Contents

CHAPTER 10

SIMPLE REGRESSION AND CORRELATION

SIMPLE REGRESSION AND CORRELATION

0.1 INTRODUCTION

The term *regression* was introduced by the English biometrician, Sir Francis Galton (1822–1911) describe a phenomenon which he observed in analyzing the heights of children and their parents. He und that, though tall parents have tall children and short parents have short children, the average height children tends to *step back* or to *regress* toward the average height of all men. This tendency toward e average height of all men was called a *regression* by Galton.

Today, the word *regression* is used in a quite different sense. It investigates the *dependence* of one riable, conventionally called the *dependent variable*, on one or more other variables, called *ependent variables,* and provides an equation to be used for estimating or predicting the average value the dependent variable from the known values of the independent variable. The dependent variable is umed to be a random variable whereas the independent variables are assumed to have *fixed* values, *i.e.* y are chosen non-randomly. The relation between the expected value of the dependent variable and the ependent variable is called a *regression relation.* When we study the dependence of a variable on a gle independent variable, it is called a *simple* or *two-variable regression*. When the dependence of a able on two or more than two independent variables is studied, it is called *multiple regression.* rthermore, when the dependence is represented by a straight line equation, the regression is said to be *ar,* otherwise it is said to be *curvilinear*.

It is relevant to note that in regression study, a variable whose variation we try to explain is a *endent variable* while an *independent variable* is a variable that is used to explain the variation in the endent variable.

Some more terminology: The dependent variable is also called the *regressand*, the *predictand,* the *nse* or the *explained variable* whereas the independent or the non-random variable is also referred to e *regressor*, the *predictor*, the *regression variable* or the *explanatory variable*.

2 DETERMINISTIC AND PROBABILISTIC RELATIONS OR MODELS

The relationship among variables may or may not be governed by an exact physical law. For enience, let us consider a set of n pairs of observation (X_i, Y_i). If the relation between the variables is *y linear*, then the mathematical equation describing the linear relation is generally written as

$$Y_i = a + bX_i.$$

e a is the value of Y when X equals zero and is called the *Y-intercept,* and b indicates the change in Y one-unit change in X and is called the *slope* of the line. Substituting a value for X in the equation, we completely determine a *unique* value of Y. The linear relation in such a case is said to be a *ministic model.* An important example of the deterministic model is the relationship between Celsius ahrenheit scales in the form of $F = 32 + \frac{9}{5}C$. Another example is the area of a circle expressed by lation, area $= \pi r^2$. Such relations cannot be studied by regression.

In contrast to the above, the linear relationship in some situations is *not exact.* For example, we t precisely determine a person's weight from his height as the relationship between them is not ed to follow an exact linear form. The weights for given values of age are reasonably assumed to e measurement of random errors. The deterministic relation in such cases is then modified to allow for the inexact relationship between the variables and we get what is called a *non-determinis* or *probabilistic model* as

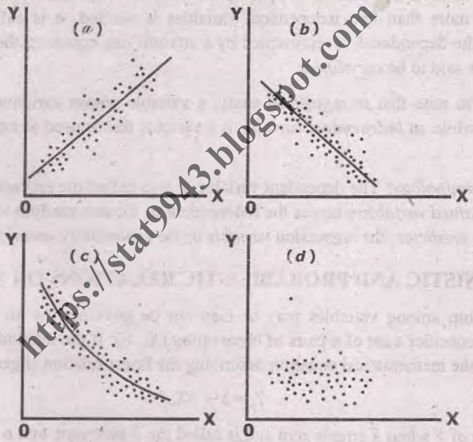
$$Y_i = a + bX_i + e_i, \qquad (i = 1, 2, \ldots, n)$$

where e_i's are the unknown random errors.

10.3 SCATTER DIAGRAM

A first step in finding whether or not a relationship between two variables exists, is to plot each pair of independent-dependent observations $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ as a point on graph paper, using the X-axis for the regression variable and the Y-axis for the dependent variable. Such a diagram is called a *scatter diagram* or a *scatter plot*. If a relationship between the variables exists, then the points in the scatter diagram will show a tendency to cluster around a straight line or some curve. Such a line or curve around which the points cluster, is called the *regression line* or *regression curve* which can be used to estimate the expected value of the random variable Y from the values of the nonrandom variable X.

The scatter diagrams shown below reveal that the relationship between two variable in (a) is positive and linear, in (b) is negative and linear, in (c) is curvilinear and in (d) there is no relationship.



10.4 SIMPLE LINEAR REGRESSION MODEL

We assume that the linear relationship between the dependent variable Y_i and the value X_i of the regressor X is

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where the X_i's are fixed or predetermined values,

the Y_i's are observations randomly drawn from a population,

the ε_i's are error components or random deviations,

α and β are population parameters, α is the intercept and the slope β is called *regression coefficient,* which may be positive or negative depending upon the direction of the relationship between X and Y.

Furthermore, we assume that

i) $E(\varepsilon_i) = 0$, *i.e.* the expected value of error term is zero, it implies that the expected value of Y is related to X in the population by a straight line;

ii) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all *i, i.e.* the variance of error term is constant. It means that the distribution of error has the same variance for all values of X. (*Homoscedasticity assumption*);

iii) $E(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, *i.e.* error terms are independent of each other (*assumption of no serial or auto correlation between* ε'*s*);

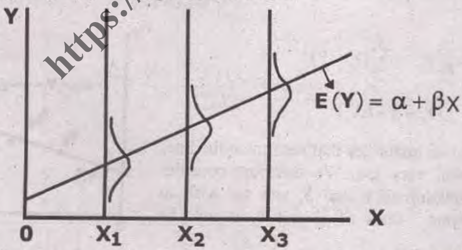
iv) $E(X, \varepsilon_i) = 0$, *i.e.* X and ε are also independent of each other;

v) ε_i'*s* are normally distributed with a mean of zero and a constant variance σ^2. This implies that Y values are also normally distributed. The distributions of Y and ε *are identical except that they have different means. This assumption is required for estimation and testing of hypothesis on linear regression.*

According to this population regression model, each Y_i is an observation from a normal distribution ... mean $= \alpha + \beta X$ and variance $= \sigma^2$. Thus the relation may be expressed alternatively as

$$E(Y) = \alpha + \beta X,$$

... implies that the expected value of Y is linearly related to X and the observed value of Y deviates ... the line $E(Y) = \alpha + \beta X$ by a random component ε, *i.e.* $\varepsilon_i = Y_i - (\alpha + \beta X_i)$. The following graph ... the assumed line, giving $E(Y)$ for the given values of X.



... in practice, we have a sample from some population, therefore we desire to estimate the ... regression line from the sample data. Then the basic relation in terms of sample data may be ...

$$Y_i = a + bX_i + e_i,$$

... b and e_i are the estimates of α, β and ε_i. The estimated regression is generally written as ... $+ bX_i$.

Many possible regression lines could be fitted to the sample data, but we choose that part which *best* fits that data. The *best* regression line is obtained by estimating the regression para the most commonly used *method of least squares* which we describe in the following subsection.

10.4.1 An Aside—The Principle of Least Squares. *The principle of least squares (LS) of determining the values of the unknown parameters that will minimize the sum of squares of residuals) where errors are defined as the differences between observed values and the corre values predicted or estimated by the fitted Model equation.*

The parameter values thus determined, will give the *least* sum of the squares of errors known as *least squares estimates.* The method of least squares that gets its name from the of a sum of squared deviations is attributed to Karl F. Gauss (1777–1855). Some people belie method was discovered at the same time by Adrien M. Legendre (1752–1833), Piere S. (1749–1827) and others. Markov's name is also mentioned in connection with its further develo recent years, efforts have been made to find better methods of fitting but the least squares metho dominant and is used as one of the important methods of estimating the population parameters.

10.4.2 Least-Squares Estimates in Simple Linear Regression. Let there be a set of obse $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$,

$$Yi = \alpha + \beta Xi + \varepsilon_i$$

or in terms of sample data as

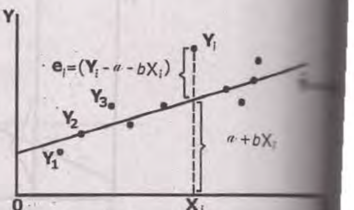
$$Yi = a + bXi + e_i,$$

where a and b are the *least-squares estimates* of α and β, e_i commonly called *residual*, is the of the observed Y_i from its estimate provided by $Y_i = a + bX_i$.

According to the principle of least-squares, we determine those values of a and b minimize the sum of squares of the residuals. In other words, the *best* regression line is the minimizes the sum of the squares of the vertical deviations between the observed values corresponding values predicted by the regression model, *i.e.* $Y_i = a + bX_i$. That is the least minimizes

$$S(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

$$= \sum (Y_i - a - bX_i)^2$$



As a and b, the two quantities that determine the line, vary, $S(a, b)$ will vary too. We therefore consider $S(a, b)$ as a function of a and b, and we wish to determine at what values of a and b, it will be minimum.

Minimizing $S(a, b)$, we need to set its partial derivatives *w.r.t* a and b equal to zero. The

$$\frac{\partial S(a, b)}{\partial a} = 2\sum(Y_i - a - bX_i)(-1) = 0, \text{ and}$$

$$\frac{\partial S(a, b)}{\partial b} = 2\sum(Y_i - a - bX_i)(-X_i) = 0$$

Simplifying, we obtain the following two equations, called the *normal equations* (the word *normal* sed here in the sense of regular or standard).

$$\sum Y_i = na + b\sum X_i \text{ and } \sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

These two normal equations are solved simultaneously for values of a and b either by direct ation or by using determinants.

i) **Direct Elimination:** Multiplying the first equation by $\sum X_i$ and the second equation by n, we get $\sum X \sum Y = na\sum X + b(\sum X)^2$ and $n\sum XY = na\sum X + nb\sum X^2$

Subtracting, we get

$$n\sum XY - \sum X \sum Y = b[n\sum X^2 - (\sum X)^2]$$

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

is the least-squares estimate of the regression co-efficient β.

Similarly, we get

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

east squares estimate of α.

lternatively, we divide the first normal equation by n, and get the least-squares estimate of α as

$$a = \overline{Y} - b\overline{X}.$$

is also shows that the estimated regression line passes through ($\overline{X}$, $\overline{Y}$), the means of the data.

By means of determinants, the solution is

$$b = \frac{\begin{vmatrix} \sum XY & \sum X \\ \sum Y & n \end{vmatrix}}{\begin{vmatrix} \sum X^2 & \sum X \\ \sum X & n \end{vmatrix}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}, \text{ and}$$

$$a = \frac{\begin{vmatrix} \sum X^2 & \sum XY \\ \sum X & \sum Y \end{vmatrix}}{\begin{vmatrix} \sum X^2 & \sum X \\ \sum X & n \end{vmatrix}} = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2},$$

se estimates give us the regression equation

$$\hat{Y}_i = a + bX_i$$
$$= \overline{Y} + b(X - \overline{X}).$$

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

Since Y is a random variable, therefore the deviations in the Y direction are taken into acc determining the *best-fitting* line.

It is very important to note that, when both X and Y are observed at random, i.e. the sample are from a bivariate population, there are two regression equations, each obtained by choosing variable as dependent whose average value is to be estimated and treating the other varia independent. In case of a single random variable, the single regression equation is used to estima values of either the dependent or the independent variable. In case of two regression lines, it is cus to denote the regression coefficients of Y on X and of X on Y by b_{yx} and b_{xy} respectively.

Example 10.1 Compute the least squares regression equation of Y on X for the followi What is the regression coefficient and what does it mean?

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50

The estimated regression line of Y on X is

$$\hat{Y} = a + bX,$$

and the two normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

To compute the necessary summations, we arrange the computations in the table below:

	X	Y	XY	X^2
	5	16	80	25
	6	19	114	36
	8	23	184	64
	10	28	280	100
	12	36	432	144
	13	41	533	169
	15	44	660	225
	16	45	720	256
	17	50	850	289
Total	102	302	3853	1308

Now $\bar{X} = \frac{\sum X}{n} = \frac{102}{9} = 11.33$, $\bar{Y} = \frac{\sum Y}{n} = \frac{302}{9} = 33.56$,

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{9(3853) - (102)(302)}{9(1308) - (102)^2}$$

$$= \frac{34677 - 30804}{11772 - 10404} = \frac{3873}{1368} = 2.831, \text{ and}$$

$$a = \overline{Y} - b\,\overline{X} = 33.56 - (2.831)\,(11.33) = 1.47.$$

ce the desired estimated regression line of Y on X is

$$\hat{Y} = 1.47 + 2.831X.$$

The estimated regression co-efficient, $b = 2.831$, which indicates that the values of Y increase by units for a unit increase in X.

Example 10.2 In an experiment to measure the stiffness of a spring, the length of the spring under t loads was measured as follows:

X=Loads (1b)	3	5	6	9	10	12	15	20	22	28
Y=length (in)	10	12	15	18	20	22	27	30	32	34

Find the regression equations appropriate for predicting

the length, given the weight on the spring;

the weight, given the length of the spring. (W.P.C.S., 1964)

The data come from a bivariate population, *i.e.* both X and Y are random, therefore there are two on lines. To find the regression equation for predicting length (Y), we take Y as dependent e and treat X as independent variable (*i.e.* non-random). For the second regression, the choice of ables is reversed.

The computations needed for the regression lines are given in the following table:

	X	Y	X^2	Y^2	XY
	3	10	9	100	30
	5	12	25	144	60
	6	15	36	225	90
	9	18	81	324	162
	10	20	100	400	200
	12	22	144	484	264
	15	27	225	729	405
	20	30	400	900	600
	22	32	484	1024	704
	28	34	784	1156	932
Total	130	220	2288	5486	3467

The estimated regression equation appropriate for predicting the length, Y, given the weight X, is

$$\hat{Y} = a_0 + b_{yx}\,X,$$

where $b_{YX} = \dfrac{n\sum XY - (\sum X)\,(\sum Y)}{n\sum X^2 - (\sum X)^2} = \dfrac{(10)\,(3467) - (130)\,(220)}{(10)\,(2288) - (130)^2}$

$$= \frac{6070}{5980} = 1.02, \text{ and}$$

$$a_0 = \overline{Y} - b_{yx}\,\overline{X} = 22 - (1.02)\,(13) = 8.74$$

Hence the desired estimated regression equation is

$$\hat{Y} = 8.74 + 1.02\,X$$

ii) The estimated regression equation appropriate for predicting the weight, X, given the length

$$\hat{X} = a_1 + b_{xy}\,Y,$$

$$\text{where } b_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} = \frac{(10)\,(3467) - (130)\,(220)}{(10)\,(5486) - (220)^2}$$

$$= \frac{6070}{6460} = 0.94, \text{ and}$$

$$a_1 = \overline{X} - b_{xy}\overline{Y} = 13 - (0.94)\,(22) = -7.68$$

Hence $\hat{X} = 0.94Y - 7.68$ is the estimated regression equation appropriate for predicting the weight given the length (Y).

10.4.3 Properties of the Least-Squares Regression Line. The least-squares linear regression has the following properties:

$b_{yx} \neq b_{xy}$

i) The least squares regression line always goes through the point ($\overline{X}$, $\overline{Y}$), the means of the
ii) The sum of the deviations of the observed values of Y_i from the least squares regression line is always equal to zero, *i.e.* $\sum(Y_i - \hat{Y}) = 0$.
iii) The sum of the squares of the deviations of the observed values from the least-squares regression line is a minimum, *i.e.* $\sum(Y_i - \hat{Y}_i)^2 = \text{minimum}$.
iv) The least-squares regression line obtained from a random sample is the line of *best fit* because a and b are the unbiased estimates of the parameters α and β.

10.4.4 Standard Deviation of Regression or Standard Error of Estimate. The observed values of (X, Y) do not all fall on the regression line but they scatter away from it. The degree of scatter (or dispersion) of the observed values about the regression line is measured by what is called the standard *deviation of regression* or the *standard error of estimate* of Y on X. For the population data, the standard deviation that measures the variation of observations about the true regression line $E(Y) = \alpha + \beta X$ is denoted by $\sigma_{Y.X}$ and is defined by

$$\sigma_{Y.X} = \sqrt{\frac{\Sigma[Y - (\alpha + \beta X)]^2}{N}}$$

where N is the population size.

For sample data, we estimate $\sigma_{Y.X}$ by $s_{y.x}$ which is defined as

$$s_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}},$$

... $\hat{Y} = a + bX$, the estimated regression line. This is actually an unbiased estimate of $\sigma_{Y.X}$, the ...tion standard deviation about the regression line. The standard error of estimate, $s_{y.x}$ will be zero ... all the observed values fall on the regression line. It is interesting to note that the ranges ..., $\hat{Y} \pm 2s_{y.x}$ and $\hat{Y} \pm 3s_{y.x}$ contain about 68%, 95.4% and 99.7% observations respectively.

To find $\sum(Y - \hat{Y})^2$, we have to calculate $\hat{Y}$ from the estimated regression line for the observed ... of X, which is not an easy task. We therefore use an alternative form obtained as below:

$$\sum(Y - \hat{Y})^2 = \sum(Y_i - a - bX_i)^2$$
$$= \sum Y_i(Y_i - a - bX_i) - a\sum(Y_i - a - bX_i) - b\sum X_i(Y_i - a - bX_i)$$
$$= \sum Y_i^2 - a\sum Y_i - b\sum X_iY_i - a[\sum Y_i - na - b\sum X_i] - b[\sum X_iY_i - a\sum X_i - b\sum X_i^2]$$

But $\sum Y_i - na - b\sum X_i = 0$ and $\sum X_iY_i - a\sum X_i - b\sum X_i^2 = 0$ as they are the normal equations. ...efore

$$\sum(Y_i - \hat{Y})^2 = \sum Y_i^2 - a\sum Y_i - b\sum X_iY_i$$

Hence $s_{y.x} = \sqrt{\dfrac{\sum Y_i^2 - a\sum Y_i - b\sum X_iY_i}{n-2}}$,

... n is the number of pairs.

Example 10.3 Using the data in Example 10.1,

... find the values of $\hat{Y}$ and show that $\sum(Y - \hat{Y}) = 0$, and

... compute the standard error of estimate $s_{y.x}$. = S.D

The calculations needed to find the values of $\hat{Y}$ and the standard error of estimate $s_{y.x}$ are given in ... below:

X	Y	$\hat{Y}$ (=1.47+2.831X)	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	Y^2
5	16	15.625	0.375	0.140625	256
6	19	18.456	0.544	0.295936	361
8	23	24.118	−1.118	1.249924	529
10	28	29.780	−1.780	3.168400	784
12	36	35.442	0.558	0.311364	1296
13	41	38.273	2.727	7.436529	1681
15	44	43.935	0.065	0.004225	1936
16	45	46.766	−1.766	3.118756	2025
17	50	49.597	0.403	0.162409	2500
102	302	301.992	0.008	15.888168	110368

i) The estimated values $\hat{Y}$ appear in the third column of the table on page 403, and
turns out to be 0.008. This small difference is due to rounding off.

ii) The standard error of estimate of Y on X is

$$s_{y.x}=\sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}}=\sqrt{\frac{15.888168}{7}}=\sqrt{2.269738}=1.51$$

Using the alternative form for the calculation of $s_{y.x}$, we get

$$s_{y.x}=\sqrt{\frac{\Sigma Y^2-a\Sigma Y-b\Sigma XY}{n-2}}$$

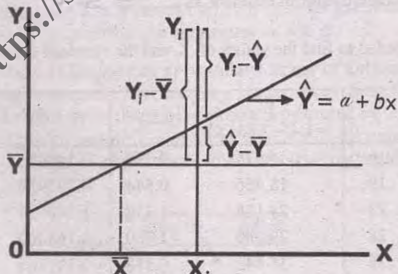
$$=\sqrt{\frac{11368-(1.47)(302)-(2.831)(3853)}{9-2}}$$

$$=\sqrt{\frac{16.217}{7}}=\sqrt{2.316714}=1.52.$$

10.4.5 Co-efficient of Determination. The variability among the values of the dependent
Y, called the *total variation*, is given by $\Sigma(Y-\overline{Y})^2$. This is composed of two parts (i) that which
explained by (associated with) the regression line, *i.e.* $\Sigma(\hat{Y}-\overline{Y})^2$, (ii) that which the regression line
to explain, *i.e.* $\Sigma(Y-\hat{Y})^2$ (see figure). In symbols

$$\Sigma(Y-\overline{Y})^2=\Sigma(Y-\hat{Y})^2+\Sigma(\hat{Y}-\overline{Y})^2$$

Total variation = Unexplained variation + Explained variation



The *co-efficient of determination* which measures the proportion of variability of the values
dependent variable (Y) explained by its linear relation with the independent variable (X), is defined
ratio of the explained variation to the total variation. We use the symbol ρ^2 for the population para

and symbol r^2 for the estimate obtained from sample. Thus the sample co-efficient of determination is given by

$$r^2 = \frac{\text{Explained variation}}{\text{Total variatioin}} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

$$= 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}.$$

An alternative form for calculating the coefficient of determination is

$$r^2 = \frac{a\sum Y + b\sum XY - (\sum Y)^2/n}{\sum Y^2 - (\sum Y)^2/n}.$$

When all the observed values fall on the regression line, then $Y = \hat{Y}$ and $\sum(Y-\bar{Y})^2 = \sum(\hat{Y}-\bar{Y})^2$, and hence $r^2 = 1$. When the observed values are such that $\hat{Y} = \bar{Y}$, then $\sum(\hat{Y}-\bar{Y})^2 = 0$, and hence $r^2 = 0$. This shows that $0 \le r^2 \le 1$. A value of $r^2 = 1$, signifies that 100% of the variability in the dependent variable is associated with the regression equation. When $r^2 = 0$, it means that none of the variability in the dependent variable is explained by X-variable. A value of $r^2 = 0.93$, indicates that 93% of the variability in Y is *explained* by its linear relationship with the independent variable X and 7% of the variation is due to chance or other factors.

Example 10.4 Taking length (Y) as dependent variable for the data in Example 10.2, calculate (i) the total variation, (ii) the unexplained variation, (iii) the explained variation, and (iv) the co-efficient of determination and interpret the coefficient.

In Example 10.2, we found that

$$\sum Y = 220,\ \sum Y^2 = 5486,\ \sum XY = 3467,\ b = 1.02,\ a = 8.74 \text{ and } n = 10.$$

We now find

(i) Total variation $= \sum(Y-\bar{Y})^2 = \sum Y^2 - (\sum Y)^2/n$

$= 5486 - (220)^2/10 = 646$

(ii) Unexplained variation $= \sum(Y-\hat{Y})^2 = \sum Y^2 - a\sum Y - b\sum XY$

$= 5486 - (8.74)(220) - (1.02)(3467)$

$= 5486 - 5459.14 = 26.86$

(iii) Explained variation $=$ Total variation $-$ unexplained variation

$= 646 - 26.86 = 619.14$

(iv) The coefficient of determination, r^2, is given by

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

$$= \frac{619.14}{646} = 0.958$$

A value of $r^2 = 0.958$ indicates that 95.8% of the variability in Y, the length of the spring, is demonstrated by its linear relationship with X, the weight on the spring.

10.5 CORRELATION

Correlation, like covariance, is a measure of the degree to which any two variables *vary together*. In other words, two variables are said to be *correlated* if they tend to simultaneously vary in some direction. If both the variables tend to increase (or decrease) together, the correlation is said to be *direct* or *positive*, *e.g.* the length of an iron bar will increase as the temperature increases. If one variable tends to increase as the other variable decreases, the correlation is said to be *negative* or *inverse*, *e.g.* the volume of gas will decrease as the pressure increases. It is worth remarking that in correlation, we assess the strength of the relationship (or interdependence) between two variables; both the variables are random variables, and they are treated symmetrically, *i.e.* there is no distinction between dependent and independent variable. In regression, by contrast, we are interested in determining the dependence of one variable that is random, upon the other variable that is non-random or fixed, and in predicting the average value of the dependent variable by using the known values of the other variable.

10.5.1 Pearson Product Moment Correlation Co-efficient. A numerical measure of strength of the *linear* relationship between any two variables is called the *Pearson's product moment correlation co-efficient* or sometimes, the *coefficient of simple correlation or total correlation*. The sample linear correlation coefficient for n pairs of observations (X_i, Y_i) usually denoted by the letter r, is defined by

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}}$$

The population correlation co-efficient for a bivariate distribution, denoted by ρ, has already been defined as

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

For computational purposes, we have an alternative form of r as

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

$$= \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

This is a more convenient and useful form, especially when $\bar{X}$ and $\bar{Y}$ are not integers. The coefficient of correlation r is a pure number (*i.e.* independent of the units in which the variables are measured) and it assumes values that can range from +1 for perfect positive linear relationship, to −1 for perfect negative linear relationship with the intermediate value of zero indicating no linear relationship between X and Y. The sign of r indicates the direction of the relationship or correlation.

It is important to note that $r = 0$ does not mean that there is no relationship at all. For example, if all the observed values lie exactly on a circle, there is a perfect *non-linear* relationship between the variables and r will have a value of zero as r only measures the linear correlation.

The linear correlation co-efficient, is also the square root of the linear co-efficient of determination,

We have $\quad \hat{Y} = \overline{Y} + b(X - \overline{X})$

or $\quad \hat{Y} - \overline{Y} = b(X - \overline{X})$

Squaring both sides, we get

$$(\hat{Y} - \overline{Y})^2 = b^2(X - \overline{X})^2$$

Substituting in the ratio, we find

$$\frac{\Sigma(\hat{Y} - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2} = \frac{b^2\,\Sigma(X - \overline{X})^2}{\Sigma(Y - \overline{Y})^2}$$

$$= \frac{\Sigma(X - \overline{X})^2}{\Sigma(Y - \overline{Y})^2}\left[\frac{\Sigma(X - \overline{X})\,(Y - \overline{Y})}{\Sigma(X - \overline{X})^2}\right]^2$$

$$= \left[\frac{\Sigma(X - \overline{X})\,(Y - \overline{Y})}{\sqrt{\Sigma(Y - \overline{Y})^2\;\Sigma(X - \overline{X})^2}}\right]^2 = r^2$$

Example 10.5 Calculate the product moment co-efficient of correlation between X and Y from the following data:

X	1	2	3	4	5
Y	2	5	3	8	7

(P.U., B.A./B.Sc. 1973)

The calculations needed to compute r are given below:

X	Y	$(X-\overline{X})$	$(X-\overline{X})^2$	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
1	2	−2	4	−3	9	6
2	5	−1	1	0	0	0
3	3	0	0	−2	4	0
4	8	1	1	3	9	3
5	7	2	4	2	4	4
15	25	0	10	0	26	13

Here $\overline{X} = \dfrac{\sum X}{n} = \dfrac{15}{5} = 3$, and $\overline{Y} = \dfrac{\sum Y}{n} = \dfrac{25}{5} = 5$

$$\therefore \quad r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2}} = \frac{13}{\sqrt{10 \times 26}} = \frac{13}{16.1} = 0.8$$

Alternatively, the following table is set up for calculation of r.

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	3	9	9	9
4	8	16	64	32
5	7	25	49	35
15	25	55	151	88

$$\therefore \quad r = \frac{\sum XY - (\sum Y)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

$$= \frac{88-(15)\,25/5}{\sqrt{[55-(15)^2/5][151-(25)^2/5]}} = \frac{13}{\sqrt{10 \times 26}} = 0.8$$

10.5.2 Correlation and Causation. The fact that correlation exists between two variabl... not imply any *cause-and-effect* relationship. Two unrelated variables such as the sale of bananas ... death rate from cancer in a city, may produce a high positive correlation which may be due to a ... unknown variable (namely, the city population). The larger the city, the more consumption of ... and the higher will be the death rate from cancer. Clearly, this is a *false* or a merely *incidental* co... which is the result of a third variable, the city size. Such a false correlation between two unco... variables is called *nonsense* or *spurious* correlation. We therefore should be very careful in inte... the correlation coefficient as a measure of relationship or interdependence between two variables.

10.5.3 Properties of r. The sample correlation co-efficient r has the following properties:

i) The correlation co-efficient r is symmetrical with respect to the variables X and Y, *i.e.* r...

ii) The correlation co-efficient lies between −1 and +1, *i.e.* $-1 \le r \le +1$.

iii) The correlation co-efficient is independent of the origin and scale.

Proof: Let u and v be the two new variables defined by $u = \dfrac{X-a}{h}$ and $v = \dfrac{Y-b}{k}$ so that $X = a + ...$ $Y = b + kv$, where a and b are the new origins and h and k are the units of measurement.

Let r_{XY} denote the correlation co-efficient between X and Y and r_{uv}, the correlation co-... between u and v.

Substituting these values in r_{XY}, *viz.*

$$r_{XY} = \frac{\Sigma(X-\overline{X})\,(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2\ \Sigma(Y-\overline{Y})^2}}, \text{ we get}$$

$$r_{XY} = \frac{\Sigma[(a+hu)-(a+h\bar{u})]\,[(b+kv)-(b+k\bar{v})]}{\sqrt{\Sigma[(a+hu)-(a+h\bar{u})]^2 . \Sigma[(b+kv)-(b+k\bar{v})]^2}},$$

where $\overline{X} = a + h\bar{u}$ and $\overline{Y} = b + k\bar{v}$. Therefore

$$r_{XY} = \frac{hk\,\Sigma(u-\bar{u})\,(v-\bar{v})}{hk\sqrt{\Sigma(u-\bar{u})^2 . \Sigma(v-\bar{v})^2}} = r_{uv}$$

This property is very useful in numerical evaluation of r, since due to this property, we can choose ... convenient origin and scale.

...) In case of a bivariate population where both X and Y are random variables, r is the geometric mean between the two regression co-efficient.

That is, if b_{yx} is the regression coefficient of the regression line of Y on X and b_{xy} is the regression ...cient of the regression line of X on Y, and r is the coefficient of correlation, then $r^2 = b_{yx}.b_{xy}$ implies ... $r = \pm\sqrt{b_{yx}.b_{xy}}$.

Since the signs of the regression coefficients depend on the same expression $\Sigma(Y-\overline{X})\,(Y-\overline{Y})$ so ... b_{yx} and b_{xy} are both positive or b_{yx} and b_{xy} are both negative. Therefore

$$r = +\sqrt{b_{yx}.b_{xy}}, \text{ if } b_{yx} \text{ and } b_{xy} \text{ are positive,}$$

$$r = -\sqrt{b_{yx}.b_{xy}}, \text{ if } b_{yx} \text{ and } b_{xy} \text{ are negative.}$$

... the value of r always takes the same sign as the regression coefficients.

The regression co-efficients and the regression lines for a bivariate population, by using the ...on of the correlation co-efficient, may be expressed as

$$\boxed{b_{yx} = r\frac{S_y}{S_x};\ b_{xy} = r\frac{S_x}{S_y}}$$

$$Y-\overline{Y} = r\frac{S_y}{S_x}(X-\overline{X});\ and\ \ X-\overline{X} = r\frac{S_x}{S_y}(Y-\overline{Y}),$$

...e letters have their usual meaning.

Example 10.6 Calculate the co-efficient of correlation between the values of X and Y given below:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Let $u = X - 69$ and $v = Y - 112$. Then $r_{xy} = r_{uv}$. The calculations needed to find r are given in the ... the next page:

X	Y	u	v	u^2	v^2	uv
78	125	9	13	81	169	117
89	137	20	25	400	625	500
97	156	28	44	784	1936	1232
69	112	0	0	0	0	0
59	107	−10	−5	100	25	50
79	136	10	24	100	576	540
68	123	−1	11	1	121	−11
61	108	−8	−4	64	16	32
600	1004	48	108	1530	3468	2160

Now $$r=\frac{\sum uv-(\sum u)(\sum v)/n}{\sqrt{\left[\sum u^2-\frac{(\sum u)^2}{n}\right]\left[\sum v^2-\frac{(\sum v)^2}{n}\right]}}$$

$$=\frac{2160-\frac{48x108}{8}}{\sqrt{\left[1530-\frac{(48)^2}{8}\right]\left[3468-\frac{(108)^2}{8}\right]}}$$

$$=\frac{2160-648}{\sqrt{(1530-288)\times(3468-1458)}}=\frac{1512}{1578}=0.96.$$

Hence the correlation coefficient between X and Y is 0.96.

Example 10.7 If b_{ij} is the regression coefficient of X_i on X_j, then calculate the product coefficient of correlation in each case, given

i) $b_{12}=-0.1,\ b_{21}=-0.4$ ii) $b_{13}=0.27,\ b_{31}=0.6$

iii) $b_{23}=0.67,\ b_{32}=0.38$.

The product moment coefficient of correlation between X_i and X_j is given by

$$r_{ij}=\sqrt{b_{ij}\times b}$$

i) Here $b_{12}=-0.1$, and $b_{21}=-0.4$

$$\therefore \quad r_{12}=-\sqrt{(-0.1)(-0.4)}=-0.20.$$

r is negative since both regression coefficients are negative.

ii) Here both regression coefficients are positive, so r is positive. Thus

$$r_{13}=+\sqrt{b_{13}\times b_{31}}=+\sqrt{(0.27)(0.6)}=+0.40.$$

iii) Here we have

$$r_{23} = \sqrt{(0.67)(0.38)} = 0.50 \quad (\because\ b_{23} \text{ and } b_{32} \text{ are positive})$$

10.5.4 Correlation Co-efficient for Grouped Data. In a simple frequency table, the data are arranged with respect to one variable only. If the arrangement is made according to two variables simultaneously in say, m columns and k rows, the frequency table thus obtained is called a *correlation table* or a *bivariate frequency table*. The number of observations falling in the $(i, j)th$ cell, is called the $(i, j)th$ cell frequency and is denoted by f_{ij}. The correlation co-efficient, if it exists, can be calculated from such a two-way frequency table by using the class midpoints as the value of the observations. The formula for r then becomes

$$r = \frac{\sum f_{ij} X_j Y_i - \frac{1}{n}(\sum f_{.j} X_j)(\sum f_{i.} Y_i)}{\sqrt{\left[\sum f_{.j} X_j^2 - \frac{1}{n}(\sum f_{.j} X_j)^2\right]\left[\sum f_{i.} Y_i^2 - \frac{1}{n}(\sum f_{i.} Y_i)^2\right]}}$$

where $f_{i.} = \sum_{j=1}^{m} f_{ij}$, the frequency of Y values, $f_{.j} = \sum_{j=1}^{k} f_{ij}$, the frequency of X values and n is the total frequency.

Example 10.8 Calculate the co-efficient of linear correlation from the table given below:

Grades in Statistics (Y)	Grades in Mathematics (X)						Total
	40–49	50–59	60–69	70–79	80–89	90–99	
90–99	--	--	--	2	4	4	10
80–89	--	--	1	4	6	5	16
70–79	--	--	5	10	8	1	24
60–69	1	4	9	5	2	--	21
50–59	3	6	6	2	--	--	17
40–49	3	5	4	--	--	--	12
Total	7	15	25	23	20	10	100

(P.U., B.A./B.Sc. 1968)

Let us introduce two new variables u and v given by the relations $u = \frac{X - 64.5}{10}$ and $v = \frac{Y - 74.5}{10}$. Then the calculations needed for finding r are arranged in the table on page (412).

Y_i	X_j / u_j \ v_i	44.5 (−2)	54.5 (−1)	64.5 (0)	74.5 (1)	84.5 (2)	94.5 (3)	$f_{i\cdot}$	$f_{i\cdot}v_i$	$f_{i\cdot}v_i^2$	$f_{ij}u_jv_i$
94.5	2	--	--	--	2 [4]	4 [16]	4 [24]	10	20	40	44
84.5	1	--	--	1 [0]	4 [4]	6 [12]	5 [15]	16	16	16	31
74.5	0	--	--	5 [0]	10 [0]	8 [0]	1 [0]	24	0	0	0
64.5	−1	1 [2]	4 [4]	9 [0]	5 [−5]	2 [−4]	---	21	−21	21	−3
54.5	−2	3 [12]	6 [12]	6 [0]	2 [−4]	---	---	17	−34	68	20
44.5	−3	3 [18]	5 [15]	4 [0]	---	---	---	12	−36	108	33
$f_{\cdot j}$		7	15	25	23	20	10	100	−55	253	125
$f_{\cdot j}u_j$		−14	−15	0	23	40	30	64			
$f_{\cdot j}u_j^2$		28	15	0	23	80	90	236			
$f_{ij}u_jv_i$		32	31	0	−1	24	39	125	← Check		

The number in the corner of each cell represents the product $f_{ij}u_jv_i$, where f_{ij} is the cell frequency. Thus $f_{1,4}\, u_4v_1 = 2(1)\,(2)=4$ and $f_{1,5}u_5v_1=4(2)\,(2) = 16$ and so on. The totals in the last column and the last row are equal and represent $\sum f_{ij}u_jv_i$.

Now $$r_{XY} = r_{uv} = \frac{n\sum fuv - (\sum fu)(\sum fv)}{\sqrt{[n\sum fu^2 - (\sum fu)^2]\,[n\sum fv^2 - (\sum fv)^2]}}$$

(subscripts dropped for convenience in printing)

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}}$$

$$= \frac{16020}{\sqrt{(19504)(22275)}} = 0.77$$

Example 10.9 (a) Correlation between X and Y is r, show that correlation between aX and bY is r or $-r$ according as a and b have the same or different signs.

b) Find correlation between X and Y connected by

$$aX + bY + c = 0.$$

(P.U., B.A. (Hons) Part-II, 19..)

a) Let $u = aX$, so that $\overline{u} = a\overline{X}$,

and $v = bY$, so that $\overline{v} = b\overline{Y}$

Then $(u-\overline{u}) = a(X-\overline{X})$ and $(v-\overline{v}) = b(Y-\overline{Y})$

By definition, we have

$$r_{uv} = \frac{\sum(u-\overline{u})(v-\overline{v})}{\sqrt{\sum(u-\overline{u})^2 \sum(v-\overline{v})^2}}$$

$$= \frac{ab\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{a^2\sum(X-\overline{X})^2 b^2 \sum(Y-\overline{Y})^2}}$$

$$= \frac{ab}{\sqrt{a^2b^2}} r_{XY}$$

$= +r$, if a and b are of the same signs.

$= -r$, if a and b are of the different signs.

b) We are given $aX + bY + c = 0$

Thus $a\sum X + b\sum Y + nc = 0$, where n is the number of pairs of values (X_i, Y_i)

Dividing by n, we get

$a\overline{X} + b\overline{Y} + c = 0$, $\overline{X}$ and $\overline{Y}$ being the means of X and Y sets of observations. Subtracting, we have

$$a(X-\overline{X}) + b(Y-\overline{Y}) = 0$$

or $$(Y-\overline{Y}) = -\frac{a}{b}(X-\overline{X})$$

Now $$r_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2}}$$

$$= \frac{-\frac{a}{b}\sum(X-\overline{X})^2}{\sqrt{\left[\sum(X-\overline{X})^2\right]\left[\frac{a^2}{b^2}\sum(X-\overline{X})^2\right]}} = \frac{-a/b}{\sqrt{\frac{a^2}{b^2}}}$$

$= -1$, if a and b are of the same signs.

$= +1$, if a and b are of the opposite signs.

RANK CORRELATION

Sometimes, the actual measurements or counts of individuals or objects are either not available or assessment is not possible. They are then arranged *in order* according to some characteristic of interest. Such an ordered arrangement is called a *ranking* and the *order given to an individual or object* is called its *rank*. The correlation between two such sets of rankings is known as *Rank Correlation*.

10.6.1 Derivation of Rank Correlation. Let a set of n objects be ranked with respect to char A as $x_1, x_2, \ldots, x_i, \ldots, x_n$, and according to character B as $y_1, y_2, \ldots, y_i, \ldots, y_n$. We assume that no two more objects are given the same ranks (*i.e.* are tied). Then obviously x_i and y_i are some two numbers 1 to n.

Since both x_i and y_i are the first n natural numbers, therefore, we have

$$\sum_{i=1}^{n} x = \sum_{i=1}^{n} y = \sum_{i=1}^{n} i = 1+2+3+\ldots+n = \frac{n(n+1)}{2},$$

$$\sum_{i=1}^{n} x^2 = \sum_{i=1}^{n} y^2 = \sum_{i=1}^{n} (i)^2 = 1^2+2^2+3^2+\ldots+n^2 = \frac{n(n+1)\,(2n+1)}{6},$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^{\,2} - \frac{(\sum y_i)^2}{n}$$

$$= \frac{n(n+1)\,(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}$$

Let d_i denote the difference in ranks assigned to the *ith* individual or object, *i.e.* $d_i = x_i - y_i$

Then $$\sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n}(x_i-y_i)^2$$

$$= \sum(x_i^{\,2}+y_i^{\,2}-2x_iy_i) = \sum x_i^{\,2} + \sum y_i^{\,2} - 2\sum x_iy_i$$

Substituting for $\sum x_i^{\,2}$ and $\sum y_i^{\,2}$, we get

$$\sum_{i=1}^{n} d_i^{\,2} = \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)(2n+1)}{6} - 2\sum x_iy_i$$

or $$\sum x_iy_i = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum d_i^{\,2}$$

The product moment co-efficient of correlation between the two sets of rankings is

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

Substitution gives

$$r_s = \frac{\left[\frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum d_i^{\,2}\right] - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}}$$

$$= \frac{\left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}\right] - \frac{1}{2}\sum d_i^{\,2}}{\frac{n(n^2-1)}{12}}$$

$$=\frac{\frac{n(n^2-1)}{12}-\frac{1}{2}\sum d_i^{\,2}}{\frac{n(n^2-1)}{12}}=1-\frac{6\sum d_i^{\,2}}{n(n^2-1)}.$$

This formula is usually denoted by r_s in order to have a distinction. It is often called Spearman's ficient of rank correlation, in honour of the psychometrician Charles Edward Spearman –1945), who first developed the procedure in 1904.

It is to be noted that $\sum d_i^{\,2}$ has the least value and is zero when the numbers are in complete ment. When they are in complete disagreement, $\sum d_i^{\,2}$ attains the maximum value and is equal to $\frac{n(n^2-1)}{3}$.

Substituting these values in the formula, we see that

$r_s = 1$ for $\sum d_i^{\,2} = 0$, and

$r_s = -1$ for $\sum d_i^{\,2} = \frac{n(n^2-1)}{3}$.

Thus r_s also lies between -1 and $+1$.

Example 10.10 Find the co-efficient of rank correlation from the following rankings of 10 in Statistics and Mathematics.

Statistics (x):	1	2	3	4	5	6	7	8	9	10
Mathematics (y):	2	4	3	1	7	5	8	10	6	9

(P.U., B.A. (Hons.) Part-I, 1964)

We calculate the co-efficient of rank correlation as follows:

x_i	y_i	$d_i(=x_i-y_i)$	d_i^2
1	2	−1	1
2	4	−2	4
3	3	0	0
4	1	3	9
5	7	−2	4
6	5	1	1
7	8	−1	1
8	10	−2	4
9	6	3	9
10	9	1	1
---	---	0	34

Hence, using Spearman's co-efficient of rank correlation, we get

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 34}{10 \times 99} = 1 - 0.2 = +0.8.$$

This indicates a high correlation between Statistics and Mathematics.

10.6.2 Rank Correlation for Tied Ranks. The Spearman's co-efficient of rank correla applies only when no ties are present. In case there are ties in ranks, the ranks are adjusted by assig the mean of the ranks which the tied *objects* or observations would have if they were ordered. example, if two *objects* or observations are tied for fourth and fifth, they are both given the mean ran 4 and 5, *i.e.* 4.5. The sum of adjusted ranks remains $\frac{n(n+1)}{2}$ but $\sum(x_i - \bar{x})^2 \neq \sum(y_i - \bar{y})^2 \neq \frac{n(n^2-}{12}$ has been shown that each set of ties involving t observations reduces the values of d^2 by a quantity to $\frac{1}{12}(t^3 - t)$. In such a situation, one of the following two methods is to be used:

First, for each tie, add a quantity $\frac{1}{12}(t^3 - t)$ to $\sum d^2$ before substituting the values Spearman's co-efficient of rank correlation in order to adjust the formula for the tied observations.

Second, use the product moment co-efficient of correlation to find the correlation between sets of adjusted ranks.

Example 10.11 Two members of a selection committee rank eight persons according suitability for promotion as follows:

Persons	A	B	C	D	E	F	G	H
Member 1	1	2.5	2.5	4	5	6	7	8
Member 2	2	4	1	3	6	6	6	8

Calculate the co-efficient of rank correlation.

We observe that both the sets of rankings contain ties. The coefficient of rank corre therefore calculated as below:

Person	Member 1	Member 2	d	d^2
A	1	2	−1	1
B	2.5	4	−1.5	2.25
C	2.5	1	1.5	2.25
D	4	3	1	1
E	5	6	−1	1
F	6	6	0	0
G	7	6	1	1
H	8	8	0	0
Σ	36	36	0	8.5

For tie between B and C, (first rankings) t=2 and for E, F and G (second rankings) t=3, the quantity to be added to $\sum d^2$ is

$$\frac{1}{12}(2^3-2)+\frac{1}{12}(3^3-3)=2.5.$$

Hence $r_s = 1 - \dfrac{6[8.5+2.5]}{8(64-1)} = 1 - \dfrac{66}{504} = 1 - 0.131 = 0.869.$

...rnative Method:

We see that the first member has tied B and C, while the second member has tied E, F and G. Let ...note the ranks given by the first member by x_i and those of second member by y_i. Then we proceed ...low:

x_i	y_i	x_i^2	y_i^2	x_iy_i
1	2	1	4	2
2.5	4	6.25	16	10
2.5	1	6.25	1	2.5
4	3	16	9	12
5	6	25	36	30
6	6	36	36	36
7	6	49	36	42
8	8	64	64	64
36	36	203.5	202	198.5

Hence the co-efficient of rank correlation is

$$r = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sqrt{[\sum x_i^2 - (\sum x_i)^2/n][\sum y_i^2 - (\sum y_i)^2/n]}}$$

$$= \frac{198.5 - (36)(36/8)}{\sqrt{[203.5-(36)^2/8][202-(36)^2/8]}}$$

$$= \frac{198.5 - 162}{\sqrt{(203.5-162)(202-162)}} = \frac{36.5}{\sqrt{(41.5)(40)}}$$

$$= \frac{36.5}{40.47} = 0.896,$$

...ndicates a high degree of agreement between the two members.

...6.3 Co-efficient of Concordance. The Spearman's co-efficient of rank correlation measures ...ement between two sets of rankings only, but in practice; the individuals or objects are sometimes ... by more than two people. We then need a co-efficient to measure agreement among more than ... of rankings. Such a co-efficient is obtained as below:

...t there be m rankings of n individuals or objects instead of two. Obviously in case of complete ..., the rank totals will form the series $m, 2m, 3m, \ldots, nm$.

...e mean of these totals is

$$\overline{X} = (m + 2m + 3\,m + \ldots + nm) \div n.$$

$$= \frac{m(1+2+3+\ldots+n)}{n} = \frac{m(n+1)}{2},$$

and the variance of these sums, which is the maximum possible, is

$$\text{Var(Total)} = \frac{1}{n}[m^2 + (2m)^2 + (3m)^2 + \ldots + (nm)^2] - \left[\frac{m(n+1)}{2}\right]^2$$

$$= \frac{m^2[1^2 + 2^2 + 3^2 + \ldots + n^2]}{n} - \left[\frac{m(n+1)}{2}\right]^2$$

$$= \frac{m^2(n+1)(2n+1)}{6} - \frac{m^2(n+1)^2}{4} = \frac{m^2(n^2-1)}{12}.$$

But the totals of observed ranks will not necessarily be the same. Let S denote the sum squares of deviations of the totals of the observed ranks from their common mean, .*i.e.* $\frac{m(n+1)}{2}$, the *Co-efficient of Concordance*, W, is defined as the ratio of the variance of the totals of the observed to the variance in case of complete agreement. Thus, we have

$$W = \frac{S}{n} \div \frac{m^2(n^2-1)}{12} = \frac{12S}{m^2(n^3-n)}.$$

This co-efficient is due to Maurice *G*. Kendall (1907–1983) and varies from 0 to 1. When represents complete agreement.

Example 10.12 The following data give rankings of six persons for their ability by three Q and R. Calculate the co-efficient of concordance.

Persons	A	B	C	D	E	F
Judge P	3	1	6	2	5	4
Judge Q	4	3	2	5	1	6
Judge R	2	1	6	5	4	3

(P.U., B.A. (Hons.), Part-

Here the totals of the observed ranks are 9, 5, 14, 12, 10 and 13; m=3 and n=6 so mean $= \frac{m(n+1)}{2} = \frac{3(6+1)}{2} = 10.5$.

Thus $S = (9-10.5)^2+(5-10.5)^2+(14-10.5)^2+(12-10.5)^2+(10-10.5)^2+(13-10.5)^2$

$$= (-1.5)^2+(-5.5)^2+(3.5)^2+(1.5)^2+(-0.5)^2+(2.5)^2 = 53.50$$

Hence $W = \frac{12S}{m^2(n^3-n)} = \frac{12 \times 53.5}{9(216-6)} = \frac{642}{1890} = +0.34$.

EXERCISES

OBJECTIVE

Answer 'True' or 'False'. If the statement is not true then replace the underlined words with words that make the statement true:

i) A high value of correlation between Y and X indicates a high likelihood of a cause and effect relationship between Y and X.

ii) Correlation analysis finds the equation of the line for two variables.

iii) The co-efficient of correlation lies between 0 and +1.

iv) The correlation co-efficient is not independent of the origin and scale.

v) Regression analysis measures the strength of the linear relationship between two variables.

vi) In regression analysis X and Y must both be normally distributed.

vii) The method of least squares gives the line of best fit.

viii) If the co-efficient of determination r^2 is equal to ½, then it indicates that 50% of the variation is due to chance or other factors.

ix) If the slope of the regression line has a negative sign, then the coefficient of determination also is negative.

x) If all the points in a scatter diagram fall on the regression line, then the standard error of estimate equals positive value.

MULTIPLE CHOICE QUESTIONS

When the slope of regression line is negative, the following statistic is also negative

a) r

b) r^2

c) Standard error of estimate

d) Standard error of slope co-efficient

If there is no linear relationship between the two variables then which one of the following does not hold?

a) a = 0

b) b = 0

c) $r^2 = 0$

d) The regression line is either vertical; or horizontal.

iii) If the correlation co-efficient r = 0.7, then the proportion of variation for Y explained by X is

a) 0.49

b) 0.50

c) 0.70

d) $\sqrt{0.70}$

iv) The dependent variable is also known as

a) Explained variable

b) Response variable

c) Predicted variable

d) All of above

v) In the regression equation $Y = \alpha + \beta x + \varepsilon$, both X and Y variables are

a) Random

b) Fixed

c) X is fixed and Y is random

d) Y is fixed and X is random

vi) The variation of the Y values around the regression line is measured by

a) $\sum(Y-\overline{Y})^2$

b) $\sum(Y-\hat{Y})^2$

c) $\sum(\hat{Y}-\overline{Y})^2$

d) None of above

vii) If both the dependent and independent variables increase simultaneously, the correlation coefficient will be in the range of

a) 0 to +1

b) 0 to −1

c) 1 to 2

d) −1 to +1

viii) Which of the following statements is incorrect about correlation coefficient?

a) It passes through the means of the data

b) It is symmetrical with respect to X and Y

c) It is independent of origin and scale

d) It is the geometric mean between the two regression coefficients

ix) If the unexplained variation between variables X and Y is 0.40 then r^2 is

a) 0.75

b) 0.60

c) 0.40

d) None of the above

x) The strength of a linear relationship between two variables Y and X is measured by

a) r^2

b) b_{yx}

c) r

d) None of above

SUBJECTIVE

a) Explain what is meant by (i) *regression*, (ii) *regressand*, (iii) *regressor*, and (iv) regression co-efficient.

b) Differentiate between a deterministic and a probabilistic relationship, giving examples.

a) What is a scatter diagram? Describe its role in the theory of regression.

b) What is a linear regression model? Explain the assumptions underlying the linear regression model.

a) Explain the *principle of least-squares*.

b) Explain briefly how the principle of least squares is used to find a regression line based on a sample of size *n*. Illustrate on a rough sketch the distances whose squares are minimized, taking care to distinguish the dependent and independent variables.

a) Find least-squares estimates of parameters in a simple linear regression model $Y_i = \alpha + \beta X_i + e_i$, where e_i's are distributed independently with mean zero and constant variance.

b) What are the properties of the least-squares regression line?

(P.U., B.A./B.Sc. 1992)

c) Show that the regression line passes through the means of observations.

(P.U., D.St. 1962)

a) Describe briefly how you would obtain the line of regression of one variable (Y) on another variable (X), using the method of least-squares.

(P.U., B.A./B.Sc. 1975)

b) What is meant by the *standard error of estimate*? If the regression line of Y on X is given by $\hat{Y} = a + bX$, prove that the standard error of estimate $s_{y.x}$ is given by

$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}.$$

10.6 Given the following set of values:

X	20	11	15	10	17	19
Y	5	15	14	17	8	9

a) Determine the equation of the least squares regression line.

b) Find the predicted values of Y for X=10, 11, 15, 17, 19, 20.

c) Use the predicted values found in (b) to find the standard error of estimate.

10.7 Given these ten pairs of (X, Y) values:

X	1	1	2	3	4	4	5	6	6	7
Y	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

a) Plot a scatter diagram for the above data.

b) Carry out the necessary computations to obtain the least-squares estimates of the param in the simple linear regression $Y_i = \alpha + \beta X_i + e_i$.

c) Compute the residuals and verify that they add to zero.

d) Use the regression equation to predict the values of Y when X=10.

10.8 For each of the following data, determine the estimated regression equation $\hat{Y} = a + bX$.

a) $\overline{X} = 10; \overline{Y} = 20; \sum XY = 1{,}000; \sum X^2 = 2{,}000; n = 10.$

b) $\sum X = 528; \sum Y = 11{,}720; \sum XY = 193{,}640; \sum X^2 = 11{,}440; n = 32.$

c) $\sum X = 1{,}239; \sum Y = 79; \sum XY = 1{,}613; \sum X^2 = 17{,}322; \ \sum Y^2 = 293; n = 100.$

d) $n = 10, \sum X = 1710, \sum Y = 760, \sum X^2 = 293{,}162, \sum Y^2 = 59{,}390, \ \sum XY = 130{,}628.$

e) $\overline{X} = 52, \overline{Y} = 237, \sum(X - \overline{X})^2 = 2800, \sum(X - \overline{X})(Y - \overline{Y}) = 9871.$

10.9 The owner of a retailing organization is interested in the relationship between price at wh commodity is offered for sale and the quantity sold. The following sample data have be collected.

Price	25	45	30	50	35	40	65	75	70	60
Quantity sold	118	105	112	100	111	108	95	88	91	96

a) Plot a scatter diagram for the above data.

b) Using the method of least squares, determine the equation for the estimated regressio Plot this line on the scatter diagram.

c) Calculate the standard deviation of regression, $s_{y.x}$. (B.Z.U., M.A. Econ.,

10.10 Given the following sets of values:

Y	6.5	5.3	8.6	1.2	4.2	2.9	1.1	3.0
X	3.2	2.7	4.5	1.0	2.0	1.7	0.6	1.9

a) Compute the least-squares regression equation for Y values on X values, that is the equation $\hat{Y} = a + bX$.

b) Compute the standard error of estimate, $s_{y.x}$.

c) Compute the least-squares regression equation for X values on Y values, that is the equation $\hat{X} = a_0 + b_0 Y$.

d) Compute the standard error of estimate, $s_{x.y}$.

11 a) Explain what is meant by the co-efficient of determination.

b) Compute the co-efficient of determination for the following data and interpret the co-efficient.

Income (X) (000)	10	20	30	40	50	60
Expenditure (Y) (000)	7	21	23	34	6	53

2 a) What is the total variation, the explained variation and the unexplained variation?

b) Compute (i) the total variation, (ii) the explained variation and (iii) the unexplained variation for the data in 10.11(b). How much of the variability in Y is explained by the linear regression model?

a) Differentiate between regression and correlation, giving examples.

(P.U., B.A./B.Sc. 1979)

b) Describe the properties of the correlation coefficient.

c) What values may r assume? Interpret the meaning when $r = -1, 0, +1$.

(P.U., B.A./B.Sc. 1980)

a) Define the terms correlation and product moment co-efficient of correlation. Prove that the correlation co-efficient is independent of the origin and scale. (P.U., B.A./B.Sc. 1981)

b) Compute the correlation co-efficient between the variables X and Y represented in the following table:

X	2	4	5	6	8	11
Y	18	12	0	8	7	5

c) Multiply each X value by 2 and add 6. Multiply each Y value by 3 and subtract 15. Find the correlation co-efficient between the two new sets of values, explaining why you do or do not obtain the same result as in part (b).

a) Show that, if r_{XY} is the correlation co-efficient calculated from a set of paired data (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n), then r_{uv}, the correlation co-efficient for $u_i = aX_i + b$ and $v_i = cY_i + d$ (with $a \neq 0$ and $c \neq 0$), is given by $r_{uv} = r_{XY}$.

b) Calculate the correlation co-efficient by first multiplying each X and Y by 10 and then subtracting 70 from each X and 60 from each Y.

X	8.2	9.6	7.0	9.4	10.9	7.1	9.0	6.6	8.4	10.5
Y	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.4

10.16 a) Explain the term correlation. It is known that $r_{XY}=0.7$. Find (i) r_{YX}, (ii) r_{uv}, where $u = -$ and $v=3Y$.

b) Calculate the coefficient of correlation for a sample of 20 pairs of observations, given tha

$$\overline{X}=2, \overline{Y}=8, \sum X^2=180, \sum Y^2=1424 \text{ and } \sum XY=404.$$ (P.U., B.Sc. Hons. 19

10.17 The following data were computed from personnel records of a manufacturing firm:

X = number of years of service, Y = weekly wage rate

$n=23$; $\sum X=2{,}433$; $\sum Y=4{,}245$; $\sum X^2=281{,}019$; $\sum Y^2=841{,}786$ and $\sum XY=482{,}788$.

i) Compute the correlation co-efficient.

ii) If the correlation co-efficient indicates that there does exist a relationship between X an compute the least-squares line of regression. What do the values of a and b signify?

(P.U., B.A./B.Sc.

10.18 Find the product moment co-efficient of correlation between traffic density and accident rate fr the following information available. Find also the coefficient of determination and interpret it.

Traffic Density	30	35	40	45	50	60	70	80	90
Accident Rate	2	4	5	5	8	15	24	30	32

10.19 Given marks as

Student	1	2	3	4	5	6	7	8	9	10	11	12
Economics paper	36	56	41	46	59	46	65	31	68	41	70	36
Physics Paper	62	42	60	53	36	50	42	66	44	58	65	71

Find the co-efficient of correlation and interpret it.

10.20 Calculate the co-efficient of correlation and obtain the lines of regression of the following data

Price (X)	3	4	5	6	7	8	9	10	11	12
Demand (Y)	25	24	20	20	19	17	16	13	10	6

(P.U., M.A. Econ.

10.21 a) Find the correlation co-efficient between X and Y, given

X	5	12	4	16	18	21	22	23	25
Y	11	16	15	20	17	19	25	24	21

(P.U., M.A. Econ.

b) Find the co-efficient of correlation between persons employed and cloth manufactured in a textile mill. Interpret the result

Persons employed	137	209	113	189	176	200	219
Cloth manufactured ('000 yds)	23	47	22	40	39	51	49

(P.U., B.A./B.Sc. 1960)

0.22 The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relation between age and blindness.

Age	No. of Persons in thousand	Blind
0 – 9	100	55
10 – 19	60	40
20 – 29	40	40
30 – 39	36	40
40 – 49	24	36
50 – 59	11	22
60 – 69	6	18
70 – 79	3	15

(P.U., B.A./B.Sc. 1983)

Hint. First calcuiate the numbers of blink per lakh and then correlate with the midpoints of age groups.

23 A computer while calculating the correlation co-efficient between two variables X and Y from 25 pairs of observations obtained the following sums:

$$\sum X = 125,\ \sum X^2 = 650,\ \sum Y = 100,\ \sum Y^2 = 460,\ \sum XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

. Obtain the correct value of the co-efficient of correlation. (P.C.S. 1972; P.U., B.A./B.Sc. 1974)

If the equations of the least squares regression lines are:

a) Y=20.8–0.219X (Y on X), and X = 16.2–0.785Y (X or Y);

b) Y=2.64+0.648X (Y on X), and X = –1.91+0.917Y (X or Y);

c) Y=1.94X+10.83 (Y on X), and X = 0.15Y+6.18 (X or Y);

d) Y=15–1.96X (Y on X), and Y = 15.91–2.22X (X or Y);

Find the product moment coefficient of correlation in each case.

10.25 Find the co-efficient of correlation for the frequency distribution of two variables given b the following table.

Practical

Y \ X	5 – 14	15 – 24	25 – 34	35 – 44	45 – 54
0 – 9	3	1	---	---	---
10 – 19	12	8	14	1	---
20 – 29	2	13	40	12	3
30 – 39	---	3	40	27	7
40 – 49	---	---	6	4	4

Practical

Also find the regression equation of Y and X.

10.26 Compute correlation co-efficient from the following correlation table for weights and heigh women students.

Height in inches	Weight in pounds					
	90	110	130	150	170	190
57	---	---	---	---	1	---
60	8	21	8	1	---	---
63	3	50	57	12	---	2
66	1	24	54	19	3	---
69	---	1	8	5	3	---
72	---	---	2	---	---	1

(P.U., B.A./B.Sc.

10.27 a) Describe in simple words the following concepts:

(i) co-efficient of correlation, (ii) scatter diagram, (iii) least squares princip (iv) estimate of regression co-efficient.

b) The following results were obtained for a bivariate frequency distribution after makin transformation $u = \dfrac{x-1250}{500}$ and $v = \dfrac{y-500}{200}$: $n = 66, \Sigma fu = -4, \Sigma fu^2 = 109, \Sigma fv = -$

$\Sigma fv^2 = 115$, $\Sigma fuv = 91$. Calculate the coefficient of correlation and obtain the equati the lines of regression in the simplest form. (P.U., B.A./B.Sc.

10.28 If X_1, X_2 and X_3 are uncorrelated variables, each having the same standard deviation, obta co-efficient of correlation between (X_1+X_2) and (X_2+X_3). (P.U., B.A. Hons. Part-II,

10.29 a) What is rank correlation? Derive Spearman's co-efficient of rank correlation.

(P.U., B.A./B.Sc. 1960, 71, 82,

b) The ranks of the same 16 students in Mathematics and Physics were as follows:

(1, 1); (2, 10); (3, 3); (4, 4); (5, 5); (6, 7); (7, 2); (8, 6), (9, 8); (10, 11); (11, 15); (12, 9 14); (14, 12); (15, 16); (16, 13); the two numbers within brackets denoting the ranks same student in Maths, and Physics respectively. Calculate the rank correlation co-eff for proficiencies of this group in two subjects. (P.U., B.A./B.Sc.

30 a) If n pairs of values of two variables a and b are given, where each variable is ranked in order (1 to n), show that the co-efficient of correlation between ranks is given by

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where d is the difference between the ranks of a and b. (P.U., B.A./B.Sc. 1989)

b) Obtain the product moment coefficient of correlation between the following values:

a	7.4	9.0	11.0	2.5	4.6	6.5
b	8.5	6.1	2.4	6.7	12.6	3.3

Rank the values and hence find a rank correlation coefficient between the two sets.

31 a) Describe circumstances in which you would use: (i) rank correlation co-efficient; (ii) product moment correlation coefficient.

b) The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both laboratory and lecture portions of a statistics course. Find the co-efficient of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

(P.U., B.A./B.Sc. 1969)

32 Ten competitors in a beauty contest are ranked by three judges in the following order.

First Judge	1	6	5	10	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6	9
Third Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation co-efficient to discuss which pair of judges have the nearest approach to common tastes in beauty. (P.U., B.A./B.Sc., 1960, B.Sc. (Hons.) Part-I, 1971)

33 In a painting competition, various entries are ranked by three judges. Use Spearman's rank correlation co-efficient to discuss which pair of judges has the nearest approach to common tastes.

Entry	A	B	C	D	E	F	G	H	K	L
Judge X	5	2	6	8	1	7	4	9	3	10
Judge Y	1	7	6	10	4	5	3	8	2	9
Judge Z	6	4	9	8	1	2	3	10	5	7

(P.U., D.St., 1964)

34 a) What are tied ranks? Explain how you would find the co-efficient of rank correlation for tied ranks.

b) Compute the co-efficient of rank correlation for the following ranks:

X	8	3	6.5	3	6.5	9	3	1	5
Y	8	9	6.5	2.5	4	5	6.5	1	2.5

10.35 Establish the formula for the 'co-efficient of concordance'. Find the same for the following data:

X	1	2	3	4	5	6	7	8	9	10
Y	7	10	4	1	6	8	9	5	2	3
Z	9	6	10	3	5	4	7	8	2	1

♦♦♦♦♦♦♦♦♦♦♦