Correlation and Regression-I

Mr. Anup Singh
Department of Mathematics
Mahatma Gandhi Central university
Motihari-845401, Bihar, India
E-mail: anup.singh254@gmail.com

Correlation

Definition: The correlation is the measure of the extent and the direction of the relationship between two variables in a bivariate distribution.

- Example: (i) Height and weight of children. (ii) An increase in the price of the commodity by a decrease in the quantity demanded.

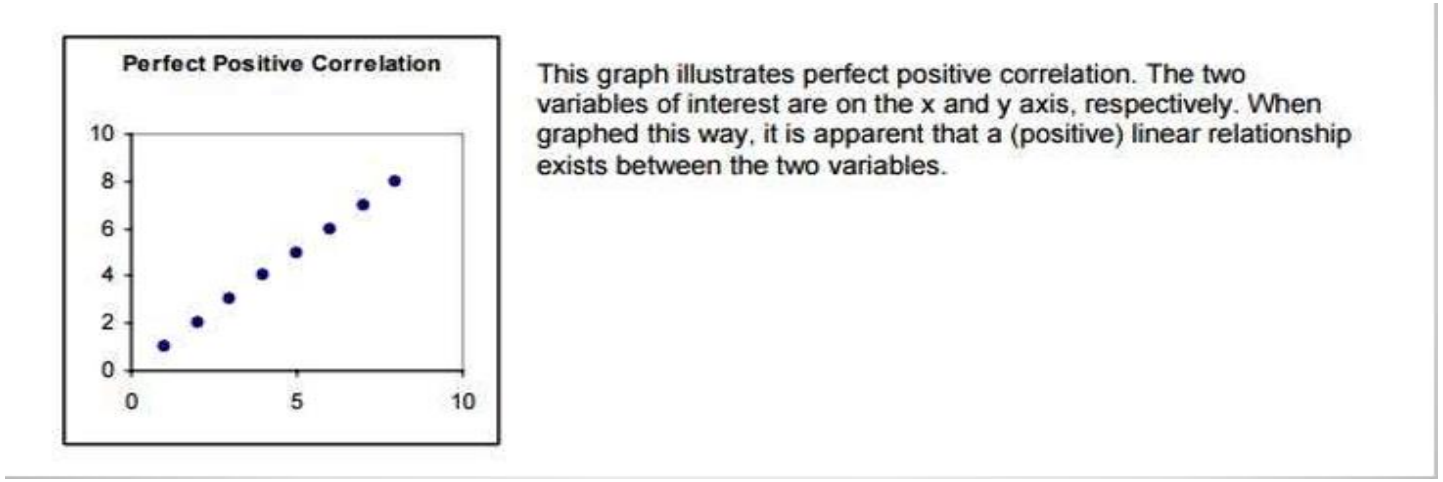
Types of Correlation: There are two important types of correlation

(i) Positive and Negative Correlation

(ii) Linear and Non-linear Correlation

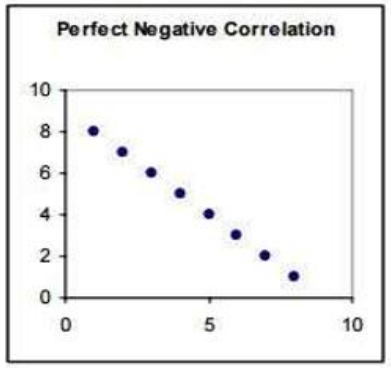
Positive correlation:

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



- Example 1: a) heights and weights (b) amount of rainfall and yields of crops (c) price and supply of a commodity (d) income and expenditure on luxury goods (e) blood pressure and age

Negative correlation: 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.



This graph illustrates perfect negative correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is apparent that a (negative) linear relationship exists between the two variables, i.e. the variables "move together".

- Example a) price and demand of commodity (b) sales of woolen garments and the days temperature.

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.

Example:

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 ; $y = a + bx$

Non-Linear Correlation : 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

$y = a + b\,x + cx^2$ or more general polynomial.

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.

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

- 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.

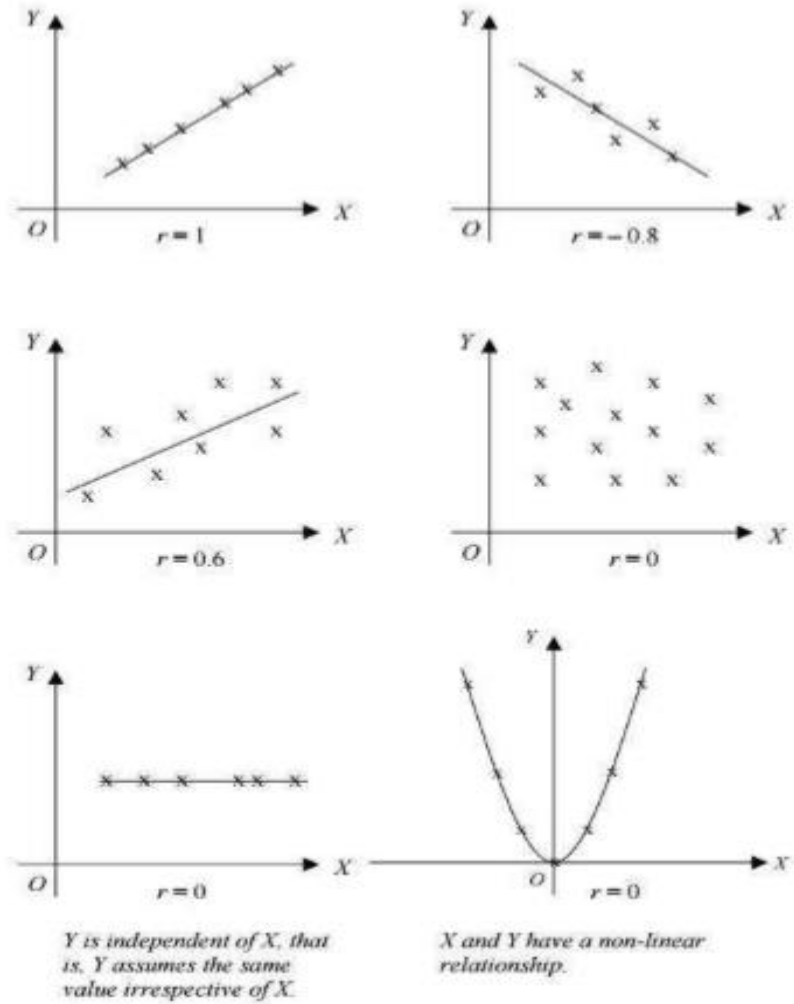


Fig: Using scattering diagrams to determine r approximately

Correlation Coefficient

Shows Strength & Direction of Correlation

Example: 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

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

r=0.5966

Reference Books:

1. Erwin Kreyszig, Advance Engineering Mathematics, 9th Edition, John Wiley & Sons, 2006.
2. Sheldon Ross, A first course in Probability, 8th Edition, Pearson Education India.
3. W. Feller An Introduction to Probability Theory and its Applications, Vol. 1, 3rd Edition, Wiley, 1968.
4. S. C. Gupta and V. . Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand & Sons.
5. P. G. Hoel, S. C. Port and C. J. Stone, Introduction to Probability Theory, Universal Book Stall, 2003.
6. A. M. Mood, F. A. Graybill and D. C. Bose, Introduction to Theory of Statistics. 3rd Edition, Tata McGraw-Hill Publication.

THANK YOU