Practical Classes 6 & 7

Correlation and Regression Analysis

Biostatistics — bivariate analysis of continuous data

chief assist. prof. Kostadin Kostadinov, MD, PhD, MPH, MEcon

Academic Year 2025/2026

Department of “Social Medicine and Public Health”

Outline

1. Foundations of Bivariate Analysis
2. Correlation Analysis
3. Regression Analysis
4. Synthesis and Pitfalls
5. Examination Questions

Foundations of Bivariate Analysis

From Univariate to Bivariate Analysis

Earlier sessions addressed **single-variable** description (means, proportions, dispersion) and **hypothesis testing** on differences between groups.

Bivariate analysis turns to a different question:

- Do two continuous variables **vary together**?
- If they do, can we describe the relationship **quantitatively**?
- Can we use one variable to **predict** the other?

These three questions are answered, in order, by **correlation**, **regression**, and the **coefficient of determination**.

Two Complementary Methods

	Correlation	Regression
Question	How strong is the linear association?	What is the model that links X to Y?
Output	Coefficient r (dimensionless, -1 to +1)	Equation $\hat{Y} = bX + a$ in original units
Symmetry	Symmetric: $r(X, Y) = r(Y, X)$	Asymmetric: predictor and outcome roles fixed
Use	Quantifying degree of association	Prediction; modelling; adjustment

Association Is Not Causation

Two variables can be statistically associated for several distinct reasons:

- **True causal effect** of X on Y
- **Reverse causation** — Y in fact influences X
- **Confounding** — a third variable causes both
- **Selection bias** in how the data were collected
- **Chance** in small samples

Correlation and regression **describe** the data; they do **not** by themselves establish a causal relationship. Establishing causation requires temporality, biological plausibility, dose–response, consistency across studies, and ideally experimental evidence.

Two Classical Reminders

Coffee and coronary heart disease. Older studies reported a positive association between coffee consumption and ischaemic heart disease. Once smoking — strongly correlated with coffee drinking — was adjusted for, much of the association disappeared. The crude correlation was largely a **confounded** one.

Firefighters and property damage. The number of firefighters at a scene correlates with the value of damaged property. Firefighters do not cause damage; both quantities reflect **fire severity** — a third variable acting as common cause.

Correlation Analysis

Correlation Coefficient — Definition

The **Pearson correlation coefficient** r measures the strength and direction of the **linear** association between two continuous variables.

$$r = \frac{\sum (d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}$$

where $d_x = X - \bar{X}$ and $d_y = Y - \bar{Y}$.

The numerator is the **covariation** of X and Y; the denominator standardises by the spread of each variable, yielding a **dimensionless** number that does not depend on measurement units.

Properties of r

- Always lies between **−1 and +1**
- $r = +1$ — **perfect positive** linear association (all points on an increasing straight line)
- $r = -1$ — **perfect negative** linear association (all points on a decreasing straight line)
- $r = 0$ — **no linear association** (but a non-linear relationship may still exist)
- r does not depend on which variable is called X and which is called Y
- r is **unitless** and unaffected by linear rescaling (e.g. converting kg to g, or mmHg to kPa)

Conventional Strength Scale

$\lvert r \rvert$	Strength	Direction by sign
0.00 – 0.30	Weak	
0.30 – 0.50	Moderate	$r > 0$: positive
0.50 – 0.70	Strong	$r < 0$: negative
0.70 – 1.00	Very strong	

Context matters: in clinical chemistry $r = 0.95$ may be unimpressive, while in social epidemiology $r = 0.20$ between a modifiable risk factor and a major outcome may be of substantial public-health importance.

Visual Patterns of Association

Pattern in scatter plot	Implied r
Tight increasing band of points	r close to +1
Tight decreasing band of points	r close to −1
Cloud with no orientation	r near 0
Curved (e.g. U-shaped) cloud	r near 0 **despite** clear association
One isolated extreme point	r may be heavily distorted

Inspect the scatter plot **before** computing r — the coefficient summarises only one feature of a pattern that may be more complex.

Worked Example — Clinical Setting

Question: In a sample of eight adult outpatients with arterial hypertension, is daily salt intake associated with systolic blood pressure?

Patient	Sodium intake X (g salt/day)	Systolic BP Y (mmHg)
1	4	132
2	5	128
3	7	144
4	8	138
5	10	150
6	11	144
7	13	156
8	14	168
Σ	72	1160

Step 1 — Compute the Means

$$\bar{X} = \frac{\sum X}{n} = \frac{72}{8} = 9 \text{ “ g/day“}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{1160}{8} = 145 \text{ “ mmHg“}$$

The patient with the lightest salt intake (4 g/day) lies **5 g below** the mean; the patient with the highest pressure (168 mmHg) lies **23 mmHg above** the mean. These individual deviations are the building blocks of the correlation coefficient.

Step 2 — Individual Deviations

$$d_x = X - \bar{X} \quad ; \quad d_y = Y - \bar{Y}$$

No.	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
1	4	132	−5	−13	65	25	169
2	5	128	−4	−17	68	16	289
3	7	144	−2	−1	2	4	1
4	8	138	−1	−7	7	1	49
5	10	150	+1	+5	5	1	25
6	11	144	+2	−1	−2	4	1
7	13	156	+4	+11	44	16	121
8	14	168	+5	+23	115	25	529
Σ	72	1160	0	0	304	92	1184

Step 3 — Compute r

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{304}{\sqrt{92 \cdot 1184}}$$

$$r = \frac{304}{\sqrt{108,928}} = \frac{304}{330.0} = 0.92$$

The deviation sums ($\Sigma d_x = 0$ and $\Sigma d_y = 0$) are useful arithmetic checks: if either column does not sum to zero, an error has been made in the means or the deviations.

Step 4 — Interpret the Result

$r = 0.92$ indicates a **strong positive** linear association between daily sodium intake and systolic blood pressure in this sample of hypertensive patients.

- Patients with higher salt intake tend to have higher systolic pressures.
- The association is **linear** — across the observed range, BP rises in roughly proportional steps with each additional gram of salt.
- The result is **descriptive**. Whether reducing salt intake would lower blood pressure in this population requires an **interventional** design — ideally a randomised trial of sodium reduction.

Assumptions Behind Pearson's r

- Both variables are **continuous** and approximately **normally distributed**
- The relationship between them is **linear**
- Observations are **independent**
- No undue influence of **extreme outliers**

When the data are markedly skewed, the relationship is monotonic but curvilinear, or one variable is ordinal, **Spearman's rank correlation coefficient** (ρ) is more appropriate. Spearman's ρ is computed on the **ranks** of the observations and is robust to outliers.

Pitfalls in Correlation

- **Outliers** — a single extreme observation can inflate or collapse r
- **Restricted range** — sampling only hypertensive patients (BP 140–180) attenuates correlations that would be visible across the full BP distribution
- **Subgroup mixing** — a positive correlation in men and a positive correlation in women may combine into a weak or even reversed overall correlation if the sexes differ in mean values (a form of **Simpson's paradox**)
- **Ecological correlations** — associations between **population averages** (e.g. national salt consumption and national stroke mortality) need not hold at the **individual** level

Beyond Pearson — Alternatives

Coefficient	When appropriate	Notes
Pearson r	Two continuous, normally distributed, linear	Most common; sensitive to outliers
Spearman ρ	Ordinal data or non-linear monotonic relationships	Computed on ranks; robust
Kendall τ	Small samples with many tied ranks	More conservative than ρ
Point-biserial	One continuous, one dichotomous variable	Mathematically equivalent to Pearson on coded data

Regression Analysis

From Association to Prediction

Correlation answers **how strongly** two variables move together. It does not allow us to **predict** the value of one from the value of the other.

Regression goes one step further:

- It models the relationship as a **mathematical equation**
- It distinguishes a **predictor** (independent variable, X) from an **outcome** (dependent variable, Y)
- It supplies a fitted line that produces a **predicted value** of Y for any chosen X
- It can be extended to several predictors at once (**multiple regression**) to adjust for confounding

Simple Linear Regression — The Model

$$\hat{Y} = b \cdot X + a$$

- $\hat{Y}$ — **predicted** value of the dependent variable
- X — value of the independent variable (predictor)
- b — **slope** (regression coefficient): expected change in Y per unit increase in X
- a — **intercept**: predicted Y when $X = 0$

Simple refers to one predictor; **multiple** to two or more.

Linear refers to a straight-line relationship; non-linear regression accommodates curves but is mathematically more demanding.

Slope and Intercept — Geometric Meaning

- The **slope** b is the **steepness** of the line: a positive b means Y rises as X rises; a negative b means Y falls as X rises.
- A **one-unit** change in X is associated with a b**-unit** change in $\hat{Y}$.
- The **intercept** a is the value of $\hat{Y}$ where the line crosses the vertical axis (i.e. at $X = 0$).
- In many medical examples the intercept has **no biological meaning** on its own — birth weight at zero gestational age, or blood pressure at zero salt intake — and serves only to anchor the line on the plot.

The Method of Least Squares

Among all possible straight lines through a scatter of points, the **best-fitting** line is the one that minimises the **sum of squared vertical distances** between the observed Y values and the line:

$$\text{minimise} \sum \left(Y - \hat{Y}\right)^2$$

These vertical distances are the **residuals** — the part of Y that the model does **not** explain.

Squaring penalises large discrepancies more than small ones and treats positive and negative residuals symmetrically. The resulting estimators of b and a are the **ordinary least squares** (OLS) estimators.

Computational Formulas

Slope:

$$b = \frac{n\sum(X \cdot Y) - \sum X \cdot \sum Y}{n\sum X^2 - (\sum X)^2}$$

Intercept:

$$a = \bar{Y} - b \cdot \bar{X}$$

The intercept formula has a useful interpretation: the regression line **always passes through the point** $(\bar{X}, \bar{Y})$. Once the slope is known, the intercept is fixed by this requirement.

Worked Example — Same Clinical Dataset

Task: Build a linear model that predicts systolic blood pressure $\hat{Y}$ from daily sodium intake X in the eight hypertensive outpatients.

Why the same dataset?

- It allows direct comparison of the two methods on identical inputs.
- The strong correlation $r = 0.92$ already confirms that a linear model is plausible.
- Where correlation summarises **strength**, regression supplies the **equation** that turns one variable into a prediction of the other.

Step 1 — Tabulate Sums for the Slope

Patient	X	Y	X · Y	X^2
1	4	132	528	16
2	5	128	640	25
3	7	144	1 008	49
4	8	138	1 104	64
5	10	150	1 500	100
6	11	144	1 584	121
7	13	156	2 028	169
8	14	168	2 352	196
Σ	72	1 160	10 744	740

Already known from Step 1 of the correlation analysis: $\bar{X} = 9$ g/day, $\bar{Y} = 145$ mmHg.

Step 2 — Compute the Slope b

$$b = \frac{n\sum XY - \sum X \cdot \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$b = \frac{8 \cdot 10,744 - 72 \cdot 1,160}{8 \cdot 740 - 72^2}$$

$$b = \frac{85,952 - 83,520}{5,920 - 5,184} = 2,\frac{432}{736} = 3.30$$

A positive slope of **3.30** indicates that, in this sample, each additional **gram of salt per day** is associated with an **increase of 3.3 mmHg** in expected systolic pressure.

Step 3 — Compute the Intercept a

$$a = \bar{Y} - b \cdot \bar{X} = 145 - 3.30 \cdot 9$$

$$a = 145 - 29.7 = 115.3$$

Fitted regression line:

$$\hat{Y} = 3.30 \cdot X + 115.3$$

In strict terms the intercept of 115.3 mmHg is the predicted SBP for $X = 0$ — an extrapolation outside the observed data. It anchors the line and should not be over-interpreted clinically.

Step 4 — Use the Model for Prediction

Sodium intake X (g/day)	Predicted SBP $\hat{Y}$ (mmHg)
4	$3.30 \cdot 4 + 115.3 = 128.5$
9 (mean)	$3.30 \cdot 9 + 115.3 = 145.0$
14	$3.30 \cdot 14 + 115.3 = 161.5$

The predicted line passes exactly through the point $(\bar{X}, \bar{Y}) = (9, 145)$, as the intercept formula guarantees. Predictions outside the observed range (here, below 4 g/day or above 14 g/day) are **extrapolations** and become progressively unreliable.

Coefficient of Determination R^2

R^2 quantifies the **proportion of variance** in the dependent variable that is **explained** by the regression model.

- Ranges from **0 to 1**; often expressed as a percentage
- In **simple linear regression**: $R^2 = r^2$
- In **multiple regression**: R^2 is the square of the multiple correlation coefficient
- Higher R^2 — model accounts for a larger share of the observed variability in Y
- $1 - R^2$ — share of variance attributable to factors **not in the model** (other determinants, measurement error, biological noise)

R^2 in the Worked Example

From the correlation step: $r = 0.92$.

$$R^2 = r^2 = (0.92)^2 = 0.85$$

Interpretation:

- Approximately **85%** of the variance in systolic blood pressure across these eight patients is **explained** by daily sodium intake.
- The remaining **15%** reflects factors not captured by the model: age, body mass, antihypertensive treatment adherence, physical activity, genetic susceptibility, and measurement variability.

What R^2 Does and Does Not Tell Us

- R^2 measures **goodness of fit**, not the **correctness** of the model. A non-linear relationship can yield a misleadingly low R^2 when fitted with a straight line.
- A high R^2 does **not** prove causation; it only quantifies how much variability the chosen predictor co-varies with.
- In fields with high measurement precision (e.g. chemistry calibration), $R^2 > 0.95$ is normal. In epidemiology and the social sciences, R^2 values of **0.10–0.40** are frequent and may still represent important relationships.
- Adding more predictors **always** increases R^2 mechanically; **adjusted** R^2 penalises model complexity and is preferred in multiple regression.

Extension — Multiple Regression

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Each **partial slope** b_i is the expected change in $\hat{Y}$ per unit increase in X_i, **holding all other predictors constant**.

This is the workhorse of medical research: it allows **adjustment for confounders**. A model relating SBP to sodium intake might adjust for age, sex, body mass index, and current antihypertensive therapy, isolating the **independent contribution** of sodium that is not explained by these covariates.

Adjusted (partial) slopes typically **differ** from unadjusted (crude) slopes — the difference quantifies the confounding.

An Illustrative Multivariable Model

Suppose, in a larger adult sample, multiple regression yields:

$$\widehat{\text{SBP}} = 90 + 0.6 \cdot \text{age} + 0.9 \cdot \text{BMI} + 0.02 \cdot \text{Na (mg/day)}$$

- SBP rises by **0.6 mmHg** per year of age, after adjusting for BMI and sodium intake.
- SBP rises by **0.9 mmHg** per BMI unit, after adjusting for age and sodium intake.
- SBP rises by **0.02 mmHg** per mg of daily sodium, after adjusting for age and BMI.

Each coefficient is **conditional** on the others — interpretation must always reference the set of variables in the model.

Assumptions of Linear Regression

- **Linearity** — the true relationship between X and Y is approximately a straight line
- **Independence** — observations are not clustered or repeated within individuals
- **Homoscedasticity** — the variance of residuals is constant across the range of X
- **Normality of residuals** — the residuals are approximately normally distributed (matters more for inference than for point estimation)
- **No undue influence of outliers** — single points should not dominate the fit

These assumptions are checked by **residual analysis**: plotting residuals against fitted values reveals non-linearity, heteroscedasticity, or influential points.

When Assumptions Fail

- **Non-linearity** — transform variables (logarithm, square root) or fit **polynomial** / **non-linear** regression
- **Non-normal residuals** — consider **generalised linear models** (e.g. logistic regression for binary outcomes, Poisson regression for counts)
- **Heteroscedasticity** — use **weighted least squares** or **robust standard errors**
- **Clustered data** — use **mixed-effects** or **generalised estimating equation** models
- **Influential outliers** — investigate the data point clinically; refit with and without it; report sensitivity

The choice of analytical model is a methodological decision, not a default — it must be justified by the data structure.

Synthesis and Pitfalls

Correlation and Regression — A Joint View

Quantity	From the worked dataset	What it conveys
r	+0.92	Strong positive linear association
b (slope)	+3.30 mmHg per g/day	Magnitude of effect in original units
a (intercept)	115.3 mmHg	Anchor of the line at $X = 0$
R^2	0.85 (85%)	Share of variance explained
$1 - R^2$	0.15 (15%)	Variance attributed to other factors

Common Mistakes to Avoid

- Reading **causation** into a correlation or a regression slope from observational data
- Reporting r without first inspecting the **scatter plot**
- Extrapolating predictions far **outside the observed range** of the predictor
- Treating the **intercept** as biologically meaningful when $X = 0$ is implausible
- Ignoring **outliers** and **influential points**
- Confusing R^2 with the **clinical importance** of the model — a low R^2 can reflect a real and modifiable risk factor; a high R^2 in a small sample can reflect overfit
- Adding predictors to inflate R^2 without theoretical justification

Reporting Standards

A complete correlation result reports: the coefficient r, its **sign**, the **sample size**, and the **p-value** (or 95% confidence interval) for the association.

A complete regression result reports: the **equation**, the **slope** with its 95% confidence interval and p-value, the **intercept**, the **coefficient of determination** R^2, and a brief statement on **assumption checks**.

Both should be accompanied by a **scatter plot** with the fitted line — visual inspection conveys what numbers alone cannot.

Examination Questions

Mandatory Open Questions — Final Exam

Given a worked correlation analysis:

- State the **value** and **sign** of the correlation coefficient.
- State the **strength** of the association (weak / moderate / strong / very strong).
- State the **direction** of the association (positive or negative) and explain what this means in clinical terms.
- State whether **causation** may be inferred from the result, and justify the answer.

Mandatory Open Questions — Final Exam

Given a worked simple linear regression model:

- State the **main finding** of the model — interpret the slope b in the units of the variables.
- State the **predicted value** of Y for a specified value of X.
- State and interpret the **coefficient of determination** R^2 — what proportion of variance in Y is explained by X?
- State the corresponding **correlation coefficient** $r = \sqrt{R^2}$ (with the same sign as the slope).

Thank you for your attention!