BP701T · B.Pharm Semester VII · Theory Notes

Biostatistics and Research Methodology

Unit III — Correlation and Regression Analysis (9 hours)

Unit III (9 hours)

Pills & Python Notes

From-zero teaching standard: every technical term is defined before use.

Syllabus coverage: Unit III (9 hours).

PILLS & PYTHON

Pharmacy • Data • AI • Healthcare

Instagram @pillsandpython · pillsandpython.com · YouTube @PillsAndPython

CHAPTER 7: Correlation Analysis — Pearson, Spearman and Multiple Correlation

Learning Objectives

- Define correlation and explain what it tells us about two variables
- Distinguish positive, negative and zero (no) correlation with examples
- Define Pearson's correlation coefficient, state its formula, assumptions, range, and interpret sign and magnitude
- Explain Spearman's rank correlation coefficient, when to use it, and the step-by-step procedure for calculating it
- Solve numerical problems on Pearson's r and Spearman's rho
- Use and read a scatter plot
- Explain the concept of multiple correlation and give its applications
- Apply correlation ideas to pharmaceutical and health-science examples
- Explain why correlation is not causation

Definitions

- **Correlation** (simple meaning): Correlation is a measure of how strongly two things change together. When one thing goes up and the other also goes up, they are positively correlated. When one goes up and the other goes down, they are negatively correlated. When the two things have no connection in how they move, there is no correlation.
- **Pearson's correlation coefficient (r)**: A number between −1 and +1 that measures the strength and direction of a *straight-line* (linear) relationship between two variables measured on a proper number scale (interval or ratio scale).
- **Spearman's rank correlation coefficient (rho, ρ or r_s)**: A number between −1 and +1 that measures the strength and direction of a relationship between two variables based on their *ranks* (order from smallest to largest) rather than their exact values. It works even when the relationship is not a straight line, or when data are ordered categories.
- **Multiple correlation**: The correlation between one variable and a *group* of two or more other variables taken together. It measures how well the group of variables, combined, predicts the single variable.
- **Scatter plot**: A picture (graph) in which each pair of values is drawn as one dot on two axes. The pattern of the dots shows whether the two variables move together.
- **Causation**: The idea that one thing directly *causes* the other. Correlation never proves causation.

Terminology

Term	Simple meaning
Variable	Any quantity that can change, e.g., dose of a drug, blood pressure
Bivariate data	Data collected in pairs — two measurements on the same item (e.g., height and weight of the same person)
Positive correlation	Both variables move in the same direction (up together, down together)
Negative correlation	The variables move in opposite directions (one up, other down)
Zero / no correlation	No clear pattern; one variable's change tells nothing about the other
Correlation coefficient	A single number (−1 to +1) that summarises strength and direction of a relationship
Linear relationship	A relationship that forms a straight line on a graph
Monotonic relationship	A relationship where one variable always increases (or always decreases) as the other increases — not necessarily in a straight line
Rank	The position of a value when data are ordered from smallest to largest (1st, 2nd, 3rd…)
Strength of relationship	How close the data points sit to a perfect pattern; weak, moderate or strong
Sign of r	The plus or minus in front of the correlation coefficient — plus means positive direction, minus means negative direction
Scatter plot	A graph with one dot for each pair of values; the "picture" of correlation
Multiple correlation coefficient (R)	A number between 0 and 1 showing how strongly one variable is related to a set of other variables combined
Outlier	A value far away from the rest of the data, which can distort a correlation value

Simple Explanations — Starting from Zero

Technical concept: Correlation is a statistical method for studying the relationship between two variables.

Simple explanation: A "variable" is anything that can take different values — the weight of a tablet, the age of a patient, the dose of a medicine. In many studies we collect two measurements from the same unit (tablet, patient, bottle). This pair of measurements is called *bivariate data* (bi = two). Correlation asks one simple question: *when the first measurement changes, does the second measurement change in a predictable way?* The correlation coefficient is a single number that answers this: its **sign** tells the direction (positive = same direction, negative = opposite direction) and its **size** (magnitude) tells the strength (close to 0 = weak, close to 1 or −1 = strong).

Example: Take the daily dose of a blood-pressure medicine and the patient's blood-pressure reading. As the dose increases, blood pressure usually decreases — a *negative correlation*. Now take the dose and the number of tablets swallowed: both increase together — a *positive correlation*. Now take the dose and the patient's height: no connection — *no correlation*.

Diagram/flowchart:

```
Two measurements from the same unit
      (dose, blood pressure)
               |
      Does changing X change Y
      in a predictable way?
               |
    +----------+----------+
    |          |          |
Same direction  Opposite   No pattern
(X up, Y up)   direction  (dots scattered
    |          (X up, Y     everywhere)
    |           down)            |
Positive       Negative      Zero correlation
correlation    correlation     r ≈ 0
```

Key takeaway: Correlation is about *pairs* of measurements moving together; the sign tells direction and the size tells strength.

Technical concept: Pearson's correlation coefficient r measures linear association.

Simple explanation: "Linear" means the pattern looks like a straight line. Pearson's r only measures how close the dots sit to one straight line. Formula:

$r = \Sigma(x - \bar{x})(y - \bar{y}) / \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$

Read this slowly: for each pair, find how far x is from the average of x ($x - \bar{x}$) and how far y is from the average of y ($y - \bar{y}$), multiply them, and add up all those products (this is the top, the "covariance part"). Then divide by a denominator that standardises the scales of x and y. The result is always between −1 and +1.

Assumptions of Pearson's r: (1) both variables are measured on a number scale with meaningful distances (interval or ratio scale — e.g., weight in mg, temperature in °C); (2) the relationship is approximately a straight line; (3) no extreme outliers distorting the picture; (4) observations are independent of each other.

Range and interpretation: r ranges from −1 to +1. r = +1 means a perfect straight-line increase; r = −1 means a perfect straight-line decrease; r = 0 means no straight-line relationship. Rough strength guide: 0 to ±0.3 weak; ±0.3 to ±0.7 moderate; ±0.7 to ±1 strong. The sign tells direction; the magnitude (size, ignoring the sign) tells strength.

Example: Suppose a textbook example gives r = +0.85 between drug concentration in blood and the pain-relief score. Sign is plus → as concentration rises, pain relief rises. Magnitude 0.85 → strong straight-line relationship.

Diagram/flowchart:

```
     r = -1          r = 0          r = +1
dots on a line    dots scattered  dots on a line
falling to the    like stars in    rising to the
right               the sky          right
\\\\\\               . . . .       //////
Perfect negative    No linear       Perfect positive
```

Key takeaway: Pearson's r answers "how straight-line is the relationship?" — always check the assumptions before trusting the number.

Technical concept: Spearman's rank correlation coefficient (ρ or r_s) measures monotonic association using ranks.

Simple explanation: A "monotonic" relationship means: as one variable keeps increasing, the other *always keeps increasing too* (or always keeps decreasing) — but not necessarily in a straight line; the curve may bend. "Rank" means replacing each value by its order: smallest gets rank 1, next rank 2, and so on. Spearman's method ranks both variables, then applies Pearson's formula to the ranks. Because it uses ranks, it works when: (1) the relationship is curved rather than straight; (2) the data are ordered categories (e.g., severity grades mild/moderate/severe); (3) there are outliers, which have less effect on ranks than on raw values.

Procedure for calculation (step-by-step):

1 List the pairs (x, y).

2 Rank the x values from smallest to largest (rank 1 = smallest). If two values are equal ("tied"), give them the average of the ranks they would share.

3 Rank the y values the same way.

4 For each pair, compute d = (rank of x) − (rank of y).

5 Square each d and add them up: Σd^2.

6 Apply the formula: $\rho = 1 - [6\Sigma d^2] / [n(n^2 - 1)]$, where n is the number of pairs.

Example: Five tablets tested for hardness (kg/cm²) and dissolution time (minutes). We care whether "harder → dissolves slower" in a general sense, even if not a straight line — Spearman is the right tool.

Diagram/flowchart:

```
Raw data (x, y)
      |
Replace values by ranks (1, 2, 3, …)
      |
Compute d = rank(x) − rank(y) for each pair
      |
Square and sum: Σd²
      |
ρ = 1 − 6Σd² / [n(n² − 1)]
      |
Same interpretation as Pearson (−1 to +1),
but valid for curves and ordered categories
```

Key takeaway: Spearman answers "do the variables consistently rise/fall together?" using order, not exact values.

Technical concept: Multiple correlation.

Simple explanation: Ordinary (simple) correlation studies *two* variables. Multiple correlation studies *one* variable against a *group* of two or more variables at once. It asks: "if I combine all these predictor variables together, how well can I predict the single outcome?" The multiple correlation coefficient R lies between 0 and 1 (it has no negative side, because combining variables can't point in a "negative direction" overall). R^2 tells what fraction of the outcome's variation is explained by the whole group.

Example: A pharmacist wants to predict a tablet's hardness (one outcome) from three factors together: compression force, binder concentration, and granule moisture. The multiple correlation R between hardness and the group of three tells how well the three factors combined predict hardness.

Key takeaway: Multiple correlation generalises the idea of correlation to one outcome versus a team of predictors.

Classifications

Types of correlation by direction:

- **Positive correlation:** both variables move in the same direction. Example: number of cigarettes smoked per day and nicotine level in blood.
- **Negative correlation:** variables move in opposite directions. Example: exercise hours per week and resting heart rate.
- **Zero (no) correlation:** no systematic pattern. Example: shoe size and marks in pharmacology (no connection).

Types of correlation by number of variables:

- **Simple correlation:** between exactly two variables (Pearson, Spearman).
- **Multiple correlation:** between one variable and a set of two or more variables.

Types of correlation by form of relationship:

- **Linear correlation:** the pattern is a straight line — measured by Pearson's r.
- **Non-linear (curvilinear) correlation:** the pattern bends (e.g., rises then levels off) — Pearson's r can be misleadingly small; Spearman's ρ or other methods suit better.

Processes

Process 1: Computing Pearson's r (step-by-step)

1 List the n pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

2 Compute the means: $\bar{x} = \Sigma x/n$, $\bar{y} = \Sigma y/n$.

3 For each pair compute $(x - \bar{x})$, $(y - \bar{y})$, their product $(x - \bar{x})(y - \bar{y})$, and the squares $(x - \bar{x})^2$ and $(y - \bar{y})^2$.

4 Add up: $\Sigma(x - \bar{x})(y - \bar{y})$, $\Sigma(x - \bar{x})^2$, $\Sigma(y - \bar{y})^2$.

5 $r = \Sigma(x - \bar{x})(y - \bar{y}) \div \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$.

6 Interpret: check sign (direction), magnitude (strength), and verify assumptions hold.

Process 2: Reading a scatter plot (step-by-step)

1 Draw x (independent variable, the "cause" candidate) on the horizontal axis, y (the "response") on the vertical axis.

2 Plot one dot per pair.

3 Look at the *shape*: does the cloud of dots slope upward (positive), downward (negative), form no pattern (no correlation), or bend (non-linear)?

4 Look at the *spread*: tightly packed around a line = strong; widely scattered = weak.

5 Note any single dot far from the rest (outlier) — investigate it before calculating r.

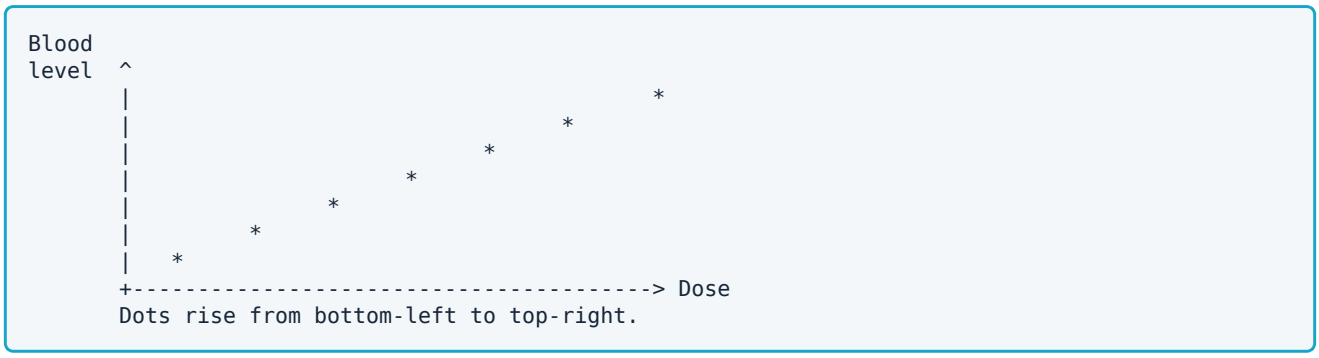
Process 3: Computing Spearman's ρ (step-by-step)

As given in the Simple Explanations section: rank x, rank y (average ranks for ties), compute d = rank difference per pair, sum the squared differences Σd^2, then $\rho = 1 - 6\Sigma d^2/[n(n^2 - 1)]$.

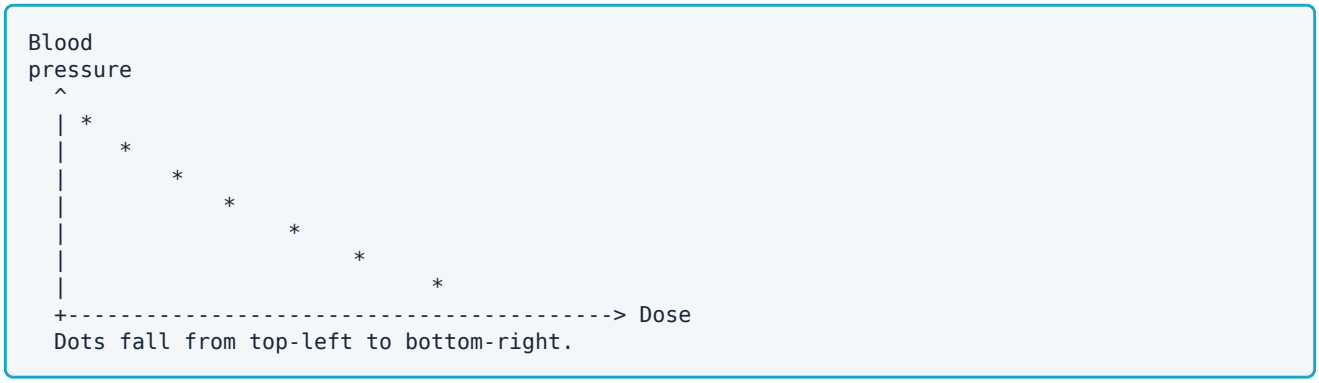
Textual Diagrams and Flowcharts

Textual scatter plot illustrations

Positive correlation (dose vs blood level of drug):



Negative correlation (dose vs blood pressure):



No correlation (height vs tablet hardness):

```
Tablet
hardness
  ^
  |   *        *     *
  |        *             *     *
  |     *     *   *         *
  |  *              *  *       *
  +----------------------------------------> Height
  Dots scattered randomly — no pattern.
```

Choosing Pearson vs Spearman — flowchart

```
        START: two variables, one pair per item
                        |
       Are both variables on a proper
       number scale (e.g., mg, mmHg, min)?
              /              \
            YES               NO (categories,
             |               e.g., mild/moderate/severe)
             |                        |
   Is the pattern approx.            USE SPEARMAN
   a straight line?                   (rank correlation)
       /       \
     YES        NO (curved pattern)
      |               |
  USE PEARSON      USE SPEARMAN
  (r)              (ρ)
   Check assumptions: no wild outliers, independent observations
```

> INFOGRAPHIC CONCEPT: Correlation Compass

The Correlation Compass. Draw a horizontal number line from −1 (far left) through 0 (centre) to +1 (far right). At −1, place a steep falling arrow labelled "Perfect negative: one up, other down". At 0, place a shapeless cloud of dots labelled "No correlation: no pattern". At +1, place a steep rising arrow labelled "Perfect positive: together up, together down". Mark three zones on each half: 0–0.3 "weak" (faint dots), 0.3–0.7 "moderate" (loose band), 0.7–1 "strong" (tight line). Below the line, a warning box: "Direction and strength ONLY — the compass never tells you who caused what." Colour scheme: cool blue for negative, grey for zero, warm green for positive.

Tables

Table 1: Positive, negative and no correlation at a glance

Feature	Positive correlation	Negative correlation	No correlation
Direction	Same direction	Opposite directions	No direction
Scatter plot	Dots rise to the right	Dots fall to the right	Dots scattered randomly
Sign of r	+	−	≈ 0
Pharmacy example	Dose vs plasma drug concentration	Dose of antihypertensive vs blood pressure	Patient's height vs tablet disintegration time
Everyday example	Study hours vs marks	Speed of fan vs room temperature drop needed	Umbrella sales vs cricket scores

Table 2: Pearson vs Spearman

Feature	Pearson's r	Spearman's ρ (rho)
Measures	Straight-line (linear) relationship	Monotonic relationship (consistently up or down, straight or curved)
Data required	Both variables on number scale (interval/ratio)	Works on number scale AND ordered categories
Calculation	Uses actual values and means	Uses ranks (1st, 2nd, 3rd...); averages tied ranks
Sensitive to outliers	Yes, strongly	Less — ranks compress extreme values
Curved relationships	Can give misleadingly small r	Handles them well
Formula	$\Sigma(x-\bar{x})(y-\bar{y})/\sqrt{[\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2]}$	$1 - 6\Sigma d^2/[n(n^2 - 1)]$
Range	−1 to +1	−1 to +1

Worked Examples and Calculation Boxes

CALCULATION BOX 1 — Pearson's r, full working

Problem: Five syrup bottles were tested. Their viscosity (x, in cP) and pouring time (y, in seconds) were measured:

Bottle	x (viscosity, cP)	y (pouring time, s)
1	10	12
2	20	18
3	30	27
4	40	33
5	50	40

Calculate Pearson's correlation coefficient and interpret it.

Solution — every step shown:

1 n = 5. $\Sigma x = 10+20+30+40+50 = 150 \rightarrow \bar{x} = 150/5 = 30$. $\Sigma y = 12+18+27+33+40 = 130 \rightarrow \bar{y} = 130/5 = 26$.

2 Build the working table:

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	12	−20	−14	280	400	196
20	18	−10	−8	80	100	64
30	27	0	1	0	0	1
40	33	10	7	70	100	49
50	40	20	14	280	400	196
Total				**710**	**1000**	**506**

1 Numerator $\Sigma(x-\bar{x})(y-\bar{y}) = 710$.

2 Denominator = $\sqrt{[1000 \times 506]} = \sqrt{506000} \approx 711.34$.

3 r = 710 ÷ 711.34 ≈ **+0.998**.

4 **Interpretation:** sign positive → as viscosity rises, pouring time rises (same direction). Magnitude 0.998 → very strong, near-perfect positive linear relationship.

Key takeaway: Pearson's r near +1 means the points sit almost exactly on a rising straight line.

CALCULATION BOX 2 — Spearman's ρ, full working

Problem: Six patients ranked their pain relief (x-rank given by a score) and the doctor gave a mobility rank (y). The data (already in ranks for simplicity) are:

Patient	Rank of x (pain relief)	Rank of y (mobility)
A	1	2
B	2	1
C	3	4
D	4	3
E	5	6
F	6	5

Calculate Spearman's rank correlation coefficient and interpret it.

Solution — every step shown:

1 n = 6. For each pair compute d = rank(x) − rank(y) and d^2:

Patient	rank x	rank y	d = x − y	d^2
A	1	2	−1	1
B	2	1	+1	1
C	3	4	−1	1
D	4	3	+1	1
E	5	6	−1	1
F	6	5	+1	1
Total				**6**

1 $\Sigma d^2 = 6$.

2 $\rho = 1 - [6 \times \Sigma d^2] / [n(n^2 - 1)] = 1 - (6 \times 6) / [6 \times (36 - 1)] = 1 - 36/210 = 1 - 0.1714 \approx$ **+0.829**.

3 **Interpretation:** positive and strong — patients with better pain relief ranks consistently also have better mobility ranks. Note the sign convention: here rank 1 = lowest, so positive ρ means "higher relief goes with higher mobility."

Key takeaway: Spearman works on order alone — it captures the consistent rise/fall even if the raw relationship is curved.

Example 3 — Multiple correlation concept (illustrative, no heavy arithmetic)

A quality-control pharmacist records three things for 20 tablet batches: compression force (kN), binder concentration (%), and the resulting tablet hardness (kp). The simple correlations are: hardness–force r = +0.60, hardness–binder r = +0.45, force–binder r = +0.20. The *multiple* correlation R between hardness and the pair (force, binder) combined is R = 0.72 — larger than either simple r, because the two predictors together explain more of the hardness variation than either alone. $R^2 = 0.72^2 = 0.5184$, so about 52% of the variation in hardness is explained by the two factors combined.

Key takeaway: Multiple R shows the joint predictive power of a team of variables for one outcome.

Analogies

- **Two friends on a see-saw:** Positive correlation is like two friends on the *same* side of a see-saw — both go up and down together. Negative correlation is like friends on *opposite* ends — one rises while the other falls. No correlation is like two children on separate swings, moving independently.
- **Shadows:** A tall tree casts a long shadow — height and shadow length are positively correlated, but the shadow doesn't *cause* the tree to be tall. A gentle first step toward "correlation is not causation."
- **Thermometer and temperature:** The mercury column and the true temperature are strongly correlated — the column *reflects* the temperature but does not create it. Many pharmacy correlations are like this: useful for prediction, not proof of cause.

Pharmacy Relevance

- **Dose vs response:** The relationship between drug dose and pharmacological response is often strongly correlated — the basis of dose–response curves.
- **Concentration vs absorbance:** In UV-visible spectrophotometry, absorbance is strongly positively correlated with concentration (Beer's law) — this is why calibration curves work (see Chapter 8).
- **Tablet hardness vs disintegration time:** Harder tablets usually take longer to disintegrate — positive correlation used in formulation optimisation.
- **Storage time vs drug content:** Drug content of a formulation usually falls as storage time increases — negative correlation that underpins stability studies and shelf-life estimation.
- **Patient compliance vs treatment success:** Higher adherence scores are positively correlated with better treatment outcomes — important in clinical pharmacy and pharmacovigilance counselling.
- **Multiple correlation in formulation:** Tablet hardness depends jointly on compression force, binder level and moisture — multiple correlation measures how well these factors together explain hardness.

India Context

- In Indian pharmaceutical QC laboratories, correlation analysis is routinely used to compare a new generic formulation's dissolution profile with the reference product's — a high correlation between the two dissolution curves supports similarity of drug release.

- Stability studies conducted as per Indian regulatory expectations track drug content against storage time; the negative correlation observed helps estimate shelf life of products sold in India's hot and humid climatic zones.
- Public-health researchers in India study correlations such as rainfall vs vector-borne disease incidence across districts to prioritise health-programme spending.
- In hospital pharmacies, pharmacists correlate the number of counselling sessions with patient adherence scores to demonstrate the value of clinical pharmacy services.

Key Points

- Correlation measures how two variables move together; it says nothing about cause.
- Positive correlation: same direction. Negative: opposite directions. Zero: no pattern.
- Pearson's r measures straight-line relationships for numeric data; range −1 to +1.
- Sign of r = direction; magnitude of r = strength (weak < 0.3, moderate 0.3–0.7, strong > 0.7).
- Pearson's assumptions: numeric scale, approximately linear, no wild outliers, independent observations.
- Spearman's ρ uses ranks; suits curved relationships, ordered categories, and outlier-prone data.
- Scatter plots should always be drawn before calculating r — the picture reveals curves, outliers and clusters that a single number hides.
- Multiple correlation R (0 to 1) measures one outcome against a group of predictors combined.

Common Confusions

- **"Correlation is not causation" (the most important confusion):** Ice-cream sales and drowning cases are positively correlated — but ice cream doesn't cause drowning. Both rise in summer (heat is the common cause). In pharmacy: a correlation between a drug and recovery does not prove the drug caused the recovery — always consider a third factor (e.g., younger patients both receive the new drug and recover faster).
- **r = 0 does not always mean "no relationship":** It only means no *straight-line* relationship. A perfect U-shaped curve (e.g., drug effect rising then falling with dose) can give r ≈ 0. Always look at the scatter plot.
- **Strength vs significance:** A "strong" correlation (large r) is not the same as a "statistically significant" one (unlikely to be due to chance). With huge samples, even tiny r values can be significant; with tiny samples, large r may be non-significant.
- **Pearson vs Spearman mix-up:** Pearson needs a straight line and numeric data; Spearman needs only a consistent rise/fall and handles ranks. Using Pearson on a curved pattern underestimates the true association.
- **Multiple R is never negative:** It runs 0 to 1 only — there is no "negative multiple correlation."

Common Mistakes

- Calculating r without first drawing a scatter plot and then missing a curved pattern or an outlier that dominates the result.
- Including one extreme outlier that artificially inflates (or deflates) r — e.g., one batch with a recording error.
- Using Pearson's r on ordered categories (e.g., pain grades 1–5 treated as exact measurements) without checking whether that is justified — Spearman is usually safer.
- Interpreting a strong correlation as proof that X causes Y, and then recommending an action on that basis.
- Forgetting that r is unit-free and symmetric: r(x, y) = r(y, x) — correlation has no "dependent" and "independent" variable (that distinction belongs to regression, Chapter 8).
- Reporting r to four decimal places from a tiny sample (n = 5) — implying false precision.

Mnemonics

- **"Sign shows the Side, Size shows the Strength":** The plus/minus tells which side the variables move; the number's size tells how strongly.
- **"Pearson is Particular, Spearman is a Sorter":** Pearson is particular — it demands numeric data and a straight line. Spearman is a sorter — it just sorts (ranks) and compares order.
- **"PLOT before you r":** Always draw the scatter PLOT before computing r.
- **"Correlation is Company, Causation is Cause":** Correlation only tells you two variables keep company; causation claims one causes the other — never confuse the two.

Exam Summary

- Definition of correlation; positive, negative and zero correlation with one example each.
- Pearson's r: definition, formula, assumptions (numeric scale, linearity, no extreme outliers, independence), range −1 to +1, interpretation of sign (direction) and magnitude (strength).
- Spearman's ρ: concept of ranks, when to use (curved/monotonic pattern, ordered categories, outliers), calculation steps, formula $1 - 6\Sigma d^2/[n(n^2-1)]$.
- Scatter plot: construction and reading (direction, strength, outliers, non-linearity).
- Multiple correlation: concept (one variable vs a group), range 0 to 1, applications in formulation and clinical prediction.
- Pharmaceutical applications: dose–response, concentration–absorbance, hardness–disintegration, stability (content vs time).
- Correlation ≠ causation — with an example.
- Numericals: compute Pearson's r from a small table; compute Spearman's ρ from ranks; interpret both.

One-Minute Revision

Correlation = two variables moving together. Positive: same direction; negative: opposite; zero: no pattern. Pearson's r = straight-line measure for numeric data, −1 to +1; sign = direction, size = strength. Spearman's ρ = rank-based, works for curves and categories. Always draw the scatter plot first. Multiple correlation = one outcome vs a team of predictors (0 to 1). Correlation never proves causation.

Questions

Short-answer questions

1 Define correlation in one sentence.

2 Give one pharmacy example each of positive, negative and zero correlation.

3 State the range of Pearson's correlation coefficient.

4 What does the sign of r tell us? What does its magnitude tell us?

5 List any two assumptions of Pearson's correlation coefficient.

6 When should Spearman's rank correlation be preferred over Pearson's?

7 What is a scatter plot and what is it used for?

8 Define multiple correlation.

9 What is the range of the multiple correlation coefficient?

10 Explain "correlation is not causation" with one example.

Long-answer questions

1 Define Pearson's correlation coefficient. Give its formula, state its assumptions, and explain how to interpret its sign and magnitude with examples.

2 Describe the procedure for calculating Spearman's rank correlation coefficient. Explain with an example when it is preferable to Pearson's method.

3 What is a scatter plot? Explain how scatter plots reveal direction, strength, non-linearity and outliers, with sketches.

4 Explain the concept of multiple correlation with two pharmaceutical applications.

5 Discuss the real-life applications of correlation analysis in pharmaceutical and health sciences.

MCQs (10) with answer key

1 Pearson's correlation coefficient ranges between: (a) 0 and 1 (b) −1 and +1 (c) −∞ and +∞ (d) 0 and 100. **Answer: (b)**

2 A correlation coefficient of −0.85 indicates: (a) weak positive (b) strong positive (c) strong negative (d) no correlation. **Answer: (c)**

3 Which plot is used to visualise correlation? (a) Bar chart (b) Scatter plot (c) Pie chart (d) Histogram. **Answer: (b)**

4 Spearman's rank correlation is most suitable when: (a) data are in ordered categories (b) the relationship is perfectly linear (c) only one variable exists (d) data are nominal. **Answer: (a)**

5 If r = 0 between dose and effect, it definitely means: (a) no relationship of any kind (b) no linear relationship (c) a negative relationship (d) a perfect curve. **Answer: (b)**

6 The multiple correlation coefficient ranges between: (a) −1 and +1 (b) 0 and 1 (c) 0 and 100 (d) −∞ and +∞. **Answer: (b)**

7 In the formula $\rho = 1 - 6\Sigma d^2/[n(n^2-1)]$, d stands for: (a) difference of values (b) difference of ranks (c) standard deviation (d) degrees of freedom. **Answer: (b)**

8 A strong positive correlation between ice-cream sales and drowning proves: (a) ice cream causes drowning (b) drowning causes ice-cream sales (c) nothing about causation — a third factor (summer heat) explains both (d) the data are wrong. **Answer: (c)**

9 Which assumption is required for Pearson's r? (a) Relationship is approximately linear (b) Data must be in ranks (c) Only categorical data allowed (d) Sample size must exceed 1000. **Answer: (a)**

10 Tied ranks in Spearman's method are handled by: (a) deleting the tied values (b) assigning the average of the ranks they share (c) assigning rank 1 to all (d) switching to Pearson's formula. **Answer: (b)**

Viva questions

1 What is the difference between positive and negative correlation?

2 Can correlation prove that a drug caused a cure? Why or why not?

3 Why do we draw a scatter plot before calculating r?

4 What happens to Pearson's r if one extreme outlier is present?

5 In what sense is multiple correlation "multiple"?

Application-based questions

1 A QC chemist finds r = +0.97 between the concentration of a standard solution and its absorbance. What does this mean, and what is the next analytical step? (Answer: very strong positive linear relationship — suitable for building a calibration curve; see Chapter 8.)

2 In a stability study, drug content falls as storage months increase, with r = −0.91. Interpret this and state one practical use. (Answer: strong negative linear relationship — supports estimating shelf life from the trend.)

3 A researcher computes Pearson's r = 0.12 for dose vs effect but the scatter plot shows a clear inverted-U shape. What went wrong and what should be done? (Answer: the relationship is non-linear, so Pearson's r is misleading; inspect the curve and consider Spearman's ρ or a non-linear model — do not report "no relationship.")

References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)

CHAPTER 8: Regression Analysis — Simple Linear, Least Squares, Multiple Regression and Over/Underfitting

Learning Objectives

- Define regression and distinguish dependent and independent variables
- Explain the concept of simple linear regression and write the simple linear regression equation
- Derive and compute the slope and intercept by the method of least squares
- Define the coefficient of determination (R^2) and interpret it
- Interpret the output of a regression analysis
- State applications of regression in pharmacy
- Explain the relationship between the regression coefficient and the correlation coefficient
- Solve problems on simple linear regression for predicting dependent values, with pharmaceutical examples
- Explain the concept and applications of multiple linear regression
- Explain the meaning of overfitting and underfitting

Definitions

- **Regression analysis** (simple meaning): A method for finding a mathematical rule (an equation) that predicts the value of one variable from the value of another variable. While correlation only says "these move together," regression gives you the actual line and lets you predict.
- **Independent variable (x)**: The variable we use to make the prediction — the input, the "cause" candidate, plotted on the horizontal axis. Also called the predictor or explanatory variable.
- **Dependent variable (y)**: The variable we want to predict — the output, the "effect," plotted on the vertical axis. It is called dependent because its predicted value depends on x.
- **Simple linear regression**: Regression with exactly one independent variable and one dependent variable, where the relationship is a straight line. Equation: $\hat{y} = a + bx$ (or $y = \beta_0 + \beta_1 x$).
- **Slope (b or β_1)**: The number in the regression equation that tells how much y changes, on average, when x increases by one unit.
- **Intercept (a or β_0)**: The predicted value of y when x = 0 — where the line crosses the vertical axis.
- **Method of least squares**: The rule for choosing the "best" line: pick the line that makes the sum of the squared vertical distances (residuals) between the actual points and the line as small as possible.
- **Residual**: The vertical gap between an actual data point and the regression line (actual y − predicted y). Small residuals = good fit.

- **Coefficient of determination (R^2)**: A number between 0 and 1 telling what fraction (proportion) of the variation in y is explained by the regression line. $R^2 = 0.81$ means 81% of y's variation is explained by x.
- **Multiple linear regression**: Regression with two or more independent variables predicting one dependent variable.
- **Overfitting**: When a model is made so complicated that it learns the noise and quirks of the training data instead of the true pattern — it fits the sample perfectly but predicts new data badly.
- **Underfitting**: When a model is too simple to capture the true pattern — it fits even the training data poorly (e.g., fitting a straight line to clearly curved data).

Terminology

Term	Simple meaning
Regression	Predicting one variable from another using an equation
Predictor (x)	The input variable used for prediction (independent variable)
Response (y)	The output variable being predicted (dependent variable)
Regression line	The straight line that best fits the data
Slope (b)	Change in y per one-unit change in x
Intercept (a)	Predicted y when x = 0
Residual	Actual y minus predicted y — the leftover error for one point
Least squares	The "best line" rule: minimise the sum of squared residuals
R^2 (coefficient of determination)	Fraction of y's variation explained by the model (0 to 1)
Fitted / predicted value ($\hat{y}$)	The y value the line gives for a given x
Regression coefficient	Another name for the slope (b) in simple linear regression
Extrapolation	Predicting y for an x far outside the observed data range — risky
Interpolation	Predicting y for an x inside the observed data range — safer
Multiple regression	One dependent variable predicted from several independent variables
Overfitting	Model too complex — memorises noise, fails on new data
Underfitting	Model too simple — misses the real pattern
Correlation (for contrast)	Measures strength of association; has no x and y roles

Simple Explanations — Starting from Zero

Technical concept: Regression analysis models the dependence of y on x.

Simple explanation: In Chapter 7, correlation treated x and y as equals — there was no "input" and "output." Regression is different: we *choose* one variable to be the input (independent variable, x) and one to be the output (dependent variable, y), and we find a formula so that, given any x, we can *predict* y. Think of a medicine vending machine: you put in x (the money), the machine's rule gives you y (the item). Regression finds the machine's rule from past data. "Simple" means one input; "linear" means the rule is a straight line: predicted $y = a + b \cdot x$.

Example: A pharmacist measures the absorbance (y) of standard drug solutions of known concentration (x). Regression finds the line absorbance = a + b·concentration. Later, for an unknown sample, the pharmacist measures only absorbance and uses the line backwards to find the unknown concentration — the daily working of calibration curves.

Diagram/flowchart:

```
Past paired data (x, y)
      |
Choose x = input (independent),
       y = output (dependent)
      |
Find best straight line: ŷ = a + bx
      |
New x arrives → plug into equation
      |
Get predicted ŷ (estimate of y)
```

Key takeaway: Regression turns "x and y move together" into "give me x and I will predict y."

Technical concept: The method of least squares.

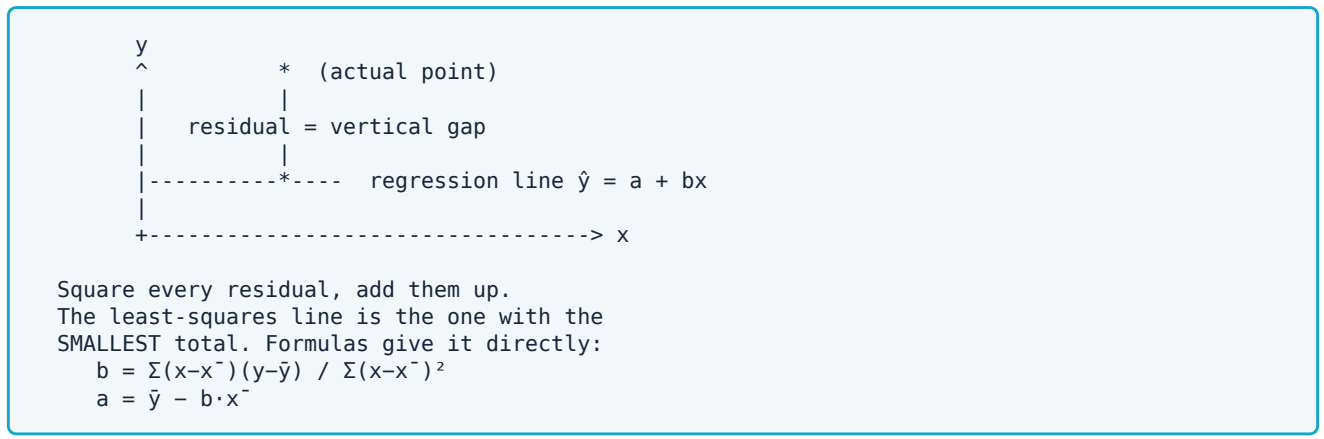
Simple explanation: Many different straight lines could be drawn through a cloud of dots. Which is "best"? For each point, measure the vertical gap between the point and the line — this gap is the *residual* (how wrong the line is for that point). Square each gap (squaring makes positive and negative gaps both count, and punishes big gaps more) and add them all up. The least-squares line is the one that makes this total — the *sum of squared residuals* — as small as possible. Mathematics gives direct formulas for the winning line's slope and intercept:

b (slope) $= \Sigma(x - \bar{x})(y - \bar{y}) / \Sigma(x - \bar{x})^2$ a (intercept) $= \bar{y} - b\cdot\bar{x}$

So: first compute the slope b from the data, then get the intercept a from the means and the slope. The resulting line always passes through the point $(\bar{x}, \bar{y})$ — the centre of the data.

Example: Using the syrup viscosity–pouring-time data from Chapter 7 (Calculation Box 1): $\bar{x} = 30$, $\bar{y} = 26$, $\Sigma(x-\bar{x})(y-\bar{y}) = 710$, $\Sigma(x-\bar{x})^2 = 1000$. Then $b = 710/1000 = 0.71$ seconds per cP, and $a = 26 - 0.71\times30 = 26 - 21.3 = 4.7$. The regression line is: predicted pouring time = 4.7 + 0.71 × viscosity. For a new bottle with viscosity 35 cP, predicted pouring time = 4.7 + 0.71×35 = 4.7 + 24.85 = 29.55 seconds.

Diagram/flowchart:



Key takeaway: Least squares picks the line with the smallest total squared vertical error; the formulas give slope first, then intercept.

Technical concept: The coefficient of determination, R^2.

Simple explanation: R^2 answers: "what fraction of the ups and downs of y is explained by x through our line?" It is a number between 0 and 1 (often written as a percentage). In *simple* linear regression, R^2 is exactly the square of Pearson's r. $R^2 = 0.81$ means 81% of the variation in y is explained by the straight-line relationship with x; the remaining 19% is due to other factors and random noise. The leftover part ($1 - R^2$) is the *unexplained* variation.

Example: If r = +0.9 between tablet compression force and hardness, then $R^2 = 0.81$ — compression force explains 81% of hardness variation. If $R^2 = 0.15$, the line explains almost nothing — x is a poor predictor of y (or the relationship is not linear).

Key takeaway: R^2 is the "report card" of the regression line — the share of y's variation the line accounts for.

Technical concept: Relationship between the regression coefficient (slope b) and the correlation coefficient r.

Simple explanation: The slope b and Pearson's r are cousins, not twins. Their formulas share the same numerator:

$b = \Sigma(x-\bar{x})(y-\bar{y}) / \Sigma(x-\bar{x})^2$, while $r = \Sigma(x-\bar{x})(y-\bar{y}) / \sqrt{[\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2]}$.

So $b = r \times (s_y / s_x)$, where s_y and s_x are the standard deviations of y and x. This tells three things: (1) b and r always have the same sign — if r is positive, the line slopes upward; (2) r is unit-free (always −1 to +1) but b carries units (e.g., "seconds per cP"); (3) if the data are standardised (converted to z-scores), the slope *equals* r.

Example: In the syrup example, r = 0.998 and b = 0.71 s/cP — same positive sign; b carries the units seconds per centipoise, while r is a pure number.

Key takeaway: Slope and correlation share direction; slope has units and scale, correlation does not.

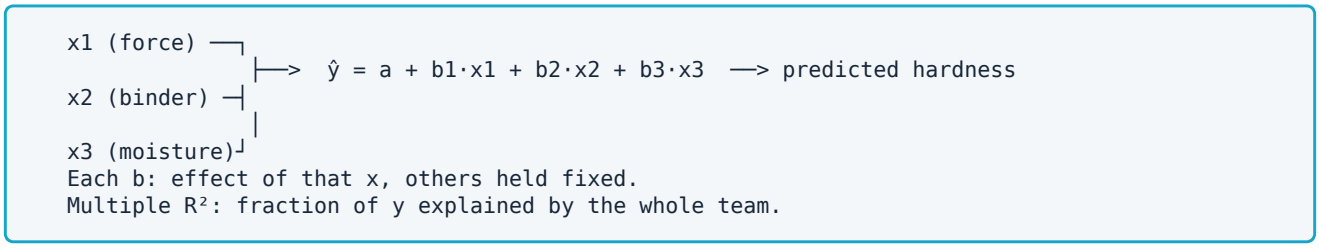
Technical concept: Multiple linear regression.

Simple explanation: Simple regression uses one predictor (one x). Real pharmacy problems usually have several influences at once. Multiple linear regression predicts y from *several* x's together: $\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$. Each slope b_1, b_2, ... tells the effect of that predictor *while holding the others fixed*. The multiple R^2 tells what fraction of y's variation the whole team of predictors explains.

Example: Predicting tablet hardness (y) from compression force (x_1), binder concentration (x_2) and granule moisture (x_3). The equation might be: hardness = 2.1 + 0.35·force + 0.80·binder − 0.12·moisture. Interpretation: raising binder by 1% raises predicted hardness by 0.80 kp, *if force and moisture stay the*

same.

Diagram/flowchart:



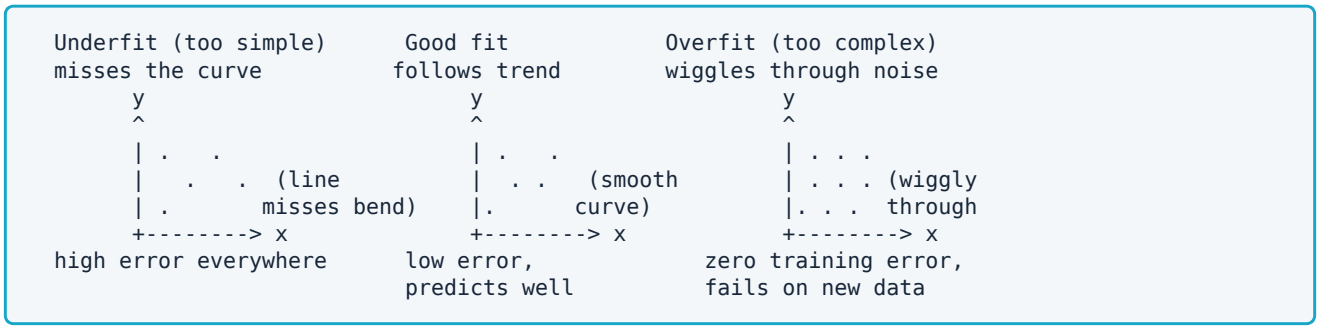
Key takeaway: Multiple regression lets a team of predictors share the prediction job; each slope is a "holding others constant" effect.

Technical concept: Overfitting and underfitting.

Simple explanation: Imagine drawing a curve through data points. **Underfitting** is drawing a straight line through data that clearly bend — the model is too simple and misses the pattern (high error even on the training data). **Overfitting** is the opposite: drawing a wiggly curve that passes exactly through *every* point, including random noise and measurement errors. It looks perfect on the training data but fails on new data, because new data won't have the same noise. The right model captures the true pattern without memorising the noise — this balance is the heart of good regression.

Example: Predicting drug content from storage time with 8 data points. A straight line (2 parameters) is reasonable. A 7th-degree polynomial (8 parameters) would pass through all 8 points exactly — perfect on paper, useless for predicting month 9, because it has memorised the measurement noise. That is overfitting. Fitting a flat horizontal line ("content never changes") to clearly falling data is underfitting.

Diagram/flowchart:



Key takeaway: Underfitting = too simple to catch the pattern; overfitting = so complex it memorises noise; good models sit between.

Classifications

Types of regression by number of predictors:

- **Simple regression:** one independent variable (this chapter's main focus).
- **Multiple regression:** two or more independent variables.

Types of regression by form:

- **Linear regression:** the equation is a straight line (or flat plane in multiple regression).
- **Non-linear regression:** the equation is a curve (e.g., exponential decay of drug concentration) — beyond this chapter's scope, but good to know it exists.

Two regression lines (a classic exam point): For the same data there are actually two least-squares lines: *y on x* (predict y from x: $\hat{y} = a + bx$) and *x on y* (predict x from y). They are different lines unless $r = \pm 1$. In pharmacy we almost always use y-on-x (e.g., predict absorbance from concentration).

Fit quality categories:

- **Underfit:** model too simple — systematic pattern left in residuals.
- **Good fit:** residuals small and random.
- **Overfit:** model too complex — fits noise, fails on new data.

Processes

Process 1: Computing slope and intercept by least squares (step-by-step)

1 Tabulate the n pairs (x, y). Decide which variable is x (predictor) and which is y (response).

2 Compute $\bar{x} = \Sigma x/n$ and $\bar{y} = \Sigma y/n$.

3 For each pair compute $(x - \bar{x})$, $(y - \bar{y})$, the product $(x - \bar{x})(y - \bar{y})$, and $(x - \bar{x})^2$.

4 Sum: $S_{xy} = \Sigma(x-\bar{x})(y-\bar{y})$ and $S_{xx} = \Sigma(x-\bar{x})^2$.

5 Slope: $b = S_{xy} / S_{xx}$.

6 Intercept: $a = \bar{y} - b \cdot \bar{x}$.

7 Write the equation: $\hat{y} = a + bx$. Check: plugging $x = \bar{x}$ must give $\hat{y} = \bar{y}$.

Process 2: Predicting a dependent value (step-by-step)

1 Confirm the regression equation $\hat{y} = a + bx$ is already fitted to relevant data.

2 Check the new x lies within (or near) the observed x-range — avoid wild extrapolation.

3 Substitute x into the equation and compute $\hat{y}$.

4 Report $\hat{y}$ with correct units, and state it is a prediction (estimate), not a measurement.

Process 3: Interpreting regression output (step-by-step)

1 Read the equation $\hat{y} = a + bx$ — identify intercept and slope.

2 Interpret the slope: "a one-unit rise in x changes predicted y by b units."

3 Interpret the intercept: predicted y at x = 0 (meaningful only if x = 0 is realistic).

4 Read R^2: "x explains $R^2 \times 100\%$ of the variation in y."

5 Check residual pattern (if available): random scatter = good; a curve in residuals = wrong model shape.

Textual Diagrams and Flowcharts

The regression workflow

```
1. Ask: which is x (input) and which is y (output)?
              |
2. Draw scatter plot — is a straight line sensible?
       /            \
     YES             NO → stop: linear regression unsuitable
      |
3. Compute b and a by least squares
      |
4. Write ŷ = a + bx; check it passes through (x̄, ȳ)
      |
5. Compute R² — how much of y is explained?
      |
6. Predict new ŷ values (interpolate, don't wildly extrapolate)
      |
7. Check residuals — random scatter means the line was a fair choice
```

Correlation vs regression — comparison diagram

```
CORRELATION (Chapter 7)              REGRESSION (this chapter)
"Do they move together?"             "Give me x, I'll predict y."
x and y are equals —                 x is input, y is output —
no input/output roles                roles matter
One number: r (−1 to +1)              An equation: ŷ = a + bx
Unit-free                            Slope carries units
r(x,y) = r(y,x) — symmetric          y-on-x line ≠ x-on-y line
Cannot predict                       Built for prediction
```

> INFOGRAPHIC CONCEPT: The Prediction Machine

The Prediction Machine. Draw a factory machine with a hopper on the left labelled "x goes IN (e.g., concentration)" and a chute on the right labelled "ŷ comes OUT (predicted absorbance)". Inside the machine, show the equation ŷ = a + bx as its gears, with the slope b drawn as a tilted ramp ("steeper ramp = bigger effect of x") and the intercept a as the machine's starting floor level ("where the line begins at x = 0"). Above the machine, a dial shows R^2 as a fuel gauge from 0% to 100% ("how much of y this machine explains"). A warning sticker reads: "Do not feed x-values far outside the training range — machine may produce nonsense (extrapolation danger)." A small side panel contrasts: "Correlation = the friendship; Regression = the machine built on it."

Tables

Table 1: Correlation vs regression

Feature	Correlation	Regression
Question answered	Do x and y move together?	What y do I predict from a given x?

Feature	Correlation	Regression
Roles of variables	None — x and y symmetric	x = independent (input), y = dependent (output)
Main result	Coefficient r (−1 to +1)	Equation $\hat{y} = a + bx$
Units	Unit-free	Slope carries units (y-units per x-unit)
Prediction	Not possible	The whole purpose
Causation	Never implied	Still not implied — prediction ≠ causation

Table 2: Interpreting regression output — quick guide

Output element	What it is	How to read it
Intercept a	Predicted y when x = 0	Starting level of the line
Slope b	Change in $\hat{y}$ per one-unit change in x	"Each extra unit of x adds b units to predicted y"
R^2	Fraction of y-variation explained	0.81 → "x explains 81% of the variation in y"
$1 - R^2$	Unexplained fraction	0.19 → "19% is due to other factors/noise"
Residual	Actual y − predicted y	Leftover error per point; should look random

Worked Examples and Calculation Boxes

CALCULATION BOX 1 — Slope, intercept and prediction, full working (pharmaceutical example)

Problem: A QC analyst prepares standard solutions of a drug and measures absorbance (y) at known concentrations (x, µg/mL) to build a calibration curve:

x (conc., µg/mL)	2	4	6	8	10
y (absorbance)	0.21	0.39	0.62	0.79	1.02

(a) Find the least-squares regression line of y on x. (b) Predict the absorbance for a 5 µg/mL solution. (c) A test sample shows absorbance 0.70 — estimate its concentration.

Solution — every step shown:

1 n = 5. $\Sigma x = 2+4+6+8+10 = 30 \rightarrow \bar{x} = 6$. $\Sigma y = 0.21+0.39+0.62+0.79+1.02 = 3.03 \rightarrow \bar{y} = 0.606$.

2 Working table:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
2	0.21	−4	−0.396	1.584	16
4	0.39	−2	−0.216	0.432	4
6	0.62	0	0.014	0.000	0
8	0.79	2	0.184	0.368	4
10	1.02	4	0.414	1.656	16

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
Total				**4.040**	**40**

1 Slope b = 4.040 / 40 = **0.101** absorbance units per µg/mL.

2 Intercept a = $\bar{y} - b \cdot \bar{x}$ = 0.606 − 0.101×6 = 0.606 − 0.606 = **0.000**.

3 Regression line: $\hat{y}$ **= 0.000 + 0.101x** (absorbance = 0.101 × concentration).

4 Check: at x = $\bar{x}$ = 6, $\hat{y}$ = 0.606 = $\bar{y}$. ✓ (The line passes through the mean point.)

5 (b) For x = 5: $\hat{y}$ = 0.101 × 5 = **0.505**. Predicted absorbance ≈ 0.51.

6 (c) Given y = 0.70, solve 0.70 = 0.101x → x = 0.70/0.101 ≈ **6.93 µg/mL**. (Using the calibration line in reverse — the standard analytical use.)

Key takeaway: This is exactly how real calibration curves work: fit the line on standards, then read unknown concentrations from measured absorbance.

CALCULATION BOX 2 — R^2 and its interpretation, full working

Problem: For the calibration data above, Pearson's r = 0.9989. Compute R^2 and interpret it. Also interpret a second case where a researcher reports R^2 = 0.42 for a regression of tablet friability on binder concentration.

Solution:

1 $R^2 = r^2 = (0.9989)^2 \approx$ **0.9978**, i.e., **99.78%**.

2 **Interpretation:** 99.78% of the variation in absorbance is explained by the straight-line relationship with concentration — an excellent calibration line; only 0.22% is unexplained noise.

3 Second case: R^2 = 0.42 → **42%** of the variation in friability is explained by binder concentration; **58%** is due to other factors (compression force, moisture, measurement noise). The line is a weak-to-moderate predictor — the researcher should look for additional predictors (multiple regression) rather than trusting this line alone.

Key takeaway: R^2 near 1 = the line explains almost everything; R^2 near 0 = the line explains almost nothing — always report and interpret it.

Example 3 — Multiple regression interpretation (illustrative)

A formulation scientist fits: predicted disintegration time (min) = 12.5 − 1.8×(binder %) + 0.6×(compression force, kN). Multiple R^2 = 0.88.

- b_1 = −1.8: each extra 1% binder *reduces* predicted disintegration time by 1.8 min, holding compression force fixed.
- b_2 = +0.6: each extra 1 kN force *increases* predicted disintegration time by 0.6 min, holding binder fixed.
- R^2 = 0.88: the two predictors together explain 88% of the variation in disintegration time.

Key takeaway: Each multiple-regression slope is a "keeping the others constant" effect; R^2 judges the whole team.

Example 4 — Overfitting vs underfitting (illustrative)

Eight monthly assay values of a stored product fall smoothly from 100% to 94%. A straight declining line fits well with small random residuals — good fit. A flat line at 97% leaves a clear downward pattern in residuals — underfit. A 7th-degree polynomial passes through all 8 points exactly but swings wildly between months and predicts an absurd 120% at month 9 — overfit (it memorised noise).

Key takeaway: Prefer the simplest model that captures the real pattern; exact fit to training data is a warning sign, not a triumph.

Analogies

- **Tailor's measuring tape:** Correlation tells you "taller people need longer trousers" (they move together). Regression is the tailor's formula: trouser length = a + b × height — it actually *predicts* the measurement. The slope b is "extra cm of trouser per cm of height."
- **Overfitting as memorising vs understanding:** A student who memorises last year's question paper word-for-word (overfit) scores 100% on it but fails a new paper. A student who learns only chapter headings (underfit) fails both. The good student learns the concepts (good fit) and handles new questions.

Pharmacy Relevance

- **Calibration curves (the flagship application):** Concentration vs absorbance (UV), peak area vs concentration (HPLC) — regression lines convert instrument readings into drug content. This is daily QC work.
- **Stability and shelf life:** Regression of drug content (y) on storage time (x) gives the degradation rate (slope); the line is extended to the specification limit to estimate shelf life.
- **Dissolution profiling:** Cumulative % drug released (y) vs time (x) — the slope reflects release rate; comparing slopes compares formulations.
- **Dose–response:** Response (y) vs log-dose (x) regression underpins potency estimation of biological products.
- **Process optimisation:** Multiple regression of tablet hardness/disintegration on compression force, binder level and moisture guides formulation scale-up.
- **Clinical predictions:** Regression of patient recovery time on age, dose and baseline severity helps set expectations and doses (with the usual caution that prediction is not causation).

India Context

- Indian pharmacopoeial and QC laboratories build calibration curves by least-squares regression for every batch assay — the regression line converts instrument readings into the reported drug content.

- Stability studies on products marketed across India's climatic zones use regression of assay value on time to fix shelf life; the slope is the degradation rate on the certificate of analysis.
- In academic pharmacy research in India, multiple regression is used to model how formulation variables jointly affect tablet properties during M.Pharm dissertation work.
- Public-health analysts use simple regression to project medicine demand (y) from population or case-load trends (x) for state drug procurement planning.

Key Points

- Regression predicts y from x using the equation $\hat{y} = a + bx$; x = independent (input), y = dependent (output).
- Least squares chooses the line minimising the sum of squared residuals; $b = \Sigma(x-\bar{x})(y-\bar{y})/\Sigma(x-\bar{x})^2$, $a = \bar{y} - b\cdot\bar{x}$; the line passes through $(\bar{x}, \bar{y})$.
- Slope b = expected change in y per one-unit change in x; intercept a = predicted y at x = 0.
- R^2 = fraction of y's variation explained by the line (in simple regression, $R^2 = r^2$); $1 - R^2$ is unexplained.
- b = r × (s_y/s_x): same sign as r; b carries units, r does not.
- Two regression lines exist (y on x, x on y); pharmacy normally uses y on x.
- Multiple regression: $\hat{y} = a + b_1x_1 + b_2x_2 + \ldots$; each slope is a holding-others-fixed effect.
- Underfitting = too simple (misses pattern); overfitting = too complex (memorises noise, fails on new data).
- Interpolate, don't recklessly extrapolate; prediction still isn't causation.

Common Confusions

- **Correlation vs regression (the big one):** Correlation = symmetric association, no prediction, no x/y roles. Regression = directional prediction with an equation. Saying "regression shows x causes y" is just as wrong as saying it for correlation.
- **Intercept always meaningful?** Only if x = 0 is realistic. "Absorbance at zero concentration = 0" makes sense; "tablet weight at zero compression force" may not — don't over-interpret a.
- **R^2 vs r:** r (−1 to +1) measures linear strength and direction; R^2 (0 to 1) measures explained variation. In simple regression $R^2 = r^2$, but they answer different questions.
- **High R^2 = good model?** Not always — an overfit wiggly model has $R^2 \approx 1$ on training data but predicts new data terribly. Check residuals and keep the model simple.
- **Which variable is x?** The one you *know* (or control) and predict *from*; y is what you *want to know*. Swapping them gives a different line (x on y) — choose deliberately.
- **Extrapolation trap:** A calibration line valid for 2–10 µg/mL may be completely wrong at 50 µg/mL (detector saturates) — never predict far outside the data.

Common Mistakes

- Fitting regression without a scatter plot, then missing curvature that invalidates the straight line.
- Extrapolating far beyond the observed x-range (e.g., predicting shelf life at 10 years from 6-month data).
- Interpreting the intercept when x = 0 is impossible or meaningless for the system.
- Reporting predictions with false precision (e.g., 6.934218 µg/mL from rough data).
- Using the x-on-y line when the y-on-x line is needed (or vice versa) when reading a calibration curve backwards.
- Adding more and more predictors until R^2 looks impressive — classic overfitting; prefer fewer, meaningful predictors.
- Forgetting that regression, like correlation, does not prove causation.

Mnemonics

- **"b before a, slope before start":** Compute the slope (b) first, then the intercept ($a = \bar{y} - b\cdot\bar{x}$).
- **"R^2 is the Report card":** R^2 reports what percentage of y the line explains.
- **"Slope has Shoes (units), r runs barefoot":** The slope carries units (y-units per x-unit); the correlation coefficient is unit-free.
- **"Under is Unfinished, Over is Overdone":** Underfitting = model unfinished (too simple); overfitting = model overdone (too fancy, memorises noise).
- **"$\hat{y} = a + bx$: a is the Anchor, b is the Bend-rate":** The intercept anchors the line at x = 0; the slope sets how fast it climbs.

Exam Summary

- Regression concept; independent (x, predictor) vs dependent (y, response) variables.
- Simple linear regression equation $\hat{y} = a + bx$ (y on x).
- Method of least squares: minimise $\Sigma(\text{residuals})^2$; formulas $b = \Sigma(x-\bar{x})(y-\bar{y})/\Sigma(x-\bar{x})^2$ and $a = \bar{y} - b\cdot\bar{x}$; line passes through $(\bar{x}, \bar{y})$.
- Coefficient of determination R^2: definition, range 0–1, interpretation as % of y-variation explained; in simple regression $R^2 = r^2$.
- Interpretation of regression output: slope ("per-unit change"), intercept, R^2, residuals.
- Relationship $b = r\cdot(s_y/s_x)$: same sign; b has units, r is unit-free.
- Pharmaceutical numericals: calibration curve (find line, predict y from x, find x from y), with full working.
- Applications: calibration curves, stability/shelf life, dissolution, dose–response, process optimisation.
- Multiple linear regression: concept, equation, "holding others fixed" interpretation of slopes, applications.

- Overfitting vs underfitting: definitions, diagrams, example, how to avoid (keep models simple, check new data).

One-Minute Revision

Regression predicts y from x: $\hat{y} = a + bx$. Least squares finds the best line — smallest total squared vertical error. Slope b = change in y per unit x (compute first); intercept $a = \bar{y} - b \cdot \bar{x}$ (predicted y at x = 0). R^2 = share of y's variation explained (r^2 in simple regression). b and r share sign; b has units. Multiple regression = team of predictors, each slope holding others fixed. Underfit = too simple; overfit = memorises noise. Interpolate, don't extrapolate; prediction ≠ causation.

Questions

Short-answer questions

1 Define regression analysis in one sentence.

2 Distinguish independent and dependent variables with a pharmacy example.

3 Write the simple linear regression equation and name each symbol.

4 State the principle of the method of least squares.

5 What is a residual?

6 Define the coefficient of determination R^2.

7 State the relationship between the slope b and Pearson's r.

8 What is multiple linear regression?

9 Define overfitting and underfitting in one line each.

10 Why is extrapolation risky?

Long-answer questions

1 Explain the method of least squares for fitting a straight line. Derive (state) the formulas for slope and intercept and explain each term.

2 What is R^2? Explain its meaning, range and interpretation with two examples (one high, one low).

3 Describe the applications of regression analysis in pharmaceutical sciences, with the calibration curve as the central example.

4 Explain multiple linear regression with a pharmaceutical example. How is each slope interpreted?

5 Explain overfitting and underfitting with diagrams. How does a good model balance the two?

MCQs (10) with answer key

1 In regression, the variable to be predicted is the: (a) independent variable (b) dependent variable (c) constant (d) error term. **Answer: (b)**

2 The least-squares line minimises: (a) sum of residuals (b) sum of squared residuals (c) sum of x values (d) the correlation coefficient. **Answer: (b)**

3 The regression line always passes through: (a) the origin (b) $(\bar{x}, \bar{y})$ (c) the highest point (d) (0, 1). **Answer: (b)**

4 If b = 2.5 in $\hat{y} = a + bx$, it means: (a) y increases 2.5 units per one-unit increase in x (b) x increases 2.5 per unit y (c) correlation is 2.5 (d) intercept is 2.5. **Answer: (a)**

5 $R^2 = 0.64$ means: (a) 64% of y's variation is explained by the line (b) correlation is 0.64 (c) slope is 0.64 (d) 64% error. **Answer: (a)**

6 The correct relation between slope b and correlation r is: (a) b = r (b) $b = r \cdot (s_y/s_x)$ (c) $b = r^2$ (d) b = 1/r. **Answer: (b)**

7 Multiple linear regression involves: (a) many dependent variables (b) two or more independent variables (c) no intercept (d) only curved lines. **Answer: (b)**

8 Overfitting means the model: (a) is too simple (b) fits training noise and predicts new data poorly (c) has $R^2 = 0$ (d) ignores all predictors. **Answer: (b)**

9 Predicting y for an x far outside the observed range is called: (a) interpolation (b) extrapolation (c) standardisation (d) ranking. **Answer: (b)**

10 In a calibration curve, concentration is normally the: (a) dependent variable (b) independent variable (c) residual (d) intercept. **Answer: (b)**

Viva questions

1 What is the difference between correlation and regression?

2 Why does the least-squares method square the residuals instead of just adding them?

3 What does the slope of a calibration curve tell you?

4 If $R^2 = 0.20$, is the regression line useful? Explain.

5 How would you recognise overfitting in practice?

Application-based questions

1 A stability study gives the regression line: assay % = 99.8 − 0.35 × (months). Interpret the slope and estimate the assay at 12 months. (Answer: potency falls 0.35% per month; at 12 months ≈ 99.8 − 4.2 = 95.6%.)

2 An analyst fits peak area = 150 + 4200 × concentration with $R^2 = 0.999$. A sample gives peak area 35,000. Estimate the concentration and comment on reliability. (Answer: x = (35000 − 150)/4200 ≈ 8.30 units; $R^2 = 0.999$ means the line explains 99.9% of variation — highly reliable within the calibrated range.)

3 A student adds predictors (room temperature, analyst's age, day of week) to a dissolution regression until $R^2 = 0.999$. Identify the problem and advise. (Answer: overfitting — junk predictors memorise noise; keep only meaningful predictors and validate on new batches.)

References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)