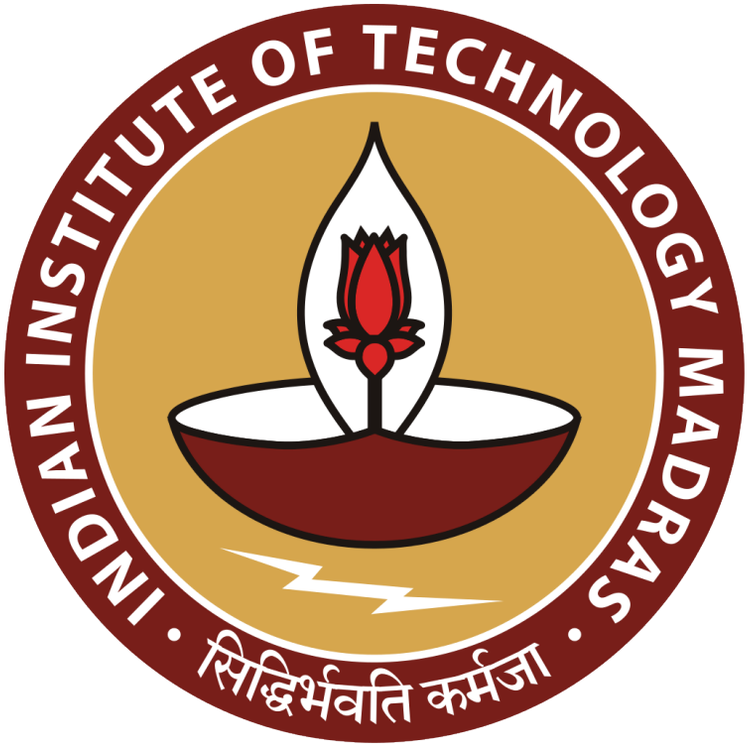

IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 4.6: Association between two numerical variables-Covariance

Usha Mohan

Indian Institute of Technology Madras

Measures of association

How do we measure the strength of association between two variables?

1. Covariance
2. Correlation

Covariance

Covariance quantifies the strength of the linear association between two numerical variables.

Covariance: Example 1

Recall, the association between age and height of a person.

Covariance: Example 1

Recall, the association between age and height of a person.

Age (years) x	Height (cms) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	75		
2	85		
3	94		
4	101		
5	108		
3	**92.6**		

Covariance: Example 1

Age (years) x	Height (cms) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	75	-2	-17.6
2	85	-1	-7.6
3	94	0	1.4
4	101	1	8.4
5	108	2	15.4

Covariance: Example 2

Variables: Age of a car and price of a car

Age (years) x	Price (INR lakhs) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	6		
2	5		
3	4		
4	3		
5	2		
3	**4**		

Covariance: Example 2

Variables: Age of a car and price of a car

Age (years) x	Price (INR lakhs) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	6	-2	2
2	5	-1	1
3	4	0	0
4	3	1	-1
5	2	2	-1
3	**4**		

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	
2	85	-1	-7.6	
3	94	0	1.4	
4	101	1	8.4	
5	108	2	15.4	

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8

Covariance: Example 2

Age x	Price y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4

Key observation

- When large (small) values of x tend to be associated with large (small) values of y- the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.
- When large (small) values of x tend to be associated with small (large) values of y- the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

Covariance

Definition

Let x_i denote the i^{th} observation of variable x, and y_i denote the i^{th} observation of variable y. Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

- *Population covariance:* $Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N}$
- *Sample covariance:* $Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
				82

- Population covariance: $\frac{82}{5} = 16.4$
- Sample covariance: $\frac{82}{4} = 20.5$

Covariance: Example 2

Age x	Price y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
				-10

- Population covariance: $\frac{-10}{5} = -2$
- Sample covariance: $\frac{-10}{4} = -2.5$

Units of Covariance

- The size of the covariance, however, is difficult to interpret because the covariance has units.

Units of Covariance

- The size of the covariance, however, is difficult to interpret because the covariance has units.
- The units of the covariance are those of the x-variable times those of the y-variable.

Section summary

1. Introduced the measure of covariance
2. How to interpret the covariance measure