Unit 19

Correlation and Regression

Contents

Learning Objectives

It is expected that after reading Unit 19 you would be able to

- Appreciate the relevance of the analysis of co-variation between two or more variables
- Describe different types of correlation
- Elaborate methods of calculating correlation of both ungrouped and grouped data
- Understand the method of regression analysis that helps in estimating the values of a variable from the knowledge of one or more variables.

19.1 Introduction

In the concluding Section of Unit 18, we mentioned the linkages between propositions. Let us now discuss the subject of correlation and regression.

Unit 19 is about correlation, that is an analysis of co-variation between two or more variables. You would notice that the statistical tool of correlation helps to measure and express the quantitative relationship between two variables. Unit 19 elaborates the ways of applying the tool. It shows the relevance of coefficient of correlation, coefficient of determination and regression analysis in the social sciences. Further, it explains regression analysis, which is the method of estimating the values of a variable from the knowledge of one or more variables. The unit tells you to use the statistical tool of correlation without fear or apprehension that its application is difficult and complex.

19.2 Correlation

Correlation® is an analysis of the co-variation between two or more variables. When the relationship between the two variables is quantitative, the statistical tool for measuring the relationship and expressing it in a brief formula is known as correlation. If a change in one variable results in a corresponding change in the other, the two variables are correlated. Let us look at types of correlation.

Types of correlation

Probing into the types of correlation, we contemplate two types : correlation:

A) Positive and Negative correlation;

B) Linear and Non-linear correlation

A) Positive and negative correlation

If the values of the two variables deviate in the same direction, i.e., if an increase in the value of one results on an average in a corresponding increase in the value of the other, or if decrease in the value of one variable results in a decrease in the value of the other, then correlation is said to be *positive or direct*. Some examples of a series of positive correlation are (i) height and weight (ii) land owned and household income. On the other hand, if the variables deviate in the opposite directions, i.e. if an increase (decrease) in the value of one variable, on an average, results in a decrease (increase) in the value of the other variable, then the correlation is *negative or indirect*. Some examples of negative correlation are (i) physical assets and the level of poverty, (ii) muscle strength and age. Figure 19.1 shows the positive and negative types of correlation.

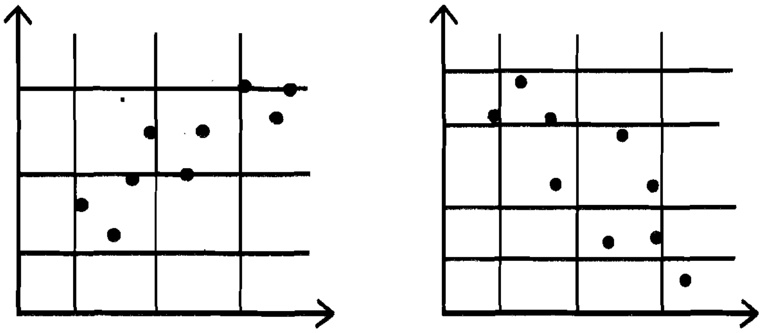

Figure 19.1 (a) Positive Correlation and (b) Negative Correlation

The values of correlation range from -1 to +1.When r = +1, it means there is perfect positive correlation between the variables. When r = -1, there is perfect negative correlation. When r = 0, it means there is no correlation between the two variables (see Figure 19.2).

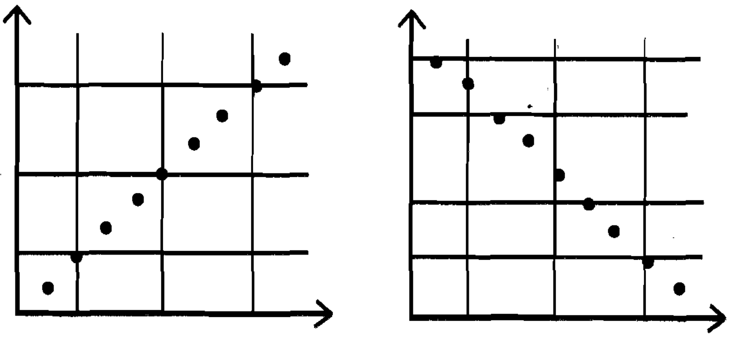

Figure 19.2 (a) Perfect Positive Correlation (r=+1) and (b) Perfect Negative correlation (r =-1)

B) Linear and non-linear correlation

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values. Consider the following data in Figure 19.3.

X	1	2	3	4	5	6
Y	3	5	7	9	11	13

Figure 19.3 Constant Change Figuring in the Entire Range of Values

In this case, the data in Figure 19.3 can be represented by the relation Y=1 + 2 X. In general, two variables are said to be linearly related if there exists a relationship of the form Y=a + b X.

On the other hand, the relationship between the two variables is said to be non-linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant but a fluctuating rate. Example of a non-linear correlation is given by the following data set in Figure 19.4.

X	1	2	3	4	5	6
Y	5	8	14	15	18	22

Figure 19.4 Non-linear Correlation

In the example in Figure 19.4, there is fluctuating (not constant) change in the value of Y corresponding to a unit change in the value of X, and thus it represents a non-linear correlation.

You would like to know how to study correlation. Let us briefly discuss the methods of studying correlation. But before going on to metods of studying correlation, let us complete Reflection and Action19.1

Reflection and Action 19.1

Relating to your hypothesis, draw the figure of its positive and negative correlations. Next draw another figure of perfect positive and perfect negative correlations. In addition, draw two more figures of constant change reflected in the entire range of values and non-linear correlation. You may take help of Figures 18.1 to 18.4 in the text above for drawing your figures.

Methods of studying correlation

The various methods to determine whether there is a correlation between two variables are (i) Scatter diagram; (ii) Graphic method; (iii) Karl Pearson's coefficient of correlation; (iv) Rank method; (v) Concurrent deviation method; and (vi) Method of least squares. Of these, the first two are based on the knowledge of diagrams and graphs and the rest on mathematical tools. Of the several mathematical tools used, the most popular is the Karl Pearson coefficient of correlation (r) and thus we will focus on this method. The procedure is different for calculating correlation from ungrouped and grouped data.

19.3 Method of Calculating Correlation of Ungrouped Data

There are various methods for the calculation of the coefficient® of correlation from ungrouped data.

i) Using actual mean
ii) Using assumed mean
iii) Direct method

The use of all these methods is illustrated with the help of the following example.

Example: Find out the correlation coefficient (Karl Pearson's) between the age at marriage of husbands and wives using the following data in Figure 19.5

Age at Marriage	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	*Case 9*	*Case 10*
Husbands	28	25	24	29	31	22	21	25	26	28
Wives	22	23	21	25	26	20	19	21	21	24

Figure 19.5 Correlation Coefficient between the Age at Marriage of Husbands and Wives

Method of calculating correlation coefficient using the actual mean

You would first learn the method of calculating correlation coefficient using the actual mean and then you would actually carry out the calculation itself.

The formula used for calculating r is:

$$r = \Sigma xy / N^{*} \acute{o}_{x}^{*} \acute{o}_{Y}$$

Where, $x = (X - M_{x})$ in which M_{x} is the mean of series of X values;

$y = (Y - M_{Y})$ in which M_{y} is the mean of series of Y values;

$\acute{o}_{x}$ = Standard deviation of series X

$\acute{o}_{Y}$ = Standard deviation of series Y

N = Number of pair of observations

This formula can also be expressed as:

$$r = \Sigma xy / \Sigma [\acute{O}x^{2*} \acute{O}y^{2}]$$

The following steps elucidate the calculation of the coefficient of correlation.

I. Take deviations of X series from the mean of X and denote them by x;

II. Square these deviations and obtain the total, i.e. Óx^2;

III. Take deviations of Y series from the mean of Y and denote them by y;

IV. Square these deviations and obtain the total, i.e. Óy^2;

V. Multiply the deviations of x and y and obtain the total Ó xy; and

VI. Substitute the values of Ó$x^{2,}$ Ó$y^{2,}$ and Ó xy in the above formula.

Calculation of correlation coefficient using actual mean

After learning the method, let us now make the calculation as reflected in Figure 19.6

X	$x = X - M_X$	X^2	Y	$y = Y - M_Y$	y^2	Xy
28	2.1	04.41	22	-0.2	00.04	-00.42
25	-0.9	00.81	23	0.8	00.64	-00.72
24	-1.9	03.61	21	-1.2	01.44	02.28
29	3.1	09.61	25	2.8	07.84	08.68
31	5.1	26.01	26	3.8	14.44	19.38
22	-3.9	15.21	20	-2.2	04.84	08.58
21	-4.9	24.01	19	-3.2	10.24	15.68
25	-0.9	00.81	21	-1.2	01.44	01.08
26	0.1	00.01	21	-1.2	01.44	-00.12
28	2.1	04.41	24	1.8	03.24	03.78
259	0	88.90	222	0	45.60	58.20

Figure 19.6 Calculation of Correlation Coefficient using Actual Mean

r = Σxy / ? [Óx^{2*} Óy^2]

M_X = 259 / 10 = 25.9 M_Y = 222 / 10 = 22.2 (Óx^2) = 88.9 (Ó y^2) =45.6 Ó xy = 58.2

r= 58.2 / v [88.9 * 45.6] = 0.914

Method of calculating correlation coefficient using assumed mean

The only difference in this method as compared to the above method is that in the former, the deviations are taken from the actual mean, and in this case from the assumed mean (i.e. by looking at the series of X and Y, assume means for X and Y and proceeding in the same manner).

Calculation of correlation coefficient using assumed mean

You would now calculate as per Figure 19.7.

X	D_X=X - Ax	d_X^2	Y	d_Y = Y- Ay	d_Y^2	d_X*d_Y
28	3	9	22	0	0	0
25	0	0	23	1	1	0
24	-1	1	21	-1	1	1
29	4	16	25	3	9	12
31	6	36	26	4	16	24
22	-3	9	20	-2	4	6
21	-4	16	19	-3	9	12
25	0	0	21	-1	1	0
26	1	1	21	-1	1	-1
28	3	9	24	2	4	6
259	9	97	222	2	46	60

Figure 19.7 Calculation of correlation coefficient using assumed mean

$N\Sigma\ d_X{}^*\ d_Y$ - (? d_X * ? d_Y)

r = ——————————————————

v {N Σ d_x^2 - (Σd $_{x)}{}^{2}$}* v {N Σ d_y^2 - (Σ d $_y)^2$

10*60 - (9* 2)

r = ——————————————————

v {10*97 - (9) ^{2}}* v {10*46 - (2) ^{2}

582

r = ————

636.697

r = 0.914

Direct method of calculating correlation coefficient

The coefficient can also be calculated by taking actual X and Y values, without taking deviations either from the actual or assumed mean. The formula for its calculation is as follows.

r = (N * ΣXY - ΣX * ΣY) / v [N * ΣX^2 - $(\Sigma X)^2$] * v [N * ΣY^2 - $(\Sigma Y)^2$]

The direct method gives the same answer as one gets when deviations are taken from the assumed or actual means. The example demonstrates this point in Figure 19.8.

X	Y	X^2	Y^2	XY
28	22	784	484	616
25	23	625	529	575
24	21	576	441	504
29	25	841	625	725
31	26	961	676	806
22	20	484	400	440
21	19	441	361	399
25	21	625	441	25
26	21	676	441	546
28	24	784	576	672
259	222	6797	4974	5808

Figure 19.8 Calculation of correlation coefficient using direct method

Let us now complete Reflection and Action 19.2 and then learn in Section 19.4 the methods of calculating correlation of grouped data.

Reflection and Action 19.2

Select one of the following two calculations and carry it out in relation to your hypothesis. You need not worry about making mistakes in your calculations. At the moment the idea is to learn the procedure. This is not to be a part of your report.

i) Calculation of correlation coefficient using assumed mean

ii) Calculation of Correlation Coefficient using Direct Method

19.4 Method Of Calculating Correlation Of Grouped Data

With a large number of observations, the data is concealed into a two-way frequency distribution called correlation table. The class intervals of Y series are written as column headings and that of the X series are written as row headings. The frequency distribution for the two variables is written in the respective cells. The formula for calculating the coefficient of correlation is:

$$r = \frac{\Sigma\, f\, {*}d_X{*}\, d_Y - (\Sigma\, f_X{*}d_X\, {*}\, \Sigma\, f_Y{*}d_Y)\ /\ N}{v\,\{\, \Sigma\, f_X{*}d_X^2 - (\Sigma\, f_X{*}d_X)^2\ /\ N\}{*}\ v\,\{\, \Sigma f_Y{*}\, d_Y^2 - (\Sigma f_Y{*}d_Y)^2 / \ N\,\}}$$

Steps:

i) Take the step deviations of variable X and denote these deviations by d_X

ii) Take the step deviations of variable Y and denote these deviations by d_Y

iii) Multiply $d_X {*}d_Y$ and the respective frequencies for each cell and write the figure obtained in the right hand upper corner of the cell.

iv) Add together all values to obtain $\Sigma\, f\, {*}d_X{*}\, d_Y$

v) Multiply all the frequencies of the variable X by the deviations of X and obtain the total $\Sigma\, f_X{*}d_X$

vi) Take the squares of the deviations of the variable X and multiply by respective frequencies to obtain $\Sigma\, f_X{*}d_x^2$

vii) Multiply all the frequencies of the variable Y by the deviations of Y and obtain the total $\Sigma\, f_Y{*}d_Y$

viii) Take the squares of the deviations of the variable Y and multiply by respective frequencies to obtain $\Sigma\, f_Y{*}d_Y^2$

ix) Substitute the values for $\Sigma\, f_Y{*}d_Y^2$, $\Sigma\, f_Y{*}d_{Y,}$ $\Sigma\, f_X{*}d_X^2$, $\Sigma\, f_X{*}d_{X,}$ $\Sigma\, f\, {*}d_X{*}d_Y$ in the above formula to get the value of r.

Let us now take an example to calculate the Karl Pearson's coefficient of correlation using the data in Figure 19.9.

Expenditure on Luxury Items	(Income in Thousand Rs.)				
	20 - 25	25 - 30	30 - 35	35 - 40	40-45
0 - 4	28	12	05		
4 - 8	41	22	09	03	
8 - 12	09	33	28	14	16
12 -16		18	22	29	37
16-20			03	09	12

Figure 19.9 Coefficient Correlation regarding Expenditure on Luxury Items

We can calculate correlation coefficient in grouped data using direct method as seen in Figure 19.10 (See figure 19.10).

Expenditure on Luxury Items	Income in Thousand Rs.)								
	20 - 25	25 - 30	30 - 35	35 - 40	40-45	fy	DY	fy*dy	fy*dy*dy
0 - 4	28	12	5			45	-2	-90	180
4 - 8	41	22	9	3		75	-1	-75	75
8 - 12	9	33	28	14	16	100	0	0	0
12 -16		18	22	29	37	106	1	106	106
16 -20			3	9	12	24	2	48	96
Fx	78	85	67	55	65	350		-11	457
dx	-2	-1	0	1	2				
fx*dx	-156	-85	0	55	130	-56			
fx*dx*dx	312	85	0	55	260	712			

Figure 19.10 Calculation of Correlation Coefficient in Grouped Data

Now we can proceed to calculate fx*dx*dy using direct method as given in Figure 19.11.

Expenditure on Luxury Items	Income in Thousand Rs.)					fx*dx*dy
	20 - 25	25 - 30	30 - 35	35 - 40	40-45	
0 - 4	112	24	0	0	0	136
4 - 8	82	22	0	-3	0	101
8 - 12	0	0	0	0	0	0
12 -16	0	-18	0	29	74	85
16 - 20	0	0	0	18	48	66
fx*dx*dy	194	28	0	44	122	388

Figure 19.11 Calculation of Correlation Coefficient of Grouped Data

N = 350 $\Sigma f *d_X * d_Y = 388$ $\Sigma f_X*d_X = -56$

$\Sigma f_Y*d_Y = -11$ $\Sigma f_X*d_X^2 = 712$ $\Sigma f_Y* d_Y^2 = 457$

$\Sigma f *d_X* d_Y - (\Sigma f_X*d_X * \Sigma f_Y*d_Y) / N$

r = ————————————————————————————————

$v \{\Sigma f_X{*}d_X^2 - (\Sigma f_X{*}d_X)^2 / N\}{*}\ v \{\Sigma f_Y{*}\ d_Y^2 - (\Sigma f_Y{*}d_Y)^2 / N\}$

$$r = \frac{388 - (-56 * -11) / 350}{v \{712 - (-56)^2 / 350\}{*}\ v \{457 - (-11)^2 / 350\}}$$

$r = 386.24 / (26.515 * 21.369) = .682$

Most of the variables show some kind of relationship. With the help of correlation one can measure the degree of relationship between two or more variables. Correlation, however, does not tell us anything about the cause and effect relationship. Even a high degree of relationship does not necessarily imply that a cause and effect relationship exists. Conversely, however the cause and effect relationship (or functional relationship) would always result in the expression of correlation.

We would now discuss regression analysis.

19.5 Regression

Regression® analysis is the method of estimating the values of a variable from the knowledge of one or more variables. The variable that the researcher tries to estimate is called dependent variable (denoted as Y), whereas the variable used for prediction is independent variable (denoted as X). In a regression equation, there may be one or more independent variables, but there is only one dependent variable. Depending on whether there are one or more independent variables, the regression equation is called simple or multiple. The term 'linear' is added if the relationship between the dependent and the independent variable is linear. Thus a simple linear regression equation is represented as

$$Y = a + b X$$

Where, Y is dependent variable

X is independent variable

'a' is regression constant

'b' is regression coefficient. It measures the change in Y corresponding to a change in X.

Similarly a multilinear regression equation is represented as

$Y = a + b_1 X_1 + b_2 X_2 + \ldots b_n X_n$

Where, Y is dependent variable

X_1, X_2, ...X_n, are independent variables

'a' is regression constant

'b_1, b_2, ...b_n' are respective regression coefficients.

Like the calculation of coefficient of the correlation, there are various methods of calculating regression equation:

1. From actual mean values of X and Y.
2. From assumed mean values of X and Y.

Calculation of regression equation using actual mean

Regression equation (of Y on X) can be calculated using the following formula:

Y- M_Y = b_{yx} * (X- M_X) or

Y- M_Y = r (ó$_Y$ / ó$_X$) * (X- M_X)

As, **b_{yx}** = **r (ó$_Y$ / ó$_X$)** = (?xy / ?x^2), the regression equation may be calculated using the following formula.

Y- M_Y = (Σ xy / Σ x^2) * (X- M_X)

Where, Y and X are dependent and independent variables respectively;

M_Y and M_X are means of Y and X variable respectively; and

y = Y-M_Y and x = X-M_X

The following example illustrates the calculation of the regression equation.

Example: Calculate the regression equation using the following data, taking age at marriage of husbands as independent variable and that of wives as dependent variable (see Figure 19.11)

Age at Marriage	***Case 1***	***Case 2***	***Case 3***	***Case 4***	***Case 5***	***Case 6***	***Case 7***	***Case 8***	***Case 9***	***Case 10***
Husbands	28	25	24	29	31	22	21	25	26	28
Wives	22	23	21	25	26	20	19	21	21	24

Calculation of regression equation using actual mean (see Figure 19.12)

Age of Wives Y	y = Y-M_Y	y^2	**Age of Husbands** X	x = X-M_X	x^2	xy
22	-0.2	00.04	28	2.1	4.41	-0.42
23	0.8	00.64	25	-0.9	0.81	-0.72
21	-1.2	01.44	24	-1.9	3.61	02.28
25	2.8	07.84	29	3.1	9.61	08.68
26	3.8	14.44	31	5.1	26.01	19.38
20	-2.2	04.84	22	-3.9	15.21	8.58
19	-3.2	10.24	21	-4.9	24.01	15.68
21	-1.2	01.44	25	-0.9	0.81	01.08
21	-1.2	01.44	26	0.1	0.01	-0.12
24	1.8	03.24	28	2.1	4.41	03.78
222	0	45.60	259	0	88.9	58.20

Figure 19.12 Calculation of Regression Equation using Actual Mean

M_Y = 222 / 10 = 22.2 M_X = 259 / 10 = 25.9

Y- M_Y = (Σxy / Σx^2) * (X- M_X)

Y- 22.2 = (58.2 / 88.9) * (X - 25.2)

Y- 22.2 = 0.655 * (X - 25.2)

Y- 22.2 = 0.655X - 16.96

Y = 5.24 + 0.655X

Calculation of regression equation using assumed mean (see Figure 19.13)

Regression equation (of Y on X) can be calculated using the following formula, taking the assumed mean:

Y- M_Y = b_{yx} * (X- M_X)

Where, b_{yx} = [Σ d_X* d_Y - (Σ d_X * Σ d_Y) / N] / [Σ d_X^2 - (Σ d_x)^2 /N]

Y and X are dependent and independent variable respectively;

M_Y and M_X are mean of Y and X variables respectively

d_Y = Y-AM_Y **and** d_X = X-AM_X

AM_Y and AM_X are the assumed mean of Y and X variable respectively; and

Calculation of regression equation using assumed mean

Age of Wives Y	d_Y = Y-AM_Y	dy^2	Age of Husbands X	d_X = X-AM_X x	dx^2	dx * dy
22	0	0	28	3	9	0
23	1	1	25	0	0	0
21	-1	1	24	-1	1	1
25	3	9	29	4	16	12
26	4	16	31	6	36	24
20	-2	4	22	-3	9	6
19	-3	9	21	-4	16	12
21	-1	1	25	0	0	0
21	-1	1	26	1	1	-1
24	2	4	28	3	9	6
222	2	46	259	9	97	60

Figure 19.13 Calculation of Regression Equation using Assumed Mean

M_Y = 222 / 10 = 22.2 M_X = 259 / 10 = 25.9

b_{yx} = [Σ d_X* d_Y - (Σd_X * Σ d_Y) / N] / [Σ d_X^2 - (Σ d_x)^2 /N]

b_{yx} = [60 - (9*2) /10] / [97 - 9*9/10]

b_{yx} = 58.2 / 88.9 = 0.655

Y- M_Y = b_{yx} * (X- M_X)

Y- 22.2 = 0.655 * (X - 25.2)

Y- 22.2 = 0.655X - 16.96

$Y = 5.24 + 0.655X$

Standard error of estimate: Perfect prediction, using a regression equation is not possible (except when correlation value is -1 or + 1). Thus the researcher is interested in finding the accuracy of estimation of a regression equation. Standard error of estimate measures the error involved in using a regression equation as a basis of estimation. It can be calculated using the following equation:

$SEE_{y..x} = v \{\Sigma(Y - Y_c)^2 / N - 2\}$

Where, $SEE_{y..x}$ is Standard error of estimate

Y is dependent variable

Y_c is predicted value of Y

N is the number of observations

It can also be calculated from the following formula

$SEE_{y..x} = v \{(\Sigma Y^2 - a?Y - b\Sigma XY) / N- 2\}$

Where, $SEE_{y..x}$ is Standard error of estimate

Y is dependent variable

X is independent variable

'a' is regression constant

'b' is regression coefficient.

N is the number of observations

Coefficient of determination: Coefficient of determination (r^2) is the square of correlation coefficient (r) and is often used in interpreting the value of the coefficient of correlation. If the value of r were 0.8 then the coefficient of determination or r^2 would be 0.64. This would mean that 64% of variance of one variable (dependent) is explained in terms of the other variable (independent).

Reflection and Action 19.3

I tried to understand how to make the calculation of regression equation using assumed mean. I could not succeed. May be you can explain it to me with an example. Write out on a separate sheet of paper your explanation with one or two examples. May be I will then follow it. You will need to send it to the co-ordinator of MSO 002.

19.6 Conclusion

Unit 19 is the last unit of Block 5 on Quantitative Methods. All five units of this block have emphasised that quantitative methods should be used in social research when they are necessary and relevant and can provide superior results. Sometimes you can use them in combination with the qualitative methods. You need not avoid the quantitative methods because of lack of information or apprehension that it is difficult to understand them. The five units of block 5 have provided you appropriate examples wherever possible and necessary to help you understand the tools that are very useful in your research project assignment.

Further Reading

Burns, Robert B. 2000. *Introduction to Research Methods*. Sage Publications: London

Cohen, Louis and Michael Holliday 1982. *Statistics for Social Research*. Harper and Row: London