Unit-V : Correlation and Regression

Principle of least squares:

Introduction: In many branches of applied mathematics, it is required to express a given data, obtained from observations, in the form of a Law connecting the two variables involved. Such a Law inferred by some scheme is known as Empirical Law. For example, it may be desired to obtain the law connecting the length and the temperature of a metal bar. At various temperatures, the length of the bar is measured. Then, an Empirical Law represents the relationship existing between temperature and length for the observed values. This relation can be used to predict the length at an arbitrary temperature. Curve Fitting is one of the best methods to obtain such an Empirical Law.

Scatter Diagram: To find a relationship between the set of paired observations x and y (say), we plot their corresponding values on the graph taking one of the variables along the X-axis and other along the Y-axis i.e., (x_1, y_1), (x_2, y_2), …, (x_n, y_n). The resulting diagram showing a collection of dots is called a *scatter diagram*. A smooth curve that approximates the above set of points is known as the *approximating curve*.

Curve Fitting: Using the method of Scatter diagram, several equations of different types can be obtained to express the given data approximately. But the problem is to find the equations of the curve of '*best fit*' which may be suitable for predicting the unknown values. The process of finding such an equation of 'best fit' is known as *curve-fitting*.

If there are n pairs of observed values in the given data, then curve fitting uses the principle of least squares to find a unique curve of 'best fit'.

Principle least squares: The principle of least squares says that, out of all the curves approximating the given set of data points i.e., (x_1, y_1), (x_2, y_2), …, (x_n, y_n), the curve having the property that $d_1^2+d_2^2+...+d_n^2$ is a minimum is the best-fitting curve. Here, d_i is the difference between x_i and y_i.

Fitting of Curves: (Working Procedures)

a. Straight Line: To fit the straight line $y=a+bx$

i. Substitute the observed set of n values in this equation.

ii. Form the normal equations for each constant

i.e, form the equations: $\sum y=na+b\sum x$ and $\sum xy=a\sum x+b\sum x^2$

iii. Solve these equations as simultaneous equations for a and b

iv. Substitute the values of a and b in the equation $y=a+bx$, which is the required line of 'best fit'.

b. Parabola: To fit the parabola $y=ax^2+bx+c$

i. Substitute the observed set of n values in this equation.

ii. Form the normal equations for each constant

i.e, form the equations:

$$\sum y=a\sum x^2+b\sum x+nc$$

$$\sum xy=a\sum x^3+b\sum x^2+c\sum x$$

and

$$\sum x^2y=a\sum x^4+b\sum x^3+c\sum x^2$$

iii. Solve these equations as simultaneous equations for a, b and c

iv. Substitute the values of a, b and c in the equation $y=ax^2+bx+c$, which is the required line of 'best fit'.

c. **Exponential Curve:** To fit the Curve of the form $y = ae^{bx}$

i. Take logarithm on both sides of this equation, we get, $\ln y = \ln a + bx$

ii. Assuming $Y = \ln y$ and $A = \ln a$, the given equation reduces to $Y = A + bx$

iii. Substitute the observed set of n values in this equation of straight line and form the normal equations, $\sum Y = nA + b\sum x$ and $\sum xY = A\sum x + b\sum x^2$

iv. Solve these equations as simultaneous equations for *A* and *b* and hence calculate *a* and *b.*

v. Substitute the values of *a* and *b* in the equation $y = ae^{bx}$, which is the required line of 'best fit'.

d. **Power Curve:** To fit the Curve of the form $y = ax^b$

i. Take logarithm on both sides of this equation, we get, $\ln y = \ln a + b\ln x$

ii. Assuming $Y = \ln y$, $A = \ln a$ and $X = \ln x$ the given equation reduces to $Y = A + bX$

iii. Substitute the observed set of n values in this equation and form the normal equations, $\sum Y = nA + b\sum X$ and $\sum XY = A\sum X + b\sum X^2$

iv. Solve these normal equations as simultaneous equations for *A* and *b* and hence calculate *a* and *b.*

v. Substitute the values of *a* and *b* in the equation $y = ax^b$, which is the required line of 'best fit'.

Correlation and Regression:

Introduction:

Very often, we have to deal with the situations where more than one variables are involved. For example, we may like to study the relationship between the heights and weights of adult males, quantum of rainfall and the yield of wheat in India over a number of years, doses of drug and a response vi2.a dose of insulin and blood sugar levels in a person, the age of individuals and their blood pressure, etc.

In such situations, our main purpose is to determine whether or not a relationship exists between the two variables. If such a relationship can be expressed by a mathematical formula, then we shall be able to use it for an analysis and hence make certain predictions.

Correlation and regression are methods that deal with the analysis of such relationships between various variables and possible predictions. In this unit, we shall confine ourselves to analysing the linear relationship between two variables. However, we can extend the methods for two variables to the situations where more than two variables are studied simultaneously.

Learning Outcomes:

After reading this, you should be able to

1. Describe the correlation between two variables
2. Compute and interpret correlation coefficient
3. Describe simple linear regression line
4. Explain how to fit a linear regression line using least squares method.

Correlation and Scatter Diagram:

In studying the linear relationship between two variables, we try to examine the question "Are the two variables mutually related to each other?" In other words, we may ask whether the changes in one variable are accompanied by some corresponding changes in the other variable. For example, to find the relationship between the heights and weights of 100 persons, we can arrange them in increasing order of their heights and see whether or not the weight increases as the height increases. In other words, we are asking, "Do taller people tend to weigh more than shorter people?" Note carefully that we are not saying that if an individual is taller than another, he has to necessarily weigh heavier. Very rarely, a taller person may weigh less than a shorter person, but quite often taller persons have higher weights than shorter persons. That is, in general, we may expect to see that as the heights of 100 individuals are arranged in increasing order and the corresponding weights written down, the weights will show a tendency to increase.

In such a situation, two variables, then, are said to be mutually related or correlated. This process of mutual relationship is called **correlation** between two variables. Note that correlation need not be only in one direction. As one variable shows an increase, the second variable may show an increase or a decrease. We know, for example, as the altitudes of places increase, the atmospheric pressure decreases.

Hence, whenever two variables are related to each other in such a way that change in the one creates a corresponding change in the other, then the variables are said to be correlated.

An easy way of studying the correlation of two quantitative variables is to plot them on a graph sheet taking one of the variables on the X-axis and the other on the Y-axis. The resulting diagram is called a scatter diagram because it shows how the pairs of observations are scattered on the graph sheet. Note that the points representing the values of x and y may lie very close to a straight line. This means that we can approximate the relationship between the values of x and y by a straight line or by some other geometrical curve. If this is a straight line, then we say that the relationship between x and y is linear. The relationship between x and y may be a curve other than a straight line. The study of such relationships is beyond the scope of the present syllabus. Hence, we shall confine our discussion to the linear relationship between x and y.

Definition: If two variables deviate in the same direction simultaneously, then they are said to be positively correlated.

In other words, if the two variables increase or decrease simultaneously (i.e., when one increases, the other also increases or. when one decreases, the other also decreases), then the correlation between the two variables is said to be a **positive correlation.**

In this case, the points of the scatter diagram follow a line of positive slope. For example, the correlation between heights and weights of a group of persons is a positive correlation.

Definition: If two variables deviate in opposite direction, then they are said to be negatively correlated or inversely correlated.

In other words, if the increase in me variable creates a decrease in the other, or the decrease in one creates an increase in the other, then the correlation between the two variables is said to be **negative correlation.** In this case, the points of the scatter diagram follow a line of negative slope. For example, the correlation between the price and demand of a commodity is a negative correlation.

<u>Coefficient of correlation</u>:

The extent degree of relationship between the two variables is measured in terms of another parameter called coefficient of correlation. Coefficient of correlation is denoted by "r".

The value of r lies between -1 and 1, i.e., $-1 \le r \le 1$.

<u>Properties:</u>

i) if r = 1 then the correlation is perfect and positive.

ii) if r = -1 then the correction is perfect and negative.

iii) if r = 0 there is no correction.

When the coefficient of correlation is perfect, both the variables x and y increase or decrease in the same proportion.

When the coefficient of correlation is negative the variables x and y are inversely proportional.

iv) if 0 < r < 1, then there is partial positive correlation between the variables.

Similarly, if -1 < r < 0, we say that there is partial negative correlation between x and y.

<u>Method of finding coefficient of correlation</u>:

1. Karl Pearson's coefficient of correlation

For a sample pair of observations of x and y the Karl Pearson's correlation coefficient is given by

$r = \dfrac{\sum\left\{(x-\bar{x})(y-\bar{y})\right\}}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$, where $\bar{x}$ is the mean of X series and $\bar{y}$ is the mean of Y series.

2. Rank correlation coefficient or Spearman's rank correlation coefficient

Is given by $r = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, d_i is the difference between ranks assigned to x_i *and* y_i ,

n is the number of pairs of data.

For repeated ranks: If two or more individuals are repeated we have to add the factor $\dfrac{m(m^2-1)}{12}$ to $\sum d_i^2$ where m is the no. of times an item is repeated. This correlation factor is to be added for each repeated value in both the X and Y series.

If the same value is there for m values average rank should be allotted, while allotting the ranks.

Ex: If two persons get same rank, the rank should be 8 and $\dfrac{8+9}{2} = 8.5$ should be allotted to each.

Regression analysis

The term "regression" literally means "stepping back" to words the average.

Def: "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data".

Regression analysis attempts to establish the "nature of the relationship between the variables that is, to study functional relationship between the variables and there by provide a mechanism for prediction or forecasting".

Regression lines: If we take the case of two variables X and Y, we shall have two regression lines as the regression of X on Y, and the regression of Y on X.

The regression equation of X on Y is used to describe the variation in the values of X for given change in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given change in X.

1. **Regression equation of X on Y**

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

where $\bar{x}$ is the mean of X series and $\bar{y}$ is the mean of Y series.

$r\frac{\sigma_x}{\sigma_y}$ is known as the regression coefficient of X on Y. The regression coefficient of X on Y is denoted by b_{xy} and $b_{xy} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{(y-\bar{y})^2}$.

2. **Regression equation of Y on X**

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

where $\bar{x}$ is the mean of X series and $\bar{y}$ is the mean of Y series.

$r\frac{\sigma_y}{\sigma_x}$ is known as the regression coefficient of Y on X. The regression coefficient of Y on X is denoted by b_{yx} and $b_{yx} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{(x-\bar{x})^2}$.

Note:

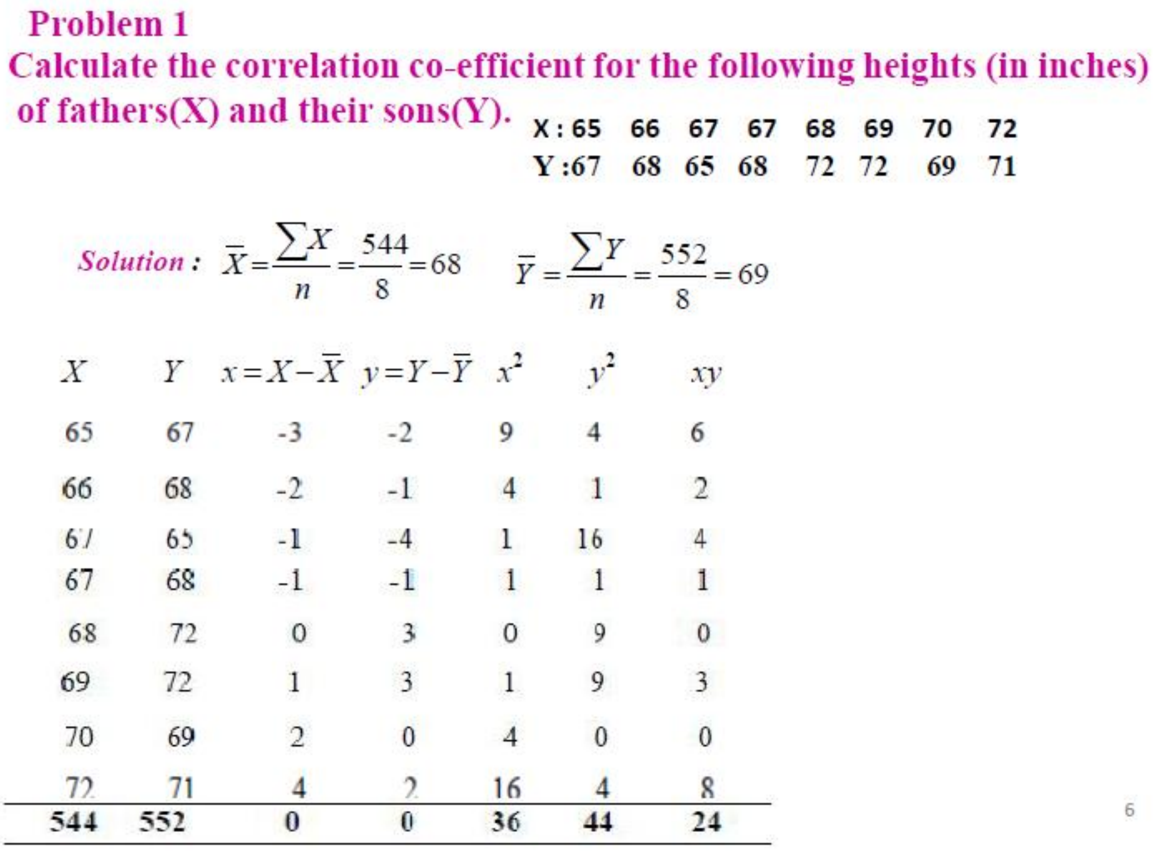
1. Correlation coefficient (r) is the root of the product of two regression coefficients.

   i.e. $r = \pm\sqrt{b_{xy}\, b_{yx}}$.

2. Both the regression coefficients will have the same sign i.e. either they will be positive or negative. It is never possible that one of the regression coefficients is –ve and the other +ve.

3. The coefficient of correlation will have the same sign as that of regression coefficients. That is if regression coefficients have negative sign, r will also we negative and regression coefficient have a positive sign, r would also be positive.

Solved Examples:



Problem 1

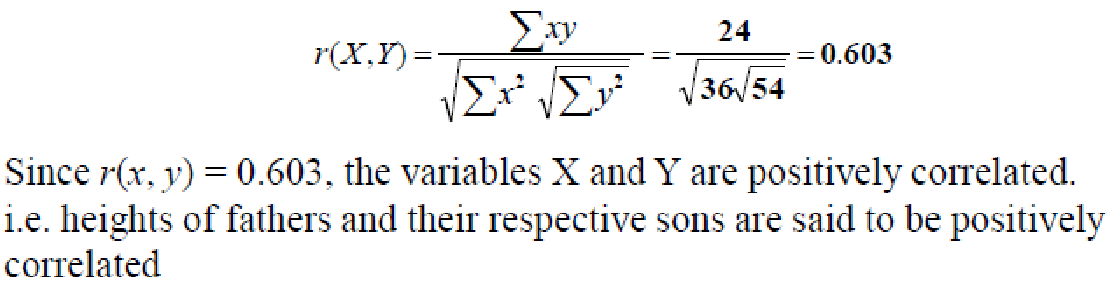
Calculate the correlation co-efficient for the following heights (in inches) of fathers(X) and their sons(Y).

X : 65	66	67	67	68	69	70	72
Y :67	68	65	68	72	72	69	71

Solution : $\overline{X}=\frac{\sum X}{n}=\frac{544}{8}=68$ $\overline{Y}=\frac{\sum Y}{n}=\frac{552}{8}=69$

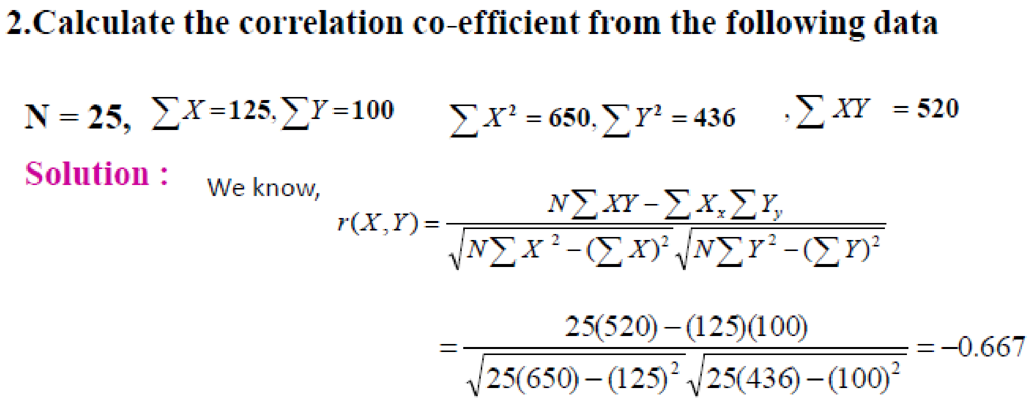
X	Y	$x=X-\overline{X}$	$y=Y-\overline{Y}$	x^2	y^2	xy
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
544	**552**	**0**	**0**	**36**	**44**	**24**

6



$$r(X,Y)=\frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}=\frac{\mathbf{24}}{\sqrt{\mathbf{36}}\sqrt{\mathbf{54}}}=\mathbf{0.603}$$

Since $r(x, y) = 0.603$, the variables X and Y are positively correlated. i.e. heights of fathers and their respective sons are said to be positively correlated



2.Calculate the correlation co-efficient from the following data

$\mathbf{N = 25}$, $\sum X=125, \sum Y=100$ $\sum X^2=650, \sum Y^2=436$, $\sum XY = 520$

Solution : We know,

$$r(X,Y)=\frac{N\sum XY-\sum X_x \sum Y_y}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}$$

$$=\frac{25(520)-(125)(100)}{\sqrt{25(650)-(125)^2}\sqrt{25(436)-(100)^2}}=-0.667$$

4.Obtain the rank correlation coeff. For the following data X:68 64 65 50 64 80 75 40 55 64

Y:62 58 68 45 81 60 68 48 50 70

Solution:

X	Y	rank in x_i	rank in y_i	$d = x_i - y_i$	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					72

In X series 75 is repeated twice which are in the position 2 & 3 in ranks

$\therefore$ common rank is 2.5(which is the average of 2&3) is to be given for each 75.Also in X series 64 is repeated 3 times which are in the position 5,6&7 in ranks.

$\therefore$ common rank is $\frac{5+6+7}{3}=6$, to be given for each 64

Similarly in Y series 68 is repeated 2 times which are in the position 3&4 in ranks

$\therefore$ common rank is 3.5(which is the average of 3&4) is to be given for each 68.

Correction factor

In X series 75 is repeated twice $\therefore$ c.f is $\dfrac{2(2^2-1)}{12} = \dfrac{1}{2}$

64 is repeated thrice $\therefore$ c.f is $\dfrac{3(3^2-1)}{12} = \dfrac{24}{12} = 2$

In Y series 68 is repeated twice $\therefore$ c.f is $\dfrac{2(2^2-1)}{12} = \dfrac{1}{2}$

$$r = 1 - \frac{6(\sum d^2 + \frac{1}{2} + 2 + \frac{1}{2})}{10(10^2-1)} = 1 - \frac{6(72+5)}{10\times 99} = 1 - \frac{450}{990} = 0.5454$$

7.Marks obtained by 10 students in Economics and Statistics are given below.

Marks In Eco. : 25 28 35 32 31 36 29 38 34 32
Marks In Stat : 43 46 49 41 36 32 31 30 33 39

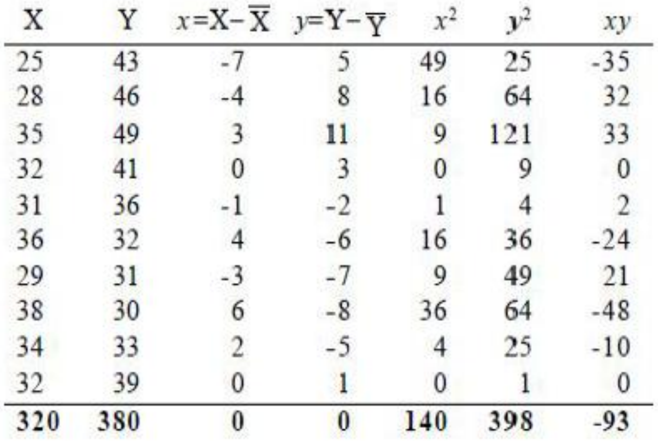
Find (i) the regression equation of Y on X

(ii) estimate the marks in statistics when the marks in Economics is 30.

Solution :

Let the marks in Economics be denoted by X and statistics by Y.

19



X	Y	$x=X-\overline{X}$	$y=Y-\overline{Y}$	x^2	y^2	xy
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	380	0	0	140	398	-93

$$\overline{X} = \frac{\sum X}{n} = \frac{320}{10} = 32 \qquad \overline{Y} = \frac{\sum Y}{n} = \frac{380}{10}$$

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{-93}{140} = -0.664$$

(i) Regression equation of Y on X is

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$Y - 38 = -0.664(X - 32)$$

$$Y = 59.25 - 0.664X$$

ii)To estimate the marks in statistics (Y) for a given marks in the Economics (X), put X = 30, in the above equation we get,

$$Y = 59.25 - 0.664(30)$$

$$= 59.25 - 19.92 = 39.33 \quad or \quad 39$$