Chapter - 10

CORRELATION AND REGRESSION

Study Strategy and Learning Objectives

Please Remember The Following Study Strategy and Learning Objectives:

Study Strategy:

1. First, read this section with the limited objective of simply trying to understand the following important key tenns and concepts: *correlation and regression, linear correlation coefficient, partial and multiple correlation, residuals, least squares estimation, scatter diagram, dependent and independent variable, simple regression and multiple regression.*
2. Second, try to understand what they accomplish, and why they are needed; and develop the ability to calculate or select them.
3. Third, learn how to interpret them.
4. Fourth, read the section once again and try to understand the underlying theory.

You will always enjoy much greater success if you understand what you are doing, instead of blindly applying mechanical steps in order to obtain an answer that may or may not make any sense.

Learning Objectives:

After careful study of this chapter, you should be able to do the following:

1. Calculate correlation coefficient and apply the correlation model.
2. Use simple linear correlation and regression for building empirical models to engineering and scientific data.
3. Understand how the method of least squares is used to estimate the parameters in a linear regression model.
4. Understand how the method of least squares extends to fitting multiple regression models.
5. Analyze residuals to determine if the regression model is an adequate fit to the data or to see if any underlying assumptions are violated.
6. Test statistical hypotheses and construct confidence intervals on regression model parameters.
7. Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval on the future observation.
8. Use simple transformations to achieve a linear regression model.

10.1 Introduction

This chapter introduces important methods for making inferences based on sample data that come in pairs. ***This chapter has the objective of determining whether there is a relationship between the two variables and, if such a relationship exists, we want to describe it with an equation that can be used for predictions.***

We begin in Section 10-2 by considering the concept of correlation, which is used to determine whether there is a statistically significant relationship between two variables. We investigate correlation using the scatterplot (a graph) and the linear correlation coefficient (a measure of the direction and strength of linear association between two variables).

In Section 10-3 we investigate regression analysis; we describe the relationship between two variables with an equation that relates them and show how to use that equation to predict values of one variable when we know values of the other variable. Then we analyze the differences between predicted values and actual observed values of a variable; use concepts of multiple regression to describe the relationship among three or more variables. Finally, we describe some basic methods for developing a mathematical model that can be used to describe the relationship between two variables.

10.2 Correlation

Under the circumstance where one variable is expected to be related to the other, the experimenter is always interested in knowing the degree of relationship between those variable with reference to the measured observations. In statistics, such a relationship is known as ***correlation***. Therefore, the correlation is defined as the degree of relationship between two or more random variables. If there are only two variables, then the correlation between them is known as ***simple correlation*** and if there are more than two variables then we have the case of ***multiple correlation***.

Main Objective: The ***main objective of this section*** is to analyze a collection of paired sample data (sometimes called bivariate data) and determine whether there appears to be a relationship between the two variables. In statistics, we refer to such a relationship as a correlation. (We will consider only linear relationships, which means that when graphed, the points approximate a straight line pattern. Also, we consider only quantitative data.)

Definition:

A ***correlation*** exists between two variables when one of them is related to the other in some way.

The problems like, the study of the relationship between input and output of a wastewater treatment plant, the relationship between the tensile strength and hardness of aluminum, or the relationship between impurities in the air and the incidence of a certain disease, etc. are referred to as problems of correlation analysis, where it is assumed that the data points (x_i, y_i) for $i = 1,2,3,\dots,n$ are values of a pair of random variables whose joint density is given by $f(x,y)$.

Assumptions:

1. The sample of paired (x, y) data is a *random sample* of quantitative data.
2. The pairs of (x, y) data have a bivariate normal distribution. (this assumption basically requires that for any fixed value of x, the corresponding values of y have a distribution that is bell-shaped, and for any fixed value of y, the values of x have a distribution that is bell-shaped.) This assumption is usually difficult to check, but a partial check can be made by determining whether the values of both x and y have distributions that are basically bell-shaped.

Notation for the Linear Correlation Coefficient

n represents the number of pairs of data present.

Σ denotes the addition of the items indicated.

Σx denotes the sum of all x-values.

Σx^2 indicates that each x-value should be squared and then those squares added.

$(\Sigma x)^2$ indicates that the x-values should be added and the total then squared.
It extremely important to avoid confusing Σx and $(\Sigma x)^2$.

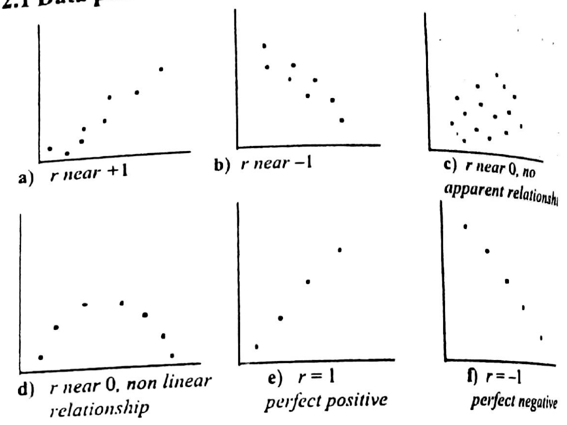
Σxy indicates that each x-value should first be multiplied by corresponding y-value. After obtaining all such products, find their sum.

r represents the linear correlation coefficient for a sample.

ρ represents the linear correlation coefficient for a population.

Remark: Sample correlation coefficient r is not an unbiased estimator of ρ, it is widely used as a point estimator whatever the form of the bivariate population.

10.2.1 Data plot for different values of r



a) *r near* +1

b) *r near* −1

c) *r near* 0, *no apparent relationshi*

d) *r near* 0, *non linear relationship*

e) $r = 1$ *perfect positive*

f) $r = -1$ *perfect negative*

10.2.2 Linear Correlation Coefficient

Because visual examinations of scatter plots are largely subjective, we more precise and objective measures. We use the linear correlation coeffici which is useful for detecting straight-line patterns.

Definition: The ***linear correlation coefficient*** r (i.e., Karl Pearson's Coeff Coefficient) measures the strength of the linear relationship between the pair and y-quantitative values in a sample. The sample correlation coefficient r for pairs $(x_1, y_1), \dots, (x_n, y_n)$ is calculated by using any one of the following formul

$$1.\quad r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \text{ where, } S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n},$$

$$S_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n},\quad S_{xy} = \Sigma(x-\bar{x}) = \Sigma xy - \frac{\Sigma x \Sigma y}{n};$$

$$2.\quad r = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}};$$

$$3.\quad r = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}};$$

$$4.\quad r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}, \text{ where } u = \frac{x-a}{h}, y = \frac{y-b}{k}.$$

[The linear correlation coefficient is sometimes referred to as the ***Pearson pr moment correlation coefficient*** in honor of Karl Pearson (1857–1936), originally developed it.]

Because the linear correlation coefficient r is calculated using sample data, sample statistic used to measure the strength of the linear correlation between. y. If we had every pair of population values for x and y, the result of (i) would population parameter, represented by ρ (*Greek rho*).

10.2.3 Properties of linear Correlation coefficient r

1. The value of r does not depend on which of the two variables under study is labeled x and which is labeled y. So the value of r is not affected by the choice of x or y.
2. The value of r is independent of the units in which x and y are measured.
3. The value of r does not change if all values of either variable are converted to a different scale.
4. The value of r is always between -1 and $+1$ inclusive. That is, $-1 \le r \le 1$.

Proof: Let us consider the sum of the squares

$$\sum\left\{\frac{x-\bar{x}}{S_x} \pm \frac{y-\bar{y}}{S_y}\right\}^2 \quad \text{where the symbols have their usual meanings}$$

Since, $S_x^2 = \frac{1}{(n-1)}\Sigma(x-\bar{x})^2 \qquad \therefore \Sigma(x-\bar{x})^2 = (n-1)\,S_x^2$

$$S_y^2 = \frac{1}{(n-1)}\Sigma(y-\bar{y})^2 \qquad \therefore \Sigma(y-\bar{y})^2 = (n-1)\,S_y^2$$

$$r = \frac{1}{n-1}\,\frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{S_x\,S_y} \qquad \therefore \Sigma(x-\bar{x})\,(y-\bar{y}) = (n-1)\,r\,S_x\,S_y$$

Now, $\sum\left\{\frac{x-\bar{x}}{S_x} \pm \frac{y-\bar{y}}{S_y}\right\}^2 = \frac{\Sigma(x-\bar{x})^2}{S_x^2} \pm \frac{2\Sigma(x-\bar{x})\,(y-\bar{y})}{S_x\,S_y} + \frac{\Sigma(y-\bar{y})^2}{S_x^2}$

$$= \frac{(n-1)S_x^2}{S_x^2} \pm \frac{2(n-1)r\,S_x\,S_y}{S_x\,S_y} + \frac{(n-1)S_y^2}{S_x^2}$$

$$=2\,(n-1) \pm 2\,(n-1)\,r \qquad = 2\,(n-1)\,(1 \pm r)$$

Since the expression on the left hand side is the sum of the squares of real quantities, so it is always non-negative.

$$\therefore 1 \pm r \ge 0 \qquad [\because 2(n-1) > 0]$$

Taking positive sign we get

$$1 + r \ge 0 \Rightarrow r \ge -1 \qquad \text{- - - } (i)$$

Taking negative sign, we get

$$1 - r \ge 0 \Rightarrow r \le 1 \qquad \text{- - - } (ii)$$

Hence combing (*i*) and (*ii*) we get

$-1 \le r \le 1$. Proved.

Alternative Method:

The regression equation of y on x is $\quad y-\bar{y} = r\frac{\sigma_x}{\sigma_y}(x-\bar{x})$

or, $\sigma_x\,(y-\bar{y}) = r\,\sigma_y(x-\bar{x})$

quaring both sides

$$\sigma_x^2\,(y-\bar{y})^2 = r^2\sigma_y^2\,(x-\bar{x})^2$$

Taking Expectation on both side

or, $\sigma_x^2\,E(y-\bar{y})^2 = r^2\sigma_y^2\,E(x-\bar{x})^2$

or, $\sigma_x^2\sigma_y^2 = r^2\sigma_y^2\,\sigma_x^2 \quad [\therefore E(y-\bar{y})^2 = \sigma_y^2,\ E(x-\bar{x})^2 = \sigma_x^2]$

$\therefore r^2 = 1$

$|r| = 1 \quad \Rightarrow -1 \le r \le 1$ proved.

5. The magnitude of r describes the strength of a linear relationship and its sign indicates the direction.
   - $r = +1$ if all pairs (x_i, y_i) lie exactly on a straight line with positive slope; and correlation is perfect positive. Here as x increases, y increases linearly and steadily.
   - $r = -1$ if all pairs (x_i, y_i) lie exactly on a straight line with negative slope; and correlation is perfect negative. Here as x increases, y decreases linearly and steadily.
   - $r > 0$ if the pattern in the scatter plot runs from lower left to upper right.
   - $r < 0$ if the pattern in the scatter plot runs from upper left to lower right.
   - $r = 0$ if there is no correlation.

   It is not designed to measure the strength of a relationship that is not linear.
6. A value of r close to zero implies that the linear association is weak.
7. Correlation coefficient is the geometric mean between two regression coefficients; i.e., $r = \sqrt{b_{yx}.\ b_{xy}}$

10.2.4 Interpreting the Linear Correlation Coefficient

The value of r must always fall between -1 and $+1$ inclusive. If r is close to 0, we conclude that there is ***no significant linear correlation*** between x and y, but if r is close to -1 or $+1$ we conclude that there is a ***significant linear correlation*** between x and y.

10.2.5 Coefficient of Determination (Explained Variation)

If we conclude that there is a significant linear correlation between x and y, we can find a linear equation that expresses y in terms of x, and that equation can be used predict values of y for given values of x. In previous section we have described a procedure for finding such equations and shown how to predict values of y when given values of x. But a predicted value of y will not necessarily be the exact result, because in addition to x, there are other factors affecting y, such as random variation and other characteristics not included in the study.

Definition: It may be noted that whenever correlation coefficient r (ρ for population) is obtained the coefficient of determination can be obtained as r^2 (ρ^2 for population) which gives how far the changes in one variable is explained by the other variable. For example if $r = 0.6$ than $r^2 = 0.36$ which means that the 36% of the changes in one variable is explained by the other variable.

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Example 1: (*Calculation of the sample correlation coefficient*): The following are the number of minute it took 10 mechanics to assemble a piece of machinery in the morning, x, and in the late afternoon, y:

$X=x$	11.1	10.3	12.0	15.1	13.7	18.5	17.3	14.2	14.8	15.3
$Y=y$	10.9	14.2	13.8	21.5	13.2	21.1	16.4	19.3	17.4	19.0

Calculate r

Solution: Using calculator we get

$$\Sigma x = 142.3,\ \Sigma y = 166.8,\ \Sigma x^2 = 2{,}085.31$$

$$\Sigma xy = 2{,}434.69,\ \Sigma y^2 = 2{,}897.80$$

So, $S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 2{,}085.31 - (142.3)^2/10 = 60.381$

$$S_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 61.126 = 61.126$$

$$S_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 115.576$$

Hence, $r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \dfrac{61.126}{\sqrt{(60.381)(115.576)}} = 0.732$

The positive value of r confirms a positive association where long assembly times tend to pair together and so do short assembly times.

Example 2: ***Calculate Karl Pearson's coefficient of correlation*** for the following data and interpret the result.

x	1	2	3	4	5	6	7
y	6	7	5	4	3	1	2

Solution: Calculation of correlation coefficient

x	Y	x^2	y^2	Xy
1	6	1	36	6
2	7	5	49	14
3	5	9	25	15
4	4	16	16	16
5	3	25	9	15
6	1	36	1	6
7	2	49	4	14

$\Sigma x = 28$. $\Sigma x = 28$, $\Sigma x^2 = 140$, $\Sigma y^2 = 140$, $\Sigma xy = 86$

n = number of observations = 7

$$\therefore \quad r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{7 \times 86 - 28 \times 28}{\sqrt{7 \times 140 - (28)^2}\sqrt{7 \times 140 - (28)+2}} = -0.923$$

The value of r indicates that there is highly negative correlation between the variables x and y.

Example 3: Find the ***coefficient of correlation*** between industrial production and export using the following data:

Production($X=x$)	55	56	58	59	60	60	62
Exports ($Y=y$)	35	38	38	39	44	43	44

Solution. Let, $u = x - 60$, $v = y - 38$. The calculations are shown in the following table.

x	Y	U	v	uv	u^2	v^2
55	35	–5	–3	15	25	9
56	38	–4	0	0	16	0
58	38	–2	0	0	4	0
59	39	–1	1	–1	1	1
60	44	0	6	0	0	36
60	43	0	5	0	0	25
62	44	2	6	12	4	36
Total		$-10 = \Sigma u$	$15 = \Sigma v$	$26 = \Sigma uv$	$50 = \Sigma u^2$	$107 = \Sigma v^2$

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}} = \frac{7 \times 26 + 10 \times 15}{\sqrt{7 \times 50 - (-10)^2}\sqrt{7 \times 107 - (15)^2}} = +0.92$$

Example 4: A computer while calculating the correlation coefficient betwee variates x and y from 25 pairs of observations obtained the following constants:

$$n = 25, \quad \Sigma x = 125, \quad \Sigma x^2 = 650,$$
$$\Sigma y = 100, \quad \Sigma y^2 = 460, \quad \Sigma xy = 508.$$

It was, however, later discovered at the time of checking that he had copie down two pairs as $\begin{array}{c|c} x & y \\ 6 & 14 \\ 8 & 6 \end{array}$ while the correct values were $\begin{array}{c|c} x & y \\ 8 & 12 \\ 6 & 8 \end{array}$. Obtain the correct value of the correlation coefficient.

Solution:

Incorrect

x	y	xy
6	14	84
8	6	48
14	20	132

Correct

x	y	xy
8	12	96
6	8	48
14	20	144

It is clear from the above tables that there is no change in the values of Σx and Σ

Corrected $\Sigma x = 125$

" $\Sigma y = 100$

" $\Sigma x^2 = 650 - 6^2 - 8^2 + 6^2 + 8^2 = 650$

" $\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

" $\Sigma xy = 508 - 132 + 144 = 520$

$$\text{Corrected } r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}} = \frac{2}{3}$$

10.2.6 Partial and Multiple Correlation

A large number of factors generally affect social and natural phenon simultaneously. The effect of these factors on one another is studied thrc correlation and regression studies. In simple correlation between two variables assumptions was that, the *effect of other factors on the phenomenon under stu ignored*. For example, when we study correlation between price and demand ther assume that other independent factors like, money exports, imports that also at price, have been completely ignored. In a study of simple correlation between p and demand we assume that not only other independent variables like money, su etc. have been ignored but also that *various other variables affecting price mutually independent of each other*. Really, there is an association between diffe independent variables affecting dependent variables.

Thus, our study of simple correlation and regression makes such assumptio which are not true, and to this extent, the relationship studied is not absolu dependable. It is necessary to study the effects of all these factors partial multiple correlations and regression analysis is done to obtain this objective.

10.2.7 Partial correlation

In multivariate study, the correlation between any two variables is not from the influence of other variables. For example, the yield of crop per acre is only depend upon quality of seed but also other so many factors like fertility of s fertilizer used, irrigation and so on.

Hence, it becomes necessary to eliminate the common association of other variable to obtain the actual association between the two variables. Broadly speaking the partial correlation is the simple correlation between two variates when the influence of other variates has been eliminated.

<u>Definition</u>: ***Partial correlation coefficient*** may be defined as a measure of degree of association between any two variables out of a set or variables eliminating the common association of remaining variables with both of them.

Let us consider three variables X_1, X_2 and X_3. Then, we have,

$r_{12.3}$ = partial correlation coefficient between X_1 and X_2 keeping X_3 constant (or by eliminating the effect of X_3)

$$= \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$r_{13.2}$ = partial correlation coeff.cient between X_1 and X_3 keeping X_2 constant (or by eliminating the effect of X_2)

$$= \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

$r_{23.1}$ = partial correlation coefficient between X_2 and X_3 keeping X_1 constant (or by eliminating the effect of X_1)

$$= \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

10.2.8 Properties of Partial Correlation Coefficient

1. Its value lies between – 1 to + 1

   i.e., $-1 \le r_{12.3} \le +1$, $-1 \le r_{13.2} \le +1$, $-1 \le r_{23.1} \le +1$

2. The position of subscript on left side of dots does not make any difference in the meaning.

   i.e., $r_{12.3} = r_{21.3}$, $r_{13.2} = r_{31.2}$ and $r_{23.1} = r_{32.1}$

<u>Coefficient of Partial Determination</u>: The square of partial correlation coefficient is known as the *coefficient of partial determination* i.e., $r_{12.3}^2$, $r_{13.2}^2$ and $r_{23.1}^2$. It is used to interpret the value of partial correlation coefficient.

For example, If $r_{23.1} = 0.9$, then $r_{23.1}^2 = 0.81$. This implies that 81 percent of the total variation in dependent variable is explained by the independent variable and is not associated with X_1.

Example 5: If $r_{12} = 0.6$, $r_{13} = 0.4$ and $r_{23} = 0.35$ find the value of

(i) Partial correlation coefficient between X_1 and X_2 keeping X_3 constant.

(ii) Partial correlation coefficient between X_2 and X_3 keeping X_1 constant.

(iii) Partial correlation coefficient between X_1 and X_3 keeping X_2 constant.

Solution: Here, we have given, $r_{12} = 0.6$, $r_{13} = 0.4$, $r_{23} = 0.35$

Now,

(i) Partial correlation coefficient between X_1 and X_2 keeping X_3 constant is given by

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.6 - 0.4 \times 0.35}{\sqrt{1 - 0.4^2}\sqrt{1 - 0.35^2}} = 0.536$$

(ii) Partial correlation coefficient between X_2 and X_3 keeping X_1 constant is given by

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}} = \frac{0.35 - 0.6 \times 0.4}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.4^2}} = 0.15.$$

(iii) Partial correlation coefficient between X_1 and X_3 keeping X_2 constant is given by

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = \frac{0.4 - 0.6 \times 0.35}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.35^2}} = 0.254$$

Example 6: The correlation between a general intelligence test and school achievement in a group of children from 6 to 15 years old is 0.86. The correlation between the general intelligence test and age in the same group is 0.65 and the correlation between school achievement and age is 0.72. What is the correlation between general intelligence and school achievement in children of the same age? Comment on the result.

Solution: Let X_1 = general intelligence test
X_2 = school achievement
X_3 = age.

Here, $r_{12} = 0.86$, $r_{13} = 0.65$ and $r_{23} = 0.72$, we need to find $r_{12.3}$.

$$\text{Now, } r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.86 - 0.65 \times 0.72}{\sqrt{1 - (0.65)^2}\sqrt{1 - (0.72)^2}} = 0.743$$

To interpret the value of $r_{12.3}$ we need to find the coefficient of partial determination. Thus, $r_{12.3}^2 = (0.743)^2 = 0.552$. This means that 55.2% of the total variation in the value of the dependent variable X_1 (general intelligence test) has been explained by the impendent variable X_2 (school achievement) where X_3 (age) is held to be constant (same).

10.2.9 Multiple Correlations

Whenever we are interested in studying the relationship between the joint effects of a group of variables upon a variable not included in that group, our study is that of multiple correlation. Therefore, multiple correlation aims at knowing how far the dependent variable is influenced by the independent variables.

<u>Definition</u>: *Multiple correlation* studies the relationship between dependent variable and joint (or combined) effects of independent variables. For example, multiple correlation gives the relationship of dependent variable (yield of paddy) and joint effect of independent variables (plot of land, labor, seed, fertilizer, pesticide, irrigation and so on.)

Let us consider three variable X_1, X_2 and X_3. Then we have,

$R_{1.23}$ = multiple correlation coefficient between dependent variable X_1 and joint effect of independent variables X_2 and X_3 on X_1.

$$= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

$R_{2.13}$ = multiple correlation coefficient between dependent variable X_2 and joint effect of independent variables X_1 and X_3 on X_2.

$$= \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{13}^2}}$$

$R_{3.12}$ = multiple correlation coefficient between dependent variable X_3 and joint effect of independent variables X_1 and X_2 on X_3

$$= \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{12}^2}}$$

10.2.10 Properties of Multiple Correlation Coefficients

1. Its value lies between 0 to +1.

   i.e., $0 \le R_{1.23} \le +1$; $0 \le R_{2.13} \le +1$; $0 \le R_{3.12} \le +1$.

2. The position of subscript on right side of dots does not make any difference in the meaning, i.e., $R_{1.23} = R_{1.32}$, $R_{2.13} = R_{2.31}$; $R_{3.12} = R_{3.21}$
3. If $R_{1.23} = 0$, then $r_{12} = 0$ and $r_{13} = 0$
4. The value of multiple correlation coefficient is not less than simple correlation coefficient, i.e., $R_{1.23} \geq r_{12}$, $R_{1.23} \geq r_{13}$, $R_{1.23} \geq r_{23}$

Definition:(*Coefficient of Multiple Determination*)

The square of multiple correlation coefficient is known as the ***coefficient of multiple determination***; i.e., $R^2_{1.23}$, $R^2_{2.13}$ and $R^2_{3.12}$. It is used to interpret the value of multiple correlation coefficient.

For example, if $R_{1.23} = 0.8$ then $R^2_{1.23} = 0.64$. This implies that 64 percent of the total variation in dependent variable (X_1) is explained by the independent variables $(X_2$ and $X_3)$.

Example-7: It is possible to get the following information from an experimental data $r_{12} = 0.5$, $r_{13} = -0.75$ and $r_{23} = 0.92$

Solution: We have, $R_{1.23} = \sqrt{\dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}}$

$$= \sqrt{\frac{0.5^2 + (-0.75)^2 - 2 \times 0.5 \times (-0.75) \times 0.92}{1 - 0.92^2}} = 3.13$$

Since $R_{1.23} > 1$, so it is not possible to get such information from an experimental data.

Example-8: (i) From the data given below find $r_{12.3}$, $R_{1.23}$, $r_{23.1}$ and $R_{2.13}$.

$$\Sigma x_1x_2 = 40,\ \Sigma x_1x_3 = 55,\ \Sigma x_2x_3 = 35$$

$$\Sigma x_2^{\,2} = 60,\ \Sigma x_3^{\,2} = 50,\ \Sigma x_1^{\,2} = 90,\ n = 6$$

where, X_1, X_2 and X_3 are variables measured from their means

Solution: (i) We have $r_{12} = \dfrac{\Sigma x_1x_2}{\sqrt{\Sigma x_1^2}\,\sqrt{\Sigma x_2^2}} = \dfrac{40}{\sqrt{90}\,\sqrt{60}} = 0.54$

$$r_{13} = \frac{\Sigma x_1x_3}{\sqrt{\Sigma x_1^2}\,\sqrt{\Sigma x_2^2}} = \frac{55}{\sqrt{90}\,\sqrt{50}} = 0.82;\quad r_{23} = \frac{\Sigma x_2x_3}{\sqrt{\Sigma x_2^2}\,\sqrt{\Sigma x_3^2}} = \frac{35}{\sqrt{60}\,\sqrt{50}} = 0.64$$

$$\text{Now, } r_{12.3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}} = \frac{0.54 - 0.82 \times 0.64}{\sqrt{1 - 0.82^2}\,\sqrt{1 - 0.64^2}} = 0.035$$

$$r_{23.1} = \frac{r_{23} - r_{12}\,r_{13}}{\sqrt{1 - r_{12}^2}\,\sqrt{1 - r_{13}^2}} = \frac{0.64 - 0.54 \times 0.82}{\sqrt{1 - 0.54^2}\,\sqrt{1 - 0.82^2}} = 0.41$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2\,r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.54^2 + 0.82^2 - 2 \times 0.54 \times 0.82 \times 0.64}{1 - 0.64^2}} = 0.82$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2\,r_{12}\,r_{13}\,r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{0.54^2 + 0.64^2 - 2 \times 0.54 \times 0.82 \times 0.64}{1 - 0.8^2}} = 0.641$$

Example-9: You are given $r_{12} = 0.93$, $r_{13} = 0.50$ and $r_{23} = 0.34$. Assuming the first variable as dependent, compute the coefficient of multiple determination. Also interpret the result.

Solution: Here, $r_{13} = 0.50$, $r_{23} = 0.34$, determination is given by

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}$$

$r_{12} = 0.93$. Since first variable is dependent, we should compute the coefficient of multiple determination $R_{1.23}^2$.

The coefficient of multiple

$$= \frac{(0.93)^2 + (0.50)^2 - 2 \times 0.93 \times 0.50 \times 0.34}{1 - (0.34)^2} = 0.903 \text{ (approx.)}$$

Since $R_{1.23}^2 = 0.903$, it shows that 90.3% of the total variation in the dependent variable X_1 has been explained by the two independent variables

10.3 Regression

Many problems in engineering and science involve exploring the relationships between two or more variables. Very often, the interest lies in establishing the actual relationship between two or more variables. This problem is dealt with ***regression analysis***. Regression analysis is a statistical technique that is very useful for these types of problems. For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes.

On the other hand, we are often not interested to know the actual relationship but are only interested in knowing the degree of relationship between two or more variables. This problem is dealt with ***correlation analysis***.

*Thus, **regression analysis shows how the variables are related while the correlation analysis measures the degree of relationship between the variables. Regression and correlation analysis thus determine the nature and strength of relationship between variables.***

Main Objective of this section: In Section correlation, we analyze paired data with the goal of determining whether there is a linear correlation between two variables. The main objective of this section is to describe the relationship between two variables by finding the graph and equation of the straight line that represents the relationship. This straight line is called the ***regression line***, and its equation is called the ***regression equation***.

The literal meaning of the word *Regression* is *stepping back* or *returning to the average value*. This term *regression* was first used by **British biometrician Sir Francis Galton (1822–1911)** on estimating the nature of relationship between the height of fathers and sons. He found in his study that

(i) The tall fathers have tall sons and short fathers have short sons.

(ii) The average height of sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is more than that of the fathers.

Galton termed the line describing the average relationship between two variables, as the line of regression. He used the word regression as the name of the general process of predicting one variable (*the height of the sons*) from another variable (*the height of the fathers*). We continue to use Galton's "regression" terminology, even though our data do not involve the same height phenomena studied by Galton.

Now days, it is one of the very important statistical tools which is extensively used in almost all sciences: natural, social and physical.

Prediction or estimation is one of the major problems in almost all spheres of human activity. The pharmaceutical concerns are interested in studying or estimating the effect of new drugs on patients. An engineer may wish to predict the amount of oxide that will form on the surface of metal baked in an oven for one hour at 200 degrees of Celsius, or in chemical process the amount of output for various values of temperature and amount of catalyst employed or the amount of deformation of a ring subject to a compressive force of 1,000 pounds etc. Regression analysis is one of the very scientific techniques for making such prediction on the basis of mathematical equations.

10.3.1 Simple regression

The regression analysis confined to the study of only two variables at a time is termed as *simple regression*. The regression analysis confined to the study of more than two variables at a time is termed as *multiple regression*.

In simple regression analysis, there are two types of variables. The variable which influences the values or is used for prediction or whose value is fixed by the experimenter is called *independent* or *regressor* or *predictor* or *explanatory* or *controlled variable*. The other variable whose value is influenced or is to be predicted is called *dependent* or *response variable*. It is random variable. If we denote this variable by Y and its observed value by y, and dependent variable by x, then we refer this relationship as the *regression of Y on x*.

Therefore, the regression equation expresses a relationship between x (called the *independent variable*, or *predictor variable*, or *explanatory variable*) and y (called the *dependent variable*, or *response variable*). The typical equation of a straight line $y = mx + b$ is expressed in the form $\hat{y} = a + b\,x$, where a is the y-intercept and b is the slope. The given notation shows that a and b are sample statistics used to estimate the population parameters β_0 and β_1. We will use paired sample data to estimate the regression equation. Using only sample data, we can't find the exact values of the population parameters β_0 and β_1, but we can use the sample data to estimate them with a and b.

Examples: the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x, the regression of the gain of weight Y of certain animals on the daily ration of food x, the regression of the heat conductivity Y of cork on the specific weight x of the cork etc.

Assumptions:

1. We are investigating only linear relationships.
2. For each x-value, y is a random variable having a normal (bell-shaped) distribution. All of these y distributions have the same variance. Also, for a given value of x, the distribution of y-values has a mean that lies on the regression line. (Results are not seriously affected if departures from normal distributions and equal variances are not too extreme.)

Definitions:

Given a collection of paired sample data, the *regression equation* $\hat{y} = a + bx$ algebraically describes the relationship between the two variables. The graph of the regression equation is called the *regression line* (or *line of best fit*, or *least-squares line*).

Notation for Regression Equation:

	Population Parameter	*Sample Statistic*
y-intercept of regression equation	β_0	a
Slope of regression equation	β_1	b
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = a + bx$

Finding the slope b and y-intercept a in the regression equation $y = a + bx$

Formula *Slope:* $b = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})}$

Formula *y-intercept:* $a = \bar{y} - b\bar{x}$

10.3.2 Regression lines

Linear relationship between two variables is represented by a straight line which is known as *regression line* or *the line of average relationship*. This line is the best estimate of one variable for any given value of the other variable. In case of two variables x and y we have two regression lines: y on x and x on y.

Definition: ***Line of regression*** of y on x is the line which gives the best estimate for the values of y for any specified value of x. It is given by $\hat{y} = a + bx$.

Similarly *line of regression* of x on y is the line which gives the best estimate for the values of x for any specified value of y. It is given by $\hat{x} = c + dy$.

10.3.3 Scatter plot or Scatter diagram

If a regression line is to be specified, we plot n paired observations on the graph paper, setting the vertical scale for the dependent variable y and horizontal scale for the independent variable x. From the plotted points, it can be visualized whether or not the plotted points lie in a straight line or appear to lie on a curve of a known type. Scatter diagram may be considered as a basis of the relationship between two variables. In case, the plotted points are scattered in a haphazard manner, i.e., no pattern is observed, then it can be inferred that the variables are not related. Now we discuss the scatter diagram with special reference to the regression line.

Let $x_1, x_2, \ldots, x_n$ denote the values of the independent variable and $y_1, y_2, \ldots, y_n$ denote the values of dependent variable associated with x_i. The available data then consists of n pairs (x_1, y_1) (x_2, y_2), ... (x_n, y_n). A first step in regression analysis involving two variables is to construct a scatter plot of the observed data as points

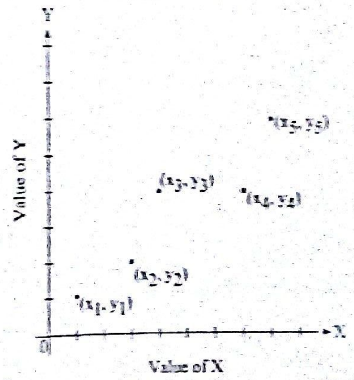
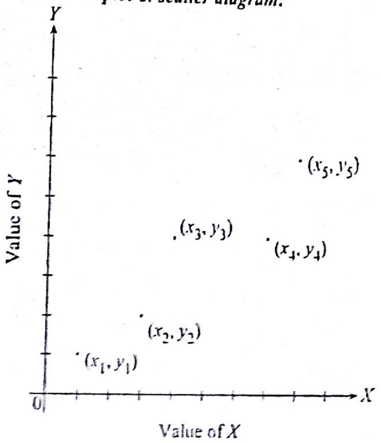


Fig: Scatter plot

We see that, there is no any simple curve which will pass through all the points. But we observe that y increase as x increase. So, this gives that there is some relationship between x and y. This kind of diagram, which shows how the data points are scattered, is called a ***scatter plot*** or ***scatter diagram.***



Definition: A *scatter plot* (or *scatter diagram*) is a graph in which the paired (x, y) sample data are plotted with a horizontal x-axis and a vertical y-axis. Each individual (x, y) pair is plotted as a single point.

In practice, when we plot various paired observations, one would hardly come across a situation in which all the points are lying exactly in a line. But if a line is drawn suitably, some points will be lying on the line and others will be lying in the close vicinity of this line as in above figure. It is observed that a few points are lying away from all the fitted lines. These extreme points may sometimes be considered outliers. As evident from the figure, a number of lines may be drawn and considered. But the ***best line will be one for which the algebraic sum of the perpendicular distances of all the points from the line is zero.*** Hence, dotted lines in the figure are not such a good fit as the smooth straight line. In short, the scatter diagram can be thought of as a device to know how closely the two variables are related and in what form i.e. linear or curvilinear.

Once it is decided on the basis of prior information or scatter diagram that the two variables are linearly related, the problem arises on deciding which of the many possible lines the best fitted line is. To cope with this problem, mathematical basis leads to the most logical and accurate solution. ***The least square method*** is most widely accepted method of fitting a straight line.

The method of Least Squares: The procedure of finding the equation of the line which best fits a given set of paired data is called the ***method of least squares.***

Least Squares Principle: "The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line (i.e., the residuals or the errors of estimates) is minimum."

Minimizing the sum of the squares of errors parallel to y–axis gives the equation of regression line of y on x and minimizing the errors parallel to x–axis gives the equation of regression line of x on y.

10.3.4 Determination of regression line by least squares method

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of observations on the two variables x and y. Let $y = a + bx + e$ . . . (i)

be the line of regression of y on x where a and b are constant are e_n the error in predicating of y corresponding to given x_i is $e_i = y_i - \hat{y}_i$

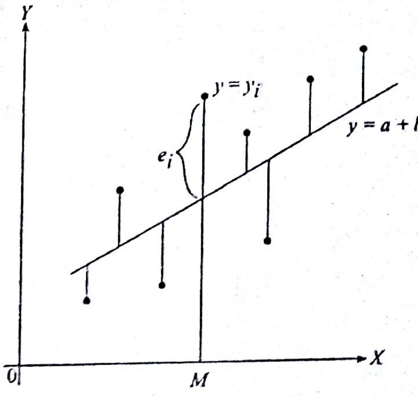


Diagram for least Squares

Sum of the squares of errors (i.e., distances of points from line) is

$e = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

According to the principle of least squares, we have to determine constants a and b such that e is minimum.

From calculus, we know that a necessary condition for minima (or maxima) is

$$\frac{\partial e}{\partial a} = 0 \text{ and } \frac{\partial e}{\partial b} = 0$$

$$\Rightarrow \Sigma 2(y - a - bx)\frac{\partial e}{\partial a}(y - a - bx) = 0 \text{ and } \Sigma 2(y - a - bx)\frac{\partial e}{\partial b}(y - a - bx) = 0$$

$$\Rightarrow \Sigma 2(y - a - bx)(-1) = 0 \text{ and } \Sigma 2(y - a - bx)(-x) = 0$$

$$\Rightarrow \Sigma y - na - b\Sigma x = 0 \text{ and } \Sigma xy - a\Sigma x - b\Sigma x^2 = 0$$

$$\Rightarrow \quad \begin{array}{l} \Sigma y = na + b\Sigma x \\ \Sigma xy = a\Sigma x + b\Sigma x^2 \end{array}$$

These equations are known as the ***normal equations*** for the least square estimators (i.e., for estimating a and b). Solving these two normal equations we get

$a = \bar{y} - b\bar{x}$

$$b = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})} = \frac{S_{xy}}{S_{xx}}$$

Putting these values in (i) we get required fitted (or estimated) regression line of y on x as

$\hat{y} = a + bx$, where $\hat{y}$ is estimated value of y for given x.

Note: (1) Least squares estimates $a = \bar{y} - b\bar{x}$ and $b = \dfrac{S_{xy}}{S_{xx}}$

where $\bar{x} = \dfrac{\Sigma x}{n}$, $\bar{y} = \dfrac{\Sigma y}{n}$; $\quad S_{xx} = \Sigma(x - \bar{x})^2 = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$

$$S_{yy} = \Sigma(y - \bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}; \quad S_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

and $\Sigma y = na + b\Sigma yx \quad \Rightarrow \dfrac{1}{n}\Sigma y = a + b\dfrac{\Sigma x}{n} \quad \Rightarrow \bar{y} = a + b\bar{x}$

(2) The numerical constant b which is slope of line of regression of y on x is called *regression coefficient of y on x*. It is denoted by b_{yx}. It represents the rate of change of y w.r. to x.

10.3.5 Regression equations

(i) Regression equation of x on y

A line of regression of y on x is the line which give the best estimate for the value of y for given value of x. The regression equation of y on x is $\hat{y} = a + bx$, where $\hat{y}$ is estimated value of y for given x, where

a = intercept of the y–axis

$b = b_{yx}$ = slope of the regression line = regression coefficient of y on x.

y = dependent variable, x = independent variable.

Normal equations are $\quad \Sigma y = na + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$

(ii) Regression equation of x on y

A line of regression of x on y is the line which give the best estimate for the value of x for given value of y. The regression equation of x on y is $\hat{x} = c + dy$

where c = intercept of the x–axis

$d = b_{xy}$ = slope of the regression line = regression coefficient of x on y.

x = dependent variable, y independent variable.

Normal equations are

$$\Sigma x = nc + d\,\Sigma y \quad \text{and} \quad \Sigma xy = c\Sigma y + d\Sigma y^2$$

Solving these two normal equations we get

$$c = \bar{x} - d\bar{y} \qquad \text{and} \quad d = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

Therefore, the estimated regression line of x on y is $\hat{x} = c + dy$ where $\hat{x}$ is estimated value of x for given value of y.

(iii) Regression equation in terms of actual means

The *regression equation y on x* is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \text{where, } \bar{y} = \frac{\Sigma y}{n}, \bar{x} = \frac{\Sigma x}{n},$$

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = r\cdot\frac{\sigma_y}{\sigma_x}$$

The *regression equation x on y* is given by

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad \text{where } b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{x})}{\Sigma(y - \bar{y})^2} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = r\cdot\frac{\sigma_x}{\sigma_y}$$

10.3.6 Coefficients of Regression

In the regression equation of y on x, *viz.*, $\hat{y} = a + bx$, the coefficient b which is the slope of the line of regression of y on x is called the coefficient of regression of y on x. It represents the change in the value of the dependent variable y for a unit change in the value of the independent variable x. In other words, it represents the rate of change of y w.r.t. x. It is denoted by b_{yx}.

Similarly, in the regression equation of x on y, viz, $\hat{x} = c + dy$ the coefficient d represents the change in the value of the dependent variable x for a unit change in the value of the independent variable y and is called the coefficient of regression of x on y. It is denoted by b_{xy}.

Example 10: (*Finding the correlation and Regression Equation*): Use the given sample data to find the linear correlation coefficient and regression equation.

x	1	1	3	5
y	2	8	6	4

Solution: Here, $\quad n = 4$, $\Sigma x = 10$, $\Sigma y = 20$, $\Sigma x^2 = 36$, $\Sigma y^2 = 120$, $\Sigma xy = 48$.

$$\text{Correlation coefficient } r = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}} = -0.135.$$

For Regression equation:

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = -0.182, \ \bar{x} = \Sigma x/n = 5, \bar{y} = \Sigma y/n = 2.5.$$

$a = \bar{y} - b\bar{x} = 5 - (-0.182)(2.5) = 5.45.$

The estimated equation of regression line is

$$\hat{y} = a + bx = 5.45 - 0.182\,x.$$

We should realize that this equation is an estimate of the true regression equation

$$y = \beta_0 + \beta_1 x.$$

10.3.7 Outliers and Influential Points

A correlation / regression analysis of bivariate (paired) data should include an investigation of outliers and influential points, defined as follows.

Definitions

In a scatterplot, an *outlier* is a point lying far away from the other data points.
Paired sample data may include one or more *influential points*, which are points that strongly affect the graph of the regression line.

An outlier is easy to identify: Examine the scatterplot and identify a point that is away from the others. Here's how to determine whether a point is an influential point: Graph the regression line resulting from the data with the point included, then graph the regression line resulting from the data with the point excluded. If the graph changes by a considerable amount, the point is influential. Influential points are often found by identifying those outliers that are *horizontally* far away from the other points.

10.3.8 Residuals and the Least-Squares Property

We have stated that the regression equation represents the straight line that fits the data "best," and we will now describe the criterion used in determining the line that is better than all others. This criterion is based on the vertical distances between the original data points and the regression line. Such distances are called residuals.

Definition: (Residuals):

For a sample of paired (x, y) data, a residual is the difference $(y - \hat{y})$ between an observed sample y-value and the value of $\hat{y}$, which is the value of y that is predicted by using the regression equation. That is,

Residual = observed y – predicted y (i.e., estimated y) $= y - \hat{y}$

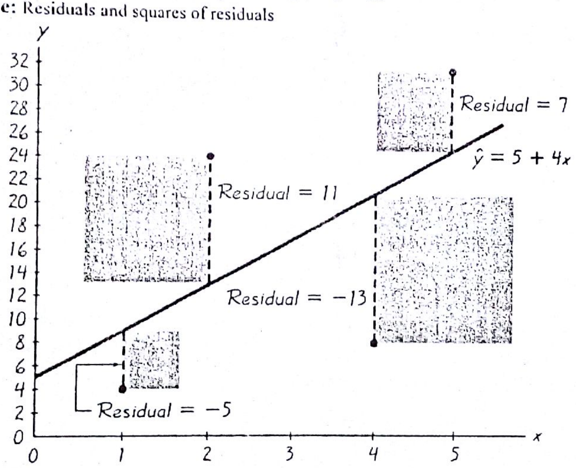
The minimum value of the sum of squares is called the ***residual sum of squares*** or ***error sum of squares***. That is

$$SSE = \text{residual sum of squares} = \sum_{i=1}^{n} (y - \hat{y})^2 = S_{yy} - S_{xy}^2 / S_{xx}$$

This definition might seem as clear as tax-form instructions, but you can easily understand residuals by referring to the following figure , which corresponds to the paired sample data listed below. In figure, the residuals are represented by the dashed lines. For a specific example, see the residual indicated as 7, which is directly above $x = 5$. If we substitute $x = 5$ into the regression equation $\hat{y} = 5 + 4x$, we get a predicted value of $\hat{y} = 25$. When $x = 5$, the *predicted value* of y is $\hat{y} = 25$, but the *actual observed* sample value is $y = 32$. The difference $y - \hat{y} = 32 - 25 = 7$ is a residual.

x	1	2	4	5
y	4	24	8	32

Figure: Residuals and squares of residuals



The regression equation represents the line that fits the points "best" according to the following *least-squares property*.

Definition: A straight line satisfies the ***least-squares property*** if the sum of the squares of the residuals is the smallest sum possible.

10.3.9 Properties of regression coefficient

Let x and y be two variables. Then there are two regression coefficients b_{yx} and b_{xy}. Some important properties are:

1. The correlation coefficient r is the geometric mean of the regression coefficients i.e., $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$

Proof: The regression co-efficient are $\dfrac{r\sigma_y}{\sigma_x}$ and $\dfrac{r\sigma_x}{\sigma_y}$

$$\text{G.M. between these two} = \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} \quad [\because \text{G.M. between } a \text{ and } b = \sqrt{ab}]$$

$$= \sqrt{r^2} = r = \text{Correlation co-efficient.}$$

2. Both the regression coefficients must have the same sign. If regression coefficients are negative then r is also negative and if they are positive then r is also positive.

Proof: Regression co-efficient of y on $x = b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$.

Regression co-efficient of x on $y = b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive, b_{yx}, b_{xy} are and r have same sign.

3. If one of the regression co-efficient is greater than unity numerically, the other must be less than unity numerically.

Proof: We know that two regression co-efficient satisfy

$$b_{yx}\, b_{xy} = r^2 < 1 \qquad [\because |r| < 1]$$

$$\therefore b_{xy} < \frac{1}{b_{yx}} < 1 \Rightarrow \frac{1}{b_{yx}} < 1 \Rightarrow 1 < b_{yx}.$$

Similarly, if $b_{yx} < 1$ then $b_{xy} > 1$.

4. The regression equations pass through their mean $(\bar{x}, \bar{y})$ i.e., regression lines interest at $(\bar{x}, \bar{y})$.

Proof: We have seen that the line of regression of y on x (which is the line of best fit when x is treated as independent variable and y as dependent variable) is

$$y = ax + b \qquad \ldots (1)$$

where a and b are given by the normal equations

$$\Sigma y = na + b\Sigma x \qquad \ldots (2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \ldots (3)$$

and n is the number of pairs of values of x and y.

Equation (2) can be written in the form

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n}$$

or, $$\bar{y} = a + b\bar{x} \qquad \ldots (4)$$

Equation (4) shows that the point $(\bar{x}, \bar{y})$ lies on (1).

Thus the line of regression passes through $(\bar{x}, \bar{y})$,

where $\bar{x}$ is the mean of x's and $\bar{y}$ the mean of y's.

Similarly we can show that the regression line $x = cy + d$ also passes through the point $(\bar{x}, \bar{y})$

5. The arithmetic mean of regression coefficients is greater than the correlation coefficient r; i.e., $\left[\dfrac{b_{yx} + b_{xy}}{2}\right] \geq r$.

Proof: We have to probe that

$$\frac{b_{yx}+b_{xy}}{2} > r \quad \text{or,} \quad \frac{\frac{r\sigma_y}{\sigma_x}+\frac{r\sigma_x}{\sigma_y}}{2} > r.$$

or, $\sigma_y^2 + \sigma_x^2 > 2\sigma_x\sigma_y$

or, $\sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y > 1$

or, $(\sigma_x - \sigma_y)^2 > 0$, which is true.

6. Regression coefficient are independent of change of origin but not of scale.

Proof: Let $u = \frac{x-a}{h}, y = \frac{y-b}{k}$ where a, b, h and k are constants.

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r\frac{k\sigma_y}{h\sigma_x} = \frac{k}{h}\left(\frac{r\sigma_y}{\sigma_x}\right) = \frac{k}{h}b_{xy}.$$

Similarly $b_{xy} = \frac{k}{h}b_{xy}$

Thus b_{yx} and b_{xy} are both independent of a and b but not of h and k.

This means regression coefficients are not affected if every values are increased or decreased by some constant multiplied by some constant value.

7. If $r = 0$, the two lines of regression are parallel to the axes.

Proof: Equations to the two lines of regression are

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \text{ and } x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y}).$$

When $r = 0$, $y - \bar{y} = 0$ and $x - \bar{x} = 0$; i.e., $y = \bar{y}$ and $x = \bar{x}$ which are equations to lines parallel to x –axis respectively.

Theorem11.1: If θ is the acute angle between the two regression lines in the case of the two variables x and y, show that

$$\tan\theta = \frac{1-r}{r}\,\frac{\sigma_x\sigma_y}{\sigma_x\sigma_y} \qquad \text{where } r, \sigma_x, \sigma_y, \text{ have their usual meanings.}$$

Explain the significance when $r = 0$ and $r = \pm 1$.

Solution: Equations to the lines of regression of y on x and x on y are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \text{ and } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Their slopes are $m_1 = r\frac{\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r\sigma_x}$

$$\tan\theta = \pm\frac{m_1 - m_2}{1 + m_2m_1} = \pm\frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}}$$

$$= \pm\frac{1-r^2}{r}\,\frac{\sigma_y}{\sigma_x}\,\frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2} = \frac{1-r^2}{r}\,\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Since $r^2 \le 1$ and σ_x, σ_y are positive

∴+ ve gives the acute angle between the lines

Hence $\tan\theta = \frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

(i) When $r = 0$, then $\theta = \infty$ or $\theta = 90°$.

Thus when $r = 0$; i.e., the variables are not correlated, the lines of regression are perpendicular to each other.

(ii) When $r = \pm 1$, that $\tan\theta = 0 \Rightarrow \theta = 0$ or π. The two lines of regression are either parallel or coincide. Hence when $r = \pm 1$ i.e., there is perfect positive or negative correlation between x and y, the lines of regression coincide.

(iii) If $\tan\theta > 0$, then θ is acute angle and
If $\tan\theta < 0$, then θ is obtuse angle

10.3.10 Using the Regression Equation for Predictions

Regression equations can be helpful when used for predicting the value of one variable, given some particular value of the other variable. If the regression line fits the data quite well, then it makes sense to use its equation for predictions, provided that we don't go beyond the scope of the available values. However, we should use the equation of the regression line only if r indicates that there is a linear correlation. In the absence of a linear correlation, we should not use the regression equation for projecting or predicting; instead, our best estimate of the second variable is simply its sample mean.

In predicting a value of y based on some given value of x...

1. ***If there is not a linear correlation, the best predicted y -value is y.***
2. ***If there is a linear correlation, the best predicted y -value is found by substituting the x-value into the regression equation.***

If r is near −1 or +1, then the regression line fits the data well, but if r is near 0, then the regression line fits poorly (and should not be used for predictions).

Guidelines for Using the Regression Equation

1. If there is no linear correlation, don't use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data. If you find a regression equation that relates women's heights and shoe sizes, it's absurd to predict the shoe size of a woman who is 10 ft tall.
3. A regression equation based on old data is not necessarily valid now. The regression equation relating used-car prices and ages of cars is no longer usable if it's based on data from the 1970s.
4. Don't make predictions about a population that is different from the population from which the sample data were drawn. If we collect sample data from men and develop a regression equation relating age and TV remote-control usage, the results don't necessarily apply to women. If we use state averages to develop a regression equation relating SAT math scores and SAT verbal scores, the results don't necessarily apply to individuals.

10.3.11 Inferences concerning least squares methods

The regression equation $y = a + bx$ is obtained on the basis of sample data. We are often interested in corresponding equation $y = \alpha + \beta x$ for the population from which the sample was drawn. The following is test concerning a normal population.

A test of hypotheses concerning the slope parameter $\beta = b$

To test the hypothesis that the regression coefficient β is equal to some specified value b, we use the fact that the statistic

$$t = \frac{(b-\beta)}{s_e}\sqrt{S_{xx}}$$

is a random having the t distribution with $n-2$ degrees of freedom.

Similarly statistics for inference about α

$$t = \frac{(a-\alpha)}{s_e}\sqrt{\frac{nS_{xx}}{S_{xx}+n(\bar{x})^2}}$$

is a random having the t distribution with $n-2$ degrees of freedom.

Definition: (***Standard error of the estimate***)

The standard error of estimate, denoted by s_e, is a measure of the differences (or distances) between the observed sample y-values and the predicted values $\hat{y}$ that are obtained using the regression equation. It is given as

$$s_e^2 = \frac{\Sigma(y-\hat{y})^2}{n-2} \quad \text{(where } \hat{y} \text{ is the predicted y-value)}$$

or an equivalent formula for this estimate of σ^2 is given by

$$\text{Estimate of } \sigma^2 = s_e^2 = \frac{S_{yy}-(S_{xy})^2/S_{xx}}{n-2}$$

Remarks:

1. Relationship between S_{xx}, S_{yy} and the respective sample variances of the x's and y's are:

$$s_x^2 = \frac{S_{xx}}{n-1} \text{ and } s_y^2 = \frac{S_{yy}}{n-1}$$

10.3.12 Confidence Interval for the Intercept and slope

$(1-\alpha)$ 100% confidence intervals using confidence limits

For intercept $\alpha : a \pm t_{\alpha/2,\,n-2} \times s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$

For slope $\beta : b \pm t_{\alpha/2,\,n-2} \times s_e\frac{1}{\sqrt{S_{xx}}}$ where $d.f. = n-2$.

Example 11: (*Fitting a straight line by least squares*)

The following are measurements of the air velocity and evaporation coefficient of burning fuel droplets in an impulse engine:

Air velocity (cm/ sec) x	*Evaporation coefficient (mm²/sec) y*
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

Fit a straight line to these data by the method of least squares, and use it to estimate the evaporation coefficient of a droplet when the air velocity is 190 cm/sec.

[TU, BE, 2064 Shrawan/ 2065 Kartik/2065 Chaitra]

Solution: For $n = 10$ pairs (x_i, y_i), we first calculate(Using Calculator)

$$\Sigma x = 2{,}000,\ \Sigma x^2 = 532{,}000;\ \Sigma y = 8.35;\ \Sigma xy = 2{,}175.40;\ \Sigma y^2 = 9.1097$$

then $S_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 532{,}000 - (2{,}000)^2/n = 132{,}000$

$$S_{xy} = \Sigma xy - (\Sigma x\Sigma y)/n = 2{,}175.40 - (2{,}000)(8.35)/10 = 505.40$$

$$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 9.1097 - (8.35)^2/10 = 2.13745.$$

So, $b = \frac{S_{xy}}{S_{xx}} = \frac{505.40}{132000} = 0.00383$

$$a = \bar{y} - b\bar{x} = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \frac{8.35}{10} - (0.00383)\frac{2000}{10} = 0.069.$$

Thus, the equation of the straight line that best fits the given data in the sen of least squares is

$\hat{y} = a + bx = 0.069 + 0.00383\,x$

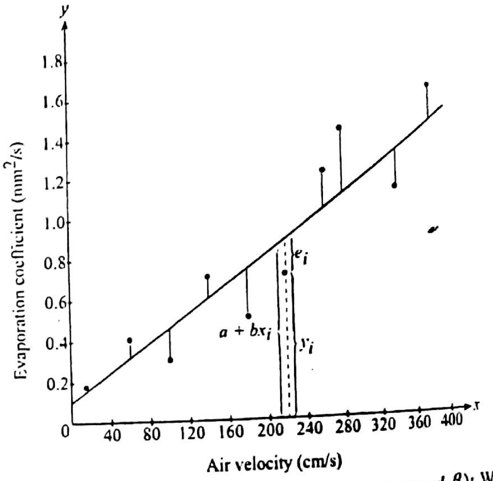
and for $x = 190$ we predict that the evaporation coefficient will be

$\hat{y} = 0.069 + (0.00383)(190) = 0.80\ m^2/s$

Finally, the residual sum of squares is

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 2.13745 - \frac{(505.40)^2}{132000} = 0.20238$$

Note 1: Diagram for least squares criterion showing the vertical deviations



Example 12: (*95% confidence interval for the intercept α and β*): With referer to the previous example construct a 95% confidence interval for the interc α.

$$\text{Here } s_e^2 = \frac{S_{yy}-(S_{xy})^2/S_{xx}}{n-2} = \frac{2.13745-(505.40)^2/132000}{10-2} = 0.0253$$

$$(1-\alpha)\ 100\% = 95\% \Rightarrow \alpha = 0.05$$

$\therefore \quad t_{\alpha 2} = t_{0.025} = 2.306$ for $10-2 = 8$ degrees of freedom and $s_e = \sqrt{0.0253} = 0.159$

Hence 95% confidence limits for intercept are given by

$$\text{C.I. for intercept} = a \pm t_{\alpha/2,n-2} \times s_e\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

$$= 0.069 \pm (2.306)(0.159)\sqrt{\frac{1}{10}+\frac{(200)^2}{132000}} = 0.069 \pm 0.233$$

Hence the required interval for intercept is (– 0.164, 0.302)

Example 13: *(A test of hypotheses concerning the slope parameter $\beta = b$)* With reference to the previous example 1 test the null hypothesis $\beta = 0$ against the alternative hypothesis $\beta \neq 0$ at the 0.05 level of significance.

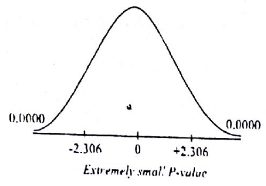
Solution:

Step 1. ***Null hypothesis*** H_0: $\beta = 0$

Alternative hypothesis H_1: $\beta \neq 0$

Step 2. ***Level of significance***: $\alpha = 0.05$

Step 3. ***Criterion***: Reject the null hypothesis if $t < -2.306$ or $t > 2.306$, where 2.306 is the value of $t_{0.025}$ for $10 - 2 = 8$ degrees of freedom, and t is given by the formula $t = \frac{(b - \beta)}{s_e}\sqrt{S_{xx}}$.



Step 4. ***Calculations***: Using the quantities obtained in the previous example

$$t = \frac{0.00383 - 0}{0.159}\sqrt{132{,}000} = 8.75.$$

Step 5. ***Decision***: Since $t = 8.75$ exceeds 2.306, the null hypothesis must rejected; we conclude that there is a relationship between air velocity and the average evaporation coefficient.

Example 14: Engineers fabricating a new transmission–type electron multiplier created an array of silicon nanopillars on a flat silicon membrane. The precise structure can influence the electrical properties. So, subsequently, the height and width of 50 nanopillars were measured in nanometers (nm) or 10^{-9} meters. The summary statistics, with x = width and y = height are $n = 50$, $\bar{x} = 88.34$, $\bar{y} = 305.58$

$S_{xx} = 7{,}239.22$, $S_{xy} = 17{,}840.1$, $S_{yy} = 66{,}975.2$

(a) Find the least squares line for predicting height from width.

(b) Find the least squares line for predicting width from height.

Solution: (a) Here y = height and the least squares estimates are

$$\text{slope} = b = \frac{S_{xy}}{S_{xx}} = \frac{17840.1}{7239.22} = 2.464$$

and $a = \bar{y} - b\,\bar{x} = 305.58 - (2.464)(88.34) = 87.88$

The fitted line is height = 87.88 + 2.464 width.

(b) Here width is the response variable and height the predictor, so x and y must be interchanged.

$\text{slope} = b = \frac{S_{xy}}{S_{yy}} = \frac{17840.1}{66975.2} = 0.266$; and $a = \bar{x} - b\bar{y} = 88.34 - (0.266)(305.58) = 6.944$

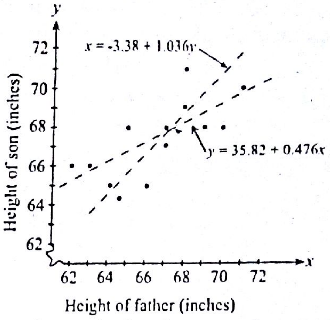
The fitted line is width = 6.944 + 0.266 height.

Example 15: The following table shows the respective heights x and y of a sample of 12 fathers and their oldest sons. (a) Construct a scatter diagram. (b) Find the least squares regression line of y on x. (c) Find the least squares regression line of x on y.

Height x of Father (inches)	65	63	67	64	68	62	70	66	68	67	69	71
Height y of son (inches)	68	66	68	65	69	66	68	65	71	67	68	70

[*TU, BE, 2068 Magh*]

Solution: (a) The scatter diagram is obtained by plotting the points (x, y) on a rectangular coordinate system as shown:



(b) The regression line of y on x is given by $y = a + bx$, where a and b are obtained by solving the normal equations

$$\Sigma y = an + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

x	y	x^2	xy	y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\Sigma x = 800$	$\Sigma y = 811$	$\Sigma x^2 = 53{,}418$	$\Sigma xy = 54{,}107$	$\Sigma y^2 = 54{,}849$

Hence the normal equations becomes

$$12\,a + 1800\,b = 811$$
$$800\,a + 53{,}418\,b = 54{,}108$$

From which we find $a = 35.82$ and $b = 0.476$ so that $y = 35.82 + 0.476\,x$.

Another method [In terms of actual means]

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = 0.476$$

$\therefore$ $y - \bar{y} = b_{yx}(x - \bar{x})$, $\bar{y} = \frac{\Sigma y}{n} = 67.583$, $\bar{x} = \frac{\Sigma x}{n} = 66.667$

$\Rightarrow$ $y = 0.476\,(x - 66.67) + 67.58 = 35.85 + 0.476\,x$

(c) Regression line of x on y is given by $x = c + dy$ where c and d are obtained by solving normal equations

$\Sigma x = cn + d\,\Sigma y$ and $\Sigma xy = c\Sigma y + d\,\Sigma y^2$

So, $12\,c + 811\,d = 800$ and $811\,c + 54{,}894\,d = 54{,}107$

From which we get $c = -3.38$ and $d = 1.036$

Hence the required line is $x = -3.38 + 1.036\,y$

Another method [In terms of actual means]

$$b_{xy} = \frac{n\Sigma xy - \Sigma y\Sigma x}{n\Sigma y^2 - (\Sigma y)^2} = 1.036$$

Therefore, $(x - \bar{x}) = b_{xy}(y - \bar{y})$

$\Rightarrow \quad x = 1.036\,(y - 67.583) + 66.667 = -3.35 + 1.036\,y.$

Example 16: Compute the standard error of estimate s_e for the data

Height x of Father (inches)	65	63	67	64	68	62	70	66	68	67	69	71
Height y of son (inches)	68	66	68	65	69	66	68	65	71	67	68	70

Solution: From above example the regression line of y on x is $y = 35.82 + 0.476\,x$. In following Table are listed in actual values of y and the estimated values of y, denoted by $\hat{y}$, as obtained from the regression line. For example, corresponding to $x = 65$, we have $\hat{y} = 35.82 + 0.476(65) = 66.67$.

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70
$\hat{Y}$	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	69.19	67.71	68.66	69.62
$y - \hat{y}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

Also listed are the values $y - \hat{y}$ which are needed in computing s_e

$$s_e^2 = \frac{\Sigma(y - \hat{y})^2}{n - 2} = \frac{(1.24)^2 + (0.19) + \ldots + (0.38)^2}{10} = 1.97 \text{ and } s_e = 1.40$$

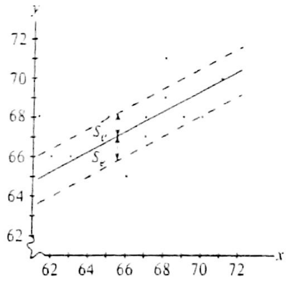
Example 17:

(a) Construct two lines parallel to the regression line of above example and having vertical distance s_e from it.

(b) Determine the percentage of data points falling between these two lines.

Solution:

(a) The regression line $y = 35.82 + 0.476\,x$ as obtained in above example is shown solid in following figure. The two parallel lines, each having vertical distance $s_e = 1.40$ from it, are shown dashed in



(b) From the figure it is seen that 9 out of 12 data points, 7 fall between the lines. Then the required percentage is

$$\frac{9}{12} \times 100 = 75\%$$

<u>**Another method**</u>: From the last line in table of previous example $y - \hat{y}$ lies between −1.28 and 1.28 (i.e., $\pm s_e$) for 9 points (x, y). Then the required percentage is

$$\frac{9}{12} \times 100\% = 75\%.$$

Example 18: Ten steel wires of diameter 0.5mm and length 2.5m were extended in a laboratory by applying vertical forces of varying magnitudes. Results are as follows:

Force (kg) x	15	19	25	35	42	48	53	56	62	65
Increase in length (mm) y	1.7	2.1	2.5	3.4	3.9	4.9	5.4	5.7	6.6	7.2

(a) Estimate the parameters of a simple linear regression model with force as explanatory variable.

(b) Find 95% confidence limits for the slope of line. *[TU BE 2068 Bhagra]*

Solution: Using calculator, we get

$\Sigma x = 420;\ \Sigma x^2 = 20518;\ \Sigma y = 43.4;\ \Sigma xy = 2128.5;\ \Sigma y^2 = 221.38$

then $S_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 2878;\ S_{xy} = \Sigma xy - (\Sigma x\Sigma y)/n = 305.7; \bar{x} = 42; \bar{y} = 4.34$

$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 33.024$

So, $b = \frac{S_{xy}}{S_{xx}} = 0.1062;\ a = \bar{y} - b\bar{x} = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = -0.1212$

(a) The parameters of a simple linear regression model with force as explanatory variable are $a = -0.1212$ and $b = 0.1062$.
The simple linear regression model with force as explanatory variable is
Magnitude = − 0.1212 + 0.1062 force.

(b) We know that $(1 - \alpha)$ 100% confidence intervals using confidence limits

For intercept $\alpha : a \pm t_{\alpha/2,\,n-2} \times s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

For slope $\quad \beta : b \pm t_{\alpha/2,\,n-2} \times s_e \frac{1}{\sqrt{S_{xx}}}$ where $d.f. = n - 2$.

For, 95% confidence limits for the slope β of line:

$(1 - \alpha)\% = 95\% \Rightarrow \alpha = 0.05$

$\therefore \quad t_{\alpha/2} = t_{0.025} = 2.306$ for $n - 2 = 10 - 2 = 8$ degrees of freedom and $s_e = \sqrt{0.0253} = 0.159$

Here $s_e^2 = \frac{S_{yy} - (S_{xy})^2/S_{xx}}{n - 2} = 0.0691$

Hence 95% confidence limits for intercept are given by

C.I. for intercept $= b \pm t_{\alpha/2,n-2} \times s_e \frac{1}{\sqrt{S_{xx}}}$

$= 0.1062 \pm (2.306)(0.0691)\frac{1}{\sqrt{2878}} = 0.1062 \pm 0.00297 = (0.10323, 0.10917)$

Hence the required interval for intercept is, (0.10323, 0.10917)

Example 19: Measurements of the resistance R to the motion of a train at different speeds gave the following results:

V(miles/hour)	10	20	30	40	50
R(lb/ton)	8	10	15	21	30

Assuming a law of the form $R = a + bV^2$, find the best values of a and b.

Solution: Put $R = y$, $V^2 = x$. Then the given equation becomes $y = a + bx$.
The following table gives the values of x and y.

$x(= V^2)$	100	400	900	1600	2500
$y(= R)$	8	10	15	21	30

Let, $\quad X = \frac{x - 900}{100}$, $Y = y - 21$. $Y = A + BX$.

The normal equations are $\Sigma Y = 5A + B\Sigma X$ and $\Sigma YX = A\Sigma X + B\Sigma X^2$.
The working is shown below.

X	Y	XY	X^2
−8	−13	104	64
−5	−11	55	25
0	−6	0	–
7	0	0	49
16	9	144	256
$\Sigma X = 10$	$\Sigma Y = -21$	$\Sigma XY = 303$	$\Sigma X^2 = 394$

Substituting in the normal equations,

$$-21 = 5A + 10B \text{ and } 303 = 10A + 39.4B$$

Solving $A = -6.1$, $B = 0.92$. Therefore, $Y = -6.1 + 0.92X$

or, $y - 21 = -6.1 + 0.92\dfrac{(x - 900)}{100}$

or, $y = 6.62 + 0.0092x$

∴ $a = 6.62$, $b = 0.0092$

Example 20: Find the equation of the regression line of y on x, if the observations (x_i, y_i) are following: (1, 4), (2, 8), (3, 2), (4, 12), (5, 10), (6, 14), (7, 16), (8, 6), (9,18).

Solution:

x_i	y_i	x_i^2	$x_i y_i$
1	4	1	4
2	8	4	16
3	2	9	6
4	12	16	48
5	10	25	50
6	14	36	84
7	16	49	112
8	6	64	48
9	18	81	162
45	90	285	530

Here, $n = 9$, $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{45}{9} = 5$, $\bar{y} = \dfrac{\Sigma y_i}{n} = \dfrac{90}{9} = 10$

$$b_{yx} = \frac{\Sigma x_i y_i - \frac{1}{n}(\Sigma x_i)(\Sigma y_i)}{\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2} = \frac{530 - \frac{1}{9}(45)(90)}{285 - \frac{1}{9}(45)^2} = \frac{530 - (5 \times 90)}{285 - (5 \times 45)}$$

$$= \frac{530 - 450}{285 - 225} = \frac{80}{60} = \frac{4}{3} = 1.33$$

The equation of regression line of Y on X is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

or, $y - 10 = 1.33(x - 5)$

or, $y = 10 + 1.33x - 6.65$

or, $y = 3.35 + 1.33x.$

Example 21: The following regression equations were obtained from a correlation table: $y = 0.516x + 33.73$, $x = 0.512y + 33.52$. Find the value of (i) the correlation t r; (ii) the mean of x's and the mean of y's.

Solution: (i) From first equation we have $b_{yx} = r\dfrac{\sigma_y}{\sigma_x} = 0.516$

and from secnd equation, $b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = 0.512.$

$$\therefore\ r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = 0.516 \times 0.512$$

or, $r^2 = 0.516 \times 0.512$ So, $r = \sqrt{0.516 \times 0.512} = 0.514$

(ii) Suppose $(\bar{x}, \bar{y})$ is the mean point in the co-ordinate plane of the two axes, then equations will be satisfied by it. In fact the regression lines pass through means and intersect each other at that point.

∴ $\bar{y} = 0.516\bar{x} + 33.73$ and $\bar{x} = 0.512\bar{y} + 32.52$

On solving (iii) and (iv), we get $\bar{x} = 67.16$, $\bar{y} = 68.61$

Example 22: In a partially destroyed laboratory record of an analysis of a correlation data, the following results are eligible: Variance of $x = 9$. Regression equations are $8x - 10y + 66 = 0$ and $40x - 18y = 214$. What were (i) the mean values of x and y (i) The standard deviation of y, and (iii) The correlation coefficient between x and y.

Solution: As the regression lines always intersect at means,

∴ Their point of intersection given the mean values

Equations are $8\bar{x} - 10\bar{y} + 66 = 0$ or, $8\bar{x} - 10\bar{y} = -66$

and $40\bar{x} - 18\bar{y} = -214$

Solving thes equations, we get $\bar{x} = 13$, $\bar{y} = 17$

Now $\sigma_x = \sqrt{\text{Variance}} = \sqrt{9} = 3$

Equations of regression lines are

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \text{or,} \quad x = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) + \bar{x}$$

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{or,} \quad y = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) + \bar{y}$$

Now given equations are

$$x = \frac{18}{40}y + \frac{214}{40} \quad \text{and} \quad y = \frac{8}{10}x + \frac{66}{10}$$

Regression coefficients are $b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = \dfrac{18}{40}$ and $b_{yx} = r\dfrac{\sigma_y}{\sigma_x} = \dfrac{8}{10}$

Coefficient of correlation

$$r = \sqrt{\text{Product of Regression coefficients}}$$

$$= \sqrt{b_{yx} \times b_{xy}} = \sqrt{\frac{8}{10} \times \frac{18}{40}} = \sqrt{\frac{144}{400}} = \frac{12}{20} = 0.6.$$

Now, $r\dfrac{\sigma_y}{\sigma_x} = \dfrac{8}{10}$; $\sigma_x = 3$; $r = \dfrac{12}{20}$. So $\dfrac{12}{20} \times \dfrac{\sigma_y}{3} = \dfrac{8}{1} \Longrightarrow \sigma_y = \dfrac{8 \times 20 \times 3}{12 \times 10} = 4.$

Hence, (i) $\bar{x} = 13$, $\bar{y} = 17$; (ii) Standard deviation of $y = 4$;

(iii) Correlation coefficient $r_{xy} = 0.6.$

Example 23: Prove that the correlation coefficient is the geometric mean between the regression coefficients. Hence, find the correlation coefficient when the regression coefficients are 0.8 and 0.2 respectively.

Solution: In regression coefficient of y on x is $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

The regression coefficient of x on y is $b_{xy} = \dfrac{\sigma_x}{\sigma_y}$

Now, $b_{yx} \times b_{xy} = r\dfrac{\sigma_y}{\sigma_x} \times \dfrac{\sigma_x}{\sigma_y} = r^2 \Rightarrow r = \sqrt{b_{yx} \times b_{xy}}$

∴ $r^2 = b_{yx} \times b_{xy} \quad \Rightarrow r = +\sqrt{1.6} = 4.$

10.3.13 Multiple Regression

So far, we have used methods of regression to investigate relationships between exactly two variables, but some circumstances require more than two variables. In predicting the price of a diamond, for example, we might consider variables such as weight (in carats), color, and clarity, so that a total of four variables are involved. This section presents a method for analyzing such relationships involving more than two variables. ***As in the previous sections of this chapter, we will work with linear relationships only.*** We begin with the multiple regression equation.

Definition:

A ***multiple regression equation*** expresses a linear relationship between a dependent variable y and two or more independent variables ($x_1, x_2, \dots, x_k$). The general form of a multiple regression equation is

$$\hat{y} = b_0 + b_1 x + b_2 x + \dots + b_k x_k .$$

We will use the following notation, which follows naturally from the notation.

Notations:

$\hat{y} = b_0 + b_1 x + b_2 x + \dots + b_k x_k .$

(General form of the estimated multiple regression equation)

n = sample size

k = number of independent variables. (The independent variables are also called ***predictor variables*** or x variables.)

$\hat{y}$ = predicted value of the dependent variable y (computed by using the multiple regression equation)

$x_1, x_2, \dots, x_k$ are the independent variables.

β_0 = the y-intercept, or the value of y when all of the predictor variables are 0 (This value is a population parameter.)

b_0 = estimate of β_0 based on the sample data (b_0 is a sample statistic.)

$\beta_1, \beta_2, \dots, \beta_k$ are the coefficients of the independent variables $x_1, x_2, \dots, x_k$

b_1 , b_2 , . . . , b_k are the sample estimates of the coefficients $\beta_1, \beta_2, \dots, \beta_k$

If there is a linear relationship between a dependent variable z and two independent variables x and y, then we would seek an equation connecting the variables that has the form.

$$\hat{z} = a + bx + cy \qquad \text{- - - (i)}$$

This is called regression equation of z on x and y.

If x is the dependent variable, a similar equation will be called a ***regression equation of x on y and z.***

The equation (i) represents a plane in three-dimensional coordinate system, it is often called a regression plane. To find the least-squares regression plane, we determine a, b, c in (i) so that

$$\Sigma z = na + b\Sigma x + c\Sigma y$$
$$\Sigma xz = a\Sigma x + b\Sigma x^2 + c\Sigma xy$$
$$\Sigma yz = a\Sigma y + b\Sigma xy + c\Sigma y^2$$

These equation are called the ***normal equations corresponding*** to equation (i).

Example 24: The following table shows the weight z to the nearest pound, heights x to the nearest inch, and ages y to the nearest year, of 12 boys

(a) fit a least squares regression plane;

(b) Estimate the weight of a boy who is 9 years old and 54 inches tall.

Weight (z)	64	71	53	67	55	58	77	57	56	51	76	68
Height (x)	57	59	49	62	51	50	55	48	52	42	61	57
Age (y)	8	10	6	11	8	7	10	9	10	6	12	9

Solution: (a) The linear regression equation z on x and y can be written as

$$\hat{z} = a + bx + cy$$

From above table we get, $\Sigma z = 753$, $\Sigma x = 643$, $\Sigma y = 106$,

$\Sigma z^2 = 48,139$, $\Sigma x^2 = 34,843$, $\Sigma y^2 = 976$,

$\Sigma xz = 40,830$, $\Sigma yz = 6796$, $\Sigma xy = 5779$

So, The normal equations are

$$12\,a + 643\,b + 106\,c = 753$$
$$643\,a + 34,843\,b + 5779\,c = 40,830$$
$$106\,a + 5779\,b + 976\,c = 6796$$

Solving above equations we get, $a = 3.6512$, $b = 0.8546$, $c = 1.5063$

The required least square regression plane is

$$\hat{z} = 3.65 + (0.855)\,(x) + (1.506)\,(y).$$

(b) Putting x =54 and y = 9 in above equation the estimated weight is

$$\hat{z} = 3.65 + (0.855)\,(54) + (1.506)\,(9) = 63.356 = 63\ lb.$$

Example 25: The following are data on the number of twists required to break a certain kind of forged alloy bar and percentage of two alloying elements present in the metal:

Number of twists (z)	*Percentage of elements A: (x)*	*Percentage of element B:(y)*
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10
58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
57	4	15
19	1	20
31	2	20
33	3	20
43	4	20

Fit a least squares regression plane and use its equation to estimate the number of twists required to break one of the bars when $x = 2.5$ and $y = 12$

Solution: From above table using calculator we get

$\Sigma z = 723$, $\Sigma x = 40$, $\Sigma y = 200$, $\Sigma x^2 = 120$, $\Sigma y^2 = 3,000$

$\Sigma xy = 500$, $\Sigma xz = 1,963$, $\Sigma yz = 8,210$

Then the normal equations are

$$723 = 16\,a + 40\,b + 200\,c$$
$$1963 = 40\,a + 120\,b + 500\,c$$
$$3210 = 200\,a + 500\,b + 3000\,c$$

Solving these equations, we get unique solution of this system as

$$a = 46.4,\ b = 7.78,\ c = -1.65$$

The equation of estimated regression plane is

$$\hat{z} = 46.4 + 7.78\,x - 1.65y$$

Putting $x = 2.5$ and y =12 into this equation

We get, $\hat{z} = 46.4 + (7.78)(2.5) - (1.65)\,(12) = 46.0$

Note that b and c are estimates of the average change in z resulting from a unit increase in the corresponding independent variable when the other independent variable in held fixed.

Example 26: Past experience shows the following result of productivity per hectare with the respective uses of fertilizers and seeds. Fit the multiple linear regression equation of Z on X and Y from the given data. [TU BE 2068 Magh]

Fertilizer (X), kgs	45	30	70	75	65	80
Seeds (Y), kgs	2	1.8	3	2.5	2	3
Productivity kgs (Z)	2000	2100	1800	1900	2400	2500

Solution: The linear regression equation Z on X and Y can be written as

$$z = a + bx + cy$$

Normal equations corresponding to equation are

$$\Sigma z = na + b\Sigma x + c\Sigma y$$
$$\Sigma xz = a\Sigma x + b\Sigma x^2 + c\Sigma xy$$
$$\Sigma yz = a\Sigma y + b\Sigma xy + c\Sigma y^2$$

z	x	y	x^2	y^2	xy	xz	yz	
2000	45	2	2025	4	90	90000	4000	
2100	30	1.8	900	3.24	54	63000	3780	
1800	70	3	4900	9	210	126000	5400	
1900	75	2.5	5625	6.25	187.5	142500	4750	
2400	65	2	4225	4	130	156000	4800	
2500	80	3	6400	9	240	200000	7500	
12700	365	14.3	24075	35.49	911.5	597500	30230	

From above table we get, $\Sigma z = 12700$, $\Sigma x = 365$, $\Sigma y = 14.3$,

$\Sigma x^2 = 24075$, $\Sigma y^2 = 35.49$, $\Sigma xz = 597500$, $\Sigma yz = 30230$, $\Sigma xy = 911.5$

So, The normal equations are

$$6a + 365b + 14.3c = 12700$$
$$365a + 24075b + 911.5c = 40{,}830$$
$$14.3a + 911.5b + 35.49c = 30230$$

Solving above equations by using calculator, we get,

$$a = -8656.43,\ b = -1136.24,\ c = 33522.07$$

The required least square regression equation (plane) is

$$z = -8656.43 - 1136.24x + 33522.07.$$

Example 27: The following table gives the results of the measurements of train resistances. V is the velocity in miles per hour, R is the resistance in pounds per ton:

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

If R is related to V by the relation, $R = a + bV + cV^2$; find a, b, c.

Solution. Here the number of observations is even. The two middle values of V are 60 and 80, the mean value of these numbers is 70. We take.

$$x = \frac{V - 70}{10};\ y = R - 22.8.\ \text{Let}\ y = A + Bx + Cx^2$$

The normal equations are

$$\Sigma y = 6A + B\Sigma x + C\Sigma x^2$$
$$\Sigma xy = A\Sigma x + B\Sigma x^2 + C\Sigma x^3$$
$$\Sigma x^2y = A\Sigma x^2 + B\Sigma x^3 + C\Sigma x^4.$$

x	y	xy	x^2	x^3	x^4	x^2y
-5	-17.3	86.5	25	-125	625	-432.5
-3	-13.7	41.1	9	-27	81	-123.3
-1	-7.9	7.9	1	-1	1	-7.9
0	0	0	1	1	1	0
3	10.5	31.5	9	27	81	94.5
5	23.2	116.0	25	125	625	580.0
$\Sigma x = 0$	$\Sigma y = -5.2$	$\Sigma xy = 283$	$\Sigma x^2 = 70$	$\Sigma x^3 = 0$	$\Sigma x^4 = 1414$	$\Sigma x^2y = 1108$

Substituting in the normal equations,

$$-5.2 = 6A + 70C,\ 283 = 70B,\ 110.8 = 70A + 1414C.$$

Solving, $B = 4.04$, $C = 0.29$, $A = -4.25$

Hence, $y = -4.25 + 4.04x + 0.29x^2$

$$\text{or,}\ R - 22.8 = -4.25 + 4.04 \times \frac{V - 70}{10} + 0.29 \times \frac{(V - 70)^2}{10}$$

$$\text{or,}\ R = 22.8 - 4.25 + 0.404V - 28.28 + 0.0029V^2 - 0.406V + 14.21$$
$$= 3.48 - 0.002V + 0.0029V^2.$$

Comparing with $R = a + bV + cV^2$, we get

$$a = 3.48,\ b = -0.002,\ c = 0.0029.$$

10.3.14 Relationship between Correlation and regression

There are two important relationships between r and least squares fit of a straight line:

1. $r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}\,\frac{S_{xy}}{S_{xx}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$

   So, the sample correlation coefficient, r, and the least squares estimate of slope, b, have the same sign.

2. The Proportion of the y variability explained by the linear relation is

$$\frac{\textit{Sum of squares due to regression}}{\textit{Total sum of squares of y}} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}\,S_{yy}} = r^2$$

10.3.15 Difference between Correlation and regression

1.	Correlation is the relationship between two or more variables which vary in sympathy with the other in the same or the opposite direction.	1.	Regression means going back and it is a mathematical measure showing the average relationship between two variables
2.	It finds out the degree of relationship between two variables and not the cause and effects of the variables.	2.	It indicates the cause and effect relationship between the variables and establishes a functional relationship
3.	The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1, inclusive.	3.	Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
4.	There may be nonsense correlation between two variables.	4.	In regression there is no such nonsense regression
5.	It has limited application because it is confined only to linear relationship between the variables.	5.	It has wide application, as it studies linear and non-linear relationship between the variables.
6.	It is not very useful for mathematical treatment	6.	It is widely used for further mathematical treatment.

Exercise10

Theoretical Questions

1. What are the properties of correlation coefficient? [TU, BE, 2058 Shrawan/ 2062 Jestha/ 2065 Kartik/ 2067 Mangsir]
2. Define correlation coefficient. Prove that r lies between −1 and +1 (inclusive) [TU, BE, 2061 Ashwin/ 2065 Chaitra/ 2065 Chaitra (BIE)/ 2067 Shrawan]
3 Prove that karl pearson's coefficient of correlation cannot exceed the limits $-1 \le r \le 1$. [TU, BE, 2064 Poush]
4. Discuss the application of correlation coefficient in contest of engineering field with suitable example. [TU, BE, 2063 Kartik]
5 Define Pearson's correlation coefficient and coefficient of determination. Discuss the linear properties of correlation coefficient. [TU BE 2068 Magh(Back)]
6. What is the regression analysis ? How does it differ from correlation ? [TU, BE 2056 Bhadra/ 2057 Bhadra/ 2063 Ashardh/ 2065 Chaitra (Re)/ 2066 Magh]
7. Explain clearly why there are usually two lines of regression? Point out the case when there is only one line of regression. Illustrate your answer by diagram also. [TU, BE, 2062 Baisakh/ 2067 Mangsir]
8. Explain the concept of regression and point out its application dealing with engineering problem with suitable example. [TU, BE, 2064 shrawan/ 2063 kartik/ 2067 Mangsir]
9. Define regression coefficient. What are the properties of regression coefficient? [TU, BE, 2064 Shrawan/2068Magh(Back)]
10. Write the basic concept of least square method of simple regression. [TU, BE, 2064 Poush]
11. What is regression analysis? Differentiate between correlation and regression.
12. Discuss briefly multiple regression analysis with suitable examples.
13. Define partial correlation. Distinguish between partial and multiple correlations.

Numerical Problems

1. Heavy metals can inhibit the biological treatment of waste in municipal treatment plants. Monthly measurements were made at a state-of-the-art treatment plant of the amount of chromium ($\mu g/l$) in both the influent and effluent.

Influent	250	290	270	100	300	410	110	130	1100
Effluent	19	10	17	11	70	60	18	30	180

(a)Make a scatter plot.

(b) Calculate the correlation coefficient r. [Ans: $r = 0.942$]

2. Calculate the Karl Pearson's coefficient of correlation between age and playing habits from the data given below.

Age	20	21	22	23	24	25
No. of students	500	400	300	240	200	160
Regular players	400	300	180	96	60	24

[Ans: $r = -0.9738$] [TU, BE, 2062 Bhadra]

3. Calculate the coefficient of correlation from the following data :

Sales (lakhs)	45	56	39	54	45	40	56	60	30	36
Advertisement cost (000 Rs)	40	36	30	44	36	32	45	42	20	36

Draw your conclusion from result. [Ans: $r = 0.823$] [TU, BE, 2062 Baisakh]

4. Calculate the Karl Pearson's coefficient of correlation from the following data regarding price and demand of certain commodity.

Price (in Rs)	21	22	23	24	25	26	27	28	29	30
Demand (in 1000 units)	18	19	19	16	17	16	16	15	13	11

[Ans: $r = -0.9108$] [TU, BE, 2063 Kartik]

5. Calculate Karl Pearson's coefficient of correlation from the following data using 4 and 26 respectively as the origin of X and Y respectively.

X	43	44	46	40	44	42	45	42	38	40
Y	29	31	19	18	19	27	27	29	41	30

[Ans: $r = -0.5217$] [TU, BE, 2064 Shrawan

6. The following table gives age and percentage of blindness in respective age interva Find out if there is any correlation between age and blindness.

Age (yrs)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
%of blindness	70	63	21	26	45	31	46	80

[Ans: $r = 0.046$, Positive correlation] [TU, BE, 2067 Mangsir

7. On 13 April 1994, the following concentrations of pollutants were recorded at eigh stations of the monitoring system for air pollution control located in the downtow area of Milan, Italy.

	Station							
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
NO_2 *mg/m*^{3}	130	130	115	120	135	142	90	116
CO_2 *mg/m*^{3}	2.9	4.4	3.6	4.1	3.3	5.7	4.8	7.3

(i) Show the relationship between NO_2 and CO_2 by graphical method.

(ii) Compute the correlation coefficient between NO_2 and CO_2.

(iii) Explain the relationship between NO_2 and CO_2.

(iv) Determine coefficient of determination between the pollutants and interpret the result using coefficient of determination.

[Ans: (ii) $r = -0.1522$, negative correlation; (iv) $r^2 = 0.02316 = 0.0232$ which means that the 2.32% of the changes in one pollutant is explained by the other pollutant] [TU, BE, 2067 Mangsir / 2068 Bhadra]

Partial and multiple correlation

8. On the basis of observation made on 39 cotton plans, the total correlation of field of cotton (X_1), the number of bolls i.e. seed vessels (X_2) and height (X_3) are found to be $r_{12} = 0.8$, $r_{13} = 0.65$, $r_{23} = 0.70$.
Compute the partial correlation between yields of cotton and the number bolls eliminating the effect of height. [Ans: $r_{12.3} = 0.6357$]

9. Compute the partial correlation coefficient from the following information by keeping the effect of the third variable X_3 constant.
The coefficient of correlation between X_1 and $X_2 = 0.80$
The coefficient of correlation between X_1 and $X_3 = 0.65$
The coefficient of correlation between X_2 and $X_3 = 0.70$ [Ans: 0.6365] [TU 2050]

10. A sample of 10 values of three variables X_1, X_2 and X_3 were obtained as [TU 2065 II]

$\Sigma X_1 = 10$	$\Sigma X_2 = 20$	$\Sigma X_3 = 30$
$\Sigma X_1^2 = 20$	$\Sigma X_2^2 = 68$	$\Sigma X_3^2 = 170$
$\Sigma X_1X_2 = 10$	$\Sigma X_1X_3 = 15$	$\Sigma X_2X_3 = 64$

Find (a) Partial correlation between X_1 and X_3 eliminating the effect of X_2.

(b) Multiple correlation between X_1, X_2 and X_3 assuming X_1 as independent.

[Ans: a) $r_{13.2} = 0.727$ b) $R_{1.23} = 0.767$

Regression

11. Calculate the coefficient of correlation and obtain the lines of regression for the following data:

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

[Ans : $y = 7.25 + 0.95\,x$, $x = -6.4 + 0.95\,y$] [TU, BE, 2056 Bhadra]

12. The two regression equations of the variable x and y are

$$y = 19.13 - 0.87y \quad \text{and } x = 0.64 - 0.50x$$

Find (i) Mean of x's; (ii) Mean of y's; (iii) Correlation coefficient between x and y.

[Ans: $\bar{x} = 15.935, y = 3.67, r = \sqrt{0.435} = 0.66$.]

13. Obtain the equations of the two lines of regression for the following data :

X	43	44	46	40	44	42	45	42	38	40
Y	29	31	19	18	19	27	27	29	41	30

[*Ans* : $y = 88.65 + (-1.454)\,x$, $x = 47.45 + (-0.187)y$] [*TU, BE, 2057 Bhadra/ 2067 Shrawan*]

14. The cost of manufacturing a lot of certain product depends on the lot size as shown by the following sample data.

Cost in Rs Y= y	30	70	140	270	530	1010	2500	5020
Lot size X = x	1	5	10	25	50	100	250	500

Fit a straight line to these data by the method of least squares using lot size as the independent variable.

[Ans : $y = 22.90 + 9.98\,x$] [*TU, BE, 2061 Ashwin*]

15. The following table gives the age of cars of a certain company and annual maintenance cost.

Age of the car in (Year)	2	4	6	8
Maintance cost in (Rs. 000)	10	20	25	30

Obtain the regression equation for cost related to age. And also estimate the cost of maintenance for 7 years old car.

[Ans : $y = 5 + 3.25x$; $y(7) = 27.75$][*TU, BE, 2062 Jestha*]

16. Suppose a statistics professor is interested in predicting final exam score (y) from SAT mathematics score (x), using the following data student.

Student	1	2	3	4	5	6	7	8	9
SAT score x	440	465	282	521	535	552	572	590	607
Final Score y	40	47	43	54	64	52	59	68	44

(i) Determine the regression equation for predicting scores on the final (Y) from SAT score (x)

(ii) From a SAT score of 500, predict the score on the final.

[Ans : $y = 24.88 + 0.054\,x$; $y(500) = 51.88$] [*TU, BE, 2062 Baisakh*]

17.. In some determination of the volume V of carbon dioxide dissolved in a given volume of water at different temperature θ the following pairs of value were obtained:

θ	0	5	10	15
V	1.8	1.45	1.18	1.00

Obtain by the method of least squares, relation of the form $V = a + b\theta$ which best fits of these observation. [Ans : $V = 1.758 + (-0.0543)\theta$] [*TU, BE, 2063 Ashada*]

18. Consider the following sample result, where the number of data point X is used to predict computer processing time Y (in seconds)

X= x	105	511	401	622	330	211	332	332
Y= y	44	214	193	299	143	112	155	131

Use the method of least squares to determine the expression for the estimated regression line. Determine the predicted processing time when the number of data points is 200. [Ans: $y = -2.44 + 0.461x$, $y(200) = 89.76$] [*TU, BE, 2063 Kartik*]

19. The insecticide commonly known as *DDT* has been banded in most countries due to its disastrous effect on the environment. The following data show the effect of thickness of the eggshells of certain birds. (Presence of higher level of *DDT* leads to thinner eggshells, which in turn make the eggs break prematurely, thus leading to a dwindling bird pollution)

DDT residue in yolk lipids (part/ml) (X)	65	98	102	117	122	247	393
Thickness of eggshell (Y)	0.52	0.53	0.50	0.49	0.49	0.41	0.37

Develop linear model to measure relationship between the level of *DDT* and eggshell thickness. [Ans : $y = 0.554 + (-0.0004997)x$] [*TU, BE, 2064 Poush*]

20. The report refuse derived fuel evaluation in an Industrial "Spreader–Stroker Boiler" reported the accompanying data on X = % refuse derived fuel (*RDF*) heat input and Y = % efficiency for certain boiler.

X	37	30	48	29	27	16	0	20
Y	78.0	77.2	74.4	77.7	76.9	79.0	82.1	76.5

Obtain the equation of the estimated regression line. Estimate the true % efficiency when % RDF heat input is 25.

[Ans : $y = 81.17 + (-0.133)x$, $y(25) = 77.85$] [*TU, BE, 2064 Poush*]

21. The following measurements show the respective heights in inches of 10 fathers and their eldest sons :

Height of father X	66	67	63	71	69	65	62	70	61	72
Height of son Y	65	68	66	65	70	67	67	71	62	63

Obtain the regression line of son's height on father's height and estimate the height of son when his father is found to be 70 inches high.

[Ans : $y = 54.71 + 0.176x$, $y(70) = 67.03$] [*TU, BE, 2065 Chaitra*]

22. From the following table, compute the line of regression for estimating blood pressure :

Blood pressure Y	147	125	160	118	149	128
Age in years X	56	42	72	36	63	47

[Ans : $y = 74.68 + 1.119\,x$] [*TU, BE 2065 Chaitra (Re)*]

23. An article in the Tappi Journal (March, 1986) presented data on green liquor Na_2S concentration (in gm/lit) and paper machine production (in tons per day). The data (read from graph) are shown below.

X	40	42	49	46	44	48	46	43
Y	820	830	890	870	890	910	950	960

(i) Fit a simple linear regression model y = green liquor Na_2S concentration and x = production.

(ii) Find the fitted value of y corresponding to $x = 915$ and the associated residual.

[Ans: (i) $y = 555.23 + 7.481x$; (ii) $y(915) = 7400.35$] [*TU, BE, 2066 Magh*]

24. Fit a straight line to the following data regarding Y as an independent variable

X	1	1.8	3.3	4.5	6.3
Y	0	1	2	3	4

[Ans : $y = -0.50 + 0.741x$] [*TU, BE, 2066 Magh*]

25. Observations on the yield of a chemical reaction taken at various temperatures were recorded as follows: [*TU, BE, 2067 Mangsir / 2068 Magh(Back)*]

X (℃)	150	150	200	250	250	300	150
Y(%)	75.4	81.2	85.5	89	90.5	96.7	75.4

(i) Plot the data

(ii) Does it appear from the plot as if the relationship is linear?

(iii) Fit a simple linear regression. [Ans: $y = 58.58 + 0.127x$]

26. Find and sketch or plot the sample regression line of y on x and plot the given data on the same axes.

(i) ***Ohm's law***

Voltage x [Volts]	30	30	60	60	90	90
Current [amperes]	3.1	3.2	6.3	6.5	10.0	10.1

Also find resistance [Ans: $y = -0.367 + 0.115x$, $R = 8.70$]

(ii) *Thermal conductivity of water*

Temperature x [°F]	32	50	100	150	212
Conductivity y [Btu/hr ft °F]	0.337	0.345	0.365	0.380	0.395

Also find y at room temperature 66 °F

[Ans: $y = 0.32923 + 0.00032x$, $y(66) = 0.35035$]

(iii) *Stopping distance of a passenger car*

Speed x [mph]	30	40	50	60
Stopping distance y [ft]	160	240	330	435

Also (a) find y at 35 mph (b) find 95% confidence interval for the regression coefficient (i.e., slope) [Ans: $y = -120.5 + 9.15x$ (a) $y(35) = 199.75$, (b) (7.44, 10.86)]

Multiple regressions

27. The following table shows the corresponding values of three variables x, y and z. (a) Find the linear least- squares regression equation of z on x and y. (b) Estimate z when $x = 10$ and $y = 6$

x	3	5	6	8	12	14
y	16	10	7	4	3	2
z	90	72	54	42	30	12

[Ans: (a) $\hat{z} = 61.40 - 3.65x + 2.54y$, (b) 40]

28. The following sample data were collected to determine the relationship between processing variables and the current gain of a Transistor in the integrated circuit

Diffusion time (hours) x	*Sheet resistance (Ω · cm) y*	*Current gain z*
1.5	66	5.3
2.5	87	7.8
0.5	69	7.4
1.2	141	9.8
2.6	93	10.8
0.3	105	9.1
2.4	111	8.1
2.0	78	7.2
0.7	66	6.5
1.6	123	12.6

Fit a regression plane and use its equation to estimate the expected current gain when the diffusion time is 2.2 hours and the sheet resistance is 90Ω · cm

[Ans: $\hat{z} = 2.266 + 0.225x + 0.0623y$, $\hat{z} = 8.37$]