Study Material
on
Bivariate samples

For UG Mathematics Students of Semester VI
Course Code-MA306, Year 2019 - 2020

1 Bivariate population:

Let X and Y be a pair of random variables defined on the event space of a random experiment E. Any performance of the E will give an observed value of the two dimensional random variable (X, Y). The totality of all such observed values of (X, Y) obtained by repeating E under uniform conditions infinite number of times, is called a bivariate population of X and Y.

The joint distribution function $F(x, y)$ of X and Y is said to determine the distribution of the bivariate population of X and Y.

2 Bivariate sample:

If E be repeated under identical conditions a finite number of times, say n times, and if the observed values of the two dimensional random variable (X, Y) be obtained as $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, then the ordered n-tuple $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ is called a bivariate sample of size n drawn from the bivariate population of X and Y.

The sample being random, the sample values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ may be regarded as observed values of n two-dimensional random variables $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ respectively, which are mutually independent all having the same distribution function as the distribution function $F(x, y)$ of the population.

The empirical distribution of the sample is obtained by placing a probability mass $\frac{1}{n}$ at each observed point $(x_i, y_i), i = 1, 2, ..., n$. Let $(\overset{o}{X}, \overset{o}{Y})$ denote the hypothetical random variable associated with the empirical distribution.

3 Marginal distributions:

Marginal disribution of $\overset{o}{X}$ is given by

$$P(\overset{o}{X} = x_i) = \sum_{j=1}^{n} P(\overset{o}{X} = x_i, \overset{o}{Y} = y_j) = P(\overset{o}{X} = x_i, \overset{o}{Y} = y_i) = \frac{1}{n},\ i = 1, 2, ..., n$$

since $P(\overset{o}{X} = x_i, \overset{o}{Y} = y_j) = 0$ for $j \neq i$.

Similarly marginal distributions of $\overset{o}{Y}$ is given by

$$P(\overset{o}{Y} = y_i) = \frac{1}{n},\ i = 1, 2, ..., n.$$

4 Sample characteristics:

The characteristics of $(\overset{o}{X}, \overset{o}{Y})$ are then called the sample characteristics by definition, the most important of which are the following.

Means: $\bar{x}$=E$(\overset{o}{X})$=$\frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y}$=E$(\overset{o}{Y})$=$\frac{1}{n}\sum_{i=1}^{n} y_i$

Variances: $S_x^2 = E\{(\overset{o}{X} - \bar{x})^2\}$=$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$; $S_y^2 = E\{(\overset{o}{Y} - \bar{y})^2\}$=$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Moments: $a_{kl} = E(\overset{o}{X}^k \overset{o}{Y}^l) = \frac{1}{n}\sum_{i=1}^{n} x_i^k y_i^l$

So $a_{k0} = a_{xk}, a_{0l} = a_{yl}, a_{00} = 1, a_{10} = \bar{x}, a_{01} = \bar{y}$.

Central moments: $m_{kl} = E\{(\overset{o}{X} - \bar{x})^k (\overset{o}{Y} - \bar{y})^l\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k (y_i - \bar{y})^l$

So $m_{k0} = m_{xk}, m_{0l} = m_{yl}, m_{10} = m_{01} = 0, m_{20} = S_x^2, m_{02} = S_y^2$.

Covariance: $m_{11} = E\{(\overset{o}{X} - \bar{x})(\overset{o}{Y} - \bar{y})\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Correlation coefficient: $r = \frac{m_{11}}{S_x S_y}$

Also we have the formulas: $S_x^2 = a_{x2} - \bar{x}^2, S_y^2 = a_{y2} - \bar{y}^2, m_{11} = a_{11} - \bar{x}\bar{y}$.

5 Scatter diagram:

Let $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ be a given bivariate sample drawn from a given bivariate population. For each ordered pair (x_i, y_i)[i=1,2,...,n] we get a point P_i in the xy plane with abscissa x_i and ordinate y_i and it is represented by a dot and thus by plotting $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ we get a diagram called the scatter diagram.

6 Properties of Sample Correlation Coefficient r:

Property 1: If r be the sample correlation coefficient of a bivariate sample $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ then $-1 \leq r \leq 1$.

Proof: Sample correlation coefficient r is given by

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

where

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

Since $x_1, x_2, ..., x_n$; $y_1, y_2, ..., y_n$ are real numbers we have

$$\left(\frac{x_i - \bar{x}}{S_x} \pm \frac{y_i - \bar{y}}{S_y}\right)^2 \geq 0 \text{ for } i = 1, 2, ..., n.$$

So we have

$$\left(\frac{x_i - \bar{x}}{S_x}\right)^2 + \left(\frac{y_i - \bar{y}}{S_y}\right)^2 \pm 2\frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} \geq 0 \text{ for } i = 1, 2, ..., n.$$

Hence we get

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)^2 + \sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{S_y}\right)^2 \pm 2\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} \geq 0,$$

or,

$$\frac{1}{S_x^2} n S_x^2 + \frac{1}{S_y^2} n S_y^2 \pm 2nr \geq 0,$$

or,

$$2 \pm 2r \geq 0,$$

or,

$$1 \pm r \geq 0.$$

Combinining both the cases of above inequation we get $-1 \leq r \leq 1$.

Property 2: If r be the correlation coeffiecient of a bivariate sample $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ and if the linear transormations $x_i' = ax_i + b$, $y_i' = cy_i + d$, (i=1,2,...,n) be used then $r' = \frac{ac}{|a||c|}r$, where a,b,c,d are constants and $a \neq 0, c \neq 0$ and r' is the correlation coeffcient of the corresponding bivariate sample $((x_1', y_1'), (x_2, y_2), ..., (x_n, y_n))$.

Proof: We have

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

where

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

Now if $\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i',\ \bar{y}' = \frac{1}{n}\sum_{i=1}^{n} y_i'$, then S_x', S_y' are given by

$$S_x'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i' - \bar{x}')^2,\ S_y'^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i' - \bar{y}')^2.$$

Then

$$\begin{aligned}
r' &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i' - \bar{x}')(y_i' - \bar{y}')}{S_x' S_y'} \text{ where } S_x', S_y' > 0 \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\bar{x} - b)(cy_i + d - c\bar{y} - d)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\bar{x} - b)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(cy_i + d - c\bar{y} - d)^2}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}[a(x_i - \bar{x})c(y_i - \bar{y})]}{\sqrt{\frac{a^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{b^2}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{ac}{|a||c|}\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{ac}{|a||c|}\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} \\
&= \frac{ac}{|a||c|}r
\end{aligned}$$

7 Regression line of the sample:

For fitting a curve of the type

$$y = g(x; c_0, c_1, ...)$$

where $c_0, c_1, ...$ are unknown parameters, to the empirical distribution of the sample, i.e., to the observed set of points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ by the principle of least squares, we have to minimise

$$S = E[\{\overset{o}{Y} - g(\overset{o}{X}; c_0, c_1, ...)\}^2].$$

Here we are concerned with the family of straight lines

$$y = c_0 + c_1 x$$

so that

$$S = E[\{\overset{o}{Y} - (c_0 + c_1\overset{o}{X})\}^2].$$

The normal equations are

$$\frac{\partial S}{\partial c_0} = 0, i.e., 2E[(\overset{o}{Y} - c_0 - c_1\overset{o}{X}) \times (-1)] = 0, i.e., E[(\overset{o}{Y} - c_0 - c_1\overset{o}{X})] = 0 \tag{1}$$

$$\frac{\partial S}{\partial c_1} = 0, i.e., 2E[(\overset{o}{Y} - c_0 - c_1\overset{o}{X}) \times (-\overset{o}{X})] = 0, i.e., E[\overset{o}{X}(\overset{o}{Y} - c_0 - c_1\overset{o}{X})] = 0 \tag{2}$$

If S is minimum for $c_0 = c_0^*, c_1 = c_1^*$, on putting these values for c_0, c_1, above equations reduce to

$$E(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X}) = 0$$

or,

$$\bar{y} = c_0^* + c_1^*\bar{x}, \tag{3}$$

and

$$E\{\overset{o}{X}(\overset{o}{Y} - c_0^* - c_1^*)\} = 0$$

or,

$$a_{11} = c_0^*\bar{x} + c_1^* a_{x2}. \tag{4}$$

Now (4)-(3)$\times\bar{x}$ gives

$$c_1^*(a_{x2} - \bar{x}^2) = a_{11} - \bar{x}\bar{y}$$

i.e.,

$$c_1^* S_x^2 = r S_x S_y$$

i.e.,

$$c_1^* = r\frac{S_y}{S_x} = b_{yx}(say).$$

Hence $c_0^* = \bar{y} - c_1^*\bar{x} = \bar{y} - r\frac{S_y}{S_x}\bar{x} = \bar{y} - b_{yx}\bar{x}$.

The regression line of $\overset{o}{Y}$ on $\overset{o}{X}$ is

$$y = c_0^* + c_1^* x = \bar{y} - b_{yx} + b_{yx} x$$

i.e.,

$$y - \bar{y} = b_{yx}(x - \bar{x}), \text{ where } b_{yx} = r\frac{S_y}{S_x}. \tag{5}$$

Similarly regression line of $\overset{o}{X}$ on $\overset{o}{Y}$ is

$$x = d_0^* + d_1^* y, \text{ where } d_0^* = \bar{x} - r\frac{S_x}{S_y}\bar{y}, \ d_1^* = r\frac{S_x}{S_y};$$

i.e.,

$$x = \bar{x} - r\frac{S_x}{S_y}\bar{y} + r\frac{S_x}{S_y} y$$

or,

$$x = \bar{x} + r\frac{S_x}{S_y}(y - \bar{y})$$

or,

$$x - \bar{x} = b_{xy}(y - \bar{y}), \text{ where } b_{xy} = r\frac{S_x}{S_y}. \tag{6}$$

$b_{yx} = r\frac{S_y}{S_x}$ and $b_{xy} = r\frac{S_x}{S_y}$ are the regression coefficient of the sample.

Here $\overset{o}{Y} = c_0 + c_1\overset{o}{X}$. Let $\overset{o}{U}_y$ be the best representation of $\overset{o}{Y}$, then

$$\overset{o}{U}_y = c_0^* + c_1^*\overset{o}{X} = \bar{y} + b_{yx}(\overset{o}{X} - \bar{x}).$$

Hence $E(\overset{o}{U}_y) = \bar{y}, \ \sigma(\overset{o}{U}_y) = |r|S_y$, and $\rho(\overset{o}{U}_y, \overset{o}{Y}) = |r| \geq 0$.

Thus we may say that the correlation coeffiecient between $\overset{o}{Y}$ and its best representation is a measure of goodness of fit of the regression lines to the observed sample points.

Measure of goodness of fit of the regression lines (5) and (6)

For regression line of $\overset{o}{Y}$ on $\overset{o}{X}$, we have $S_{min} = E[\{\overset{o}{Y} - (c_0^* + c_1^*\overset{o}{X})\}^2]$, which is a mesure of dispersion of the regression line (5) to the observed samples points.

Now, we have

$$\begin{aligned}(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X})^2 &= \left\{\overset{o}{Y} - \bar{y} - r\tfrac{S_y}{S_x}(\overset{o}{X} - \bar{x})\right\}^2 \\ &= (\overset{o}{Y} - \bar{y})^2 + r^2\tfrac{S_y^2}{S_x^2}(\overset{o}{X} - \bar{x}) - 2r\tfrac{S_y}{S_x}(\overset{o}{Y} - \bar{y})(\overset{o}{X} - \bar{x}).\end{aligned}$$

So,

$$\begin{aligned} S_{min} &= E\{(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X})^2\} \\ &= E\{(\overset{o}{Y} - \bar{y})^2 + r^2\frac{S_y^2}{S_x^2}(\overset{o}{X} - \bar{x})^2 - 2r\frac{S_y}{S_x}(\overset{o}{Y} - \bar{y})(\overset{o}{X} - \bar{x})\} \\ &= E\{(\overset{o}{Y} - \bar{y})^2\} + r^2\frac{S_y^2}{S_x^2}E\{(\overset{o}{X} - \bar{x})^2\} - 2r\frac{S_y}{S_x}E\{(\overset{o}{Y} - \bar{y})(\overset{o}{X} - \bar{x})\} \\ &= S_y^2 + r^2\frac{S_y^2}{S_x^2}S_x^2 - 2r\frac{S_y}{S_x}rS_xS_y \\ &= S_y^2 + r^2S_y^2 - 2r^2S_y^2 \\ &= S_y^2 - r^2S_y^2 \\ &= S_y^2(1 - r^2) \end{aligned} \tag{7}$$

Similarly for regression line of $\overset{o}{X}$ on $\overset{o}{Y}$ we will obtain

$$S_{min} = E\{(\overset{o}{X} - d_0^* - d_1^*\overset{o}{Y})^2\} = S_x^2(1 - r^2). \tag{8}$$

From equations (7) and (8) we find that for given values of S_x and S_y, S_{min} for both case decreases as $|r|$ increases. Hence goodness of fit of two regression lines to the bivariate sample increases as $|r|$ increases.

So $|r|$ can be used as a direct measure of goodness of fit of the regression lines to a bivariate sample $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where the sample correlation coefficient r is given by

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

8 Parabolic curve fitting:

For fitting a curve of kth degree

$$y = c_0 + c_1x + c_2x^2 + ... + c_kx^k$$

where $c_0, c_1, c_2, ..., c_k$ are unknown parameters, to the empirical sample distribution, i.e., to the observed set of points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ by the principle of least squares, we have to minimise

$$S = E[\{\overset{o}{Y} - (c_0 + c_1\overset{o}{X} + c_2\overset{o}{X^2} + ... + c_k\overset{o}{X^k})\}^2].$$

The normal equations are

$$\frac{\partial S}{\partial c_0} = 0,\ \frac{\partial S}{\partial c_1} = 0,\ \frac{\partial S}{\partial c_2} = 0,\ ..., \frac{\partial S}{\partial c_k} = 0.$$

Now,

$$\begin{aligned}
\frac{\partial S}{\partial c_0} &= 0 \\
\Rightarrow 2E\{(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k}) \times (-1)\} &= 0 \qquad (1.1) \\
\Rightarrow E\{(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k})\} &= 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial S}{\partial c_1} &= 0 \\
\Rightarrow 2E\{(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k}) \times (-\overset{o}{X})\} &= 0 \qquad (1.2) \\
\Rightarrow E\{\overset{o}{X}(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k})\} &= 0
\end{aligned}$$

$$...$$

$$\begin{aligned}
\frac{\partial S}{\partial c_k} &= 0 \\
\Rightarrow 2E\{(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k}) \times (-\overset{o}{X^k})\} &= 0 \qquad (1.k) \\
\Rightarrow E\{\overset{o}{X^k}(\overset{o}{Y} - c_0 - c_1\overset{o}{X} - c_2\overset{o}{X^2} - ... - c_k\overset{o}{X^k})\} &= 0
\end{aligned}$$

If S is minimum for $c_0 = c_0^*, c_1 = c_1^*, c_2 = c_2^*, ..., c_k = c_k^*$ on putting these values for $c_0, c_1, c_2, ..., c_k$, above equations reduce to

$$\begin{aligned}
&E\{(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X} - c_2^*\overset{o}{X^2} - ... - c_k^*\overset{o}{X^k})\} &= 0 \\
&E\{\overset{o}{X}(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X} - c_2^*\overset{o}{X^2} - ... - c_k^*\overset{o}{X^k})\} &= 0 \qquad (2) \\
&... \\
&E\{\overset{o}{X^k}(\overset{o}{Y} - c_0^* - c_1^*\overset{o}{X} - c_2^*\overset{o}{X^2} - ... - c_k^*\overset{o}{X^k})\} &= 0
\end{aligned}$$

In terms of sample moments equation (2) can be written as follows

$$\begin{aligned}
c_0^*a_{00} + c_1^*a_{10} + c_2^*a_{20} + ... + c_k^*a_{k0} &= a_{01} \\
c_0^*a_{10} + c_1^*a_{20} + c_2^*a_{30} + ... + c_k^*a_{k+1,0} &= a_{11} \qquad (3) \\
... & \\
c_0^*a_{k0} + c_1^*a_{k+1,0} + c_2^*a_{k+2,0} + ... + c_k^*a_{2k,0} &= a_{k1}
\end{aligned}$$

Multiplying all the equations of (3) by n we obtain

$$\begin{aligned}
nc_0^* + c_1^*\sum_{i=1}^{n} x_i + c_2^*\sum_{i=1}^{n} x_i^2 + ... + c_k^*\sum_{i=1}^{n} x_i^k &= \sum_{i=1}^{n} y_i \\
c_0^*\sum_{i=1}^{n} x_i + c_1^*\sum_{i=1}^{n} x_i^2 + c_2^*\sum_{i=1}^{n} x_i^3 + ... + c_k^*\sum_{i=1}^{n} x_i^{k+1} &= \sum_{i=1}^{n} x_iy_i \qquad (4) \\
... & \\
c_0^*\sum_{i=1}^{n} x_i^k + c_1^*\sum_{i=1}^{n} x_i^{k+2} + c_2^*\sum_{i=1}^{n} x_i^{k+3} + ... + c_k^*\sum_{i=1}^{n} x_i^{2k} &= \sum_{i=1}^{n} x_i^ky_i
\end{aligned}$$

These equations determine the least square estimates $c_0^*, c_1^*, c_2^*, ..., c_k^*$ of the parameters. The best fitting parabola of degree k is then

$$y = c_0^* + c_1^* x + c_2^* x^2 + ... + c_k^* x^k.$$

The residual of $\overset{o}{Y}$, $\overset{o}{V}_y$ is given by

$$\overset{o}{V}_y = \overset{o}{Y} - c_0^* - c_1^* \overset{o}{X} - c_2^* \overset{o}{X^2} - ... - c_k^* \overset{o}{X^k} \tag{5}$$

and the residual of the sample, i.e., the values v_{yi} which $\overset{o}{V}_y$ takes are given by

$$v_{yi} = y_i - c_0^* - c_1^* x_i - c_2^* x_i^2 - ... - c_k^* x_i^k. \tag{6}$$

1. The first equation of (4) states that $\sum v_{yi} = 0$, i.e., the sum of the residuals is zero.
2. The normal equations may also be written as

$$\sum v_{yi} = 0, \ \sum x_i v_{yi} = 0, \ ..., \ \sum x_i^k v_{yi} = 0. \tag{7}$$

Measure of goodness of fit of the least square regression parabola to the bivariate sample

If S^* be the minimum value of S then

$$S^* = \frac{1}{n} \sum_{i=1}^{n} [y_i - (c_0^* + c_1^* x_i + c_2^* x_i^2 + ... + c_k^* x_i^k)]^2.$$

Now we consider the expression S_1 given by

$$S_1 = \frac{1}{n} \sum_{i=1}^{n} [t y_i - c_0 - c_1 x_i - c_2 x_i^2 - ... - c_k x_i^k]^2.$$

We see that S_1 is a continuously differentiable function of $t, c_0, c_1, ..., c_k$ and it is also homogeneous function of degree 2 in $t, c_0, c_1, ..., c_k$. Then by Euler's theorem we get

$$t\frac{\partial S_1}{\partial t} + c_0 \frac{\partial S_1}{\partial c_0} + c_1 \frac{\partial S_1}{\partial c_1} + c_k \frac{\partial S_1}{\partial c_k} = 2S_1.$$

Now we find that S^*=value of S_1 for $t = 1, c_0 = c_0^*, c_1 = c_1^*, ..., c_k = c_k^*$ and consequently we get

$$\left[t\frac{\partial S_1}{\partial t} + c_0 \frac{\partial S_1}{\partial c_0} + c_1 \frac{\partial S_1}{\partial c_1} + c_k \frac{\partial S_1}{\partial c_k} \right]_{(t=1, c_0=c_0^*, c_1=c_1^*, ..., c_k=c_k^*)} = 2S^* \tag{8}$$

Again we see that for $t = 1, c_0 = c_0^*, c_1 = c_1^*, ..., c_k = c_k^*$, $\frac{\partial S_1}{\partial c_0}$, $\frac{\partial S_1}{\partial c_1}$,..., $\frac{\partial S_1}{\partial c_k}$ all vanish due to the normal equations $\frac{\partial S}{\partial c_0} = 0$, $\frac{\partial S}{\partial c_1} = 0$,..., $\frac{\partial S}{\partial c_k} = 0$ which are satisfied by $c_0 = c_0^*, c_1 = c_1^*, ..., c_k = c_k^*$.

Then by equation (8) we get

$$\left[\frac{\partial S_1}{\partial t}\right]_{(t=1, c_0=c_0^*, c_1=c_1^*, ..., c_k=c_k^*)} = 2S^* \tag{9}$$

Now

$$\frac{\partial S_1}{\partial t} = \frac{2}{n}\sum_{i=1}^{n}[ty_i - c_0 - c_1x_i - c_2x_i^2 - ... - c_kx_i^k]y_i.$$

So

$$\left[\frac{\partial S_1}{\partial t}\right]_{(t=1, c_0=c_0^*, c_1=c_1^*, ..., c_k=c_k^*)} = \frac{2}{n}\sum_{i=1}^{n}[y_i - c_0^* - c_1^*x_i - c_2^*x_i^2 - ... - c_k^*x_i^k]y_i.$$

Hence using (9) we get

$$S^* = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - c_0^*\frac{1}{n}\sum_{i=1}^{n}y_i - c_1^*\frac{1}{n}\sum_{i=1}^{n}x_iy_i - c_2^*\frac{1}{n}\sum_{i=1}^{n}x_i^2y_i - ... - c_k^*\frac{1}{n}\sum_{i=1}^{n}x_i^ky_i. \tag{10}$$

Now we observe that goodness of fit of the best fitting parabola $y = c_0^* + c_1^*x + c_2^*x^2 + ... + c_k^*x^k$ to the bivariate sample becomes high if S^* is low and vice-versa. So S^* gives an inverse measure of goodness of fit where by eqution (10)

$$nS^* = \sum_{i=1}^{n}y_i^2 - c_0^*\sum_{i=1}^{n}y_i - c_1^*\sum_{i=1}^{n}x_iy_i - c_2^*\sum_{i=1}^{n}x_i^2y_i - ... - c_k^*\sum_{i=1}^{n}x_i^ky_i.$$

Now let $\overset{o}{U}_y = c_0^* + c_1^*\overset{o}{X} + c_2^*\overset{o}{X}^2 + ... + c_k^*\overset{o}{X}^k$. If $R_y = \rho(\overset{o}{U}_y, \overset{o}{Y})$, then we have

$$S^* = S_y^2(1 - R_y^2)$$

where $0 \leq R_y \leq 1$ and S_y is the standard deviation of $\overset{o}{Y}$. Then we have

$$R_y^2 = 1 - \frac{S^*}{S_y^2}, \text{ where } S_y > 0. \tag{11}$$

Now from equations (10) and (11) we get

$$R_y^2 = 1 - \frac{1}{nS_y^2}\left[\sum_{i=1}^{n}y_i^2 - c_0^*\sum_{i=1}^{n}y_i - c_1^*\sum_{i=1}^{n}x_iy_i - c_2^*\sum_{i=1}^{n}x_i^2y_i - ... - c_k^*\sum_{i=1}^{n}x_i^ky_i\right]. \tag{12}$$

Now we see that S^* decreases as $R_y(0 \leq R_y \leq 1)$ increases. So R_y can be taken as a direct measure of goodness of fit of the least square regression parabola to the given bivariate sample where R_y is given by equation (12).

9 Problems:

Problem 1:

A bivariate sample of size 11 gave the results $\bar{x} = 7$, $S_x = 2$, $\bar{y} = 9$, $S_y = 4$ and $r = 0.5$. It was later found that one pair of sample values $(x = 7, y = 9)$ was inaccurate and was rejected. How would the original value of r be affected by this rejection?

Solution: Let $((x_1, y_1), (x_2, y_2), ..., (x_{11}, y_{11}))$ be the given sample. The original value of r is 0.5. Then we have

$$0.5 = \frac{\frac{1}{11}\sum_{i=1}^{11} x_i y_i - \bar{x}\bar{y}}{S_x S_y}$$

or,

$$0.5 = \frac{\frac{1}{11}\sum_{i=1}^{11} x_i y_i - 7 \times 9}{2 \times 4}$$

or,

$$\sum_{i=1}^{11} x_i y_i = 737.$$

Now $\bar{x} = 7 \Rightarrow \sum_{i=1}^{11} x_i = 77$; $\bar{y} = 9 \Rightarrow \sum_{i=1}^{11} y_i = 99$; $S_x = 2 \Rightarrow \frac{1}{11}\sum_{i=1}^{11} x_i^2 - \bar{x}^2 = 4 \Rightarrow \sum_{i=1}^{11} x_i^2 = 583$; $S_y = 4 \Rightarrow \frac{1}{11}\sum_{i=1}^{11} y_i^2 - \bar{y}^2 = 4 \Rightarrow \sum_{i=1}^{11} y_i^2 = 1067$.

Let $((x_1', y_1'), (x_2', y_2'), ..., (x_{10}', y_{10}'))$ be the sample after regecting the pair $(7, 9)$ and $\bar{x}', \bar{y}', S_x'^2, S_y'^2$ be the corresponding values of $\bar{x}, \bar{y}, S_x^2, S_y^2$. Then we have,

$$\bar{x}' = \frac{1}{10}\sum_{i=1}^{10} x_i' = \frac{77 - 7}{10} = 7; \bar{y}' = \frac{1}{10}\sum_{i=1}^{10} y_i' = \frac{99 - 9}{10} = 9;$$

$$S_x' = \sqrt{\frac{1}{10}x_i'^2 - \bar{x}'^2} = \sqrt{\frac{583 - 49}{10} - 49} = \sqrt{4.4}; S_y' = \sqrt{\frac{1}{10}y_i'^2 - \bar{y}'^2} = \sqrt{\frac{1067 - 81}{10} - 81} = \sqrt{17.6}.$$

The correlation coefficient r' after rejecting the pair is given by

$$r' = \frac{\frac{1}{10}\sum_{i=1}^{10} x_i' y_i' - \bar{x}'\bar{y}'}{S_x' S_y'} = \frac{\frac{737-63}{10} - 7 \times 9}{\sqrt{4.4}\sqrt{17.6}} = \frac{4.4}{\sqrt{4.4}\sqrt{17.6}} = \frac{\sqrt{4.4}}{\sqrt{17.6}} = \sqrt{\frac{1}{4}} = 0.5.$$

So the original value of r will not be affected by the rejection of the pair $(7, 9)$.

Problem 2:

The regression lines for a bivariate sample are gievn by $x + 2y - 5 = 0,\ 2x + 3y - 8 = 0$ and $S_x^2 = 12$. Calculate the values of $\bar{x},\ \bar{y},\ S_y$ and r

Solution: Two regression lines intersect at the point $(\bar{x}, \bar{y})$. The point of intersection of the given lines are (1,2). So $\bar{x} = 1, \bar{y} = 2$.

Following two cases may happen

1. $x + 2y - 5 = 0$ is the regression line of $\overset{o}{Y}$ on $\overset{o}{X}$ and $2x + 3y - 8 = 0$ is the regession line $\overset{o}{X}$ on $\overset{o}{Y}$.
2. $x + 2y - 5 = 0$ is the regression line of $\overset{o}{X}$ on $\overset{o}{Y}$ and $2x + 3y - 8 = 0$ is the regession line $\overset{o}{Y}$ on $\overset{o}{X}$.

Case 1: We have $r\frac{S_y}{S_x} = -\frac{1}{2}$ and $r\frac{S_x}{S_y} = -\frac{3}{2}$ from which we get $r^2 = \frac{3}{4} \Rightarrow r = \pm\frac{\sqrt{3}}{2}$. Since $S_x > 0$, $S_y > 0$ and $r\frac{S_y}{S_x} = -\frac{1}{2}$, so $r < 0$ and hence $r = -\frac{\sqrt{3}}{2}$.
Case 2: $r\frac{S_x}{S_y} = -2$ and $r\frac{S_y}{S_x} = -\frac{2}{3}$ from which we get $r^2 = \frac{4}{3} > 1$, which is impossible since $-1 \leq r \leq 1$.

So $r = -\frac{\sqrt{3}}{2}$. Now $r\frac{S_y}{S_x} = -\frac{1}{2}$ gives $-\frac{\sqrt{3}}{2}\frac{S_y}{\sqrt{12}} = -\frac{1}{2} \Rightarrow S_y = 2$.

Problem 3:

Two random variable X and Y are connected by the relation $3X+4Y+5 = 0$. A sample $(x_i, y_i), i = 1, 2, ..., n$ is taken from the bivariate population of (X, Y); obtain the correlation coefficient of the sample.

Homewrok

Problem 4:

Calculate the correlation coefficient and determine the regression lines of Y on X and X on Y for the sample

X	8	10	5	8	9
Y	1	3	1	2	3

Homewrok

Problem 5:

Find the most likely price in Bombay corresponding to the price Rs. 70 in Calcutta from the following data obtained from 25 observations on the price of a commodity in each of two cities.
Average price in Calcutta=Rs.65
Average price in Bombay= Rs.67
Standard deviation of price in Calcutta=Rs.2.5
Standars deviation of price in Bombay=Rs.3.5
Correlation coefficient between the prices of the commodity in two cities is 0.8.

Homewrok

Acknowledgement

This study material has been prepared with the help of the following books.

References

1. S.K. De, S. Sen, Mathematical Statistics, U. N. Dhur & Sons Private Ltd.
2. Amritava Gupta, Groundwork of Mathematical Probability and Statitics, Academic Publishers.