Correlation and regression

Correlation :

We know that the quantity of measure of relationship between X and Y is given by Covariance between X and Y

Here, Covariance between X and $Y = Cov.(X,Y) = E(XY) - E(X)\, E(Y)$

and $\quad -\infty < Cov.(X,Y) < \infty.$

Correlation between X and $Y = r(X,Y) = \dfrac{Cov.(X,Y)}{\sigma_x\ \sigma_y}$

where $\qquad \sigma_x =$ s.d. of X and $\sigma_y =$ s.d. of Y

Limits for correlation coefficient :

Let $\quad E(X) = \mu_X$ and $E(Y) = \mu_Y$

Now consider $E\left[\left\{\dfrac{X-\mu_X}{\sigma_X} \pm \dfrac{Y-\mu_Y}{\sigma_Y}\right\}^2\right] \geq 0$

$$\Rightarrow \quad E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}\right] \geq 0$$

$$\Rightarrow \quad E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right] + E\left[\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right] \pm 2E\left[\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}\right] \geq 0$$

$$\Rightarrow \quad 1+1 \pm 2r(X,Y) \geq 0 \quad \Rightarrow \quad 1 \pm r(X,Y) \geq 0 \quad \Rightarrow \quad -1 \leq r(X,Y) \leq 1$$

Note : Correlation coefficient is independent of origin and scale.

Let $\quad U = \dfrac{X-a}{h},\ V = \dfrac{Y-b}{k}$ where a, b, h, k are constants, then $r(X,Y) = r(U,V)$

Formulas to calculate correlation coefficient :

For ungrouped data :

If X takes the values x_i, $i = 1,2,\ldots,n$ and Y takes the values y_i, $i = 1,2,\ldots,n$, then (x_i, y_i), $i = 1,2,\ldots,n$ are the n pair of observations.

$$E(X)=\frac{\sum_{i=1}^{n} x_i}{n}\ ,\quad E(Y)=\frac{\sum_{i=1}^{n} y_i}{n}\ ,\quad E(XY)=\frac{\sum_{i=1}^{n} x_i y_i}{n}\ ,$$

$$\sigma_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}-\left(E(X)\right)^2}\ ,\quad \sigma_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2}-\left(E(Y)\right)^2}$$

Correlation coefficient between X and $Y = r(X,Y)=\dfrac{E(XY)-E(X)E(Y)}{\sigma_X\sigma_Y}$

For grouped data (bivariate frequency distribution table) :

Let x_i, $i = 1,2,\ldots,n$ be the mid-points of the n class intervals defined for X with respective frequencies f_{x_i}, $i = 1,2,\ldots,n$.

Let y_j, $j = 1,2,\ldots,m$ be the mid-points of the m class intervals defined for Y with respective frequencies f_{y_j}, $j = 1,2,\ldots,m$.

Let f_{ij} be the frequency of i^{th} class interval of X correlated with j^{th} class interval of Y.

Let $N=\sum_{i=1}^{n} f_{x_i}=\sum_{j=1}^{m} f_{y_j}=\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}$

Now we can calculate

$$E(X)=\frac{\sum_{i=1}^{n} f_{x_i}x_i}{N},\quad E(Y)=\frac{\sum_{j=1}^{m} f_{y_j}y_j}{N},\quad E(XY)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}x_i y_j}{N}$$

$$\sigma_X=\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_{x_i}x_i^{\,2}-[E(X)]^2},\quad \sigma_Y=\sqrt{\frac{1}{N}\sum_{j=1}^{m} f_{y_j}y_j^{\,2}-[E(Y)]^2}$$

Correlation coefficient between X and $Y = r(X,Y)=\dfrac{E(XY)-E(X)E(Y)}{\sigma_X\sigma_Y}$

Problems on correlation coefficient :

1. Calculate the correlation coefficient for the following heights in inches of fathers (X) and their children (Y).

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

2. To calculate the correlation coefficient between two variables X and Y from 25 pairs of observations, the following data is obtained.

$$\sum X = 125, \sum Y = 100, \sum X^2 = 650, \sum Y^2 = 460, \sum XY = 508$$

Later it was discovered that the two pairs of data values (6,14), (8,6) are considered instead of correct pair of values (8,12), (6,8).

Hence, obtain the correlation coefficient for the correct data.

3. Calculate the correlation coefficient for the following bivariate frequency distribution table.

Y X	15-25	25-35	35-45	45-55	55-65	65-75
15-25	1	1				
25-35	2	12	1			
35-45		4	10	1		
45-55			3	6	1	
55-65				2	4	2
65-75					1	2

4. Calculate the correlation coefficient for the following bivariate frequency distribution table.

Y X	59-62	63-66	67-70	71-74	75-78
100	2	1			
120	7	8	4	2	
140	5	15	22	7	1
160	2	12	63	19	5
180		7	28	32	12
200		2	10	20	7
220			1	4	2

Lines of regression :

Let (x_i, y_i), $i = 1,2,\ldots,n$ be the given set of n points.

Aim : We want to get a straight line passing through the given set of n points.

If not, we want to get a best fit of straight line for the given set of n points.

Let $Y = a + bX$ be the best fit of straight line for the given set of n points if Y is dependent on X and this best fit of straight line is known as regression line of Y on X.

That means, we want to find the values of a and b such that

$$S = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \quad \text{is minimum.}$$

Here S is a function of a and b.

For S to be minimum, the necessary conditions are $\dfrac{\partial S}{\partial a} = 0, \ \dfrac{\partial S}{\partial b} = 0$

$$\frac{\partial S}{\partial a}=0 \quad \Rightarrow \quad na+b\sum x_i=\sum y_i \qquad (1)$$

$$\frac{\partial S}{\partial b}=0 \quad \Rightarrow \quad a\sum x_i+b\sum x_i^{\;2}=\sum x_i y_i \qquad (2)$$

Solving (1) and (2), we get *a* and *b*.

By solving, we get

$$b=\frac{n\Sigma x_i y_i-(\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^{\;2}-(\Sigma x_i)^2}=\frac{\frac{1}{n}\Sigma x_i y_i-\left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma y_i}{n}\right)}{\frac{1}{n}\Sigma x_i^{\;2}-\left(\frac{\Sigma x_i}{n}\right)^2}$$

$$=\frac{Cov(X,Y)}{\sigma_x^{\;2}}=\frac{Cov(X,Y)}{\sigma_x\sigma_y}\frac{\sigma_y}{\sigma_x}=r\frac{\sigma_y}{\sigma_x}$$

$$a=\frac{1}{n}(\Sigma y_i-b\Sigma x_i)$$

For these values of (*a*, *b*), S will be minimum if

$$\left.\frac{\partial^2 S}{\partial a^2}\right|_{(a,b)}>0 \qquad \text{and} \qquad \left[\frac{\partial^2 S}{\partial a^2}\frac{\partial^2 S}{\partial b^2}-\left(\frac{\partial^2 S}{\partial a\partial b}\right)^2\right]_{(a,b)}>0$$

Here

$$\left.\frac{\partial^2 S}{\partial a^2}\right|_{(a,b)}=n>0 \qquad \text{and}$$

$$\left[\frac{\partial^2 S}{\partial a^2}\frac{\partial^2 S}{\partial b^2}-\left(\frac{\partial^2 S}{\partial a\partial b}\right)^2\right]_{(a,b)}=n\sum x_i^{\;2}-\left(\sum x_i\right)^2=n^2\left[\frac{1}{n}\sum x_i^{\;2}-\left(\frac{\sum x_i}{n}\right)^2\right]=n^2\sigma_x^{\;2}>0$$

Therefore, at the values of (*a*, *b*) (the solution of solving the equations (1) and (2)), S is minimum.

Note :

1. If $\bar{x} = \frac{\sum x_i}{n}, \bar{y} = \frac{\sum y_i}{n}$, then regression line of Y on X is given by

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

2. If X is dependent on Y and the best fit of straight line $X = c + dY$ is known as regression line of X on Y.

The values of c and d can be found by using relations

$$d = r\frac{\sigma_X}{\sigma_Y} \quad \text{and} \quad c = \frac{1}{n}\left(\sum x_i - d\sum y_i\right)$$

Also we can get the regression line of X on Y by using the equation

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

3. The point $(\bar{x}, \bar{y})$ is the point intersection of both the regression lines.

4. The correlation coefficient is the G.M. (geometric mean) of two regression coefficients (slopes of the regression lines).

$$\therefore \quad r = \pm\sqrt{\left(r\frac{\sigma_Y}{\sigma_X}\right)\left(r\frac{\sigma_X}{\sigma_Y}\right)}$$

If both the regression coefficients are positive, then we will take $r > 0$.

If both the regression coefficients are negative, then we will take $r < 0$.

Problems on regression lines :

1. Find the correlation coefficient between X and Y for the following data. Also find the regression lines.

X	1	2	3	4	5	6	7	8	9	10
Y	10	12	16	28	25	36	41	49	40	50

2. Develop the estimating equation that best describes the following data and predict Y for X = 10, 15, 20.

X	13	16	14	11	17	9	13	17	18	12
Y	6.2	8.6	7.2	4.5	9.0	3.5	6.5	9.3	9.5	5.7

3. Cost accountants often estimate overhead based on the level of production. At the Standard Knitting Co., they have collected information on overhead expenses and units produced at different plants and want to estimate a regression equation to predict future overhead.

Overhead expenses	191	170	272	155	280	173	234	116	153	178
Units	40	42	53	35	56	39	48	30	37	40

Develop the regression equation for the cost accountants.

4. It is given that variance of $X = 9$ and the two regression lines are $8X - 10Y + 66 = 0$, $40X - 18Y - 214 = 0$. Calculate

(i) the mean values of X and Y

(ii) the correlation coefficient between X and Y

(iii) the standard deviation of Y

5. The variables X and Y are connected by the equation $aX + bY + c = 0$. Show that the correlation coefficient between them is -1 if signs of a and b are alike and +1 if they are different.