Sri SAI RAM ENGINEERING COLLEGE
INSTITUTE OF TECHNOLOGY
West Tambaram, Chennai - 44

Sairam INSTITUTIONS

SAIRAM DIGITAL RESOURCES

YEAR II SEM IV

MA8391

PROBABILITY AND STATISTICS

UNIT II

TWO DIMENSIONAL RANDOM VARIABLES

2.4 LINEAR REGRESSION

SCIENCE & HUMANITIES

NBA NAAC (A+) nirf

Regression Lines

Regression is the study of the relationship between the variables X and Y.

The Regression line Y on X is $Y - \bar{Y} = \frac{r\sigma_Y}{\sigma_X}(X - \bar{X})$ --.(1)

Where $\bar{X}$ is mean of X, $\bar{Y}$ is mean of Y, r is the correlation coefficient between X and Y.

$$\sigma_X = standard\ deviation\ of\ X$$
$$\sigma_Y = standard\ deviation\ of\ Y$$

The Regression line X on Y is $X - \bar{X} = \frac{r\sigma_X}{\sigma_Y}(Y - \bar{Y})$ -----(2)

The Regression line X on Y is $X - \bar{X} = \frac{r\sigma_X}{\sigma_Y}(Y - \bar{Y})$ --(2)

The Coefficient of X in equation (1), $\frac{r\sigma_Y}{\sigma_X}$ is called regression coefficient of Y on X and it is dentoted by b_{YX}.

$b_{YX} = \frac{r\sigma_Y}{\sigma_X}$.

Similarly, the Coefficient of Y in equation (2) , $\frac{r\sigma_X}{\sigma_Y}$ is called regression coefficient of X on Y and it is dentoted by b_{XY}. $b_{XY} = \frac{r\sigma_X}{\sigma_Y}$.

Properties of Regression coefficients

1. $b_{XY}\ b_{YX} = \frac{r\sigma_Y}{\sigma_X}\frac{r\sigma_X}{\sigma_Y} = r^2.$

$$\therefore r = \pm\sqrt{b_{XY}\ b_{YX}}$$

If $b_{XY}\ and\ \ b_{YX}$ are positive, then $r = +\sqrt{b_{XY}\ b_{YX}}$

If $b_{XY}\ and\ \ b_{YX}$ are negative, then $r = -\sqrt{b_{XY}\ b_{YX}}$

Therefore, r is the geometric mean of the regression coefficients.

2. If $b_{XY}\ and\ \ b_{YX}$ have opposite sign, r does not exist.

1. The regression coefficients are unaffected by change of origin, but are affected by change of scale.
2. If $r = 1\ or\ -1$, the regression lines coincide.
3. If r=0, the regression lines are perpendicular.
4. If θ is the angle between the lines, then

$$\tan\theta = \frac{1-r^2}{r}\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}$$

Distinguish between correlation and regression Analysis

Solution:

1. Correlation means relationship between two variables and Regression is a Mathematical

Measure of expressing the average relationship between the two variables.

2. Correlation need not imply cause and effect relationship between the variables. Regression

analysis clearly indicates the cause and effect relationship between Variables.

3. Correlation coefficient is symmetric i.e. $r_{xy} = r_{ix}$ where regression coefficient is not symmetric

4. Correlation coefficient is the measure of the direction and degree of linear relationship between two variables. In regression using the relationship between two variables we can predict the dependent variable value for any given independent variable value.

3. Find the regression lines:

X	6	8	10	18	20	23
Y	40	36	20	14	10	2

Solution:

X	Y	X^2	Y^2	XY
6	40	36	1600	240
8	36	64	1296	288
10	20	100	400	200
18	14	324	196	252
20	10	400	100	200
23	2	529	4	46
X= 85	Y= 122	X^2=1453	Y^2 =3596	XY=1226

$$\sum X = 85 \quad \sum y = 122 \quad \sum X^2 = 1453 \quad \sum Y^2 = 3596 \quad \sum XY = 1226$$

$$= \frac{\sum x}{n} = \frac{85}{6} = 14.17 \, , \overline{Y} = \frac{\sum y}{n} = \frac{122}{6} = 20.33$$

$$\sigma_X = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{1453}{6} - \left(\frac{85}{6}\right)^2} = 6.44$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{3596}{6} - \left(\frac{122}{6}\right)^2} = 13.63$$

$$\sum X = 85 \quad \sum y = 122 \quad \sum X^2 = 1453 \quad \sum Y^2 = 3596 \quad \sum XY = 1226$$

$$= \frac{\sum x}{n} = \frac{85}{6} = 14.17 \, , \overline{Y} = \frac{\sum y}{n} = \frac{122}{6} = 20.33$$

$$\sigma_X = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\frac{1453}{6} - \left(\frac{85}{6}\right)^2} = 6.44$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{3596}{6} - \left(\frac{122}{6}\right)^2} = 13.63$$

$$r = \frac{\frac{\sum xy}{n} - \overline{xy}}{\sigma_X \sigma_v} = \frac{\frac{1226}{6} - .(14.17).(20.33)}{(644)(1363)} = -0.95$$

$$b_{\eta)} = r\frac{\sigma_x}{\sigma_y} = -0.95 \times \frac{6.44}{13.63} = -0.45$$

$$b_{vx} = r\frac{\sigma_y}{\sigma_x} = -0.95 \times \frac{13.\cdot 63}{644} = -2.01$$

The regression line X on Y is

$$x - \overline{x} = b_{xy}(y - \overline{y}) \Rightarrow x - 14.17 = -0.45(y - \overline{y})$$

$$\Rightarrow x = -0.45y + 23.32$$

The regression line y on X is

$$y - \overline{y} = b_{yx}(x - \overline{x}) \Rightarrow y - 20.33 = -2.01(x - 14.17)$$

$$\Rightarrow y = -2.01x + 48.81$$

3. Using the given information given below compute $\bar{x}, \bar{y}$ and r. Also compute σ_y when $\sigma_x = 2$,

$$2x + 3y = 8 \text{ and } 4x + y = 10.$$

Solution:

When the regression equation is Known the arithmetic means are computed by solving the equation.

$2x + 3y = 8$---------$-(1)$

$4x + y = 10$---------$-(2)$

$(1) \times 2 \Rightarrow 4x + 6y = 16$----$-(3)$

$(2) - (3) \Rightarrow -5y = -6$

$$\Rightarrow y = \frac{6}{5}$$

Equation (1) $\Rightarrow 2x + 3\left(\frac{6}{5}\right) = 8 \Rightarrow 2x = 8 - \frac{18}{5} \Rightarrow x = \frac{11}{5}$

i.e. $\overline{x} = \frac{11}{5} \& \overline{y} = \frac{6}{5}$

To find r , Let $2x + 3y = 8$ be the regression equation of X on Y.

$$2x = 8 - 3y \Rightarrow x = 4 - \frac{3}{2}y$$

$\Rightarrow b_{\mathrm{w}}$ =Coefficient of y in the equation of X on y $= -\frac{3}{2}$

Let $4x + y = 10$ be the regression equation of Y on X

$$\Rightarrow y = 10 - 4x$$

$\Rightarrow b_{yx}$ =coefficient of X in the equation of Y on $X = -4$.

$$r = \pm\sqrt{b_{xy}b_{yx}}$$

$$= -\sqrt{\left(-\frac{3}{2}\right)}(-4) \ (\because b_{xy} \text{ \& } b_{yx} \text{ are negative})$$

$$= -2.45$$

Since r is not in the range of $(-1 \leq r \leq 1)$ the assumption is wrong.

Now let equation (1) be the equation of Y on X

$$\Rightarrow y = \frac{8}{3} - \frac{2x}{3}$$

$\Rightarrow b_{yx}$ = Coefficient of X in the equation of Y on X

$$b_{yx} = -\frac{2}{3}$$

from equation (2) be the equation of X on Y

$$b_{xy} = -\frac{1}{4}$$

$$r = \pm\sqrt{b_{xy} b_{yx}} = \sqrt{-\frac{2}{3} \times -\frac{1}{4}} = 0.4081$$

To compute σ_y from equation (4) $b_{yx} = -\frac{2}{3}$

But we know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$\Rightarrow -\frac{2}{3} = 0.4081 \times \frac{\sigma_y}{2}$$

$$\Rightarrow \sigma_y = -3.26$$

4. Given $f(x,y) = xe^{-x(v+1)}$, $x \geq 0, y \geq 0$. Find the regression curve of Y on X.

Solution:

Regression curve of Y on X is $E(y/x)$

$$E(y/_x) = \int_{-\infty}^{\infty} y\, f(y/_x)dy$$

$$f(y/x) = \frac{f(x,y)}{f_X(X)}$$

Marginal density function $f_X(x) = \int_0^{\infty} f\,(x,y)dy$

$$= x\int_0^{\infty} e^{-x(y+1)}\,dy$$

$$= x\left[\frac{e^{-x(y+1)}}{-x}\right]_0^{\infty} = e^{-x}, x \geq 0$$

Conditional pdf of Y on X is $f(y/_x) = \frac{f(x,y)}{f_X(x)} = \frac{xe^{-x\tau-x}}{e^{-x}} = xe^{-x\tau}$

The regression curve of y on X is given by

$$E(y/_x) = \int_0^\infty y\, xe^{-x\tau} dy$$

$$= x\left[y\frac{e^{-xy}}{-x} - \frac{e^{-\eta'}}{x^2}\right]_0^\infty$$

$E(y/x) = \frac{1}{x} \Rightarrow y = \frac{1}{x}$ and hence $xy = 1$.

5. Given $f(x,y) = \begin{cases} \frac{x+y}{3}, & 0 < x < 1, 0 < y < 2 \\ 0, & otherwise \end{cases}$ **obtain the regression of y on X and X on Y**

Solution: Regression of Y on X is $E(Y/X)$

$$E(y/_X) = \int_{-\alpha}^{\alpha} y\, f(y/_x)\, dy$$

$$f(y/_X) = \frac{f(x,y)}{f_X(x)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f\,(x,y)\, dy$$

$$= \int_0^2 \left(\frac{x+y}{3}\right) dy = \frac{1}{3}\left[xy + \frac{y^2}{2}\right]_0^2$$

$$= \frac{2(x+1)}{3}$$

$$f(y/_X) = \frac{f(x,y)}{f_X(x)} = \frac{x+y}{2(x+1)}$$

Regression of Y on $X = E(^y/_X) = \int_0^2 \frac{y(x+y)}{2(x+1)} dy$

$$= \frac{1}{2(x+1)}\left[\frac{xy^2}{2} + \frac{y^3}{3}\right]_0^2$$

$$= \frac{1}{2(x+1)}\left[2x + \frac{8}{3}\right] = \frac{3x+4}{3(x+1)}$$

$$E(X/Y) = \int_{-\infty}^{\infty} x\, f\left(/_{y}^{X}\right) dx$$

$$f(X/Y) = \frac{f(x,y)}{f_Y(y)}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f\,(x,y)dx = \int_{0}^{1} \left(\frac{x+y}{3}\right) dx = \frac{1}{3}\left[\frac{x^2}{2} + xy\right]_0^1$$

$$= \frac{1}{3}\left[\frac{1}{2} + y\right]$$

$$f(\mathrm{X/Y}) = \frac{2(x+y)}{2y+1}$$

Regression of X on $Y = E(X/\mathrm{Y}) = \int_0^1 \frac{x+y}{2y+1} dx = \frac{1}{2y+1}\left[\frac{x^2}{2} + xy\right]_0^1$

$$= \frac{\frac{1}{2} + y}{2y+1} = \frac{1}{2}$$

6. In a partially destroyed laboratory record only the lines of regressions and variance of X are available. The regression equations are $8x - 10y + 66 = 0$ and $40x - 18y = 214$ and variance of $X = 9$. Find (i) the correlation coefficient between X and Y (ii) Mean values of X and Y (iii) variance of Y.

Solution:

Given $8x - 10y = -66$ ... ... (1)

$40x - 18y = 214$ ... ... (2)

Let (1) be the regression line of y on x and (2) be the regression line of x on y.

$\therefore 10y = 8x + 66 \Rightarrow y = \frac{8x}{10} + \frac{66}{10}$. $\therefore$ the regression coefficient of y on x is $b_{yx} = \frac{8}{10} = \frac{4}{5}$

$\therefore 40x = 18y + 214 \Rightarrow x = \frac{18y}{40} + \frac{214}{40}$. $\therefore$ the regression coefficient of x on y is $b_{xy} = \frac{18}{40} = \frac{9}{20}$

$\therefore b_{yx}b_{xy} = \left(\frac{4}{5}\right)\left(\frac{9}{20}\right) = \frac{9}{25} < 1$

Let r be the correlation between x and y.

$\therefore r = \sqrt{b_{yx}b_{xy}} = \sqrt{\frac{9}{25}} = \frac{3}{5} = 0.6$ [Since both regression coefficients are positive, r is positive]

Let $(\overline{x}, \overline{y})$ be the point of intersection of the two regression lines.

Solving (1) and (2) we get x , y

$5\,x(1) \Rightarrow 40x - 50y = -330$

$$40x - 18y = 214$$

Subtracting $-32y = -544$

$\therefore y = 17$

Now, $8x - 10y = -66 \Rightarrow 8x - 10(17) = -66 \Rightarrow 8x = 170 - 66 \Rightarrow 8x = 104 \Rightarrow x = 13$

$\therefore (\overline{x}, \overline{y}) = (13\ ,\ 17)$ is the mean of X and Y.

We know, $\frac{\sigma_y^2}{\sigma_X^2} = \frac{b_1}{b_2} \Rightarrow \sigma_y^2 = \frac{b_1}{b_2}\sigma_X^2 \Rightarrow \sigma_y^2 = \frac{\frac{4}{5}}{\frac{9}{20}}(9) \Rightarrow \sigma_y^2 = 16$

$\therefore$ Variance of Y is 16