UNIT-V

STATISTICS

Topic Learning Objectives:

Upon Completion of this unit, students will be able to:

- Expand their knowledge and skills of the Statistical Concepts and a personal development experience towards the needs of statistical data analysis.
- Understand the Central Moments, Skewness and Kurtosis.
- Describe & evaluate the concept of correlation and regression coefficients.
- Investigate the strength and direction of a relationship between two variables by collecting measurements and using appropriate statistical analysis.
- To model a linear relationship between a dependent variable and two or more independent variables.

Introduction:

In many fields of Applied Mathematics and Engineering we face some problems and do experiments involving two variables. In this chapter, we consider the Mathematical theory of statistics, by presenting an elementary treatment of Central moments, mean, variance, coefficients of skewness and kurtosis in terms of moments, curve fitting, correlation and regression. In mathematics, a moment is a specific quantitative measure of the shape of a function. It is used in both mechanics and statistics. If the function represents physical density, then the zeroth moment is the total mass, the first moment divided by the total mass is the center of mass, and the second moment is the rotational inertia. If the function is a probability distribution, then the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second central moment is the variance, the third standardized moment is the skewness, and the fourth standardized moment is the kurtosis.

Moments:

In mechanics, moment refers to the turning or the rotating effect of a force whereas it is used to describe the peculiarities of a frequency distribution in statistics. We can measure the central tendency of a set of observations by using moments. Moments also help in measuring the scatteredness, asymmetry and peakedness of a curve for a particular distribution. Moments refers to the average of the deviations from mean or some other value raised to a certain power. The arithmetic mean of various powers of these deviations in any distribution is called the moments of the distribution about mean. Moments about mean are generally used in statistics.

Moments for ungrouped data:

Now we first define the moments for ungrouped data. The r[th] moment about origin is denoted by μ_r' and defined by,

$$\mu_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \quad r = 1, 2, 3 \ldots \tag{1}$$

Here the μ_r' is the r[th] moment when we are dealing with the n observations denoted by x_1, x_2... x_n. Thus, for r =1, 2, 3 and 4 we get the first four raw moments about the origin.

$$\mu_1' = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu_2' = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \mu_3' = \frac{1}{n}\sum_{i=1}^{n} x_i^3 \text{ and } \mu_4' = \frac{1}{n}\sum_{i=1}^{n} x_i^4.$$

Similarly, we can define the r[th] moment about the arithmetic mean $\bar{x}$ or this is also called the r [th] central moment and it is denoted by the notation μ_r and it is defined as:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r, \quad r = 1, 2, 3 \ldots \tag{2}$$

Thus, for r =1, we get the first central moment about the mean as $\mu_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Similarly for r = 2, we get the second central moment about the mean as $\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ which is equal to variance.

Moments for grouped data:

Suppose we are having observations $x_1, x_2, \ldots, x_n$ which are the mid points of the class-intervals and $f_1, f_2, \ldots, f_n$ are their corresponding frequencies then the r[th] moment about origin is denoted by μ_r' and defined by,

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r, \quad r = 1, 2, 3 \ldots \text{ and } \quad N = \sum_{i=1}^{n} f_i \tag{3}$$

Similarly, the r[th] moment about arithmetic mean is denoted by μ_r and defined by,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^r, \quad r = 1, 2, 3 \ldots \tag{4}$$

Also, the r[th] moment about any point A is denoted by μ_r' and defined by,

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r, \quad r = 1, 2, 3 \ldots \tag{5}$$

Note: If $d_i = \frac{(x_i - A)}{h}$ or $d_i = \frac{(x_i - \bar{x})}{h}$, Then r[th] order moments about an arbitrary point A and mean $\bar{x}$ are defined respectively by $\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^r h^r$ & $\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i d_i^r h^r$ r = 1, 2, 3 …

Relation between raw (Moments about origin or any point) and Central Moments

The central moments can be expressed in terms of raw moments and vice-versa. The general relation between the moments about mean in terms of moments about any point is given by,

$$\mu_r = \mu_r' - {}^rC_1\, \mu_{r-1}'\, \mu_1' + {}^rC_2\, \mu_{r-2}'\, \mu_1'^2 - {}^rC_3\, \mu_{r-3}'\, \mu_1'^3 + \cdots + (-1)^r\, \mu_1'^r, \text{ r =1,2, …} \tag{6}$$

In particular, on putting r = 2, 3 and 4 in equation (6), we get

$$\mu_2 = \mu_2' - \mu_1'^2, \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \text{ and } \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4.$$

Conversely,

$\mu_r' = \mu_r + {}^rC_1\, \mu_{r-1}\, \mu_1' + {}^rC_2\, \mu_{r-2}\, \mu'^2{}_1 + \ldots + \mu'^r{}_1,$ r = 1, 2, 3 … (7)

In particular, on putting r = 2, 3 and 4 in equation (7), we get

$$\mu_2' = \mu_2 + \mu_1'^2, \mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 \text{ and } \mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu'^2{}_1 + \mu_1'^4.$$

Example 1: The first four moments of a distribution about the value 4 of the variables are -1.5, 17, -30 and 108. Find the moments about the mean.

Solution: Given A = 4, μ_1'=-1.5, μ_2'=17, μ_3'=-30 and μ_4'=108.

Moments about mean:

$\mu_2 = \mu_2' - \mu'^2{}_1 = 17 - (-1.5)^2 = 14.75$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu'^3{}_1 = -30 - 3(17)(-1.5) + 2(-1.5)^2 = 39.75$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu'^2{}_1 - 3\mu'^4{}_1$

$= 108 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125.$

Example 2: Calculate the first four moments of the following distribution about the mean.

x:	0	1	2	3	4	5	6	7	8
f:	1	8	28	56	70	56	28	8	1

Solution:

x	f	d = (x - $\bar{x}$)	fd	fd^2	fd^3	fd^4
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	∑f=N=256		**∑ fd = 0**	**∑ fd^2 = 512**	**∑ fd^3 = 0**	**∑ fd^4 = 2816**

Moments about the mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{1024}{256} = 4$ are

$$\mu_1 = \frac{\sum fd}{N} = 0,\ \mu_2 = \frac{\sum fd^2}{N} = 2,\ \mu_3 = \frac{\sum fd^3}{N} = 0,\ \mu_4 = \frac{\sum fd^4}{N} = 11$$

Example 3: Wages of workers are given in the following table:

1.5 - 2.5	2.5 - 3.5	3.5 - 4.5	4.5 - 5.5	5.5 - 6.5
1	3	7	3	4

Calculate the first four central moments of the following distribution.

x	Mid value x	f	fx	$d = x - \bar{x}$	fd	fd^2	fd^3	fd^4
1.5-2.5	2	1	2	-2	-2	4	-8	16
2.5-3.5	3	3	9	-1	-3	3	-3	3
3.5-4.5	4	7	28	0	0	0	0	0
4.5-5.5	5	3	15	1	3	3	3	3
5.5-6.5	6	4	34	2	8	16	32	64
Total		$\sum f = 18$	$\sum fx = 72$		$\sum fd = 0$	$\sum fd^2 = 26$	$\sum fd^3 = 24$	$\sum fd^4 = 86$

Mean of x values $(\bar{x}) = \frac{\sum fx}{\sum f} = 4$

First central moment $(\mu_1) = 0$

Second central moment $(\mu_2) = 1.4444$

Third central moment $(\mu_3) = 1.3333$

Fourth central moment $(\mu_4) = 4.7778$

Example 4: Wages of workers are given in the following table:

1.5 - 2.5	2.5 - 3.5	3.5 - 4.5	4.5 - 5.5	5.5 - 6.5
1	3	7	3	3

Calculate the first four central moments of the following distribution.

Class (1)	Mid value (x) (2)	f (3)	$f \cdot x$ (4) = (2) × (3)	$(x - \bar{x})$ (5)	$f \cdot (x - \bar{x})^2$ (6) = (3) × (5)	$f \cdot (x - \bar{x})^3$ (7) = (5) × (6)	$f \cdot (x - \bar{x})^4$ (8) = (5) × (7)
1.5 - 2.5	2	1	2	-2.2353	4.9965	-11.1687	24.9654
2.5 - 3.5	3	3	9	-1.2353	4.5779	-5.655	6.9856
3.5 - 4.5	4	7	28	-0.2353	0.3875	-0.0912	0.0215
4.5 - 5.5	5	3	15	0.7647	1.7543	1.3415	1.0259
5.5 - 6.5	6	3	18	1.7647	9.3426	16.4869	29.0945
---	---	---	---	---	---	---	---
--	--	$n = 17$	$\sum f \cdot x = 72$	`--	= 21.0588	= 0.9135	= 62.0928

Mean of x values ($\bar{x}$) =$\frac{\sum \mathbf{fx}}{\sum \mathbf{f}} = 4.2353$

First central moment (μ_1) = 0

Second central moment (μ_2) = 1.2388

Third central moment (μ_3) = 0.0537

Fourth central moment (μ_4) = 3.6525

Skewness and Kurtosis:

Averages tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items around a central value. These measures do not reveal whether the dispersal of value on either side of an average is symmetrical or not. If observations are arranged in a symmetrical manner around a measure of central tendency, we get a symmetrical distribution; otherwise, it may be arranged in an asymmetrical order which gives asymmetrical distribution.

Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristics of a frequency distribution. Thus, skewness is a measure that studies the degree and direction of departure from symmetry.

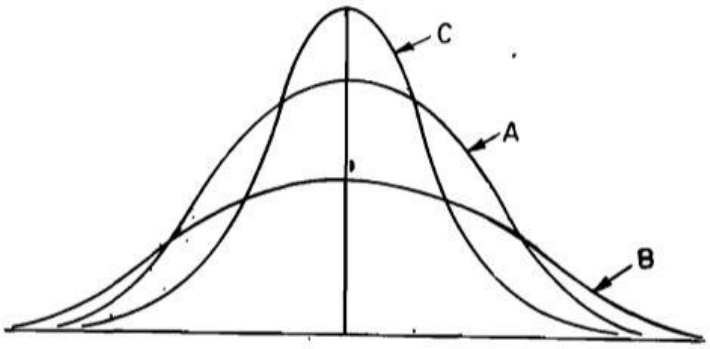
A symmetrical distribution, gives a 'symmetrical curve', where the value of mean, median and mode are exactly equal. On the other hand, in an asymmetrical distribution, the values of mean, median and mode are not equal. When two or more symmetrical distributions are compared, the difference in them is studied with 'Kurtosis'. On the other hand, when two or more symmetrical distributions are compared, they will give different degrees of Skewness. These measures are mutually exclusive i.e. the presence of skewness implies absence of kurtosis and vice-versa.

Measures of Kurtosis:

Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the Karl Pearson co-efficient β_2 and given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Kurtosis studies the concentration of the items at the central part of a series. The following figure in which all the three curves A, B and C are symmetrical about the mean.



Curve of the type 'A' which is neither flat nor peaked is called the normal curve or 'MESOKURTIC' curve ($\beta_2 = 3$). If items concentrate too much at the center (more peaked than the normal curve), the curve of the type 'C' becomes 'LEPTOKURTIC' curve ($\beta_2 > 3$).

If the concentration at the center is comparatively less (flatter than the normal curve), the curve of the type 'B' becomes 'PLATYKURTIC' curve ($\beta_2 < 3$).

Measures of Skewness:

Literally, skewness means 'lack of symmetry'. A distribution is said to be skewed if

(i) Mean, Median and Mode fall at different points.

(ii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Karl Pearson's coefficient of Skewness: The method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is as follows:

$S_k = \frac{\text{Mean - Mode}}{\sigma}$, where σ is the standard deviation of the distribution.

Based upon moments, co-efficient of skewness is defined as follows:

$$S_k = \frac{\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}, \text{ where } \beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Nature of Skewness:

Skewness can be positive or negative or zero. The direction of skewness is determined by observing whether the mean is greater than the mode (positive skewness) or less than the mode (negative skewness).

(i) When the values of mean, median and mode are equal, there is no skewness.

(ii) When mean > median > mode, skewness will be positive.

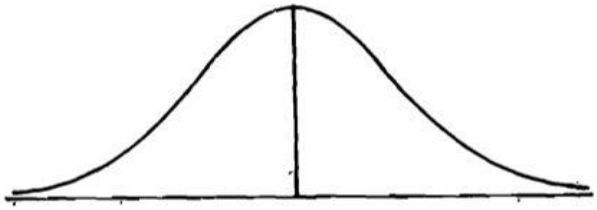
(iii) When mean < median < mode, skewness will be negative.

Characteristic of a good measure of skewness:

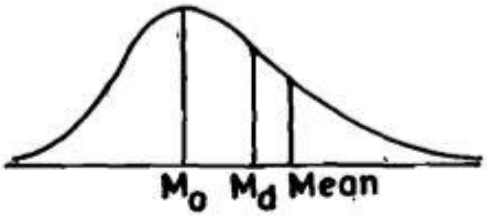
1. It should be a pure number in the sense that its value should be independent of the unit of

the series and also degree of variation in the series.

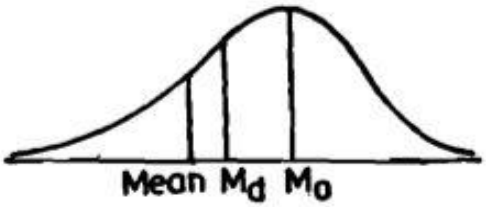
2. It should have zero-value, when the distribution is symmetrical.
3. It should have a meaningful scale of measurement so that we could easily interpret the measured value.



$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)



(Positively Skewed Distribution)



(Negatively Skewed Distribution)

Note:

From $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ ……(*) we observe the following:

- μ_3^2 is always positive whether μ_3 is positive or negative.
- μ_2^3 is always positive as μ_2 is variance.

∴ from (*) β_1 is always positive which is not so always as skewness may be negative also.

To overcome this, the measure of skewness is defined by

$$\gamma_1 = \pm\sqrt{\beta_1}$$

Here sign of γ_1 depends on the sign of μ_3.

Similarly, the measure of kurtosis is defined by $\gamma_2 = \beta_2 - 3$.

Example 5: Wages of workers are given in the following table:

10-12	12-14	14-16	16-18	18-20	20 - 22	22 - 24
1	3	7	12	12	4	3

Calculate the first four central moments of the following distribution. Also compute β_1 and β_2.

Solution:

Wages	f	Mid-point x	d = (x -17) / 2	fd	fd^2	fd^3	fd^4
10-12	1	11	-3	-3	9	-27	81
12-14	3	13	-2	-6	12	-24	48
14-16	7	15	-1	-7	7	-7	7
16-18	12	17	0	0	0	0	0
18-20	12	19	1	12	12	12	12
20-22	4	21	2	8	16	32	64
22-24	3	23	3	9	27	81	243
				$\sum$ = 13	**$\sum$ = 83**	**$\sum$ = 67**	**$\sum$ =455**

$$\mu_1' = \frac{\sum fd}{N} \times h = 0.52,\ \mu_2' = \frac{\sum fd^2}{N} \times h^2 = 2.16,\ \mu_3' = \frac{\sum fd^3}{N} \times h^3 = 10.72,$$
$$\mu_4' = \frac{\sum fd^4}{N} \times h^4 = 145.6$$

Moments about mean:

$\mu_1 = 0, \mu_2 = \mu_2' - \mu'^2_1 = 2.16 - 0.2704 = 1.8896$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu'^3_1 = 10.72 - 3(2.16)(0.52) + 2(0.52)^2 = 7.491$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu'^2_1 - 3\mu'^4_1 = 145.6 - 4(0.52)(10.72) + 6(2.56)(0.52)^2 - 3 \times 0.07312$
$= 126.5874.$

So, we have $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 8.317,\ \beta_2 = \frac{\mu_4}{\mu_2^2} = 35.4527.$

Exercise:

1. The first four raw moments of a distribution are 2, 136, 320 and 40,000. Find the coefficients of skewness and kurtosis.

Ans. $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.0904,\ \beta_2 = \frac{\mu_4}{\mu_2^2} = 2.333.$

2. Find the second, third and fourth central moments of the frequency distribution given below. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

Class limits	Frequency
110 – 115	5
115 – 120	15
120 – 125	20
125 – 130	35
130 – 135	10
135 – 140	10
140 – 145	5

Ans.

$\mu_2 = 54, \mu_3 = 100.5, \mu_4 = 7827$

$\gamma_1 = \sqrt{\beta_1} = \sqrt{0,0641} = 0.2532; \gamma_2 = \beta_2 - 3 = -0.3158$

3. Find the second, third and fourth central moments of the frequency distribution given below. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

5	10	15	20	25	30	35
4	10	20	36	16	12	2

Ans.

$$\mu_2 = 44.41,\ \mu_3 = -12.504,\ \mu_4 = 5423.5057,\ \beta_1 = 0.001785,$$
$$\beta_2 = 2.7499,\ \gamma_1 = \sqrt{\beta_1} = 0.25298;\ \gamma_2 = \beta_2 - 3 = -0.317.$$

4. Compute the first four moments about mean from the following data. Hence, find (i) a measure of skewness and (ii) a measure of kurtosis.

Class Intervals:	0 -10	10 – 20	20 – 30	30 – 40
Frequency:	1	3	4	2

Ans.

$$\mu_1 = 0,\ \mu_2 = 81,\ \mu_3 = -144,\ \mu_4 = 14817,\ \beta_1 = 0.03902,$$
$$\beta_2 = 0.01909,\ \gamma_1 = \sqrt{\beta_1} = 0.1975;\ \gamma_2 = \beta_2 - 3 = -2.9809.$$

Correlation and Regression:

The word correlation is used in everyday life to denote some form of association. In statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarize the association.

Correlation:

Correlation means simply a relation between two or more variables.

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other.

Ex: 1. x: supply y: price

2. x: demand y: Price.

Positive correlation:

If **an** increase or decrease in one variable corresponds to an increase or decrease in the other then the correlation is said to be positive correlation or direct correlation.

Ex: 1. Demand and price of commodity. 2. Income and expenditure.

Negative correlation:

If an increase or decrease in one variable corresponds to an decrease or increase in the other then the correlation is said to be negative correlation or inversely correlated.

Ex: 1.Supply and Price of a commodity.

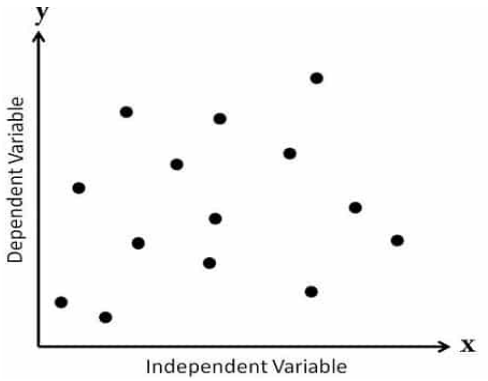
2. Correlation between Volume and pressure of a perfect gas.

No correlation

If there exist no relationship between two variables then they are said to be non correlated.

Scatter diagram

To obtain a measure of relationship between two variables x and y we plot their corresponding values in the xy - plane. The resulting diagram showing the collection of the dots is called the dot diagram or scatter diagram.



Correlation Coefficient (Karl Pearson correlation coefficient)

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Karl Pearson's correlation coefficient and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

Let $x_1, x_2, x_3, \ldots\ldots, x_n$ be n values of x and $y_1, y_2, y_3, \ldots\ldots y_n$ be the corresponding n values of y, then the coefficient of correlation between x and y is

$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$, where ${\sigma_x}^2$ - variance of the x series, ${\sigma_y}^2$- variance of the y series, $\bar{x} = \frac{\sum x}{n}$

→ Mean of the x series $\bar{y} = \frac{\sum y}{n}$ → mean of the y series.

For computation purpose we can use the formula

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

Limits for correlation coefficient

The coefficient of correlation numerically does not exceed unity ($-1 \le r \le 1$).

Proof:

We have $r = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}}$, i =1,2,……n,

Taking $x_i - \bar{x} = a_i$ and $y_i - \bar{y} = b_i$

$$r = \frac{\frac{1}{n}\sum a_i b_i}{\sqrt{\frac{1}{n}\sum {a_i}^2}\sqrt{\frac{1}{n}\sum {b_i}^2}} \quad r^2 = \frac{(\sum a_i b_i)^2}{\sum {a_i}^2 \ \sum {b_i}^2}. \qquad (1)$$

By Schwartz inequality, which states that if a_i, b_i i =1, 2… n are real quantities then

$(\sum a_i b_i)^2 \leq \sum {a_i}^2 \ \sum {b_i}^2$ and the sign of equality holding if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3} =$ …… $= \frac{a_n}{b_n}$.

Using this equation (1) becomes $r^2 \leq 1$,

$$\Rightarrow |r| \leq 1,$$
$$\Rightarrow -1 \leq |r| \leq 1.$$

Hence correlation coefficient cannot exceed unity numerically.

Note:

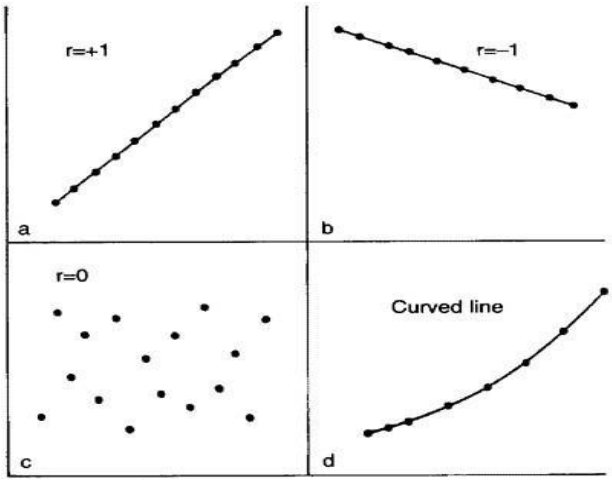


Figure 1.1 Correlation illustrated.

1. If r =-1 there is a perfect negative correlation.
2. If r =1 there is a perfect positive correlation.
3. If r =0 then the variables are non-correlated.

RANK CORRELATION

In many practical situations, characters are not measurable.

They are qualitative characteristics and individuals or items can be ranked in order of their merits. This type of situation occurs when we deal with the qualitative study such as honesty, beauty, voice, etc. For example, contestants of a singing competition may be ranked by judge according to their performance. In another example, students may be ranked in different subjects according to their performance in tests.

Arrangement of individuals or items in order of merit or proficiency in the possession of a certain characteristic is called ranking and the number indicating the position of individuals or items is known as rank.

If ranks of individuals or items are available for two characteristics then correlation between ranks of these two characteristics is known as rank correlation.

With the help of rank correlation, we find the association between two qualitative characteristics. As we know that the Karl Pearson's correlation coefficient gives the intensity of linear relationship between two variables andSpearman's rank correlation coefficient gives the concentration of association between two qualitative characteristics. In fact, Spearman's rank correlation coefficient measures the strength of association between two ranked variables.Derivation of the Spearman's rank correlation coefficient formula is discussedin the following section.

RANK CORRELATION COEFFICIENT FORMULA

Suppose we have a group of n individuals and let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be the ranks of n individuals in characteristics A and B respectively. Then rank correlation coefficient r_s is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

Here d_i is difference between ranks assigned in characteristics A and B. and n is number of pairs of data.

This formula was given by Spearman and hence it is known as Spearman'srank correlation coefficient formula.

Note 1: When two or more observations have equal values, if there is a tie, it is difficult to assign ranks to them. In such cases, the observations are given the average of the ranks they would have received. Then, a different formula is used to calculate the rank correlation coefficient.

The Spearman's correlation coefficient for tied ranks can be calculated using the formula

$$r_s = 1 - \frac{6\left[\sum_{i=1}^{n} {d_i}^2 + \frac{1}{12}[(m_1^3 - m_1) + (m_2^3 - m_2) + (m_3^3 - m_3) + \cdots]\right]}{n(n^2-1)}$$

Where $m_1, m_2, \ldots\ldots$ are number of repetitions of ranks and $\frac{1}{12}\sum(m_i^3 - m_i)$ are the corresponding correction factors.

Note 2: r_s lie between -1 and 1.

Examples:

1. If r is the correlation coefficient between x and y and z= ax+by. Show that

   $r = \frac{{\sigma_z}^2 - (a^2{\sigma_x}^2 + b^2{\sigma_y}^2)}{2ab\sigma_x\sigma_y}$.

   Solution: Let z = ax + by $\Rightarrow \frac{1}{n}\sum z = \frac{a}{n}\sum x + \frac{b}{n}\sum y \Rightarrow \bar{z} = a\bar{x} + b\bar{y}$,

   $$\frac{1}{n}\sum(z-\bar{z})^2 = a^2\frac{1}{n}\sum(x-\bar{x})^2 + b^2\frac{1}{n}\sum(y-\bar{y})^2 + 2ab\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}),$$

   $$\Rightarrow \quad {\sigma_z}^2 = a^2{\sigma_x}^2 + b^2{\sigma_y}^2 + 2ab r\sigma_x\sigma_y,$$

$\Rightarrow$ r = $\frac{\sigma_z^2-(a^2\sigma_x^2+b^2\sigma_y^2)}{2ab\sigma_x\sigma_y}$.

2. While calculating the correlation coefficient between x and y from 25 pairs of observations a person obtained the following values. $\sum x_i = 125, \sum {x_i}^2 = 650$, $\sum y_i = 100, \sum {y_i}^2 = 460, \sum x_i y_i = 508$. It was later discovered that he had copied down the pairs (8,12) and (6,8) as (6,12) and (8,6) respectively. Obtain the correct value of the correlation coefficient.

Solution: To get the correct values, we subtract the incorrect values and add the corresponding correct values.

Therefore, correct values of sums

$\sum x_i = 125 - 6 - 8 + 8 + 6 = 125$, $\sum {x_i}^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$,

$\sum y_i = 100 - 12 - 6 + 12 + 8 = 102$, $\sum {y_i}^2 = 460 - 12^2 - 6^2 + 12^2 + 8^2 = 488$,

and $\sum x_i y_i = 508 - (6X12) - (8X6) + (8X12) + (6X8) = 532$,

given n = 25,

$r = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{\{n\sum x^2-(\sum x)^2\}\{n\sum y^2-(\sum y)^2\}}}$ = 0.51912.

3. The following Table gives the age (in years) of 10 married couples. Calculate the coefficient of correlation between these ages.

Age of Husband(x)	23	27	28	29	30	31	33	35	36	39
Age of wife(y)	18	22	23	24	25	26	28	29	30	32

Solution: Here n=10, we find $\bar{x} = \frac{1}{n}\sum x_i = \frac{311}{10} = 31.1$ $\bar{y} = \frac{1}{n}\sum y_i = \frac{257}{10} = 25.7$.

x_i	y_i	$X_i = x_i - \bar{x}$	${X_i}^2$	$Y_i = y_i - \bar{y}$	${Y_i}^2$	X_iY_i
23	18	-8.1	65.61	-7.7	59.29	62.37
27	22	-4.1	16.81	-3.7	13.69	15.17
28	23	-3.1	9.61	-2.7	7.29	8.37
29	24	-2.1	4.41	-1.7	2.89	3.57
30	25	-1.1	1.21	-0.7	0.49	0.77
31	26	-0.1	0.01	0.3	0.09	-0.03
33	28	1.9	3.61	2.3	5.29	4.37
35	29	3.9	15.21	3.3	10.89	12.87
36	30	4.9	24.01	4.3	18.49	21.07
39	32	7.9	62.41	6.3	39.69	49.77
			$\sum {X_i}^2 = 202.9$	$\sum {Y_i}^2 = 158.10$		$\sum X_iY_i$=178.3

$$r = \frac{\sum X_i Y_i}{\sqrt{\sum {X_i}^2 \sum {Y_i}^2}} = 0.9955 \approx 1.$$

i.e, the ages of husbands and wives are almost perfectly correlated.

4. Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extentthe knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

Let us denote the rank of students in Statistics by R_X and rank in Mathematicsby R_y. For the calculation of rank correlation coefficient, we have to find $\sum_{i=1}^{n} {d_i}^2$ which is obtained through the following table:

Rank in Statistics (Rx)	**Rank in Mathematics (Ry)**	**Difference of Ranks $d_i = R_x - R_y$**	${d_i}^2$
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum_{i=1}^{8} {d_i}^2$ =22

Here, n = number of paired observations = 8

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)} = 1 - \frac{6X22}{8X63} = 0.74$$

Thus, there is a positive association between ranks of Statistics andMathematics.

5. Suppose we have ranks of 5 students in three subjects Computer,Physics and Statistics and we want to test which two subjects have the same trend.

Rank in Computer	2	4	5	1	3
Rank in Physics	5	1	2	3	4
Rank in Statistics	2	3	5	4	1

Solution: In this problem, we want to see which two subjects have same trend i.e., which two subjects have the positive rank correlation coefficient.

Here we have to calculate three rank correlation coefficients

r_{12s} = Rank correlation coefficient between the ranks of Computer and Physics

r_{23s} = Rank correlation coefficient between the ranks of Physics and Statistics

r_{13s}= Rank correlation coefficient between the ranks of Computer andStatistics

Let R_1 , R_2 and R_3 be the ranks of students in Computer, Physics and Statistics respectively.

Rank in Computer(R_1)	**Rank in Physics (R_2)**	**Rank in Statistics (R_3)**	d_{12} = R_1-R_2	d_{12}^2	d_{23}= R_2-R_3	d_{23}^2	d_{13}= R_1-R_3	d_{13}^2
2	5	2	−3	9	3	9	0	0
4	1	3	3	9	−2	4	1	1
5	2	5	3	9	−3	9	0	0
1	3	4	−2	4	−1	1	−3	9
3	4	1	−1	1	−3	9	2	4
Total				32		32		14

Thus $\sum {d_{12}}^2 = 32$, $\sum {d_{23}}^2 = 32$, $\sum {d_{13}}^2 = 14$

Now,

$r_{12s} = 1 - \frac{6\sum {d_{12}}^2}{n(n^2-1)}$ =-0.6

$r_{23s} = 1 - \frac{6\sum {d_{23}}^2}{n(n^2-1)}$ =-0.6

$r_{13s} = 1 - \frac{6\sum {d_{13}}^2}{n(n^2-1)}$ =-0.3

r_{12s} is negative which indicates that Computer and Physics have opposite trend. Similarly, negative rank correlation. r_{23s} shows the opposite trend in Physics and Statistics. r_{13s} = 0.3 indicates that Computer and Statistics have same trend.

Sometimes we do not have rank but actual values of variables are available. If we are interested in rank correlation coefficient, we find ranks from the given values. Considering this case we are taking a problem and try to solve it.

Example 6: Calculate rank correlation coefficient from the following data:

x	78	89	97	69	59	79	68
y	125	137	156	112	107	136	124

Solution: We have some calculation in the following table:

x	y	Rank of x (R_x)	Rank of y (R_y)	d = R_x-R_y	d^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	-1	1
59	107	7	7	0	0
79	136	3	3	0	0
68	124	6	5	1	1
					$\sum_{i=1}^{n} {d_i}^2 = 2$

Assign the rank to x-series and y-series in descending order of the given numbers i.e. in x-series maximum is 97, assign rank 1, next number is 89, assign rank 2 and so on. Similarly rank can be assigned for the y-series.

Spearman's Rank correlation formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)} = 1 - \frac{6 \times 2}{7(49-1)} = 0.96$$

Example 7: Calculate rank correlation coefficient from the following data:

x	81	78	73	73	69	68	62	58
y	10	12	18	18	18	22	20	24

Solution: We have some calculation in the following table:

x	y	Rank of x (R_x)	Rank of y (R_y)	d = R_x-R_y	d^2
81	10	1	8	7	49
78	12	2	7	5	25
73	18	3.5	5	1.5	2.25
73	18	3.5	5	1.5	2.25
69	18	5	5	0	0
68	22	6	2	-4	6
62	20	7	3	-4	16
58	24	8	1	-7	49
					$\sum_{i=1}^{n} {d_i}^2 = 159.50$

Spearman's Rank correlation formula is

$$r_s = 1 - \frac{6\{\sum_{i=1}^{n} {d_i}^2 + \frac{1}{12}[\,(m_1^3 - m_1) + (m_2^3 - m_2)]\}}{n(n^2-1)}$$

Where $m_1 = 2$ (the two items of x have equal value i,e. 73) and $m_2 = 3$ (three items of y having value i,e 18)

$$= 1 - \frac{6 \times \left[159.50 + \frac{1}{12}\left((8-2) + (27-3)\right)\right]}{8(64-1)} = -0.9286$$

Regression:

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of data.

Line of regression:

Line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. So the line of regression is the line of best fit.

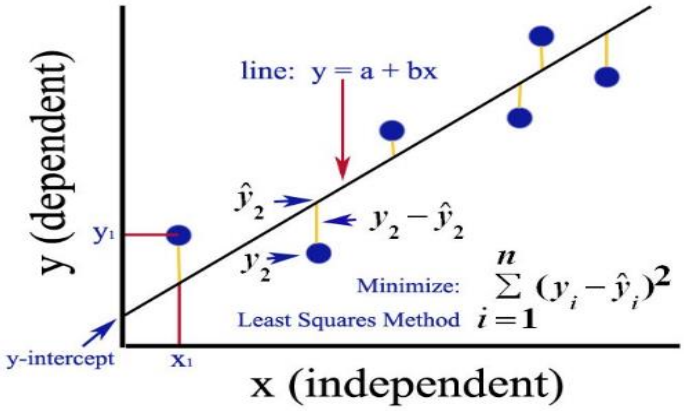
Method of Least squares:

Suppose we are given n values of $x_1, x_2, x_3, \ldots, x_n$ of an independent variable x and the corresponding values $y_1, y_2, y_3, \ldots, y_n$ of a variable y depending on x. Then the pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ give us n- points in the xy-plane. Generally, it is not possible to find the actual curve y = f(x) that passes through these points. Hence, we try to find a curve that serves as best approximation to the curve y = f(x). Such a curve is referred to as the curve of best fit. The process of determining a curve of best fit is called curve fitting. A method to find curve of best fit is called method of least squares.

The method of least squares tells that the curve should pass as closely as possible to meet all the points. Let y = f(x) be an approximate relation that fits into the data (x_i, y_i) then y_i are

called observed values $Y_i = f(x_i)$ are called the expected values. The expected values $E_i = y_i - Y_i$ are called the estimated error or residuals.

The method of least squares provides a relationship y = f(x) such that sum of the squares of the residues is least. Such a curve is known as least square curve.



Regression line of y on x:

Let regression line of y on x be y = a + bx.

The normal equations by the method of least squares is

$\sum y = na + b\sum x,$

$\sum xy = a\sum x + b\sum x^2,$

$\frac{1}{n}\sum y = a + \frac{b}{n}\sum x.$

$\bar{y} = a + b\bar{x}$ is the regression line passing through $((\bar{x}, \bar{y})$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(XY)}{\sum X^2} = \frac{\sum(XY)}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x},$$

$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \Rightarrow Y = b_{yx}X$ is the regression line of y on x.

Regression line of x on y:

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \Rightarrow X = b_{xy}Y$$

Note:

1. Regression coefficient of y on x

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

2. Regression coefficient of x on y

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

Properties of Lines of Regression (Linear Regression)

1. The two regression lines x on y and y on x always intersect at their means $(\bar{x}, \bar{y})$.
2. If θ is the angle between two regression lines then $tan\,\theta = \frac{1-r^2}{r}\,\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$.

(i) When r = 0 (the variables are independent), $tan\,\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$. Then the two lines of regression are perpendicular to each other.

(ii) When $r = \pm 1$(variables are perfectly correlated), $\theta = 0$ or π. Then the lines of regression coincide.

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

Note: If b_{yx} and b_{xy} both are positive then r is positive. Similarly if b_{yx} and b_{xy} both are negative then r is negative.

Examples:

1. If two regression equations of the variables x and y are x = 19.13 – 0.87y, y = 11.6 – 0.5x, find

(a) mean of x

(b) mean of y

(c)The correlation coefficient between x and y.

Soln: Since $\bar{x}$ and $\bar{y}$ lie on two regression lines,

$$\bar{x} = 19.13 - 0.87\bar{y},\ \bar{y} = 11.64 - 0.5\bar{x},$$

Solving we get $\bar{x} = 15.99, \bar{y} = 3.6$.

$b_{yx} = -0.5, b_{xy} = -0.87, r = \sqrt{-0.5 \times -0.87} = -0.66$ (since b_{yx} and b_{xy} have negative sign)

2. In the following table data is showing the test scores made by salesman on an intelligent test and their weekly sales.

Test scores(x)	1	2	3	4	5	6	7	8	9	10
sales(y)	2.5	6	4.5	5	4.5	2	5.5	3	4.5	3

Calculate the regression line of sales on test scores and estimate the most possible weekly volume if a sales man scores 70.

Soln:

$\bar{x} = 5.5,\ \bar{y} = 4.05$, Regression line of y on x is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$,

y = 0.06x + 0.45.

When x = 70, y = 4.65.

3. In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are legible. Variance of x=9, Regression equations 8x -10y + 66 = 0, 40x - 18y = 214. What are (i) the mean values of x and y

(ii) the correlation coefficient between x and y

(iii) the standard deviation of y.

Soln:(i) Since both the lines of regression pass through the point $(\bar{x},\bar{y})$

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

$40\bar{x}$ -18$\bar{y}$ - 214 = 0.

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

(ii) $\sigma_x{}^2 = 9.$ therefore, $\sigma_x = 3$

Let 8x - 10y + 66 = 0 and 40x - 18y = 214 be the lines of regression of y on x and x on y respectively

$b_{yx} = \frac{4}{5}$,$b_{xy} = \frac{18}{40} = \frac{9}{20}$, Hence $r^2 = b_{yx}\, b_{xy} = \frac{9}{25}$, $r = \pm\frac{3}{5} = \pm 0.6$.

Since both the regression coefficients positive we take r = 0.6.

Standard deviation of y = 4.

4. The following table gives the stopping distance y in meters of a motor bike Moving at a speed of x Kms/hour when the breaks are applied

x	16	24	32	40	48	56
y	0.39	0.75	1.23	1.91	2.77	3.81

Find the correlation coefficient between the speed and the stopping distance, and the equations of regression lines. Hence estimate the maximum speed at which the motor bike could be driven if the stopping distance is not to exceed 5 meters.

Soln:

$\bar{x} = 36,\ \bar{y} = 1.81,\ , \sigma_x = 13.663, \sigma_y = 1.1831,$

$b_{yx} = 0.0851, b_{xy} = 11.352,$

$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}}$ =0.983.

The equation of the line of regression of y on x is y = 0.0851x - 1.2536 (i)

and the equation of the line of regression of x on y is x =11.352y + 15.453. (ii)

For y = 5, equation (ii) gives x = 72.213.

Accordingly, for the stopping distance not to exceed 5 meters, the speed must not exceed 72 Kms/hour.

Multivariate Regression Analysis using least squares estimation of the parameters

When several independent variables are used to estimate the value of the dependent variable it is called multiple regression. The multiple linear regression model is just an extension of the simple linear regrssion model. In simple linear regression, we used "x" to represent the explanatory variable. In multiple linear regression, we will have more than one explanatory variable.

Let an experiment be conducted n times, and the data is obtained as follows:

Observation number	Response Y	Explanatory variables $X_1 \quad X_2 \quad \dots \quad X_k$
1	y_1	$x_{11} \quad x_{12} \dots x_{1k}$
2	y_2	$x_{21} \quad x_{22} \dots x_{2k}$
⋮	⋮	⋮ ⋮ ⋱ ⋮
n	Y_n	$x_{n1} \quad x_{n2} \dots x_{nk}$

Assuming that the model is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \ ,$$

where, y is an observed value of variable for a particular observation in the population. $\beta_0, \beta_1, \beta_2 \dots \beta_k$ are parameters which are to be determined.

the n-tuples of observations are also assumed to follow the same model. Thus they satisfy

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k}$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k}$$
$$\vdots \qquad\qquad \vdots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} \ .$$

These n equations can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik}, \quad i = 1, \dots, n.$$
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Using least squares principle, we get the following normal equations:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2} + \cdots \beta_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots \beta_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

.....

.....

$$\beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \beta_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots \beta_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

Solving the above normal equations, we get the values of $\beta_0, \beta_1, \beta_2 \dots \dots \beta_k$.

Example:

1. A company produces two different items *A* and *B*. The data below shows the sale of these items in one day and the profit made by the company on that day.

x_1*(Sales of item A)*	8	11	9	8	6	10	7
x_2 *(Sales of item B)*	6	4	5	7	1	1	0
Profit (y)	93.26	89.76	60.78	79.34	28.23	75.83	32.74

Fit the best multilinear model that represents the relationship between sales of A and B and the profit.

Solution: The normal equations corresponding to the regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ are:

$$n\beta_0 + \beta_1 \sum x_1 + \beta_2 \sum x_2 = \sum y$$
$$\beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 = \sum y x_1$$
$$\beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 = \sum y x_2$$

	x_1	x_2	y	x_1^2	x_2^2	$x_1 * y$	$x_2 * y$	$x_1 * x_2$
	8	6	93.26	64	36	746.08	559.56	48
	11	4	89.76	121	16	987.36	359.04	44
	9	5	60.78	81	25	547.02	303.9	45
	8	7	79.34	64	49	634.72	555.38	56
	6	1	28.23	36	1	169.38	28.23	6
	10	1	75.83	100	1	758.3	75.83	10
	7	0	32.74	49	0	229.18	0	0
$\sum$ =	59	24	459.94	515	128	4072.04	1881.94	209

$$7\beta_0 + 59\beta_1 + 24\beta_2 = 459.94$$
$$59\beta_0 + 515\beta_1 + 209\beta_2 = 4072.04$$
$$24\beta_0 + 209\beta_1 + 128\beta_2 = 1881.94$$

$$\beta_0 = -28.5193, \beta_1 = 9.0031, \beta_2 = 5.3496$$
$$=> \quad y = -28.5193 + 9.0031x_1 + 5.3496x_2$$

2. A set of experimental runs was made to determine a way of predicting cooking time y at various values of oven width x_1 and flue temperature x_2. The coded data were recorded as follows:

y	6.40	15.05	18.75	30.25	44.85	48.94	51.55	61.50	100.44	111.42
x_1	1.32	2.69	3.56	4.41	5.35	6.20	7.12	8.87	9.80	10.65
x_2	1.15	3.40	4.10	8.75	14.82	15.15	15.32	18.18	35.19	40.40

Estimate the multiple linear regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Solution:

The normal equations corresponding to the regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ are:

$$n\beta_0 + \beta_1 \sum x_1 + \beta_2 \sum x_2 = \sum y$$
$$\beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 = \sum y x_1$$
$$\beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 = \sum y x_2$$

For the given data

$$n = 10, \sum x_1 = 59.97, \sum x_1^2 = 446.9965, \sum y = 489.15, \sum y x_1 = 3875.9365$$
$$\sum x_2 = 156.46, \sum x_2^2 = 3991.1208, \sum y x_2 = 11749.8781, \sum x_1 x_2 = 1282.5215$$

Substituting these values in the above normal equations and solving we get,

$$\beta_0 = 0.58, \beta_1 = 2.7122, \beta_2 = 2.0497$$

Hence the required multiple linear regression equation is

$$y = 0.58 + 2.7122 x_1 + 2.0497 x_2$$

Exercise:

1. If the coefficient of correlation between the variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$. Find the ratio of the standard deviation of x and y.
   Ans. $\frac{\sigma_x}{\sigma_y} = \frac{1}{2}$ or . $\frac{\sigma_x}{\sigma_y} = \frac{2}{1}$.
2. Prove the following formulas for the coefficient of correlation r (in the usual notation)
   a) $r = 1 - \frac{1}{2n}\sum\left(\frac{X_i}{\sigma_x} - \frac{Y_i}{\sigma_y}\right)^2$, $r = -1 + \frac{1}{2n}\sum\left(\frac{X_i}{\sigma_x} + \frac{Y_i}{\sigma_y}\right)^2$.
3. Find the rank correlation coefficient for the following data:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

4. Ten participants in a contest are ranked by two judges as follows:

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

   Calculate the rank correlation coefficient.

5. The following table shows the ages x and the systolic pressures of 12 persons.

Age (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood Pressure (**y**)	147	125	160	118	149	128	150	145	115	140	152	155

lculate the coefficient of correlation between x and y. Estimate the blood pressure of a person whose age is 45 years.

Ans. r = 0.8961, y = 80.78 + 1.138 x , when x = 45, y = 132.

6. The height (inches) and weight (pounds) of baseball players are given below:

(76, 212), (76, 224), (72, 180), (74, 210), (75, 215), (71, 200), (77, 235), (78, 235), (77, 194), (76, 185).

(i) Estimate the coefficient of correlation between weight and height of baseball players.

(ii) Find the regression line between weight and height. Use the regression equation to find the weight of a baseball player that is 68 inches tall.

Ans. r = 0.5529, y = 4.737 x – 147.227, x = 0.064 y + 61.712, when x = 68, y = 97.37.

7. The equations of regression lines of two variables x and y are 4 x – 5y + 33 = 0 and 20x - 9y = 107, Find the correlation coefficient and the means of x and y.

Ans. r = 0.6, Mean of x = 13 and Mean of y = 17.

8. If the tangent of the angle between the lines of regression of y on x and x on y is 0.6 and the standard deviation of y is twice the standard deviation of x. find the coefficient of correlation between x and y.

Ans. r = 0.5.

9. The chemistry grade, intelligence test score and number of classes missed data of 12 students are given.

Chemistry grade (y)	85	74	76	90	85	87	94	98	81	91	76	74
Test score(x_1)	65	50	55	65	55	70	65	70	55	70	50	55
Classes missed(x_2)	1	7	5	2	6	3	2	5	4	3	1	4

a) Fit the best multilinear model that represents the relationship of the form
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and missed 4 classes

10. An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final weight(y)	95	77	80	100	97	70	50	80	92	84
Initial weight(x_1)	42	33	33	45	39	36	32	41	40	38
Feed weight(x_2)	272	226	259	292	311	183	173	236	230	235

a) Fit the best multilinear model that represents the relationship of the form
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

b) Predict the final weight of an animal having an initial weight of 35 kilograms that is given 250 kilograms of feed.

Resources:

1. https://nptel.ac.in/courses/111105042/
2. http://www.nptelvideos.in/2012/12/regression-analysis.html
3. https://nptel.ac.in/courses/111104074/