Covariance And Correlation

13 October 2023 21:13

Covariance and correlation are essential tools in various fields, such as statistics, data science, machine learning, and data analysis. They serve as useful measures for determining the relationship between two variables.

These concepts are particularly significant in artificial intelligence and machine learning, as they are frequently employed in linear regression and neural networks to model and predict the relationship between variables.

However, they have different properties and may be used in different contexts depending on the research question and the data being analysed.

Understanding the relationship between two variables is a critical aspect of data analysis. Two commonly used measures of relationships are correlation and covariance.

While both provide useful insights into the relationship between two variables, they have distinct properties and may be used in different contexts.

Correlation and covariance are sensitive to outliers, so checking for outliers is important before calculating these measures.

While correlation measures the linear relationship between two variables, it may not capture the full extent of the relationship if it is not linear. In cases where the relationship is not linear, other measures, such as nonparametric correlation coefficients or nonlinear regression, may be more appropriate.

Sometimes, a high correlation coefficient may not necessarily imply causality between the two variables. The correlation only measures the association between two variables, and other factors may affect the relationship between the two variables.

Covariance and correlation can be calculated using different methods, such as raw data, deviations from the mean, or data ranks. The choice of method can affect the resulting correlation or covariance coefficient.

Example

Suppose we have two variables, X and Y, and we want to measure their relationship. We calculate the covariance and correlation coefficients and obtain the following results:

Covariance: 500
Correlation: 0.8

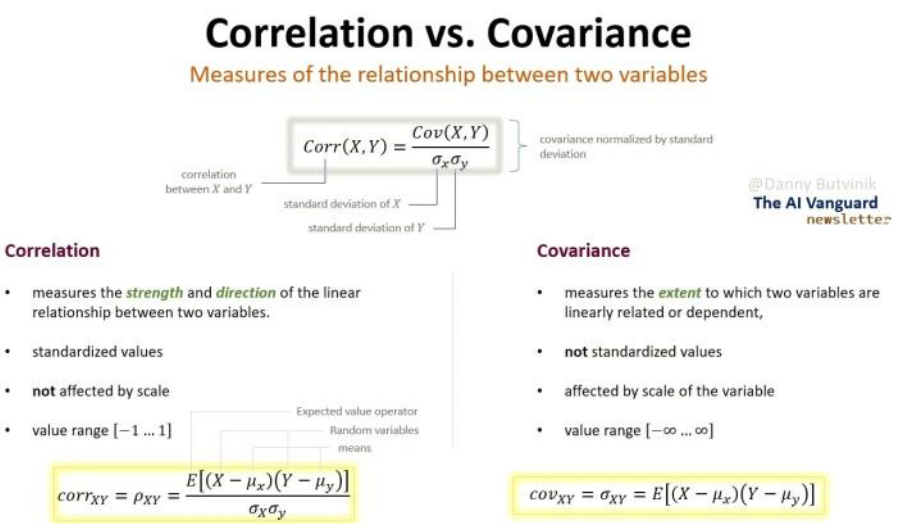
At first glance, X and Y appear to have a strong, positive relationship. However, upon further inspection, we find one outlier in the data driving the results. After removing the outlier, we recalculate the covariance and correlation coefficients and obtain the following results:

Covariance: 200
Correlation: 0.6

We can see that the correlation coefficient decreased, indicating a weaker relationship between X and Y, and the covariance decreased even more. This illustrates the importance of checking for outliers and the potential for spurious correlations when working with real-world data.

It's crucial to carefully examine the data and ensure that no outliers or other elements that might

skew the results are responsible for the findings.



..... it doesn't have relationship

Covariance: one of the very topic Data Preprocessing, quantity a Relation

2D vars: Dependent variable

Size	Price
1200 sqm	1000 $
1800 sqm	2000 $
1800 sqm	3000 $

Quantifying

S↑ P↑ +ve
S↓ P↓ +ve

Size ⇔ Price

$$\text{Covariance}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x - u_x) \times (y_i - u_y)$$

(x: Size, y: Price)

$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_1 - u_x)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x - u_x) \times (x - u_x)$$

$$\Rightarrow \text{Cov}(x, x) = \text{Var}(x)$$

Covariance is +ve

2D & 1D

Random variables

Means

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_1 - u_x) \times (y_1 - u_y)$$

x↑ y↑ = □ +ve

x↑ y↓ = -ve

In covariance we find direction of relationship but don't know how much the +ve (value)

We use Pearson cc

wk	M	T	W	T	F	S	S
40				1	2	3	4
41	5	6	7	8	9	10	11
42	12	13	14	15	16	17	18
43	19	20	21	22	23	24	25
44	26	27	28	29	30	31	

October 2020

~~correlation~~

→ Covariance

Pearson Correlation Coefficient → direction of relationship b/w x & y

9.00

+ve corr ↑ — ID 2 — y

(x↑ y↑) weight — Weight

(x↑ y↓)

-ve

10.00

11.00

12.00

1.00

Pearson CC = $\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

2.00

3.00

In covariance we dont know strength only direction of Relation

4.00

Range b/w $\rho(x,y) = \{-1, 1\}$

↓ we divide by variance so it ranges

5.00

6.00

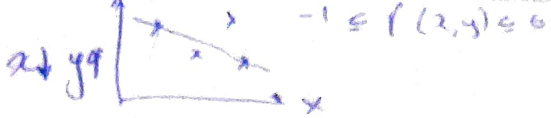
why? the -1 to 1 Because it measures the linear relationship b/w two variables

7.00

-1 = Perfect negative linear relationship b/w 2 variables. It means 1 variable increases the other decreases in perfectly constant manner

NOTES

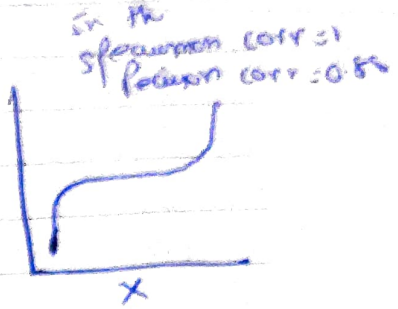
1 = Perfect +ve linear relationship between the two variables. This means 1 variable increases other also increases



$-1 \le f(x,y) \le 0$

0 → no linear relationship

SPEARMAN'S RANK CORELATION



$r_s = f r_{gx} ; r_{gy} = \frac{cov(r_{gx}, r_{gy})}{\sigma_{rgx} \sigma_{rgy}}$

* $r = r_{cont}$

First find rank

Steps

(1) Sort the data by the first column (x_i) create a new column and assign it the ranked value 1, 2, 3, ..

(2) Next Sort the data by the Second Column y_i and Similiarly assign it ranked Value 1, 2, 3, ..

(3) Create 5th column d_i to hold difference b/w two rank column (x_i and y_i)

(4) Create one final Column d_i^2 to hold the value of column d_i squared.

NOTES

example

	Iq (x)	Hours of TV per week	rank (xi)	rank (yi)	di	di^2
9.00						
10.00	86	0	1	1	0	0
	97	20	2	6	-4	16
11.00	99	28	3	8	-5	25
	100	27	4	7	-3	9
12.00	101	50	5	10	-5	25
	103	29	6	9	-3	9
1.00	106	7	7	3	4	16
	110	17	8	5	3	9
2.00	112	6	9	2	7	49
	113	12	10	4	9	31
3.00						

4.00 Spearman corr

5.00 $r_s = 1 - \frac{6 \Sigma di^2}{n(n^2-1)}$

6.00 $\rho = 1 - \frac{6 \times 194}{10(10^2-1)}$

7.00

11 Sunday $= \rho = \frac{-29}{165} = -0.175$ X↑ Y↓

relation b/w Iq & TV is low

NOTES

Advantage :- We can find relation with respect to Non linear Data.

Ans
The spearman corelation is less Sensitive than the pearson corelation to strong outlier that are in both Samples
That is because spearman's limits the outlier to the value of rank.

NOTES