Chapter-02. Correlation And Regression.

Correlation:

Correlation analysis is used as a statistical tool to the association b/w two variables.

The problem in analysing the association b/w two variables can be broken down into two steps We tried to know whether the two variables are related or independent of each other.

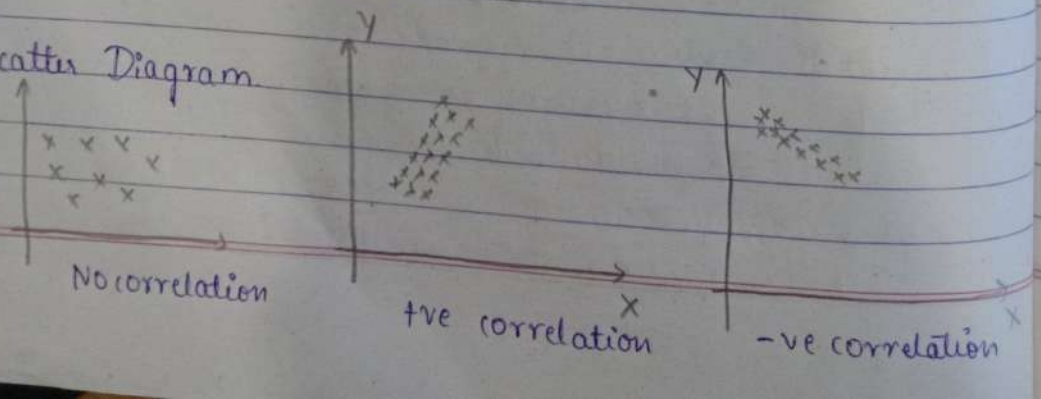
If we find that there is a relationship b/w two variables We try to know its nature and strength. This means whether these variables have +ve or -ve relationship and how close that relationship is.

Types Of Correlation:-

1) Positive Correlation indicating that the movement of the two variables is in the same direction i.e. both are variables are either increasing or decreasing
Ex: Demand and Supply
Income and Expenditure are positively correlated

2) Negative Correlation indicating that the movement of the two variable is in the opposite direction i.e if one variable increases the other decreases and vice-versa.
Ex: volume and pressure
Price and demand

Scatter Diagram



'r' is called the co-efficient of correlation b/w x and y and is defined by:-

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

It can be proved that $-1 \le r \le 1$ if $r = \pm 1$ we say that x and y are perfectly correlated if $r = 0$ we say that x and y are non correlated.

Regression

It is an estimation of unknown from known values. The best fitting straight line of the form

$y = ax + b$ (regression line y on x)

$x = ay + b$ (regression line x on y)

Equation of the regression line:-

⇒ Regression line y on x.

$$y - \bar{y} = byx\,(x - \bar{x})$$

where byx is the regression co-efficient

$$byx = r\frac{\sigma y}{\sigma x}$$

-OR-

$$byx = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

2) Regression line x on y.

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

-OR-

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{-OR-} \quad b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

NOTE:

1) R is the geometric mean of the regression coefficients

i.e $$r = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

2) The sign of r will be +ve or -ve according as the regression coefficients r+ or r- ie the correlation coefficients and the two regression coefficients have same sign

3) One of the regression coefficient is greater than unity the other must be less than unity ie

$b_{yx} > 1$ then $b_{xy} < 1$

so that

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

$$-1 \leq r \leq 1$$

4) Correlation describes the strength of the linear relationship b/w two variables regression tells us how to

ii) TRIBE are own... Retri...

draw the straight line described by the correlation

[P4]

Sales (Rs Crore)	10	11	13	15	16	19	14
Adv Exp (in lakh)	60	62	65	70	73	75	71

i) Regression line x on y.

$$(x-\bar{x}) \cdot b_{xy} (y-\bar{y})$$

$\bar{x} = 14.0000$ $\quad \sigma x = 2.82$

$\bar{y} = 68.0000$ $\quad \sigma y = 5.29150$

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

$r = 0.95$

$\Sigma xy = 6764$ $\quad \Sigma x^2 = 1428$

$\Sigma x = 98$ $\quad \Sigma y^2 = 32564$

$\Sigma y = 478$

$$(x-14) = r \frac{\sigma x}{\sigma y} (y-68)$$

$$(x-14) = 0.95 \times \frac{2.828}{5.2915} (y-68).$$

$$x-14 = 0.5077 (y-68)$$

when $y = 90$

$x = 0.5077 (90-68) + 14$

$x = 25.1694$

(ii) regression line y on x

$(y-\bar{y}) = b_{yx}(x-\bar{x})$.

$y-68 = b_{yx}(x-14)$

$y-68 = 1\,\frac{\sigma_y}{\sigma_x}(x-14)$

$y-68 = 0.95 \times \frac{5.2915}{2.828}(x-14)$.

$y-68 = 1.7775(x-14)$.

when $x=25$

$y = \frac{35021}{400} = 87.64$ lakh

$r = 0.95$ so they are strongly correlated

LP 6]. Find the correlation $r=?$

Production	55	56	58	59	60	60	62
Export	35	38	38	39	44	48	45

$\Sigma x^2 = 24050$ $\Sigma xy = 16568$ $\Sigma x = 410$

$\Sigma y^2 = 11444$ $\Sigma y = 282$

$$r = \frac{7(16568) - (410)(282)}{\sqrt{(7 \times 24050 - (410)^2 \times (7 \times 11444 - (282)^2}}$$

$$= \frac{356}{\sqrt{(68350 - (68100) * (80108 - 79524)}}$$

$$= \frac{356}{\sqrt{250 \times 584}} \quad = \frac{356}{\sqrt{146000}} \quad = \frac{356}{382.099} \quad = 0.9316$$

Production and Exports are strongly correlated i.e Correlation is positive

(e) Use $\tan\theta = \left(\frac{1 - r^2}{r}\right)\left(\frac{\sigma x \, \sigma y}{\sigma x^2 + \sigma y^2}\right)$ to interpret the relation b/w the two variables when $r = 0$, $r = 1$ and $r = -1$

/* θ is angle b/w to regression line $\nearrow m$

$\nearrow byx$

$y = mx + c$

$\nearrow bxy$

$x = my + c$

$byx = r\frac{\sigma y}{\sigma x} \Rightarrow$ slope of y on x.

$byx = r\frac{\sigma x}{\sigma y} \Rightarrow$ slope of x on y.

Slope of x on y is $= \frac{1}{m}$.

$$m_2 = \frac{1}{bxy} = \frac{1}{r\frac{\sigma x}{\sigma y}}$$

$$\tan\theta = \left|\frac{m_1 + m_2}{1 - m_1 m_2}\right|,$$

is the angle b/w two lines.

NOTE:

$y = mx + c$ slope $= m$

$x = my + c$ slope $= \frac{1}{m}$

*/

Regression line y on x

$$m_1 = r\frac{\sigma_y}{\sigma_x}$$

$$m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$$

Angle between two regression lines

$$\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$$

$$\tan\theta = \left|\frac{r\sigma_y/\sigma_x - 1/r\,\sigma_y/\sigma_x}{1 + \left(r\frac{\sigma_y}{\sigma_x} \times \frac{\sigma_y}{r\sigma_x}\right)}\right|$$

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

$$\tan\theta = \left|\frac{\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma_y}{\sigma_x}\right)}{\left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}\right)}\right| = \left|\frac{\left(\frac{r^2-1}{r}\right)\sigma_y\,\sigma_x}{\sigma_x^2 + \sigma_y^2}\right|$$

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

when $r = 0$

$$\tan\theta = \infty$$
$$\theta = \frac{\pi}{2}$$

∴ the two regression lines are $\perp^r$ to each other. Hence the estimated value of y is same for all values of x and vice-versa i.e. there is no relation b/w x and y

when $r = \pm 1$

$$\tan\theta = 0$$
$$\theta = 0, \pi - -$$
or
$$\theta = n\pi, \quad n = 0 \text{ to } \infty.$$

∴ Hence the lines of regression coincide and there is perfect correlation b/w the two variables x and y.

Multiple linear Regression:-

The multiple regression is an extension of linear regression allowing a response variable y to be modelled as a linear function of multidimensional feature vector

The multiple linear regression model

Note Answer any TWO

express the mean of the response variable y as a function of one or more distinct predictor variable $x_1, x_2 \ldots x_k$. It take the form $y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k$

Apply the method of least squares to estimate the parameters $a_0, a_1, a_2 \ldots a_k$.

Normal equations from the least square method.

$$\Sigma y = a_0 n + a_1 \Sigma x_1 + a_2 \Sigma x_2 + \ldots + a_k \Sigma x_k$$

$$\Sigma x_1 y = a_0 \Sigma x_1 + a_1 \Sigma x_1 x_1 + a_2 \Sigma x_1 x_2 + \ldots + a_k \Sigma x_1 x_k$$

$$\Sigma x_2 y = a_0 \Sigma x_2 + a_1 \Sigma x_2 x_1 + a_2 \Sigma x_2 x_2 + \ldots + a_k \Sigma x_2 x_k$$

$$\Sigma x_k y = a_0 \Sigma x_k + a_1 \Sigma x_k x_1 + a_2 \Sigma x_k x_2 + \ldots + a_k \Sigma x_k x_k$$

Normal Equation for $y = a_0 + a_1 x_1 + a_2 x_2$

$$\Sigma y = a_0 n + a_1 \Sigma x_1 + a_2 \Sigma x_2$$

$$\Sigma x_1 y = a_0 \Sigma x_1 + a_1 \Sigma x_1^2 + a_2 \Sigma x_1 x_2$$

$$\Sigma x_2 y = a_0 \Sigma x_2 + a \Sigma x_1 x_2 + a_2 \Sigma x_2^2$$

Strength of Association

Below 80% <Not good>

The coefficient of determination r^2 measures the strength of association and is the ratio of explained variation in y to the total variation in y

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Ex: The correlation co-efficient of number of times absent and final grade is $r = -0.975$
And the coefficient of determination $r^2 = 0.9506$
i.e 95%

About 95% of the variation in final grades can be explained by the no. of times a student is absent The other 5% is unexplained and can be due to sampling error or other variables such as intelligence, amount of time studied etc.

SSY: We shall now refer to this term as the corrected total sum of squares. It measures the total variability in the data

SSR: It is the regression sum of squares and measures the variability in y attributed to the linear association between the mean of y and the prediction variables

SSE: It is the sum of squares of the residuals. It is the measure of the random -OR- unexplained variability in the response variable

$$SSE = SSY - SSR$$

$$R^2 = \frac{SSR}{SSY}$$

$$SSY = [n\Sigma y^2 - (\Sigma y)^2] / n$$

$$SSR = [a_0\Sigma y + a_1 \Sigma y x_1 + a_2 \Sigma y x_2 + \cdots + a_k \Sigma y x_k \{(\Sigma y)^2/n\}$$

LP 8] x: Test 1
y: Test 2

y on x

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

when $x = 85$ $y = ?$

LP 9:] $y = a + b_1 x_1 + b_2 x_2$

$\Rightarrow y = a_0 + a_1 x_1 + a_2 x_2$

y	7	12	17	20
x_1	4	7	9	12
x_2	1	2	5	8

Normal equations are

$\Sigma y = a_0 n + a_1 \Sigma x_1 + a_2 \Sigma x_2$

$\Sigma x_1 y = a_0 \Sigma x_1 + a_1 \Sigma x_1^2 + a_2 \Sigma x_1 x_2$.

$\Sigma x_2 y = a_0 \Sigma x_2 + a_1 \Sigma x_1 x_2 + a_2 \Sigma x_2^2$

$\Sigma x_1 = 32$ $\quad \Sigma x_1^2 = 290$ $\quad \Sigma x_1 x_2 = 159$

$\Sigma y = 56$ $\quad \Sigma y^2 = 882$ $\quad \Sigma x_1 y = 505$

$\Sigma x_2 = 16$ $\quad \Sigma x_2^2 = 94$ $\quad \Sigma x_2 y = 276$.

$56 = a_0(4) + a_1(32) + a_2(16)$

$505 = a_0(32) + a_1(290) + a_2(159)$

$276 = a_0(16) + a_1(159) + a_2(94)$

$a_0 = \frac{38}{59}$ $\quad a_1 = \frac{98}{59}$ $\quad a_2 = \frac{1}{59}$

$a_0 = 0.644$ $\quad a_1 = 1.661$ $\quad a_2 = 0.0169$

$\therefore \Rightarrow y = 0.644 + 1.661 x_1 + 0.0169 x_2$ → MODEL

Co-efficient of multiple determination = $\frac{SSR}{SSY}$

$SSY = [n \Sigma y^2 - (\Sigma y)^2]/n = [4 \times (882) - (56)^2]/4 = 98$

$SSR = [a_0 \Sigma y + a_1 \Sigma x_1 y + a_2 \Sigma x_2 y - \{(\Sigma y)^2/n\}]$

$SSR = [36.064 + 838.805 + 4.6644 - 784]$

$SSR = 95.5334$

$R^2 = \frac{95.5334}{98} = 0.9748 = 97.4\%$ can be explained by the model.

∴ MODEL is good
(i.e. estimation of sales on basis of newspaper and radio advertisement)

Conclusion: The model explains 97.4% of the sales and 2.6% is the error in the model.

Predicted value:

$\underline{\underline{7.30}} = 0.644 + (4 \times 1.661) + (1 \times 0.0169)$

But the actual value is $\underline{\underline{7}}$

Imp:-

Box Plot
qq-norm
qq-plot
Regression
Multiple Regression

Ogive
Histogram
Outliers → Boxplot
median → ogive
mode → histogram

Example:-

Two regression lines are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$ and $\sigma_x^2 = 9$

Find $\bar{x}$, $\bar{y}$ and σ_y and r.

→ Since $(\bar{x}, \bar{y})$ lies on regression lines

$$8\bar{x} - 10\bar{y} = -66$$
$$40\bar{x} - 18\bar{y} = +214$$

- Solving the simultaneous equations-

$\bar{x} = 13 \quad \bar{y} = 17$

$$8x = 10y - 66$$
$$x = \frac{10}{8}y - \frac{66}{8}.$$
$$b_{xy} = \frac{10}{8} > 1.$$

$$18y = 40x - 214$$
$$y = \frac{40x}{18} - \frac{214}{4}$$
$$b_{yx} = \frac{40}{18} > 1$$

b_{xy} and b_{yx} both are greater 1. So extracted values are wrong

$$10y = 8x + 66$$
$$y = \frac{8x}{10} + \frac{66}{10}$$
$$b_{yx} = \frac{8}{10} < 1.$$

$40x = 18y + 214$

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

$b_{xy} = \frac{18}{40} > 1$

b_{yx} is < 1 and $b_{xy} > 1$ So the extracted values are correct

Now,

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

$r = \sqrt{\frac{18}{40} \times \frac{8}{10}}$

$\boxed{r = 0.6}$

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

$\frac{8}{10} = 0.6 \times \frac{\sigma_y}{3}$

$\boxed{\sigma_y = 4}$