

Correlation

Correlation

- Developed by **Francis Galton** in **1885**.
- A correlation is a **measure** of the **association** or the **relationship** that exists between **two variables**.
- It measures three characteristics of the relationship between two variables.

1. Direction of relationship (+ve or -ve)

2. Form of relationship (Linear or non-linear)

3. Degree of relationship(Value/Magnitude)

Definition of Correlation

- “Correlation is a **joint relationship** between two variables.”
(**Lathrop**)
- “Correlation is a relationship or **dependence**. It is the fact that two things or variables are so related **that change in one is accompanied by a corresponding** or parallel change in the other.” (**Garrett**)
- “A coefficient of correlation is a **single number** that tells us to **what extent two things are related**, to what extent variation in one go with variations in the other.” (**Guilford**)

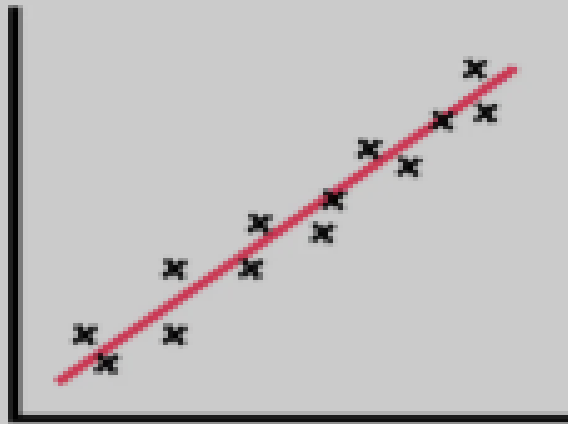
Correlation Coefficient

- **Correlation coefficient** gives **magnitude** and **direction** of the relationship.
- Correlation coefficient can vary between **-1.0 to +1.0**.

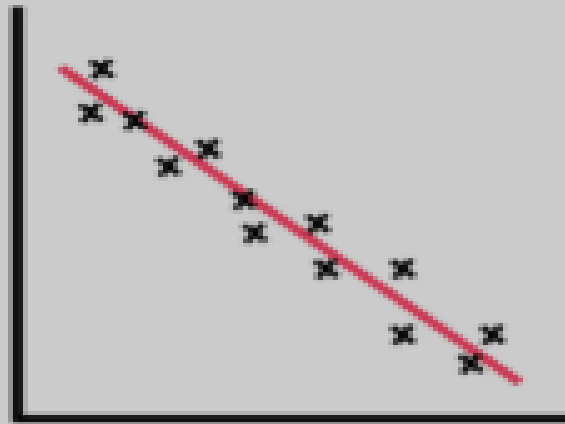


Types of Correlation

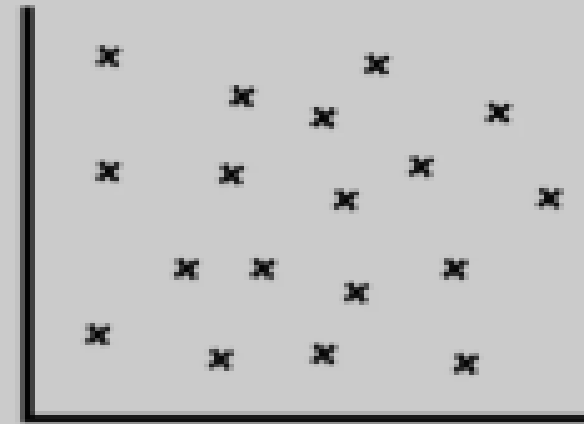
On the basis of Direction of relationship (+ve or -ve)



Positive
Correlation



Negative
Correlation



No
Correlation

Coefficient Range	Relationship
-0.91 a -1.00	Perfect negative correlation
-0.76 a -0.90	Very strong negative correlation
-0.51 a -0.75	Considerable negative correlation
-0.11 a -0.50	Moderate negative correlation
-0.01 a -0.10	Weak negative correlation
0.00	No correlation
+0.01 a +0.10	Weak positive correlation
+0.11 a +0.50	Moderate positive correlation
+0.51 a +0.75	Considerable positive correlation
+0.76 a +0.90	Very strong positive correlation
+0.91 a +1.00	Perfect positive correlation

Source: Own elaboration based on Hernández et al. 2014, Page 305

Pearson's Correlation Coefficient Formula

Pearson's Correlation Coefficient Formula is added below:

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$r_c = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

$$\text{Cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}}$$

$$r_c = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

How to Find Pearson's Correlation Coefficient?

Follow the steps added below to find the Pearson's Correlation Coefficient of any given data set.

Step 1: Firstly make a chart with the given data like subject, x , and y and add three more columns in it xy , x^2 and y^2 .

Step 2: Now multiply the x and y columns to fill the xy column. For example:- in x we have 24 and in y we have 65 so xy will be $24 \times 65 = 1560$.

Step 3: Now, take the square of the numbers in the x column and fill the x^2 column.

Step 4: Now, take the square of the numbers in the y column and fill the y^2 column.

Step 5: Now, add up all the values in the columns and put the result at the bottom. Greek letter sigma (Σ) is the short way of saying summation.

Step 6: Now, use the formula for Pearson's correlation coefficient:

Correlation Coefficient Formula Problems

Problem 1: Calculate the correlation coefficient from the following table:

SUBJECT	AGE (X)	GLUCOSE LEVEL (Y)
1	42	98
2	23	68
3	22	73
4	47	79
5	50	88
6	60	82

Solution:

Make a table from the given data and add three more columns of XY , X^2 , and Y^2 .

<i>SUBJECT</i>	<i>AGE (X)</i>	<i>GLUCOSE LEVEL (Y)</i>	<i>XY</i>	<i>X²</i>	<i>Y²</i>
1	42	98	4116	1764	9604
2	23	68	1564	529	4624
3	22	73	1606	484	5329
4	47	79	3713	2209	6241
5	50	88	4400	2500	7744
6	60	82	4980	3600	6724
Σ	244	488	20379	11086	40266

$$\Sigma xy = 20379$$

$$\Sigma x = 244$$

$$\Sigma y = 488$$

$$\Sigma x^2 = 11086$$

$$\Sigma y^2 = 40266$$

$$n = 6.$$

$$R = 6(20379) - (244)(488) / \sqrt{[6(11086) - (244)^2][6(40266) - (488)^2]}$$

$$R = 3202 / \sqrt{[6980][3452]}$$

$$R = 0.6439$$

$$R = 3202/4972.238$$

It shows that the relationship between the variables of the data is a strong positive relationship.

Find coefficient of correlation of the following data :

x	y
100	30
200	50
300	60
400	80
500	100
600	110
700	130

$$r = \frac{n \sum dx dy - \sum dx \cdot \sum dy}{\sqrt{n \sum dx^2 - (\sum dx)^2} \sqrt{n \sum dy^2 - (\sum dy)^2}}$$

$$\underline{n}, \sum dx, \sum dy, \sum dx dy, \sum dx^2, \sum dy^2$$

$$dx = \frac{x - A}{h}$$

$$dy = \frac{y - A}{h}$$

Middle of series for x,
A=400 and for y, A=80

For x, h=100 as there is gap of 100 in each entry whereas for y, h=10 all entries are divisible by 10 and it can be any no. which can divide all the entries.

x	y	dx	dy	dx ²	dy ²	dx dy
100	30	-3	-5	9	25	15
200	50	-2	-3	4	9	6
300	60	-1	-2	1	4	2
400	80	0	0	0	0	0
500	100	1	2	1	4	2
600	110	2	3	4	9	6
700	130	3	5	9	25	15
		<u>0</u>	<u>0</u>	<u>28</u>	<u>76</u>	<u>46</u>

$$r = \frac{7(46) - 0(0)}{\sqrt{7(28) - (0)^2} \sqrt{7(76) - (0)^2}}$$

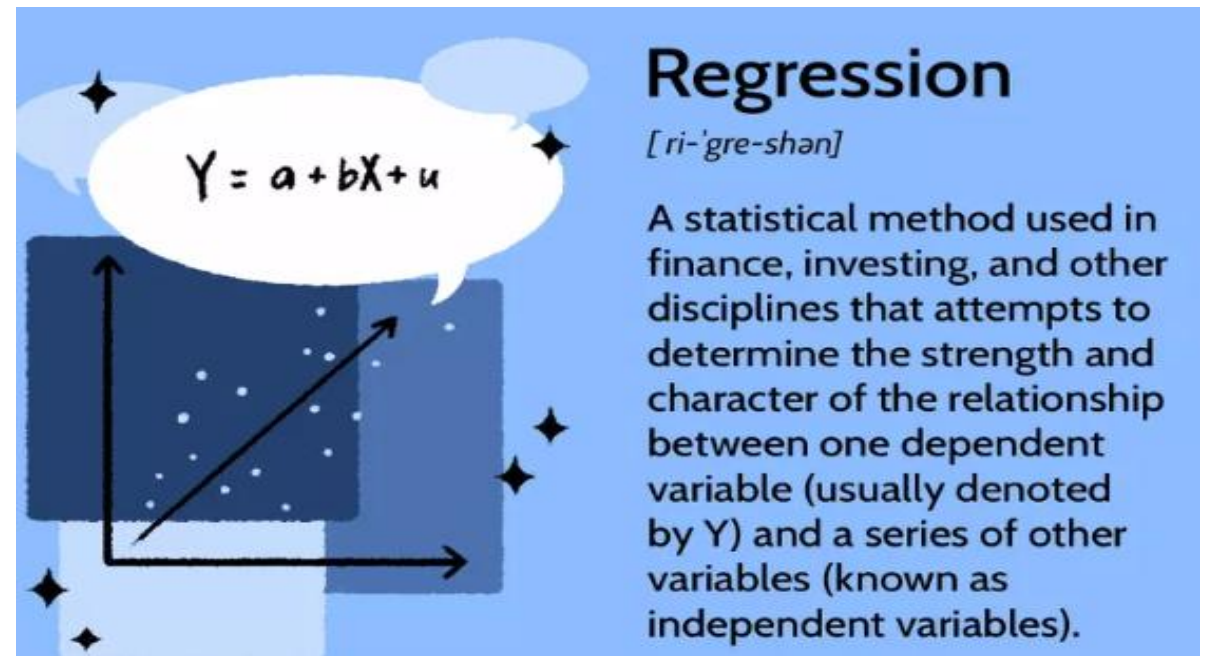
$$\begin{aligned} \frac{7(46)}{\sqrt{7(28)} \sqrt{7(76)}} &= \frac{46}{\sqrt{28 \times 76}} \\ &= \frac{46}{46.13} = \boxed{0.99} \end{aligned}$$

Regression

What Is Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between a dependent variable and one or more independent variables.

Linear regression is the most common form of this technique. Also called simple regression or ordinary least squares (OLS), linear regression establishes the linear relationship between two variables. Regression captures the correlation between variables observed in a data set and quantifies whether those correlations are statistically significant or not.



Regression

[ri-'gre-shən]

A statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Simple linear regression: $Y = a + bX + u$

Multiple linear regression:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_tX_t + u$$

where: Y = The dependent variable you are trying to predictor explain

X = The explanatory (independent) variable(s) you are using to predict or associate with Y

a = The y intercept

b = (beta coefficient) is the slope of the explanatory variable(s)

u = The regression residual or error term

Discriminant analysis

- One dependent variable (nominal)
- One or more independent variable(s) (interval or ratio)

Formula for linear regression equation is given by:

$$y = a + bx$$

a and b are given by the following formulas:

$$a \text{ (intercept)} = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b \text{ (slope)} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Where,

x and y are two variables on the regression line.

b = Slope of the line.

a = y -intercept of the line.

x = Values of the first data set.

y = Values of the second data set.

Question: Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Solution:

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x$ = 20	$\sum y$ = 25	$\sum x^2$ = 120	$\sum xy$ = 144

$$b$$

$$=$$

$$\frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$b$$

$$=$$

$$\frac{4 \times 144 - 20 \times 25}{4 \times 120 - 400}$$

$$b = 0.95$$

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{25 \times 120 - 20 \times 144}{4(120) - 400}$$

$$a = 1.5$$

Linear regression is given by:

$$y = a + bx$$

$$y = 1.5 + 0.95x$$

Correlation	Regression
1. It indicates only the nature and extent of linear relationship	It is the study about the impact of the independent variable on the dependent variable. It is used for predictions.
2. If the linear correlation coefficient is positive / negative, then the two variables are positively / or negatively correlated	The regression coefficient is positive, then for every unit increase in x , the corresponding average increase in y is b_{yx} . Similarly, if the regression coefficient is negative, then for every unit increase in x , the corresponding average decrease in y is b_{yx} .
3. One of the variables can be taken as x and the other one can be taken as the variable y .	Care must be taken for the choice of independent variable and dependent variable. We can not assign arbitrarily x as independent variable and y as dependent variable.
4. It is symmetric in x and y , i.e., $r_{xy}=r_{yx}$	It is not symmetric in x and y , that is, b_{xy} and b_{yx} have different meaning and interpretations.