UNIT III

Correlation and Regression

Correlation analysis is the statistical tool used to measure the degree to which two variables are linearly related to each other. Correlation measures the degree of association between two variables.

If the quantities (X, Y) vary in such a way that change in one variable corresponds to change in the other variable, then the variables X and Y are correlated.

Example: Price of commodity and amount of demand.

Correlation can be studied using various methods like

1. Scatter diagram

2. Karl Pearson's coefficient of correlation

3. Spearman's rank correlation coefficient.

Correlation**:**

If the change in one variable affects a change in the other variable, the variables are said to be correlated.

Two types of correlations are Positive correlation, Negative correlation

Positive Correlation:

If the two variables deviate in the same direction

Eg: Height and Weight of a group of persons, Income and Expenditure.

Negative Correlation:

If the two variables constantly deviate in opposite directions.

Eg: Price and Demand of a commodity, the correlation between volume and pressure of a perfect gas.

Measurement of Correlation:

We can measure the correlation between the two variables by using Karl – Pearson's coefficient of correlation.

Karl - Pearson' s coefficient of correlation

Correlation coefficient between two random variables X and Y, usually denoted by

$$r(X,Y) = \frac{COV(X,Y)}{\sigma_X.\sigma_Y}$$

Where $COV(X,Y) = \frac{1}{n}\sum XY - \bar{X}\,\bar{Y}$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X^2}}\ , \bar{X} = \frac{\sum X}{n}$$

$$\sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y^2}}\ , \bar{Y} = \frac{\sum Y}{n}$$

(n is the number of items in the given data)

Note:

1. Correlation coefficient may also be denoted by $\rho(X,Y) or\ \rho_{XY}$
2. If $\rho(X,Y) = 0$, We say that X and Y are uncorrelated.
3. Correlation coefficient does not exceed unity.

Note:

Types of correlation based on 'r'

Value of 'r'	Correlation is said to be
$r = 1$	Perfect and positive
$0 < r < 1$	Positive
$-1 < r < 0$	Negative
$r = 0$	Uncorrelated

Note: Method for finding correlation coefficient (discrete case)

$$r(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n\sigma_X\sigma_Y}$$

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

Regression

Definition

Regression is the measure of the average relationship between two or more variables in terms of original units of data. Example: If the sales and advertising are correlated we can find out the expected amount of sales for a given advertising expenditure or the amount needed for attaining the given amount of sales,

Lines of regression

If two variables X and Y are correlated i.c., there exists an association between them, we can see that the scatter diagram will be more or less concentrated around a curve. This curve is called Curve of regression.

If the curve is a straight line, it is called the line of regression and the regression is a linear regression. We shall have two regression lines as the regression line of X and Y and the regression line of Y and X. The regression line of Y and X gives the most probable value of Y for given values of X and the regression line of X and Y gives the most probable values of X for given values of Y.

Relation between Correlation Analysis and Regression Analysis

Sl. No	Correlation Analysis	Regression Analysis
1	Correlation coefficient r between X and Y is a measure of linear relationship between X and Y	The regression coefficients are mathematical measures expressing the average relationship between the two variables.
2	The correlation coefficient does not reflect upon the nature of variable (independent or dependent variable)	Regression coefficients reflect on the nature of variable i.e, which is dependent variable. In other words, it estimates the value of dependent variable for any given value of in dependent variable.
3	It does not imply cause and effect relationship between the variables under study	It indicates the cause and ef fect relationship between the vari ables. The variable corresponding to cause is taken as independent variable, whereas corresponding to effect is taken as dependent variable
4	It is a relative measure and is in- dependent of the units of measurement	Regression coefficients are absolute measures of finding out the relationship between two or more variables
5	It indicates the degree of association.	It is used to forecast the nature of dependent variable when the value of independent variable is known.

Uses of Regression Analysis

1. The cause and effect relations are indicated from the study of regression analysis.
2. It establishes the rate of change in one variable in terms of the changes in

another variable.

3. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
4. It helps in prediction and thus it can estimate the value of unknown quantities.
5. It enables us to study the nature of relationship between the variables.
6. It can be useful to all natural, social and physical sciences, where the data are in functional relationship

Regression Equations:

The line of regression of Y on X is given by

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

Where r is the correlation coefficient, σ_Y and σ_X are standard deviations.

The line of regression of Y on X is given by

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

Note :

1. The regression coefficients can be denoted by

$b_{yx} = r.\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r.\frac{\sigma_x}{\sigma_y}$

2. The regression co-efficients are obtained by the following expressions for discrete values of X and Y

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

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

3. Both the regression lines pass through the point $(\bar{x}, \bar{y})$where $\bar{x}$ and $\bar{y}$ are means of X and Y respectively.
4. Correlation coefficient is the Geometric mean between the regression coefficients. $b_{xy}.b_{yx} = r^2 \Longrightarrow r = \pm\sqrt{b_{xy}.b_{yx}}$
5. If one of the regression coefficients is greater than unity the other must be less than unity.
6. Regression coefficients are independent of the change of origin but not of scale.

Angle between regression lines

The angle θ between the two lines of regression is given by

$$\boldsymbol{tan\theta = \frac{1 - r^2}{r}\left(\frac{\sigma_Y \sigma_X}{{\sigma_Y}^2 + {\sigma_X}^2}\right)}$$

Problems under Correlation

1. Calculate the correlation coefficient for the following heights (in inches) of father X and their sons Y.

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

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\sum(X) = 544$	$\sum(Y) = 552$	$\sum(XY) = 37560$	$\sum(X^2) = 37028$	$\sum(Y^2) = 38132$

$\bar{X} = \frac{544}{8} = 68$, $\bar{Y} = \frac{552}{8} = 69$

$\bar{X}\,\bar{Y} = 68\,X\,69 = 4692$

$$r(X,Y) = r_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{8\,(37560) - (544)(552)}{\sqrt{8\,(37028) - (544)^2}\sqrt{\sqrt{8\,(38132) - (552)^2}}} = 0.6031$$

2. Find the correlation coefficient between industrial production and export using the following data:

Production (X)	**55**	**56**	**58**	**59**	**60**	**60**	**62**
Export (Y)	**35**	**38**	**37**	**39**	**44**	**43**	**44**

Solution:

U	**V**	$\boldsymbol{U = X - 58}$	$\boldsymbol{V = Y - 40}$	**UV**	**U^2**	**V^2**
55	35	-3	-5	15	9	25
56	38	-2	-2	4	4	4
58	37	0	-3	0	0	9
59	39	1	-1	-1	1	1
60	44	2	4	8	4	16
60	43	2	3	6	4	9
62	44	4	4	16	16	16

		$\sum(U) = 4$	$\sum(V) = 0$	$\sum(UV) = 48$	$\sum(U^2) = 38$	$\sum(V^2) = 80$

Now $\bar{U} = \frac{\sum U}{n} = \frac{4}{7} = 0.5714$

$$\bar{V} = \frac{\sum V}{n} = 0$$

$Cov\,(U,V) = \frac{\sum UV}{n} - \bar{U}\bar{V} = 6.857$

$\sigma_X = \sqrt{\frac{1}{n}\sum U^2 - \overline{U^2}} = 2.2588$

$\sigma_Y = \sqrt{\frac{1}{n}\sum V^2 - \overline{V^2}} = 3.38$

$r\,(U,V) = \frac{COV(U,V)}{\sigma_U.\sigma_V} = \frac{\frac{1}{n}\sum UV - \bar{U}\bar{V}}{\sigma_U.\sigma_V} = 0.898$

3. Given number of pairs of observations of X and Y series = 8.
   X series arithmetic mean = 74.5
   X series assumed mean = 69.0
   X series standard deviation = 13.07
   Y series arithmetic mean = 125.5
   Y series assumed mean = 112.0
   Y series standard deviation = 15.85
   Summation of products of corresponding deviations of X and Y series = 2176.
   Calculate the coefficient of correlation between X and Y series.
   Solution:

$$r = \frac{\sum xy - n(a_x - \bar{x})(a_y - \bar{y})}{n\sigma_x.\sigma_y}$$

$$= \frac{2176 - 8(69.0 - 74.5)(112.0 - 125.5)}{8 * 13.07 * 15.85} = 0.95$$

4. From the following data, calculate the coefficient of correlation between X and y series.

Mean of X series = 75
Assumed mean of X series = 70
Mean of Y series = 126
Assumed mean of Y series = 113
Standard deviation of X series = 13.5
Standard deviation of Y series = 15.8
Sum of products of corresponding deviations of X and Y series =2186
No. of pairs = 8

Solution

$$r = \frac{\sum xy - n(a_x - \bar{x})(a_y - \bar{y})}{n\sigma_x . \sigma_y}$$

$$= \frac{2186 - 8(70 - 75)(113 - 126)}{8 * 13.5 * 15.8} = 0.9763$$

5. the following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.

Age in year /Marks	**18**	**19**	**20**	**21**	**Total**
10-20	**4**	**2**	**2**		**8**
20-30	**5**	**4**	**6**	**4**	**19**
30-40	**6**	**8**	**10**	**11**	**35**
40-50	**4**	**4**	**6**	**8**	**22**
50-60		**2**	**4**	**4**	**10**

60-70		**2**	**3**	**1**	**6**
Total	**19**	**22**	**31**	**28**	**100**

Calculate the correlation coefficient.

Solution:

$$u = x - 19, v = \frac{y - 35}{10}$$

		u	-1	0	1	2	Total			
v	**Mid Value**	**Age X** / Mark Y	**18**	**19**	**20**	**21**	**f**	f_v	f_{v^2}	f_{uv}
-2	**15**	**10-20**	**4** 8	**2** 0	**2** -4	**-**	**8**	**-16**	**32**	**4**
-1	**25**	**20-30**	**5** 5	**4** 0	**6** -6	**4** -8	**19**	**-19**	**19**	**-9**
0	**35**	**30-40**	**6** 0	**8** 0	**10** 0	**11** 0	**35**	**0**	**0**	**0**
1	**45**	**40-50**	**4** -4	**4** 0	**6** 6	**8** 16	**22**	**22**	**22**	**18**
2	**55**	**50-60**	**-**	**2** 0	**4** 8	**4** 16	**10**	**20**	**40**	**24**
3	**65**	**60-70**	**-**	**2** 0	**3** 9	**1** 6	**6**	**18**	**54**	**15**
	Total	**f**	**19**	**22**	**31**	**28**	**N=100**	**25**	**167**	**52**
		f_u	**-19**	**0**	**31**	**56**	**68**			
		f_{u^2}	**19**	**0**	**31**	**112**	**162**			

f_{uv}	9	0	13	30	52

$$r = \frac{N\sum f_{xy} - (\sum f_x)(\sum f_y)}{\sqrt{N\sum f_{x^2} - (\sum f_x)^2}\ \sqrt{N\sum f_{y^2} - (\sum f_y)^2}}$$

$$r = \frac{N\sum f_{uv} - (\sum f_u)(\sum f_v)}{\sqrt{N\sum f_{u^2} - (\sum f_u)^2}\ \sqrt{N\sum f_{v^2} - (\sum f_v)^2}}$$

$$r = \frac{(100 * 52) - (68 * 25)}{\sqrt{(100 * 162) - (68)^2}\ \sqrt{(100 * 167) - (25)^2}} = 0.2566$$

7. From the following table given below calculate the coefficient of correlation between the ages of husband and wives.

Age in year /Marks	18	19	20	21	Total
10-20	4	2	2		8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60		2	4	4	10
60-70		2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Solution:

$$u = x - 19, v = \frac{y - 35}{10}$$

		u	-1	0	1	2	Total			
v	Mid Value	Age X / Mark Y	18	19	20	21	f	f_v	f_{v^2}	f_{uv}
-2	15	10-20	4 (8)	2 (0)	2 (-4)	-	8	-16	32	4
-1	25	20-30	5 (5)	4 (0)	6 (-6)	4 (-8)	19	-19	19	-9
0	35	30-40	6 (0)	8 (0)	10 (0)	11 (0)	35	0	0	0
1	45	40-50	4 (-4)	4 (0)	6 (6)	8 (16)	22	22	22	18
2	55	50-60	-	2 (0)	4 (8)	4 (16)	10	20	40	24
3	65	60-70	-	2 (0)	3 (9)	1 (6)	6	18	54	15
	Total	f	19	22	31	28	N=100	25	167	52
		f_u	-19	0	31	56	68			
		f_{u^2}	19	0	31	112	162			
		f_{uv}	9	0	13	30	52			

$$r = \frac{N\sum f_{xy} - (\sum f_x)(\sum f_y)}{\sqrt{N\sum f_{x^2} - (\sum f_x)^2}\ \sqrt{N\sum f_{y^2} - \left(\sum f_y\right)^2}}$$

$$r = \frac{N\sum f_{uv} - (\sum f_u)(\sum f_v)}{\sqrt{N\sum f_{u^2} - (\sum f_u)^2}\ \sqrt{N\sum f_{v^2} - (\sum f_v)^2}}$$

$$r = \frac{(100 * 52) - (68 * 25)}{\sqrt{(100 * 162) - (68)^2}\ \sqrt{(100 * 167) - (25)^2}} = 0.2566$$