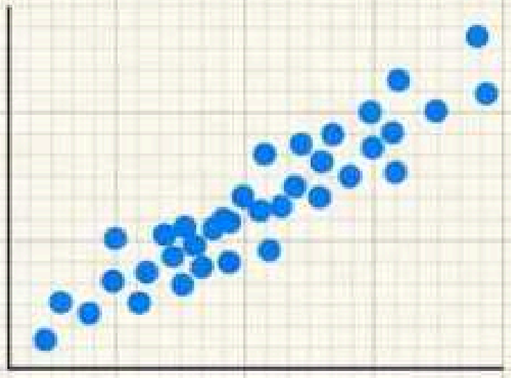
4.1

Two quantitative variables:
From visualization to prediction—
and the limits of cause.

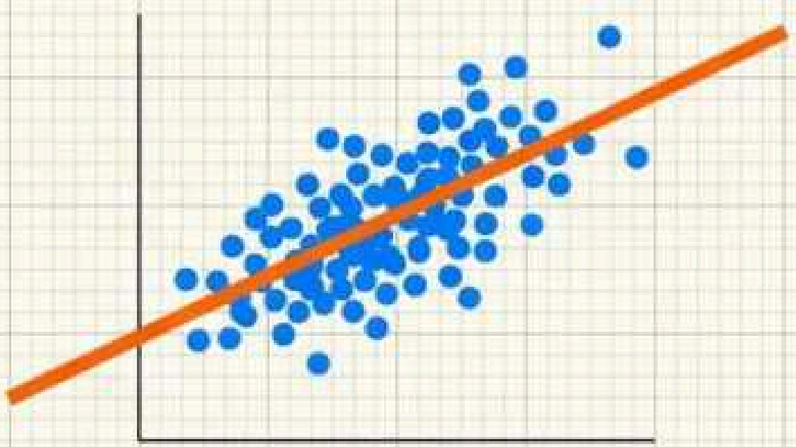
Visualizing relationships



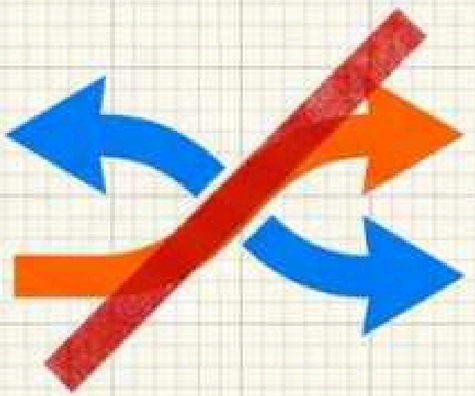
Measuring correlation

$$\{ r \}$$

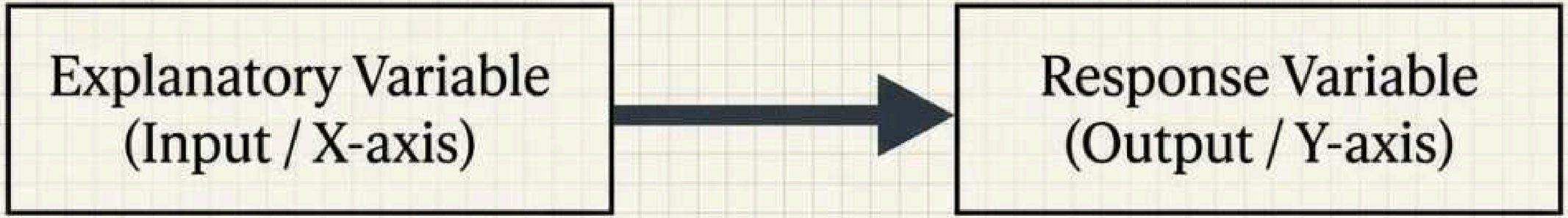
Making predictions



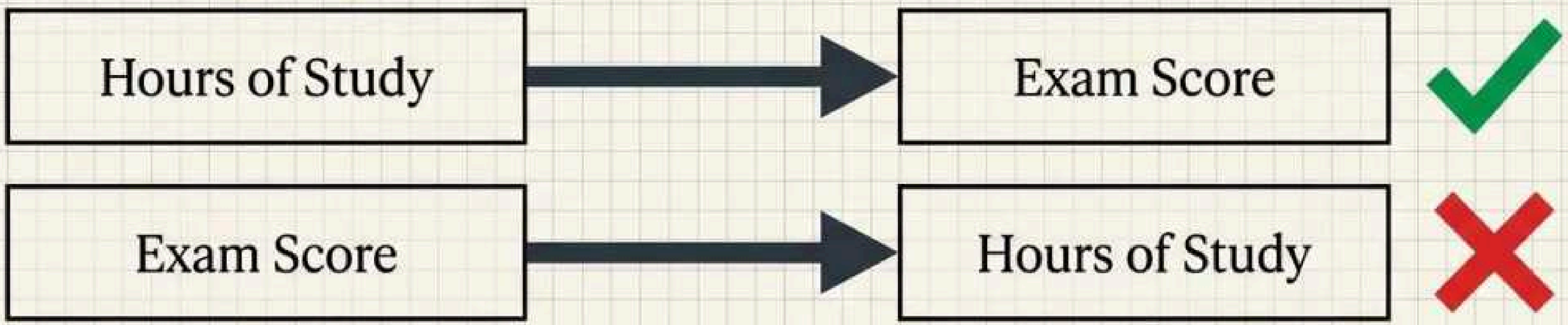
Association ≠ Causation



4.2

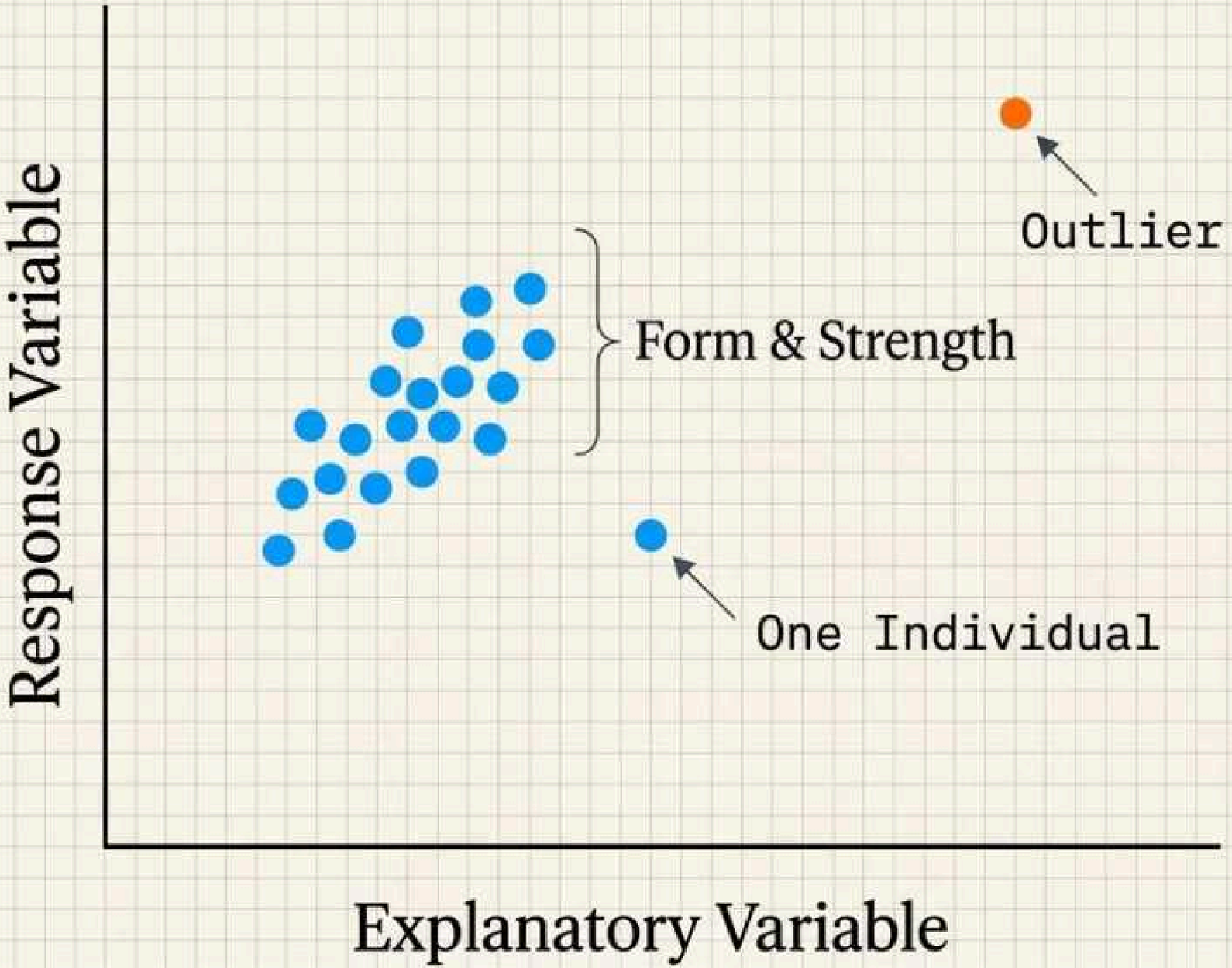


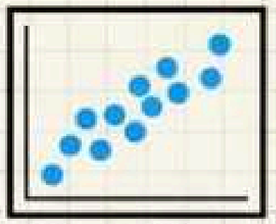
Case Study Mapping



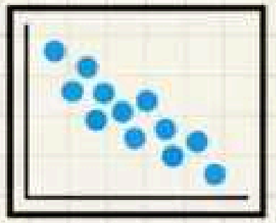
Labels reflect our specific question, not a proven cause.

4.3





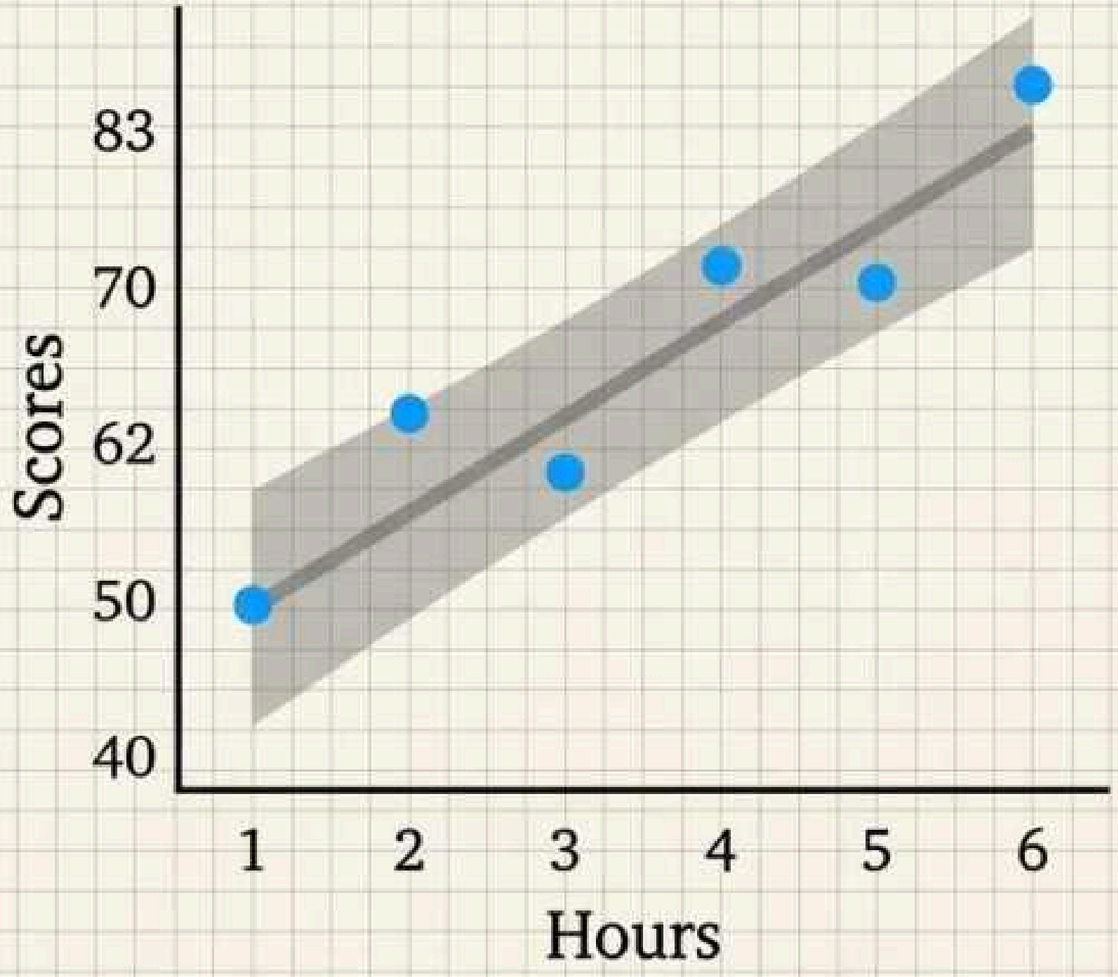
Positive Direction



Negative Direction

4.4

Hours	Scores
1	50
2	62
3	58
4	71
5	70
6	83



[Positive] [Linear] [Fairly Strong]

No point lies exactly on a line.

4.5

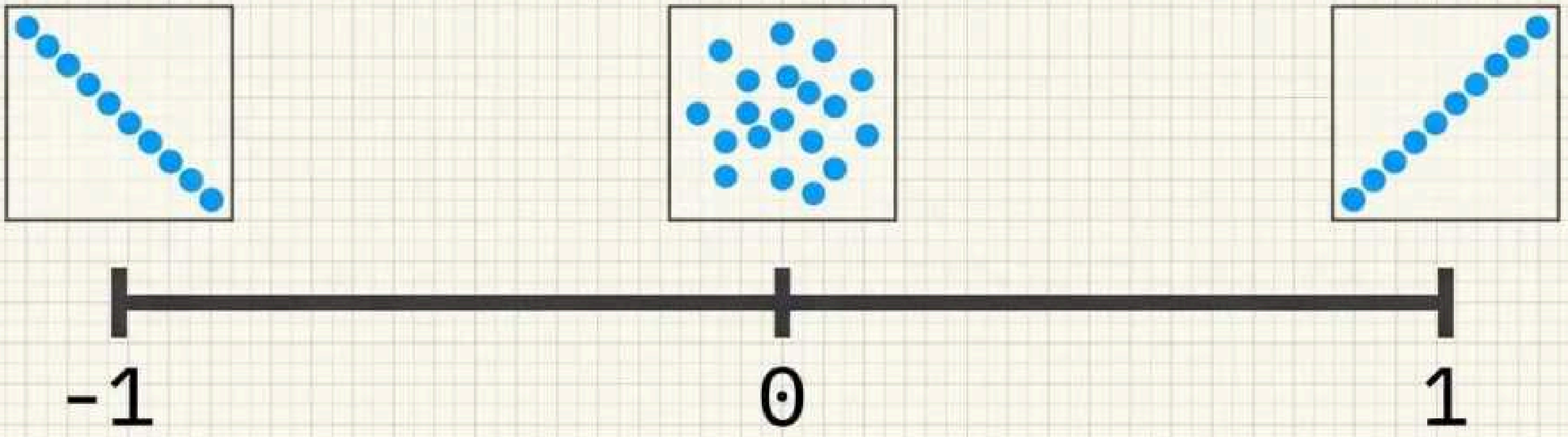
$$r = \frac{\text{Average of } (z\text{-score}_x \times z\text{-score}_y)}{\text{Summed over all individuals and divided by } n - 1}$$

$+ r$ = Positive Association

$- r$ = Negative Association

Measures the strength and direction of a straight-line relationship with one number.

4.6



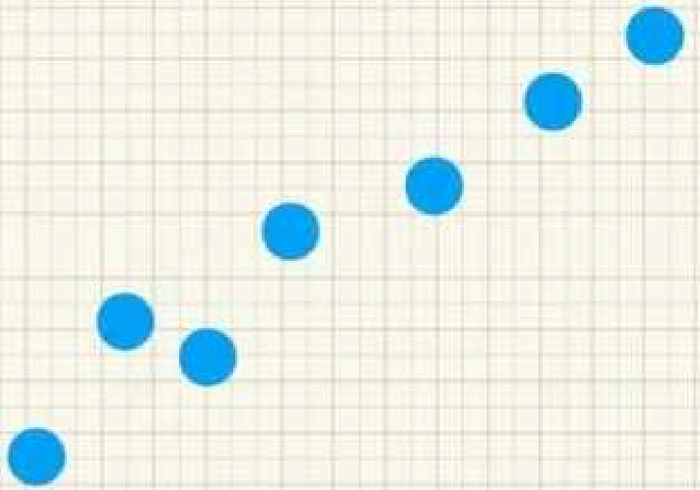
Unitless

Changing inches to cm does not change r.

Symmetrical

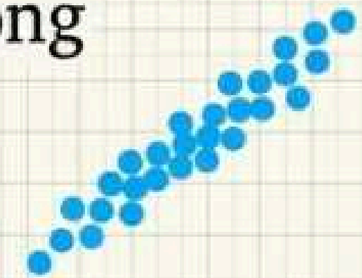
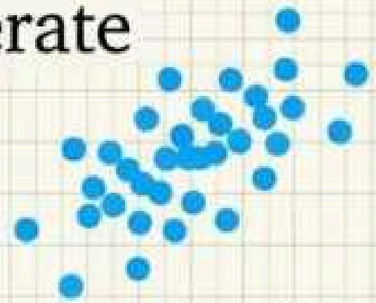
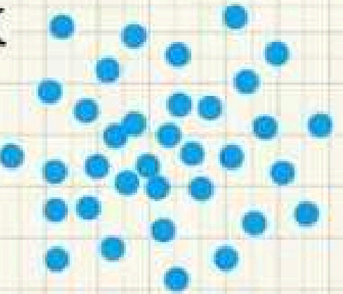
Swapping X and Y yields the same r.

4.7



$r \approx 0.94$

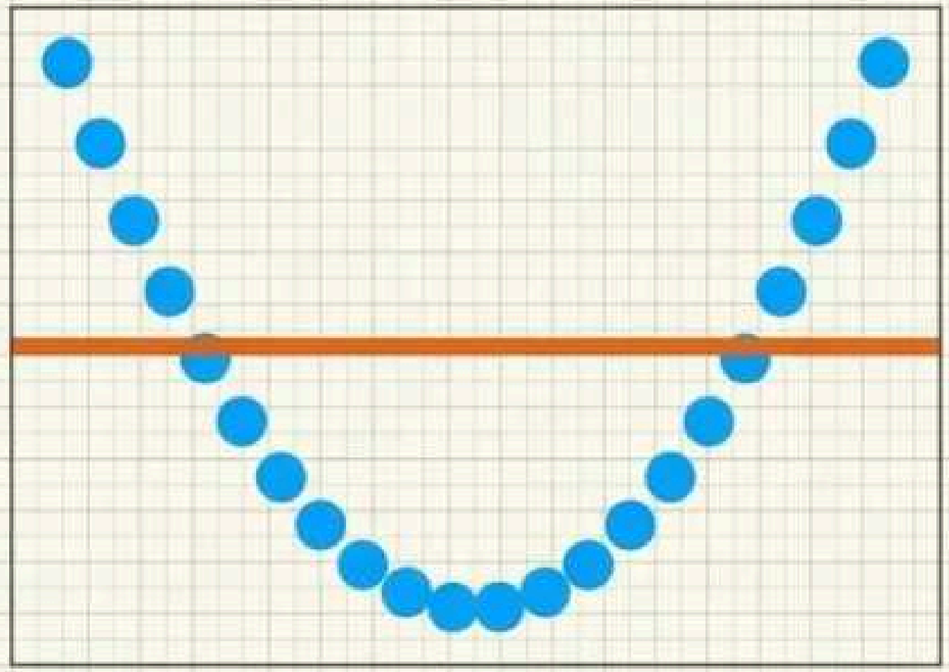
Strong Positive Linear.

$r > 0.8$	$r \sim 0.5$	$r < 0.3$
Strong 	Moderate 	Weak 

Judge by the scatterplot as well—a single number can hide the shape.

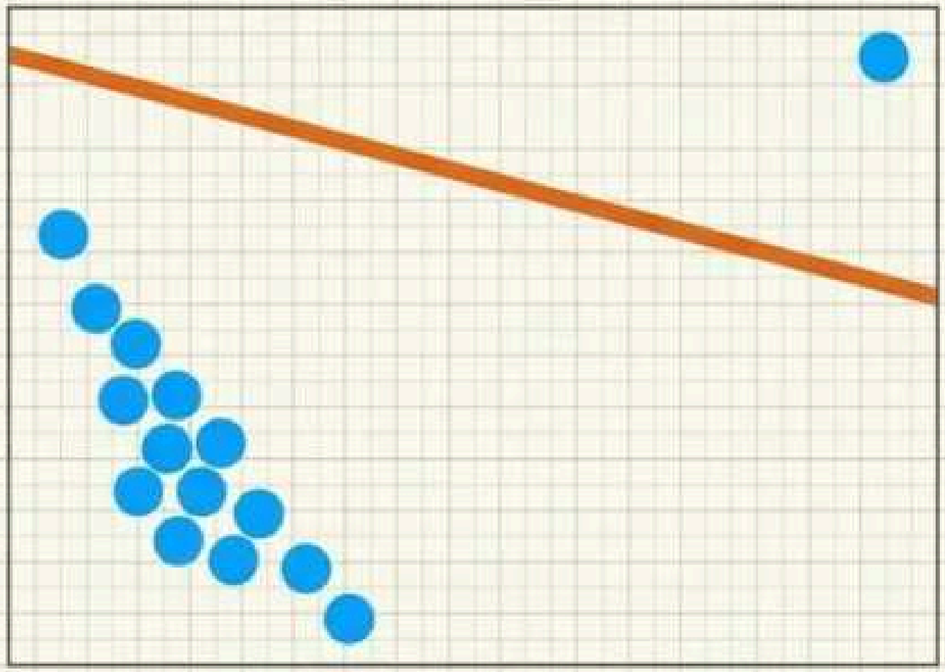
4.8

Non-linear



`r = 0` (Rising half cancels the falling half)

Outliers

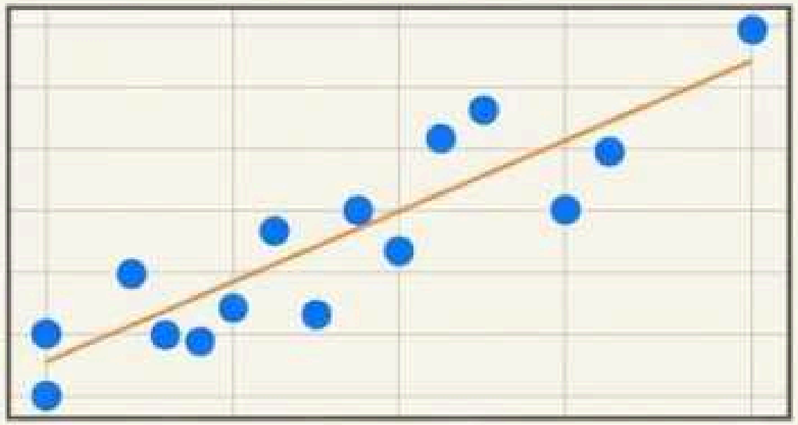


Sensitive to a single far-off point

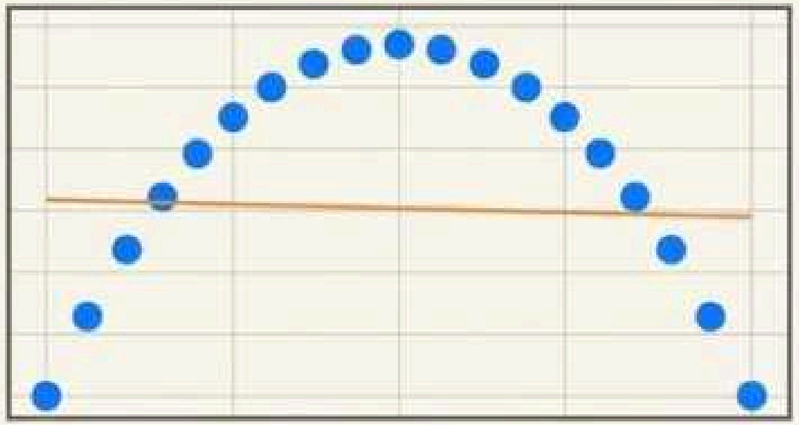
Correlation has limits. Always draw the picture.

4.9

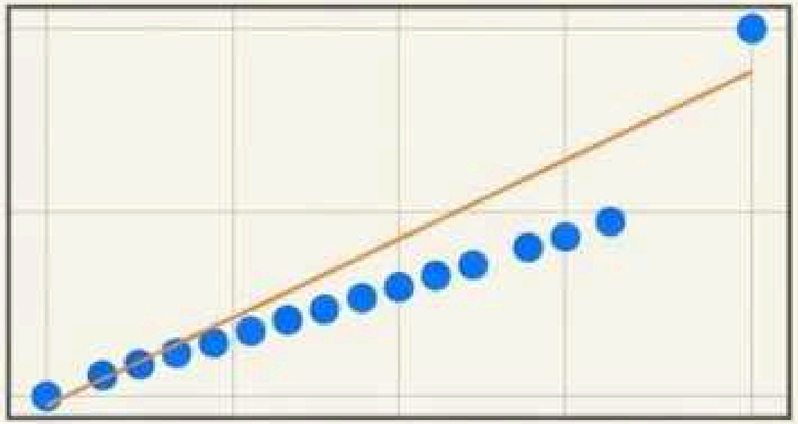
Francis Anscombe (1973). All four sets share identical means, spreads, and $r \approx 0.82$.



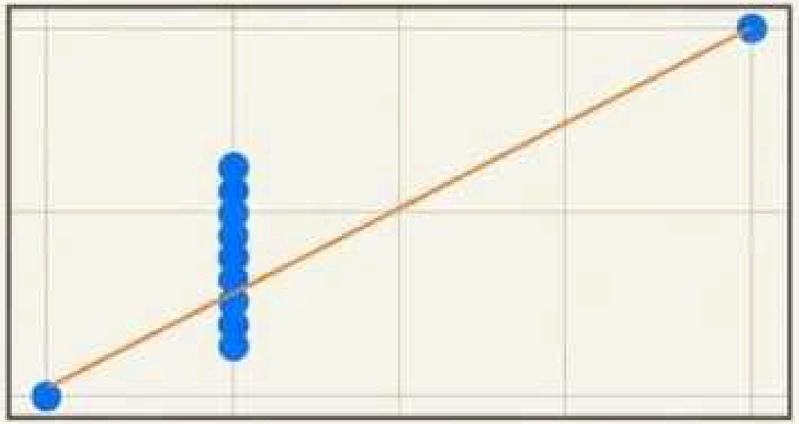
Simple Linear Relationship



Strong Non-linear Pattern



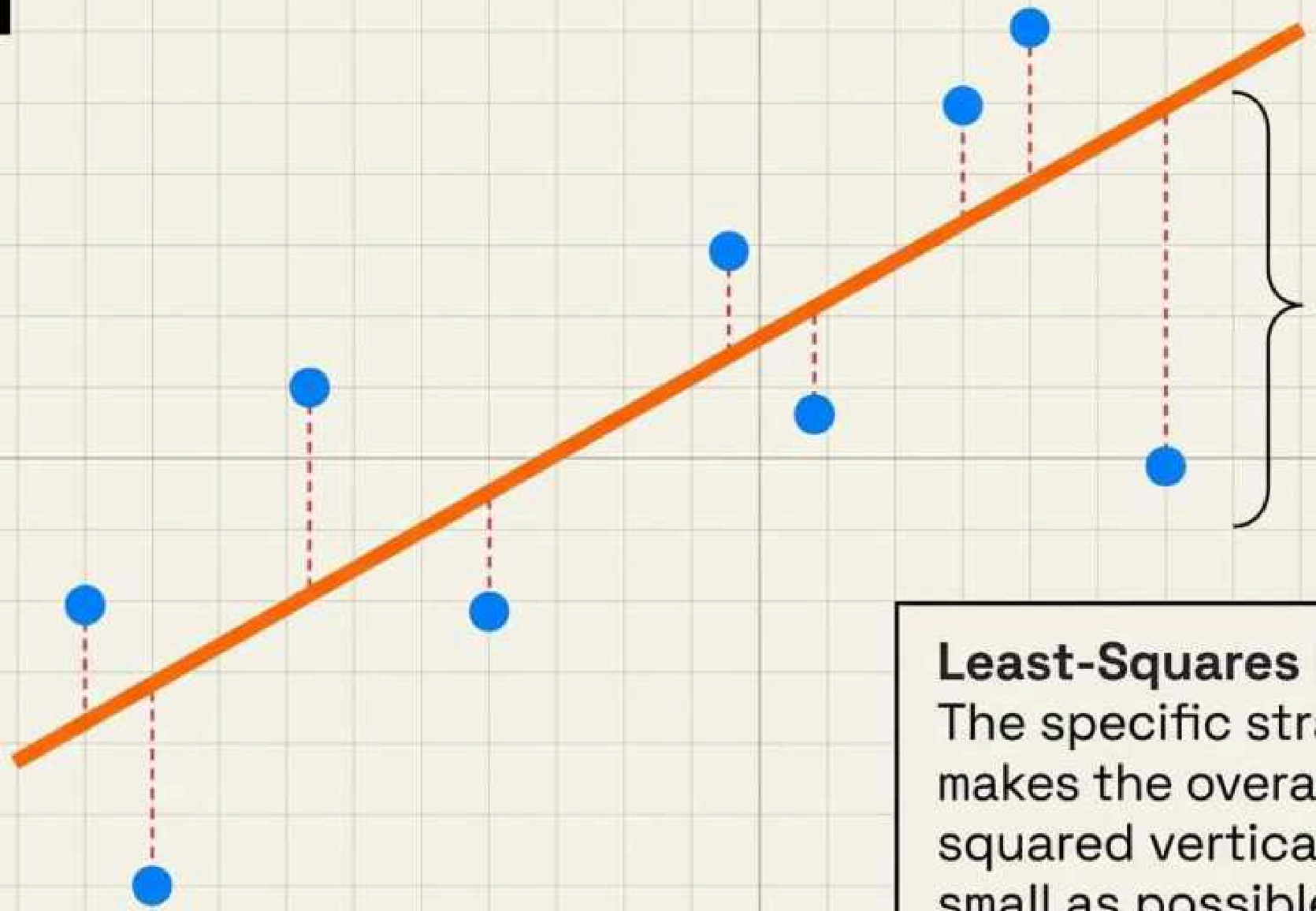
Impact of a Single Outlier



Outlier Drives the Line

Plot before you trust a summary.

4.10



Least-Squares Regression:
The specific straight line that makes the overall sum of these squared vertical distances as small as possible.

4.11

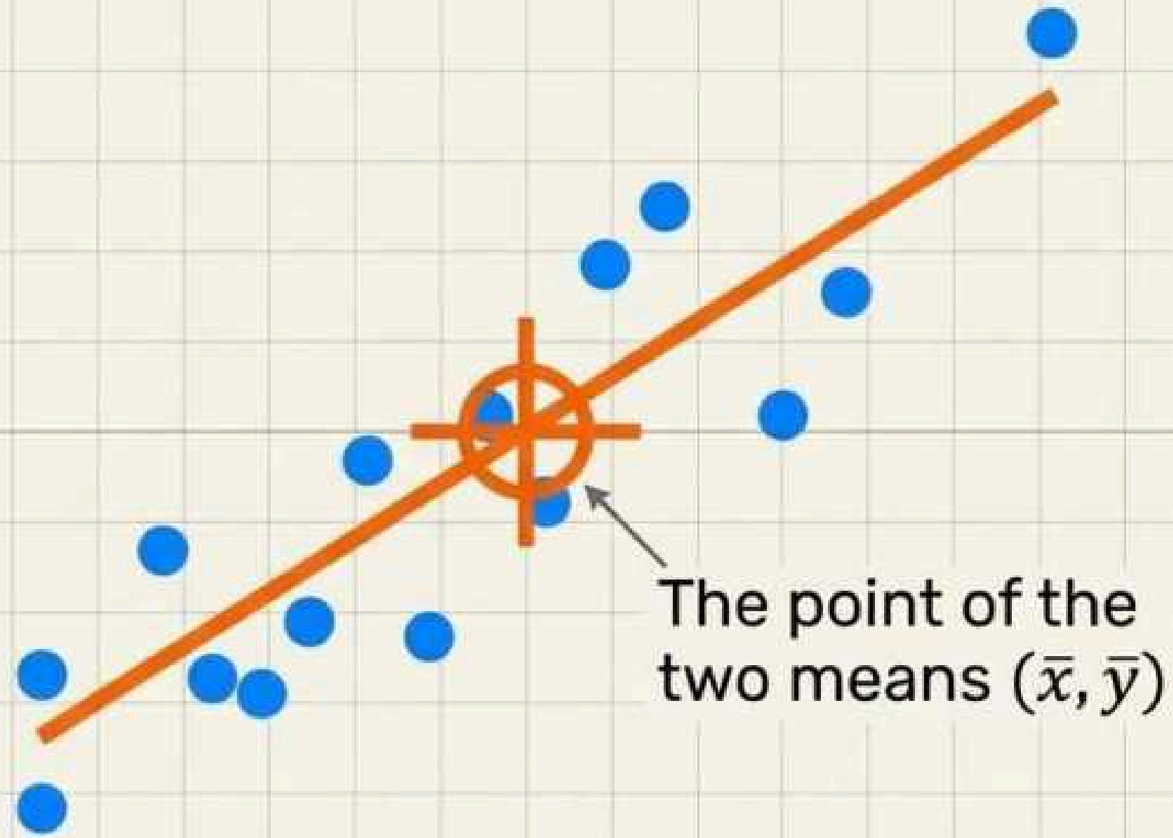
$$\hat{y} = a + bx$$

Predicted Response

Intercept

Slope

$$\text{Slope (b)} = r \times \frac{SD_y}{SD_x}$$



Can be computed with just 5 summary numbers.

4.12

1. The 5 Summary Numbers

Hours: Mean = 3.5, SD = 1.87

Score: Mean = 65.7, SD = 11.54

Correlation: r = 0.936

↓

2. Calculate Slope (b)

0.936 × (11.54 / 1.87) = 5.77

↓

3. Calculate Intercept (a)

65.7 - (5.77 × 3.5) = 45.5

↓

Predicted Score = 45.5 + 5.77 × Hours

4.13

5.77

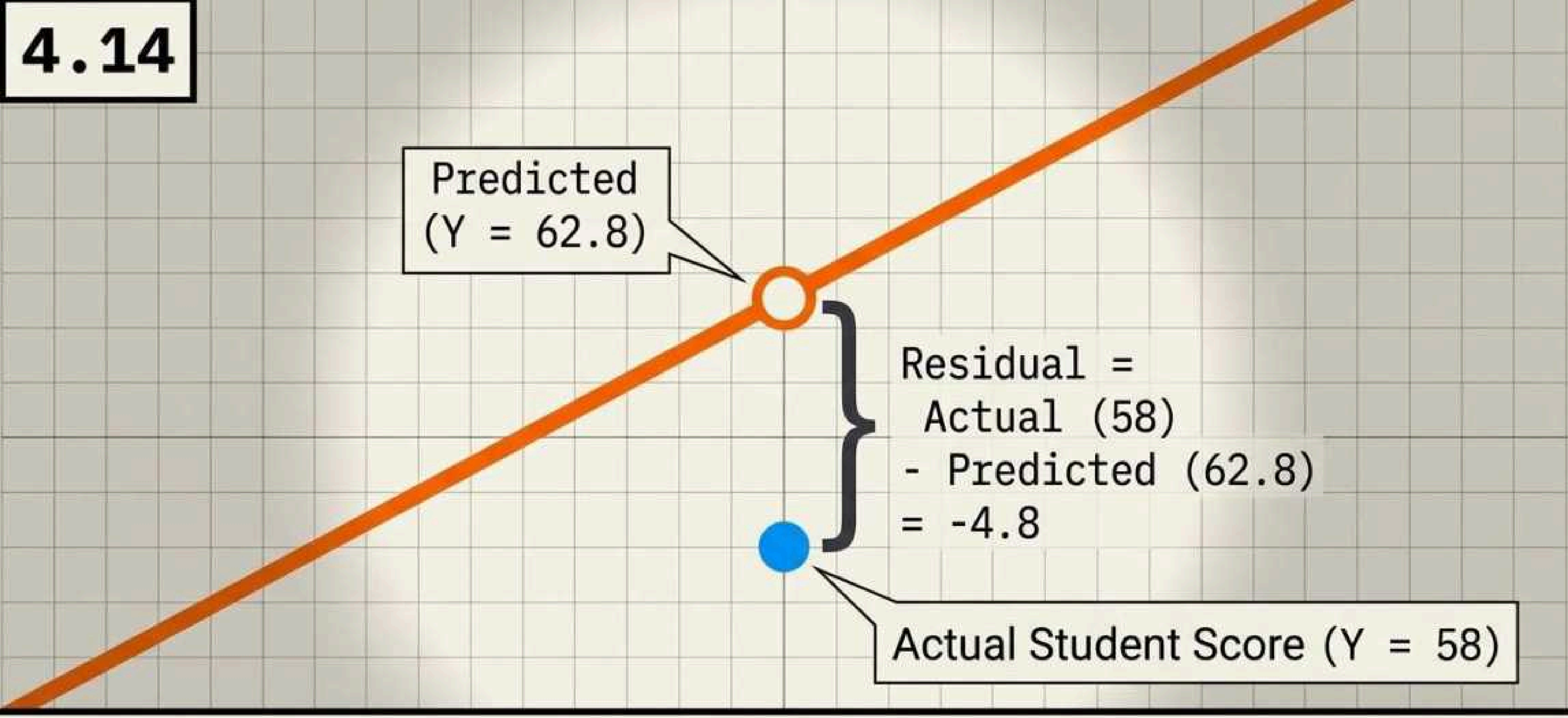
Each additional hour of study goes with about 5.8 more points on quiz, on average.

45.5

The predicted score for 0 hours of study.

(Interpret with care—may not always be meaningful.)

Always speak Always speak in the units of the problem.
Use words like “predicted” or “on average”—never “exactly”.



Residuals reveal how far each point lies from the line. Plotting them can expose hidden curves.

4.15

How good is the line?

$r^2 \approx 0.876$



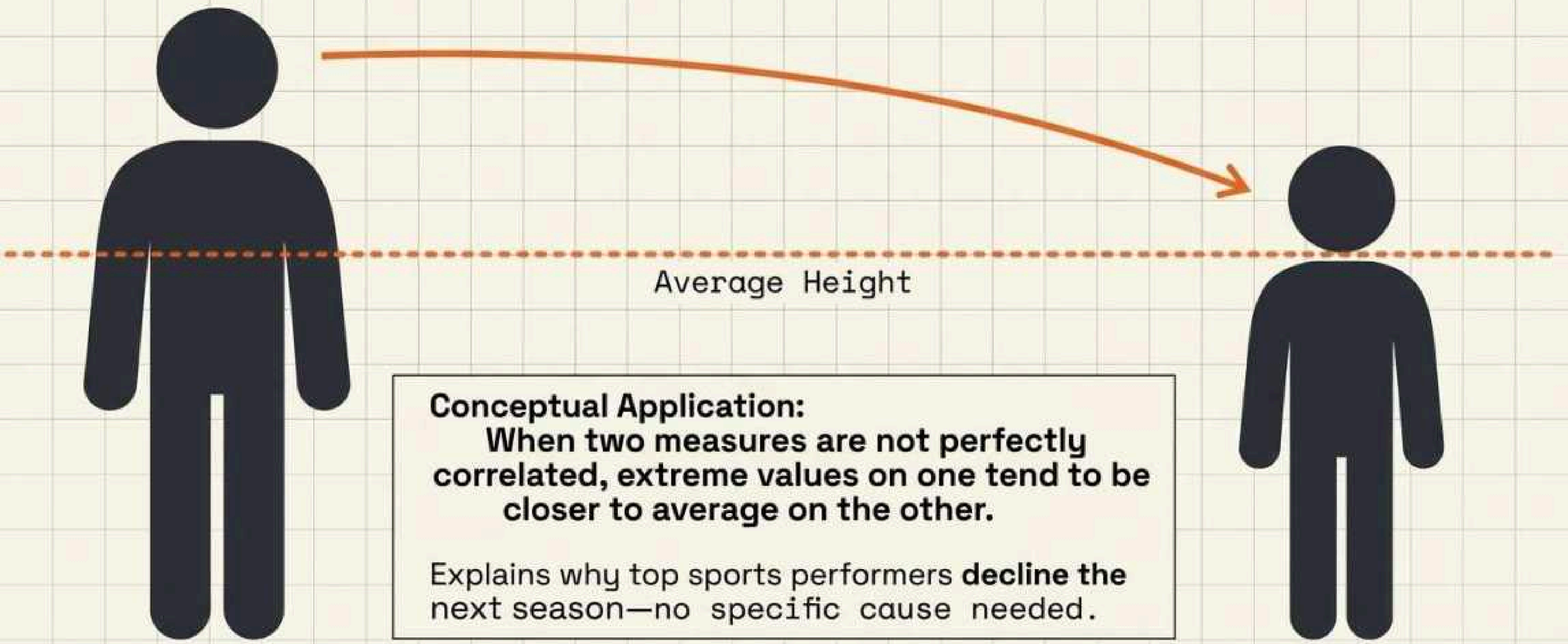
4.16



Extrapolation Warning: Relationships that look straight over a range often bend outside it. The line is only trustworthy within the data range.

4.17

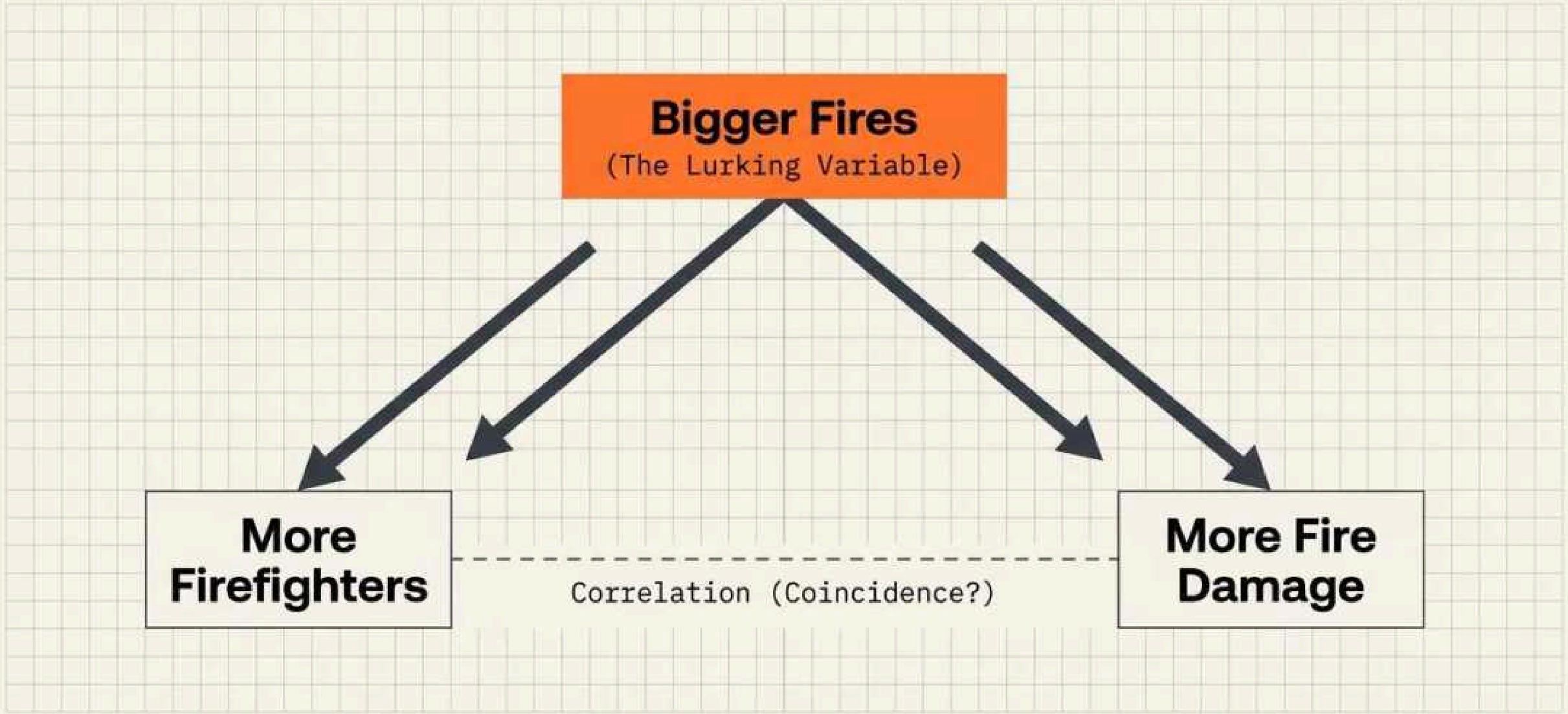
Why is it called ‘Regression’?



Conceptual Application:
When two measures are not perfectly correlated, extreme values on one tend to be closer to average on the other.

Explains why top sports performers **decline the** next season—no specific cause needed.

4.18 Correlation ≠ Causation



Experiments with Random Assignment

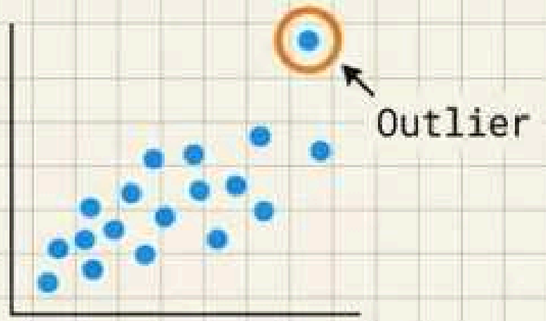
(The definitive route to establish cause. Studied in Lecture 5.)

When experiments are impossible (e.g., smoking and cancer), a case builds from:

- Consistent results across multiple studies.
- Strong and graded association.
- A plausible biological/physical mechanism.
- The rigorous exclusion of rival explanations.

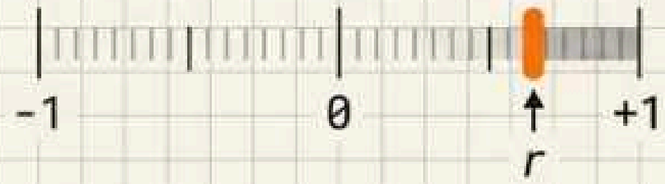
4.20

Draw



Describe direction, form, strength, and outliers.

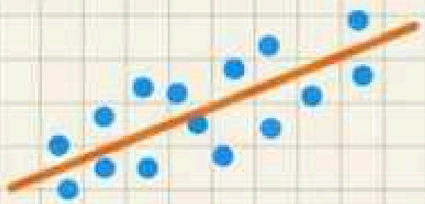
Measure



Correlation r measures straight-line strength.

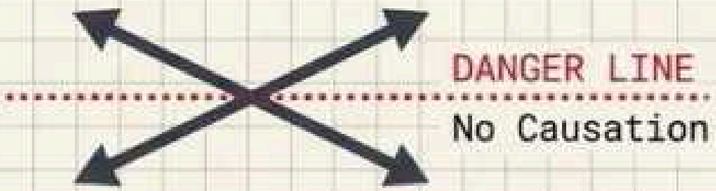
Predict

$y = a + bx$



r^2 shows accounted variation. Slope = response units per explanation unit.

Warn



Do not extrapolate. Correlation alone is never cause.

Up Next in Lecture 5: Where should data come from?