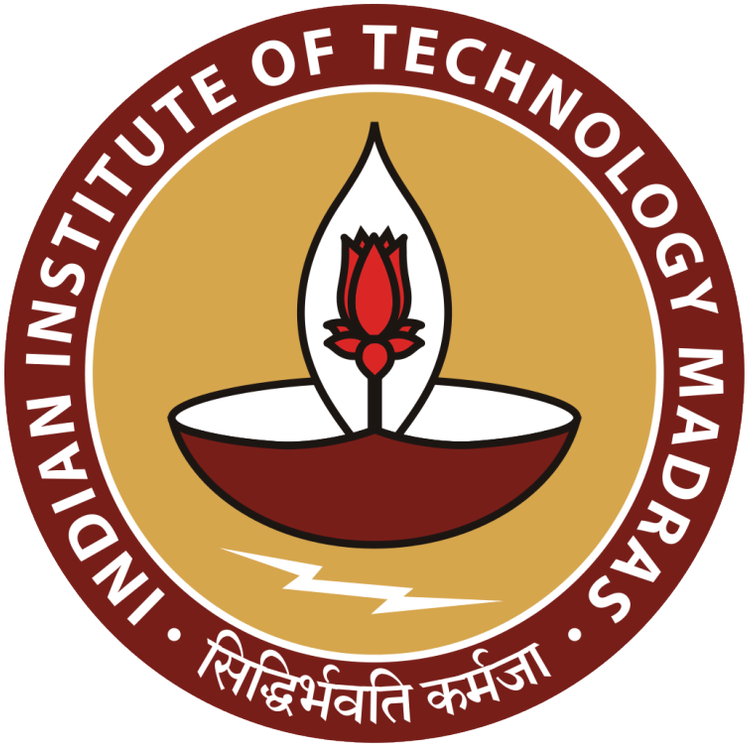


IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 4.7: Association between two numerical variables-Correlation

Usha Mohan

Indian Institute of Technology Madras

Learning objectives

1. Understand the measure of correlation.
2. Interpret correlation to quantify the strength of association between two numerical variables.

Correlation

Correlation

- A more easily intepreted measure of linear association between two numerical variables is correlation

Correlation

- A more easily intepreted measure of linear association between two numerical variables is correlation
- It is derived from covariance.
- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by

Correlation

- A more easily intepreted measure of linear association between two numerical variables is correlation
- It is derived from covariance.
- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by

$$r =$$

Correlation

- A more easily intepreted measure of linear association between two numerical variables is correlation
- It is derived from covariance.
- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by

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

Correlation

- A more easily intepreted measure of linear association between two numerical variables is correlation
- It is derived from covariance.
- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x, y)}{s_x s_y}$$

Remark

The units of the standard deviations cancel out the units of covariance

Remark

The units of the standard deviations cancel out the units of covariance

Remark

It can be shown that the correlation measure always lies between -1 and +1

Correlation: Example 1

Correlation: Example 1

Age x	Height y	sq.Devn of x $(x_i - \bar{x})^2$	sq.Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	4	309.76	35.2
2	85	1	57.76	7.6
3	94	0	1.96	0
4	101	1	70.56	8.4
5	108	4	237.16	30.8
		10	**677.2**	**82**

- $s_x = 1.58$, $s_y = 13.01$
- $r = \frac{82}{\sqrt{10 \times 677.2}}$ OR $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Correlation: Example 2

Age x	Price y	sq. Devn of x $(x_i - \bar{x})^2$	sq. Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	4	4	-4
2	5	1	1	-1
3	4	0	0	0
4	3	1	1	-1
5	2	4	4	-4
		10	**10**	**-10**

- $s_x = 1.58$, $s_y = 1.58$
- $r = \frac{-10}{\sqrt{10} \times \sqrt{10}}$ OR $\frac{-2.5}{1.58 \times 1.58} = -1$

Correlation using google sheets

Step 1 The function CORREL(series1, series2) will return the value of correlation.

For example: If the data corresponding to x-variable (series1) is in cell A2:A6 and data corresponding to y-variable (series2) is in cells B2:B6; then CORREL(A2:A6,B2:B6) returns the value of the Pearson Correlation coefficient.

Section summary

Section summary

1. Introduced measure of correlation.
2. Interpreting correlation between variables.

Learning objectives

1. Summarize the linear association between two variables using the equation of a line.
2. Understand the significance of R^2

Summarizing the association with a line

Summarizing the association with a line

- The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

Summarizing the association with a line

- The strength of linear association between the variables was measured using the measures of Covariance and Correlation.
- The linear association can be described using the equation of a line.

Equation of line using google sheets

Equation of line using google sheets

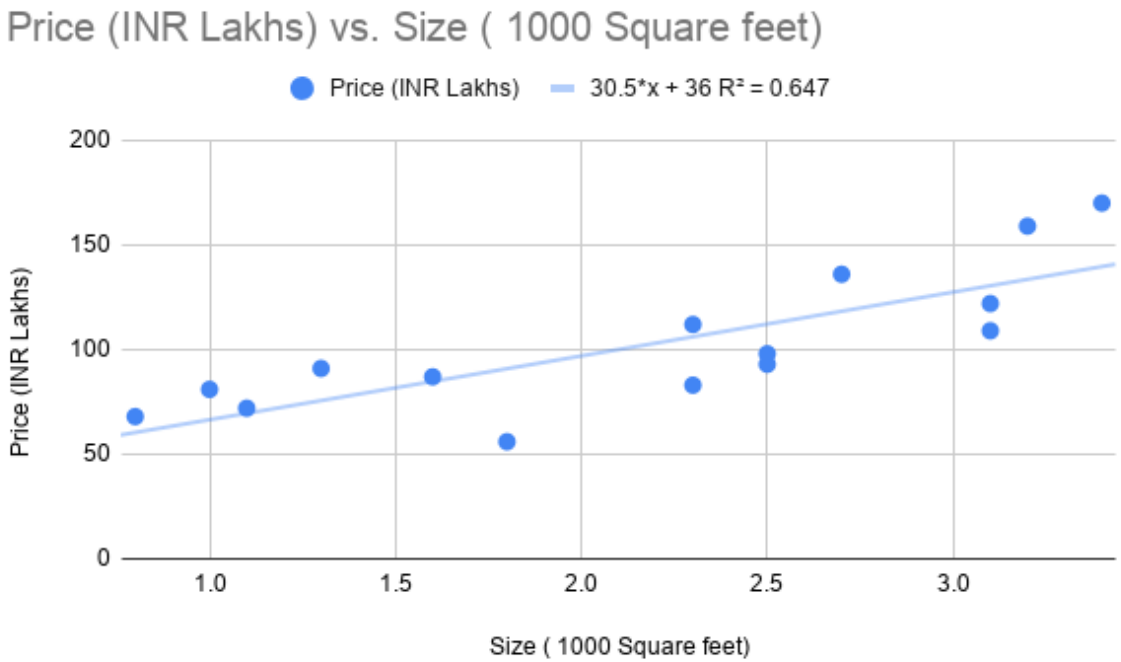
Step 1 Open the scatter plot

Step 2 Under customize tab, click on series

Step 3 Click on trendline

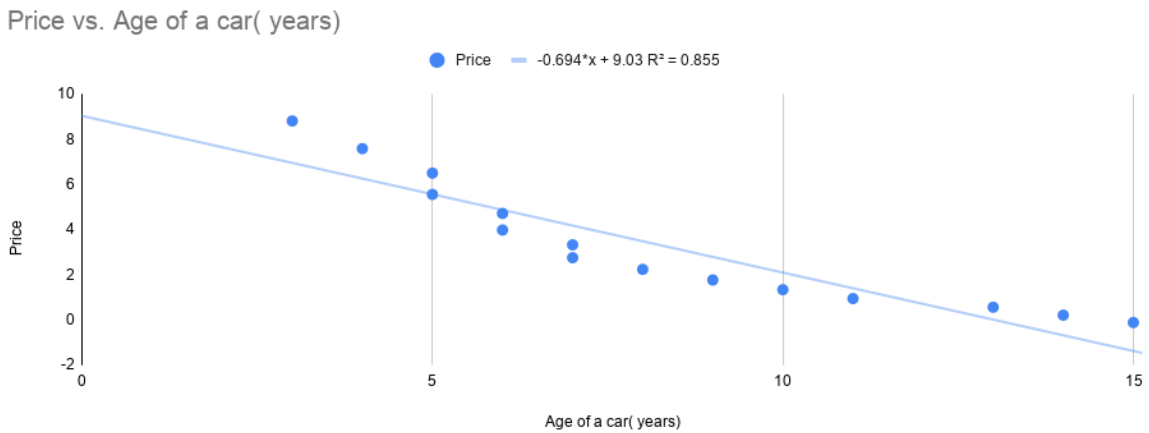
Step 4 Under label tab, click on use equation, and click the show R^2 button.

Example 1: Size versus Price of homes: Equation



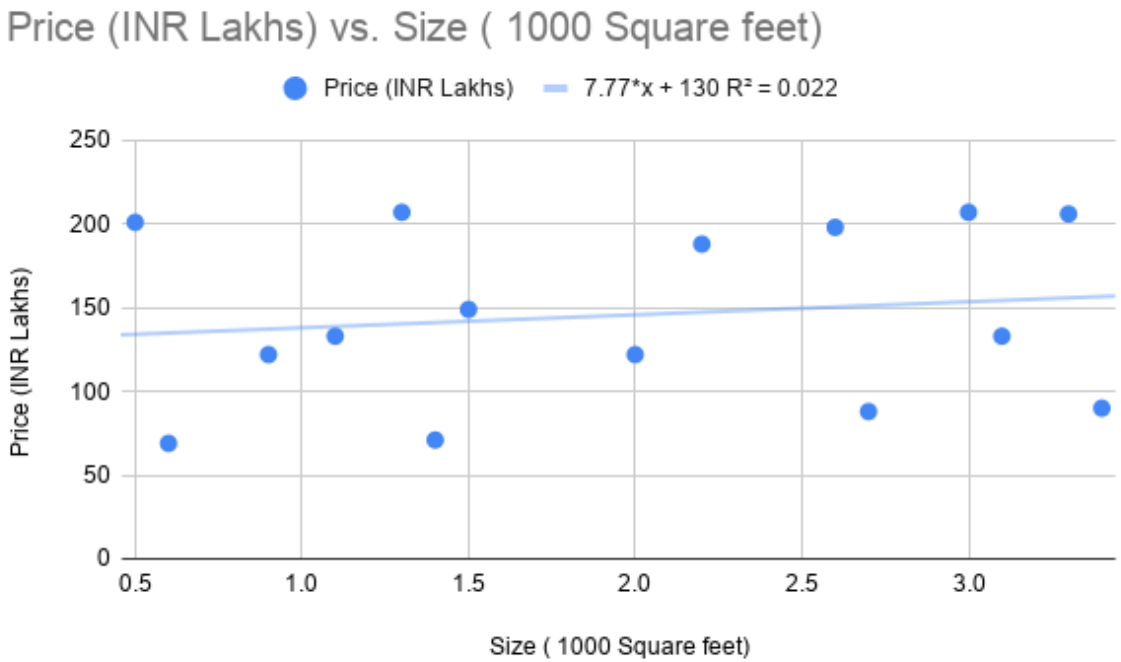
Equation of the line: $Price = 30.5 \times Size + 36$;
$R^2 = 0.647$; $r = 0.804$

Example 2: Age versus Price of cars: Equation



Equation of the line: $Price = -0.694 \times Age + 9.03$;
$R^2 = 0.855$; $r = -0.9247$

Example 3: Size versus Price of homes: Equation



Equation of the line: $Price = 7.77 \times Size + 130$;
$R^2 = 0.022$; $r = 0.149$

Section summary

1. Equation of a line describing linear relationship between two variables.
2. Interpreting slope, R^2 of the line.