Correlation and Regression

CORRELATION

Correlation : Relation between two variable or more Variable. The primory objective of correlation analysis is to measure the strength or degree of relationship between two or more variables. If the change in one variable affects a change in the ~~onde~~ other variable, the variables are said to be correlated.

Types of correlation : ① Positive or negative
② Simple or multiple
③ Linear or non-linear

Positive or negative: If the two variables deviate in the same direction, that is if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive.

But if they constantly deviate in the opposite directions, that is if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be inverse or negative.

(iii) If the Variables are independent, there cannot be any correlation and the variables are said to be Zero correlation.

Example: The correlation between (1) the heights and weights of a group of persons.

Simple correlation and Multiple correlation:

⇒ Correlation only between two variables is called simple correlation. Example: Correlation between income and expenditure.

⇒ correlation between three or more than three Variable is called multiple correlation.

Ex: $Qd = f(P, Pc, Ps, t, y)$

Linear and Non Linear Correlation:

⇒ Correlation is said to be linear when the amount of change in one Variable tends to bear a constant ratio to the amount of change in the other.

The graph of the variables having a linear relationship will from a straight line.

Ex: X = 1, 2, 3, 4, 5, 6, 7, 8

Y = 5, 7, 9, 11, 13, 15, 17, 19,

$Y = 3+2x$

⇒ The correlation would be non linear if the amount of change in one Variable does not bear a constant ratio to the amount of change in the other Variable.

Methods of Studying simple Correlation:

1. Scatter Diagram Method
2. karl pearson's coefficient of correlation
3. Spearman's Rank correlation

Scatter diagram method: The diagrammetic way of representing bivariate data is called scatter diagram.

Interpret of r

$r = +1$, indicates a perfect positive relationship between x and y.

$r = -1$, indicates a perfect negative relationship between x and y

$r = 0$, means there is no linear relationship between x and y. Here two variables are linearly independent.

$0 < r < 1$, indicates a positive relationship between x and y

$-1<r<0$; indicates a negative relationship between x and y.

Correlation Coefficient: The numerical value by which we measure the strength of linear relationship between two or more variables is called correlation coefficient.

Let, $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ be the pairs of n observations. Then the correlation coefficient between x and y is denoted by r_{xy} and defined as,

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

The formula is given by karl pearson.

Assumptions of pearson's correlation coefficient:

① There is linear relationship between two variables. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.

② Cause and effect relation exists between different forces operating on the item of the two variable series.

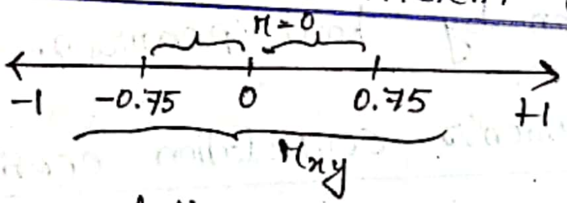
Properties of correlation coefficient:

1. Correlation coefficient is independent of change of origin and scale of measurement.
2. Correlation coefficient lies between -1 to $+1$

$$-1 \le r_{xy} \le 1$$

3. Correlation coefficient is symmetric. i.e $r_{xy} = r_{yx}$.
4. Correlation coefficient is the geometric mean of regression coefficients i.e. $r = \sqrt{b_{yx} \times b_{xy}}$
5. For two independent Variable correlation coefficient is zero.
6. It is always unit free.

Comments on correlation coefficient (r_{xy}):



1. $r=0$, no correlation
2. $0<r<0.75$ – Simple positive correlation coefficient
3. $0.75 \le r < 1$ – Strongly positive correlation coefficient
4. $r=1$ – Perfect positive correlation coefficient
5. $-0.75<r<0$ – Negative Correlation Coefficient
6. $-1<r \le -0.75$ – Strongly negative correlation coefficient
7. $r=-1$ – Perfect negative correlation coefficient.

Problem: 1 (calculate the correlation coefficient between temperature of water and reduction in pulse rate)

Temperature of water	68	65	70	62	60	55	58	65	69	63
Reduction in pulse rate	2	5	1	10	9	13	10	3	4	6

Solution:

x	y	x^2	y^2	xy
68	2	4624	4	136
65	5	4225	25	325
70	1	4900	1	70
62	10	3844	100	620
60	9	3600	81	540
55	13	3025	169	715
58	10	3364	100	580
65	3	4225	~~81~~9	195
69	4	4761	16	276
63	6	3969	36	378
$\Sigma x = 635$	$\Sigma y = 63$	$\Sigma x^2 = 40537$	$\Sigma y^2 = 541$	$\Sigma xy = 3835$

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

$$= \frac{10 \times 3835 - 635 \times 63}{\sqrt{\{10\times 40537 - (635)^2\}\{10\times 541 - (63)^2\}}}$$

$n = 10$

$$= -0.94$$

The result -0.94 indicates that the correlation coefficient between temperature of water and reduction in pulse rate is highly negatively correlated.

REGRESSION ANALYSIS

Regression: Regression is the functional relationship between two variables and of the two variables one may represent cause and the other may represent effect. The variable representing cause is known as independent variable and is denoted by x.

x = independent variable also known as Predictor Variable or regressor.

Y = dependent variable also known as predicted Variable.

a = constant term / intercept term

b = b_{xy}, b_{yx}

$$y = a + bx.$$

Regression coefficient: The mathematical measures of regression are called the coefficient of regression.

Let, $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ be the pairs of n observations. Then the regression coefficient of y on x is denoted by b_{yx} and defined ~~by~~ by,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\{n \Sigma x^2 - (\Sigma x)^2\}} \quad \Big| \quad b_{yx} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Again, the regression coefficient of x on y is denoted by b_{xy} and defined by,

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\{n \Sigma y^2 - (\Sigma y)^2\}} \quad \Big|$$ ~~$b_{xy} =$~~ -

Properties of regression coefficient:

① Regression Coefficient is independent of change of origin but not of scale.

② Regression coefficient lines between $-\infty < b_{yx} < \infty$

③ Regression coefficient is not symmetric. $b_{yx} \neq b_{xy}$.

④ The geometric mean of regression coefficient is equal to correlation coefficient, $r_{xy} = \sqrt{b_{yx} * b_{xy}}$

⑤ The arithmetic mean of two regression coefficient is greater than correlation coefficient.

$$\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r$$

⑥ $b_{xy} \geq 1$ and $b_{yx} < 1$

⑦ It is not unit free.

Regression equation: The regression equation of y on x is expressed as follows:

$$y = a+bx$$
$$\therefore a = y-bx$$
$$= \bar{y} - b\bar{x}$$

Here, y = dependent Variable
x = independent Variable
a = intercept term
b = slope of the line

and $$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\{n \Sigma x^2 - (\Sigma x)^2\}}$$

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

Similarly, the regression equation of x on y is expressed as follows:

$$x = a+by$$
$$\therefore a = x-by$$
$$= \bar{x} - b\bar{y}$$

Here,
x = dependent Vooriable
y = independent variable
a = intercept term
b = slope of the line

and $$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\{n\Sigma y^2 - (\Sigma y)^2\}}$$

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

Difference between correlation coefficient and Regression Coefficient :

Correlation Coefficient	Regression Coefficient
① The numerical value by which we measure the strength of linear relationship between two or more variables is called correlation coefficient.	① The mathematical measures of regression are called the coefficient of regression.
② Correlation coefficient is independent of change of origin and scale of measurement.	② Regression coefficient is independent of change of origin but not scale.
③ It lies between -1 to +1, $-1 \leq r_{xy} \leq 1$	③ It lies between $-\infty < b_{yx} < \infty$
④ It is symmetric. i.e, $r_{xy} = r_{yx}$.	④ It is not symmetric. i.e, $b_{xy} \neq b_{yx}$
⑤ It is always unit free.	⑤ It is not pure number.
⑥ When $r=0$ then the variables are correlated.	⑥ When $r = 0$ then two lines of regression are perpendicular to each other.

Consider the following data.

x:	1	2	3	4	5	6
y:	6	4	3	5	4	2

① Calculate karl pearson corrnelation coefficient and Comment.

② Draw Scatter diagram.

③ Compute the regression equation of y on x.

④ Estimate the value of y when $x = 4.5$.

⑤ Compute the regression equation of x on y.

⑥ Predict the value of x when $y = 3$.

Solution:

Computing Table

x	y	x^2	y^2	xy
1	6	1	36	6
2	4	4	16	8
3	3	9	9	9
4	5	16	25	20
5	4	25	16	20
6	2	36	4	12
$\Sigma x = 21$	$\Sigma y = 24$	$\Sigma x^2 = 91$	$\Sigma y^2 = 106$	$\Sigma xy = 75$

$n = 6$.

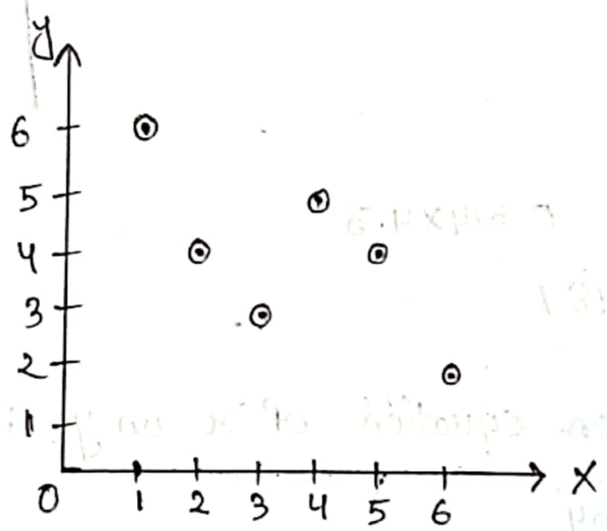
① Correlation coefficient, $r_{xy} = \dfrac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}}$

$$= \frac{6\times 75 - 21\times 24}{\sqrt{(6\times 91 - 21^2)(6\times 106 - 24^2)}}$$

$$= -0.68$$

∴ Comment: There exists negative correlation coefficient between x and y.

② Scatter diagram:



③ The regression equation of y on x is $y = a + bx$ —①

Here, $a = \bar{y} - b\bar{x}$

$$= \frac{\Sigma y}{n} - b\frac{\Sigma x}{n}$$

$$= \frac{24}{6} - b\frac{21}{6} \quad \text{—②}$$

$$\therefore b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{6\times 75 - 21\times 24}{6\times 91 - (21)^2}$$

$$= -0.514$$

③ From ②

$$a = \frac{24}{6} - (-0.514)\left(\frac{21}{6}\right)$$
$$= 5.79$$
$$= 5.8$$

From ①

$$y = a + bx$$
$$= 5.8 + (-0.514)x$$
$$= 5.8 - 0.514x$$

Therefore, the regression equation of y on x is

$$\hat{y} = 5.8 - 0.514x$$

$\hat{y}$ → Estimate
এই sign দেওয়া
হয় a ও b এর
মান জানা থাকলে,

④ When $x = 4.5$

$$\hat{y} = 5.8 - 0.514 \times 4.5$$
$$= 3.487$$

⑤ The regression equation of x on y is $x = a + by$ — ①

Here, $a = \bar{x} - b\bar{y}$

$$= \frac{\Sigma x}{n} - b\frac{\Sigma y}{n}$$
$$= \frac{21}{6} - b\frac{24}{6} \quad \text{—②}$$

$$\therefore b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$
$$= \frac{6 \times 75 - 21 \times 24}{6 \times 106 - (24)^2}$$
$$= -0.9$$

From ②,

$$a = \frac{21}{6} - (-0.9)\frac{24}{6}$$

$$= 7.1$$

From ①,

$$x = a + by$$

$$= 7.1 + (-0.9)y$$

$$= 7.1 - 0.9y$$

Therefore the regression equation of ~~y~~ ~~on x i~~ x on y is

$$\hat{x} = 7.1 - 0.9y.$$

⑤ When $y = 3$

$$\hat{x} = 7.1 - 0.9 \times 3$$

$$= 4.4.$$