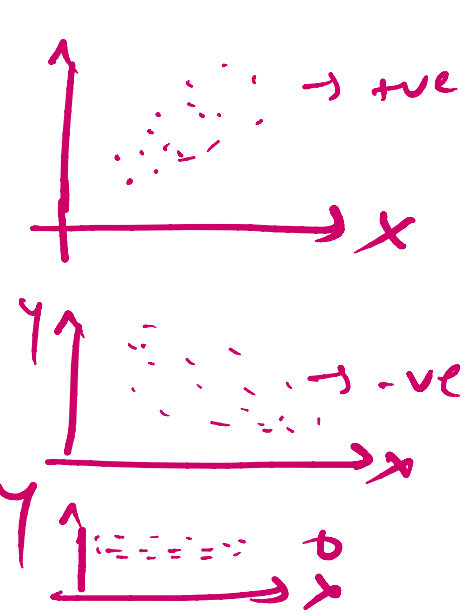
07 November 2023 10:36 PM

Correlation and Regression

- measure the degree of association between variables (relation)
- bivariate data ⇒ relation between two variable
  - one indp → x
  - one dep → y

Graphical representation of this degree of association between x and y ⇒ Scatter Diagram.

1. +ve correlation
2. -ve correlation
3. Uncorrelated (0) (No relation between x & y).



— ✱ —

Mathematically, the measure of degree of association between two variable x and y is correlation coefficient denoted by

is correlation coefficient denoted by

$$r_{x,y} = \frac{\text{cov}(x,y)}{\sqrt{v(x)}\sqrt{v(y)}}$$

Now, $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

or, $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}$

s.d, $\sigma_x = \sqrt{v(x)} = \sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2} = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$\sigma_y = \sqrt{v(y)} = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

Properties: (i) $-1 \le r_{x,y} \le +1$

(ii) correlation coefficient do not depend on the change in origin and scale of variable.

ie if x and y are changed as $u = \frac{x-c}{d}$ and $v = \frac{y-c'}{d'}$

then $r_{x,y} = r_{u,v}$

Proof

(i) $-1 \le r_{x,y} \le 1$.

Let x and y assumes the value (x_i, y_i),

Let x and y assumes the value (x_1, y_1), $(x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, where

$\bar{x} = \frac{1}{n}\Sigma x_i$ and $\bar{y} = \frac{1}{n}\Sigma y_i$ the resp mean of x and y

and σ_x and σ_y are the resp s.d.

Now let us write

$u_i = \frac{x_i - \bar{x}}{\sigma_x}$ and $v_i = \frac{y_i - \bar{y}}{\sigma_y}$

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 = \frac{\Sigma (x_i - \bar{x})^2}{n \sigma_x^2}$$

$$\Sigma u_i^2 = \frac{\sigma_x^2}{\sigma_x^2} = 1$$

and $\Sigma v_i^2 = \Sigma \left(\frac{y_i - \bar{y}}{\sigma_y}\right)^2 = \frac{\frac{1}{n}\Sigma (y_i - \bar{y})^2}{\sigma_y^2}$

$$\Sigma v_i^2 = \frac{\sigma_y^2}{\sigma_y^2} = 1$$

and $\Sigma u_i v_i = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$

$$= \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \cdot \sigma_y}$$

$$= \frac{\text{cov}(x, y)}{}$$

$$\sqrt{v(x)}\sqrt{v(y)}$$

$$\boxed{\Sigma u_i v_i = r_{x,y}}$$

So, $\Sigma (u_i + v_i)^2 \geq 0$

$$\Sigma u_i^2 + \Sigma v_i^2 + 2\Sigma u_i v_i \geq 0$$

$$1 + 1 + 2 \cdot r_{x,y} \geq 0$$

$$2 + 2 r_{x,y} \geq 0$$

$$2 (1 + r_{x,y}) \geq 0$$

$\Rightarrow$ $1 + r_{x,y} \geq 0$

$\therefore$ $\boxed{r_{x,y} \geq -1}$ —— (1)

Again $\Sigma (u_i - v_i)^2 \geq 0$

$$\Sigma u_i^2 + \Sigma v_i^2 - 2\Sigma u_i v_i \geq 0$$

$$1 + 1 - 2 r_{x,y} \geq 0$$

$$2 (1 - r_{x,y}) \geq 0$$

$\Rightarrow$ $1 - r_{x,y} \geq 0$

$\Rightarrow$ $\boxed{r_{x,y} \leq 1}$ — (2)

Comparing (1) and (2) $-1 \leq r_{x,y} \leq 1$ (Proved)

Proof (ii) Let x and y be the variables and $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the n pair of observations.

Now let us change the origin and scale of X variable by 'a' and 'b' units

then $u_i = \dfrac{x-a}{b}$

Similarly if Y variable is changed by 'c' units of origin and 'd' units of scale then,

$$v_i = \frac{y-c}{d}$$

Here from the given observation of (x_i, y_i) we have $\bar{x}$, $\bar{y}$ as mean and σ_x, σ_y as s.d of x and y resp

$$\therefore r_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

Since $u_i = \dfrac{x_i - a}{b}$

$\Rightarrow x_i = a + b u_i$

$\therefore \boxed{\bar{x} = a + b\bar{u}}$

Similarly $v_i = \dfrac{y_i - c}{d}$

$\boxed{\bar{y} = c + d\bar{v}}$

$$\sigma_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\sigma_x^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$= \frac{1}{n}\Sigma(a + bu_i - a - b\bar{u})^2$$

$$= \frac{1}{n} \cdot b^2 \Sigma(u_i - \bar{u})^2$$

$$\therefore \sigma_x^2 = b^2 \sigma_u^2$$

$$\boxed{\sigma_x = |b|\,\sigma_u}$$

Similarly $\boxed{\sigma_y = |d|\,\sigma_v}$

$$\text{Cov}(x,y) = \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\Sigma(a + bu_i - a - b\bar{u})(c + dv_i - c - d\bar{v})$$

$$= \frac{bd}{n}\Sigma(u_i - \bar{u})(v_i - \bar{v})$$

$$= \boxed{b \cdot d\ \text{cov}(u,v)}$$

∴ Correlation coefficient between x, y

is $r_{x,y} = \dfrac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \dfrac{b \cdot d \cdot \text{cov}(u,v)}{|b|\sigma_u \cdot |d|\sigma_v}$

$$r_{x,y} = \frac{bd}{|b||d|} \cdot r_{u,v}$$

if b and d are of same sign

then $r_{x,y} = r_{u,v}$

and if b and are of opposite sign

then $r_{x,y} = -r_{u,v}$

Q $\sum_{i=1}^{100} x_i = 280$ $\sum_{i=1}^{100} y_i = 60$ } Calculate

Q. $\sum_{i=1} x_i = 280$ $\quad \sum_{i=1} y_i = 60$

$\sum x_i^2 = 2384$ $\quad \sum y_i^2 = 117$

$\sum x_i y_i = 438$

Calculate $r_{x,y}$.

$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{100} \times 280 = 2.8$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{60}{100} = 0.6$

$\sigma_x = \sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \sqrt{\frac{2384}{100} - 2.8^2}$

$= \sqrt{16} = 4.$

$\sigma_y = \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2} = \sqrt{0.81} = 0.9$

$cov(x,y) = \frac{1}{n}\sum xy - \bar{x}\bar{y}$

$= \frac{438}{100} - (2.8 \times 0.6) = 4.38 - (1.68)$

$= 2.70$

$\therefore r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{2.7}{4 \times 0.9} = \underline{\underline{ans}}.$

Q

x	65	63	67	64	68	62	70	66
y	68	66	68	65	69	66	68	65

x	65	63	67	64	68	62	70	66
y	68	66	68	65	69	66	68	65

Calculate r_{xy}.

x	y	x^2	y^2	xy
Σx	Σy	Σx^2	Σy^2	Σxy

mean — s.d — cov. $\therefore r_{xy} =$ ans.

Q While calculating the coefficient of correlation between two variables x and y, the following results were obtained :

$$n = 25 \quad \Sigma x = 125 \quad \Sigma y = 100$$

$$\Sigma x^2 = 650 \quad \Sigma y^2 = 460$$

$$\Sigma xy = 508$$

It was however later discovered at the time of checking that two pairs of obs (x, y) were copied $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$

0 copied (6,7) and (8,8) wrongly. the correct values were (8,12) and (6,8) respectively. Determine the correct coefficient of correlation.

From the question we have

$\Sigma x = 125$

We will subtract incorrect obs and add the correct obs, we get.

corrected Σx = $125 - 6 - 8 + 8 + 6$

$= 125$

$\therefore$ corrected mean, $\bar{x} = \frac{1}{25} \times 125 = 5$ ✓

Again $\Sigma y = 100$

after substitution ....., we get

corrected $\Sigma y = 100 - 14 - 6 + 12 + 8$

$= 100$

$\therefore$ corrected mean $\bar{y} = \frac{1}{25} \times 100 = 4$ ✓

given $\Sigma x^2 = 650$

corrected $\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

$\therefore \sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = \sqrt{\frac{650}{25} - 5^2} = \sqrt{26 - 25}$

$= \sqrt{1} = 1$

and $\Sigma y^2 = 460$

corrected $\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2$

$= 436$

$$\sigma_y = \sqrt{\frac{1}{25} \times 936 - 4^2}$$

$$= \sqrt{1.44}$$

$$\sigma_y = 1.2$$

given

$\Sigma xy = 508$

correct $\Sigma xy = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8)$

$\Sigma xy = 520$

$$\therefore \text{cov}(x,y) = \frac{1}{5} \times 520 - (5 \times 4)$$

$$= \frac{104}{5} - 20$$

$$= \frac{104 - 100}{5}$$

$$= \frac{4}{5}$$

$$\therefore r_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{4}{5}}{1 \times 1.2} = \frac{4/0.6}{5 \times 1.2} = \frac{8}{12} = \frac{2}{3}$$

(ans)

— x —