

UNIT 3

Correlation and Regression Analysis

(6 Hours) — Probability and Statistics (CACS202)

1. Correlation

1.1 Definition

Correlation is a statistical technique that measures the **strength and direction of the linear relationship** between two quantitative variables, e.g., height and weight, study hours and marks, price and demand. Correlation shows association only — it does **not** prove that one variable causes the other.

- **Positive correlation** — both variables move in the same direction (study time $\uparrow$, marks $\uparrow$).
- **Negative correlation** — variables move in opposite directions (price $\uparrow$, demand $\downarrow$).
- **Zero correlation** — no linear relationship.
- **Simple, multiple and partial** correlation depending on the number of variables; **linear** and **non-linear** depending on the form.

1.2 Scatter Diagram

A **scatter diagram** is a graph in which each pair (X, Y) is plotted as a point. The pattern of points shows the type of correlation at a glance: points rising from left to right indicate positive correlation, falling points indicate negative correlation and a shapeless cloud indicates no correlation. The closer the points are to a straight line, the stronger the relationship.

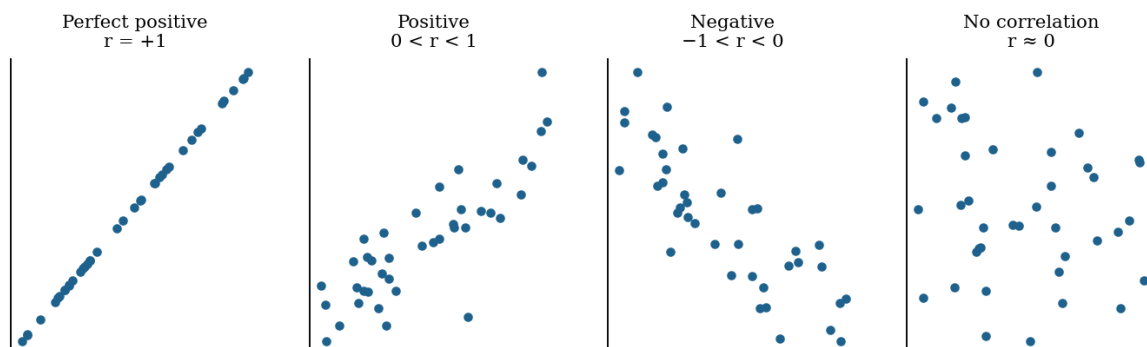


Figure 3.1 — Scatter diagrams showing different types of correlation

1.3 Karl Pearson's Coefficient of Correlation

Karl Pearson's **coefficient of correlation** r (also called the product-moment correlation) is the ratio of the covariance of X and Y to the product of their standard deviations:

$$r = \text{Cov}(X, Y) / (\sigma_x \sigma_y), \text{ where } \text{Cov}(X, Y) = \Sigma(X - \bar{X})(Y - \bar{Y})/n = \Sigma XY/n - \bar{X}\bar{Y}$$

A convenient **computational formula** for raw data is:

$$r = [n\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{[n\Sigma X^2 - (\Sigma X)^2] [n\Sigma Y^2 - (\Sigma Y)^2]}$$

Value of r	Interpretation
$r = +1$	Perfect positive correlation
$0.7 \leq r < 1$	Strong positive
$0.3 \leq r < 0.7$	Moderate positive
$0 < r < 0.3$	Weak positive
$r = 0$	No linear correlation
$-0.3 < r < 0$	Weak negative
$-1 < r \leq -0.3$	Moderate to strong negative
$r = -1$	Perfect negative correlation

Properties of r : (i) $-1 \leq r \leq +1$; (ii) r is a pure number with no units; (iii) r is independent of change of origin and scale (if the changes have the same sign); (iv) $r_{xy} = r_{yx}$ (symmetric); (v) if X and Y are independent, $r = 0$, but $r = 0$ does not always mean independence (there can be a non-linear relation). The **coefficient of determination** r^2 gives the proportion of variation in Y explained by X .

Example 3.1 — Calculating r

Find the correlation coefficient between advertising expenditure X (Rs. 000) and sales Y (Rs. lakh): $X = 2, 4, 6, 8, 10$; $Y = 3, 7, 5, 10, 11$.

X	Y	X^2	Y^2	XY
2	3	4	9	6
4	7	16	49	28
6	5	36	25	30
8	10	64	100	80
10	11	100	121	110

X	Y	X ²	Y ²	XY
$\Sigma X = 30$	$\Sigma Y = 36$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 304$	$\Sigma XY = 254$

Solution

$n = 5$. Numerator = $n\Sigma XY - \Sigma X\Sigma Y = 5 \times 254 - 30 \times 36 = 1270 - 1080 = \mathbf{190}$

$n\Sigma X^2 - (\Sigma X)^2 = 5 \times 220 - 900 = 200$; $n\Sigma Y^2 - (\Sigma Y)^2 = 5 \times 304 - 1296 = 224$

$r = 190 / \sqrt{(200 \times 224)} = 190 / 211.66 = \mathbf{0.90}$

Interpretation: **strong positive** correlation; $r^2 = 0.81$, i.e., about 80.6% of the variation in sales is explained by advertising.

Example 3.2 — Effect of change of origin and scale

If r between X and Y is 0.8, find r between $U = 2X + 3$ and $V = 5 - 3Y$.

Solution

U has a positive scale factor (+2) but V has a negative one (−3). Signs differ, so the sign of r reverses: $r(U, V) = \mathbf{-0.8}$.

2. Regression Analysis

2.1 Definition, Dependent and Independent Variables

Regression analysis studies the **average functional relationship** between variables so that the value of one variable can be **estimated (predicted)** from the other. The variable being predicted is the **dependent (response, explained) variable** Y ; the variable used for prediction is the **independent (predictor, explanatory) variable** X . Sir Francis Galton first used the term in his study of heights of fathers and sons.

Correlation	Regression
Measures the degree and direction of association	Gives the form of the relation and enables prediction
Symmetric: $r_{xy} = r_{yx}$	Not symmetric: Y on X differs from X on Y
No distinction between dependent and independent variables	One variable is dependent, the other independent
r is a pure number between -1 and 1	Regression coefficient has the units of Y per unit of X

2.2 Least Squares Method

The **method of least squares** fits the straight line $\hat{Y} = a + bX$ so that the **sum of squares of the vertical deviations** (residuals) between observed Y and estimated $\hat{Y}$ is minimum. Minimizing $\sum(Y - a - bX)^2$ gives the **normal equations**:

$$\sum Y = na + b\sum X \text{ and } \sum XY = a\sum X + b\sum X^2$$

Solving them, the **regression coefficient** (slope) and intercept are:

$$b_{yx} = [n\sum XY - \sum X\sum Y] / [n\sum X^2 - (\sum X)^2] = r \sigma_y / \sigma_x, a = \bar{Y} - b \bar{X}$$

Similarly, the regression line of **X on Y** is $X = a' + b'Y$ with $b_{xy} = [n\sum XY - \sum X\sum Y] / [n\sum Y^2 - (\sum Y)^2] = r \sigma_x / \sigma_y$.

- Both regression lines pass through the point $(\bar{X}, \bar{Y})$.
- $r^2 = b_{yx} \times b_{xy}$, so $r = \pm\sqrt{(b_{yx} b_{xy})}$; the sign of r is the common sign of both coefficients, and both coefficients have the same sign.
- If $r = \pm 1$ the two lines coincide; if $r = 0$ they are at right angles.
- Arithmetic mean of b_{yx} and b_{xy} is $\geq r$ (since the geometric mean is $\leq$ the arithmetic mean).
- Regression coefficients are independent of change of origin but not of scale.

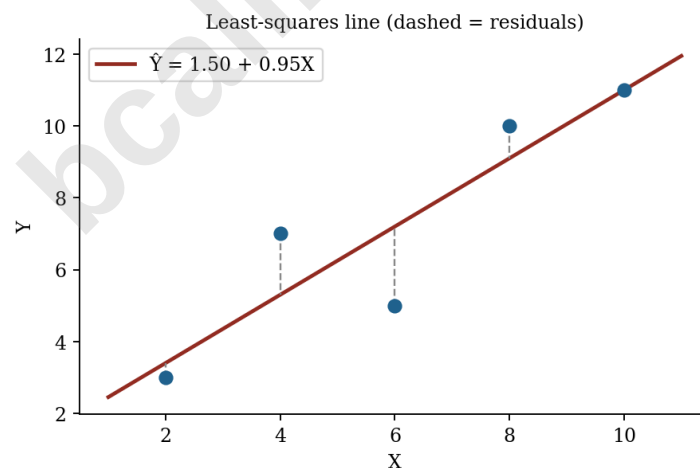


Figure 3.2 — Least-squares line fitted to the data of Example 3.1

Example 3.3 — Regression lines

For the data of Example 3.1, (a) fit the regression line of Y on X , (b) fit X on Y , (c) estimate sales when advertising is Rs. 7 (thousand) and (d) verify r .

Solution

$$\bar{X} = 30/5 = 6.00, \bar{Y} = 36/5 = 7.20$$

$$(a) b_{yx} = 190/200 = \mathbf{0.95}; a = 7.20 - 0.95 \times 6.00 = 1.50. \hat{Y} = \mathbf{1.50 + 0.95 X}$$

$$(b) b_{xy} = 190/224 = \mathbf{0.848}; a' = 6.00 - 0.848 \times 7.20 = -0.11. \mathbf{X = -0.11 + 0.848 Y}$$

$$(c) \text{ For } X = 7: \hat{Y} = 1.50 + 0.95 \times 7 = \mathbf{8.15} \text{ (lakh).}$$

$$(d) r = \sqrt{(b_{yx} \times b_{xy})} = \sqrt{(0.950 \times 0.848)} = \sqrt{0.806} = \mathbf{0.90} \text{ (positive since both b's are positive) — matches Example 3.1.}$$

Example 3.4 — Using summary statistics

$\bar{X} = 50, \bar{Y} = 60, \sigma_x = 4, \sigma_y = 6, r = 0.75$. Find the regression equation of Y on X and estimate Y for $X = 54$.

Solution

$$b_{yx} = r \sigma_y / \sigma_x = 0.75 \times 6/4 = \mathbf{1.125}$$

$$Y - \bar{Y} = b_{yx} (X - \bar{X}) \rightarrow Y - 60 = 1.125(X - 50) \rightarrow \hat{Y} = \mathbf{3.750 + 1.125X}$$

$$\text{For } X = 54: \hat{Y} = 60 + 1.125 \times 4 = \mathbf{64.50}.$$

Note: Predictions are reliable only within (or near) the observed range of X, and only when r is reasonably high. Extrapolation far outside the data is risky.

Key Takeaways

- ✓ Correlation measures the strength and direction of a linear relationship; a scatter diagram shows it visually.
- ✓ Karl Pearson's $r = \text{Cov}(X, Y) / (\sigma_x \sigma_y) = [n \sum XY - \sum X \sum Y] / \sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}$, with $-1 \leq r \leq 1$.
- ✓ r^2 (coefficient of determination) is the proportion of variation in Y explained by X; correlation does not imply causation.
- ✓ Least squares fits $\hat{Y} = a + bX$ by minimizing the sum of squared residuals; $b = [n \sum XY - \sum X \sum Y] / [n \sum X^2 - (\sum X)^2]$, $a = \bar{Y} - b\bar{X}$.
- ✓ Regression coefficients satisfy $r^2 = b_{yx} \cdot b_{xy}$; both lines pass through $(\bar{X}, \bar{Y})$.

— End of Unit 3 —