ? Name it specifically (temperature, age, fire size), not just 'something else'.
  - 3 Check the direction.** Could  $y$  be driving  $x$  instead of the other way round?
  - 4 Check who is in the data.** Random sample? Volunteers? Say what group the result applies to.
  - 5 Write the claim with the right verbs.** Use 'is associated with', 'tends to', 'in this sample' for observational data.
- ◆ Cause-and-effect words (causes, leads to, improves, because of) need random assignment to back them.
  - ◆ Always quote  $r$  or  $r^2$  with context: what percent of the variation in which variable.
  - ◆ Honest reporting also means showing the full axis, reporting all analyses you ran, and protecting the privacy of the people in the data.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

**Common traps**

1) Writing 'causes' after seeing a high  $r$ . 2) Naming a lurking variable that only affects one of the two variables. 3) Applying a conclusion from group averages to individuals. 4) Saying random sampling allows a causal conclusion (it allows generalizing; random assignment allows causation).

**Mistake Clinic****Spot the error 1**

A student reads  $r = 0.994$  between cones sold and sunburn cases and writes: 'Selling more ice cream causes sunburn.' The conclusion mistakes association for causation. Both variables rise with temperature, a lurking variable. The right statement: 'Months with more cone sales also tend to have more sunburn cases; both are likely driven by warm, sunny weather.'

**Spot the error 2**

A student compares Program A and Program B overall (40% vs 79.1%) and says B is better for every student. The overall rates are distorted because A serves mostly advanced students. Within each group A has the higher pass rate (90% vs 85%, 30% vs 20%). The correct reading: A is better for each kind of student, and the overall difference comes from group sizes.

**Spot the error 3**

A random sample of 400 adults shows that those who exercise more report better mood ( $r = 0.45$ ). A student writes: 'Since the sample was random, exercise improves mood.' Random **sampling** lets us generalize the **association** to the population. It does not make this a causal study because the adults chose their own exercise levels. Causal claims need random **assignment** of exercise levels.

**Real-World Connection**

A news story reports that cities with more libraries have fewer crimes. Before concluding that building a library lowers crime, a careful reader asks about lurking variables such as city size, income and age distribution, and asks how the data were collected. Good citizens and good scientists separate what the data show from what the headline wants them to show.

**Challenge Corner**

A group of 30 volunteers is randomly split into two groups of 15. Group 1 sleeps 9 hours for a week and group 2 sleeps 6 hours; afterwards reaction times are tested. Group 1's mean reaction time is faster. (a) Can we conclude that more sleep causes faster reactions? (b) For whom? (c) What would improve the study? Solution: (a) Yes, a causal conclusion is supported because treatments were randomly assigned, so lurking variables are balanced on average (if the difference is too large to be explained by chance in a significance test). (b) Only for volunteers like these, since they were not a random sample of a larger population. (c) Blinding the testers to the group, a larger sample, and a random sample from the target population would strengthen it.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

## Practice

### **PRACTICE** Correlation vs Causation and Data Ethics

Show your work on paper. Use a calculator only when a problem says so, or if it is allowed at your level.

- A study finds a strong positive correlation between daily minutes of exercise and a health score. Which statement is best supported by the correlation alone?
  - Exercise minutes and health score are associated.
  - More exercise causes a higher health score.
  - A higher health score causes people to exercise.
  - Exercise is the only factor affecting health.
- Among elementary school children, shoe size and reading score are strongly positively correlated. Which lurking variable most plausibly explains this?
  - Age
  - Number of books at home
  - Height of the teacher
  - Classroom temperature
- Which study design can best support the conclusion that a new teaching method causes higher scores?
  - Students choose which method to use, then compared
  - Students randomly assigned to the new or old method, then compared
  - A survey asks teachers which method they prefer
  - Scores compared across schools that happen to use each method
- Across 50 states the correlation between average income and average test score is 0.9. Which conclusion commits the ecological fallacy?
  - States with higher average income tend to have higher average test scores.
  - The relationship between state averages is strong.
  - Other variables may influence both averages.
  - A person with a higher income will almost surely have a higher test score.
- Program A has pass rates of 90% and 30% in two groups. Program B has 85% and 20%. Program A serves 20 and 100 students and B serves 100 and 10. What is true?
  - A is higher in each group and higher overall.
  - B is higher in each group.
  - A is higher in each group but lower overall (Simpson's paradox).
  - The overall rate for A is the average **60%**.
- What is the main purpose of random assignment to treatment groups in an experiment?
  - To balance lurking variables across the groups on average
  - To ensure the sample is representative of the population
  - To guarantee that every subject improves
  - To make the study double-blind
- A researcher finds a correlation of 0.3 between hours of video games and weekly hours of homework in a sample. Which wording is most appropriate?
  - Playing video games causes students to do more homework.
  - In this sample, more video game time is weakly associated with more homework time.
  - Doing homework causes students to play more video games.
  - There is no relationship between the two.
- A website invites visitors to vote on a question and reports that 88% of the votes say yes. Why should the result not be generalized to all people?
  - The sample is too large.
  - Percentages cannot be generalized.
  - It is a voluntary response sample, which is likely biased.
  - Random sampling always gives lower percentages.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

9. An analyst runs 40 different tests and reports only the one that gave a striking pattern. Which statement is most accurate?
  - (A) This is good practice because it highlights the best result.
  - (B) This is fine because the results are objective.
  - (C) This only matters for experiments, not for surveys.
  - (D) This is cherry-picking: a pattern can appear by chance in many tests, so all tests should be reported.
10. A study of  $r = 0.7$  is reported. What percent of the variation in  $y$  is accounted for by the linear relationship with  $x$ ?
11. Program A passed 18 of 20 beginners and 30 of 100 advanced students. Find the overall pass rate for A as a percent.
12. Cities with more firefighters tend to have more fire damage. Name a lurking variable.
13. Class averages on two variables have  $r = 0.99$ . Can we say the correlation among individual students is about 0.99? Explain briefly.
14. An observational study finds  $r = -0.6$  between screen time and sleep. Give  $r^2$  as a percent and a correct one-sentence interpretation.
15. Using the town's LSRL  $\hat{y} = -13.14 + 0.0689x$  for sunburn cases, predict cases for  $x = 600$  cones to one decimal, and say whether selling fewer cones would reduce sunburn.

## Answers and Explanations

Complete solutions for every practice problem, numbered to match. Read the explanation even when your answer is right.

### 35.1 Linear Regression in Depth

#### PRACTICE · LINEAR REGRESSION IN DEPTH

1. A. The slope is the predicted change in  $y$  per 1 unit of  $x$ , on average. Choice B reverses the roles and says exactly, C describes the intercept wrongly, and D claims causation.
2. D.  $30 + 55(2.4) = 30 + 132 = 162$ . Choice B forgets to add the intercept, and A adds 55 to 30 without using  $x$ .
3. B.  $b = r(sy/sx) = -0.6(10/5) = -1.2$ . Choice A flips the ratio, C loses the sign, and D forgets the ratio.
4. C.  $a = ybar - b \cdot xbar = 40 - 2(7) = 26$ . Choice A is just  $b$  times  $xbar$ , B adds instead of subtracting, and D has the wrong sign.
5. B. Residual = observed - predicted =  $58 - 64.5 = -6.5$ . Choice A reverses the subtraction.
6. A. The value 45 is outside the range 10 to 30, so it is extrapolation. The other three are inside the range.
7. C. The line passes through  $(xbar, ybar)$ :  $3 + 0.5(10) = 8$ . Choice A leaves out the intercept, and D adds 10 to 3.
8. D. A minute is  $1/60$  of an hour so the slope per minute is  $6/60 = 0.1$ , and the intercept ( $x = 0$ ) stays 20. Choice A multiplies instead of dividing.
9. B. Least squares minimizes the sum of squared residuals. Passing through many points is not the criterion, and for the LSRL the plain residuals sum to 0.
10.  $r = 0.75$ . From  $b = r \frac{s_y}{s_x}$ :  $1.5 = r \cdot 2$ .
11.  $\hat{y} = 1.5 + 1.7x$ . Means 5 and 10;  $\sum(x - \bar{x})(y - \bar{y}) = 34$ ,  $\sum(x - \bar{x})^2 = 20$ , so  $b = 1.7$  and  $a = 10 - 1.7(5) = 1.5$ .
12. 88.4.  $b = -0.8 \cdot 2 = -1.6$ ,  $a = 82 + 22.4 = 104.4$ ,  $\hat{y} = 104.4 - 16 = 88.4$ .
13. 24.5. Slope  $b' = r \frac{s_x}{s_y} = 0.3$ ,  $a' = 20 - 0.3(50) = 5$ ,  $\hat{x} = 5 + 0.3(65) = 24.5$ .
14.  $\hat{y} = 8 + 24x$ . Replace  $x$  by  $60x$ :  $0.4 \cdot 60 = 24$ .
15.  $\bar{y} = 70$  and  $r = 0.5$ .  $\bar{y} = 40 + 2(15) = 70$ . From  $2 = r \cdot 4$ ,  $r = 0.5$ .

### 35.2 Residual Analysis and Influential Points

#### PRACTICE · RESIDUAL ANALYSIS AND INFLUENTIAL POINTS

1. A. Random scatter with constant spread gives no evidence against a linear model. Choice D goes too far: absence of a pattern does not prove linearity, and B and C do not follow.
2. C. A U-shape means the data bend away from the line, so a nonlinear model is better. Choice A and B are not what a U-shape shows.
3. B.  $r$  squared = 0.81, and it measures variation explained. Choice A forgets to square, while C and D misstate what  $r$  squared means.
4. D. Predicted =  $5 + 2(4) = 13$ , and residual =  $15 - 13 = 2$ . Choice A reverses the subtraction and B is the prediction, not the residual.
5. A. Its  $x$ -value is far from the mean, so leverage is high, but it follows the pattern, so removing it hardly changes the line. Choices B and C need a large residual, and D contradicts the far-right  $x$ -value.
6. B. A mid-range outlier has little leverage, so the slope barely moves, but it lifts the line and adds scatter, so  $r$  drops. A has the opposite effect on  $r$ , and C ignores the extra scatter.
7. C.  $s = \sqrt{\text{SSE}/(n-2)} = \sqrt{36/9} = 2$ . Choice B divides by 10, which is  $n - 1$ , and A forgets the division.
8. D.  $r$  squared =  $1 - \text{SSE}/\text{SST} = 1 - 0.25 = 0.75$ . Choice A is  $\text{SSE}/\text{SST}$ , the unexplained fraction.
9. B. A point far in  $x$  (high leverage) and far from the old line's prediction of 16.2 has the greatest effect, so (12,7). (12,16) is far in  $x$  but lies near the line, and the points at  $x = 3$  have no leverage.
10. 3. Predicted  $4 + 1.5(6) = 13$ , residual  $16 - 13 = 3$ .
11. 64%.  $r^2 = 0.36$ , so 36% is accounted for and  $100 - 36 = 64\%$  is not.
12.  $s \approx 2.708$ .  $\text{SSE} = 4 + 1 + 9 + 4 + 4 = 22$ ,  $s = \sqrt{22/3} \approx 2.708$ .

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**

13.  $SSE = 396$ .  $SST = (n - 1)s_y^2 = 1100$  and  $SSE = (1 - r^2)SST = 0.36 \cdot 1100 = 396$ .
14.  $r = -0.9$ .  $\sqrt{0.81} = 0.9$  and  $r$  has the sign of the slope.
15.  $0.310$ .  $\bar{x} = 4.5$ ,  $\bar{y} = 17/3$ ,  $\sum(x - \bar{x})(y - \bar{y}) = 24$ ,  $\sum(x - \bar{x})^2 = 77.5$ , slope  $= \frac{24}{77.5} \approx 0.310$ .

### 35.3 Nonlinear Models and Transformations

#### PRACTICE · NONLINEAR MODELS AND TRANSFORMATIONS

- A. A constant ratio is exponential,  $y = A b^x$ , and  $\ln y = \ln A + x \ln b$  is linear in  $x$ . Choice C is for power models.
- B. Taking  $\ln$  of both sides gives  $\ln y = \ln A + p \ln x$ , a line in  $(\ln x, \ln y)$ . Choice A linearizes exponential data.
- C. Exponentiate:  $y = e^{1.2} (e^{0.3})^x$ , and  $e^{1.2}$  is about 3.32 and  $e^{0.3}$  is about 1.350. Choice A skips the exponentiation, B uses 1.2 and 0.3 as the base values, D adds instead of multiplying.
- D.  $\ln \hat{y} = 2 + 0.5(2) = 3$ , so  $y\text{-hat} = e^3 = 20.09$ . Choice A forgets to exponentiate.
- A. A patternless residual plot and a higher  $r$  squared both favor the transformed model. Choices B and C ignore the evidence of curvature.
- B. The factor for doubling  $x$  is  $2^p = 4$ , so  $p = 2$ . Choice A mistakes the factor for the exponent.
- C. Only 2, 6, 18, 54 has a constant ratio of 3. Choice A has a constant difference (linear), and the other two have neither.
- D.  $10^{0.5} = 3.162$ , so  $y\text{-hat} = 3.162 x^2$ . Choice C uses 0.5 itself as the multiplier, forgetting to raise 10 to the intercept.
- A. The  $r$  squared refers to the transformed variable  $\ln y$ . Choice B drops that qualifier, and C and D misstate what  $r$  squared means.
- $66.69$ .  $\ln \hat{y} = 3.2 + 1 = 4.2$  and  $\hat{y} = e^{4.2} \approx 66.69$ .
- $28.4\%$ . The growth factor is  $e^{0.25} \approx 1.284$ .
- $1585$ .  $\log \hat{y} = 0.2 + 1.5(2) = 3.2$  and  $\hat{y} = 10^{3.2} \approx 1584.9$ .
- $y = 2.5(2)^x$ . The ratio is  $b = 2$  each step and  $A = 5/2 = 2.5$ .
- $y = 4x^2$ . At  $x = 1$ :  $A = 4$ . Then  $36/4 = 9 = 3^p$  so  $p = 2$ .
- $y = e^{-1}x^3 \approx 0.368x^3$ , so  $y(2) \approx 2.94$ . Slope  $= \frac{8-2}{3-1} = 3$ , intercept  $= 2 - 3(1) = -1$ .

### 35.4 Correlation vs Causation and Data Ethics

#### PRACTICE · CORRELATION VS CAUSATION AND DATA ETHICS

- A. Correlation shows association only. Choices B and C claim a direction of cause that the data cannot establish, and D is not implied.
- A. Older children have larger feet and also read better. The other variables do not plausibly affect both.
- B. Random assignment balances lurking variables, allowing causal conclusions. Choices A, C and D are observational and could be confounded.
- D. The group-level correlation does not apply to individuals; individual correlations are typically weaker. The other three statements are about states or are cautious.
- C. Overall A is  $18 + 30 = 48$  of  $120 = 40\%$ ; B is  $85 + 2 = 87$  of  $110 = 79.1\%$ . Choice D averages the two rates without weighting by group size.
- A. Random assignment spreads lurking variables evenly across groups. Choice B describes random sampling, D is a different feature (blinding).
- B. It states a weak association in the sample without a causal claim. Choices A and C claim causation, D contradicts  $r = 0.3$ .
- C. Voluntary response samples over-represent people with strong opinions. Choices A and B are false statements about sampling.
- D. With many tests some striking patterns appear by chance. Ethical reporting discloses everything that was tried; A, B and C ignore that.
- $49\%$ .  $r^2 = 0.49$ .
- $40\%$ .  $\frac{18+30}{20+100} = \frac{48}{120} = 0.40$ .
- The size of the fires (or the size of the city). Larger fires cause both more damage and more firefighters sent.
- No. Averages hide individual variation, so individual correlations are usually weaker; assuming otherwise is the ecological fallacy.
- $36\%$ . About  $36\%$  of the variation in sleep time is accounted for by the linear relationship with screen time (association, not cause).
- $28.2$  cases.  $-13.14 + 0.0689(600) = 28.2$ . No: the link is through temperature (a lurking variable), not a cause.

Want a plan built around **your** student?

Book a free consultation. The short intake takes about a minute and opens right away.

**BOOK NOW →**