MA 2302: Introduction to Probability and Statistics

Correlation and Regression

Instructor
Prof. Gopal Krishna Panda
Department of Mathematics
NIT Rourkela

Correlation and Regression

Correlation

In the last chapter, we discussed about the joint distribution of two random variables X and Y. We come across random variables X and Y that are independent and dependent. X and Y that dependent means, there exists some sort of relation between them. Correlation deals with the amount of dependence between X and Y. In our discussion, we only discussion about the extent of linearity between X and Y. It is measured by the correlation coefficient $\boldsymbol{r}$ defined by

$$\boldsymbol{r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}}$$

where σ_X and σ_Y are standard deviations of X and Y and $Cov(X,Y)$ is the covariance between X and Y. We will see that r measures the amount of linear dependence between X and Y. Recall that

$$\boldsymbol{\sigma_X^2 = E(X-\mu_X)^2, \sigma_Y^2 = E(Y-\mu_Y)^2}$$

and

$$\boldsymbol{Cov(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\}}.$$

Example 1: *Let X and Y be random variables such that* $f(0,0) = 0.1, f(0,1) = 0.2,\ f(1,0) = 0.3$ *and* $f(1,1) = 0.4$. *Find the correlation coefficient between X and Y.*

Ans. To find the correlation coefficient r, we need to calculate the variances of X and Y and the covariance between X and Y. In tabular form, the joint distribution of X and Y and the marginal distributions of X and Y are given in the following tables:

$x \downarrow y \rightarrow$	**0**	**1**	$f_1(x) \downarrow$
0	0.1	0.2	**0.3**
1	0.3	0.4	**0.7**
$f_2(y) \rightarrow$	0.4	0.6	**1**

Thus,

$$\mu_X = E(X) = 0 \times 0.3 + 1 \times 0.7 = 0.7, E(X^2) = 0^2 \times 0.3 + 1^2 \times 0.7 = 0.7,$$

$$\mu_Y = E(Y) = 0 \times 0.4 + 1 \times 0.6 = 0.6, E(Y^2) = 0^2 \times 0.4 + 1^2 \times 0.6 = 0.6,$$

and

$$E(XY) = \sum_x \sum_y xy\, f(x,y) = 0 \times 0 \times 0.1 + 0 \times 1 \times 0.2 + 1 \times 0 \times 0.3$$

$$+1 \times 1 \times 0.4 = 0.4.$$

Hence,

$$\sigma_X^2 = E(X^2) - \mu_X^2 = 0.7 - 0.7^2 = 0.21,$$

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = 0.6 - 0.6^2 = 0.24,$$

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y = 0.4 - 0.7 \times 0.6 = -0.02.$$

Hence,

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{-0.02}{\sqrt{0.21 \times 0.24}} = -0.0891.$$

Example 2: *If* (X, Y) *has the pdf* $f(x,y) = \frac{1}{32}$ *if* $x \geq 0, y \geq 0$, $x + y \leq 8$ *and* $f(x,y) = 0$ *otherwise, find* r.

Ans. It is easy to see that the marginal densities of X and Y are

$$f_1(x) = \begin{cases} \dfrac{8-x}{32} & \text{if } 0 \leq x \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_2(y) = \begin{cases} \dfrac{8-y}{32} & \text{if } 0 \leq y \leq 8 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\mu_X = \int_{-\infty}^{\infty} x\, f_1(x) dx = \int_0^8 x \cdot \frac{8-x}{32} dx = \frac{8}{3}, \quad \mu_Y = \frac{8}{3}.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2\, f_1(x) dx = \int_0^8 x^2 \cdot \frac{8-x}{32} dx = \frac{32}{3}, E(Y^2) = \frac{32}{3}.$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dxdy = \frac{1}{32}\int_0^8\int_0^{8-y} xy\, dxdy$$

$$= \frac{1}{32}\int_0^8 y\left\{\int_0^{8-y} x\, dx\right\} dy = \frac{1}{64}\int_0^8 y(8-y)^2\, dy = \frac{16}{3}.$$

Hence,

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \frac{32}{3} - \left(\frac{8}{3}\right)^2 = \frac{32}{9}, \quad \sigma_Y^2 = \frac{32}{9}.$$

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = \frac{16}{3} - \frac{8}{3}\times\frac{8}{3} = -\frac{16}{9}.$$

Hence,

$$r = \frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y} = \frac{-\frac{16}{9}}{\sqrt{\frac{32}{9}\times\frac{32}{9}}} = -\frac{1}{2}.$$

Limits of the correlation coefficient

Theorem 1: $-1 \leq r \leq 1$.

Proof: If X is any random variable, then $E(X^2) \geq 0$. In particular,

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

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

$$\Rightarrow \frac{1}{\sigma_X^2}E(X-\mu_X)^2 + \frac{1}{\sigma_Y^2}E(Y-\mu_Y)^2 \pm 2\cdot\frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X\cdot\sigma_Y} \geq 0$$

$$\Rightarrow \frac{1}{\sigma_X^2}\times\sigma_X^2 + \frac{1}{\sigma_Y^2}\times\sigma_Y^2 \pm 2\times\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y} \geq 0$$

$$\Rightarrow \frac{1}{\sigma_X^2} \times \sigma_X^2 + \frac{1}{\sigma_Y^2} \times \sigma_Y^2 \pm 2 \times \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \geq 0$$

$$\Rightarrow 1 + 1 \pm 2r \geq 0 \Rightarrow 1 + r \geq 0 \text{ and } 1 - r \geq 0$$

Thus,

$$r \geq -1 \text{ and } r \leq 1 \Rightarrow -1 \leq r \leq 1.$$

Observe that $\boldsymbol{r = 1}$ or $\boldsymbol{r = -1}$ according as

$$E\left[\frac{X - \mu_X}{\sigma_X} - \frac{Y - \mu_Y}{\sigma_Y}\right]^2 = 0 \text{ or } E\left[\frac{X - \mu_X}{\sigma_X} + \frac{Y - \mu_Y}{\sigma_Y}\right]^2 = 0$$

which happens iff $\left[\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y}\right]^2 = 0$ or $\left[\frac{X-\mu_X}{\sigma_X} + \frac{Y-\mu_Y}{\sigma_Y}\right]^2 = 0$ with probability 1, or equivalently, $\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y} = 0$ or $\frac{X-\mu_X}{\sigma_X} + \frac{Y-\mu_Y}{\sigma_Y} = 0$ with probability 1, which is a clear indication that ***X* and *Y* are linearly related** with probability 1. Notice that if $r = 1$, the slope is positive and if $r = -1$, the slope is negative.

Theorem 2: *If X and Y are linearly related, then* $|r| = 1$.

Proof: Assume that X and Y are linearly related, say $Y = mX + c$. Then,

$$\mu_Y = E(mX + c) = mE(X) + c = m\mu_X + c.$$

$$\sigma_Y^2 = E(Y - \mu_Y)^2 = E\{(mX + c) - (m\mu_X + c)\}^2$$

$$= E(mX - m\mu_X)^2 = E\{m^2(X - \mu_X)^2\}$$

$$= m^2 E(X - \mu_X)^2 = m^2\sigma_X^2 \Rightarrow \sigma_Y = |m|\sigma_X.$$

$$Cov(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\} = E\{(X - \mu_X)(mX - m\mu_X)\}$$

$$= E\{m(X - \mu_X)(X - \mu_X)\} = mE(X - \mu_X)^2 = m\sigma_X^2.$$

Hence,

$$r = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{m\sigma_X^2}{\sigma_X \cdot |m|\sigma_X} = \frac{m}{|m|} = \begin{cases} 1 & \text{if } m > 0, \\ -1 & \text{if } m < 0. \end{cases}$$

- Thus, if X and Y are linearly related with positive slope, then $r = 1$.
- If if X and Y are linearly related with negative slope, then $r = -1$.
- Indeed, if an increase (decrease) in X, results in a corresponding increase (decrease) in Y, then r is positive.
- If an increase (decrease) in X, results in a corresponding decrease (increase) in Y, then r is negative.
- The maximum correlation is $|r| = 1$.
- No linearity means r is close to zero.
- $|r|$ denotes the amount of linear dependence between X and Y.
- The sign of r and $Cov(X, Y)$ are same.
- X and Y are uncorrelated (i.e. $r = 0$) if and only if $Cov(X, Y) = 0$.
- $r = 0$ does not mean that X and Y are not related, this simply means that there is no linearity between X and Y.

Theorem 3: *The correlation coefficient r has no unit and is independent of change of origin and scale.*

Proof: The first part follows from the definition. Now assume that $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$ where h and k are positive. We will prove that $r(X,Y) = r(U,V)$. Observe that $X = a + hU, Y = b + kV$. Hence,

$$\mu_X = E(a + hU) = a + hE(U) = a + h\mu_U,$$

$$\mu_Y = E(b + kV) = b + kE(V) = b + k\mu_V,$$

$$\sigma_X^2 = E(X - \mu_X)^2 = E\{(a + hU) - (a + h\mu_U)\}^2 = E(hU - h\mu_U)^2$$

$$= E\{h^2(U - \mu_U)^2\} = h^2E(U - \mu_U)^2 = h^2\sigma_U^2 \Rightarrow \sigma_X = h\sigma_U.$$

Similarly,

$$\sigma_Y^2 = E(Y - \mu_Y)^2 = E\{(b + kV) - (b + k\mu_V)\}^2 = E(kV - k\mu_V)^2$$

$$= E\{k^2(V - \mu_V)^2\} = k^2 E(V - \mu_V)^2 = k^2\sigma_V^2 \Rightarrow \sigma_Y = k\sigma_V.$$

Now,

$$Cov(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\} = E\{h(U - \mu_U)k(V - \mu_V)\}$$

$$= hkE\{(U - \mu_U)(V - \mu_V)\} = hk\ Cov(U, V).$$

Hence,

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{hk\ Cov(U, V)}{h\sigma_U \cdot k\sigma_V} = \frac{Cov(U, V)}{\sigma_U \cdot \sigma_V} = r(U, V).$$

Hence, the correlation coefficient is independent of change of origin and scale.

Theorem 4: *If X and Y are independent, then $r = 0$. The converse is not true.*

Proof: If X and Y are independent, then

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.$$

Hence,

$$r = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = 0.$$

Moreover, let X be uniformly distributed in the interval $[-1,1]$ and $Y = X^2$. Hence

$$f(x) = \begin{cases} \frac{1}{2} \text{ if } -1 \leq x \leq 1 \\ 0 \qquad \text{otherwise.} \end{cases}$$

Then,

$$\mu'_{2r-1} = E(X^{2r-1}) = \frac{1}{2}\int_{-1}^{1} x^{2r-1}\, dx = 0, r = 1,2,3, \ldots$$

In particular, $\mu = E(X) = 0$. Hence, $E(X^{2r-1}) = \mu'_{2r-1} = \mu_{2r-1} = 0, r = 1,2,3$ .... Now,

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0 \Rightarrow r = 0.$$

However, X and Y are not independent since $Y = X^2$.

Example 3: *If X and Y are independent random variables with variances σ_X^2 and σ_Y^2, $U = X + Y$ and $V = X - kY$, then find k such that U and V are uncorrelated.*

Ans. If X and Y are independent, then $r(X,Y) = 0$ and hence $Cov(X,Y) = 0$. U and V will be uncorrelated if and only if $Cov(U,V) = 0$. But,

$$\begin{aligned} Cov(U,V) &= Cov(X+Y, X-kY) \\ &= E[\{(X+Y) - (\mu_X + \mu_Y)\}\{(X-kY) - (\mu_X - k\mu_Y)\}] \\ &= E[\{(X-\mu_X) + (Y-\mu_Y)\}\{(X-\mu_X) - k(Y-\mu_Y)\}] \\ &= E\{(X-\mu_X)^2\} - kE\{(X-\mu_X)(Y-\mu_Y)\} \\ &\qquad + E\{(X-\mu_X)(Y-\mu_Y)\} - kE\{(Y-\mu_Y)^2\} \\ &= \sigma_X^2 - kCov(X,Y) + Cov(X,Y) - k\sigma_Y^2 = \sigma_X^2 - k\sigma_Y^2. \end{aligned}$$

Thus, U and V are uncorrelated if

$$\sigma_X^2 - k\sigma_Y^2 = 0 \Rightarrow k = \frac{\sigma_X^2}{\sigma_Y^2}.$$

Example 4: *If* $X \sim N(0,1)$ *find the correlation between* X *and* $Y = X^3$.

Ans. If $X \sim N(0,1)$, then $E(X) = 0$, hence the moments about origin are the central moments. In particular,

$$\mu_{2r-1} = E(X^{2r-1}) = 0,$$

$$\mu_{2r} = E(X^{2r}) = 1 \cdot 3 \cdot \cdots \cdot (2r-1) = \frac{(2r)!}{2^r r!}, r = 1,2, \dots$$

Hence,

$$\mu_Y = E(Y) = E(X^3) = 0, \sigma_X^2 = 1.$$

$$\sigma_Y^2 = E(Y^2) - E^2(Y) = E(X^6) - E^2(X^3) = E(X^6) = 1 \cdot 3 \cdot 5 = 15.$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^4) = 1 \cdot 3 = 3.$$

Hence,

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{3}{\sqrt{1 \cdot 15}} = \sqrt{3/5} = 0.7746.$$

Correlation coefficient for a sample

If a sample of a bivariate distribution is given, then the sample correlation coefficient can be calculated by a similar formula. Assume that the marks of a group of n students in two tests are as follows: (X → mark in first test, Y → mark in second test)

X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n

Then,

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

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

and

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Correlation coefficient for a sample

In some texts, the sample correlation coefficient is calculated slightly differently. It is defined as

$$r = \frac{S_{X,Y}}{S_X \cdot S_Y}$$

where,

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right], S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

$$S_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right]$$

Here, S_X^2, S_Y^2 and $S_{X,Y}$ are respectively called the sample variance of X, the sample variance of Y and the sample covariance of X and Y. Actually, S_X^2, S_Y^2 and $S_{X,Y}$ are the unbiased estimate of their corresponding population parameters. However, the sample correlation coefficient calculated by either formula gives the same value. Hence, we follow the former formula.

Example 5: *The following table gives the marks of 5 students in two tests out of 20* (X → mark in first test, Y → mark in second test). *Find the correlation coefficient.*

X	7	15	12	11	9
Y	10	14	8	15	13
X^2	49	225	144	121	81
Y^2	100	196	64	225	169
XY	70	210	96	165	117

$$\sum x_i = 54, \sum y_i = 60, \sum x_i^2 = 620, \sum y_i^2 = 754, \sum x_i y_i = 658$$

$$\bar{x} = \frac{1}{5}\sum x_i = 10.8,\ \bar{y} = \frac{1}{5}\sum y_i = 12,$$

From the table,

$$\sum x_i = 54, \sum y_i = 60, \sum x_i^2 = 620, \sum y_i^2 = 754, \sum x_i y_i = 658.$$

Hence,

$$\bar{x} = \frac{1}{5}\sum x_i = 10.8,\ \bar{y} = \frac{1}{5}\sum y_i = 12,$$

$$\sigma_X^2 = \frac{1}{5}\sum x_i^2 - \bar{x}^2 = \frac{1}{5} \times 620 - 10.8^2 = 7.36,$$

$$\sigma_Y^2 = \frac{1}{5}\sum y_i^2 - \bar{y}^2 = \frac{1}{5} \times 754 - 12^2 = 6.8,$$

$$Cov(X,Y) = \frac{1}{5}\sum x_i y_i - \bar{x}\bar{y} = \frac{1}{5} \times 658 - 10.8 \times 12 = 2.$$

Hence,

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{2}{\sqrt{7.36 \times 6.8}} = \mathbf{0.2827}.$$

Example 6: Find r *and verify that in the former case,* r *is positive and in the later case* r *is negative.*

X	1	2	3	4	5	6	7	8	9	10
Y	1	4	9	16	25	36	49	64	81	100

Verify that for the above data $r = 0.9746$. For the second data $r = -0.9746$.

X	1	2	3	4	5	6	7	8	9	10
Y	100	81	64	49	36	25	16	9	4	1

For the following data prove that $r = 0$.

X	-4	-3	-2	-1	0	1	2	3	4	
Y	16	9	4	1	0	1	4	9	16	

Assignment: From $S = \{1,2,3,4,5\}$ consider all possible samples of size 3. They are $\binom{5}{3} = 10$ in number. List all of them. Then calculate $\bar{x}$, S^2 and σ^2 for all the samples. Do the same thing for S. Then take arithmetic means of all the sample values, compare with corresponding population values. Give your conclusion.

Rank correlation:

We know that for the data

$\boldsymbol{X}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\cdots$	$\boldsymbol{x_n}$
$\boldsymbol{Y}$	$\boldsymbol{y_1}$	$\boldsymbol{y_2}$	$\boldsymbol{y_3}$	$\cdots$	$\boldsymbol{y_n}$

the correlation coefficient is given by

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

where

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

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Now, consider the situation where in a class of size n, the students are arranged according to their SGPA in first semester and accordingly, ranks are assigned, Rank 1 being assigned to the student with highest SGPA, Rank 2 to the next highest SGPA and so on. The same thing is also done in the second semester, i.e. the students are again ranked according to their SGPA in second semester. It may so happen that a student who secures highest SGPA need not do so in the second semester or a student whose rank is 5 in first semester, may remain fixed, may improve or may go down. Sometimes, it is necessary to know if there is any correlation between these ranks. It can be done by calculating the correlation coefficient between ranks of SGPA of first and second semester. We denote by X, the rank in first semester and Y, the rank in second semester. Thus, the possible values of both X and Y are $1,2,\ldots,n$ but need not in the same order. We first calculate the correlation coefficient between X and Y assuming that there are no repeated ranks using the usual formula for correlation.

Observe that

$$\sum x_i = \sum y_i = \frac{n(n+1)}{2} \Rightarrow \overline{x} = \bar{y} = \frac{n+1}{2},$$

$$\sum x_i^2 = \sum y_i^2 = \frac{n(n+1)(2n+1)}{6} \Rightarrow \sigma_X^2 = \sigma_Y^2 = \frac{n^2-1}{12}$$

Let $d_i = x_i - y_i, i = 1,2, \dots, n$. Then

$$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum (x_i - y_i)^2 = \frac{1}{n}\sum \{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \frac{1}{n}\sum (x_i - \bar{x})^2 + \frac{1}{n}\sum (y_i - \bar{y})^2 - 2 \times \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sigma_X^2 + \sigma_Y^2 - 2Cov(X,Y) = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y.$$

Since

$$\sigma_X^2 = \sigma_Y^2 = \frac{n^2-1}{12},$$

it follows that

$$\frac{1}{n}\sum d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y = 2 \times \frac{n^2-1}{12} - 2r \times \frac{n^2-1}{12}$$

$$= (1-r) \times \frac{n^2-1}{6}.$$

$$\Rightarrow 1 - r = \frac{6\sum d_i^2}{n(n^2-1)}.$$

Hence,

$$\boldsymbol{r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}}.$$

Example 7: *The following table gives the marks of 10 students in two tests out of 100* (X → mark in first test, Y → mark in second test). *Find the rank correlation coefficient.*

Test I Marks	45	87	55	67	97	25	75	48	52	17
Test II Marks	65	70	78	82	81	32	67	55	60	37

Ans. To calculate the rank correlation coefficient, we proceed as follows:

Rank of Test I (x)	**8**	**2**	**5**	**4**	**1**	**9**	**3**	**7**	**6**	**10**
Rank of Test II (y)	**6**	**4**	**3**	**1**	**2**	**10**	**5**	**8**	**7**	**9**
$\boldsymbol{d = x - y}$	$\mathbf{2}$	$\mathbf{-2}$	$\mathbf{2}$	$\mathbf{3}$	$\mathbf{-1}$	$\mathbf{-1}$	$\mathbf{-2}$	$\mathbf{-1}$	$\mathbf{-1}$	$\mathbf{1}$
$\boldsymbol{d^2}$	**4**	**4**	**4**	**9**	**1**	**1**	**4**	**1**	**1**	**1**

Thus, $n = 10, \sum d_i^2 = 30$. Hence, the rank correlation coefficient is given by

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 30}{10(100-1)} = \frac{9}{11}.$$

Repeated ranks

When two or more X or Y values are same, the rank assigned to these values are the average values in the case they are different. For each such repeated ranks, the factor

$$\frac{m(m^2-1)}{12}$$

is to be added to $\sum d_i^2$. The following example illustrates this correction.

Example 7: *The following table gives the marks of 10 students in two tests out of 100* (X → mark in first test, Y → mark in second test). *Find the rank correlation coefficient.*

Test I Marks	45	87	55	61	97	25	75	48	61	17
Test II Marks	65	65	78	82	81	32	65	55	60	37

Ans. To calculate the rank correlation coefficient, we proceed as follows:

Rank of Test I (x)	**8**	**2**	**6**	**4.5**	**1**	**9**	**3**	**7**	**4.5**	**10**
Rank of Test II (y)	**5**	**5**	**3**	**1**	**2**	**10**	**5**	**8**	**7**	**9**
$\boldsymbol{d = x - y}$	$\mathbf{3}$	$\mathbf{-3}$	$\mathbf{3}$	$\mathbf{-3.5}$	$\mathbf{-1}$	$\mathbf{-1}$	$\mathbf{-2}$	$\mathbf{-1}$	$\mathbf{-2.5}$	$\mathbf{1}$
$\boldsymbol{d^2}$	**9**	**9**	**9**	**12.25**	**1**	**1**	**4**	**1**	**6.25**	**1**

Thus, $n = 10, \sum d_i^2 = 53.5$. There are 2 repeated x-rank and 3 y-rank. Hence, the factor to be added to $\sum d_i^2$ is

$$\frac{2(2^2 - 1)}{12} + \frac{3(3^2 - 1)}{12} = \frac{30}{12} = 2.5.$$

Hence, the rank correlation coefficient is given by

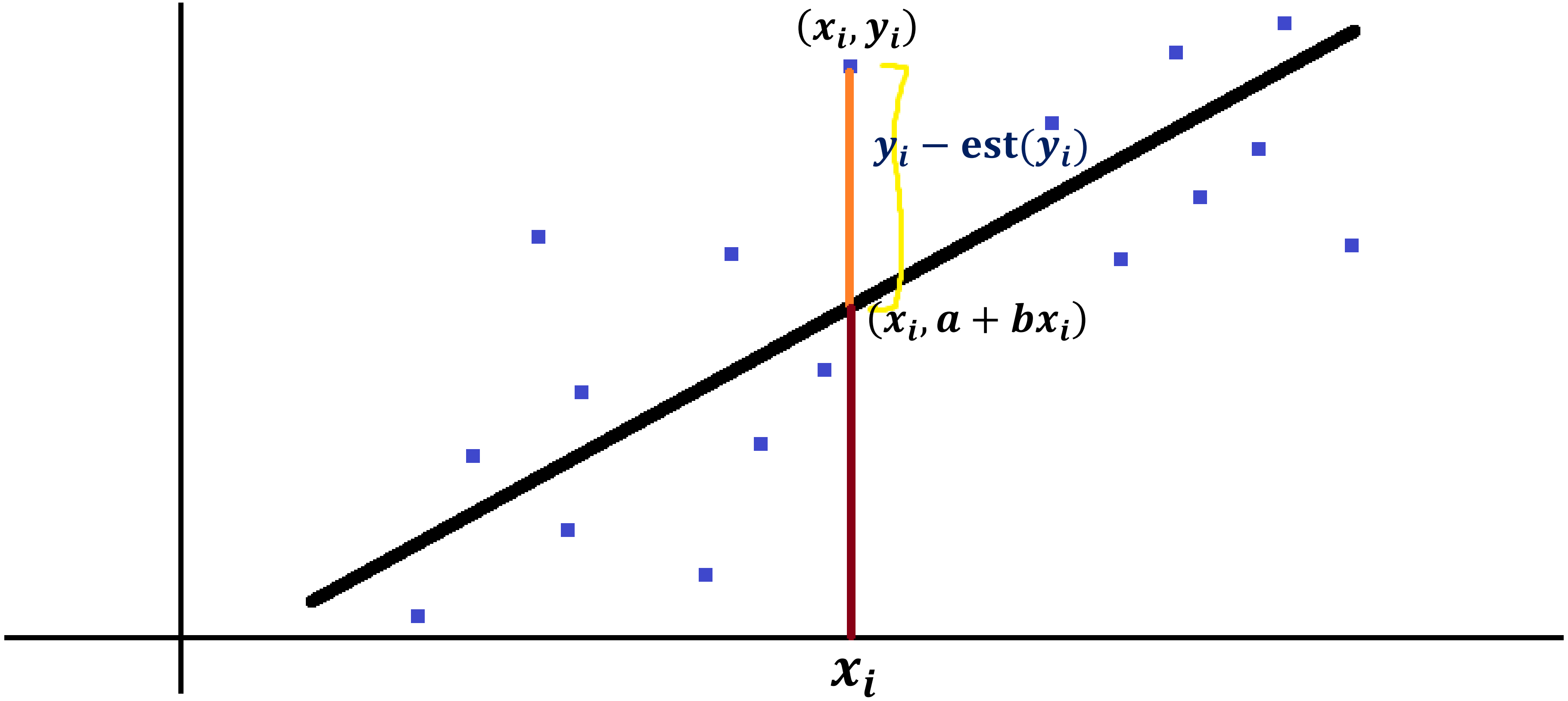
$$r = 1 - \frac{6\left(\sum d_i^2 + \sum \frac{m(m^2 - 1)}{12}\right)}{n(n^2 - 1)} = 1 - \frac{6(53.5 + 2.5)}{10(100 - 1)} = 1 - \frac{56}{165} = \frac{109}{165}.$$

Regression

As we have discussed, the correlation coefficient between X and Y expresses the amount of linearity between X and Y numrically. The purpose of regression is to approximate the relationship between X and Y by linearity. This is achieved in two ways: estimating Y without disturbing X such that the points $(X, est(Y))$ are on a straight line and estimating X without disturbing Y such that $(est(X), Y)$ are on a straight line.

Assume that we are given with n pairs of values of (X, Y) say $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ which are not on a straight line and we want to find a line which passes as close as possible to these points. Let the line be of the form $Y = a + bX$ which means that we want to replace y_i by $\mathrm{est}(y_i) = a + bx_i$, $i = 1,2, \dots, n$ such that $\left(x_i, \mathrm{est}(y_i)\right)$ is on $Y = a + bX$.

In the following figure, the error $\boldsymbol{E_i = y_i - \mathbf{est}(y_i) = y_i - a - bx_i}$ that occurs



due to the replacement of $(\boldsymbol{x_i, y_i})$ by means of $(\boldsymbol{x_i, a + bx_i})$ is shown. There are n such errors which need to be minimized in some way for the best possible line.

It is clear from the figure that the error in replacing the point (x_i, y_i) by $\left(x_i, \text{est}(y_i)\right) = (x_i, a + bx_i)$ is equal to $E_i = y_i - a - bx_i, i = 1,2, \dots, n$. We need to minimize these errors in some way. We can't make all errors zero since the points are not on a line. Since positive and negative errors of equal magnitude are equally important, we need to consider some function of the absolute error to be minimized. Since it is difficult to minimize sum of absolute errors we consider minimizing the sum of squares of the errors

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

with respect to a and b such that the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ lie as close as possible to the line $Y = a + bX$. This minimization is know as the **principle of least squares** which requires the calculus of functions of two variables. Since S has no maximum (e.g., $S \to \infty$ if $b = 0$ and $a \to \infty$), the minimum of S corresponds to the values of a and b satisfying

$$\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0.$$

Thus,

$$\sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0, \qquad \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0.$$

The above equations are known as normal equations and can be simplified as

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0, \qquad \sum_{i=1}^{n} \left(x_i y_i - a x_i - b x_i^2\right) = 0.$$

On rearrangement, we get

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i\,, \qquad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2\,.$$

Dividing both the equations by n, we get

$$\bar{y} = a + b\bar{x}, \qquad \frac{1}{n} \sum_{i=1}^{n} x_i y_i = a\bar{x} + b \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2\,.$$

$$\bar{y} = a + b\bar{x}, \qquad \frac{1}{n}\sum_{i=1}^{n} x_i y_i = a\bar{x} + b \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2 .$$

Eliminating a from the first equation and substituting in the second equation, we get

$$a = \bar{y} - b\bar{x}, b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{Cov(X,Y)}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}.$$

The value of b so obtained is denoted by

$$b_{YX} = \frac{Cov(X,Y)}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

and is known as the *regression coefficient of Y on X*. Now, $a = \bar{y} - b_{YX}\bar{x}$ and the line of the form $Y = a + bX$ with minimum mean square error is

$$Y = a + bX = \bar{y} - b_{YX}\bar{x} + b_{YX}X$$

and can be written as a more convenient form as and can be written as a more convenient form as

$$\boldsymbol{Y-\bar{y}=b_{YX}(X-\bar{x})}$$

and is known as the *line of regression of Y on X*, or simply *LR of Y on X*.

The *line of regression of Y on X* , that is $Y=a+bX$, has been obtained keeping X fixed and estimating Y such that the values of $(X,est(Y))$ are on a straight line. We can similarly consider the problem of finding a line of the form $X=a+bY$, where we keep Y fixed and estimate X such that the values of $(est(X),Y)$ lie on a straight line. For this we have to repeat the entire calculation we have done for fitting the line $Y=a+bX$. However, the can also be obtained from fitting we have already done just by interchanging X and Y. In doing so, we obtain the line

$$\boldsymbol{X-\bar{x}=b_{XY}(Y-\bar{y})}$$

which is known as the *line of regression of X on Y*, or simply *LR of X on Y*.

The multiplier b_{XY} of

$$\boldsymbol{X - \bar{x} = b_{XY}(Y - \bar{y})}$$

is known as the *regression coefficient of X on Y* and is equal to

$$b_{XY} = \frac{Cov(X, Y)}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}.$$

- In LR of Y on X, X is fixed, Y is estimated so that $(X, est(Y))$ are on a straight line.
- In LR of X on Y, Y is fixed, X is estimated so that $(est(X), Y)$ are on a straight line.
- The LR of Y on X and LR of X on Y can also be written as

$$\boldsymbol{Y - \bar{y} = b_{YX}(X - \bar{x})}, \qquad \boldsymbol{Y - \bar{y} = \frac{1}{b_{XY}}(X - \bar{x})}.$$

- The slope of LR of Y on X is b_{YX}, the regression coefficient of X on Y while the slope of LR of X on Y is $1/b_{XY}$ the reciprocal of the regression coefficient.

- Since the signs of b_{YX} , b_{XY} and r are same as the sign of $Cov(X,Y)$, hence both the regression coefficients and the correlation coefficient are of the same sign.

- Hence, the slopes of both regression coefficients are of the same sign. Thus, either both the slopes are the same sign.

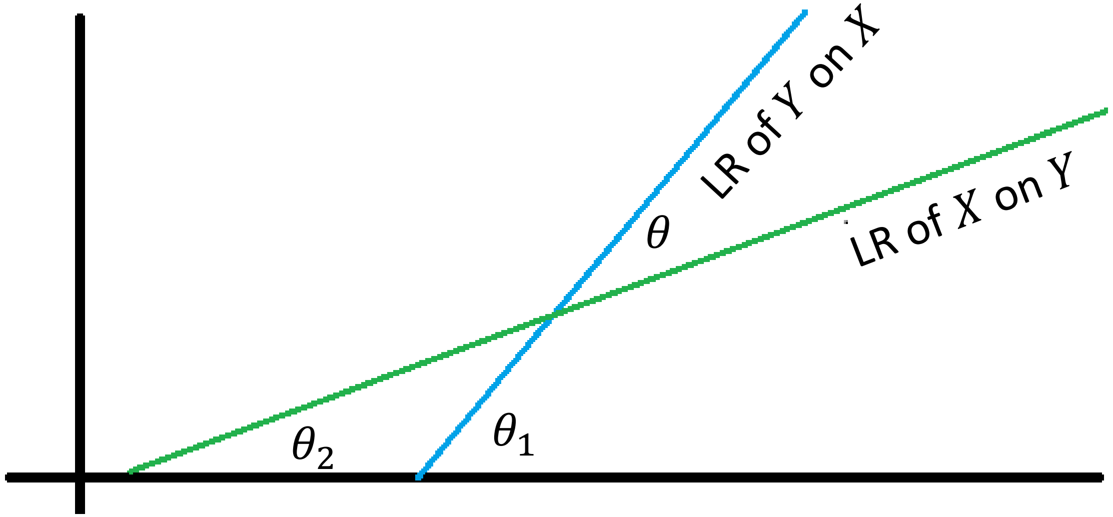
- The two regression lines are either both inclined towards the positive x-axis or both are inclines towards the negative x-axis (if they are not perpendicular to each other).

- $b_{YX} \cdot b_{XY} = r\frac{\sigma_Y}{\sigma_X} \cdot r\frac{\sigma_X}{\sigma_Y} = r^2$. Hence, the correlation coefficient is the geometric mean of the regression coefficients. However, mind the signs of the b_{YX}, b_{XY} and r.

- The point $(\bar{x}, \bar{y})$ is common to both the regression lines. Hence, the point of intersection of the two regression lines is $(\bar{x}, \bar{y})$.

- The two regression lines $\left(Y - \bar{y} = b_{YX}(X - \bar{x}),\ Y - \bar{y} = 1/b_{XY}(X - \bar{x})\right)$ coincide if only if $b_{YX} = 1/b_{XY}$ which is equivalent to $r^2 = 1$ and $r = \pm 1$. Thus, when the n points are on a straight line, there is just one regression line.

- **Angle of intersection of the two regression line:**



Observe that $\theta = |\theta_1 - \theta_2|$, $0 \le \theta \le \pi/2$, $\tan\theta_1 = b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ and $\tan\theta_2 = 1/b_{XY} = \frac{1}{r} \cdot \frac{\sigma_Y}{\sigma_X}$. Hence,

$$\tan\theta = \tan|\theta_1 - \theta_2| = \frac{|\tan\theta_1 - \tan\theta_2|}{1 + \tan\theta_1 \cdot \tan\theta_2}$$

$$= \frac{\left|r - \frac{1}{r}\right| \cdot \frac{\sigma_Y}{\sigma_X}}{1 + \frac{\sigma_Y^2}{\sigma_X^2}} = \frac{\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \left|r - \frac{1}{r}\right|.$$

Hence, the angle of intersection of the two regression lines is

$$\theta = \tan^{-1}\left[\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\cdot\left|r-\frac{1}{r}\right|\right].$$

- Case I: $\theta = 0$ iff $\tan\theta = 0$ (since $0 \le \theta \le \pi/2$) and hence

$$\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\cdot\left|r-\frac{1}{r}\right| = 0 \Rightarrow r-\frac{1}{r} = 0 \Rightarrow r = \pm 1.$$

  Thus, $\theta = 0$ iff X and Y are linearly related.

- $\theta = \pi/2$ iff $\tan\theta = \infty$ which is possible iff $r = 0$. Thus, $\theta = \pi/2$ iff X and Y are uncorrelated.

Example 8: *The following table gives the marks of 5 students in two tests out of 20* (X → *mark in first test, Y* → *mark in second test*). *Find the two lines of regression. Estimate the mark of a student in the second test if his mark in the first test is* 10 *and the also estimate the mark of a student in the first test if his mark in the second test is 12.*

X	7	15	12	11	9
Y	10	14	8	15	13

Ans: Referring to Example 5, one can see that

$$\bar{x} = 10.8, \bar{y} = 12, \sigma_X^2 = 7.36, \sigma_Y^2 = 6.8, Cov(X, Y) = 2.$$

Hence,

$$b_{YX} = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{2}{7.36} = 0.2717$$

$$b_{XY} = \frac{Cov(X, Y)}{\sigma_Y^2} = \frac{2}{6.8} = 0.2941$$

Now, the LR of Y on X is given by

$$Y - \bar{y} = b_{YX}(X - \bar{x})$$

$$\Rightarrow Y - 12 = 0.2717(X - 10.8)$$

$$\Rightarrow 0.2717X - Y = -9.0656$$

and the LR of X on Y is given by

$$X - \bar{x} = b_{XY}(Y - \bar{y})$$

$$\Rightarrow X - 10.8 = 0.2941(Y - 12)$$

$$\Rightarrow X - 0.2941Y = 7.2708$$

Estimated mark of a student in the second test if his mark in the first test is 10 is to be obtained from LR of Y on X. Hence, when $X = 10$, $Y = 0.2717 \times 10 + 9.0656 = 11.78$.

Estimated mark of a student in the first test if his mark in the second test is 12 is to be obtained from LR of X on Y. Hence, if $Y = 12$, then $X = 0.2941 \times 12 + 7.2708 = 10.8$.

Example 9: *The two lines of regression are given by* $3.40X - Y = 24.72$ *and* $X - 3.68Y = -33.37$. *Find the means of* X *and* Y, *the two regression coefficients, the correlation coefficient and the ratio of variances of* X *and* Y.

Ans. The two regression lines intersects at $(\bar{x}, \bar{y})$. Hence, $3.40\bar{x} - \bar{y} = 24.72$ *and* $\bar{x} - 3.68\bar{y} = -33.37$. On solving, we get $\bar{x} = 10.8$ $\bar{y} = 12$.

To identify which line is LR of Y on X and which is LR of X on Y, we have to notice that the slope of one line is the regression coefficient of Y on X and reciprocal of slope of other line is regression coefficient of X on Y. Further, the product of regression coefficients is equal to $r^2 \leq 1$. Hence, if m_1 and m_2 are slope of the two lines respectively, then

$$m_1 \times \frac{1}{m_2} \leq 1 \; or \; m_2 \times \frac{1}{m_1} \leq 1$$

(But both are not $\leq$ 1.)

If $m_1 \times \frac{1}{m_2} \leq 1$, then $m_1 = b_{YX}$ and $\frac{1}{m_2} = b_{XY}$, and if $m_2 \times \frac{1}{m_1} \leq 1$, then $m_2 = b_{YX}$ and $\frac{1}{m_1} = b_{XY}$.

The two lines are of regression are

$$3.40X - Y = 24.72, \qquad X - 3.68Y = -33.37.$$

Hence, $m_1 = 3.40$ and $m_2 = 0.272$. Observe that $\frac{m_2}{m_1} \leq 1$. Thus,

$$m_2 = 0.272 = b_{YX}, \qquad m_1 = 3.40 = \frac{1}{b_{XY}} \Rightarrow b_{XY} = 0.294.$$

$$r = \sqrt{b_{XY} \cdot b_{YX}} = 0.08 \Rightarrow r = 0.283.$$

Observe that r is positive since both b_{XY} and b_{YX} are positive. Since

$$b_{XY} = r \cdot \frac{\sigma_X}{\sigma_Y} \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{b_{XY}}{r} = \frac{0.294}{0.283} = 1.039.$$

Hence

$$\frac{\sigma_X^2}{\sigma_Y^2} = 1.08.$$