KIN 610: Quantitative Methods in Kinesiology

Chapter 11: Correlation and Bivariate Regression

Ovande Furtado Jr., PhD.

2026-02-21

1 FYI

This presentation is based on the following books. The references are coming from these books unless otherwise specified.

Main sources:

- Moore, D. S., Notz, W. I., & Fligner, M. (2021). *The basic practice of statistics* (9th ed.). W.H. Freeman.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms

ClassShare App

You may be asked in class to go to the ClassShare App to answer questions.

- https://classsharedrfurtado.netlify.app/

SPSS Tutorial

- SPSS Tutorial: Correlation

2 Intro Question

- A coach wants to know: does greater leg strength predict higher vertical jump performance? You collect leg strength (kg) and jump height (cm) from 30 athletes. How would you describe and model this relationship?

Click to reveal answer

We need tools that (1) **quantify the strength** of the relationship and (2) allow us to **make predictions**. Correlation tells us how strongly two variables co-vary; regression produces a mathematical equation for prediction. Together they form the foundation of bivariate analysis in Movement Science.

- **Correlation** quantifies the strength and direction of a linear relationship between two variables.
- **Bivariate regression** models that relationship mathematically and enables prediction.
- Both depend on visualizing data with a **scatterplot** first.

3 Learning Objectives

By the end of this chapter, you should be able to:

- Explain what correlation measures and how it quantifies linear relationships
- Compute and interpret Pearson's r and r^2
- Distinguish between correlation and causation
- Construct and interpret scatterplots for bivariate data
- Fit a bivariate regression model and interpret the slope, intercept, and R^2
- Assess assumptions: linearity, homoscedasticity, independence, normality of residuals
- Recognize the influence of outliers on correlation and regression results
- Apply and report correlation and regression results appropriately in Movement Science

4 Symbols

Symbol	Name	Pronunciation	Definition
r	Pearson's correlation	"r"	Strength and direction of the linear relationship
ρ	Population correlation	"rho"	True correlation in the population
r^2	Coefficient of determination	"r squared"	Proportion of variance in Y explained by X
R^2	Coefficient of determination (regression)	"R squared"	Proportion of variance explained by the regression model

Symbol	Name	Pronunciation	Definition
$\hat{y}$	Predicted value	"y hat"	Value of Y predicted by the regression equation
a	Intercept	"a"	Predicted Y when $X = 0$
b	Slope	"b"	Change in $\hat{y}$ for a one-unit increase in X
e	Residual	"residual"	Difference between observed and predicted Y

5 What is Correlation?

Correlation measures the strength and direction of the **linear relationship** between two continuous variables[1,2].

Key properties:

- Dimensionless (no units) and standardized
- Ranges from -1 to $+1$
- Symmetric: $r_{XY} = r_{YX}$

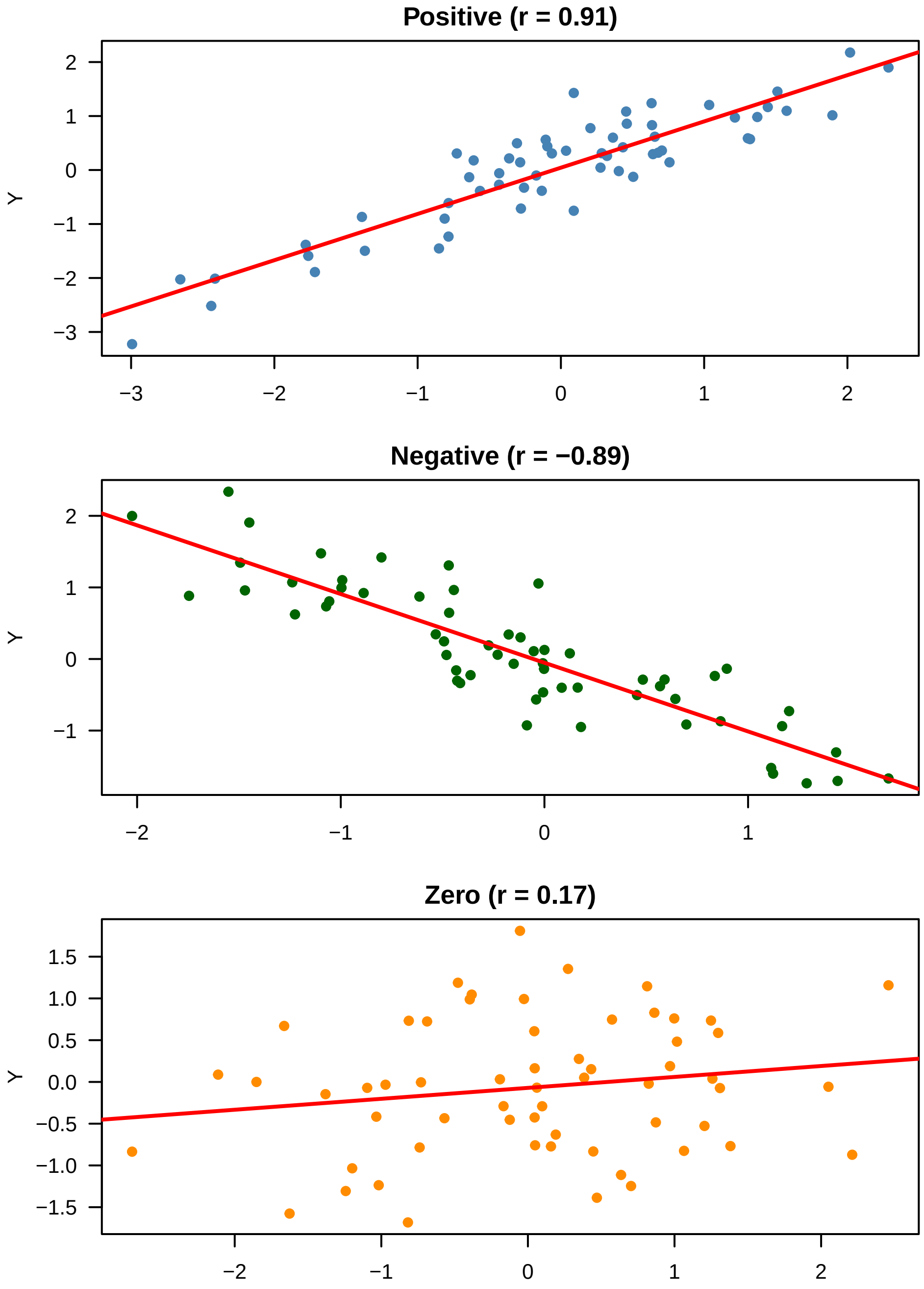
Directions:

- **Positive**: Higher X → Higher Y (e.g., leg strength and jump height)
- **Negative**: Higher X → Lower Y (e.g., body mass and endurance performance)
- **Zero**: No linear relationship

Benchmarks[1]:

$\lvert r \rvert$	Strength
> 0.7	Strong
0.4–0.7	Moderate
< 0.4	Weak

- Remind students that Pearson's r is unitless — you can correlate variables measured in completely different units.
- The benchmarks are approximate; what's "strong" depends on the field.



Positive (r = 0.91)

Negative (r = −0.89)

Zero (r = 0.17)

Figure 1: Three types of correlation: positive (top), negative (middle), and zero (bottom)

6 Pearson's Formula for r

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

- The **numerator** measures covariation: how X and Y vary together
- The **denominator** standardizes by the variability of each variable individually

Equivalently, using z-scores[1]:

$$r = \frac{1}{n-1} \sum z_x \, z_y$$

Intuition

When above-average X tends to pair with above-average Y (both deviations have the same sign), products are mostly positive → $r > 0$. When they pair with opposite signs → $r < 0$.

- Most students will use SPSS or R to compute r. The formula helps build intuition.
- Focus on the z-score version as the conceptual bridge.

7 Worked Example: Computing r

Data: leg strength (kg) and vertical jump height (cm) in 8 athletes.

Athlete	X (kg)	Y (cm)
1	80	45
2	90	50
3	70	40
4	100	55
5	85	48
6	95	52
7	75	42
8	88	49

Means: $\bar{x} = 85.375$ kg, $\bar{y} = 47.625$ cm

Sums of squares and cross-products:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 354.12$$
$$\sum(x_i - \bar{x})^2 = 707.88$$
$$\sum(y_i - \bar{y})^2 = 177.88$$

Pearson's r:

$$r = \frac{354.12}{\sqrt{707.88 \times 177.88}} = \frac{354.12}{354.90} = \mathbf{0.998}$$

Interpretation:

An extremely strong positive linear relationship between leg strength and vertical jump height.

> **!** Important
>
> Always plot first — verify this is indeed linear!

- Walk through the computation carefully.
- Point out that r = 0.998 is unusually high for real data — this is a constructed example.

8 Scatterplots: Always Plot Your Data

Scatterplots are the essential first step — never compute r without visualizing the data first[3,4].

What scatterplots reveal:

- **Shape of the relationship**: Indicates whether a linear model is appropriate or if a nonlinear (curved) pattern exists.
- **Strength and direction**: Shows how closely points cluster together and whether they follow a positive or negative trend.
- **Outliers or influential points**: Highlights extreme values that could disproportionately skew r and regression results.
- **Heteroscedasticity**: Reveals if the spread of data points changes across the range of values (e.g., forming a funnel shape).

> ⚠ Anscombe's Quartet
>
> Four datasets with **identical** $r = 0.816$, $\bar{x}$, $\bar{y}$, and regression lines — but completely different patterns when plotted. Correlation alone can be deeply misleading[1].

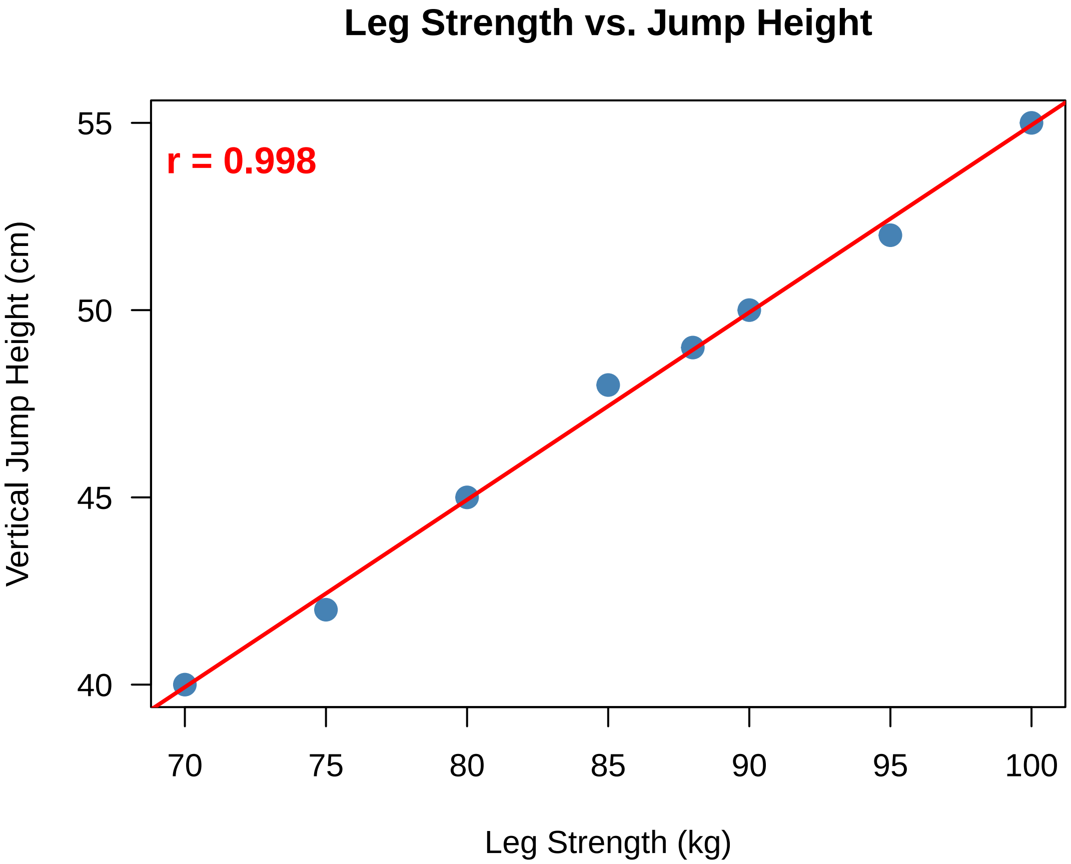


Figure 2: Strong positive linear relationship between leg strength and vertical jump height (r = 0.998)

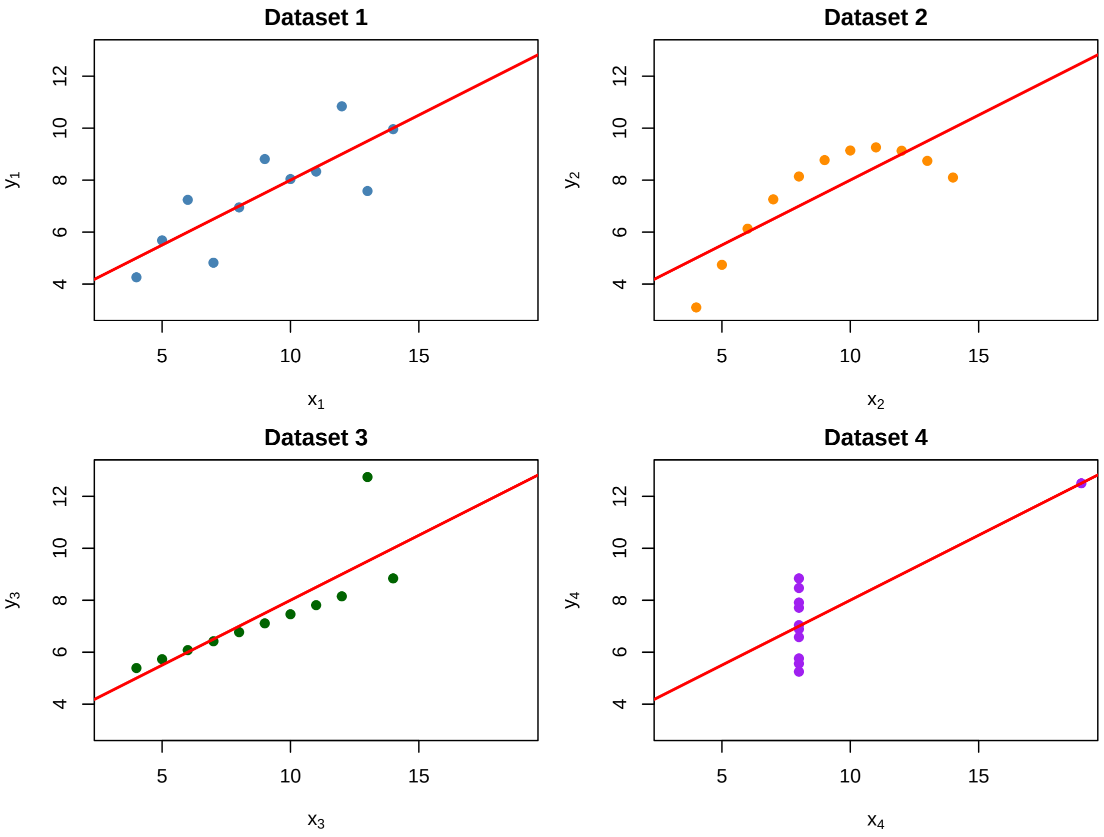


Figure 3: Anscombe's Quartet: Four datasets with identical statistics but different distributions.

- Emphasize Anscombe's Quartet: same statistics, wildly different data. This is the strongest argument for always plotting first.

9 Critical Limitation 1: Linearity

Pearson's r measures **only linear** associations[1,3]. Two variables can have a strong, meaningful relationship yet produce $r \approx 0$ if the relationship is nonlinear.

Movement Science examples of nonlinearity:

- **Lactate–intensity**: exponential rise at high intensities
- **Arousal–performance**: inverted-U (Yerkes-Dodson law)
- **Fatigue–time**: rapid initial decline, then plateau

What to do if nonlinear:

1. Transform data (e.g., log transformation)
2. Fit a nonlinear model
3. Use Spearman's rank correlation (r_s)

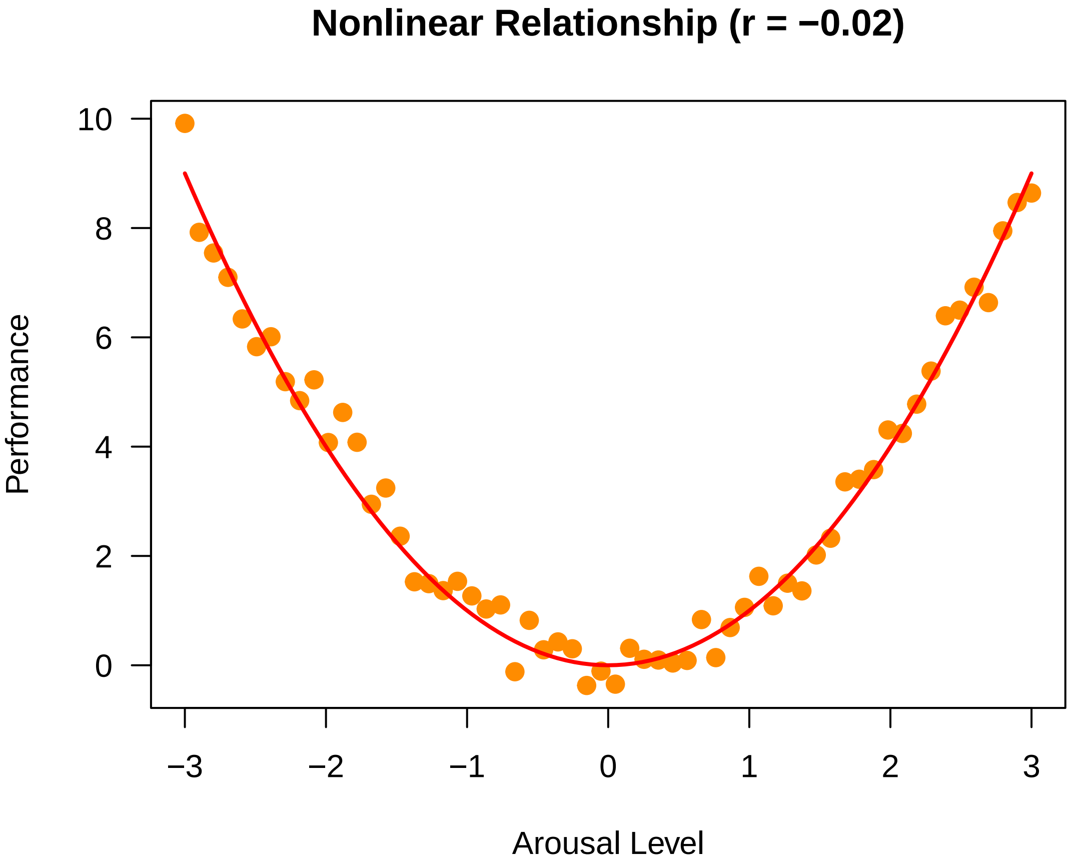


Figure 4: A nonlinear relationship: Pearson's r 0 despite a clear U-shaped pattern

- The inverted-U example is particularly relatable for kinesiology students.
- Key message: r 0 does NOT mean no relationship.

10 Critical Limitation 2: Correlation Causation

A strong correlation does not prove that one variable causes the other[1,5].

Three reasons correlations can be misleading:

Explanation	Description	Example
Confounding variable	Third variable drives both	Hot weather → ice cream sales AND drownings
Reverse causation	Direction assumed backwards	Do fit people exercise, or does exercise make people fit?
Spurious correlation	Coincidence	Spelling bee winner length spider deaths

Establishing causation requires:

1. Temporal precedence (cause precedes effect)
2. Covariation (variables must correlate)
3. Elimination of alternatives (RCT or experimental control)

! Language matters

Use: "X is *associated with* Y" or "X and Y are *related*"
Avoid: "X *causes* Y" (unless you have experimental evidence)

ℹ Movement Science example

A strong negative correlation between physical activity and cardiovascular disease does **not** prove that activity prevents heart disease[5]. Healthier individuals may simply be more likely to exercise (reverse causation), or genetic factors may influence both (confounding). Only RCTs can establish causation[6].

- The ice cream/drowning example is always memorable.
- Remind students that in Movement Science, most observational studies can only show association.

11 Coefficient of Determination: r^2

Squaring r gives r^2, the **coefficient of determination**: the proportion of variance in Y explained by X[1,3].

$$r^2 = (0.998)^2 = 0.996$$

Interpretation:

- **99.6%** of the variability in jump height is accounted for by leg strength
- The remaining **0.4%** is unexplained (technique, fiber type, measurement error, etc.)

Practical benchmarks[7]:

r^2	Interpretation
< 0.10	Weak ($<$ 10% shared variance)
0.10–0.30	Moderate
≥ 0.30	Strong

! Statistical significance   practical importance

With very large samples, even $r = 0.05$ can be statistically significant — yet $r^2 = 0.0025$ means only 0.25% of variance is explained[4,6].
Always report r, r^2, and confidence intervals — not just p-values.

Significance test for r:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad df = n - 2$$

For our example: $t = 38.8$, $df = 6$, $p < .001$

- r² is the more interpretable quantity for practical purposes.
- The significance test just tells us if r is distinguishable from zero — not how meaningful it is.

12 Bivariate Linear Regression: The Model

Regression goes beyond correlation: it fits a **mathematical equation** to predict Y from X[1,3].

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

Components:

Symbol	Name	Meaning
$\hat{y}$	Predicted value	Estimated Y for a given X
a	Intercept	Predicted Y when $X = 0$

Symbol	Name	Meaning
b	Slope	Change in $\hat{y}$ per 1-unit increase in X

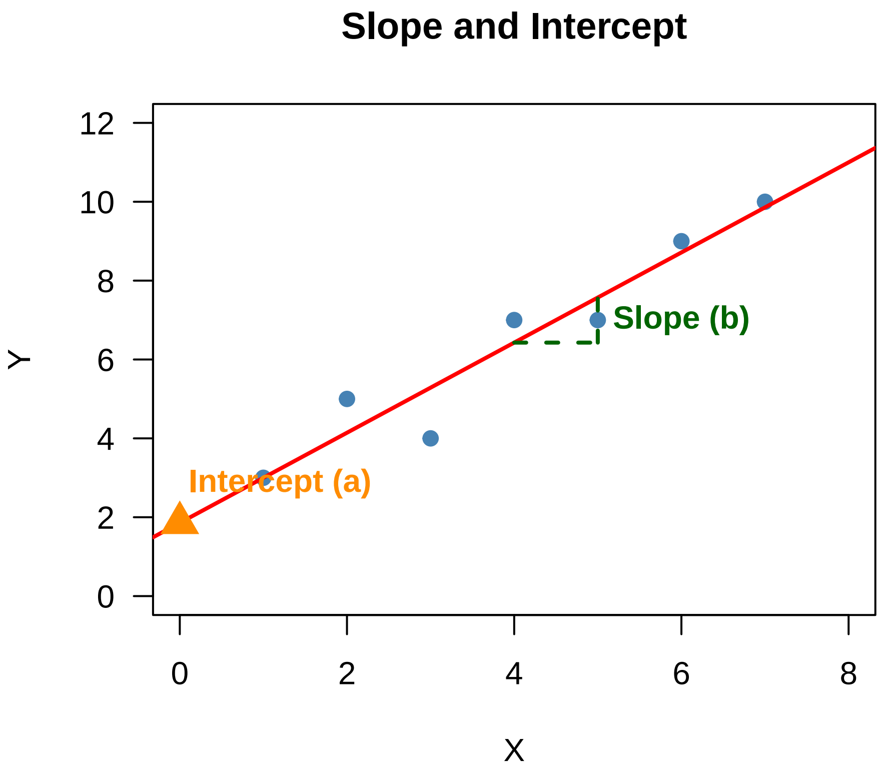


Figure 5: Visualizing the regression line components

Correlation vs. Regression

	Correlation	Regression
Goal	Quantify association	Predict Y from X
Output	r, r^2	Equation $\hat{y} = a + bx$
Symmetric?	Yes ($r_{XY} = r_{YX}$)	No (predicting Y from X vice versa)
When to use	Describe relationship	Make predictions

- Key distinction: correlation is symmetric, regression is not (you must choose a predictor and an outcome).
- Emphasize that $\hat{y}$ is the *predicted* value, not the actual observed value.

13 Understanding Slope and Intercept

To build the regression equation $\hat{y} = a + bx$, we must calculate the **slope** (b) and **intercept** (a).

The Slope (b)

- Determines the **steepness** of the regression line.
- Formula: $b = r\frac{s_y}{s_x}$
- If b is positive, the line goes up; if negative, the line goes down.
- Represents the **rate of change**: how much does Y change when X increases by exactly one unit?
- In real-world terms (e.g., strength → jump height), a slope of 0.5 means a 1kg increase in strength yields a 0.5cm increase in jump height.

The Intercept (a)

- The point where the regression line crosses the Y-axis.
- Formula: $a = \bar{y} - b\bar{x}$
- Represents the predicted value of Y when X is exactly exactly 0.
- Often, a **is just a mathematical anchor**. For instance, estimating jump height for someone with 0kg of leg strength is absurd! Never over-interpret the intercept outside the plausible range of your data[1].
- The regression line **always passes exactly through** the means of the data: $(\bar{x}, \bar{y})$.

- Walk students through interpreting the slope carefully; it's the most practically useful number.
- Remind them why the intercept might be completely meaningless in physical contexts like kinesiology.

14 Worked Example: Regression Equation

Using the leg strength data: $\bar{x} = 85.375$, $\bar{y} = 47.625$, $s_x = 10.056$, $s_y = 5.041$, $r = 0.998$.

Step 1: Compute the slope

$$b = r\frac{s_y}{s_x} = 0.998 \times \frac{5.041}{10.056} = 0.998 \times 0.501 = \mathbf{0.500} \text{ cm/kg}$$

Step 2: Compute the intercept

$$a = \bar{y} - b\bar{x} = 47.625 - (0.500 \times 85.375) = \mathbf{4.938} \text{ cm}$$

Step 3: Write the equation

$$\hat{y} = 4.938 + 0.500x$$

Making a prediction: If leg strength = 92 kg:

$$\hat{y} = 4.938 + 0.500(92) = 50.9 \text{ cm}$$

Slope ($b = 0.500$)

For every **1 kg increase** in leg strength, predicted jump height increases by **0.5 cm** on average.

Intercept ($a = 4.938$)

Predicted jump height when leg strength = 0. This value is **not meaningful** here — no one has zero leg strength. Do not over-interpret intercepts outside the data range.

Extrapolation

Never predict outside the observed range of X (70–100 kg in this example). The linear relationship may not hold beyond that range[1].

- Emphasize: slope is the most important and interpretable coefficient.
- Intercept is often just a mathematical anchor, not directly meaningful.

15 Residuals and Model Fit

A **residual** is the difference between the observed and predicted value[1]:

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

What residuals tell us:

- **How well the model fits:** Residuals show the error for each prediction.

- **Whether assumptions are met:** We want errors to be pure, unpredictable noise. If they are randomly scattered around zero (no pattern), it means the linear model successfully captured the relationship and variance is constant.

R^2 (in bivariate regression = r^2):

$$R^2 = 0.996$$

- 99.6% of variance in jump height explained by leg strength
- 0.4% is residual (unexplained) variance

Reading the Residual Plot:

Pattern	Diagnosis
Random scatter	**Assumptions met:** Errors are random (linearity) with constant spread (homoscedasticity).
Funnel shape	**Heteroscedasticity:** Spread of errors changes over time, violating constant variance.
Curved pattern	**Nonlinearity:** The linear model missed a curved relationship.
Outliers	**Influential points:** Specific extreme values that might distort the model.

- Random scatter in residual plot = good. Any systematic pattern = problem.
- Always show residual plots alongside the scatterplot when reporting regression.

16 Assumptions of Correlation and Regression

Both methods rely on five key assumptions[1,3]:

1. Linearity The X–Y relationship must be approximately linear. → Check: scatterplot and residual plot.

2. Homoscedasticity Variance in Y is constant across all values of X. Violations produce a funnel shape in residual plots. → Check: residual plot.

3. Independence Each observation must be independent (one data point per participant, or use appropriate repeated-measures methods). → Check: study design.

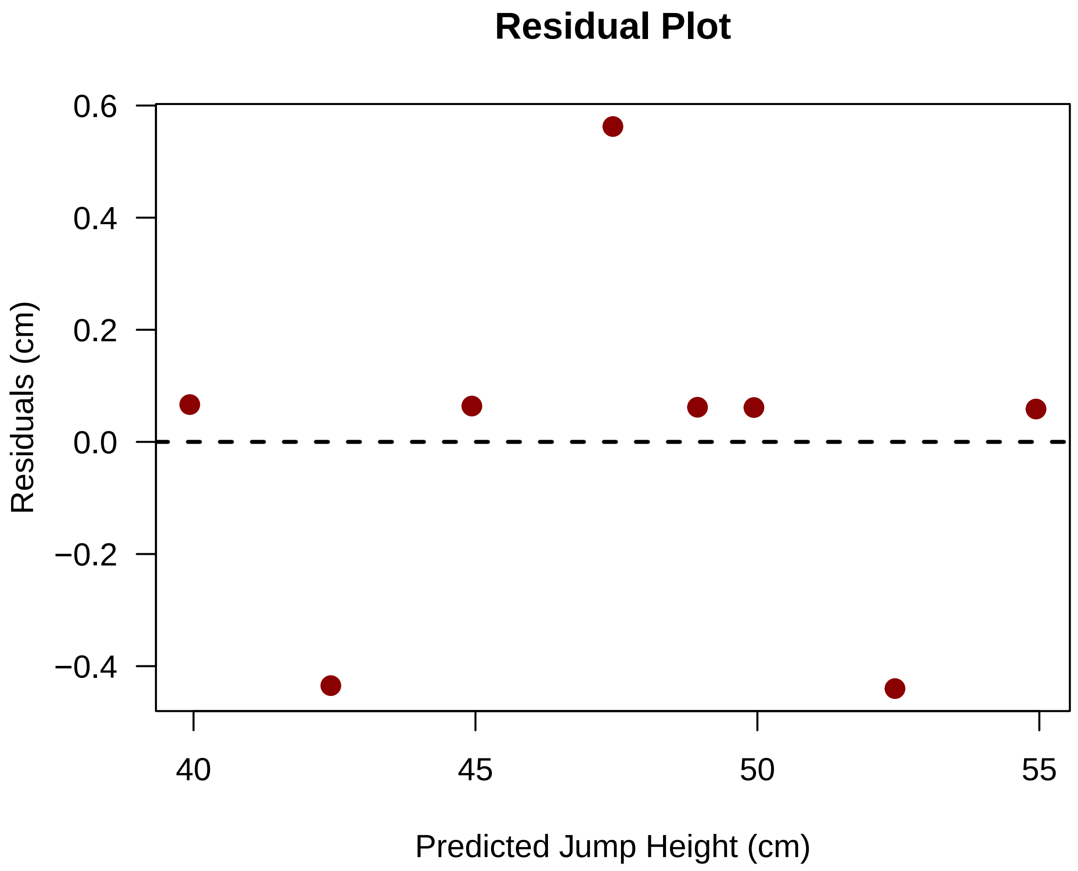


Figure 6: Residual plot: random scatter around zero indicates assumptions are met

4. Normality of residuals Residuals should be approximately normally distributed (for inference). Less critical for large samples (Central Limit Theorem). → Check: histogram or Q-Q plot of residuals.

5. No extreme outliers Outliers — especially those with high leverage (extreme X) and large residuals — can distort r and regression coefficients. → Check: scatterplot, residual plot, Cook's distance.

> ⚠ Key principle
>
> Violating assumptions — especially linearity and homoscedasticity — can produce misleading r values, biased slope estimates, and incorrect standard errors[3,8].

- Assumptions 1 and 2 are the most practically important for kinesiology students.
- Most SPSS output includes residual plots — students should know how to read them.

17 Outliers and Influential Points

Outliers can have a **disproportionate** influence on r and regression coefficients[1,8].

Types of problematic points:

- **Outlier in Y**: Large residual; pulls regression line up or down
- **High leverage point**: Extreme X value; pulls regression line toward it
- **Influential point**: Extreme X AND large residual; both distorts slope and r

What to do with outliers:

1. Check for **data entry errors** first
2. If legitimate, report results **with and without** the outlier
3. Consider **robust methods** (e.g., Spearman's r_s, robust regression)
4. Never delete outliers automatically — they may represent real biological variability

- A single influential point can change r from ~1.0 to a much lower value or inflate it artificially.
- This motivates why we always inspect scatterplots and residual plots.

18 Statistical vs. Practical Significance

Just as with hypothesis testing, a **statistically significant** correlation may not be **practically meaningful**[4,6].

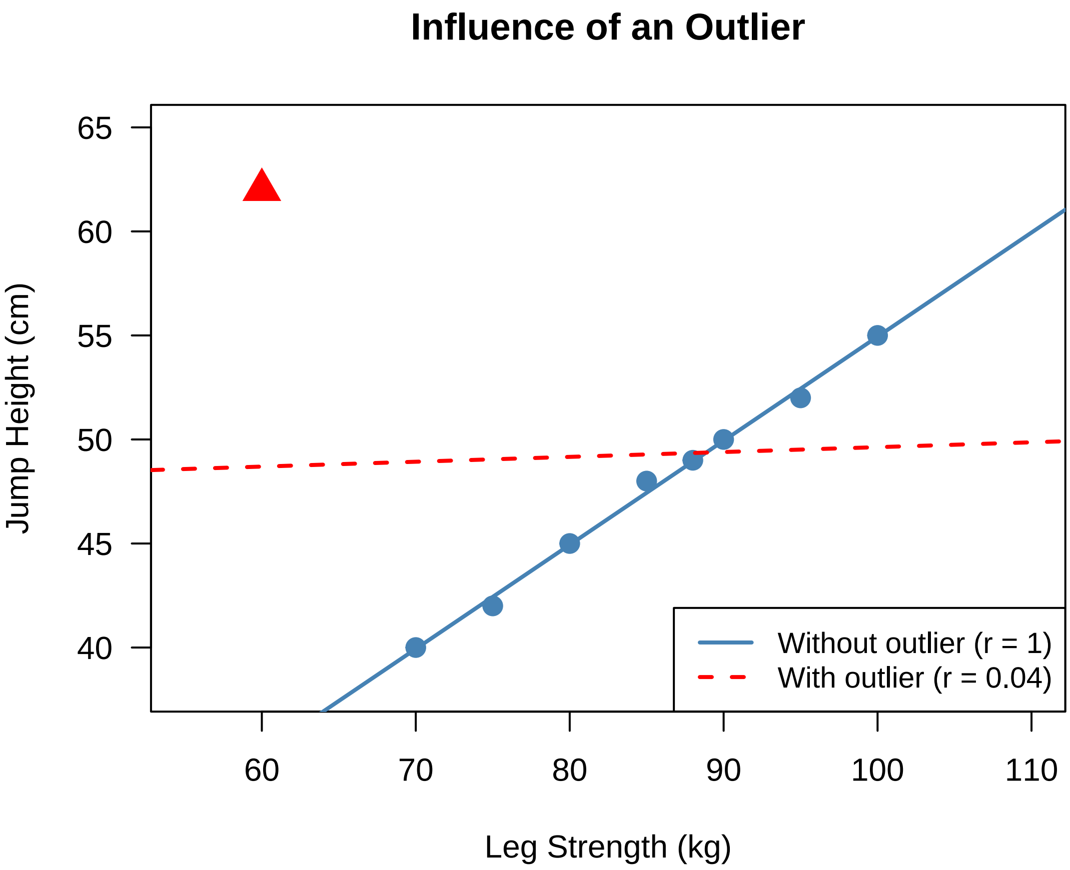


Figure 7: Effect of an influential outlier on the regression line. Blue = original data; red = with outlier added.

Examples in Movement Science:

Scenario	r	p	r^2	Practical interpretation
New training program vs. VO2max	.10	.02	1%	Stat. sig. but **trivial** (n = 400)
Strength vs. jump height	.85	.08	72%	Large effect, **underpowered** (n = 8)
Balance vs. fall risk	.55	.001	30%	Stat. sig. AND **meaningful**

Key principle:

- Large samples can make tiny r values statistically significant
- Small samples may fail to detect large real correlations
- **Always report r, r^2, and confidence intervals** alongside p-values[2,9]

! Effect size benchmarks for r^2 [@cohen1988]

- $r^2 \approx 0.01 \rightarrow$ Small effect
- $r^2 \approx 0.09 \rightarrow$ Medium effect
- $r^2 \approx 0.25 \rightarrow$ Large effect

In elite sport contexts, even very small correlations ($r \approx 0.10$–0.30) can be practically important — a 1% improvement can separate medal positions[6,7].

- This is the same issue as in hypothesis testing — sample size drives p-values.
- Encourage students to always anchor their interpretation in r and r^2 magnitudes.

19 Reporting Results in APA Style

Correlation:

> "Leg strength was significantly and positively correlated with vertical jump height, $r(6) = .998$, $p < .001$."

Note: $df = n - 2 = 6$ in parentheses.

Regression:

> "A bivariate linear regression revealed that leg strength significantly predicted vertical jump height ($b = 0.50$, $\beta = .998$), $R^2 = .996$, $F(1, 6) = 1502.1$, $p < .001$. For every 1 kg increase in leg strength, jump height increased by 0.50 cm."

Always include:

- The regression equation
- R^2 and its interpretation
- A scatterplot with the regression line
- Residual plots to support assumptions

APA formatting rules

- Use lowercase r italicized for Pearson's correlation
- Report degrees of freedom in parentheses: $r(df)$
- Report $p < .001$ when the p-value is very small
- Include confidence intervals for r when possible: $r = .998$, 95% CI $[.990, 1.000]$
- Use cautious language: "associated with," not "causes"

- APA reporting is a skill students need for their lab reports and theses.
- Emphasize the confidence interval for r — more informative than just the p-value.

20 Common Misconceptions

Misconception 1

"$r = 0$ means there is no relationship between the variables."
Correct: $r = 0$ means there is no **linear** relationship. A strong nonlinear (curved) relationship can produce $r \approx 0$. Always check a scatterplot[1].

Misconception 2

"A significant correlation proves causation."
Correct: Correlation quantifies association only. Causation requires experimental design (random assignment, manipulation, control of confounds)[1,2].

Misconception 3

"$r = 0.90$ is twice as strong as $r = 0.45$."
Correct: r is not a ratio scale. Compare using r^2: $0.90^2 = 81\%$ vs. $0.45^2 = 20\%$ variance explained — a 4× difference, not 2×.

Misconception 4

"Non-overlapping confidence intervals for r confirm the correlations are different."
Correct: Use Fisher's z-transformation to formally test whether two r values differ — visual overlap of CIs is not a reliable test.

Misconception 5

"I can use the regression equation to predict values for athletes much stronger than any in my sample."
Correct: That would be **extrapolation** — the linear relationship may not hold outside the observed range of X[3,8].

- These align with the most common student errors on lab reports and exams.

21 Workflow Summary

Use this sequence whenever examining the relationship between two continuous variables[1,3]:

Step	Action	Tool
1	**Create a scatterplot**	Visualize pattern, outliers, linearity
2	**Compute** r	Quantify strength and direction
3	**Test significance**	$t = r\sqrt{n-2}/\sqrt{1-r^2}$, $df = n-2$
4	**Fit regression model** (if prediction needed)	$\hat{y} = a + bx$; report slope, intercept, R^2
5	**Check assumptions**	Residual plot, Q-Q plot
6	**Interpret cautiously**	Correlation causation; report effect sizes

> Important
>
> The goal is not just a number — it is **understanding the nature of the relationship** and communicating it honestly, including its limitations.

22 Summary: Key Takeaways

1. **Correlation (r)** quantifies the strength and direction of a linear relationship; ranges from -1 to $+1$
2. **Always plot your data first** — r cannot detect nonlinear relationships
3. **Correlation does not imply causation** — confounding, reverse causation, and spurious correlations are always possible
4. r^2 represents the proportion of variance explained — more interpretable than r alone
5. **Bivariate regression** produces a prediction equation $\hat{y} = a + bx$; the slope tells you how much Y changes per unit of X
6. **Check assumptions**: linearity, homoscedasticity, independence, normality of residuals, no extreme outliers
7. **Extrapolation is risky** — restrict predictions to the observed range of X
8. **Statistical significance practical importance** — always report r, r^2, CI, and p together

> Important
>
> Correlation and regression are powerful descriptive tools — but responsible use requires knowing their limits.

23 Practice Questions

1. What does it mean if $r = 0$ for two variables in a Movement Science study?
2. A researcher finds $r = 0.60$ between weekly training volume and 1-RM bench press. What percentage of variance in bench press is explained by training volume?
3. Why must you always create a scatterplot before computing a correlation coefficient?
4. Explain the difference between a high-leverage point and an influential point.
5. A regression equation predicting VO2max from resting heart rate is: $\hat{y} = 80 - 0.5x$. Predict VO2max for an athlete with resting HR = 60 bpm.
6. What does a funnel shape in a residual plot indicate, and how might you address it?
7. Why is it inappropriate to conclude causation from a significant correlation between ice cream sales and sports injuries?
8. When would you prefer to report Spearman's r_s instead of Pearson's r?

24 References

1. Moore, D. S., McCabe, G. P., & Craig, B. A. (2021). *Introduction to the practice of statistics* (10th ed.). W. H. Freeman; Company.
2. Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
3. Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.
4. Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.
5. Vincent, W. J. (1999). *Statistics in kinesiology*. Human Kinetics.
6. Batterham, A. M., & Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, *1*(1), 50–57.
7. Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1–15. https://doi.org/10.2165/00007256-200030010-00001
8. Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Academic Press.
9. Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.
10. Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms/