

MAT2001

Statistics for Engineers

Module 3

Correlation and Regression

Syllabus

Correlation and regression:

Correlation and Regression – Partial and Multiple correlation- Multiple regression.

Covariance

$$\text{Var}(X) = E\left[(X - E(X)) \cdot (X - E(X))\right]$$

$$\text{Covar}(X, Y) = E\left[(X - E(X)) \cdot (Y - E(Y))\right]$$

Correlation

CORRELATION COEFFICIENT

As the variance $E\{X - E(X)\}^2$ measures the variations of the R.V. X from its mean value $E(X)$, the quantity $E\{[X - E(X)][Y - E(Y)]\}$ measures the simultaneous variation of two R.V.'s X and Y from their respective means and hence it is called the *covariance of X, Y* and denoted as $\text{Cov}(X, Y)$.

$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$ is also called the *product moment of X and Y* and is also denoted as $p(X, Y)$.

$\frac{p(x, y)}{\sigma_x \sigma_y}$ is a measure of intensity of linear relationship between X and Y and is called *Karl-Pearson's Product Moment Correlation Coefficient* or simply *correlation coefficient* between X and Y . It is denoted by $r(X, Y)$ or r_{XY} or simply r .

Thus

$$r_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{E\{X - E(X)\}^2 E\{Y - E(Y)\}^2}} \quad (1)$$

since σ_x , the standard deviation of X is the positive square root of the variance of X .

$$r_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{E\{X - E(X)\}^2 E\{Y - E(Y)\}^2}}$$

$$r_{XY} = \frac{E(XY) - E(X) \cdot E(Y)}{\sqrt{\{E(X^2) - E^2(X)\} \{E(Y^2) - E^2(Y)\}}}$$

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

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

Properties of Correlation Coefficient

1. $-1 \leq r_{XY} \leq 1$ or $|\text{Cov}(X, Y)| \leq \sigma_X \cdot \sigma_Y$.

Note: When $0 < r_{XY} \leq 1$, the correlation between X and Y is said to be *positive* or *direct*.

When $-1 \leq r_{XY} \leq 0$, the correlation is said to be *negative* or *inverse*.

When $-1 \leq r_{XY} \leq -0.5$ or $0.5 \leq r_{XY} \leq 1$, the correlation is assumed to be high, otherwise the correlation is assumed to be poor.

2. Correlation coefficient is independent of change of origin and scale.

Example:

Compute the coefficients of correlation between X and Y using the following data:

X :	65	67	66	71	67	70	68	69
Y :	67	68	68	70	64	67	72	70

Comment about the nature of correlation.

Solution:

We effect change of origin in respect of both X and Y . The new origins are chosen at or near the average of extreme values. Thus we take $\frac{65+71}{2} = 68$ as the new origin for X and $\frac{64+72}{2} = 68$ as the new origin for Y . viz., we put $u_i = (x_i - 68)$ and $v_i = y_i - 68$ and find r_{UV} .

$X = x_i$	$Y = y_i$	$u_i = x_i - 68$	$v_i = y_i - 68$	u_i^2	v_i^2	$u_i v_i$
65	67	-3	-1	9	1	3
67	68	-1	0	1	0	0
66	68	-2	0	4	0	0
71	70	3	2	9	4	6
67	64	-1	-4	1	16	4
70	67	2	-1	4	1	-2
68	72	0	4	0	16	0
69	70	1	2	1	1	2
Total		-1	2	29	39	13

$$\begin{aligned} r_{XY} = r_{UV} &= \frac{n \sum uv - \sum u \cdot \sum v}{\sqrt{\{n \sum u^2 - (\sum u)^2\} \{n \sum v^2 - (\sum v)^2\}}} \\ &= \frac{8 \times 13 - (-1) \times 2}{\sqrt{(8 \times 29 - 1)(8 \times 39 - 4)}} = \frac{106}{\sqrt{231 \times 308}} \approx 0.3974 \end{aligned}$$

Exercise:

Find the coefficient of correlation between X and Y using the following data:

X :	5	10	15	20	25
Y :	16	19	23	26	30

Rank Correlation Coefficient

Sometimes the actual numerical values of X and Y may not be available, but the positions of the actual values arranged in order of merit (ranks) only may be available. The ranks of X and Y will in general, be different and hence may be considered as random variables. Let them be denoted by U and V . The correlation coefficient between U and V is called *the rank correlation coefficient* between (the ranks of) X , Y and denoted by ρ_{XY} .

Let us now derive a formula for ρ_{XY} or r_{UV} . Since U represents ranks of n values of X , U takes the values $1, 2, 3, \dots, n$.

Similarly V takes the same values $1, 2, 3, \dots, n$ in a different order.

$$D = U - V$$

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

[Note: The formula for the rank correlation coefficient is known as *spearman's formula*. The values of r_{XY} and ρ_{XY} (or r_{UV}) will be, in general, different.

Example:

Ten students got the following percentage of marks in Mathematics and Physical sciences:

Students:	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics:	78	36	98	25	75	82	90	62	65	39
Marks in Phy. Sciences:	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation coefficient.

Solution:

Denoting the ranks in Mathematics and in Phy. Sciences by U and V , we have the following values of U and V :

U :	4	9	1	10	5	3	2	7	6	8
V :	3	9	1	7	4	6	2	8	5	10
D :	1	0	0	3	1	-3	0	-1	1	-2
D^2 :	1	0	0	9	1	9	0	1	1	4

$\therefore \Sigma d^2 = 26$

$$\begin{aligned}\rho_{XY} = r_{UV} &= 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \times 26}{10 \times 99} = 0.8424\end{aligned}$$

Exercise:

Ten competitors in a beauty contest were ranked by three judges as follows:

	Competitors									
Judges	1	2	3	4	5	6	7	8	9	10
A:	6	5	3	10	2	4	9	7	8	1
B:	5	8	4	7	10	2	1	6	9	3
C:	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges have the nearest approach to common taste of beauty.

Regression

When the random variables X and Y are linearly correlated, the points plotted on the scatter diagram, corresponding to n pairs of observed values of X and Y , will have a tendency to cluster round a straight line. This straight is called *the regression line*. The regression line can be taken as the best fitting straight line for the observed pairs of values of X and Y in the least square sense, with which the students are familiar.

When two R.V.'s X and Y are linearly correlated, we may not know which variable takes independent values. If we treat X as the independent variable and hence assume that the values of Y depend on those of X , the regression line is called the *regression line of Y on X* . If we assume that the values of X depend on those of the independent variable Y , the *regression line of X on Y* is obtained. Thus in situations where the distinction cannot be made between the R.V.'s X and Y as to which is the independent variable and which is the dependent variable, there will be two regression lines. However, when the value of $Y(X)$ is to be predicted corresponding to a specified value of $X(Y)$, we should make use of the regression line of $Y(X)$ on $X(Y)$.

Equation of the Regression Line of Y on X:

By the principle of least squares, the normal equations which give the values of a and b .

are
$$\sum y_i = a \sum x_i + nb \quad (2)$$

and
$$\sum x_i y_i = a \sum x_i^2 + b \sum x_i \quad (3)$$

Dividing equation (2) by n , we get

$$\bar{y} = a \bar{x} + b \quad (4)$$

the equation of the regression line of Y on X as

$$y - \bar{y} = \frac{p_{XY}}{\sigma_X^2} (x - \bar{x})$$

$$y - \bar{y} = \frac{r_{XY} \sigma_Y}{\sigma_X} (x - \bar{x})$$

$$\left[\because r_{XY} = \frac{p_{XY}}{\sigma_X \sigma_Y} \right]$$

Equation of the Regression Line of X on Y:

In a similar manner, assuming the equation of the regression line of X and Y as $x = ay + b$ and using the equations

we can get the equation of the regression line of X on Y as

$$x - \bar{x} = \frac{P_{XY}}{\sigma_Y^2} (y - \bar{y})$$

or

$$x - \bar{x} = \frac{r_{XY} \sigma_X}{\sigma_Y} (y - \bar{y})$$

$$X_{on Y}; (y - \bar{y}) = \frac{r_{XY} \sigma_Y}{\sigma_X} (x - \bar{x})$$

$$X_{on Y}; (x - \bar{x}) = \frac{r_{XY} \sigma_X}{\sigma_Y} (y - \bar{y})$$

Example:

Obtain the equations of the lines of regression from the following data:

X:	1	2	3	4	5	6	7
Y:	9	8	10	12	11	13	14

Example:

Obtain the equations of the lines of regression from the following data:

X:	1	2	3	4	5	6	7
Y:	9	8	10	12	11	13	14

Find \bar{y} when $x = 3.4$
 $y = 10.9$
 $\bar{x} = 3.4$

x_i y_i x_i^2 y_i^2 $x_i y_i$

1
2
3
4
5
6
7

Total $\sum x_i$ $\sum y_i$ $\sum x_i^2$ $\sum y_i^2$ $\sum x_i y_i$

$$\bar{x} = \frac{\sum x_i}{n} \quad \sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2}$$

$$\bar{y} = \frac{\sum y_i}{n} \quad \sigma_y = \sqrt{\frac{1}{n} \sum y_i^2 - \left(\frac{\sum y_i}{n}\right)^2}$$

$$\text{Cov}(X, Y) = \frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i$$

$$r_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

Regression line of Y on X

$$(y - \bar{y}) = \frac{r_{xy} \sigma_y}{\sigma_x} (x - \bar{x})$$

$y = a + bx$ —→ ①
 Regression of X on Y

$$(x - \bar{x}) = \frac{r_{xy} \sigma_x}{\sigma_y} (y - \bar{y})$$

$x = a + by$ —→ ②
 \downarrow
 10.9

Solution:

X	Y	U = X - 4	V = Y - 11	U ²	V ²	UV
1	9	-3	-2	9	4	6
2	8	-2	-3	4	9	6
3	10	-1	-1	1	1	1
4	12	0	1	0	1	0
5	11	1	0	1	0	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
Total		0	0	28	28	26

$$\bar{x} = E(X) = 4 + \frac{1}{n} \sum u = 4$$

$$\bar{y} = E(Y) = 11 + \frac{1}{n} \sum v = 11$$

$$\sigma_X^2 = \frac{1}{n} \sum u^2 - \left(\frac{1}{n} \sum u \right)^2 = \frac{1}{7} \times 28 = 4$$

$$\sigma_Y^2 = \frac{1}{n} \sum v^2 - \left(\frac{1}{n} \sum v \right)^2 = \frac{1}{7} \times 28 = 4$$

$$C_{XY} = \frac{1}{n} \sum uv - \left(\frac{1}{n} \sum u \right) \left(\frac{1}{n} \sum v \right) = \frac{1}{7} \times 26 = 3.7$$

The regression line of Y on X is

$$y - \bar{y} = \frac{P_{XY}}{\sigma_X^2} (x - \bar{x})$$

$$\text{i.e., } y - 11 = \frac{3.7}{4} (x - 4)$$

$$\text{i.e., } 3.7x - 4y + 29.2 = 0$$

The regression line of X on Y is

$$x - \bar{x} = \frac{P_{XY}}{\sigma_Y^2} (y - \bar{y})$$

$$\text{i.e., } x - 4 = \frac{3.7}{4} (y - 11)$$

$$\text{i.e., } 4x - 3.7y + 24.7 = 0$$

Exercise:

Find the equations of the regression lines from the following data. Also estimate the value of Y when $X = 71$ and the value of X when $Y = 70$.

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

Exercise:

Obtain the equations of the regression lines from the following data, using the method of least squares.

Also estimate the value of (i) Y , when $X = 38$ and (ii) X , when $Y = 18$.

$X:$	22	26	29	30	31	31	34	35
$Y:$	20	20	21	29	27	24	27	31

$(X_1, X_2, X_3) \rightarrow$ Trivariate Distribution

$$r(X_1, X_2) = r_{X_1 X_2} = r_{12} = r_{21}$$

$$r(X_1, X_3) = r_{X_1 X_3} = r_{13} = r_{31}$$

$$r(X_2, X_3) = r_{X_2 X_3} = r_{23} = r_{32}$$

$$r_{x_1} = \frac{\text{Cov}(X_1, X)}{\sigma_X \cdot \sigma_X} = r_{YX}$$

Multiple and Partial Correlation

Multiple Correlation

Suppose one variable may be influenced by many other variables. Such a correlation is called multiple correlation.

Multiple Correlation Co-efficient (R)

In a trivariate distribution (X_1, X_2, X_3) , the multiple correlation co-efficient of X_1 on $X_2 \& X_3$ is denoted and defined as

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

||| $R_{2.13}$ & $R_{3.12}$

Note:

$$* R_{1.23}^2 \leq 1$$

$$|r_{xy}| \leq 1$$

$$* 0 \leq R_{1.23} \leq 1$$

Partial Correlation

The correlation between 2 Variables X_1 and X_2 may be partly due to the correlation of a third Variable X_3 with both X_1 and X_2 . In such a situation, the effect of X_3 on each of X_1 and X_2 were eliminated. Such a correlation is called Partial Correlation.

Partial Correlation Co-efficient:

The partial correlation co-efficient between X_1 & X_2 after eliminating the linear effect of X_3 , is denoted & defined as

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} \quad \text{and} \quad r_{23.1} = \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

Example:

In a trivariate distribution : $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$

Find the the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0.77 - 0.72 \times 0.52}{\sqrt{[1 - (0.72)^2][1 - (0.52)^2]}} = 0.62$$

$$\begin{aligned} R_{1.23}^2 &= \frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2} \\ &= \frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2} = 0.7334 \end{aligned}$$

$$\therefore R_{1.23} = \pm 0.8564$$

Exercise:

In a trivariate distribution : $r_{12} = 0.7, r_{23} = r_{31} = 0.5.$

Find (i) $r_{23.1},$ (ii) $R_{1.23},$

Multiple Regression

Regression equation of X_1 on X_2 and X_3 is given by

$$(X_1 - \bar{X}_1) \frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2) \frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3) \frac{\omega_{13}}{\sigma_3} = 0$$

where $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

$$\omega_{12} = - \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13} r_{23} - r_{21}$$

$$\omega_{13} = r_{23} r_{12} - r_{13}$$

Example:

Find the regression equation of X_1 on X_2 and X_3 given the following results :—

Trait	Mean	Standard deviation	r_{12}	r_{23}	r_{31}
X_1	28.02	4.42	+ 0.80	—	—
X_2	4.91	1.10	—	-0.56	—
X_3	594	85	—	—	- 0.40

**where X_1 = Seed per acre; X_2 = Rainfall in inches
 X_3 = Accumulated temperature above 42°F.**

Solution. Regression equation of X_1 on X_2 and X_3 is given by

$$(X_1 - \bar{X}_1) \frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2) \frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3) \frac{\omega_{13}}{\sigma_3} = 0$$

where $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0.56)^2 = 0.686$$

$$\omega_{12} = - \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13} r_{23} - r_{21} = -0.576$$

$$\omega_{13} = r_{23} r_{12} - r_{13} = (-0.56)(0.80) - (-0.40) = -0.048$$

\therefore Required equation of plane of regression of X_1 on X_2 and X_3 is given by

$$\frac{0.686}{4.42} (X_1 - 28.02) + \frac{(-0.576)}{1.10} (X_2 - 4.91) + \frac{(-0.048)}{85.00} (X_3 - 594) = 0$$

