BLOCK III:
CORRELATION, REGRESSION AND ASSOCIATION OF ATTRIBUTES

Unit 1 : Basic Concepts of Partial and Multiple Correlation and Regression

Unit 2 : Various Formulae, Problems, Uses and Limitation of Partial and Multiple Correlations and Regressions

Unit 3 : Coefficient of Multiple Determinations, Association of Attributes: Concept, Order of a Class, Class Frequency, Consistency of Data

Unit 4 : Kinds of Association of Attributes, Methods of Measuring Association Between Two Attributes, Partial Association

Unit-1
Basic Concepts of Partial and Multiple Correlation and Regression

Unit Structure:

1.1 Introduction

So far we have confined our discussion to univariate distribution which involve only one variable . However, in business, the key to decision-making often lies in the understanding of the relationships between two or more variables. A distribution where each unit of the series assumes two values is called a Bivariate Distribution. For example, a company in the distribution business may determine that there is a relationship between the price of crude oil and its own transportation costs. A marketing executive might want to know how strong the relationship is between advertising dollars and sales dollars for a product or a company. In this chapter, we will study the concept of correlation and how it can be used to estimate the relationship between two variables using regression.

1.2 Objective

After going through this unit, you will be able to-

- understand the concept of partial correlations,
- understand the concept of multiple correlations,
- understand the significance of multiple regression.

1.3 Basic Concepts of partial correlation

1.3.1 Partial Correlation

The partial correlation coefficient describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically. In partial correlation, the linear association between a dependent variable and one particular independent variable is studied, holding other independent variables constant.

If a dependent variable X_1 and two independent variables X_2 and X_3 are included in the partial correlation analysis, then the partial correlation between X_1 and X_2holding X_3 constant is denoted by $r_{12.3}$. Similarly, partial correlation betweenX_1 and X_3 holding X_2constant is denoted by $r_{13.2}$

Depending upon the number of independent variables which are held constant partial correlation coefficients are often called as zero-order, first-order, second-order correlation coefficients.

The partial correlation between X_1 and X_2 holding X_3 constant is determined by

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \quad ;$$

Similarly partial correlation between X_1 and X_3 holding X_2 constant

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{32}^2}} \quad \text{and}$$

Partial correlation between X_2 and X_3 holding X_1 constant is given by -

$$r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{1 - r_{21}^2} \sqrt{1 - r_{31}^2}}$$

Again we have the following relations:

(a) $$r_{12.3} = \sqrt{R_{12.3}^2} = \sqrt{1 - \frac{S_{1.23}^2}{S_{1.3}^2}} \quad and \quad r_{13.2} = \sqrt{R_{13.2}^2} = \sqrt{1 - \frac{S_{1.23}^2}{S_{1.2}^2}}$$

(b)
$$r_{12.3} = \sqrt{b_{12.3} x b_{21.3}} \; ;$$
$$r_{13.2} = \sqrt{b_{13.2} x b_{31.2}} \quad \textit{and}$$
$$r_{23.1} = \sqrt{b_{23.1} x b_{32.1}}$$

Stop to Consider

The partial correlation coefficient describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically. For e.g. the relationship between the yield of wheat and fertiliser when all other variables such as nature of soil, irrigation, climate, seed and techniques of cultivation are kept constant is termed as partial correlation.

Using these relations the partial correlation between X_1 and X_2 holding X_3 constant can also be determined as:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{13}^2} \left(\frac{S_1}{S_2}\right) \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{13}^2} \left(\frac{S_2}{S_1}\right)$$

$$= \frac{\left(r_{12} - r_{13} r_{23}\right)^2}{\left(1 - r_{13}^2\right)\left(1 - r_{13}^2\right)}$$

Note

(1) $r_{xy} = r_{yx}$

(2) The value of the partial correlation coefficient lies between -1 and 1.

1.3.2 Partial Correlation Coefficient in four Variables

If there are four variables X_1,X_2,X_3 and X_4 under consideration for the joint study, then the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and X_4 is given by

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} r_{23.4}}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2}} \qquad (1)$$

Or alternatively

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \qquad (2)$$

Formula (1) and (2) are identical and hence whatever expression we take for $r_{12.34}$we shall get the same value . In the similar manner formula for other partial correlation coefficient can be calculated.$\mathbf{r}_{12.34}$is known as the second order partial correlation coefficient .

1.3.3 Partial Correlation Coefficient in four Variables

If there are four variables X_1,X_2,X_3 and X_4 under consideration for the joint study, then the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and $X_4$is given by

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} r_{23.4}}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2}} \qquad (1)$$

Or alternatively

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \qquad (2)$$

Formula (1) and (2) are identical and hence whatever expression we take for $r_{12.34}$we shall get the same value . In the similar manner formula for other partial correlation coefficient can be calculated.$\mathbf{r}_{12.34}$is known as the second order partial correlation coefficient .

<u>Example 1</u>: In a trivariate distribution it is found that r_{12}=0.70 , r_{13}=0.61 , r_{23}=0.40. Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$.

Solution: The partial correlation between X_2 and X_3 holding X_1 constant is determined by

Substituting the given values we get

$$r_{23.1}=\frac{0.40-0.70\times 0.61}{\sqrt{1-(0.70)^2\sqrt{1-(0.61)^2}}}$$

$$=\frac{0.40-0.427}{\sqrt{0.51\sqrt{0.6279}}}$$

$$=\frac{0.027}{0.714\times 0.7924}$$

$$=\frac{0.027}{0.5657}$$

$$=0.0477$$

Check Your Progress

1. Given the following values: r_{23}=0.4

   r_{13}=0.61 , r_{12}=0.7

   Find the partial correlation coefficients: $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

2. In a tri-variate distribution. it was found r_{12}=0.75 , r_{13}= 0.9 , r_{23}=0.6

   Find the values of $r_{12.3}$, $r_{23.1}$ and $r_{1.23}$

<u>**Exercise 2:**</u> The correlation between intelligence test scores and school achievement in a group of school students is 0.80. The correlation between intelligence test scores and age in the same age group is 0.70 and the score between intelligence test scores and age is 0.60. Find out the correlation between intelligence test scores and school achievement in children of the same age. Comment on the result.

Solution: Let x_1= intelligence test scores; x_2=school achievement; x_3= age of children

Given that r_{12}=0.80 , r_{13}=0.70 , r_{23}=0.60

Then the correlation between intelligence test scores and school achievement, keeping the influence of age as constant, is given by

$$r_{12.3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{12}{}^{2}}\sqrt{1-r_{23}{}^{2}}}$$

$$=\frac{0.8-0.7\times0.6}{\sqrt{1-(0.7)^{2}}\sqrt{1-(0.6)^{2}}}$$

$$=\frac{0.8-0.42}{\sqrt{0.51\sqrt{0.64}}}=\frac{0.38}{0.57}=0.667$$

Hence it can be concluded that intelligence test scores and school achievement are associated to each other to the extent of of $r_{12.3}=0.667$ while the influence of childrens' age is held constant.

Exercise 3: Given $r_{12.4}=0.60$, $r_{13.4}=0.50$, $r_{23.4}=0.70$ find $r_{12.34}$ and $r_{13.24}$

Solution:

$$r_{12.34}=\frac{r_{12.4}-r_{13.4}r_{23.4}}{\sqrt{1-r^{2}13.4}\sqrt{1-r_{23.4}{}^{2}}}$$

$$=\frac{0.6-(0.5\times0.7)}{\sqrt{\{1-(0.5)^{2}\}}\sqrt{\{1-(0.7)^{2}\}}}$$

$$=\frac{0.6-0.35}{\sqrt{0.75\times0.51}}$$

$$=\frac{0.25}{\sqrt{0.3825}}$$

$$=\frac{0.25}{0.62}\quad=0.403$$

$$r_{13.24}=\frac{r_{13.4}-r_{12.4}r_{23.4}}{\sqrt{1-r^{2}12.4}\sqrt{1-r_{23.4}{}^{2}}}$$

$$=\frac{0.5-(0.6\times0.7)}{\sqrt{\{1-(0.6)^{2}\}}\sqrt{\{1-(0.7)^{2}\}}}$$

$$=\frac{0.5-0.42}{\sqrt{0.64\times0.51}}$$

$$=\frac{0.08}{0.57}$$

$$=0.14$$

Exercise 4 : Suppose a computer has found for a given set of variables X_1,X_2 and X_3 the correlation coefficients are $r_{12=0.91}$, $r_{13=0.33}$ and r_{23}= 0.81. Explain whether these computations may be said to be free from error.

Solution : For determining whether the given computations are correct or not , we calculate the value of the partial correlation coefficient $r_{12.3}$ for variables 1 and 2 keeping the influence of variable 3 constant . If the value of $r_{12.3}$ is less than one , then the computation may be said to be free from error .

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{0.91-(0.33\times 0.81)}{\sqrt{1-(0.33)^2}\sqrt{1-(0.81)^2}}$$

$$= \frac{0.91-0.2673}{\sqrt{1-0.1089}\sqrt{1-0.6561}}$$

$$= \frac{0.6427}{\sqrt{0.8911\times 0.3439}}$$

$$= 1.161$$

Since the calculated value of $r_{12.3}$ is more than one, the computation given in the question are not free from error.

1.4 Multiple Correlation

The aim of the theory of multiple correlation is to study the joint effect of a group of variables not included in the group. In general, the coefficient of multiple correlation measures the extent of the association between the dependent variable and several independent variables taken together. Thus while studying multiple correlation, the effect of certain independent factors on a dependent factor is studied without treating any factor constant.

A linear multiple correlation is denoted by the symbol R and the necessary subscripts are added to it. The subscript before the decimal represents the dependent variable and the subscripts after the decimal represent the independent variables which affect the dependent variable. For example $R_{1.23}$ indicates that the variable X_1 is associated with the variables X_2 and X_3 and the formula for it is:

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

By symmetry we may also write

$$R_{1.23} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \text{ and } R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Note:

(1) The value of multiple correlation coefficient always lie between 0 and 1.

(2) If r_{12}=r_{13}=0 then $R_{1.23}$=0 implying no linear relationship between the variables.

(3) $R_{1.23}$=$R_{1.32}$

(4) If $R_{1.23}$=1 , the correlation is called perfect

<u>Exercise 5 :</u> In a three variate multiple correlation analysis, the following results were obtained r_{12}=0.59 r_{13}=0.46 and r_{23}=0.77. Find $R_{1.23}$

Solution: Multiple Correlation Coefficient is defined as

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

$$= \sqrt{\frac{(0.50)^2 + (0.46)^2 - 2(0.59 \times 0.46 \times 0.77)}{1 - (0.77)^2}}$$

$$= \sqrt{\frac{0.3481 + 0.2116 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.5597 - 0.418}{0.4071}}$$

$$= \sqrt{\frac{0.1417}{0.4071}} = 0.589$$

Stop to Consider

The multiple correlation may be defined as a statistical tool designed to measure the degree of relationship existing among three or more variables. it is denoted by the symbol R

$R_{1.23}$ stands for the coefficients of multiple correlation between X_1, X_2 and X_3. The subscript to the left stands for dependent variable while the subscripts to the right of the dot represent independent variables.

1.5 Multiple Regression:

Simple regression analysis is bivariate linear regression in which one dependent variable, y, is predicted by one independent variable, x. Examples of simple regression applications include models to predict retail sales by population dens.

Simple regression analysis (discussed in Chapter 12) is bivariate linear regression in which one dependent variable, y, is predicted by one independent variable, x. Examples of simple regression applications include models to predict retail sales by population density. Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis.

Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable. If a regression model characterises the relationship between a dependent y and only one independent x, then such a regression model is called a simple regression model given by

$$\hat{Y} = a + bX \qquad (1)$$

But if more than one independent variable is associated with a dependent variable then such a regression model is called a multiple regression model. In multiple regression analysis, the dependent variable, y, is sometimes referred to as the response variable

A multiple regression equation is an equation which is used for estimating a dependent variable say X_1 from a set of independent variables X_2, X_3,... and is called the regression equation of X_1 on X_2, X_3,.... For two independent variables X_2, X_3 we have the following simplest regression equation of X_1 on X_2 and X_3 and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, $b_{12.3}$ and $b_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1), a is called the X- intercept, $b_{12.3}$ indicates the slope of the regression line of X_1 on X_2called partial regression coefficient ofX_1 on $X_2$when X_3 is held constant. $b_{13.2}$ indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3keeping X_2constant.

If we take the deviations of the variables from their respective means and denote these deviations by x_1, x_2 andx_3 i.e. if

$$x_1 = X_1 - \overline{X}_1$$
$$x_2 = X_2 - \overline{X}_2$$
$$x_3 = X_3 - \overline{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\overline{X}_1 = a + b_{12.3}\overline{X}_2 + b_{13.2}\overline{X}_3$$
$$X_1 - \overline{X}_1 = b_{12.3}(X_2 - \overline{X}_2) + b_{13.2}(X_3 - \overline{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2 x_1 = b_{12.3}\Sigma\, x_2^2 + b_{13.2}\Sigma\, x_3 x_2 \qquad (4)$$
$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\, x_3^2$$

Solving the two normal equations in (4) the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_2} \text{ and } b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \overline{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{23}} \right) \frac{S_1}{S_2} (X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2{}_{23}} \right) \frac{S_2}{S_1} \left(X_1 - \overline{X}_1 \right) \qquad (6)$$

Similarly the multiple regression equation of X$_2$ on X$_3$ and X$_1$ as

$$X_2 - \overline{X}_2 = \left(\frac{r_{23} - r_{12}r_{13}}{1 - r^2{}_{13}} \right) \frac{S_2}{S_3} (X_3 - \overline{X}_3) + \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{13}} \right) \frac{S_2}{S_1} \left(X_1 - \overline{X}_1 \right) \qquad (7)$$

<u>Exercise 6 :</u> Find the multiple linear regression equation of X$_1$ on X$_2$ and X$_3$ from the data relating to three variables given below:

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

Solution:

The regression equation of X$_1$ on X$_2$ and X$_3$ is

$$\sum X_1 = a + b_{12.3} \sum x_2 + b_{13.2} \sum x_3$$

$$\sum x_2 x_1 = b_{12.3} \sum x_2{}^2 + b_{13.2} \sum x_3 x_2$$

$$\sum x_2 x_3 = b_{12.3} \sum x_3 x_2 + b_{13.2} \sum x_3{}^2$$

Calculating the required values we get:

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_2{}^2$	$X_3{}^2$	$X_1{}^2$
4	15	30	60	120	450	225	900	16
6	12	24	72	144	288	144	576	36
7	8	20	56	140	160	64	400	49
9	6	14	54	126	64	36	196	81
13	4	10	52	130	40	16	100	169
15	3	4	45	60	12	9	16	225
$\sum X_{1=54}$	$\sum X_{2=48}$	$\sum X_3$=102	$\sum X_1X_{2=339}$	$\sum X_1X_{3=}720$	$\sum X_2X_{3=}1034$	$\sum X_2{}^{2=}494$	$\sum X_3{}^{2=}2188$	$\sum X_1{}^{2=}576$

Substituting the values in the normal equations

48a+494 b$_{12.3}$+1034b$_{13.2}$ =54 (2)

102a +1034 b$_{12.3}$+ 2188 b$_{13.2}$=720 (3)

Multiplying equation (1) by 8 we have

48a +384 b$_{12.3}$+816 b$_{13.2}$=432 (4)

Subtracting Equation (2) from equation (4) we get

110 b$_{12.3}$+218 b$_{13.2}$ =93 (5)

Multiplying Equation (1) by 17 we get

$102\ a + 816\ b_{12.3} + 1734\ b_{13.2} = 918$ (6)

Subtracting Equation (3) from equation (4) we get

$218\ b_{12.3} + 454\ b_{13.2} = -198$ (7)

Multiplying equation (5) by 109 we obtain

$11990\ b_{12.3} + 24970\ b_{13.2} = 10137$ (8)

Multiplying equation (7) by 55 , we get

$11990\ b_{12.3} + 23762\ b_{13.2} = -10890$ (9)

Subtracting equation (8) from equation (9) we get

$1208\ b_{12.3} = -753$

$b_{12.3} = -0.623$

Substituting the value of $b_{12.3}$ in equation (5) we get

$110\ b_{12.3} = 218\ (-0.623) = -93$

$110\ b_{12.3} = 135.814 - 93$

$b_{12.3} = 42.814/110 = 0.389$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (1) we get

$6a + 48\ (0.389) + 102\ (-0.623) = 54$

$6a = 54 + 63.546 - 18.672 = 98.874$

$a = 16.479$

Thus the required equation is

$X_1 = 16.479 + 0.389X_2 - 0.623\ X_3$

<u>Exercise 7:</u> Given $r_{12}=0.28$ $r_{23}=0.49$ $r_{31}=0.51$

Solution The required equation of X_3 on X_1 and X_2 is

$X_{3=}\ b_{31.2}X_1 + b_{31.2}\ X_2$ (1)

Where $b_{31.2} = \dfrac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2} \cdot \dfrac{s_3}{s_1}$ and $b_{32.1} = \dfrac{r_{32} - r_{13}r_{12}}{1 - r_{12}^2} \cdot \dfrac{s_3}{s_2}$

The symbols s_1, s_2 and s_3 are the standard deviations of X_1 , X_2 and X_3respectively. However this represents the sample standard deviation and the population standard deviation and is given by symbol σ. But some authors, without making this distinction use the symbol σ for both sample s.d and population S.D

$$b_{31.2} = \frac{0.51 - 0.49x0.28}{1-(0.28)^2} x \frac{2.7}{2.7}$$

$$= 0.405$$

$$b_{32.1} = \frac{0.49 - 0.51x0.28}{1-(0.28)^2} x \frac{2.7}{2.4}$$

$$= 0.424$$

Equation (1) gives

X_3= 0.405 X_1+0.424 X_2

Check Your Progress

3. The simple correlation coefficients between temperature temperature(x_1) , yield of corn (x_2) and rainfall (x_3) are r_{12}=0.59 , r_{13}=0.46 and r_{23}=0.77. Calculate the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$

1.6 Summing Up

- Partial and multiple correlations describe the relationship between two variables when influence of one or more other variables is held constant or involved.
- Multiple regressions establish an equation for estimating value of a dependent variable given the value of two or more independent variables.
- The coefficient of partial correlation between X_1 and X_2 keeping X_3 constant is denoted by the symbol $r_{12.3}$
- In case of four variables the partial correlation coefficient between X_1 and X_2 eliminating the influence of X_3 and X_4 is denoted by $r_{12.34}$
- The coefficient of multiple correlation is denoted by the symbol $R_{1.23}$

- The coefficient of multiple determination is denoted by the symbol R^2
- The value of partial correlation coefficient lies between -1 and +1
- The value of partial multiple correlation coefficient lies between 0 and 1

1.7 Key words

- Partial correlation coefficient:

  It is a measure of the degree of linear association between a dependent variable and one particular independent variable and one particular independent variable when all other independent variables are kept constant.

- Multiple correlation coefficient:

  It gives the effects of all the independent variables on a dependent variable.

- A multiple regression equation:

  It is an equation which is used for estimating a dependent variable from a set of independent variables

1.8 Model Questions

Short answer questions:

1. What is the importance of partial correlation analysis?
2. Define multiple correlation.
3. What are partial regression coefficients?
4. Given $r_{12} = 0.5$, $r_{13} = 0.4$ and $r_{23} = 0.1$, find the values of $r_{12.3}$.
5. If $r_{12} = 0.9$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find the values of $R_{1.23}$.
6. The range of partial correlation coefficient r12.3 is

   (a) -1 to 1 (b) 0 to ?

   (c) 0 to 1 (d) none of these
7. If multiple correlation coefficient R1.23 =1, then it implies a

   (a) lack of linear relationship (b) perfect relationship

(c) reasonably good relationship (d) none of these

8. A coefficient which examines the association between a dependent variable and an independent variable after factoring out the effect of other independent variables is known as ________.

9. A statistical technique that develops an equation that relates a dependent variable to one or more independent variables is called _______.

Long answer questions:

1. Define the following terms:

   a) Coefficient of partial and multiple correlation.

   b) Coefficient of multiple determination.

2. What are normal equations and how are they used in multiple regression analysis?

Distinguish between partial and multiple correlation and point out their usefulness in statistical analysis

3. The simple correlation coefficients between profits (X_1), sales (X_2) and advertising expenditure (X_3) of a factory are $r_{12} = 0.69$, $r_{13} = 0.45$ and $r_{23} = 0.58$ find the partial correlation coefficients $r_{12.3}$ and $r_{13.2}$ and interprete them.

4. In a trivariate distribution:

   $\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5, r_{23} = 0.4, r_{31} = 0.6, r_{13} = 0.7$

   Determine the regression equation of X_1 on X_2 and X_3 if the variates are measures from their means.

5. The following data are given:

   $x_1 = 6, x_2 = 7, x_3 = 8; s_1 = 1 s_1 = 2, s_1 = 3; r_{23} = 0.8, r_{12} = 0.6, r_{13} = 0.7$

   a) Find the regression equation of x_3 on x_1 and x_2

   b) Estimate the value of x_3 when $x_1 = 4$ and $x_2 = 5$

1.9 References and Suggested Readings

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.

2. Business Statistics: S. Saha, New Central Book Agency.
3. Basic Statistics:B. L. Agarwal, New Age International Limited.
4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
5. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited.

******

Unit 2

Various Formulae, Problems, Uses and Limitation of Partial and Multiple Correlations and Regressions

Unit Structure:

2.1 Introduction

In this unit we describe about partial correlation coefficient which describes the relationship between one of the independent variables and the dependent variable, given that the other independent variables are held constant statistically. In partial correlation, the linear association between a dependent variable and one particular independent variable is studied, holding other independent variables constant.

2.2 Objective

After going through this unit, you will be able to-

- understand the formulas and problems of partial correlations,
- understand the formulas and problems of multiple correlations,
- understand the problems of multiple regression.

2.3 Various formulae and Problems of

2.3.1 Partial Correlation

The partial correlation between X_1 and X_2 holding X_3 constant is determined by

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \quad ;$$

Similarly partial correlation between X_1 and X_3 holding X_2 constant

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}} \quad \text{and}$$

Partial correlation between X_2 and X_3 holding X_1 constant is given by -

$$r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

Again we have the following relations:

(a) $$r_{12.3} = \sqrt{R^2_{12.3}} = \sqrt{1 - \frac{S^2_{1.23}}{S^2_{1.3}}} \quad and \quad r_{13.2} = \sqrt{R^2_{13.2}} = \sqrt{1 - \frac{S^2_{1.23}}{S^2_{1.2}}}$$

(b) $$r_{12.3} = \sqrt{b_{12.3}} x b_{21.3} \ ;$$

$$r_{13.2} = \sqrt{b_{13.2}} x b_{31.2} \quad and$$

$$r_{23.1} = \sqrt{b_{23.1}} x b_{32.1}$$

Using these relations the partial correlation between X_1 and X_2 holding X_3 constant can also be determined as:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{13}^2}\left(\frac{S_1}{S_2}\right) \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{13}^2}\left(\frac{S_2}{S_1}\right)$$

$$= \frac{\left(r_{12} - r_{13} r_{23}\right)^2}{\left(1 - r_{13}^2\right)\left(1 - r_{13}^2\right)}$$

Note

(1) $r_{xy}=r_{yx}$

(2) The value of the partial correlation coefficient lies between -1 and 1.

2.3.2 Multiple Correlations

A linear multiple correlation is denoted by the symbol R and the necessary subscripts are added to it. The subscript before the decimal represents the dependent variable and the subscripts after the decimal represent the independent variables which affect the dependent variable. For example $R_{1.23}$ indicates that the variable X_1 is associated with the variables X_2 and X_3 and the formula for it is:

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

By symmetry we may also write

$$R_{1.23} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \quad \text{and}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Note:

(1) The value of multiple correlation coefficient always lie between 0 and 1.

(2) If $r_{12}=r_{13}=0$ then $R_{1.23}=0$ implying no linear relationship between the variables.

(3) $R_{1.23}=R_{1.32}$

(4) If $R_{1.23}=1$, the correlation is called perfect

2.3.3 Regressions

A multiple regression equation is an equation which is used for estimating a dependent variable say X$_1$ from a set of independent variables X$_2$, X$_3$,... and is called the regression equation of X$_1$ on X$_2$, X$_3$,.... For two independent variables X$_2$, X$_3$ we have the following simplest regression equation of X$_1$ on X$_2$ and X$_3$ and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, b$_{12.3}$ and b$_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1) , a is called the X- intercept , b$_{12.3}$ indicates the slope of the regression line of X$_1$ on X$_2$called partial regression coefficient ofX$_1$ on X$_2$when X$_3$ is held constant . b$_{13.2}$ indicates the slope of the regression line of X$_1$ on X$_3$ when X$_2$ is held constant. It is called the partial regression coefficient of X$_1$ on X$_3$ when X$_2$ is held constant. It is called the partial regression coefficient of X$_1$ on X$_3$keeping X$_2$constant.

If we take the deviations of the variables from their respective means and denote these deviations by x$_1$, x$_2$ andx$_3$ i.e. if

$$x_1 = X_1 - \overline{X}_1$$
$$x_2 = X_2 - \overline{X}_2$$
$$x_3 = X_3 - \overline{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\overline{X}_1 = a + b_{12.3}\overline{X}_2 + b_{13.2}\overline{X}_3$$
$$X_1 - \overline{X}_1 = b_{12.3}(X_2 - \overline{X}_2) + b_{13.2}(X_3 - \overline{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2x_1 = b_{12.3}\Sigma\ x_2{}^2 + b_{13.2}\Sigma\ x_3x_2 \qquad (4)$$
$$\Sigma x_2x_3 = b_{12.3}\Sigma x_3x_2 + b_{13.2}\Sigma\ x_3{}^2$$

Solving the two normal equations in (4) the partial regression coefficients b$_{12.3}$ and b$_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_2} \quad \text{and} \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \overline{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{23}}\right)\frac{S_1}{S_2}(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2{}_{23}}\right)\frac{S_2}{S_1}\left(X_1 - \overline{X}_1\right) \qquad (6)$$

Similarly the multiple regression equation of X_2 on X_3 and X_1 as

$$X_2 - \overline{X}_2 = \left(\frac{r_{23} - r_{12}r_{13}}{1 - r^2{}_{13}}\right)\frac{S_2}{S_3}(X_3 - \overline{X}_3) + \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{13}}\right)\frac{S_2}{S_1}\left(X_1 - \overline{X}_1\right) \qquad (7)$$

2.4 Uses and limitation of

2.4.1 Partial Correlation

Advantages

(1) Partial correlation is especially useful in the analysis of interrelated variables where the linear relation between a dependent and independent variable can be studied by eliminating the influence of other independent variables.

(2) In partial correlation, relationships are expressed concisely in a few well defined coefficients.

(3) Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kind of phenomenon.

Limitation

(1) While calculating partial correlation coefficient it is assumed that the simple correlation or zero order correlation from which partial correlation is studied have linear relationship between the variables. In actual practice, particularly in social science this assumption is not desirable because linear relationship doesnot generally exist in such situations.

(2) The effects of the independent variables are studied additively not jointly. The various independent variables considered in the study are assumed to be independent of each other, however in actual practice this may not be true and there may be a relationship among the variables.

(3) The computation of partial correlation coefficients especially beyond first order is laborious and time consuming.

Check Your Progress

1. What are the advantage and limitation of partial correlation?
2. What are the advantage and limitation of multiple correlation?
3. What is regression?

2.4.2 Multiple Correlations

The advantages and limitations of multiple correlation :

(1) Through multiple correlation analysis we intend to measure the degree of association between a single dependent variable and a number of independent variables taken together as a group.

(2) It expresses the type and degree of relationship in a few concise coefficients. For example $_{R1.234}{}^{2}$ =0.85 means that three factors X_2, X_3, and X_4 explain 85%of the squared variability in X_1.

(3) It gives the best predictions that can be computed linearly from the independent variables

Limitations

(1) In multiple correlation it is assumed that the relationships between variables are linear which does not hold good .For example in the field of agriculture most relationships are non-linear, therefore, linear regression coefficients cannot describe non -linear relations accurately.

(2) Also it is assumed that the effects of the independent variables upon the dependent variable are separate, distinct and additive. But it is not possible if interrelation exist among the variables.

(3) The calculation involved in multiple correlation is quite tedious and therefore should be interpreted with caution.

2.4.3 Regressions

Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable. If a regression model characterises the relationship between a dependent y and only one independent x, then such a regression model is called a simple regression model given by

$$\hat{Y} = a + bX \qquad (1)$$

But if more than one independent variable is associated with a dependent variable then such a regression model is called a multiple regression model. In multiple regression analysis, the dependent variable, y, is sometimes referred to as the response variable

A multiple regression equation is an equation which is used for estimating a dependent variable say X_1 from a set of independent variables X_2, X_3,... and is called the regression equation of X_1 on X_2, X_3,.... For two independent variables X_2, X_3 we have the following simplest regression equation of X_1 on X_2 and X_3 and is of the form :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad (2)$$

Where a, $b_{12.3}$ and $b_{13.2}$ are the parameters to be estimated by the Principle of Least Square. The equations so obtained are known as the normal equations

In equation (1) , a is called the X- intercept , $b_{12.3}$ indicates the slope of the regression line of X_1 on X_2called partial regression coefficient ofX_1 on $X_2$when X_3 is held constant . $b_{13.2}$ indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3 when X_2 is held constant. It is called the partial regression coefficient of X_1 on X_3keeping X_2constant.

If we take the deviations of the variables from their respective means and denote these deviations by x_1 , x_2 andx_3 i.e. if

$$x_1 = X_1 - \overline{X}_1$$
$$x_2 = X_2 - \overline{X}_2$$
$$x_3 = X_3 - \overline{X}_3$$

Note :

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$
$$\overline{X}_1 = a + b_{12.3}\overline{X}_2 + b_{13.2}\overline{X}_3$$
$$X_1 - \overline{X}_1 = b_{12.3}(X_2 - \overline{X}_2) + b_{13.2}(X_3 - \overline{X}_3)$$

We have

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad (3)$$

The normal equations in (2) reduces to

$$\Sigma x_2 x_1 = b_{12.3}\Sigma\ x_2^2 + b_{13.2}\Sigma\ x_3 x_2 \qquad (4)$$

$$\Sigma x_2 x_3 = b_{12.3}\Sigma x_3 x_2 + b_{13.2}\Sigma\ x_3^2$$

Solving the two normal equations in (4) the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ can be obtained as

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_2} \text{ and } b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{s_1}{s_3} \qquad (5)$$

Where S_1 , S_2 and S_3 are the standard deviations of X_1, X_2 and X_3 respectively. Thus using equation (3) and equation (5) we get the multiple regression equation of X_1 on X_2 and X_3 as

$$X_1 - \overline{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{23}}\right)\frac{S_1}{S_2}(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2{}_{23}}\right)\frac{S_2}{S_1}\left(X_1 - \overline{X}_1\right) \qquad (6)$$

Similarly the multiple regression equation of X_2 on X_3 and X_1 as

$$X_2 - \overline{X}_2 = \left(\frac{r_{23} - r_{12}r_{13}}{1 - r^2{}_{13}}\right)\frac{S_2}{S_3}(X_3 - \overline{X}_3) + \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2{}_{13}}\right)\frac{S_2}{S_1}\left(X_1 - \overline{X}_1\right) \qquad (7)$$

2.5 Illustration and Examples

Example 1: In a trivariate distribution it is found that r_{12}=0.70 , r_{13}=0.61 , r_{23}=0.40. Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$.

Solution: The partial correlation between X_2 and X_3 holding X_1 constant is determined by

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

Substituting the given values we get

Given $r_{12.4}$=0.60 , $r_{13.4}$=0.50 , $r_{23.4}$=0.70 find $r_{12.34}$ and $r_{13.24}$

Solution:

$$
\begin{aligned}
r_{12.34} &= \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1-r^{2}{}_{13.4}}\sqrt{1-r_{23.4}{}^{2}}} \\
&= \frac{0.6-(0.5\times 0.7)}{\sqrt{\{1-(0.5)^{2}\}}\sqrt{\{1-(0.7)^{2}\}}} \\
&= \frac{0.6-0.35}{\sqrt{0.75\times 0.51}} \\
&= \frac{0.25}{\sqrt{0.3825}} \\
&= \frac{0.25}{0.62} \quad = 0.403
\end{aligned}
$$

$$
\begin{aligned}
r_{13.24} &= \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{1-r^{2}{}_{12.4}}\sqrt{1-r_{23.4}{}^{2}}} \\
&= \frac{0.5-(0.6\times 0.7)}{\sqrt{\{1-(0.6)^{2}\}}\sqrt{\{1-(0.7)^{2}\}}} \\
&= \frac{0.5-0.42}{\sqrt{0.64\times 0.51}} \\
&= \frac{0.08}{0.57} \\
&= 0.14
\end{aligned}
$$

<u>Example 2</u>: In a three variate multiple correlation analysis, the following results were obtained r_{12}=0.59 r_{13}=0.46 and r_{23}=0.77. Find $R_{1.23}$
Solution: Multiple Correlation Coefficient is defined as

$$
\begin{aligned}
R_{1.23} &= \sqrt{\frac{r_{12}^{2} + r_{13}^{2} - 2r_{12}r_{13}r_{23}}{1-r_{23}^{2}}} \\
&= \sqrt{\frac{(0.50)^{2} + (0.46)^{2} - 2(0.59\times 0.46\times 0.77)}{1-(0.77)^{2}}} \\
&= \sqrt{\frac{0.3481+0.2116-0.418}{0.4071}} \\
&= \sqrt{\frac{0.5597-0.418}{0.4071}} \\
&= \sqrt{\frac{0.1417}{0.4071}} = 0.589
\end{aligned}
$$

Stop to Consider

The multiple correlation may be defined as a statistical tool designed to measure the degree of relationship existing among three or more variables. it is denoted by the symbol R

$R_{1.23}$ stands for the coefficients of multiple correlation between X_1, X_2 and X_3. The subscript to the left stands for dependent variable while the subscripts to the right of the dot represent independent variables.

2.6 Summing Up

- Partial and multiple correlations describe the relationship between two variables when influence of one or more other variables is held constant or involved.
- Multiple regressions establish an equation for estimating value of a dependent variable given the value of two or more independent variables
- The value of partial correlation coefficient lies between -1 and +1
- The value of multiple correlation coefficient lies between 0 and 1
- Multiple regression equation: It is an equation which is used for estimating a dependent variable from a set of independent variables

2.7 Model Questions

Short answer questions:

1. Given $r_{12} = 0.5$, $r_{13} = 0.4$ and $r_{23} = 0.1$, find the values of $r_{12.3}$.
2. If $r_{12} = 0.9$, $r_{13} = 0.75$ and $r_{23} = 0.7$, find the values of $R_{1.23}$.

Long answer questions:

1. In a trivariate distribution:

   $\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5,\ r_{23} = 0.4,\ r_{31} = 0.6,\ r_{13} = 0.7$

   Determine the regression equation of X_1 on X_2 and X_3 if the variates are measured from their means.
2. The following data are given:

   $x_1 = 6,\ x_2 = 7,\ x_3 = 8;\ s_1 = 1 s_1 = 2,\ s_1 = 3;\ r_{23} = 0.8,\ r_{12} = 0.6,\ r_{13} = 0.7$

   a) Find the regression equation of x_3 on x_1 and x_2

   b) Estimate the value of x_3 when $x_1 = 4$ and $x_2 = 5$

2.8 References and Suggested Readings

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.

2. Business Statistics: S. Saha, New Central Book Agency.

3. Basic Statistics:B. L. Agarwal, New Age International Limited.

4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.

5. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.

6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited.

******

Unit 3

Coefficient of Multiple Determinations, Association of Attributes: Concept, Order of a Class, Class Frequency, Consistency of Data

Unit Structure:

3.1 Introduction

In this unit you will learn about the concept of multiple determination and adjusted coefficient of determination. . In multiple regressions, the coefficient of multiple determinations represents the proportion of the variation in Y that is explained by the set of independent variables selected. The Adjusted Coefficient of Determination (Adjusted R-squared) is an adjustment for the Coefficient of Determination that takes into account the number of variables in a data set. The unit also discusses the concept of association of attributes and discusses about consistency of data**.** Qualitative variables are called attributes and theory of attributes deals with the measurement of data whose magnitude cannot be directly measured numerically. Though, the qualitative data can be quantified but for the sake of clear understanding and convenience, the statistical methodologies for the analysis of qualitative data have been separately developed. Association of two attributes measure the degree of relationship between two phenomena whose sizes cannot be measured but one can only determine the presence or absence of a particular attribute or quality. The unit also discusses about the consistency of the data, the conditions for consistency and the independence of attributes. A data is said to be consistent if no class frequency turns out to be negative.

3.2 Objectives

After going through this unit, you will be able to

- Learn about the coefficient of multiple determination
- Learn about the association of attributes
- Describe the class and class frequency
- Discuss about the consistency of data

3.3 Coefficient of Multiple Determination (R^2):

The coefficient of determination can be thought of as a percent which gives an idea of how many data points fall within the line formed by the regression equation. It is represented by the symbol R and represents the proportion of the total variation in the dependent variable y , accounted for or explained by the independent variables in the multiple regression model . The value of R^2 lies between 0 and 1. It is given by

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

The higher the coefficient, the higher percentage of points pass through the line when the data points and the line are plotted. If the coefficient is 0.80, then 80% of the points should fall within the regression line. A higher coefficient is an indicator of a better goodness of fit for the observations.

3.4 Adjusted Coefficient of Determination

Sometimes additional independent variables added to multiple regression model increase R^2 while their contribution is insignificant in explaining the variation in the dependent variable. To correct for this defect, R^2 is adjusted by taking into account the degrees of freedom which decreases with inclusion of additional independent or explanatory variables in the model.

The following equation shows the relationship between adjusted $\bar{R}^2$ and R^2.

$$\bar{R}^2 = 1-(1-R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

$$\begin{aligned}&where\ n = sample\ size\\ &\quad k\ = the\ number\ of\ independent\ \text{var}iables\ in\ the\ regression\ equation\end{aligned}$$

One major difference between R-square and adjusted $\bar{R}-square$ is that R-square supposes that every independent variable in the model explains the variation in the dependent variable. It gives the percentage of explained variation as if all independent variables in the model affect the dependent variable, whereas adjusted R-square gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable.

Example 3.1 Suppose R^2= 0.944346527, k=2 and n=8. Calculate Adjusted R^2.

Solution:

$$\bar{R}^2 = 1-(1-R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

$$\begin{aligned}&where\ n = sample\ size\\ &\quad k\ = the\ number\ of\ independent\ \text{var}iables\ in\ the\ regression\ equation\end{aligned}$$

$$\begin{aligned}&= 1-(1-0.944346527)\left[\frac{8-1}{8-(2+1)}\right]\\ &= 0.922085138\end{aligned}$$

Based on the value of R^2, the proportion of variation explained by the estimated regression line is approximately 0.922 or 92.2 percent.

3.5 Concept of Association of Attributes

Correlation as a statistical tool is used to measure the degree of relationship between two phenomenon which are capable of direct quantitative measurement. Data regarding such phenomenon are known as statistics of variables. On the other hand, certain phenomenon like blindness, deafness etc. are not capable of direct quantitative measurements are called attributes. Such phenomenon is studied only on the presence or absence of the attribute. Thus, method of association of attributes is employed to measure the degree of relationship between two phenomena which we cannot measure and where we can only determine the presence or absence of a particular attribute.

3.6 Classes and Class Frequencies

While dealing with statistics of attributes, the data has to be classified .The classification is done on the basis of presence or absence of a particular attribute or characteristic. When we are studying only one attribute two classes are formed- one possessing that attribute and another not possessing it. When two attributes are studied four classes are formed. The number of observations assigned to any class is termed as 'class frequency'. Class frequencies are represented by enclosing the corresponding class symbols in brackets. Hence (A) stands for the number of items or individuals who possess the attribute A. (AB) stands for the number of items who possess both the attributes A and B , (Aβ) denotes the number of items possessing attribute A but not possessing attribute B and so on . The classes having one or more positive attributes are called positive classes and classes having one or more negative attributes are called negative classes.

Number of Classes: The total number of classes comprising of the various attributes are given by 3^n , n representing the number of attributes . If one attribute is studied , then thre will be 3^1 =3 classes . Thus if literacy is studied , the presence of literacy is represnted by A, its absence by α and total by N. Therefore there will be 3 classes i.e A, α and N. If two attributes are studied , the number of classes will be 3^2 =9 . The 9 classes are A, α , B, β , AB, α B, A β , α β and N.

If three attributes are studied, the number of classes will be 3^3=27 classes.

Class Frequencies

The number of observations assigned to any class is called the '**frequency of the class' or 'class frequency'. Class frequencies are generally denoted by enclosing the corresponding class symbols in small brackets ( ).**

Let 'A' be an attribute say 'Honesty'

(A) denotes the number of persons possessing 'A' attribute

i.e (A) =10 means that there are ten persons who are honest

Also 10 is the frequency of the class

If 'B' stands for male , then β would mean female

If 'A' stands for honest , then α would mean non-honest

Therefore ,

(AB)= 30, means that there are 30 males who are honest

$(A\beta)$ = 10 , means there are 10 females who are honest

$(\alpha\beta)$ =15 , means there are 15 females who ar not honest and so on

3.6.1 Order of a Class

The order of a class depends upon the number of attributes under study . A class having one attribute is known as the class of first order. A class having two attributes is called the class of second order and so on.

The total number of observations denoted by the symbol N is called class or frequency of zero order. Since no attributes are specified. Thus we have:

N	: Class (or Frequency) of Zero Order
(A), (B), (α), (β)	: Class (or Frequency) of First Order
(AB), $(A\beta)$, (αB) , $(\alpha\beta)$	: Class (or Frequency) of Second Order

3.6.2 Class Frequencies

The four groups (AB), $(A\beta)$, (αB) , $(\alpha\beta)$ are called ultimate class frequencies . They are represented by enclosing the corresponding class symbols viz, AB, α B, A β and $\alpha\beta$ in the small brackets ( ).

The number of ultimate classes is determined by 2^n, where n is the number of attributes. Thus for one attribute , there will be 2^1=2 ultimate classes; for two attributes, there will be 2^2=4 ultimate classes ; and for three attributes , there will be 2^3= 8 ultimate classes.

The total number of observations is equal to the positive and negative frequencies of the same class of the first order i.e ,

$$N=(A)+(\alpha) \quad \text{or } N=(B)+(\beta)$$

Moreover,

$$(A) = (AB)+(A\beta) \; ; \qquad (B)=(AB)+(\alpha B)$$

$$(\alpha) = (\alpha B)+(\alpha\beta) \; ; \qquad (\beta)=(A\beta)+(\alpha\beta)$$

The frequencies of the positive , negative and ultimate classes can be known from the following table which is known as the the contingency table of order (2x2) for two attributes A and B is shown below:

	A	α	Total
B	(AB)	(α B)	(B)
β	(A β)	(α β)	(β)
Total	(A)	(α)	N

3.6.3 Relationship Between the Class Frequencies

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies as follows:

N=(A) +(α)
(A) = (AB)+(A β)
(B) = (AB)+(α B)
(α) = (α B)+(α β)
(β) = (A β) +(α β)

If the number of attributes is n, there will be 3^n classes and 2^n class frequencies.

Stop to Consider

- Phenomenon like blindness, deafness etc. not capable of direct quantitative measurements are called attributes.
- The number of observations assigned to any class is termed as 'class frequency'.
- Method of association of attributes is employed to measure the degree of relationship between two phenomena which we cannot measure and where we can only determine the presence or absence of a particular attribute.

Check Your Progress

1. What do you understand by co-efficient of Multiple Determination (R^2)
2. What is the major differences between R^2 and adjusted R^2?
3. What is association of Attributes?

3. A coefficient which examines the association between a dependent variable and an independent variable after factoring out the effect of other independent variables is known as ________.

4. A statistical technique that develops an equation that relates a dependent variable to one or more independent variables is called _______.

5. The total variation explained by a regression model is given by _______.

6. In statistics, ______ means a quality on characteristic which are not related to quantitative measurements and hence cannot be measured directly.

Ex1. From the following data find the missing frequencies

(AB)= 50 , (αβ)=25 , (α)= 100, N=250

Solution : The missing frequencies are (Aβ) , (β), (αB) , (A) and (B).

Putting these values in the contingency table , we have

	A	α	Total
B	50	75	125
β	100	25	125
Total	150	100	250

(α)= (αB)+(αβ)

100= (αB)+25

(αB)=100-25=75

(B)= (AB)+(αB)

= 50+75

= 125

N=(B)+(β)

250= 125+ (β)

(β)= 250-125 ==125

(α)= (αB)+(αβ)

(β)=(Aβ)+(αβ)

125= (Aβ) + 25

(Aβ)=125-25=100

(A)= (AB)+(Aβ)

=50+100

=150

Hence (A)= 150 , (Aβ)=100 , (β)= 125, (αB)=75 and (B)= 125

3.7 Consistency of Data.

Consistency of Data: A given set of data is said to be consistent if none of the class frequencies is negative. If any class frequency is negative, the given set of data is said to be inconsistent.

Exercise 4.0 Test the consistency in the data given below:

(i)	N=1200	(A)=600	(AB)=400	(B)= 500
(ii)	N=1000	(AB)=200	(A)=150	(B)=300

Solution:

Case I We are given (A) =600 (AB) =400 (B) =500 and N=1200

We substitute these values in the following contingency table:

	A	α	Total
B	(AB) =400	(α B)=100*	(B)=500
B	(A β) =200*	(α β) =500*	(β) =700*
Total	(A) =600	(α)=600	N =1200

From the table

(A β) =(A) - (AB) =600-400=200

(α B)= (B)-(AB) = 500-400= 100

(α β) =(α)- (α B) =600=100=500

Since all the ultimate class frequencies are positive we conclude that the given daa is consistent.

Case II:

We are given (A)=150 (AB)=200 (B)= 300 and N=1000

We substitute these values in the following contingency table:

	A	α	Total
B	$(AB) = 200$	$(\alpha B) = 100^*$	$(B) = 300$
B	$(A\beta) = -50$	$(\alpha\beta) = 150$	$(\beta) = 700$
Total	$(A) = 150$	$(\alpha) = 850$	$N = 1000$

From the table

$(A\beta) = (A) - (AB) = 150 - 200 = -50$

$(\alpha B) = (B) - (AB) = 300 - 200 = 100$

$(\alpha\beta) = (\alpha) - (\alpha B) = 850 - 100 = 750$

We find that one of the ultimate class frequencies i.e. $(A\beta)$ is negative and hence the given data is inconsistent.

3.8 Summing Up

- In statistics, an attribute means a quality on characteristic which are not related to quantitative measurements and hence cannot be measured directly.
- Association of attributes relates to the study of relationship between two or more attributes.
- Class frequency denotes the number of individuals who possess that attribute.
- A given set of data are said to be consistent if none of the class frequencies is negative.
- Two attributes A and B are said to be independent in case the presence or absence of one attribute has no relationship with the presence or absence of other attribute.

3.9 Key Words

The coefficient of determination is multiple regression represents the proportion of the total variation in the multiple values of the dependent variable y accounted for or explained by the independent variables in the multiple regression model.

Attribute: In statistics, an attribute means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.

Classes refer to different attributes, their subgroups and combinations. The number of observations assigned to them is called their class frequencies.

The attributes are said to be independent when the presence or absence of one attribute does not affect the presence or absence of the other.

In order to find the degree of association between two or more sets of attributes, the coefficient of association is used.

3.10 Model Questions

A. Multiple Choice Questions

1. A quality on characteristic which are not related to qualitative measurement is called

   (a) attribute (b) class frequency

   (c) classes (d) none of these

2. The coefficient is used to

   (a) find the degree of association between two or more sets of attributes.

   (b) find the degree of association between two subgroups.

   (c) find the degree of association between the frequencies

   (d) none of these

B. State whether True or False:

1. The closer the coefficient of multiple correlation is to 1, the better the relationship between the variables.
2. The coefficient of multiple correlation is the square root of coefficient of multiple determination.
3. Classes refer to different attributes, their sub-groups and combinations.
4. The number of attributes attached to any class is termed its class frequency.

C. Short answer questions:

1. Define association of attributes.
2. How is the coefficient of association calculated?
3. When is a set of data said to be consistent?

D. Long answer questions:

1. Under what condition is it important to use the adjusted multiple coefficient of determination ?.

2. Test the consistency of the data given below:

   Case I : (AB) = 200, (A) = 300, (α) = 200, (B) = 250, (αβ) = 150, (N) = 500

   Case II : (AB) = 250, (A) = 150, (α) = 1000, (B) = 500, (αβ) = 600, (N) = 1750

3. From the following ultimate frequencies , find the frequencies of the positive and negative classs, and the total number of observations

   (AB)=100 ; (αB)=8 , (Aβ)=50 , (αβ)=40

3.11 References and Suggested Readings

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.
2. Business Statistics: S. Saha, New Central Book Agency.
3. Basic Statistics:B. L. Agarwal, New Age International Limited.
4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.
5. Quantitative Techniques for Decision Making:Anand Sharma, Himalaya Publishing House.
6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited

******

Unit 4

Kinds of Association of Attributes, Methods of Measuring Association Between Two Attributes, Partial Association

Unit Structure:

4.1 Introduction

In this unit, you will learn about the association of attributes. Two attributes are said to be associated if they are not independent but are related .The method of association of attributes is employed to measure the degree of relationship between two phenomena which cannot be measured but its presence or absence can only be determined. The phenomena's which cannot be measured quantitatively, i.e beauty, honesty, insanity, deafness etc. The observations possessing a particular quality, say , honesty are grouped together , counted and given numerical shape i.e they are quantified. There are different kinds of association between attributes and a quantitative measure for nature and the degree of association can be found using different methods.

4.2 Objectives

After going through this unit, you will be able to

- learn about the association of attributes,
- learn about the kinds of association of attributes,
- explore the different methods of association of attributes.

4.3 Kinds of Association of Attributes

There can be three kinds of association between attributes

(1) **Positive Association :** When two attributes are present or absent together in the data, it is known as positive association. Such association is found between literacy and employment, smoking and cancer, vaccination and immunity from a disease, etc.

(2) **Negative Association:** When the presence of an attribute is associated with the absence of the other attribute, it is called negative association. Such association is found between vaccination and attack of a disease, cleanliness and ill-health, etc.

(3) **Independence:** When there exists no association between two attributes or when they have no tendency to be present together or the presence of one attribute does not affect the other attribute, the two attributes are regarded as independent.

Thus two attributes A and B may be (i) positively associated (ii) negatively associated or (iii) independent based on the following conditions :

Thus

If $(XY) > \frac{(X)(Y)}{N}, then\ X\ and\ Y\ are\ positively\ associated$

$(XY) < \frac{(X)(Y)}{N}, then\ X\ \ and\ \ Y\ are\ \ negatively\ associated$

$(XY) = \frac{(X)(Y)}{N}, then\ X\ and\ Y\ are\ independent$

N being the number of items

Exercise 4.1 Show whether A and B are independent, positively associated or negatively associated for the following values:

(AB) =128 (α B) = 384 (A β)=24 (α β)=72

Solution: Using the given values we have

(A) = (AB)+(Aβ)

= 128+24

= 152

(B) = (AB) + (α B)

= 128+384

= 512

$(\alpha) = (\alpha B) + (\alpha\beta)$

$= 384+72$

$= 456$

$(N) = (A)+(\alpha) = 152+456 = 608$

Thus

$$\frac{(A)\times(B)}{N} = \frac{152\times152}{608} = 128$$

$(AB) = 128$

Therefore $(AB) = \frac{(A)\times(B)}{N}$

Hence A and B are independent.

Check Your Progress

1. When is a set of data said to be consistent ?
2. Define positive and negative association.
3. When are two attributes independent ?
4. In statistics, ______ means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.
5. When observed and expected frequencies are equal, then both attributes are __________

4.4 Methods of Studying Association Between Two Attributes

The nature and degree of association between two or more attributes can be studied by using the following methods

(1) Proportion Method : This method consists in comparing the presence or absence of a given attribute in the other

Two attributes A and B are said to be :

Positively associated if $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)} \quad or \quad \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$

Negatively associated if $\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)} \quad or \quad \frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$

However , if $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad or \quad \frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$,

$then\ A\ and\ B\ are\ independent$

This method can only determine the nature of association between attributes that is whether it is positive or negative or no association but it does not study the degree of association whether it is high or low.

(2) Comparison of Observed and Expected Frequencies

Two attributes A and B are said to be

Positively associated if $(AB) > \frac{(A)(B)}{N}$

Negatively associated if $(AB) < \frac{(A)(B)}{N}$

And independent if $(AB) = \frac{(A)(B)}{N}$

Similar expressions can be obtained for other ultimate class frequencies such as (Aβ) , (αB) and (αβ)

The main limitation of this method is that with the help of this method we can only find out the nature of association between the attributes, whether the association between them is Positive, Negative or Independent. We cannot determine the degree of association.

(3) Yule's Coefficient of Association : In order to understand properly the significance of association or the relationship between two or more attributes ,it is necessary to find the degree of association between them.

The nature and degree of association between two attributes can be found out by applying the following formula given by Prof.G.V Yule

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha\beta)}{AB)(\alpha\beta) + (A\beta)(\alpha\beta)}$$

Where Q denotes coefficient of association. The value of Q lies between -1 and 1.

If the attributes are independent of each other, the coefficient of association will be zero.

If the attributes are perfectly or positively associated, the coefficient will be +1.

If they are completely negatively associated or disassociated , the coefficient will be -1. thus the value of coefficient of association ranges from -1 to +1.

The degree of association is measured by the coefficient of association given by Prof. YULE is as follows:

$$Q = (AB) \times (\alpha\beta) - (A\beta) \times (\alpha\beta)$$

$$(AB) \times (\alpha\beta) + (A\beta) \times (\alpha\beta)$$

Where : Q is coefficient of Association.

CHARACTERISTIS OF YULE'S COEFFICIENT OF ASSOCIATION

1) If $Q = 0$ there is no association.

   $Q = +1$ the association is positive and perfect.

   $Q = - 1$ the association is negative and perfect.

   Generally Q lies between +1 and -1.

   Yule's coefficient is independent of the relative proportion of A's and α's in the data. The value of the coefficient remains the same if all the terms containing A, α, B, β are multiplied by a constant.

   Yule's coefficient of association is superior because it provides information not only on the nature, but also on the degree of association .

(2) Coefficient of Colligation

   Prof. YULE has given another important coefficient which is also independent of the relative proportion of A's and α's is known as coefficient of colligation and is denoted by γ **(gamma)** which can be calculated with the help of following formula:

   This measure of association is given by Prof. Yule. The measure is denoted by the symbol Y and is given by

$$Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha B)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha B)}}}$$

4.5 Partial Association

Partial association is also known as association in a sub universe. If two attributes A & B are associated with each other it is likely that this association may be due to the association of attributes A with C and attributes of B with C. Thus association of A & B in the sub population C is known as Partial Association.

4.6 Summing Up

Positive Association: When two attributes are present or absent together in the data, it is known as positive association.

Negative Association: When the presence of an attribute is associated with the absence of the other attribute, it is called negative association.

Independence. When there exists no association between two attributes or when they have no tendency to be present together or the presence of one attribute does not affect the other attribute, the two attributes are regarded as independent.

In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Yule's coefficient of association is superior because it provides information not only on the nature, but also on the degree of association.

Generally Q lies between +1 and -1.

Stop to Consider

- Attribute: In statistics, an attribute means a quality or characteristic which are not related to quantitative measurements and hence cannot be measured directly.
- There can be three kinds of association between attributes.
- Yule's Coefficient of Association measures the degree of association between attributes.

4.7 Model Questions

(A) Multiple Choice Questions

1. The coefficient is used to

   (a) find the degree of association between two or more sets of attributes.

   (b) find the degree of association between two subgroups.

   (c) find the degree of association between the frequencies

   (d) none of these

2. The value of Q lies between

   (a) 0 and 1

   (b) -1 and 1

   (c) -1 and 0

   (d) None of the above

B. State whether True or False:

1. Classes refer to different attributes, their sub-groups and combinations.
2. The number of attributes attached to any class is termed its class frequency.

C. Short answer questions:

1. Define association of attributes.
2. How is the coefficient of association calculated?
3. What is partial association?

D. Long answer questions:

1. Compute the following table and find Yule's coefficient of association

   N = 800, (A) = 470, (B) = 450 and (AB) = 230

2. Test the consistency of the data given below:

   Case I : (AB) = 200, (A) = 300, (α) = 200, (B) = 250, ($\alpha\beta$) = 150, (N) = 500

   Case II : (AB) = 250, (A) = 150, (α) = 1000, (B) = 500, ($\alpha\beta$) = 600, (N) = 1750

3. A teacher examined 280 students in Economics and Auditing and found that 160 failed in Economics, 140 failed in Auditing and 80 failed in both the subjects. Is there any association between failure in Economics and Auditing?

4. Eighty–eight residents of an Indian city, who were interviewed during a sample survey and classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of Association and comment on its value.

	Smokers	Non-Smokers
Drinking Tea	40	33
Non- Drinking	3	12

4.8 References and Suggested Readings

1. Essential statistics for Economics and Business Studies: Padmalochan Hazarika, Akansha Publishing House, New Delhi.

2. Business Statistics: S. Saha, New Central Book Agency.

3. Basic Statistics: B. L. Agarwal, New Age International Limited.

4. Statistics for Management: Prentice Hall of India Limited, Levin Richard & Rubin David.

5. Quantitative Techniques for Decision Making: Anand Sharma, Himalaya Publishing House.

6. Methods of Statistical Analysis : P.S Grewal, Sterling Publishers Private Limited.

******