Unit-5

correlation and ~~recursion~~ Regression

correlation :- (Linear correlation or karl pearson co-efficient of correlation)

correlation co-efficient between two random variables x and y. It is denoted by $r(x,y)$ (or) r_{xy} (or) ρ_{xy}

(i) $r(x,y) = \dfrac{cov(x,y)}{\sigma_x \sigma_y}$ To find $cov(x,y)$

$cov(x,y) = N\Sigma xy - \Sigma x \Sigma y$

(ii) $r(x,y) = \dfrac{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y_i-\bar{y})^2}}$

(iii) $r = \dfrac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2-(\Sigma x)^2}\sqrt{N\Sigma y^2-(\Sigma y)^2}}$ r (or) ρ

$var(x) = N\Sigma x^2 - (\Sigma x)^2$

$var(y) = N\Sigma y^2 - (\Sigma y)^2$

$(r = -1$ to 1 (or) $-1 \le r \le 1)$

① problems based on correlation co-effient for the following data.

x:	65	66	67	67	68	69	70	72
y:	67	68	65	68	72	72	69	71

Soln:

To find correlation co-efficient

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2-(\Sigma x)^2}\sqrt{N\Sigma y^2-(\Sigma y)^2}}$$

$N = 8$

x	y	x^2	y^2	xy
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\Sigma x = 544$	$\Sigma y = 552$	$\Sigma x^2 = 37028$	$\Sigma y^2 = 38132$	$\Sigma xy = 37560$

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{8(37560) - (544)(552)}{\sqrt{8(37028) - (544)^2}\sqrt{8(38132) - (552)^2}}$$

$$= \frac{300480 - 300288}{\sqrt{296224 - 295936}\sqrt{305056 - 304704}}$$

$$= \frac{192}{\sqrt{288}\sqrt{352}}$$

$$= \frac{192}{\sqrt{101376}} = \frac{192}{318.3959} = 0.6030$$

$r = 0.6030$

$\therefore r = 0.603$

② Find the correlation co-efficient from the following data.

x :	10	12	14	16	18	20	22	24
y :	14	18	16	22	26	28	27	30

Soln :-

To find correlation co-efficient

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$N = 8$

X	Y	X^2	Y^2	XY
10	14	100	196	140
12	18	144	324	216
14	16	196	256	224
16	22	256	484	352
18	26	324	676	468
20	28	400	784	560
22	27	484	729	594
24	30	576	900	720
$\Sigma X = 136$	$\Sigma Y = 181$	$\Sigma X^2 = 2480$	$\Sigma Y^2 = 4349$	$\Sigma XY = 3274$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{8(3274) - (136)(181)}{\sqrt{8(2480) - (136)^2}\sqrt{8(4349) - (181)^2}}$$

$$= \frac{26192 - 24616}{\sqrt{19840 - 18496}\sqrt{34792 - 32761}}$$

$$= \frac{1576}{\sqrt{1344}\sqrt{2031}}$$

$$= \frac{1576}{\sqrt{2729664}}$$

$$= \frac{1576}{1652.1694} = 0.9538$$

$\therefore r = 0.9538$

$Var(X) = N\Sigma x^2 - (\Sigma x)^2 = 1344$

$Var(Y) = N\Sigma y^2 - (\Sigma y)^2 = 2031$

$cov(X,Y) = N\Sigma XY - \Sigma X \Sigma Y = 1576$

$S.D(X) = \sqrt{N\Sigma x^2 - (\Sigma x)^2} = 36.6606$

$S.D(Y) = \sqrt{N\Sigma y^2 - (\Sigma y)^2} = 45.0666$

③ Find the correlation co-efficient for following data

X:	10	14	18	22	26	30
Y:	18	12	24	6	30	36

Soln:-

To Find: correlation co-efficient

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$N = 6$

X	Y	X^2	Y^2	XY
10	18	100	324	180
14	12	196	144	168
18	24	324	576	432
22	6	484	36	132
26	30	676	900	780
30	36	900	1296	1080
$\Sigma X = 120$	$\Sigma Y = 126$	$\Sigma X^2 = 2680$	$\Sigma Y^2 = 3276$	$\Sigma XY = 2772$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma x^2 - (\Sigma X)^2}\sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{6(2772) - (120)(126)}{\sqrt{6(2680) - (120)^2}\sqrt{6(3276) - (126)^2}}$$

$$= \frac{16632 - 15120}{\sqrt{16080 - 14400}\sqrt{19656 - 15876}}$$

$$= \frac{1512}{\sqrt{1680}\sqrt{3780}}$$

$$= \frac{1512}{\sqrt{6350400}}$$

$$= \frac{1512}{2520} = 0.6$$

$r = 0.6$

$var(X) = N\Sigma x^2 - (\Sigma x)^2 = 1680$

$var(Y) = N\Sigma y^2 - (\Sigma y)^2 = 3780$

$cov(X,Y) = N\Sigma XY - \Sigma X \Sigma Y = 1512$

$S.D(X) = \sqrt{N\Sigma x^2 - (\Sigma x)^2} = 40.9878$

$S.D(Y) = \sqrt{N\Sigma y^2 - (\Sigma y)^2} = 61.4817$

limits for co-efficient (or)

prove that $-1 \leq \gamma_{xy} \leq 1$

$$\gamma(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2\right]^{1/2}}$$

$$\therefore \gamma^2(x,y) = \frac{(\Sigma a_i b_i)^2}{(\Sigma a_i)^2 (\Sigma b_i)^2} \longrightarrow (1)$$

where, $a_i = x_i - \bar{x}$

$b_i = y_i - \bar{y}$

we have the schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots n$ are real quantities.

then,

$$\left[\sum_{i=1}^{n} a_i b_i\right]^2 \leq \left[\sum_{i=1}^{n} a_i^2\right]\left[\sum_{i=1}^{n} b_i^2\right]$$

The sign of equality holding if and only if; $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$

using schwartz inequality, we get from (1)

$$\gamma^2(x,y) \leq 1$$

That is $|\gamma(x,y)| \leq 1 \Rightarrow -1 \leq \gamma(x,y) \leq 1$.

Hence correlation co-efficient cannot exceed unity numerically.

it always lies between -1 and +1.

if $\gamma = +1$, the correlation is perfect and positive and if $\gamma = -1$, correlation is perfect and negative.

properties of correlation co-efficient :-

(i) correlation co-efficient lies between 1 and -1; (i.e) $-1 \leq \gamma \leq 1$

(ii) correlation co-efficient is independent of change of scale. for eg, if the terms of a series are 25, 50, 100, 150, 200, 225; we can simplify these terms are 1, 2, 4, 6, 8, 9; calculate γ easily. (Dividing by 25)

(iii) γ is independent of unit of measurement.

(iv) If b_{xy} and b_{yx} are two regression co-efficients; correlation co-efficient is $\sqrt{b_{xy} \times b_{yx}}$.

(v) correlation coefficient works both ways, or $\gamma_{xy} = \gamma_{yx}$. That is we may take any series dependent and other as independent, its value remains the same.

Regression

Definition: Regression

Regression analysis is a mathematical we asure of the average relationship between two or more variables in terms of the original units of the data.

In Regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for predictio is called independent variable. In regression analysis independent variable is also known as regressor or perdictor or explanatory variable while the dependent variable is also known as regressed or explained variable.

Line of Regression:-

Definition: Curve of regression

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round same curve called the

"Curve of regression".

Definition: Linear regression

If the curve is a straight line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

Derive lines of Regression formula:-

Let us suppose that in the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots n$; y is dependent variable and x is independent variable. Let the line of regression of y on x be $Y = a + bX$.

According to the principle of least squares, the normal equations for estrimatily a and b are.

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \longrightarrow (1)$$

and $$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \longrightarrow (2)$$

From (1) on dividing by n, we get

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

Thus the line of regression of y on x passes through the point $(\bar{x}, \bar{y})$.

Now

$$\mu_{11} = cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y} \longrightarrow (4)$$

Also

$$\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_{x^2} + \bar{x}^2 \longrightarrow (5)$$

Dividing (2) by n and using (4) and (5), we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_{x^2} + \bar{x}^2) \longrightarrow (6)$$

Mutiplying (3) by $\bar{x}$ and then subtracting from (6), we get

$$\mu_{11} = b\sigma_{x^2} \Rightarrow b = \frac{\mu_{11}}{\sigma_{x^2}} \longrightarrow (7)$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_{x^2}}(X - \bar{x}) \longrightarrow (8)$$

$$Y - \bar{y} = \gamma \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \longrightarrow (9)$$

Starting with the equation $X = A + BY$ and proceeding similarly or by simply inter changing the variables X and Y in (8) and (9),

the equation of the line of regression of x on y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_{y^2}} (Y - \bar{y}) \longrightarrow (10)$$

$$x - \bar{x} = \gamma \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \longrightarrow (11)$$

Regression Coefficients :-

'b' the slope of the line of regression of y on x is also called the coefficient of regression of y on x.

b_{yx} = Regression coefficient of y on x

$$= \frac{\mu_{11}}{\sigma_{x^2}} = \gamma \frac{\sigma_y}{\sigma_x}$$

Similarly, the coefficient of regression of x on y.

b_{xy} = Regression coefficient of x on y

$$= \frac{\mu_{11}}{\sigma_{y^2}} = \gamma \frac{\sigma_x}{\sigma_y}$$

Properties of Regression Coefficient

property 1 :-

Correlation coefficient is the geometric mean between the regression coefficients.

proof:-

We know that

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \longrightarrow (1)$$

and $$b_{yx} = r\frac{\sigma_y}{\sigma_x} \longrightarrow (2)$$

From (1) and (2)

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x}$$

$$= r^2$$

(ie)., $$r^2 = b_{xy} \times b_{yx}$$

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

property 2:-

If one of the regression coefficients is greater than unity, the other must be less than unity

(or)

coefficient

If the correlation between two variables is 0 or ±1 how will the regression lines be?

proof :-

Let one of the regression coefficients (say) b_{xy} be greater than unity, than we have to show that $b_{xy} < 1$.

Now $b_{yx} > 1 \Rightarrow \frac{1}{b_{yx}} < 1 \longrightarrow (1)$

Also $\gamma^2 \leq 1 \Rightarrow b_{yx} b_{xy} \leq 1$

Hence $b_{xy} \leq \frac{1}{b_{yx}} < 1$ [from (1)]

property 3:-

Arithmetic mean of the regression coefficients is greater than the correlation coefficient γ, provided $\gamma > 0$.

proof :-

We have to prove that

$$\frac{1}{2}(b_{yx} + b_{xy}) \geq \gamma$$

(or) $$\frac{1}{2}\left(\gamma \frac{\sigma_y}{\sigma_x} + \gamma \frac{\sigma_x}{\sigma_y}\right) \geq \gamma$$

(or) $$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2 \quad (\because \gamma > 0)$$

$$\Rightarrow \sigma_{y^2} + \sigma_{x^2} - 2\sigma_x \sigma_y \geq 0$$

That is $(\sigma_y - \sigma_x)^2 \geq 0$ which is always true, since the square of real quantities is ≥ 0.

Property 4:-

Regression coefficients are independent of the change of origin but not of scale.

proof:-

Let $U = \frac{X-a}{h}$

and $V = \frac{Y-b}{k}$

$\Rightarrow X = a + hU$

$Y = b + kV$.

where $a, b, h > 0$ and $k > 0$ are constants. Then

$$cov(X,Y) = hk\, cov(U,V).$$

$$\sigma_{x^2} = h^2 \sigma_{U^2} \text{ and } \sigma_{y^2} = k^2 \sigma_{V^2}$$

$$b_{yx} = \frac{\mu_{11}}{\sigma_{x^2}} = \frac{hk\, cov(U,V)}{h^2 \sigma_{U^2}}$$

$$= \frac{k}{h} \cdot \frac{cov(U,V)}{\sigma_{U^2}} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that

$$b_{xy} = \left(\frac{h}{k}\right) b_{UV}$$

Derive the formula for the angle between two lines of regression:-

Equations of the lines of regression of y on x, and x on y are $y-\bar{y} = r\cdot\frac{\sigma_y}{\sigma_x}(x-\bar{x})$ and $x-\bar{x} = r\cdot\frac{\sigma_x}{\sigma_y}(y-\bar{y})$.

Slopes of these lines are $r\cdot\frac{\sigma_y}{\sigma_x}$ and $\frac{\sigma_y}{r\sigma_x}$ respectively. If θ is the angle between the two lines of regression than

$$\tan\theta = \frac{\frac{r\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1+r\frac{\sigma_y}{\sigma_x}\cdot\frac{\sigma_y}{r\sigma_x}}$$

$$= \frac{r^2-1}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right)$$

$$= \frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right)$$

$$\therefore \theta = \tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_{x^2}+\sigma_{y^2}}\right)\right\}$$

(1) Given $\text{var}(x) = 9$

problems based on regression.

regression equation $8x-10y+66=0$ and $40x-18y=214$. To Find

(i) To find mean ($\bar{x}, \bar{y}$) value of x and y

(ii) the correlation coefficient between x and y.

(iii) S.D (σ_y) of y when variance(x) = 9.

Soln :-

Given : $8x - 10y + 66 = 0$, $40x - 18y = 214$

Since, both the lines of regression passes through the point $(\bar{x}, \bar{y})$

(i)

$$8x - 10y = -66$$

$$40x - 18y = 214$$

let $x = \bar{x}$

$y = \bar{y}$

$$8\bar{x} - 10\bar{y} = -66 \quad \text{—①}$$

$$40\bar{x} - 18\bar{y} = 214 \quad \text{—②}$$

① × 5 $\Rightarrow 40\bar{x} - 50\bar{y} = -330$

② $\Rightarrow 40\bar{x} - 18\bar{y} = 214$

$$-32\bar{y} = -544$$

$$\bar{y} = \frac{544}{32}$$

$$\bar{y} = 17$$

$\bar{y} = 17$ sub in ①

$$8\bar{x} - 10\bar{y} = -66$$

$8\bar{x} - 10(17) = -66$

$8\bar{x} - 170 = -66$

$8\bar{x} = -66 + 170$

$8\bar{x} = 104$

$\bar{x} = \frac{104}{8}$

$\bar{x} = 13$

(ii) ② $\Rightarrow 40\bar{x} - 18\bar{y} = 214$

$40\bar{x} = 214 + 18\bar{y}$

$\bar{x} = \frac{214}{40} + \frac{18}{40} y$

$\bar{x} = 5.35 + 0.45y$

$b_{xy} = 0.45$ (+ve)

① $\Rightarrow 8\bar{x} - 10\bar{y} = -66$

$-10\bar{y} = -66 - 8x$

$10\bar{y} = 66 + 8x$

$\bar{y} = \frac{66}{10} + \frac{8}{10} x$

$\bar{y} = 66 + 0.8x$

$b_{yx} = 0.8$ (+ve)

$\therefore r = \pm\sqrt{b_{xy} \cdot b_{yx}} = \pm\sqrt{0.45 \times 0.8}$

$= \pm\sqrt{0.36} = \pm 0.6$

$\therefore r = 0.6$

(iii) To Find S.D of y

Given : $Var(X) = 9 = \sigma_x^2$

$$\sigma_x = \sqrt{9} = 3$$

W.K.T $b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$0.45 = 0.6 \times \frac{3}{\sigma_y}$$

$$0.45 \times \sigma_y = 1.8$$

$$\sigma_y = \frac{1.8}{0.45} = 4$$

S.D of $y = 4$

$$\sigma_y = 4$$

Formula :-

$$b_{xy} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

$$b_{yx} = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

① Obtain the equation of two lines of regression for the following data and also find estimate of x for y = 23

x	10	12	14	16	18	20	22	24
y	14	18	16	22	26	28	27	30

$\bar{x} = 17$

$n = 8$

$\bar{y} = 22.625$

$b_{xy} = 0.776$

$b_{yx} = 1.1726$

x on y $\Rightarrow x = 0.776y - 0.557$ (last step 23)

y on x $\Rightarrow y = 1.1726x + 2.6908$