ICS Part-II Statistics — Punjab Board

CHAPTER 14

Simple Linear Regression and Correlation

Statistics • 2nd Year • ICS • Punjab Board • Free PDF Notes

Chapter 14: Simple Linear Regression and Correlation

Chapters 12 and 13 dealt with a single variable at a time -- estimating or testing claims about one population mean or proportion. This chapter opens a new area of statistics devoted to relationships BETWEEN two variables. Two very different but closely related questions are asked about any pair of variables: is there a mathematical rule that lets us predict one variable from the other (regression), and how strong or close is the linear association between them (correlation)? Regression analysis fits the 'best' straight line through a scatter of paired observations and uses it to estimate or predict the value of one variable (the dependent, or regressand) from a given value of the other (the independent, or regressor). Correlation analysis, by contrast, treats both variables symmetrically and measures the degree to which they move together, without singling out either one as dependent.

The chapter builds the whole machinery from first principles: the distinction between a functional and a statistical relation, the population regression model and its sample counterpart, the method of least squares used to fit the best line, the sample correlation coefficient r and its properties, and finally the tight algebraic relationship that ties regression coefficients and the correlation coefficient together. A recurring and important theme is that regression and correlation are mathematically linked but conceptually distinct -- and that neither one, however strong, proves that one variable causes changes in the other.

Learning Objectives

- Distinguish a functional relation from a statistical relation between two variables
- Define regressor, regressand, regression function, and regression curve
- State and apply the least squares principle to fit a straight line to paired data
- Derive and use the least squares estimates a and b of the population regression line

- Use coding and scaling to simplify regression computations for equally spaced x-values
- State the limitations of linear regression
- Define positive, negative, and no correlation between two variables
- Compute and interpret the sample correlation coefficient r and list its key properties
- Relate the correlation coefficient r to the two regression coefficients b_yx and b_xy
- Explain why a high correlation does not, by itself, establish causation

Key Concepts

14.1 Relations Between Variables

A functional relation between two variables x and y is a perfect relation: the value of the dependent variable y is uniquely and exactly determined by the value of the independent variable x through a mathematical formula, y=f(x). Observations plotted on a graph all fall exactly on the line or curve of the relationship -- this exactness is the defining feature of a functional relation. A statistical relation, by contrast, is not exact: the value of y is not uniquely determined merely from knowledge of x, and plotted observations scatter around, rather than fall exactly on, a line or curve. Most real relationships in business, economics, and the social sciences -- income and expenditure, fertilizer and crop yield, advertising cost and sales, a father's height and his son's height -- are statistical, not functional, relations.

A separate question from the mathematical form of a relationship is whether it is causal: does a genuine cause-and-effect mechanism link the two variables, or do they merely tend to move together? Regression analysis and correlation analysis, on their own, make no assertions whatsoever about causality -- establishing cause requires evidence and reasoning that goes well beyond fitting a line or computing a correlation coefficient.

14.2 Regression Analysis

Regression analysis provides a method for estimating an average (often linear) relationship between two or more variables, allowing the investigator to explain

and predict the value of one variable from known values of the others. In the simplest case, only two variables are involved: the regressor (also called the predictor, independent, controlled, or explanatory variable, usually denoted x) is the variable that forms the basis for estimation, its values fixed in advance by the experimenter and free of random error; the regressand (also called the response, predictand, dependent, or explained variable, usually denoted Y) is the variable whose value is being estimated or predicted, and which is subject to random variation even for a fixed value of x.

If mu_Y|x = E(Y|x) denotes the expected value of the distribution of Y for a given, fixed value of x, then the simple regression is defined as mu_Y|x = f(x), where f(x) may be linear, quadratic, exponential, or any other functional form describing how the regressor relates to the response. Because the points only are given and the function f(x) must be 'worked backwards' or regressed from them, this function is called the regression function. The regression curve is the locus of these expected values mu_Y|x traced out as x varies -- the curve joining the expected value of Y's distribution at each possible value of x.

14.3 Curve Fitting

Curve fitting is the process of estimating, from an observed sample, the parameters of the population regression function relating a response variable to a regressor variable. The least squares principle, due to the French mathematician Adrien Legendre, states that among all possible curves that could be fitted to a set of observed data, the sum of squares of the residuals (the observed values minus their corresponding fitted/estimated values) should be made as small as possible; the curve achieving this minimum is called the least squares fit. Using this principle avoids the personal bias that would otherwise creep in if a curve were fitted to data purely 'by eye'.

Before selecting the form of curve to fit, a scatter diagram is drawn: a set of points in a rectangular coordinate system, with x plotted horizontally and y plotted vertically, where each point represents one observed pair (x_i, y_i). A visual inspection of the scatter diagram reveals the apparent nature of the relationship -- whether the points tend to run from lower-left to upper-right (a direct relationship, Y tends to increase as x increases), from upper-left to lower-right (an inverse relationship, Y tends to decrease as x increases), whether the

pattern looks straight (linear) or bent (curvilinear), and how tightly or loosely the points cluster around the apparent pattern.

14.4 Simple Linear Regression

When the simple regression describes the expected value of the dependent variable Y as a LINEAR function of the independent variable x, the regression is called simple linear regression, written mu_Y|x = alpha + beta.x. Here alpha is the intercept of the line along the y-axis (the value of mu_Y|x when x=0), and beta -- the simple linear regression coefficient -- is the change in the mean of Y's probability distribution per unit increase in x, i.e. the slope of the line, which remains constant at every value of x. Geometrically, beta is measured by tan(theta), where theta is the angle the line makes with the positive x-axis: if beta is positive the line slopes upward, and if beta is negative the line slopes downward.

14.5 The Simple Linear Regression Model

For a fixed value x_i of the regressor, the associated response is a random variable Y_i with mean mu_Y|xi = alpha + beta.x_i and constant variance sigma^2 (assumed the same for every value of x -- the variance does not depend on x, even though the mean does). The deviation of Y_i from its mean is the random error, epsilon_i = Y_i - (alpha + beta.x_i), and the resulting population regression model is Y_i = alpha + beta.x_i + epsilon_i, where alpha and beta are the (unknown) population intercept and slope. This model is 'simple' because there is only one regressor, 'linear in parameters' because no parameter appears as an exponent or is multiplied by another parameter, and 'linear in the independent variable' because x appears only to the first power.

Since alpha and beta are unknown population parameters, they must be estimated from sample data (x1,y1), (x2,y2), ..., (xn,yn). Let a be the best estimate of alpha and b be the best estimate of beta; then the sample simple linear regression line is y-hat = a + b.x, and for any specific x_i, y-hat_i = a + b.x_i is the best point estimate of mu_Y|xi. The residual (or deviation, or prediction error) for observation i is e_i = y_i - y-hat_i = y_i - a - b.x_i, giving the sample regression model y_i = a + b.x_i + e_i.

The covariance of two variables X and Y, denoted s_xy, measures their linear mutual variability: s_xy = [sum(x_i - xbar)(y_i - ybar)]/n = [sum(x_i.y_i)]/n - xbar.ybar. Its sign shows the direction of mutual variability -- positive when the variables tend to move together, negative when they move in opposite directions. If z_i = x_i +/- y_i, then s_z^2 = s_x^2 + s_y^2 +/- 2.s_xy in general, reducing to s_z^2 = s_x^2 + s_y^2 when X and Y are independent (s_xy=0).

Theorem 14.1 (Least Squares Point Estimation): given n observed pairs (x_i, y_i), the least squares line y-hat = a + b.x has slope b = [n.sum(x_i.y_i) - (sum x_i)(sum y_i)] / [n.sum(x_i^2) - (sum x_i)^2] = [sum(x_i-xbar)(y_i-ybar)] / [sum(x_i-xbar)^2] = s_xy/s_x^2, and intercept a = [sum y_i - b.sum x_i]/n = ybar - b.xbar. The least squares line always passes through the point of means (xbar, ybar), and the least squares estimate of mu_Y|xi is y-hat_i = ybar + b(x_i - xbar).

Coding and Scaling (Theorem 14.2): the sample regression coefficient b is independent of a change of origin but NOT independent of a change of scale; if u_i=(x_i-p)/h and v_i=(y_i-q)/k, then b_yx = (k/h).b_vu. When the x-values are equally spaced at interval h, a special coding takes the origin at the middle value (odd n) or the average of the two middle values (even n), and rescales by h (or h/2), making sum(u_i)=0 and greatly simplifying the arithmetic; the resulting line of Y on u, y-hat=a+bu with a=ybar and b=sum(u_i.y_i)/sum(u_i^2), is then transformed back to the original x-scale.

Properties of the least squares line: the sum of residuals is zero (sum e_i = 0); the sum of the observed y-values equals the sum of the fitted values (sum y_i = sum y-hat_i), so ybar equals the mean of the fitted values too; the sum of squared residuals, sum(e_i^2) = sum(y_i^2) - a.sum(y_i) - b.sum(x_i.y_i), is the smallest possible value achievable by ANY straight line through the data; and the line always passes through the point of means (xbar, ybar) -- the 'centre of gravity' of the data.

Limitations of Linear Regression: (i) it applies only to relationships that can genuinely be described by a straight line -- a scatter diagram should be examined first to check this; (ii) the least-squares procedure will always produce a 'best-fit' line, even when no real linear relationship exists at all, so a formal test of significance is needed to confirm the slope b is 'real' and not just sampling noise; (iii) the regression equation is asymmetrical -- the equation predicting Y

from x cannot simply be algebraically rearranged to predict x from Y, a separate regression of x on Y must be fitted; (iv) the regression equation is reliable only over the range of x actually observed in the sample, and should not be extrapolated far beyond that range.

14.6 Simple Linear Correlation

Simple linear correlation measures the strength or closeness of a linear relationship between two variables, addressing two related questions: are the two variables related at all, and how closely does one variable's regression line fit the scatter of observed points? Positive (direct) correlation occurs when the two variables tend to move in the same direction (both increase or both decrease together), corresponding to regression lines with positive slopes; it is called perfect positive correlation if every point lies exactly on a straight line with positive slope. Negative (inverse) correlation occurs when the variables move in opposite directions, corresponding to negative-sloped regression lines; perfect negative correlation means every point lies exactly on a line with negative slope. No correlation exists when one least squares regression line is horizontal and the other is vertical -- equivalently, if X and Y are independent, Cov(X,Y)=0, which implies the correlation coefficient rho=0.

14.7 Correlation Analysis

In pure correlation problems, both variables X and Y are random, and the relationship is considered simultaneously and symmetrically -- unlike regression, where one variable is singled out as dependent. Typical correlation examples include heights and weights of persons, ages of husbands and wives at marriage, and marks in two different subjects. Given a random sample of n pairs (x_i,y_i) from a bivariate population, the sample correlation coefficient r (or r_xy) is defined as r = s_xy/(s_x.s_y), which expands into several algebraically equivalent computing formulas, most commonly r = [n.sum(x_i.y_i) - (sum x_i)(sum y_i)] / sqrt{[n.sum(x_i^2)-(sum x_i)^2][n.sum(y_i^2)-(sum y_i)^2]}.

Properties of r: (1) r is symmetrical in X and Y, i.e. r_xy = r_yx; (2) r equals the covariance of the two variables measured in standard units, r=Cov(z_x,z_y); (3) the MAGNITUDE |r| is unaffected by a change of origin or scale in either variable, though r changes sign if the variables are multiplied/divided by constants of

opposite sign; (4) r always lies between -1 and +1 inclusive; (5) |r| is the geometric mean of the two regression coefficients, r = +/-sqrt(b_yx . b_xy), with the sign matching the common sign of b_yx and b_xy; (6) r is zero whenever one of the two variables is constant. Theorem 14.3 formalises property (3): the correlation coefficient is independent of the origin and the scale of measurement of the variables (up to a possible sign flip if the scale change reverses direction).

Goodness of Fit and the Regression-Correlation Link: for a bivariate sample there are two distinct least squares lines -- Y on X, y-hat=a_yx+b_yx.x, and X on Y, x-hat=a_xy+b_xy.y -- with b_yx=s_xy/s_x^2=r.s_y/s_x and b_xy=s_xy/s_y^2=r.s_x/s_y. Because s_x and s_y are always positive, s_xy, b_yx, b_xy, and r always share the same sign, and if any one of these four quantities is zero, all the others must be zero too. Theorem 14.4 states the two regression lines always intersect at the point of means (xbar, ybar). Theorem 14.5 states that r is the slope of the regression lines when both variables are expressed in standard (z-score) units. Theorem 14.6 states that the two regression lines coincide (become identical) exactly when every sample point lies exactly on a single straight line, i.e. when |r|=1.

Correlation and Causation: a high correlation coefficient shows only that the observed data are consistent with a linear association -- it does NOT, by itself, establish that one variable causes changes in the other. A high correlation between two variables may arise because X causes Y, because Y causes X, because some third factor Z affects both X and Y simultaneously (a common cause, sometimes called a spurious or nonsense correlation), or purely by chance. Only further, careful investigation beyond the correlation coefficient itself -- controlled experiments, theoretical reasoning, checking for confounding variables -- can establish whether a causal relationship genuinely exists.

Important Definitions

What is a functional relation between two variables?: A perfect, exact relation y=f(x) in which the value of the dependent variable is uniquely determined by the independent variable; all plotted observations fall exactly on the line or curve.

What is a statistical relation between two variables?: A relation in which the value of the dependent variable is not uniquely determined by the independent variable; plotted observations scatter around, rather than fall exactly on, the line or curve of the relationship.

What are the regressor and the regressand?: The regressor (x) is the independent/predictor variable forming the basis of estimation, fixed in advance and free of random error; the regressand (Y) is the dependent/response variable being predicted, subject to random variation for any fixed x.

What is the least squares principle?: The principle, due to Adrien Legendre, that the best-fitting curve to a set of data is the one for which the sum of squares of the residuals (observed minus estimated values) is as small as possible.

What is a scatter diagram?: A plot of points in a rectangular coordinate system, one point per observed pair (x_i, y_i), used to visually assess the nature (linear/curvilinear, direct/inverse) and strength of the relationship between two variables before fitting a curve.

What is the simple linear regression coefficient, beta?: The relative change in the expected value of the dependent random variable Y per unit increase in the independent variable x; it is the (constant) slope of the population regression line mu_Y|x = alpha + beta.x.

What is the covariance of two variables, s_xy?: A measure of the linear mutual variability of two variables, s_xy = [sum(x_i-xbar)(y_i-ybar)]/n; positive when the variables move together, negative when they move oppositely.

What is the sample correlation coefficient, r?: A number between -1 and +1, r = s_xy/(s_x.s_y), measuring the strength and direction of the linear relationship between two random variables X and Y.

What does 'no correlation' mean between two variables?: One least squares regression line is horizontal and the other is vertical; equivalently, if X and Y are independent then Cov(X,Y)=0 and the correlation coefficient rho=0.

Why doesn't a high correlation prove causation?: A high correlation only shows the data are consistent with a linear association; it may instead arise

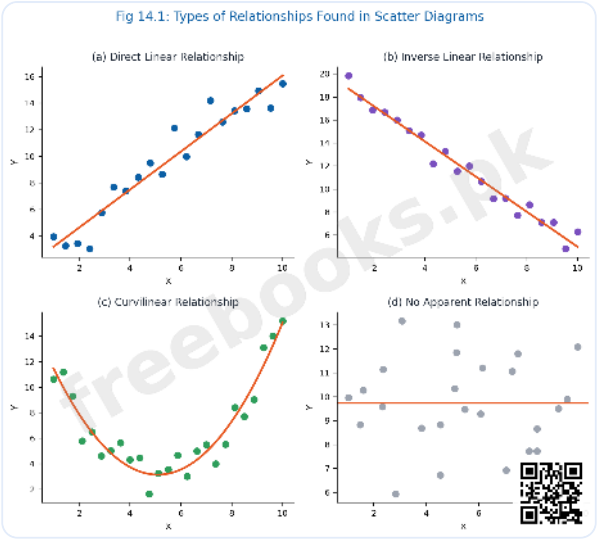
because Y causes X, because a third factor affects both variables, or purely by chance -- further investigation is needed to establish a genuine causal link.

Key Facts and Relations

Topic	Key Fact / Relation
Population simple linear regression	mu_Y|x = alpha + beta.x; Y_i = alpha + beta.x_i + epsilon_i
Sample simple linear regression	y-hat = a + bx; y_i = a + b.x_i + e_i, e_i = y_i - y-hat_i
Least squares slope, b	b = [n.sum(x_i y_i) - (sum x_i)(sum y_i)] / [n.sum(x_i^2) - (sum x_i)^2] = s_xy/s_x^2
Least squares intercept, a	a = [sum y_i - b.sum x_i]/n = ybar - b.xbar
Covariance, s_xy	s_xy = [sum(x_i-xbar)(y_i-ybar)]/n = [sum x_i y_i]/n - xbar.ybar
Sample correlation coefficient, r	r = s_xy/(s_x.s_y) = [n.sum(x_i y_i)-(sum x_i)(sum y_i)] / sqrt{[n.sum x_i^2-(sum x_i)^2][n.sum y_i^2-(sum y_i)^2]}
Regression coefficients via r	b_yx = s_xy/s_x^2 = r.s_y/s_x; b_xy = s_xy/s_y^2 = r.s_x/s_y
r as geometric mean of regression coefficients	r = (+/-) sqrt(b_yx . b_xy), sign = common sign of b_yx and b_xy
Regression equation of Y on X (via r)	y-hat = ybar + (r.s_y/s_x)(x - xbar)
Regression equation of X on Y (via r)	x-hat = xbar + (r.s_x/s_y)(y - ybar)
Coding and scaling relation	u_i=(x_i-p)/h, v_i=(y_i-q)/k => b_yx = (k/h).b_vu
Sum of squared residuals identity	sum(e_i^2) = sum(y_i^2) - a.sum(y_i) - b.sum(x_i y_i) [minimum value over all straight lines]

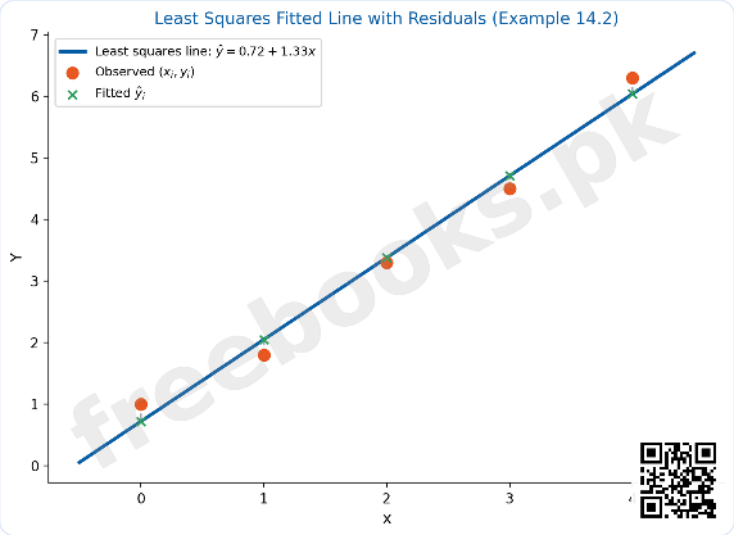
Diagrams

Scatter Diagram Patterns: A 2x2 panel of scatter diagrams illustrating the main relationship patterns covered in the chapter: direct linear relationship (positive slope, points rising left to right), inverse linear relationship (negative slope, points falling left to right), a curvilinear relationship (points following a bending curve rather than a line), and no apparent relationship (points scattered with no discernible pattern), matching Fig 14.1 of the textbook



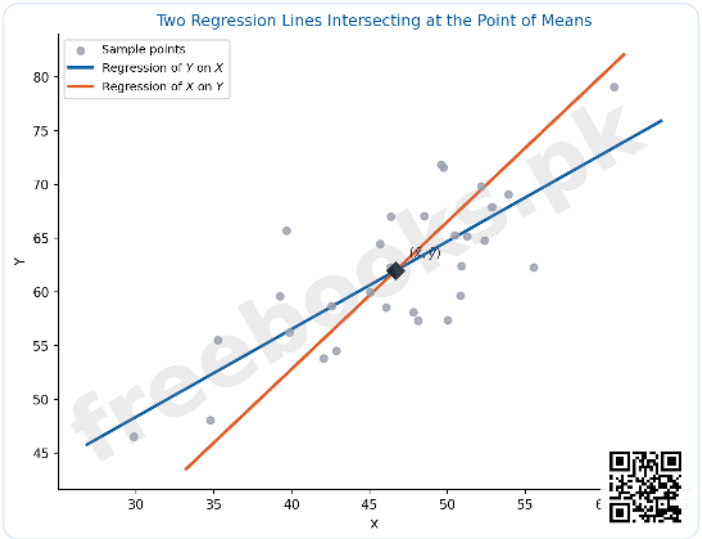
Scatter Diagram Patterns

Least Squares Fitted Line with Residuals: A scatter diagram of Example 14.2's data (x=0,1,2,3,4; y=1.0,1.8,3.3,4.5,6.3) with the least squares regression line y-hat=0.72+1.33x drawn through the points, and a vertical dashed segment from each observed point down (or up) to the fitted line representing its residual e_i, illustrating the least squares principle of minimizing the sum of squared residuals



Least Squares Fitted Line with Residuals

Two Regression Lines Intersecting at the Point of Means: A scatter diagram with both least squares lines drawn -- the regression of Y on X and the regression of X on Y -- shown crossing exactly at the point (xbar, ybar), illustrating Theorem 14.4 and how the angle between the two lines narrows as the correlation |r| approaches 1 (lines nearly coincide) and widens as |r| approaches 0 (lines nearly perpendicular)



Two Regression Lines Intersecting at the Point of Means

Short Questions & Answers

Distinguish between a functional relation and a statistical relation.

A functional relation is exact -- y=f(x) uniquely determines y from x, and all points fall exactly on the curve; a statistical relation is not exact -- y is not

uniquely determined by x, and observed points scatter around, rather than lie exactly on, the relationship.

Define the regressor and the regressand.

The regressor (x) is the independent, predictor variable, fixed in advance and free of random error; the regressand (Y) is the dependent, response variable whose value is predicted from x and which is subject to random variation.

State the least squares principle.

Among all possible curves that could be fitted to a set of observed data, the least squares fit is the one for which the sum of squares of the residuals (observed values minus their fitted/estimated values) is the smallest possible.

What does it mean for the least squares line to 'pass through the point of means'?

The fitted regression line y-hat=a+bx always passes exactly through the point (xbar, ybar) -- when x_i=xbar, the fitted value y-hat_i equals ybar exactly; this is one of the guaranteed properties of the least squares solution.

Why is r said to be the geometric mean of the two regression coefficients?

Algebraically, r = +/-sqrt(b_yx . b_xy), because b_yx=s_xy/s_x^2 and b_xy=s_xy/s_y^2, so their product equals s_xy^2/(s_x^2.s_y^2) = r^2; the sign of r matches the common sign of b_yx and b_xy.

Why can a strong correlation not, by itself, prove causation?

A high correlation only shows the data are consistent with a linear association; the same pattern could arise because X causes Y, because Y causes X, because a third factor affects both, or purely by chance -- further investigation beyond the correlation coefficient is required to establish causation.

Long Questions & Answers

Trace the full development of the simple linear regression model from the population regression line through to the sample least squares line, explaining the role of each symbol and using Example 14.1 to illustrate the computations.

Regression analysis begins by recognising that in most real-world situations the relationship between two variables is statistical rather than functional -- for a given value of the independent, non-random variable x, the dependent variable Y does not take on one single fixed value but instead behaves as a random variable, possessing its own probability distribution with a mean mu_Y|x=E(Y|x) and a variance sigma_Y|x^2=Var(Y|x). It is assumed that this mean, mu_Y|x, changes in some constant, systematic way as x changes, while the variance sigma_Y|x^2 is assumed constant across all values of x, equal to a single common value sigma^2 -- an assumption sometimes called homoscedasticity. When it is reasonable to assume that all these conditional means mu_Y|x lie along a single straight line, the population regression line is written mu_Y|x = alpha + beta.x, where alpha is the population y-intercept (the value of mu_Y|x when x=0) and beta is the population slope, also called the population regression coefficient, representing the change in the mean of Y per unit increase in x. For a specific value x_i drawn from x's domain, the associated random variable Y_i has mean mu_Y|xi=alpha+beta.x_i and constant variance sigma^2; the deviation of Y_i from this mean is called the random error or disturbance, epsilon_i=Y_i-(alpha+beta.x_i), leading to the population regression model Y_i=alpha+beta.x_i+epsilon_i, which is simple (one regressor), linear in its parameters, and linear in x. In practice, alpha, beta, mu_Y|x, and sigma^2 are all unknown population parameters that must be ESTIMATED from a sample of n observed pairs (x_1,y_1), (x_2,y_2), ..., (x_n,y_n). It is customary to let a denote the best (least squares) estimate of alpha, b denote the best estimate of beta, and y-hat denote the resulting estimate of mu_Y|x, giving the sample simple linear regression line y-hat=a+bx, with y-hat_i=a+b.x_i being the best point estimate of mu_Y|xi for any specific x_i. The difference between an actually observed y_i and its fitted value y-hat_i is the residual, e_i=y_i-y-hat_i=y_i-a-b.x_i, an estimate of the unobservable population error epsilon_i, giving the sample regression model y_i=a+b.x_i+e_i. The values a and b that make this

fitted line the 'best' one are found using the method of least squares, formalised in Theorem 14.1: b=[n.sum(x_iy_i)-(sum x_i)(sum y_i)]/[n.sum(x_i^2)-(sum x_i)^2], which can also be written as s_xy/s_x^2, and a=[sum y_i - b.sum x_i]/n=ybar-b.xbar. Example 14.1 of the textbook illustrates the full computation concretely: eight students' matriculation scores x_i (480,490,510,510,530,550,610,640) are paired with their college grade point averages y_i (2.7,2.9,3.3,2.9,3.1,3.0,3.2,3.7). Summing gives sum x_i=4320, sum y_i=24.8, sum(x_i^2)=2355800, and sum(x_iy_i)=13492, with n=8. Substituting into the least squares formula gives b=[8(13492)-(4320)(24.8)]/[8(2355800)-(4320)^2]=0.00435, and then a=[24.8-0.00435(4320)]/8=0.751, so the best fitted line is y-hat=0.751+0.00435x. This estimated line can now be used for prediction: for a student scoring x=600 on matriculation, the estimated mean GPA is y-hat=0.751+0.00435(600)=3.361. Several important properties are guaranteed to hold for any least squares line, and can be checked against this example: the sum of the residuals sum(e_i) is exactly zero (any apparent nonzero result is due only to rounding); the sum of the observed y-values equals the sum of the fitted y-hat values, so the mean of the fitted values equals ybar; the sum of squared residuals sum(e_i^2) achieves the smallest possible value of any straight line through this data (that is the entire point of the least squares principle); and the fitted line always passes exactly through the point of means (xbar, ybar) -- here xbar=4320/8=540 and ybar=24.8/8=3.1, and indeed y-hat=0.751+0.00435(540)=3.10 confirms this.

Define the sample correlation coefficient r, list its key properties, explain how it connects algebraically to the two regression coefficients b_yx and b_xy, and explain why correlation does not establish causation.

Correlation analysis, unlike regression analysis, treats both variables X and Y symmetrically as jointly random -- there is no distinguished 'independent' or 'dependent' variable, only a bivariate population from which pairs (x_i,y_i) are randomly sampled, such as the heights and weights of a group of people or the marks obtained by students in two different subjects. Given a random sample of n such pairs, the sample correlation coefficient, denoted r or r_xy, is defined as r=s_xy/(s_x.s_y), the covariance of X and Y divided by the product of their individual standard deviations; this expands algebraically into several

computationally convenient equivalent forms, the most commonly used being r=[n.sum(x_iy_i)-(sum x_i)(sum y_i)]/sqrt{[n.sum(x_i^2)-(sum x_i)^2][n.sum(y_i^2)-(sum y_i)^2]}. This coefficient r has a set of well-established properties that make it such a useful summary measure. First, r is symmetrical with respect to X and Y, meaning r_xy=r_yx -- unlike the two regression coefficients, which generally differ depending on which variable is treated as dependent. Second, r can be interpreted as the covariance between the two variables after each has been converted into standard (z-score) units, r=Cov(z_x,z_y). Third, and very usefully in practice, the MAGNITUDE of r, |r|, is completely unaffected by a change of origin (adding or subtracting a constant from either variable) or a change of scale (multiplying or dividing by a constant of the SAME sign in both variables) -- this is precisely why, in Example 14.5 of the textbook, computing r directly from the raw height and weight data gives exactly the same value, 0.956, as computing it from deviations taken from arbitrarily assumed means of 70 and 120 respectively; only if the variables are rescaled by constants of OPPOSITE sign does r flip in sign. Fourth, r always lies within the closed interval [-1,+1], with the extreme values +1 and -1 occurring only when every single sample point lies exactly on a straight line with positive or negative slope respectively (perfect correlation), and a value near zero indicating little or no linear association (though a strong NONLINEAR association could still be present even when r is near zero, since r only measures LINEAR association). Fifth, and most directly connecting correlation back to regression, |r| is the geometric mean of the two regression coefficients b_yx (the slope of Y regressed on X) and b_xy (the slope of X regressed on Y): r=+/-sqrt(b_yx . b_xy), with the sign of r matching the common sign shared by b_yx and b_xy (since s_x and s_y are always positive, the sign of s_xy determines the sign of both regression coefficients and of r together, and if any one of s_xy, b_yx, b_xy, or r is zero then all four must be zero). This tight relationship can also be written directly in terms of r: b_yx=r.s_y/s_x and b_xy=r.s_x/s_y, which lets the two regression equations be rewritten using r explicitly as y-hat=ybar+(r.s_y/s_x)(x-xbar) and x-hat=xbar+(r.s_x/s_y)(y-ybar). A further set of theorems ties the geometry together neatly: the two regression lines always intersect at the point of means (xbar,ybar) regardless of the value of r; r itself equals the common slope of both regression lines when the data are expressed in standardised (z-score) units; and the two regression lines become IDENTICAL (coincide exactly) precisely when every

sample point lies on a single straight line, that is, when |r|=1 -- conversely, when r=0, the two lines are at right angles to one another (one horizontal, one vertical), reflecting 'no correlation'. Despite this rich and precise mathematical structure, it is essential to recognise that even a correlation coefficient very close to +1 or -1 does NOT, by itself, prove that one variable causes changes in the other. A high correlation is consistent with (but does not distinguish between) several very different underlying explanations: X may genuinely cause Y; Y may instead cause X; some third, unmeasured factor Z may independently influence both X and Y, generating a 'spurious' correlation between them even though neither causes the other directly (the textbook's own example is a high correlation between city temperature and birth rate, which plainly reflects no direct causal mechanism at all); or the observed correlation may simply be a product of chance, especially in small samples. Establishing genuine causation requires evidence well beyond the correlation coefficient itself -- controlled experiments, a plausible causal mechanism, ruling out confounding third factors, and, ideally, replication of the finding in independent samples.

Multiple Choice Questions (MCQs)

A relation y=f(x) in which the value of y is uniquely and exactly determined by x is called: (A) A statistical relation (B) A functional relation (C) A correlation (D) A regression curve

Correct answer: (B) A functional relation. An exact relation, where every point falls precisely on the curve, is called a functional relation.

The independent, predictor variable in a regression problem is called the: (A) Regressand (B) Response variable (C) Regressor (D) Residual

Correct answer: (C) Regressor. The regressor (x) forms the basis of estimation/ prediction; the regressand (Y) is the variable being predicted.

The least squares principle chooses the fitted curve that minimizes: (A) The sum of the residuals (B) The sum of squares of the residuals (C) The correlation coefficient (D) The sample size

Correct answer: (B) The sum of squares of the residuals. Least squares minimizes sum(e_i^2), the sum of squared residuals, not the (unsquared) sum which is always zero for the optimal line.

In the population regression model Y_i = alpha + beta.x_i + epsilon_i, beta represents: (A) The y-intercept (B) The change in mean of Y per unit increase in x (C) The variance of Y (D) The correlation coefficient

Correct answer: (B) The change in mean of Y per unit increase in x. Beta is the population regression coefficient / slope -- the change in mu_Y|x per unit increase in x.

The least squares regression line always passes through the point: (A) (0,0) (B) (xbar, ybar) (C) (max x, max y) (D) (median x, median y)

Correct answer: (B) (xbar, ybar). One of the guaranteed properties of the least squares line is that it passes exactly through the point of means (xbar, ybar).

If the values of x are multiplied by a positive constant h and y by a positive constant k, the regression coefficient b_yx is: (A) Unaffected (B) Multiplied by k/h (C) Multiplied by h/k (D) Set to zero

Correct answer: (B) Multiplied by k/h. By the coding/scaling theorem, b_yx = (k/h).b_vu -- the regression coefficient DOES change with a change of scale (unlike r's magnitude).

The sample correlation coefficient r always lies in the range: (A) 0 to 1 (B) -1 to 0 (C) -1 to +1 (D) -infinity to +infinity

Correct answer: (C) -1 to +1. By its mathematical construction, r always satisfies -1<=r<=1.

If b_yx = 0.8 and b_xy = 0.45 (both positive), the correlation coefficient r equals: (A) 1.25 (B) 0.6 (C) 0.36 (D) -0.6

Correct answer: (B) 0.6. r = +sqrt(b_yx . b_xy) = sqrt(0.8*0.45) = sqrt(0.36) = 0.6 (positive since both regression coefficients are positive).

The two regression lines (Y on X, and X on Y) coincide exactly when: (A) r = 0 (B) r = 0.5 (C) |r| = 1 (D) The sample size is large

Correct answer: (C) |r| = 1. The two regression lines become identical only when every point lies exactly on one line, i.e. when |r|=1 (perfect correlation).

A high correlation coefficient between two variables: (A) Always proves X causes Y (B) Never has any practical meaning (C) Does not, by itself, establish a causal relationship (D) Means the two regression lines are at right angles

Correct answer: (C) Does not, by itself, establish a causal relationship. Correlation and regression make no assertion about causality; a high r may reflect direct causation, reverse causation, a common third cause, or chance.

Quick Revision Summary

- Functional relation = exact, y=f(x) | Statistical relation = not exact, points scatter around the curve
- Regressor (x) = independent/predictor, fixed in advance | Regressand (Y) = dependent/response, random
- Least squares principle: minimize sum(e_i^2), the sum of SQUARED residuals
- Population model: mu_Y|x = alpha + beta.x, Y_i = alpha + beta.x_i + epsilon_i
- Sample model: y-hat = a + bx; b = s_xy/s_x^2; a = ybar - b.xbar
- Least squares line ALWAYS passes through (xbar, ybar); sum(e_i)=0; sum(y_i)=sum(y-hat_i)
- Covariance s_xy: positive = variables move together, negative = move oppositely
- Correlation r = s_xy/(s_x s_y); |r| unaffected by change of origin/scale; -1<=r<=1
- r = +/-sqrt(b_yx . b_xy); b_yx = r.s_y/s_x; b_xy = r.s_x/s_y
- Two regression lines meet at (xbar, ybar); coincide when |r|=1; right angles when r=0
- Regression/correlation never proves causation -- could be X causes Y, Y causes X, a third factor, or chance

Exam Tips

- Always draw or imagine the scatter diagram FIRST -- it tells you whether a straight-line fit even makes sense before you compute anything
- Remember b (slope) uses s_xy/s_x^2, NOT s_xy/(s_x.s_y) -- that second formula is for r, a very common mix-up
- When x-values are equally spaced, use the coding trick (origin at the middle, unit = the common interval) to avoid large, error-prone numbers
- Check your answer: does the fitted line pass through (xbar, ybar)? If not, recheck your arithmetic

- |r| is unaffected by adding/subtracting or multiplying/dividing by a POSITIVE constant -- use this to simplify awkward raw data before computing r
- Never claim causation from r alone in an exam answer -- always note that regression/correlation analysis makes no assertion about cause and effect