CORRELATION THEORY AND REGRESSION ANALYSIS

The primary 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 other variable, the variables are said to be correlated.

For example, the production of paddy is dependent on the rainfall. Here production of paddy is considered to be a dependent variable.

Types of Correlation

- Positive or negative
- Simple or multiple
- 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 director 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. If the variables are independent, there cannot be any correlation and the variables are said to be zero correlation.

For example, the correlation between (1) the heights and weights of a group of persons, (2) the income and expenditure is positive and the correlation between (1) price and demand of a commodity, (2) the volume and pressure of a perfect gas is negative. And there is no correlation between income and height.

Simple correlation and Multiple Correlation

Correlation only between two variables is called simple correlation. For example, correlation between income and expenditure.

Under Multiple Correlation three or more than three variables are studied. Ex. Qd= f (P,PC, PS, t, y)

Linear correlation 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 form a straight line.

Example: 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

Scatter Diagram method;

Karl Pearson's Coefficient of correlation;

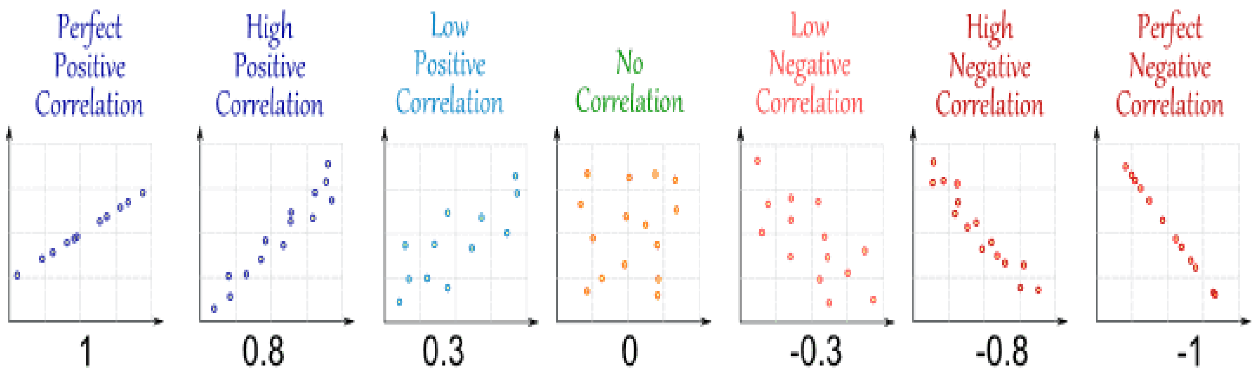
Spearman's Rank Correlation and

Scatter diagram method

The diagrammatic way of representing bivariate data is called scatter diagram.

Suppose, (x_1,y_1), (x_2,y_2)……….(x_n,y_n) are n pairs of observations. If the values of the variables x and y be plotted along the x-axis and y-axis respectively in the xy-plane, the diagram of dots so obtained is known as scatter diagram.

Scatter diagrams for different values of r are as follows:



Interpret of r

r= +1, indicates a perfect positive relationship between x and y. the scatter diagram will be as in fig. 1.1

r=-1, indicates a perfect negative relationship between x and y. the scatter diagram will be as in fig. 1.2

r=0, means there is no linear relationship between x and y. In this case the two variables are linearly independent. the scatter diagram will be as in fig. 1.5 and 1.6

0 < r <1, indicates a positive relationship between x and y. In this case the scatter diagram will be as in fig. 1.3

-1< r <0, indicates a negative relationship between x and y. In this case the scatter diagram will be as in fig. 1.4

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)………..(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,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \text{……………..(1)}$$

Equation (1) is also called Karl pearson’s coefficient of correlation formula given by 1890.

Algebraically (1) reduces to

$$r = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

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

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

Comment on Correlation Coefficient

1 = Perfect positive correlation

0.7 ≤ c < 1 = Strong positive correlation

0.4 ≤ c < 0.7 = Fairly positive correlation

0 < c < 0.4 = Weak positive correlation

0 = No correlation

0 > c > -0.4 = Weak negative correlation

-0.4 ≥ c > -0.7 = Fairly negative correlation

-0.7 ≥ c < -1 = Strong negative correlation

-1 = Perfect negative correlation

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. i.e, -1< r_{xy} < 1.
3. Correlation coefficient is symmetric. i.e, $r_{xy} = r_{yx}$
4. Correlation coefficient is the geometric mean of regression coefficients i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$
5. For two independent variable correlation coefficient is zero
6. It is always unit free.

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also

Limitation of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

Coefficient of Determination

The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination.

The Coefficient of Determination = r^2.

Suppose: r = 0.9, r^2 = 0.81 this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.

The maximum value of r^2 is 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it.

Coefficient of Determination = Explained variation / Total variation

An example of Coefficient of Determination

When r = 0.60, r^2 = 0.36 -----(1)

r = 0.30, r^2 = 0.09 -----(2)

This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained .

Theorm: Show that Correlation coefficient lies between -1 to +1 i.e, $-1 \le r_{xy} \le 1$.

Proof: Let, (x_1,y_1), (x_2,y_2)……….(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,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \text{ ……………..(1)}$$

Suppose, $(x_i - \bar{x}) = X$ and $(y_i - \bar{y}) = Y$ therefore

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

Let us consider the following expression which is always positive.

i.e, $\sum\left(\frac{X}{\sqrt{\sum X^2}} \pm \frac{Y}{\sqrt{\sum Y^2}}\right)^2 \ge 0$

or, $\sum\left(\frac{X^2}{\sum X^2} \pm 2\frac{X}{\sqrt{\sum X^2}}\frac{Y}{\sqrt{Y^2}} + \frac{Y^2}{\sum Y^2}\right) \ge 0$

or, $\left(\frac{\sum X^2}{\sum X^2} \pm 2\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} + \frac{\sum Y^2}{\sum Y^2}\right) \ge 0$

or, $1 \pm 2r + 1 \ge 0$

or, $2(1 \pm r) \ge 0$

or, $(1 \pm r) \ge 0$ ……(i)

From (i), $1+r \ge 0$ [considering +ve sign.]

or, $r \ge -1$

or, $-1 \le r$ …………(ii)

and $1-r \ge 0$

or, $1 \ge r$

or, $r \le 1$ …………..(iii)

From (ii) and (iii) we get, 1< r < 1.

i.e, coefficient lies between -1 to +1.

Theorem: Show that for two independent variable correlation coefficient is zero.

Proof: Let, $(x_1,y_1), (x_2,y_2)$………..(x_n,y_n) be the pairs of n observations. Then the arithmetic mean of x_i is $\bar{x}$ and y_i is $\bar{y}$. Since x and y are independent therefore,

$$\text{Covariance, Cov(x,y)}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}=0$$

$$\text{or, } \sum(x_i-\bar{x})(y_i-\bar{y})=0$$

$$\text{We Know, } \quad r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{0}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

= 0 (proved)

Application Problem-1: If y = mx + c, then find the correlation coefficient between x and y.

Solution: Let, $(x_1,y_1), (x_2,y_2)$………..(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,

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \text{ ……………..(1)}$$

Now, y = mx + c………..(ii)

Therefore, $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(mx_i + c - m\bar{x} - c)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(mx_i + c - m\bar{x} - c)^2}}$ ……………..(1)

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(mx_i - m\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(mx_i - m\bar{x})^2}}$$

$$= \frac{m\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{m\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1$$

Procedure for computing the correlation coefficient

Calculate the mean of the two series ‘x’ &’y’

Calculate the deviations ‘x’ &’y’ in two series from their respective mean.

Square each deviation of ‘x’ &’y’ then obtain the sum of the squared deviation i.e.Σx2& .Σy2

Multiply each deviation under x with each deviation under y & obtain the product of ‘xy’.Then obtain the sum of the product of x , y i.e. Σxy

Substitute the value in the formula.

Application Problem-1: A research physician recorded the pulse rates and the temperatures of water submerging the faces of ten small children in cold water to control the abnormally rapid heartbeats. The results are presented in the following table. 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: Calculating table of correlation coefficient.

x_i	y_i	x_i^2	y_i^2	x_iy_i
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	9	195
69	4	4761	16	276
63	6	3969	36	378
$\sum x_i = 635$	$\sum y_i = 63$	$\sum x_i^2 = 40537$	$\sum y_i^2 = 541$	$\sum x_i y_i = 3835$

We know, $r_{xy} = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$

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

$= -0.94$

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

Assignment problem-1: Compute r for the for the following paired sets of values:

i.(x, y): (1,2) , (2, 3), (3, 5), (4, 4), (5, 7)

ii. (x, y): (1,1) , (2, 3), (3, 5), (4, 7), (5, 9)

iii.(x, y): (1,10) , (2, 8), (3, 6), (4, 4), (5, 2)

iv.(x, y): (2,9) , (3, 5), (4, 6), (5, 2), (6, 1)

v.(x, y): (-2,4) , (-1, 1), (0, 0), (1, 1), (2, 4)

Solution 1: (x, y): (1,2) , (2, 3), (3, 5), (4, 4), (5, 7)

The formula for finding correlation coefficient is

$$r_{xy} = \frac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

Let us make a table to calculate correlation coefficient.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	4	16	16	16
5	7	25	49	35

$\sum x_i = 15$	$\sum y_i = 21$	$\sum x_i^{\ 2} = 55$	$\sum y_i^{\ 2} = 103$	$\sum x_i y_i = 74$

$$r_{xy} = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^{\ 2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^{\ 2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

$$= \frac{74 - \frac{15 \times 21}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{103 - \frac{(21)^2}{5}\right\}}}$$

$$= 0.90$$

Comment: There exists a strong positive relationship between x and y.

Problem: above ii-v (Assignment)

Assignment Problem-2: The following table gives the ages and blood pressure of 10 women:

Age in years x	56	42	36	47	49	42	72	63	55	60
Blood pressure y	147	125	118	128	125	140	155	160	149	150

Draw a scatter diagram

Find correlation coefficient between x and y and comment.

Ans: Try your-self

Assignment Problem-3: The scores of 12 students in their mathematics and physics classes are:

Mathematics	2	3	4	4	5	6	6	7	7	8	10	10
Physics	1	3	2	4	4	4	6	4	6	7	9	10

Find the correlation coefficient distribution and interpret it.

Comment on the followings:

(i) r=0 (ii) r=-1 (iii) r=1 (iv) r≥ 1 (v) r<1

(i) r=0, indicates that the correlation coefficient between x and y is zero.

(ii) r=-1, indicates that the correlation coefficient between x and y is perfect negative.

(iii) r= 1, indicates that the correlation coefficient between x and y is perfect positive.

(iv) r≥ 1i.e, r=1 and r>1 i.e, r>1, is not possible, because the Correlation coefficient lies between -1 to +1.

(v) r<1, not possible because, the Correlation coefficient lies between -1 to +1.

Uses of correlation coefficient.

1. To find the relationship between two variables.
2. To find the relationship between dependent variable and combined influence of a group of independent variables.
3. To solve many problem in biology.
4. In social studies like relationships between crime and educations, correlation analysis has got definite role to play.
5. In economies this is used specially.

RANK CORRELATION

Rank correlation: In some situation it is difficult to measure the values of the variables from bivariate distribution numerically, but they can be ranked. The correlation coefficient between these two ranks is usually called rank correlation coefficient, given by Spearman (1904). It is denoted by R. this is the only method for finding relationship between two qualitative variables like beauty, honesty, intelligence, efficiency and so on.

When there are no ties, the formula for computing the spearman's rank correlation coefficient

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

Here, R= rank correlation coefficient, n = number of pairs of observations being ranked.

d = difference between rank of x and rank of y.

Remarks:

(i) We always have $\sum d_i = \sum (R_1 - R_2) = 0$

(ii)Like simple correlation coefficient, rank correlation coefficient lies between -1 to +1.

Note: For finding rank correlation coefficient, we may have two types of data:

Actual observations are given

Actual ranks are given

Interpretation of Rank Correlation Coefficient (R)

The value of rank correlation coefficient, R ranges from -1 to +1

If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction

If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction

If R = 0, then there is no correlation

Application Problem-1: Obtain the rank correlation co-efficient for the following data:

A:	80	75	90	70	65	60
B:	65	70	60	75	85	80

Solution: Here ranks of the score are not given. Let us start ranking from the highest value for both the variables as shown in the table given below:

A	B	Rank of A (x)	Rank of B (y)	d = x-y	d^2
80	65	2	5	-3	9
75	70	3	4	-1	1
90	60	1	6	-5	25
70	75	4	3	1	1
65	85	5	1	4	16
60	80	6	2	4	16
Total				$\sum d_i == 0$	$\sum d_i^2 == 68$

$$R = 1- \frac{6\sum d^2}{n(n^2-1)} = 1- \frac{6\times 4}{6(6^2-1)} \quad = -\,0.94$$

Conclusion: There exist strongly negative relationship between A and B.

Application Problem -2: Obtain the rank correlation co-efficient for the following data:

Examiner	A	B	C	D	E
I	1	2	3	4	5
II	2	4	1	5	4

Solution: Here ranks of the score are given:

Ranking by examiner-I: R_1	Ranking by examiner-II: R_2	$d = R_1 - R_2$	d^2
1	2	-1	1
2	3	-1	1
3	1	2	4
4	5	-1	1
5	4	1	1
Total		$\sum d_i == 0$	$\sum d_i^2 == 8$

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 8}{5(5^2-1)} = 0.6$$

Comment: There is a positive rank correlation coefficient between the rankings of two examiners.

Repeated ranks or ties observations:

When ranks are repeated the following formula is used for finding rank correlation coefficient:

$$R = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\ldots\ldots\ldots\right\}}{n(n^2-1)}$$

Problems of equal ranks or tie in ranks:

Application Problem -3: The following data refer to the marks obtained by 8 students in mathematics and statistics:

Marks in mathematics	20	80	40	12	28	20	15	60
Marks in statistics	30	60	20	30	50	30	40	20

Compute rank correlation coefficient and comment.

Solution: let the marks obtained by mathematics be x and the marks obtained by statistics be y.

Table for computation of rank correlation.

x	y	Rank of x (R_1)	Rank of y (R_2)	$d = R_1 - R_2$	d^2

20	30	3.5	4	-0.5	0.25
80	60	8	8	0	0
40	20	6	2	4	16
12	30	1	4	-3	9
28	50	5	7	-2	4
20	30	3.5	4	-0.5	0.25
15	40	2	6	-4	16
60	10	7	1	6	36
					$\sum d_i^2 == 81.5$

Here, $m_1 = 2$, $m_2 = 3$, $n = 8$

$$R = 1 - \frac{6\left\{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right\}}{8\left(8^2 - 1\right)}$$

$= 0$

Merits Spearman's Rank Correlation

- This method is simpler to understand and easier to apply compared to karl pearson's correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Limitation Spearman's Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.
- This method should be applied where N exceeds 30.

Assignment problem-4:

The following figures relate to advertisement expenditure and profit:

Profit (Tk.Crore):x	25	28	27	33	31	10	16	16	18	23
Adv. Exp.(Tk. Lakh):y	87	91	92	95	93	52	68	72	78	86

(i)Draw a scatter diagram and comment

(ii) Calculate Karl Pearson's and Spearman rank correlation coefficients and comment.

Assignment problem-5:

The following figures relate to advertisement expenditure and sales of a company:

Adv. Exp. (Tk. Lac)	62	67	73	78	85	78	91	92	96	98
Sales (Tk.Crore)	11	13	17	18	21	24	21	27	26	21

Calculate Karl Pearson's correlation coefficient and Spearman rank correlation

Coefficient and comment.

Website:
http://www.pindling.org/Math/Statistics/Textbook/Examples/Chapter3/chapter3_examples.htm

REGRESSION ANALYSIS

What is regression?

Ans: The probable movement of one variable in terms of the other variables is called regression.

In other words the statistical technique by which we can estimate the unknown value of one variable (dependent) from the known value of another variable is called regression.

The term "regression" was used by a famous Biometrician Sir. F. Galton (1822-1911) in 1877.

Example: The productions of paddy of amount y is dependent on rainfall of amount x. Here x is independent variable and y is dependent variable.

Regression analysis.

Ans: Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of data.

Regression coefficient.

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

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

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

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

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

Regression lines:

If we consider two variables X and Y, we shall have two regression lines as the regression line of Y on X and the regression line of X on Y. The regression line of Y on X gives the most probable values of Y for given values of X and The regression line of X on Y gives the most probable values of X for given values of Y. Thus we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression lines will coincide i.e, we will have one line.

Regression equation:

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

y = a + bx , where y is the dependent variable to be estimated and x is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

Here, a = $\bar{y} - b\bar{x} = \dfrac{\sum y}{n} - b\dfrac{\sum x}{n}$ and b= $\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$= \frac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

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

x = a + by, where x is the dependent variable to be estimated and y is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

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

And b = $\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} {y_i}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

Properties of regression coefficient.

Ans: 1. Regression coefficient is independent of change of origin but not of scale.

2. Regression coefficient lies between $-\infty$ to $+\infty$. i.e, $-<\infty \quad b_{yx} < \infty$.

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

4. The geometric mean of regression coefficients is equal to correlation coefficient

i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

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

Coefficient. i.e, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

6. If one of regression coefficient is greater than unity the other must be less than

unity. i.e, $b_{yx} \geq 1$ and $b_{xy} < 1$

7. Regression coefficient is not pure number.

Coefficient of Determination, r^2 or R^2 :

The coefficient of determination, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The coefficient of determination is the ratio of the explained variation to the total variation.

The coefficient of determination is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y.

- The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if r = 0.922, then r^2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

- The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

Show that correlation coefficient is the geometric mean of regression coefficients. i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

Proof: Let, (x_1,y_1), (x_2,y_2)……….(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,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \ldots\ldots\ldots\ldots\ldots..(1)$$

Again, the regression coefficient of y on x is, $b_{yx} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Again, the regression coefficient of x on y is, $b_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

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

$$\sqrt{b_{yx} \times b_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$= r_{xy}$ (proved)

The arithmetic mean of two regression coefficient is greater than correlation coefficient. i.e, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

Proof: Let, (x_1,y_1), (x_2,y_2)……….. (x_n,y_n) be the pairs of n observations. Then the regression coefficient of y on x is denoted by b_{yx} and the regression coefficient of x on y is denoted by b_{xy}.

The arithmetic mean of b_{yx} and b_{xy} is A.M= $\left(\frac{b_{yx} + b_{xy}}{2}\right)$ and the geometric mean is

G.M= $\sqrt{b_{yx} \times b_{xy}}$

We know, Correlation coefficient is the geometric mean of regression coefficients.

i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

Since, A.M $\geq$ G.M

or, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq \sqrt{b_{yx} \times b_{xy}}$

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

Uses of regression.

Ans: (i) Whether a relationship exists or not.

(ii) To find the strength of relationship.

(iii) Determination of mathematical equation.

(iv) Prediction the values of the dependent variables.

Distinguish 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 of scale.
3. Correlation coefficient lies between -1 to +1. i.e, $-1< r_{xy} < 1$.	3. Regression coefficient lies between $-\propto$ to $+\propto$. i.e, $-<\propto b_{yx} < \propto$.
4. Correlation coefficient is symmetric. i.e, $r_{xy}= r_{yx}$ 5. It is always unit free.	4. Regression coefficient is not symmetric. i.e, $b_{xy} \neq b_{yx}$ 5. Regression coefficient 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.

Application problem-1: A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of Husband(in years):x	39	25	29	35	32	27	37
Age of wife(in years):y	37	18	20	25	25	20	30

(a) Compute the regression line of y on x.

(b) Predict the age of wife whose husband's age in 45 years.

(c) Find the regression line of x on y and estimate the age of husband if the age of his wife is 28 years.

(d) Compute the value of correlation coefficient with the help of regression coefficients.

Solution: The equation of the best –fitted regression line of y on x is $\hat{y}$ = a + bx

Where, $b = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$ and $a = \bar{y} - b\bar{x}$

Computation table

x	y	x^2	y^2	xy
39	37	1521	1369	1443
25	18	625	324	450
29	20	841	400	580
35	25	1225	625	875
32	25	1024	625	800
27	20	729	400	540
37	30	1369	900	1110
$\sum x = 224$	$\sum y = 175$	$\sum x^2 = 7334$	$\sum y^2 = 4643$	$\sum xy = 5798$

(a) Here,

$$b = \frac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{5798 - \dfrac{(224)(175)}{7}}{7334 - \dfrac{(224)^2}{7}} = 1.193$$

And $a = \bar{y} - b\bar{x}$

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

$= \frac{175}{7}$ -(1.193) $\frac{(224)}{7}$ = 25-38.176 = -13.176

Hence the fitted regression line is $\hat{y}$ = a + bx = -13.176 + 1.193x

(b) Hence, if the age of husband is 45, the probable age of wife would be

$\hat{y}$ = -13.176 + 1.193x = -13.176 + 1.193 $\times$ 45 = 40.51 years.

(c) The equation of the best –fitted regression line of y on x is $\hat{x}$ = a + by

$$\text{Where, } b = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

$$= \frac{5798 - \frac{(224)(175)}{7}}{4643 - \frac{(175)^2}{7}} = 0.739$$

And a = $\bar{x}$ - $b\bar{y}$

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

$$= \frac{224}{7} - 0.739\frac{175}{7} = 13.525$$

Hence the fitted regression line is $\hat{x}$ = a + by = 13.525+ 0.739y

Hence, if the age of wife is 28 years, the estimate age of husband is

$\hat{x}$ = a + by

= 13.525+ (0.739)(28) = 34.22 years.

Application problem-2: A research physician recorded the pulse rates and the temperatures of water submerging the faces of ten small children in cold water to control the abnormally rapid heartbeats. The results are presented in the following table. Calculate the correlation coefficient and regression coefficients 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

Also show that (i) $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

Solution: Calculating table of correlation coefficient and regression coefficients.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
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	9	195
69	4	4761	16	276
63	6	3969	36	378
$\sum x_i = 635$	$\sum y_i = 63$	$\sum x_i^2 = 40537$	$\sum y_i^2 = 541$	$\sum x_i y_i = 3835$

We know, $r_{xy} = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$

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

= -0.94

We know, the regression coefficient of y on x is, $b_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{3835 - \frac{635 \times 63}{10}}{40537 - \frac{(635)^2}{10}} = \ = \frac{-1655}{2145} = \text{-0.77}$$

Again, the regression coefficient of x on y is, $b_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}} = \frac{3835 - \frac{635 \times 63}{10}}{541 - \frac{(63)^2}{10}} = \frac{-1655}{1441} = \text{-1.1}$$

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

Here, $\left(\frac{b_{yx} + b_{xy}}{2}\right) = \frac{(-0.77) + (-1.1)}{2} =$ -0.94 = r_{xy}

Assignment Problem-1: The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store:

Test Scores: y	14	19	24	21	26	22	15	20	19
Sales(in lakh Taka)	31	36	48	37	50	45	33	41	39

(a) Find the regression equation of test scores on sales.

Ans: $\hat{y}$ = -2.4 + 0.56x

(b) Find the test scores when the sale is Tk. 40 lakh.

Ans: 20 lakh

(c) Find the regression equation of sales on test scores.

Ans: $\hat{x}$ = 7.8 + 1.61y

(d) Predict the value of sale if the test score is 30

Ans: 56.1 lakh

(e) Compute the value of correlation coefficient with the help of regression coefficients.

Assignment Problem-2: The following table gives the ages and blood pressure of 10 women:

Age in years x	56	42	36	47	49	42	72	63	55	60
Blood pressure y	147	125	118	128	125	140	155	160	149	150

(i) Obtain the regression line of y on x. Ans: $\hat{y}$ = 83.76+ 1.11x

(ii) Estimate the blood pressure of a women whose age is 50 years. Ans: 139.26

(iii) Obtain the regression line of x on y.

(iv) Find correlation coefficient between x and y and comment.

Assignment Problem-3: Consider the following data set on two variables x and y:

x : 1 2 3 4 5 6

y : 6 4 3 5 4 2

(a) Find the equation of the regression line y on x. Ans: $\hat{y}$ = 5.799-0.541x

(b) Graph the line on a scatter diagram.

(c) Estimate the value of y when x = 4.5 Ans: $\hat{y}$ = 3.486

(d) Predict the value of y when x = 8. Ans: $\hat{y}$ =1.687

Assignment Problem-4: Cost accountants often estimate overhead based on production. At the standard knitting company, they have collected information on overhead expenses and units produced at different plants and what to estimate a regression equation to predict future overhead.

Units	56	40	48	30	41	42	55	35
Overhead	282	173	233	116	191	171	274	152

(i)Draw a scatter diagram and comment

(ii)Fit a regression equation.

(iii)Estimate overhead when 65 units are produced.

Assignment Problem-5: The following data refer to information about annual sales

(Tk.'000) and year of experience of a super store of 8 salesmen:

Salesmen	1	2	3	4	5	6	7	8
Annual sales (Tk.'000)	90	75	78	86	95	110	130	145
Year of experience	7	4	5	6	11	12	13	17

(i)Fit two regression lines.

(ii)Estimate sales for year of experience is 10

(iii)Estimate year of experience for sales 100000

Assignment Problem (1-5): Same as solution-1

Final examination

SPRING-14 (CSE)

(a) Define regression line of y on x. Mention the properties of regression coefficients.

(b) The following data give the hardness(X) and tensile strength(Y) of 7 samples of metal in certain units.

X	146	152	158	164	170	176	182
Y	75	78	77	89	82	85	86

(i)Obtain the regression equation of y on x (ii)Estimate the value y when x is 69.

Spring-2012 (EEE)

(a) What is regression analysis? Write down the properties of regression coefficient.

(b) Prove that correlation coefficient is the geometric mean of regression coefficient.

(c) The regression coefficient of y on x is 0.5 and that of x on y is 1.9. Find the coefficient of correlation and also show that $r_{xy} \leq \left(\frac{b_{yx} + b_{xy}}{2}\right)$

Spring-2012 (ETE)

(a)Define regression analysis. What does regression coefficient measures?
(b) The regression coefficient of y on x is -0.8 and that of x on y is -0.6. Find the coefficient of correlation and comment.

Autumn-12 (CSE)

(a) Distinguish between correlation coefficient and regression coefficient.
(b) A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of Husband(in years):x	39	25	29	35	32	27	37
Age of wife(in years):y	37	18	20	25	25	20	30

(i) Find the regression line of wife on husband.
(ii) Estimate the probable age of wife if the age of husband is 30.
(iii) Compute the value of correlation coefficient with the help of regression coefficients.

(c) (i) Who had coined the term ‘regression’?
(ii) What are the limits of coefficient of correlation?
(iii) When will regression coefficients become coefficient of correlation?

Autumn-13(CSE)

(a) What are regression coefficients? Point out the properties of regression coefficients.

(b) The following data give the hardness(X) and tensile strength(Y) of 7 samples of metal in certain units.

X	146	152	158	164	170	176	182
Y	75	78	77	89	82	85	86

(i) Obtain the regression equation of y on x.

(ii)Estimate the x when y is 79.

(iii)

(c)What is the use of studying regression? Distinguish between correlation and regression.

Spring-13 (CSE)

(b) A researcher wants to find out if there is any relationship between the heights of the sons and the heights of the fathers. He took a random sample of six fathers and their six sons. Their heights in inches are given below:

Height of father(In inches): y	68	63	66	67	65	67
Height of Son(In inches): x	70	66	65	69	68	67

(i) Fit a regression line of the height of father y on the height of son x.

(ii) Predict the height of father if son's height is 65 inches.

(a) Define regression analysis. What does regression coefficient measures?

(b) If $\sum x = 56$, $\sum y = 40$, $\sum x^2 = 524$, $\sum y^2 = 256$, $\sum xy = 364$ and n = 8 then calculate the correlation coefficient and comment.

(c) The regression coefficient of y on x is -0.8 and that of x on y is -0.6. Find the coefficient of correlation and comment.

"THE END"