Correlation and Regression

Simple Regression and Correlation

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".

Simple Regression and Correlation

We will discuss a powerful statistical technique for examining whether or not two variables are related. Specifically, we are going to talk about the ideas of simple regression and correlation.
One reason why regression is powerful is that we can use it to demonstrate causality; that is, we can show that an independent variable causes a change in a dependent variable.

Scatter diagrams

The simplest thing we can do with two variables that we believe are related is to draw a scatter diagram. A scatter diagram is a simple graph that plots values of our dependent variable Y and independent variable X.

Normally we plot our dependent variable on the vertical axis and the independent variable on the horizontal axis.

Just from our scatter diagram, we can sometimes get a fairly good idea of the relationship between our variables.

Types of Correlation

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

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.

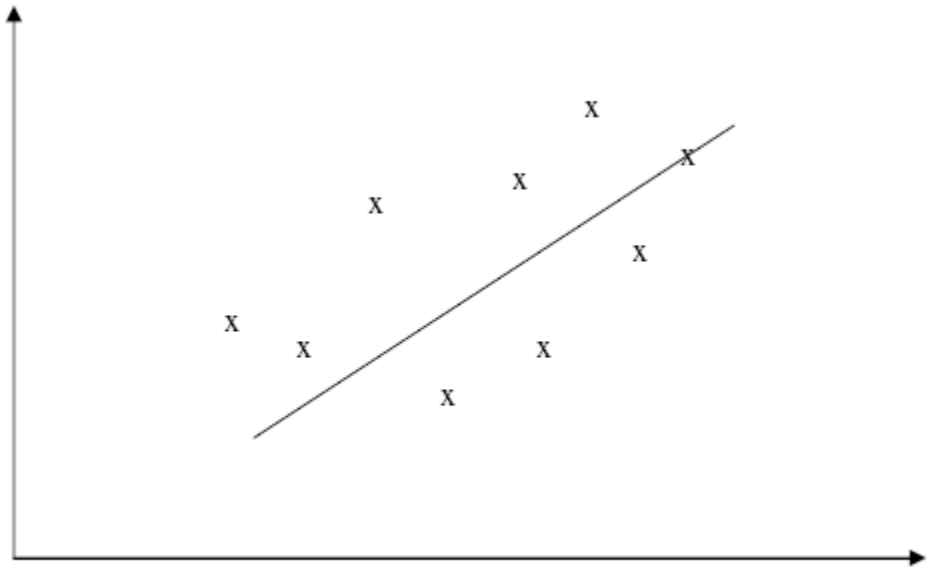

Figure for positive correlation

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.

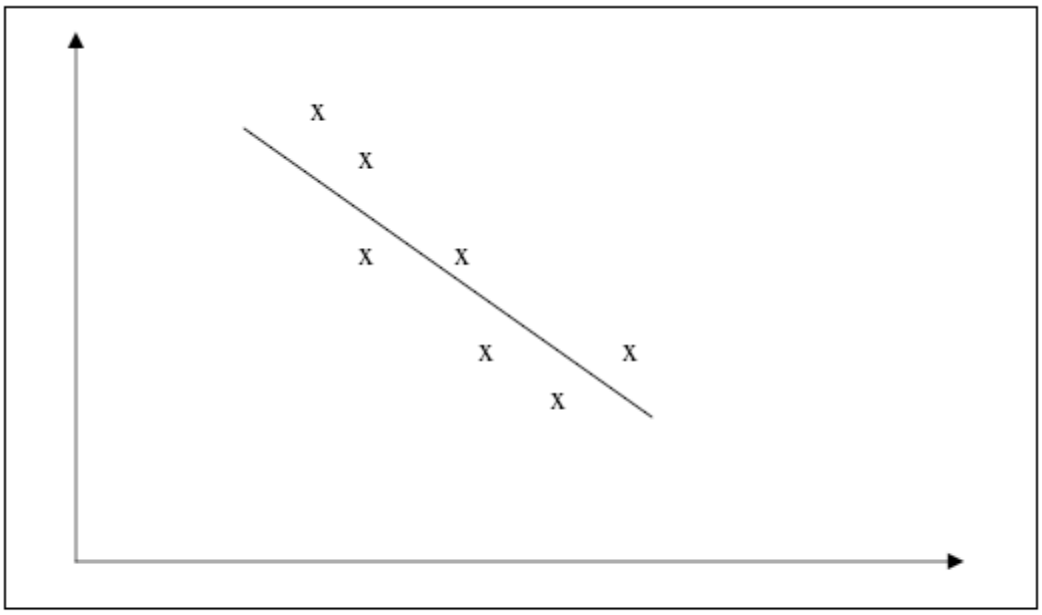

Figure for negative correlation

Some examples of series of negative correlation are:
(i) Volume and pressure of perfect gas;
(ii) Current and resistance [keeping the voltage constant] (R=V/I)
(iii) Price and demand of goods.

While our scatter diagram gives us a fairly good idea of the relationship between the variables, and even some idea of how the regression line should look, we need to do the math to figure out exactly where it goes.
That is, regression expresses the mathematical relationship between two variables that have a non-zero correlation coefficient.

Karl Pearson's Coefficient of correlation r

The Karl Pearson's coefficient of correlation between X and Y is given as

$r(x,y) = r = \dfrac{Cov(X,Y)}{\sigma_x \sigma_y}$ **where**

$$\begin{aligned} Cov(X,Y) &= E(X - E(X))E(Y - E(Y)) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

is the Covariance of X and Y and

$\sigma_x,\ \sigma_y$ are the standard deviations of X and Y respectively.

We have

$$\begin{aligned} r &= \frac{Cov(X,Y)}{\sigma_x \sigma_y} \\ &= \frac{E(XY) - E(X)E(Y)}{\sqrt{\left(E(X^2) - \{E(X)\}^2\right)\left(E(Y^2) - \{E(Y)\}^2\right)}} \end{aligned}$$

Given the set of values (X_i, Y_i) we can obtain the correlation coefficient as:

$$r = \frac{\frac{1}{n}\sum X_i Y_i - \frac{1}{n}\sum X_i \frac{1}{n}\sum Y_i}{\sqrt{\left(\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2\right)\left(\frac{1}{n}\sum Y_i^2 - \left(\frac{1}{n}\sum Y_i\right)^2\right)}}$$

Remarks.

1. $-1 \le r \le 1$

If the value of r is negative, it denotes there is an inverse relation between X and Y – that is, if X increases, then Y decreases and if X decreases, then Y increases. Similarly, if the value of r is positive, it denotes there is a direct relation between X and Y – that is, if X increases, then Y increases and if X decreases, then Y decreases. If the value of r is zero, it means there is no correlation between the two variables, i.e. X and Y are not related at all. Further, if r is positive, the closer it is to one, the stronger the correlation. Similarly, a negative r closer to 1 will mean a stronger correlation than if it is closer to zero.

2. The correlation coefficient is unaffected by change of origin and scale.

Examples

1. Obtain the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y):

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

Solution: We have the Karl Pearson's Coefficient of correlation r to be given by

$$r = \frac{Cov(X,Y)}{\sqrt{\sigma_x \sigma_y}}$$

$$= \frac{\frac{1}{n}\sum X_i Y_i - \frac{1}{n}\sum X_i \frac{1}{n}\sum Y_i}{\sqrt{\left(\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2\right)\left(\frac{1}{n}\sum Y_i^2 - \left(\frac{1}{n}\sum Y_i\right)^2\right)}}$$

$$n = 8;\ \ \sum X_i = 544;\ \sum Y_i = 552;\ \sum X_i Y_i = 37560$$
$$\sum X_i^{\,2} = 37028;\ \sum Y_i^2 = 38132$$

$$\therefore r = \frac{\frac{1}{n}\sum X_i Y_i - \frac{1}{n}\sum X_i \frac{1}{n}\sum Y_i}{\sqrt{\left(\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2\right)\left(\frac{1}{n}\sum Y_i^2 - \left(\frac{1}{n}\sum Y_i\right)^2\right)}}$$

$$= \frac{4695 - 4692}{\sqrt{(4.5)(5.5)}}$$

$$= \frac{3}{\sqrt{24.75}}$$

$$= 0.603$$

Practice Problem

The following data gives the growth of employment in lakhs in organized sector in India between 1988 and 1995. Find the Karl Pearson's coefficient of correlation between the employment in Public and Private sectors.

Year	1988	1989	1990	1991	1992	1993	1994	1995
Public	98	101	104	107	113	120	125	128
Private	65	65	67	68	69	69	68	68

Solution: Let the random variables X and Y respectively denote the employment in Public and Private sectors.

We have the Karl Pearson's Coefficient of correlation r to be given as

$$r = \frac{Cov(X,Y)}{\sqrt{\sigma_x \sigma_y}} = \frac{\frac{1}{n}\sum X_i Y_i - \frac{1}{n}\sum X_i \frac{1}{n}\sum Y_i}{\sqrt{\left(\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2\right)\left(\frac{1}{n}\sum Y_i^2 - \left(\frac{1}{n}\sum Y_i\right)^2\right)}} \quad ...(*)$$

$$n = 8; \quad \sum X_i = 896; \ \sum Y_i = 539; \ \sum X_i Y_i = 60,460$$
$$\sum X_i^2 = 1,01,248; \ \sum Y_i^2 = 36,333$$

Substituting these values on (*) we get $\therefore r = 0.7269$

Prof. Namcy Sinollin