Covariance And Correlation

Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

Covariance

Definition: Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

to : [Quantify the Relationship between X and Y]

X	y
→ 2	3
→ 4	5
→ 6	7
→ 8	9

weather it is:

X↑ y↑

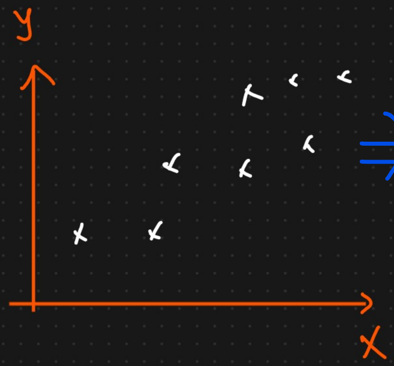
X↓ y↑

X↑ y↓

X↓ y↓

Dataset

↓ ↑ Size of House	Price ↑ ↓
1200	45 lakhs
1300	50 lakh
1500	75 lakh

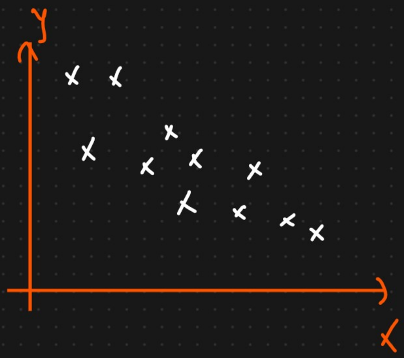


it can be seen :

X↑ y↑

X↓ y↓

⇒ +ve Covariance ⇒ +ve value



it can be seen that:

X↓ y↑

X↑ y↓

X	y
7	10
6	12
5	14
4	16

⇒ -ve Covariance ⇒ -ve value

Covariance of x to y :

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

thus , cov of x to x will be :

$$\Rightarrow \mathrm{Cov}(x,x) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(x_i-\bar{x})}{n-1}$$

$$\boxed{\mathrm{Cov}(x,x) = \mathrm{Var}(x)} \qquad \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$$

where :

$x_i \rightarrow$ Datapoint of random variable x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of random variable y

$\bar{y} \rightarrow$ Sample mean of y

thus the covariance of x to x
is variance of x

e.g :

Students

Hour Studied (X)	Exam Score (Y)
2	50
3	60
4	70
5	80
6	90

as , we can see that :

X↑ Y↑

X↓ Y↓

it is :

⇒ +ve Covariance

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

proof :

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$ //

② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$ //

$$Cov(X,Y) = \frac{(2-4)(50-70)+(3-4)(60-70)+(4-4)(70-70)+(5-4)(80-70)+(6-4)(90-70)}{4}$$

$$\boxed{Cov(X,Y) = \underline{\underline{20}}}.$$

⇒ The positive covariance indicates the no. of hours studied increand the exam store also.

$$\left\{\begin{array}{cc} X & Y \\ 7 & 10 \\ 6 & 12 \\ 5 & 14 \end{array}\right\} \Rightarrow$$

cov is : $\underline{\underline{-ve}}$.

X↑ Y↓

X↓ Y↑

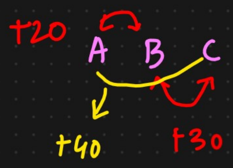
0.96	0.98	-200 +100 20	-300 +300 30
Cov(A,B)	Cov(B,C)	Cov(A,B)	Cov(B,C)

Advantages :

to - ① Quantify the Relationship between X and Y

|-1 to 1|

Disadvantage



① Covariance does not have a specific limit value.

↓

Cov(X,Y) ⇒ -∞ to ∞

② Correlation

- i. → Pearson Correlation Coefficient
- ii. → Spearman Rank Correlation

① Pearson Correlation Coefficient in : ⇒ [-1 to 1] always.

$$\boxed{\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}} = \frac{20}{\sigma_x . \sigma_y .} \Rightarrow 0 \text{ to } 1$$

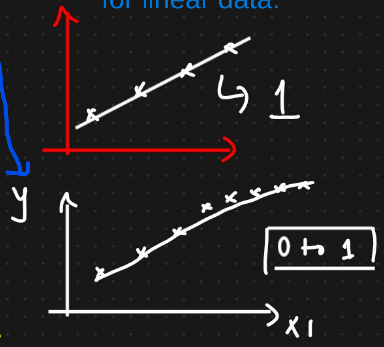
✓ The more the value towards +1 the more +ve correlated X & Y is.

✗ The more the value towards -1 the more -ve correlated it is (x,y)

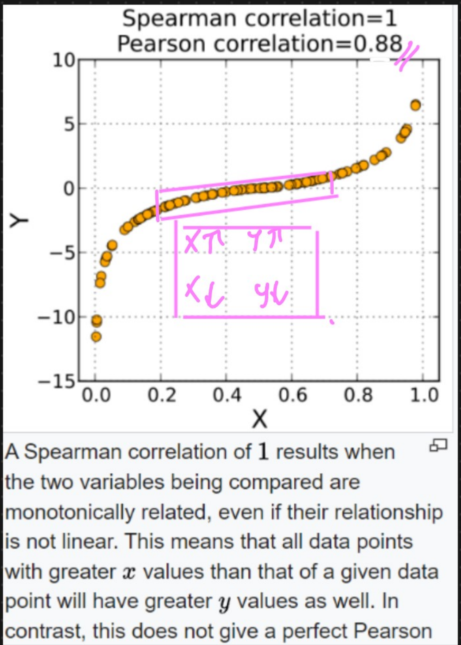
(ii) Spearman Rank Correlation

for linear data:

for non linear data:



Pearson correlation not able to capture the correlation for non linear data



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson

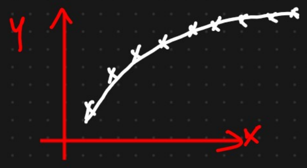
relationship :

⇓

⇒ X↑ Y↑
⇒ X↓ Y↓ }

⇓

Pearson Correlation = 0.88



x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

$$\boxed{r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}} \Leftarrow$$

R(x) means the rank of x , according to the values . eg. R(1) is 2 for x values.

these concepts are used for :

Feature Selection

+ve Correlated

Size of House ↑

No.of Room ↑

Location ↑

~~No.of People stays in the house~~

$\boxed{\approx 0}$

No

Haunted

o/p

Price ↑↑ vv

-ve Correlation

as there is norelation between this and the price is not possible

note :
t works well for non-linear but monotonic relationships.
If the relationship is non-monotonic, Spearman's rank correlation might not capture the relationship accurately.