

1 CHAPTER OVERVIEW

Your visual roadmap through the chapter

Association Between Variables

Contingency tables, scatter plots, positive/negative/no association

1

Correlation & the Correlation Coefficient

Strength & direction of a relationship; Spearman's rank correlation

2

Linear Regression Analysis

The line of best fit; forecasting with $\hat{y}=a+bx$

3

WHAT IT'S ABOUT

Establishing whether two variables are associated, measuring the strength of that association (correlation), and modelling the functional relationship between them for prediction (regression).

WHY IT MATTERS FOR QAB

Businesses use these tools to understand relationships (e.g. cost vs output, advertising vs sales) and to forecast future values.

EXAM SKILLS REQUIRED

Construct and read scatter plots; calculate the correlation coefficient and Spearman's rank correlation; derive and use a regression line to forecast.

2 KEY DEFINITIONS

Source-based terminology, exam-ready

Contingency Table

Displays two categorical variables in rows and columns, showing the frequency of observations at each intersection. Also called cross-tabulation.

Key word: categorical, cross-tab

Scatter Plot

A graphical display for two quantitative variables, using the x-axis for the independent variable and y-axis for the dependent variable.

Key word: x independent, y dependent

Correlation

A statistical relationship or association between two variables; changes in one are associated with changes in the other.

Key word: association

Correlation Coefficient (r)

A numerical value measuring the degree, strength and direction of association between two quantitative variables; ranges from -1 to +1.

Key word: -1 to +1

Coefficient of Determination (r^2)

The square of the correlation coefficient; shows how much of the variation in y is explained by variation in x. Ranges 0 to +1.

Key word: % variation explained

Spearman's Rank Correlation

A correlation coefficient calculated when variables are based on rank/order rather than magnitude.

Key word: ranked data

Linear Regression / Line of Best Fit

A statistical technique for calculating the line that minimizes the sum of squared deviations (residuals) between the line and actual data points.

Key word: least squares

Residual

The vertical distance between an actual data point and the regression line.

Key word: actual - predicted

3 CORE CONCEPT

Explained in exam-oriented language

CONCEPT → Correlation

MEANING

Measures the strength and direction of an association — it does NOT prove causation.

PURPOSE

Determine whether, and how strongly, two variables move together.

APPLICATION

Testing whether two quantitative variables (or two rankings) are related.

CONCEPT → Regression

MEANING

Provides a functional (algebraic) relationship between a dependent and an independent variable.

PURPOSE

Predict/forecast the dependent variable from a given value of the independent variable.

APPLICATION

Estimating fixed/variable costs, forecasting sales, trend analysis.

4 KEY RULES & PRINCIPLES

Every source rule, exam-ready

RULE 1 Interpreting r

- $r = +1$: perfect positive; $r = -1$: perfect negative; $r = 0$: no association
- Closer to ± 1 = stronger linear association; closer to 0 = weaker
- r between +0.90 and +1 generally indicates good linear correlation

RULE 4 Coefficient of determination (r^2)

- $r^2=0$: independent variable explains NONE of the variation in y
- $r^2=1$: independent variable explains ALL of the variation in y
- Between 0 and 1: that % of variation in y is explained by x

RULE 2 Properties of r

- Unit-less (a ratio); unaffected by the units the variables are measured in
- Symmetric: $r_{xy} = r_{yx}$
- Always lies between -1 and $+1$
- Is the geometric mean of the two regression coefficients

RULE 5 Calculating the regression line

- **b must be calculated first**, since a 's formula depends on b
- The regression line always passes through the point $(\bar{x}, \bar{y})$
- $b = r \times (S_y \div S_x)$ links the regression coefficient to r

RULE 3 Correlation $\neq$ causation

- A strong correlation does not mean the independent variable causes the dependent variable to change
- Other factors may influence the relationship

RULE 6 Limitations to remember

- Scatter plots: only show 2 variables at once; few/atypical points mislead; can't quantify strength precisely
- Regression: poor fit for non-linear data; distorted by outliers; unsafe to extrapolate beyond the observed data range

5 FORMULA ZONE — CORRELATION

Every formula supported by the source text

CORRELATION COEFFICIENT (METHOD 1 — RAW SUMS)

$$r = \frac{[n\sum xy - \sum x \sum y] \div \sqrt{[(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)]}}$$

x, y=paired data values, **n**=number of pairs. Most direct to compute from raw data.

CORRELATION COEFFICIENT (METHOD 2 & 3 — DEVIATIONS)

$$r = \frac{\sum (x - \bar{x})(y - \bar{y}) \div \sqrt{[\sum (x - \bar{x})^2 \times \sum (y - \bar{y})^2]}}$$

$\sum (x - \bar{x})(y - \bar{y}) \div n$ is the **covariance** (s_{xy}). Equivalent to $r = s_{xy} \div (n \cdot s_x s_y)$ using standard deviations.

COEFFICIENT OF DETERMINATION

$$r^2 = (\text{correlation coefficient})^2$$

Simply square the value of r . Always between 0 and 1.

SPEARMAN'S RANK CORRELATION COEFFICIENT

$$r = 1 - \frac{[6\sum d^2 \div n(n^2 - 1)]}{n}$$

d=difference between ranks in a pair, **n**=number of pairs. The "6" is a fixed constant, not related to n .

5 FORMULA ZONE — LINEAR REGRESSION

Every formula supported by the source text

REGRESSION LINE

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

$\hat{y}$ =predicted value of y , **a**=intercept, **b**=slope (regression coefficient). Calculate **b** BEFORE **a**.

SLOPE (B) — TWO EQUIVALENT FORMS

$$b = \frac{[n\sum xy - \sum x \sum y] \div [n\sum x^2 - (\sum x)^2]}{n} \text{ or } \frac{\sum (x - \bar{x})(y - \bar{y}) \div \sum (x - \bar{x})^2}{n}$$

Use whichever form matches the data already tabulated (raw sums, or deviations from the mean).

INTERCEPT (A)

$$a = (\sum y \div n) - b(\sum x \div n) \text{ or } a = \bar{y} - b\bar{x}$$

Requires **b** to already be calculated. The line always passes through $(\bar{x}, \bar{y})$.

LINK BETWEEN B AND R

$$b = r \times (S_y \div S_x) \mid r = b \times (S_x \div S_y)$$

S_x, S_y are the standard deviations of x and y .

6 CALCULATION ROADMAP

General path for any correlation/regression question

1- Read Question

Identify the paired data and what's required

2- Identify Method

Correlation (r), rank correlation, or regression?

3- Tabulate

Build the x, y, xy, x^2 , y^2 (or deviation) columns

4- Substitute

Put the column totals into the formula

5- Calculate & Interpret

Solve, interpret strength/direction or forecast

7 STEP-BY-STEP METHODS

Kept separate — do not mix methods

METHOD A · Constructing a Scatter Plot

- 1 Identify the independent (x) and dependent (y) variable
- 2 Draw the axes with appropriate scales
- 3 Plot each (x, y) pair as a point
- 4 Observe the pattern: positive, negative, or no association

METHOD C · Spearman's Rank Correlation

- 1 Rank each variable's values (if not already ranked)
- 2 Find d = difference between the two ranks for each pair
- 3 Square each d and sum: $\sum d^2$
- 4 Substitute into $r = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$

METHOD B · Correlation Coefficient (Raw Sums)

- 1 Tabulate x, y, xy, x^2 , y^2 for every pair
- 2 Sum each column: $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$
- 3 Substitute into $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$
- 4 Interpret the sign (direction) and size (strength) of r

METHOD D · Regression Line & Forecasting

- 1 Tabulate the required sums or deviations (as for correlation)
- 2 Calculate b FIRST using the slope formula
- 3 Calculate a using $a = \bar{y} - b\bar{x}$ (or the sums version)
- 4 Write $\hat{y} = a + bx$, then substitute any x to forecast $\hat{y}$

8 DIFFERENT CASES / SCENARIOS

Recognise the data, pick the right tool

IF both variables are categorical
→ USE a **contingency table**

IF both variables are quantitative
→ USE a **scatter plot** and/or the **correlation coefficient**

IF data is based on rank/order, not magnitude
→ USE **Spearman's rank correlation**

IF you need to predict/forecast a value
→ USE the **regression line** ($\hat{y}=a+bx$), not correlation alone

IF r is close to +1 or -1
→ Strong linear association exists

IF r is close to 0
→ Little to no linear association

IF asked what % of variation in y is explained by x
→ USE r^2 (coefficient of determination)

IF forecasting a value OUTSIDE the range of the original data
→ Treat with caution — extrapolation may not be valid

9 COMPARISON TABLES

Don't confuse these!

	Correlation	Regression
What it measures	Degree/strength of association between two variables	A functional relationship, for predicting one variable from another
Provides causation?	No	Identifies dependent/independent roles, but still not proof of causation
Output	A single value, r (−1 to +1)	An equation, $\hat{y}=a+bx$

	Pearson's r (Correlation Coefficient)	Spearman's Rank Correlation
Data type	Actual quantitative values	Ranks/order (e.g. preference, quality)
Formula basis	Raw values or deviations from the mean	Differences between ranks (d)

10 READING SCATTER PLOTS & REGRESSION LINES

How to interpret the source's graphs

Reading a scatter plot

Points trending upward (left to right) = positive association. Points trending downward = negative association. Points scattered with no visible pattern = no association. Isolated points far from the rest may be outliers.

Reading a regression line graph

The line of best fit runs through the data cloud, minimizing the total squared vertical distance (residuals) from each point to the line. The line always passes through the point $(\bar{x}, \bar{y})$.

11 WORKED CALCULATION LOGIC

Method students must reproduce in the exam

REGRESSION LINE: GIVEN → FORMULA → SUBSTITUTION → CALCULATION → ANSWER

A company's monthly output/cost data gives: $n=6$, $\Sigma x=41.8$, $\Sigma y=268.3$, $\Sigma xy=1,885.91$, $\Sigma x^2=295.88$. Find the regression line and forecast cost at 10,000 units ($x=10$).

GIVEN: $n=6$, $\Sigma x=41.8$, $\Sigma y=268.3$, $\Sigma xy=1,885.91$, $\Sigma x^2=295.88$

FORMULA: $b = \frac{[n\Sigma xy - \Sigma x\Sigma y]}{[n\Sigma x^2 - (\Sigma x)^2]}$ then $a = (\Sigma y/n) - b(\Sigma x/n)$

SUBSTITUTION: $b = \frac{[6(1,885.91) - (41.8)(268.3)]}{[6(295.88) - 41.8^2]} = \frac{100.52 + 28.04}{3.585} = 3.585$

CALCULATION: $a = 44.72 - 3.585(6.97) = 44.72 - 24.98 = 19.74 \rightarrow \hat{y} = 19.74 + 3.585x$

FINAL ANSWER: At $x=10$ (10,000 units), $\hat{y} = 19.74 + 3.585(10) = \text{Rs. } 55.6 \text{ million}$

12 EXAM QUESTION PATTERNS

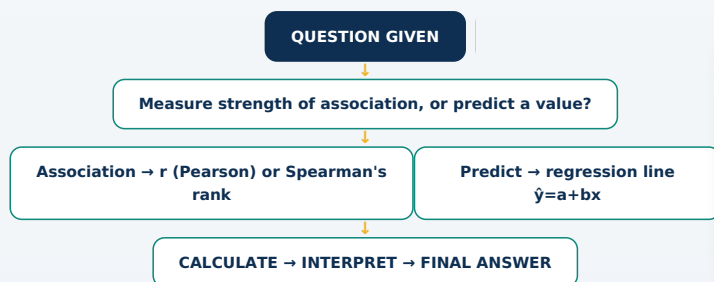
- **Correlation coefficient** — calculate and interpret r
- **Coefficient of determination** — interpret r^2 as % variation explained
- **Spearman's rank correlation** — ranked/ordinal data
- **Regression line derivation** — find a and b , write $\hat{y} = a + bx$
- **Forecasting** — substitute a given x to predict y
- Conceptual — correlation vs causation, limitations of each tool

13 EXAM TRAPS

⚠ WATCH OUT

- ✗ **Correlation implies causation** — a strong r never proves one variable causes the other.
- ✗ **Wrong correlation formula** — using Pearson's r on ranked data instead of Spearman's.
- ✗ **Calculating a before b** — a 's formula depends on b ; always find b first.
- ✗ **r vs r^2 confusion** — forgetting to square r , or re-squaring an already-squared r^2 .
- ✗ **Extrapolating too far** — forecasting well beyond the observed x -range is unreliable.

14 QUICK DECISION TREE



CHAPTER 10 IN 60 SECONDS

$r = \frac{[n\Sigma xy - \Sigma x\Sigma y]}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$
 $r^2 = r \times r$ (% variation explained)
Spearman's: $1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$
Regression: $\hat{y} = a + bx$
 b first, then $a = \bar{y} - b\bar{x}$
 $b = r \times (\Sigma y / \Sigma x)$
 Line passes through $(\bar{x}, \bar{y})$
Trap: correlation $\neq$ causation