Using rank correlation technique, find which pair of judges have more or less the same taste in music.

26. If X, Y, Z are uncorrelated R.V.'s having the same variance, find the correlation coefficient between $(X + Y)$ and $(Y + Z)$.
27. If X and Y are two uncorrelated R.V.'s with zero means, prove that $U = X \cos \alpha + Y \sin \alpha$ and $V = X \sin \alpha - Y \cos \alpha$ are also uncorrelated.
28. X and Y are independent R.V.'s with means 5 and 10 and variances 4 and 9 respectively. Obtain the correlation coefficient between U and V, where $U = 3X + 4Y$ and $V = 3X - Y$.
29. If X_1, X_2, X_3 are three uncorrelated R.V.'s having variances v_1, v_2, v_3 respectively, obtain the coefficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$.
30. Show that (i) $E\{aX + bY\} = aE(X) + bE(Y)$ and (ii) $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2abC(X, Y)$, where $C(X, Y)$ is the covariance of (X, Y).
31. It two R.V.'s are uncorrelated, prove that the variance of their sum is equal to the sum of their variances.
32. If the joint density fucntion of (X, Y) is given by $f(x, y) = 2 - x - y$, $0 \le x, y \le 1$, find $E(X)$, $E(Y)$, var (X), var (Y) and r_{XY}.
33. If the two dimensional R.V. (X, Y) is uniformly distributed in $0 \le x < y \le 1$, find $E(X)$, $E(Y)$, var (X), var (Y) and r_{XY}.
34. If the two dimensional R.V. (X, Y) is uniformly distributed over R, where R is defined by $\{(x, y)/x^2 + y^2 \le 1, y \ge 0\}$, find r_{XY}.
35. If the joint pdf of (X, Y) is given by $f(x, y) = x + y$, $0 \le x, y \le 1$, find r_{XY}.
36. Let X be a R.V. with mean value = 3 and variance = 2. Find the second moment of X about the origin. Another R.V. Y is defined by $Y = -6X + 22$. Find the mean value of Y and the correlation of X and Y.

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

The regression line of Y on X is the best-fitting straight line for the observed of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, based on the assumption that x independent variable and y is the dependent variable. Hence, let the equation the regression line of Y on X be assumed as $y = ax + b$.

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

are $$\Sigma y_i = a \Sigma x_i + nb \quad (1)$$

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

Dividing equation (2) by n, we get

$$\bar{y} = a \bar{x} + b \quad (3)$$

where $\bar{x} = E(X)$ and $\bar{y} = E(Y)$. (1)–(4) gives the required equation as

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

Eliminating b between equations (2) and (3)

we get $$a = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

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

or $$a = \frac{E(XY) - E(X)\cdot E(Y)}{E(X^2) - E^2(X)} = \frac{p_{XY}}{\sigma_X^2} \quad (5)$$

Using (6) in (5), we get the equation of the regression line of Y on X as

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

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

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

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

$\Sigma x_i = a$

we can get the equation of t

$x - \bar{x} =$

or $x - \bar{x} =$

Note:

1. $\frac{p_{XY}}{\sigma_X^2}$ or $\frac{r_{XY}\,\sigma_Y}{\sigma_X}$ is c

   by b_1 or b_{YX}. $\frac{p_{XY}}{\sigma_Y^2}$

   Y and denoted by b
2. Clearly $b_1 b_2 = r^2_{XY}$

   $\therefore$ r_{XY}

   The sign of r_{XY} is t

   $\frac{\sigma_Y}{\sigma_X}$ have the sam

   Also $\frac{b_1}{b_2}$
3. When there is per
   ± 1, the two regre
4. The point of inte
   whose co-ordinat
5. When there is no
   the equations of t
   right angles.

Standard Error

Although we use the re
sponding to a specified
corresponding to an ob
this manner need not, i
Y, namely, y_i. Hence th
mate of Y. This error w
variable. The standard
error of estimate of Y

$$\Sigma x_i = a\,\Sigma y_i + nb \text{ and } \Sigma x_i y_i = a\,\Sigma y_i^2 + b\,\Sigma y_i,$$

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

$$x - \bar{x} = \frac{p_{XY}}{\sigma_Y^2}(y - \bar{y}) \tag{9}$$

or

$$x - \bar{x} = \frac{r_{XY}\,\sigma_X}{\sigma_Y}(y - \bar{y}) \tag{10}$$

Note:

1. $\frac{p_{XY}}{\sigma_X^2}$ or $\frac{r_{XY}\,\sigma_Y}{\sigma_X}$ is called *the regression coefficient of Y on X* and denoted by b_1 or b_{YX}. $\frac{p_{XY}}{\sigma_Y^2}$ or $\frac{r_{XY}\,\sigma_X}{\sigma_Y}$ is called *the regression coefficient of X on Y* and denoted by b_2 or b_{XY}.
2. Clearly $b_1\,b_2 = r_{XY}^2$, i.e., r_{XY} is the geometric mean of b_1 and b_2.

   $\therefore \quad r_{XY} = \pm\sqrt{b_1\,b_2}$

   The sign of r_{XY} is the same as that of b_1 or b_2, as $b_1 = r_{xy}\frac{\sigma_Y}{\sigma_X}$ and $b_2 = r_{XY}\frac{\sigma_Y}{\sigma_X}$ have the same sign as r_{XY} (Θ σ_X and σ_Y are positive).

   b_1 Y on X
   b_2 X on Y

   Also $\quad \frac{b_1}{b_2} = \frac{\sigma_Y^2}{\sigma_X^2}$
3. When there is perfect linear correlation between X and Y, viz., when $r_{XY} = \pm 1$, the two regression lines coincide.
4. The point of intersection of the two regression lines is clearly the point whose co-ordinates are $(\bar{x}, \bar{y})$.
5. When there is no linear correlation between X and Y, viz., when $r_{XY} = 0$, the equations of the regression lines become $y = \bar{y}$ and $x = \bar{x}$, which are at right angles.

Standard Error of Estimate of Y

Although we use the regression line of Y on X to predict the value of Y corresponding to a specified value of X we may also use it to estimate the value of Y corresponding to an observed value of $X = x_i$, say. The value of Y estimated in this manner need not, in general, be equal to the corresponding observed value of Y, namely, y_i. Hence the difference between Y and Y_E is called *the error of estimate of Y*. This error will vary from one observed value to the other and a random variable. The standard deviation of this R.V. $(Y - Y_E)$ is called *the standard error of estimate of Y* and denoted by S_Y.

Now $E\{Y - Y_L\} = E\left[Y - \left\{\bar{y} + \frac{r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})\right\}\right]$

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

$= 0$

$\sigma^2_{(Y - Y_L)} = E\{(Y - Y_L)^2\} - E^2(Y - Y_L)$

$= E\left[Y - \left\{\bar{y} + \frac{r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})\right\}\right]^2$

$= E\left[(Y - \bar{y})^2 + \frac{r^2_{XY}\,\sigma^2_Y}{\sigma^2_X}(X - \bar{x})^2 - \frac{2r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})(Y$

(i.e.) $S^2_Y = \sigma^2_Y + \frac{r^2_{XY}\,\sigma^2_Y}{\sigma^2_X}\sigma^2_X - \frac{2r_{XY}\,\sigma_Y}{\sigma_X}\text{Cov}(X, Y)$

$= \sigma^2_Y + r^2_{XY} \cdot \sigma^2_Y - 2r_{XY}\,\sigma^2_Y$

[Θ Cov $(X, Y) = r_{XY}$

$= (1 - r^2_{XY})\,\sigma^2_Y$ or $S_Y = \sqrt{1 - r^2_{XY}}\,\sigma_Y$

Similarly, the standard error of estimate of X, denoted by S_X is given by

$S^2_X = (1 - r^2_{XY})\,\sigma^2_X$ or $S_X = \sqrt{1 - r^2_{XY}}\,\sigma_X$

[**Note:** We may use (1) or (2) to prove that $|r_{XY}| \le 1$.

From (1), $S_Y = \sqrt{1 - r^2_{XY}}\,\sigma_Y$

Since S_Y and σ_Y are positive, $1 - r^2_{XY} \ge 0$

$\therefore$ $r^2_{XY} \le 1$

i.e., $|r_{XY}| \le 1$ or $-1 \le r_{XY} \le 1$]

Worked Example 4(C)

Example 1

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

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

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

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

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

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

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

The regression line of Y on X is

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

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

i.e., $$3.7x - 4y + 29.2 = 0$$

The regression line of X on Y is

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

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

i.e., $$4x - 3.7y + 24.7 = 0$$

Example 2

Obtain the equations of the regression lines from the following data, using the method of least squares. Hence find the coefficient of correlation between X and Y. Also estimate the value of (i) Y, when $X = 38$ and (ii) X, when $Y = 18$.