Dr Mariamma Antony
Asst.Professor in Statistics
Little Flower College,Guruvayur
2020-2021

CORRELATION and REGRESSION

➢ CORRELATION

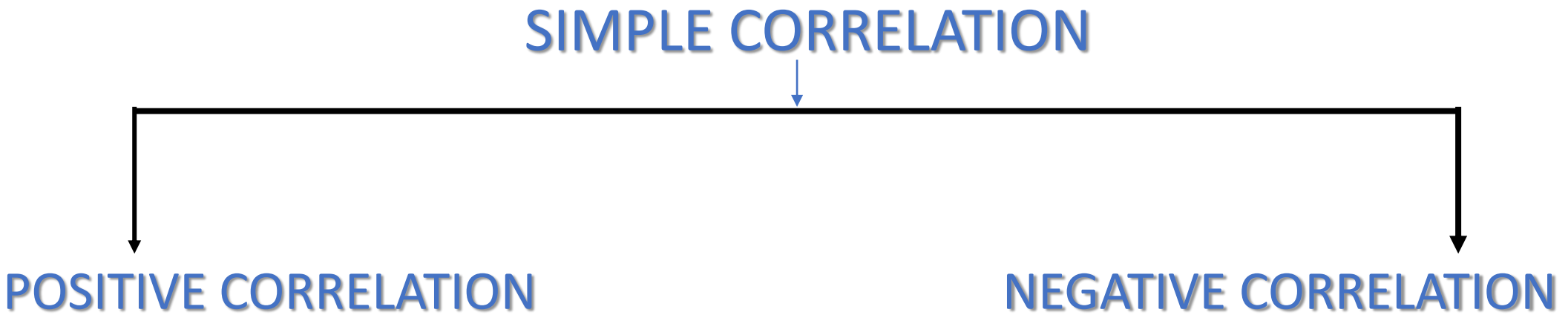
- It is the measure of degree of linear or non-linear association between two or more variables.

▪ TYPES OF CORRELATION

1. **SIMPLE CORRELATION**
2. **MULTIPLE CORRELATION**
3. **PARTIAL CORREALTION**

1 . SIMPLE CORRELATION

- It is based on linear relationship between the variables.



- If the linear relation is in such a way that the increment in one variable results in the increment of other also , then there is a **positive or direct correlation**.
- If the linear relation is in such a way that the increment in one variable results in the decrease of the other , and then there is a **negative or inverse correlation.**

2. MULTIPLE CORRELATION

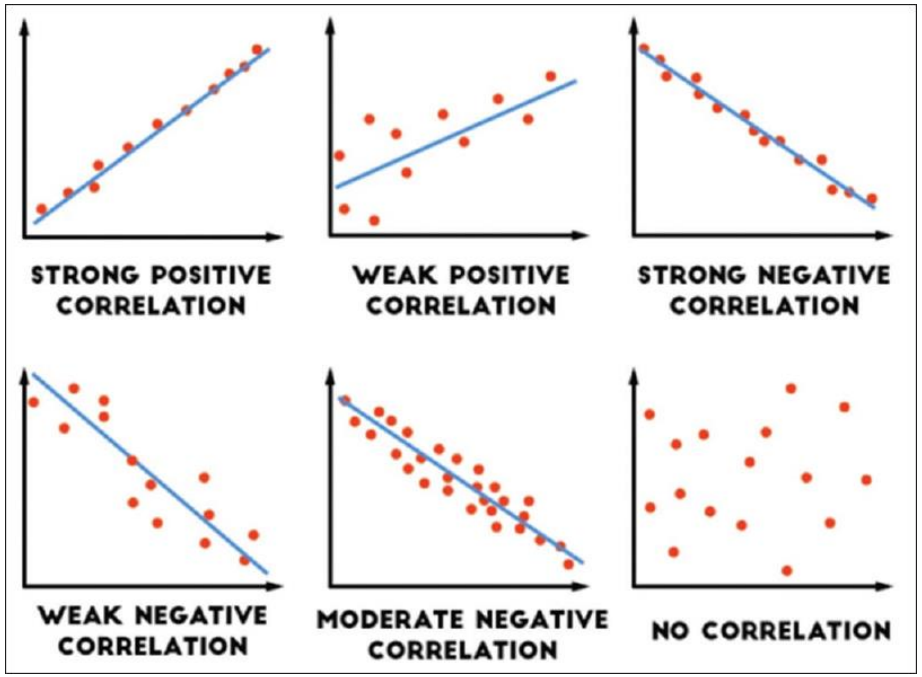
- When we consider more than two variables at a time and their association , the degree of association of one variable to the remaining variables is coming under multiple correlation.

3. PARTIAL CORRELATION

- When we consider the degree of association between two variables by assuming all other variables as constants is coming under partial correlation

SCATTER DIAGRAM

- A graph in which the values of two variables are plotted along two axes , the pattern of the resulting points revealing any correlation present.



➢ CURVE FITTING

- The process of determining the best values of the parameters involved in a proposed relation between the variables considered statistically is known as **curve fitting**.
- The value of parameters are estimated using the ***Principle of least squares*** **.**

▪ PRINCIPLE OF LEAST SQUARES

The Principle of least squares states that the best estimates of $a_1, a_2, \ldots, a_n$ in the relation $y_i = f(x_i, a_1, a_2, \ldots, a_n) + \varepsilon_i$ are those values of $a_1, a_2, \ldots, a_n$ which minimize the sum of squares of the residual errors. Hence to find the values of $a_1, a_2, \ldots, a_n$ such that

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i, a_1, a_2, \ldots, a_n)]^2 \text{ is minimum.}$$

FITTING OF A STRAIGHT LINE y=ax+b

1. Fitting a straight line by the method of least squares:

Let $(x_i, y_i), i = 0,1,2, \ldots\ldots, n$ be the n sets of observations and let the related relation by y = ax + b. Now we have to select a and b so that the straight line is the best fit to the data.

As explained earlier, the residual at $x = x_i$ is

$$d_i = y_i - f(x_i) = y_i - (ax_i + b), i = 1,2, \ldots, n$$

$$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

By the principle of least squares, E is minimum.

$$\frac{\partial E}{\partial a} = 0 \; and \; \frac{\partial E}{\partial b} = 0$$

i.e., $2\sum[y_i - (ax_i + b)](-x_i) = 0$ & $2\sum[y_i - (ax_i + b)](-1) = 0$

i.e., $\sum_{i=1}^{n}(x_i y_i - ax_i^2 - bx_i) = 0$ & $\sum_{i=1}^{n}(y_i - ax_i - b) = 0$

i.e., $a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$ .... (eq.1)

And $a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$ ..... (eq.2)

Since, x_i, y_i are known, equations (1) & (2) give two equations in a & b. Solve for a & b from (1) & (2) & obtain the best fit y= ax + b.

Note:

- Equations (1) & (2) are called normal equations.
- Dropping suffix i from (1) & (2), the normal equations are

  $$a\sum x + nb = \sum y \; \& \; a\sum x^2 + b\sum x = \sum xy$$

  Which are get taking $\sum$ on both sides of y = ax + b & also taking $\sum$ on both sides after multiplying by x both sides of y = ax + b.

- Transformation like $X = \frac{x-a}{h}, Y = \frac{y-b}{h}$ reduce the linear equation y = ax + b to the form Y = AX + B. Hence, a linear fit is another linear fit in both systems of coordinates.

FITTING OF A CURVE y=ax^2+bx+c

- Normal equations are:

$$1.\ \sum y = an + b\sum x + c\sum x^2$$

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

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

FITTING OF y=ab^x

Taking log on both sides we get

```
log(y) = log(a) + x*log(b) ----- (2)

Now let Y = log(y), A = log(a) and B = log(b)

then equation (2) becomes,

Y = A + Bx ----- (3),

Now we fit equation (3) using least square regression as:

1. Form normal equations:

    ∑Y = nA +  B ∑x
```

FITTING OF A CURVE $y=ax^b$

Taking log on both sides , we get

$\log(y) = \log(ab^x)$

$\log(y) = \log(a) + \log(b^x)$

$\log(y) = \log(a) + x*\log(b)$ ----- (2)

Now let $Y = \log(y)$, $A = \log(a)$ and $B = \log(b)$

then equation (2) becomes,

$Y = A + Bx$ ----- (3),

Now we fit equation (3) using least square regression as:

1. Form normal equations:

$$\sum Y = nA + B \sum x$$
$$\sum xY = A\sum x + B\sum x^2$$

2. Solve normal equations as simulataneous equations for A and B

3. We calculate afrom A and bfrom B as:
   a = exp(A)
   b = exp(B)

4. Substitute the value of a and b in $y= ab^x$ to find line of best fit.

FITTING OF A CURVE y=ae^bx

The method illustrated above can be used in the case of fitting of $y = ae^{bx}$ also. Taking logarithm on both sides, the curve becomes, $\log y = \log a + x \times b \log e$. Let $Y = \log y$; $A = \log a$ and $B = b \log e$, we get, $Y = A + Bx$ or $Y = Bx + A$. From the given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ values, taking the logarithm of y_i values Y_i values are obtained. Then use the following normal equations to obtain A and B.

$$\sum_{i=1}^{n} x_i Y_i = B \sum_{i=1}^{n} x_i^2 + A \sum_{i=1}^{n} x_i \quad \text{----- (1) and}$$

$$\sum_{i=1}^{n} Y_i = B \sum_{i=1}^{n} x_i + nA \quad \text{----- (2)}$$

Now, a is the antilogarithm of A and $b = \dfrac{B}{\log e}$.

KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl pearson (1867-1936) a British Biometrician, developed the coefficient of correlation to express the degree of linear relationship between two variables

Correlation co-efficient between two random variables X and Y denoted by $r(X, Y)$, is given by

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$ Where

$$Cov(X,Y) = \frac{1}{n}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$ (covariance between X and Y)

$$\sigma_X = \sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}$$ (standard deviation of X)

$$\sigma_Y = \sqrt{\frac{1}{n}\sum_i (Y_i - \bar{Y})^2}$$ (standard deviation of Y)

Theorem : For two variables x and y , $-1 \le r(x, y) \le +1$, where r(x ,y) is the Pearson's coefficient of correlation.

Proof

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observations on x and y.

Consider $\frac{(x_i - \bar{x})}{\sigma_x}$ and $\frac{(y_i - \bar{y})}{\sigma_y}$, where $\bar{x}$ and $\bar{y}$ are the means and σ_x and σ_y are the standard deviations of x and y respectively.

We have, $\left[\frac{(x_i - \bar{x})}{\sigma_x} \pm \frac{(y_i - \bar{y})}{\sigma_y}\right]^2 \ge 0$, because it is the square of a real number.

Adding all such terms for I = 1,2,...,n and dividing by n,

$$\Rightarrow \frac{1}{n}\sum_i \left[\frac{(x_i - \bar{x})}{\sigma_x} \pm \frac{(y_i - \bar{y})}{\sigma_y}\right]^2 \ge 0$$

On expansion,

$$\Rightarrow \frac{1}{n}\sum_i \left[\frac{(x_i - \bar{x})^2}{\sigma_x^2}\right] + \frac{1}{n}\sum_i \left[\frac{(y_i - \bar{y})^2}{\sigma_y^2}\right] \pm 2\frac{1}{n}\sum_i \frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} \ge 0$$

$$\Rightarrow \frac{1}{\sigma_x^2}\frac{1}{n}\sum_i [(x_i - \bar{x})^2] + \frac{1}{\sigma_y^2}\frac{1}{n}\sum_i [(y_i - \bar{y})^2] \pm 2\frac{1}{\sigma_x\sigma_y}\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \ge 0$$

$$\Rightarrow \frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_y^2} \pm 2\frac{Cov(x, y)}{\sigma_x\sigma_y} \ge 0.$$

That is, $1 + 1 \pm 2\frac{P_{xy}}{\sigma_x\sigma_y} \ge 0$

$\Rightarrow 2 \pm 2 r_{xy} \ge 0$. That is, $1 \pm r_{xy} \ge 0$

This gives, $1 + r_{xy} \ge 0$ or $1 - r_{xy} \ge 0$

That is $r_{xy} \ge -1$ or $r_{xy} \le 1$

$\Rightarrow -1 \le r_{xy} \le +1$

Theorem : (Invariance of correlation coefficient under linear transformation):A transformation on the variables x and y to u and v in the form u=(x-A)/C and v=(y-B)/D is making no change in the coefficient of correlation between the variables . That is r(x , y)=r(u , v)

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observations on x and y.

Then, $$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Let, $u = \frac{x - A}{C}$ and $v = \frac{y - B}{D}$;

Then, Pearson's coefficient of correlation between u and v,

$$r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - A}{C} - \left(\frac{\bar{x} - A}{C}\right)\right]\left[\frac{y_i - B}{D} - \left(\frac{\bar{y} - B}{D}\right)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - A}{C} - \left(\frac{\bar{x} - A}{C}\right)\right]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - B}{D} - \left(\frac{\bar{y} - B}{D}\right)\right]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{C}\right]\left[\frac{y_i - \bar{y}}{D}\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{C}\right]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{D}\right]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{CD} \times \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}][y_i - \bar{y}]}{\frac{1}{CD} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{CD} \times P_{xy}}{\frac{1}{CD} \times \sigma_x \sigma_y} = \frac{P_{xy}}{\sigma_x \sigma_y}$$

$$\Rightarrow r_{uv} = r_{xy}.$$

➢ REGRESSION LINES

For the pair of values of (X, Y), where X is an independent variable and Y is the dependent variable the line of regression of Y on X is given by

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

where b_{yx} is the regression co-efficient of Y on X and given by

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

where r is the correlation co-efficient between X and Y and σ_x and σ_y are the standard deviations of X and Y respectively

Similarly when Y is treated as an independent variable and X as dependent variable, the line of regression of X on Y is given by

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$ where b_{xy}

is the regression co-efficient of X on Y and given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$ Where $x = X - \bar{X},\ y = Y - \bar{Y}$

PROPERTIES OF REGRESSION COEFFICIENTS

i)The geometric mean between the regression coefficient is the correlation coefficient i.e

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

Note: sign of b_{xy} , b_{yx} **&** r are same always

Both the lines of reg. intersect at $(\bar{x}, \bar{y})$

ii)If one of the reg.coe. is greater than unity,the other must be less than unity

iii)Arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient

iv)Regression coef. Are independent of change of origin but dependent on change of scale.

ANGLE BETWEEN THE REGRESSION LINES

v)Angle between 2 reg .lines

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

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

Note:1. If r=0 then 2 var. are uncorrelated and lines of reg. are perpendicular to each other.

2. If $r = \pm 1$ then the 2 lines are parallel. But the lines intersect at $(\bar{x}, \bar{y})$ implies the lines must coincide

Theorem : The point of intersection of two regression lines is $(\bar{x}, \bar{y})$

We have regression equation y on x;

$$(y-\bar{y}) = r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \quad \text{---- (1)}$$

and the regression equation x on y;

$$(x-\bar{x}) = r_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \quad \text{---- (2)}$$

Put (2) in (1) gives,

$$(y-\bar{y}) = r_{xy}\frac{\sigma_y}{\sigma_x} r_{xy}\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

$$\Rightarrow (y-\bar{y}) = r_{xy}^2(y-\bar{y})$$

$$\Rightarrow (1-r_{xy}^2)y = (1-r_{xy}^2)\bar{y}$$

$$\Rightarrow y = \bar{y}$$

Put $y = \bar{y}$ in (2) $\Rightarrow (x-\bar{x}) = 0$

$$\Rightarrow x = \bar{x}$$

Hence the point of intersection of the regression lines is $(\bar{x}, \bar{y})$

IDENTIFICATION OF REGRESSION LINES AND DETERMINATION OF CORRELATION COEFFICIENT

- Let the two lines be $a_1x + b_1y + c_1=0$ and $a_2x + b_2y + c_2=0$.
- Assume the first to be regression line y on x and the second regression line x on y.
- Then the regression coefficient y on x is $-a_1/b_1$ and regression coefficient x on y is $-b_2/a_2$.
- If their geometric mean is less than or equal to one , then our assumption is correct . Otherwise , the first one is regression line x on y and the second regression line y on x
- Then regression coefficient y on x is$-a_2/b_2$ and regression coefficient x on y is $-b_1/a_1$.