# Correlation

Correlation coefficient b/w variables $x$ & $y$ is denoted by $r(x,y)$ or $r_{xy}$ $\boxed{-1 \leq r \leq +1}$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2} \times \sqrt{\sum (y-\bar{y})^2}}$$

$$r = \frac{\frac{1}{n}\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum (x-\bar{x})^2}{n}} \times \sqrt{\frac{\sum (y-\bar{y})^2}{n}}}$$

$$\boxed{r = \frac{Cov(x,y)}{\sigma_x \sigma_y}}$$

Working formula

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

Q1. find the coeff. of correlation

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ : | 1 | 3 | 5 | 7 | 8 | 10 |
| $y$ : | 8 | 12 | 15 | 17 | 18 | 20 |

Sol

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 1 | 8 | 1 | 64 | 8 |
| 3 | 12 | 9 | 144 | 36 |
| 5 | 15 | 25 | 225 | 75 |
| 7 | 17 | 49 | 289 | 119 |
| 8 | 18 | 64 | 324 | 144 |
| 10 | 20 | 100 | 400 | 200 |
| 34 | 90 | 248 | 1446 | 582 |

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

$$= \frac{3492 - 3060}{18.22 \times 24} =$$

$$= 0.9879$$

# Rank correlation Coeff.

$$\boxed{r = 1 - \frac{6\Sigma D^2}{n(n^2-1)}}$$ if Ranks are non repeated

$$\boxed{r = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12} m_1(m_1^2-1) + \frac{1}{12} m_2(m_2^2-1) + \cdots\right]}{n(n^2-1)}}$$ if ranks are repeated

| X | Y | $R_x$ | $R_y$ | $D = R_x - R_y$ | $D^2$ |
|---|---|---|---|---|---|
| 7 | 8 | 2 | 1 | 1 | 1 |
| 9 | 2 | 1 | 3 | -2 | 4 |
| 2 | 3 | 3 | 2 | 1 | 1 |
| | | | | | $\Sigma D^2 - 6$ |

Q Calculate rank correlation Coeff.

| X | Y | $R_x$ | $R_y$ | D | $D^2$ |
|---|---|---|---|---|---|
| 68 | 62 | 4 | 5 | -1 | 1 |
| 64 | 58 | 6 | 7 | -1 | 1 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 50 | 45 | 9 | 10 | -1 | 1 |
| 64 | 81 | 6 | 1 | 5 | 25 |
| 80 | 60 | 1 | 6 | -5 | 25 |
| 75 | 68 | 2.5 | 3.5 | -1 | 1 |
| 40 | 48 | 10 | 9 | 1 | 1 |
| 55 | 50 | 8 | 8 | 0 | 0 |
| 64 | 70 | 6 | 2 | 4 | 16 |
| | | | | | $\Sigma D^2 = 72$ |

$m_1 = 2$   $75 < {2 \atop 3}$   mean $= 2.5$   $m_3 = 2$   $68 < {3 \atop 4}$

$m_2 = 3$   $64 < {5 \atop {6 \atop 7}}$   $m = 6$

$$r = 1 - \frac{6\left[72 - \frac{1}{12}\times 2(4-1) - \frac{1}{12}\times 3(9-1) - \frac{1}{12}\times 2(4-1)\right]}{10(100-1)}$$

$$= 1 - \frac{6\left[72 + \frac{1}{2} + 2 + \frac{1}{2}\right]}{990}$$

$$= 1 - \frac{6[72+3]}{990} = 1 - \frac{6\times 75}{990} = 1 - \frac{\cancel{420}}{990}$$

$$= 1 - \frac{450}{990} \qquad = 1 - \cancel{4.2515}$$

$$= -\cancel{3.2515}$$

$$= 0.5454$$

Q Obtain rank correlat$^n$ coeff.

| X | Y | $R_x$ | $R_y$ | D | $D^2$ |
|---|---|---|---|---|---|
| 15 | 50 | 7 | 3 | 4 | 16 |
| 20 | 30. | 5.5 | 5 | 0.5 | 0.25 |
| 27 | 55 | 4 | 2 | 2 | 4 |
| 13 | 30 | 8 | 5 | 3 | 9 |
| 45 | 25 | 3 | 7 | -4 | 16 |
| 60 | 10 | 2 | 8 | -6 | 36 |
| 20 | 30 | 5.5 | 5 | 0.5 | 0.25 |
| 75 | 70 | 1 | 1 | 0 | 0 |
| | | | | | $\Sigma D^2 = 81.5$ |

$m_1 = 2$ , $m_2 = 3$

$$r = 1 - \frac{6\left[81.5 + \frac{1}{12}\times 2(4-1) + \frac{1}{12}\times 3(9-1)\right]}{8(64-1)} = 1 - \frac{6(81.5+0.5+2)}{8\times 63}$$

$$= 1 - \frac{\cancel{6}^{3}\times \cancel{84}^{21}}{\cancel{8}^{4}\times \cancel{63}^{21}}$$

$$r = 1 - 1 = 0$$

Q Calculate rank correlatⁿ

| X | Y | $R_x$ | $R_y$ | D | $D^2$ |
|---|---|---|---|---|---|
| 10 | 30 | 9 | 9 | 0 | 0 |
| 15 | 42 | 5 | 3 | 2 | 4 |
| 12 | 45 | 8 | 2 | 6 | 36 |
| 17 | 46 | 3 | 1 | 2 | 4 |
| 13 | 33 | 7 | 8 | -1 | 1 |
| 16 | 34 | 4 | 7 | -3 | 9 |
| 24 | 40 | 1 | 4 | -3 | 9 |
| 14 | 35 | 6 | 6 | 0 | 0 |
| 22 | 39 | 2 | 5 | -3 | 9 |
| | | | | | 72 |

$$r = 1 - \frac{\overset{2}{\cancel{6}} \times \overset{24}{\cancel{72}}}{\underset{3}{\cancel{9}}(81-1)} = 1 - \frac{24}{40} = 1 - \frac{6}{10}$$

$$\boxed{r = 0.4}$$ A

# Correlation and Regression

Regression line y on x $(y-\bar{y}) = b_{yx}(x-\bar{x})$

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

$$\bar{x} = \frac{\Sigma x}{n}, \quad \bar{y} = \frac{\Sigma y}{n}$$

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

Regression coeff. $b_{xy}$ & $b_{yx}$

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

$\boxed{b_{xy} \times b_{yx} = r^2}$ $\boxed{r = \sqrt{b_{xy} \times b_{yx}}}$

$-1 \leq r \leq +1$

- If $b_{xy}$ & $b_{yx}$ are positive then $r$ is positive
- If $b_{xy}$ & $b_{yx}$ are negative then $r$ is negative

Q1. The foll. table gives the age ($x$) in years of cars and annual maintainance cost ($y$) in hundred rupees

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1 | 3 | 5 | 7 | 9 |
| $y$ | 15 | 18 | 21 | 23 | 22 |

estimate the maintainance cost for a 4 years old car after foll. regression eq$^n$

Sol. Regression line y on x

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

| $x$ | $y$ | $xy$ | $x^2$ |
|---|---|---|---|
| 1 | 15 | 15 | 1 |
| 3 | 18 | 54 | 9 |
| 5 | 21 | 105 | 25 |
| 7 | 23 | 161 | 49 |
| 9 | 22 | 198 | 81 |
| 25 | 99 | 533 | 165 |

$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5$

$\bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$

~~$(y - 19.8)$~~ =

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

$$= \frac{5 \times 533 - 25 \times 99}{5 \times 165 - 625} = \frac{190}{200} = 0.95$$

$$(y - 19.8) = 0.95\,(x - 5)$$

$$y = 0.95x + 15.05$$

$$y = 0.95 \times 4 + 15.05 = 18.85 \text{ hundred rupees}$$
$$= 1885 \text{ rupees}$$

2. find correlat$^n$ coef and obtain the eq$^n$ to the line of regression for the foll. data

| | | | | | |
|---|---|---|---|---|---|
| $x$ : | 6 | 2 | 10 | 4 | 8 |
| $y$ : | 9 | 11 | 5 | 8 | 7 |

Sol.

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

| $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 6 | 9 | 54 | 36 | 81 |
| 2 | 11 | 22 | 4 | 121 |
| 10 | 5 | 50 | 100 | 25 |
| 4 | 8 | 32 | 16 | 64 |
| 8 | 7 | 56 | 64 | 49 |
| 30 | 40 | 214 | 220 | 340 |

$$\bar{x} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{50}{5} = 8$$

$$b_{yx} = \frac{(5 \times 214) - (40 \times 30)}{(5 \times 220) - (30)^2} = -\frac{130}{200}$$

$$= -0.65$$

$$b_{xy} = \frac{-130}{(5 \times 340) - (40)^2} = \frac{-130}{100} = -1.3$$

$$(y-8) = -0.65 \times (x-6)$$

$$\boxed{y = -0.65x + 11.9}$$

$$(x-6) = -1.3 \times (y-8)$$

$$\boxed{x = -1.3y + 16.4}$$

$$r = \sqrt{(-0.65) \times (-1.3)} = -0.9192$$

Imp

Q3 Two lines of regression is given by

$x + 2y - 5 = 0$, $2x + 3y - 8 = 0$ & $\sigma_x^2 = 12$

variance of $x$

Calc. (a) Mean value of $x$ & $y$

(b) Variance of $y$

(c) The coeff. of correlatⁿ

Sol. $x + 2y - 5 = 0$ $\quad$ $2x + 3y - 8 = 0$

Check $y$ on $x$ $\quad$ $x$ on $y$

$2y = -x + 5$ $\quad$ $2x = -3y + 8$

$y = -\frac{1}{2}x + \frac{5}{2}$ $\quad$ $x = -\frac{3}{2}y + 4$

$b_{yx} = -\frac{1}{2}$ $\quad$ $b_{xy} = -\frac{3}{2}$

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}$$

ⓒ $$\boxed{r = -\frac{\sqrt{3}}{2}}$$

ⓐ To find $\bar{x}$ & $\bar{y}$

$\bar{x} + 2\bar{y} = 5$
$2\bar{x} + 3\bar{y} = 8$

} Mean always satisfy regression Eq$^n$

$\boxed{\bar{y} = 2}$ $\boxed{\bar{x} = 1}$

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

$$-\frac{1}{\cancel{2}} = -\frac{\sqrt{3}}{\cancel{2}} \times \frac{\sigma_y}{\sqrt{12}}$$

$$\boxed{\sigma_y = \sqrt{\frac{12}{3}} = 2}$$ Ans

Variance of $y$ $(\sigma_y^2) = 2^2 = 4$

V. Imp.

4. If $4x - 5y + 33 = 0$ & $20x - 9y = 107$ are 2 lines of regression. find the mean values of $x$ & $y$, coeff. of correlation, std. deviation of $y$ if the variance of $x$ is 9.

Sol. $y$ on $x$

$$y = \frac{4}{5}x + \frac{33}{5}$$

$$b_{yx} = 4/5$$

$x$ on $y$

$$x = \frac{9}{20}y + \frac{107}{20}$$

$$b_{xy} = 9/20$$

$$r = \sqrt{\frac{4}{5} \times \frac{9}{20}} = \frac{3}{5} = 0.6$$

Hint: Bigger $x$ coeff means $x$ on $y$

(b) To find $\bar{x}$ & $\bar{y}$

$$4\bar{x} - 5\bar{y} = -33$$
$$20\bar{x} - 9\bar{y} = 107$$

$$\boxed{\bar{x} = 13, \bar{y} = 17}$$

(c)

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

$$\frac{4}{5} = \frac{3}{5} \times \frac{\sigma_y}{3}$$

$$\boxed{\sigma_y = 4}$$

5. For two random variables X & Y with the same mean the 2 regression eq^n are $y = ax + b$ & $x = \alpha y + \beta$

Show that $\frac{b}{\beta} = \frac{1-a}{1-\alpha}$. Also find the common mean.

$\left(m = \frac{\beta - b}{a - \alpha}\right)$ Ans.

[AKTU 2017]

Sol. $\bar{x} = \bar{y} = m$

$m = am + b \Rightarrow b = m(1-a)$ — (I)

$m = \alpha m + \beta \Rightarrow \beta = m(1-\alpha)$ — (II)

$$\boxed{\frac{b}{\beta} = \frac{m(1-a)}{m(1-\alpha)} = \frac{(1-a)}{(1-\alpha)}}$$

proved

$① = ⑪$

$am + b = \alpha m + \beta$

$am - \alpha m = \beta - b$

Common mean $\boxed{m = \dfrac{\beta - b}{a - \alpha}}$ ~~proved~~ Ans

**Linear regression, Non-linear regression, Multiple regression**

1. find the multiple linear regression eqⁿ of $X_1$ on $X_2$ & $X_3$

| | | | | | | |
|---|---|---|---|---|---|---|
| $X_1$ | 3 | 5 | 6 | 8 | 12 | 10 |
| $X_2$ | 10 | 10 | 5 | 7 | 5 | 2 |
| $X_3$ | 20 | 25 | 15 | 16 | 15 | 2 |

Sol. fit $X_1 = a + bX_2 + cX_3$

Normal eqⁿ

$\Sigma X_1 = an + b\Sigma X_2 + c\Sigma X_3$

$\Sigma X_1X_2 = a\Sigma X_2 + b\Sigma X_2^2 + c\Sigma X_2X_3$

$\Sigma X_1X_3 = a\Sigma X_3 + b\Sigma X_2X_3 + c\Sigma X_3^2$

| $X_1$ | $X_2$ | $X_3$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ | $X_2^2$ | $X_3^2$ |
|---|---|---|---|---|---|---|---|
| 3 | 10 | 20 | 30 | 60 | 200 | 100 | 400 |
| 5 | 10 | 25 | 50 | 125 | 250 | 100 | 625 |
| 6 | 5 | 15 | 30 | 90 | 75 | 25 | 225 |
| 8 | 7 | 16 | 56 | 128 | 112 | 49 | 256 |
| 12 | 5 | 15 | 60 | 180 | 75 | 25 | 225 |
| 10 | 2 | 2 | 20 | 20 | 4 | 4 | 4 |
| 44 | 39 | 93 | 246 | 603 | 716 | 303 | 1735 |

$44 = 6a + 39b + 93c$

$246 = 39a + 303b + 716c$

$603 = 93a + 716b + 1735c$

$a = 12.36$

$b = -1.3986$

$c = 0.2621$

fit: $\boxed{X_1 = 12.36 - 1.3986X_2 + 0.2621X_3}$ Ans

[ARTU 2022]

2. fit a parabolic curve of regression of y on x to the foll. Data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ : | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| $y$ : | 1.1 | 1.3 | 1.6 | 2 | 2.7 | 3.4 | 4.1 |

Sol. Parabolic : $y = a + bx + cx^2$ —— (1)

$\Sigma y = an + b\Sigma x + c\Sigma x^2$

$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$

| $x$ | $y$ | $xy$ | $x^2$ | $x^2y$ | $x^3$ | $x^4$ |
|---|---|---|---|---|---|---|
| 1 | 1.1 | 1.1 | 1 | 1.1 | 1 | 1 |
| 1.5 | 1.3 | 1.95 | 2.25 | 2.925 | 3.375 | 5.0625 |
| 2 | 1.6 | 3.2 | 4 | 6.4 | 8 | 16 |
| 2.5 | 2 | 5.0 | 6.25 | 12.5 | 15.625 | 39.0625 |
| 3 | 2.7 | 8.1 | 9 | 24.3 | 27 | 81 |
| 3.5 | 3.4 | 11.9 | 12.25 | 41.65 | 42.875 | 150.0625 |
| 4 | 4.1 | 16.4 | 16 | 65.6 | 64 | 256 |
| 17.5 | 16.2 | 47.65 | 50.75 | 154.475 | 161.875 | 548.1875 |

$7a + 17.5b + 50.75c = 16.2$

$17.5a + 50.75b + 161.875c = 47.65$

$50.75a + 161.875b + 548.1875c = 154.475$

$a = 1.0357$, $b = -0.1928$, $c = 0.2428$

$$\boxed{y = 1.0357 - 0.1928x + 0.2428x^2}$$

Ans

To prove :- $\tan\theta = \frac{1-r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}$

Angle b/w 2 slope '$\theta$'

Proof :- $\tan\theta = \frac{m_2 - m_1}{1+m_1 m_2}$

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

compare $\quad y = mx + c$

$$m_1 = r\frac{\sigma_y}{\sigma_x} \quad \text{—— (1)}$$

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

$$(y-\bar{y}) = \frac{1}{r}\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x} \quad \text{—— (2)}$$

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

$$= \left(\frac{1-r^2}{r}\right)\frac{\sigma_y/\sigma_x}{\frac{(\sigma_x^2+\sigma_y^2)}{\sigma_x^2}} = \left(\frac{1-r^2}{r}\right)\times\frac{\sigma_y\sigma_x}{\sigma_x^2+\sigma_y^2}$$

<u>proved</u>

* If $r=0$; then $\tan\theta = \infty$

$\theta = \frac{\pi}{2}$ means slopes are $\perp$ar.

* If $r = \pm 1$; then $\tan\theta = 0 \Rightarrow \theta = 0$

slopes are ||el or overlap each other.

Q Obtain a regression plane by using multiple linear regression to fit the data

| | | | | |
|---|---|---|---|---|
| X | 1 | 2 | 3 | 4 |
| Z | 0 | 1 | 2 | 3 |
| Y | 12 | 18 | 24 | 30 |

} On mentⁿ dependent then last value suppose as dependent

Sol. let
$$y = a + bx + cz$$
$$\Sigma y = an + b\Sigma x + c\Sigma z$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma xz$$
$$\Sigma zy = a\Sigma z + b\Sigma xz + c\Sigma z^2$$

| $x$ | $z$ | $y$ | $x^2$ | $z^2$ | $xy$ | $yz$ | $xz$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 12 | 1 | 0 | 12 | 0 | 0 |
| 2 | 1 | 18 | 4 | 1 | 36 | 18 | 2 |
| 3 | 2 | 24 | 9 | 4 | 72 | 48 | 6 |
| 4 | 3 | 30 | 16 | 9 | 120 | 90 | 12 |
| 10 | 6 | 84 | 30 | 14 | 240 | 156 | 20 |

$$4a + 10b + 6c = 84$$
$$10a + 30b + 20c = 240$$
$$6a + 20b + 14c = 156$$
} Infinite solⁿ

$$\boxed{y = 10 + 2x + 4z}$$ Ans

Imp

Q If the coeff. of correlatⁿ b/w $x$ & $y$ is 0.5 and the acute angle b/w their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$ show that $\sigma_x = \frac{1}{2}\sigma_y$

Sol. $\tan\theta = \frac{1-r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}$

$\tan\theta = \frac{3}{5} \qquad r = \frac{1}{2}$

$$\frac{3}{5} = \frac{1-1/4}{1/2} \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

$$\frac{\cancel{3}}{5} = \frac{\cancel{3}}{2} \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

$$\boxed{2\sigma_x^2+2\sigma_y^2 = 5\sigma_x\sigma_y}$$

$$2\sigma_x^2 - 5\sigma_x\sigma_y + 2\sigma_y^2 = 0$$

$$2\sigma_x^2 - 4\sigma_x\sigma_y - \sigma_x\sigma_y + 2\sigma_y^2 = 0$$

$$2\sigma_x(\sigma_x - 2\sigma_y) - \sigma_y(\sigma_x - 2\sigma_y) = 0$$

$$(2\sigma_x - \sigma_y)(\sigma_x - 2\sigma_y) = 0$$

we get, $2\sigma_x - \sigma_y = 0$

$$\boxed{\sigma_x = \frac{1}{2}\sigma_y}$$ proved

Imp

Q Two random variable have the regression line with equations $3x+2y=26$, $6x+y=31$ find the mean value of $x$ & $y$ correlatn coeff. $x$ & $y$.

Sol. $6\bar{x} + \bar{y} = 31$, $3\bar{x} + 2\bar{y} = 26$ } $\bar{x} = 4$, $\bar{y} = 7$

y on x

$3x + 2y = 26$

$y = 13 - \frac{3}{2}x$

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

x on y

$6x + y = 31$

$x = \frac{31}{6} - \frac{y}{6}$

$b_{xy} = -\frac{1}{6}$

$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{\frac{3}{2} \times \frac{1}{6}} = \sqrt{\frac{3}{12}} = \sqrt{\frac{1}{4}} = \frac{1}{2} = 0.5$

$\boxed{r = -0.5}$ Ans