Chapter 5

CORRELATION AND REGRESSION

5.1: Introduction

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

Suppose we have a set of 30 students in a class and we want to measure the heights and weights of all the students. We observe that each individual (unit) of the set assumes two values – one relating to the height and the other to the weight. Such a distribution in which each individual or unit of the set is made up of two values is called a bivariate distribution. The following examples will illustrate clearly the meaning of bivariate distribution.

(i) In a class of 60 students the series of marks obtained in two subjects by all of them.

(ii) The series of sales revenue and advertising expenditure of two companies in a particular year.

(iii) The series of ages of husbands and wives in a sample of selected married couples.

Thus in a bivariate distribution, we are given a set of pairs of observations, wherein each pair represents the values of two variables.

In a bivariate distribution, we are interested in finding a relationship (if it exists) between the two variables under study.

The concept of 'correlation' is a statistical tool which studies the relationship between two variables and Correlation Analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

"Two variables are said to be in correlation if the change in one of the variables results in a change in the other variable".

5.2: Types of Correlation

There are two important types of correlation. They are (1) Positive and Negative correlation and (2) Linear and Non – Linear correlation.

5.2.1: Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

(i) Heights and weights;
(ii) Household income and expenditure;
(iii) Price and supply of commodities;
(iv) Amount of rainfall and yield of crops.

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.

Some examples of series of negative correlation are:

(i) Volume and pressure of perfect gas;
(ii) Current and resistance [keeping the voltage constant] $(R=\frac{V}{I})$;
(iii) Price and demand of goods.

Graphs of Positive and Negative correlation:

Suppose we are given sets of data relating to heights and weights of students in a class. They can be plotted on the coordinate plane using x – axis to represent heights and y – axis to represent weights. The different graphs shown below illustrate the different types of correlations.

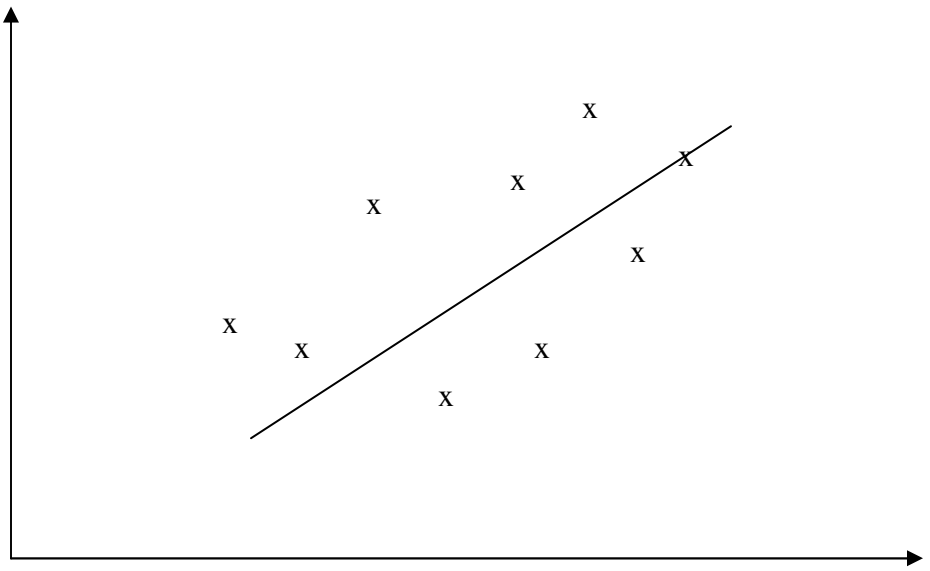

Figure for positive correlation

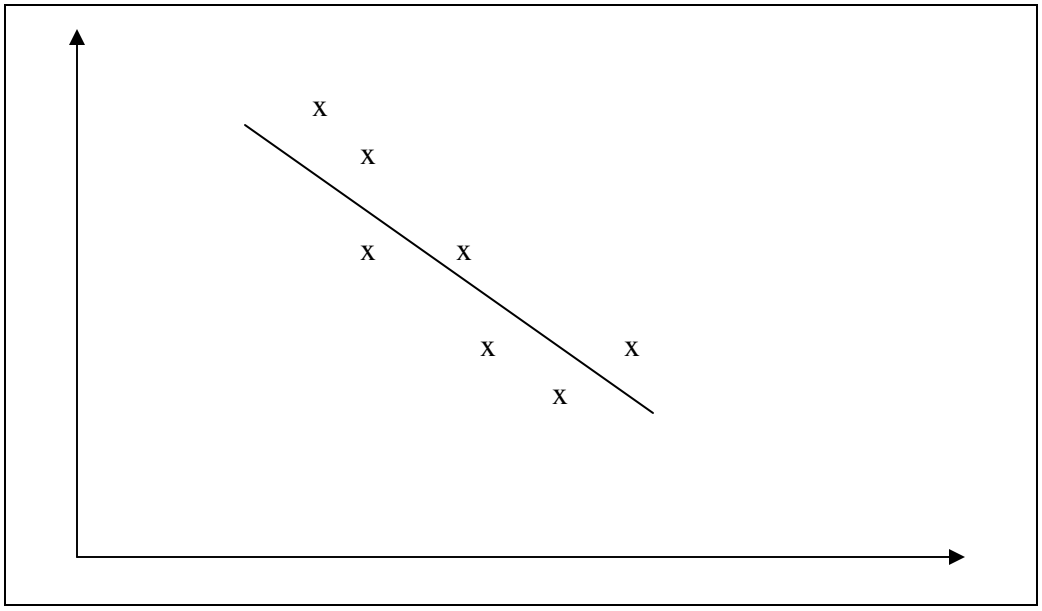

Figure for negative correlation

Note:

(i) If the points are very close to each other, a fairly good amount of correlation can be expected between the two variables. On the other hand if they are widely scattered a poor correlation can be expected between them.

(ii) If the points are scattered and they reveal no upward or downward trend as in the case of (d) then we say the variables are uncorrelated.

(iii) If there is an upward trend rising from the lower left hand corner and going upward to the upper right hand corner, the correlation obtained from the graph is said to be positive. Also, if there is a downward trend from the upper left hand corner the correlation obtained is said to be negative.

(iv) The graphs shown above are generally termed as **scatter diagrams**.

Example:1: The following are the heights and weights of 15 students of a class. Draw a graph to indicate whether the correlation is negative or positive.

Heights (cms)	Weights (kgs)
170	65
172	66
181	69
157	55
150	51
168	63
166	61
175	75
177	72
165	64
163	61
152	52
161	60
173	70
175	72

Since the points are dense (close to each other) we can expect a high degree of correlation between the series of heights and weights. Further, since the points reveal an upward trend, the correlation is positive. Arrange the data in increasing order of height and check that , as height increases, the weight also increases, except for some (stray) cases..

EXERCISES

(1) A Company has just brought out an annual report in which the capital investment and profits were given for the past few years. Find the type of correlation (if it exists).

Capital Investment (crores)	10 16 18 24 36 48 57
Profits (lakhs)	12 14 13 18 26 38 62

(2) Try to construct more examples on the positive and negative correlations.
(3) Construct the scattered diagram of the data given below and indicate the type of correlation.

(Average Value in Lakhs of Rs.)

Years	1965	1970	1975	1980	1985	1990
Raw cotton import	42	60	112	98	118	132
Cotton manufacture exports	68	79	88	86	106	114

5.3: Linear and Non – Linear Correlation

The correlation between two variables is said to be **linear** if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.

For example consider the following data.

X	2	4	6	8	10
Y	7	13	19	25	31

Thus, for a unit change in the value of x, there is a constant change in the corresponding values of y and the above data can be expressed by the relation

$$y = 3x + 1$$

In general two variables x and y are said to be **linearly related**, if there exists a relationship of the form

$$y = a + bx$$

where 'a' and 'b' are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

The relationship between two variables is said to be **non – linear** if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form

$$y = a + bx + cx^2$$

or more general polynomial.

5.4: The Coefficient of Correlation

One of the most widely used statistics is the **coefficient of correlation** 'r' which measures the degree of association between the two values of related variables given in the data set. It takes values from + 1 to – 1. If two sets or data have r = +1, they are said to be perfectly **correlated positively** if r = -1 they are said to be perfectly **correlated negatively**; and if r = 0 they are **uncorrelated.**

The coefficient of correlation 'r' is given by the formula

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

The following example illustrates this idea.

Example:2: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

Solution:

x	y	x^2	y^2	xy
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372
796	1546	63,776	243036	1242069

Then

$$r=\frac{10(124206)-(796)(1546)}{\sqrt{\left[(10)63776-(796)^2\right]\left[(10)(243036)-(1546)^2\right]}}$$

$$= \frac{11444}{\sqrt{(1144)(40244)}}$$
$$= 0.5966$$

5.4: Rank Correlation

Data which are arranged in numerical order, usually from largest to smallest and numbered 1,2,3 ---- are said to be in **ranks** or **ranked data.**. These ranks prove useful at certain times when two or more values of one variable are the same. The coefficient of correlation for such type of data is given by **Spearman rank difference correlation coefficient** and is denoted by R.

In order to calculate R, we arrange data in ranks computing the difference in rank 'd' for each pair. The following example will explain the usefulness of R. R is given by the formula

$$R = 1 - 6\frac{(\sum d^2)}{n(n^2 - 1)}$$

Example:3: The data given below are obtained from student records. Calculate the rank correlation coefficient 'R' for the data.

Subject	Grade Point Average (x)	Graduate Record exam score (y)
1.	8.3	2300
2.	8.6	2250
3.	9.2	2380
4.	9.8	2400
5.	8.0	2000
6.	7.8	2100
7.	9.4	2360
8.	9.0	2350
9.	7.2	2000
10.	8.6	2260

Note that in the G. P. A. column we have two students having a grade point average of 8.6 also in G. R. E. score there is a tie for 2000.

Now we first arrange the data in descending order and then rank 1,2,3,---- 10 accordingly. In case of a tie, the rank of each tied value is the mean of all positions they occupy. In x, for instance, 8.6 occupy ranks 5 and 6. So each has a rank $\frac{5+6}{2}=5.5$;

Similarly in 'y' 2000 occupies ranks 9 and 10, so each has rank $\frac{9+10}{2}=9.5$.

Now we come back to our formula $R=1-\frac{6\sum d^2}{n(n^2-1)}$

We compute 'd' , square it and substitute its value in the formula.

Subject	x	y	Rank of x	Rank of y	d	d^2
1.	8.3	2300	7	5	2	4
2.	8.6	2250	5.5	7	-1.5	2.25
3.	9.2	2380	3	2	1	1
4.	9.8	2400	1	1	0	0
5.	8.0	2000	8	9.5	-1.5	2.25
6.	7.8	2100	9	8	1	1
7.	9.4	2360	2	3	-1	1
8.	9.0	2350	4	4	0	0
9.	7.2	2000	10	9.5	0.5	0.25
10.	8.6	2260	5.5	6	-0.5	0.25

So here, n = 10, sum of d^2 = 12. So

$$R=1-\frac{6(12)}{10\,(100-1)}$$
$$=1-0.0727 \quad = 0.9273$$

Note: If we are provided with only ranks without giving the values of x and y we can still find Spearman rank difference correlation R by taking the difference of the ranks and proceeding in the above shown manner.

EXERCISES

1. A horse owner is investigating the relationship between weight carried and the finish position of several horses in his stable. Calculate r and R for the data given

Weight Carried	Position Finished
110	2
113	6
120	3
115	4
110	6
115	5
117	4
123	2
106	1
108	4
110	1
110	3

2. The top and bottom number which may appear on a die are as follows

Top	1	2	3	4	5	6
bottom	5	6	4	3	1	2

Calculate r and R for these values. Are the results surprising?

3. The ranks of two sets of variables (Heights and Weights) are given below. Calculate the Spearman rank difference correlation coefficient R.

	1	2	3	4	5	6	7	8	9	10
Heights	2	6	8	4	7	4	9.5	4	1	9.5
Weights	9	1	9	4	5	9	2	7	6	3

5.5: Regression

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy.

Regression analysis was explained by M. M. Blair as follows: "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

5.5.1: Regression Equation

Suppose we have a sample of size 'n' and it has two sets of measures, denoted by x and y. We can predict the values of 'y' given the values of 'x' by using the equation, called the REGRESSION EQUATION.

$$y^* = a + bx$$

where the coefficients a and b are given by

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

$$a = \frac{\sum y - b\sum x}{n}$$

The symbol y* refers to the predicted value of y from a given value of x from the regression equation.

Example: 4 : Scores made by students in a statistics class in the mid - term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

STUDENT	MID – TERM	FINAL
1.	98	90
2.	66	74
3.	100	98
4.	96	88
5.	88	80
6.	45	62
7.	76	78
8.	60	74
9.	74	86
10.	82	80

Solution:

We want to predict the final exam scores from the mid term scores. So let us designate ‘y’ for the final exam scores and ‘x’ for the mid – term exam scores. We open the following table for the calculations.

Stud	x	y	X^2	xy
1	98	90	9604	8820
2	66	74	4356	4884
3	100	98	10,000	9800
4	96	88	9216	8448
5	88	80	7744	7040
6	45	62	2025	2790
7	76	78	5776	5928
8	60	74	3600	4440
9	74	86	5476	6364
10	82	80	6724	6560
Total	785	810	64,521	65,071

Numerator of b = 10 * 65,071 – 785 * 810 = 6,50,710 – 6,35,850 = 14,860
Denominator of b = 10 * 64, 521 – $(785)^2$ = 6,45,210 – 6,16,225 = 28,985

Therefore, b = 14,860 / 28,985 = 0.5127

Numerator of a = 810 – 785 * 0.5127 = 810 – 402.4695 = 407.5305
Denominator of a = 10

Therefore a = 40.7531

Thus , the regression equation is given by

$$y^* = 40.7531 + (0.5127)\, x$$

We can use this to find the projected or estimated final scores of the students.
For example, for the midterm score of 50 the projected final score is
$y^* = 40.7531 + (0.5127)\, 50 = 40.7531 + 25.635 = 66.3881$
which is a quite a good estimation.
To give another example, consider the midterm score of 70. Then the projected final score is
$y^* = 40.7531 + (0.5127)\, 70 = 40.7531 + 35.889 = 76.6421,$
which is again a very good estimation.

This brings us to the end of this chapter. We close with some problems for you.

EXERCISES

1. The data given below are obtained from student records. Calculate the regression equation and compute the estimated GRE scores for GPA = 7.5, 8.5..

Subject	Grade Point Average (x)	Graduate Record exam score (y)
11.	8.3	2300
12.	8.6	2250
13.	9.2	2380
14.	9.8	2400
15.	8.0	2000
16.	7.8	2100
17.	9.4	2360
18.	9.0	2350
19.	7.2	2000
20.	8.6	2260

2. A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

3. A horse was subject to the test of how many minutes it takes to reach a point from the starting point. The horse was made to carry luggage of various weights on 10 trials.. The data collected are presented below in the table.

Trial No.	Weight (in Kgs)	Time taken (in mins)
1	11	13
2	23	22
3	16	16
4	32	47
5	12	13
6	28	39
7	29	43
8	19	21
9	25	32
10	20	22

Find the regression equation between the load and the time taken to reach the goal. Estimate the time taken for the loads of 35 Kgs , 23 Kgs, and 9 Kgs. Are the answers in agrrement with your intuitive feelings? Justify.

$$$