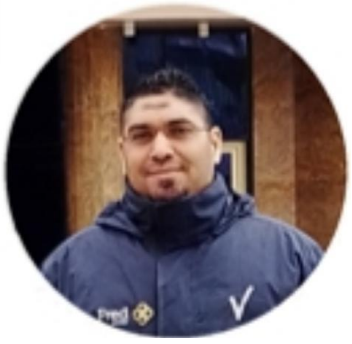
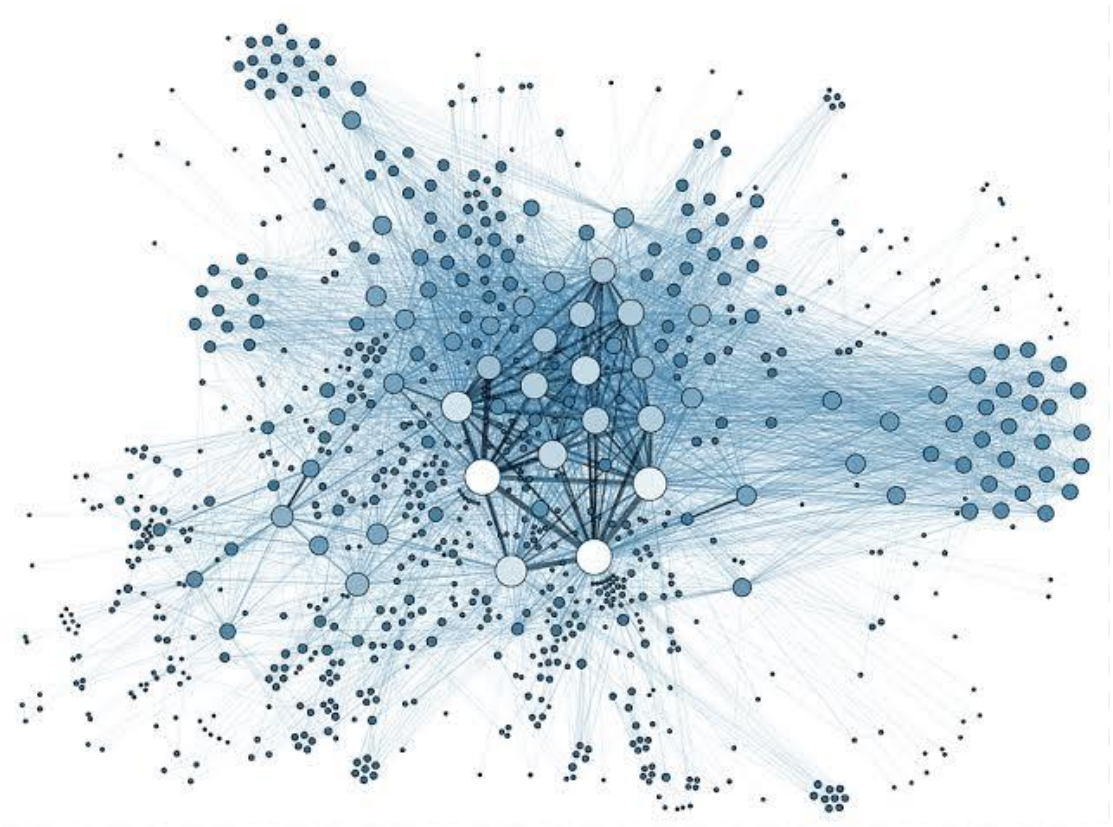


Statistics for Big Data

Inferential statistics

Correlation analysis 2

Part 17



Correlation analysis

2. Rank correlation

Rank correlation is used to study the relationship between variables Probability Random variables when the data of these factors Arranged or can be arranged in certain ranks. In this form is calculated Correlation factor (r) using the following relationship:

$$r = 6 * \sum (x_i - y_i)^2 / n(n^2 - 1)$$

Where:

yi. Rank the first factor or the first quality

Xi Rank the second factor or the second quality

Example

We have six students who have the following degrees in statistics and economics:

Student	Statistics	Economics
1	52	46
2	34	57
3	64	53
4	57	26
5	57	43
6	36	56

Solution

We sort the data in ascending order, and we get the following table:

Statistics ranks	Statistics	Economics ranks	Economics
6	34	6	26
5	36	5	43
4	52	4	46
2.5	57	3	53
2.5	57	2	56
1	64	1	57

As for the subject of statistics, we note that it has the highest value, so the first rank was given, while the score 57 is repeated twice, so its rank is located in the middle between the two ranks, i.e. 2.5 = 2 / (2 + 3) and so the ranks are completed for the rest of the grades .

- **Now we rewrite the data as follows**

Student	Statistics ranks	Statistics	Economics ranks	Economics
1	6	34	6	26
2	5	36	5	43
3	4	52	4	46
4	2.5	57	3	53
5	2.5	57	2	56
6	1	64	1	57

We apply the above relationship to calculate the correlation factor r to get the following .

$$r = 1 - 6 * \Sigma (x_i - y_i) / n(n^2 - 1)$$

$$= 1 - 6*(56.5) / 6*(36 - 1)$$

$$= 1 - (56.5 / 35)$$

$$= 1 - 1.61$$

$$= - 0.61$$

We note from the above result that the correlation factor of - 0.61 = r, which means that r Correlation is negative but strong. In other words, the student's ranking in statistics Contrary to his order in economics, this leads us to say that if the student's mark in The subject of statistics is high, its mark will be in the subject of economics low and vice versa Correct.

3. Partial correlation

Partial correlation is used to study the relationship between random variables when there are several variables and we want to calculate the correlation between two variables or two specific characteristics, Note that the correlation with the rest of the variables is known, so if we have several variables X1, x2,.... Xn and we wanted to calculate the correlation between two specific variables or traits X1 and X2 knowing that the correlation with the rest of the variables X3,....Xn is known; Then we call the desired correlation the partial correlation, and to calculate the partial correlation we use the following general relation:

$$r_{12,3} = r_{12} - r_{13} * r_{23} / \sqrt{(1 - r_{13}^{\ 2}) * (1 - r_{23}^{\ 2})}$$

where

r12,3. Partial correlation between two specific variables or traits x1 and x2

r12. Simple correlation factor between two variables or traits x1 and x2

r13. Simple correlation factor between two variables or traits X1 and x2

r23. Simple correlation factor between two variables or traits x2 and x3

It is worth saying that the partial correlation is a retroactive relationship, that is, to calculate the partial correlation of two other known factors, we need to calculate the required Simple correlation factors.

Example

When studying the relationship between the characteristics of the number of grains per spike, the number of spikes per plant, and the production characteristic of one of the crops, which is high-yielding wheat, we have the following:

Number of spike in a plant	The number of grains in the spikes	Production
X2	X1	X3
3	9	5
2	7	13
6	5	16
5	3	23
4	1	33

We apply the above relationship to calculate the partial correlation factors $(r_{12,3}, r_{13,2}, r_{23,1})$ **and we get the following:**

	X2	*X1*	*X3*
X2	1		
X1	-0.5	1	
X3	0.358568583	0.986063603	1

From the previous table above we find that:

-0.5 = r12,3 partial correlation between the two variables X1 * X2

0.986063603-= r13,2 partial correlation between variables X1 * X3

0.358568583 = r12,3 partial correlation between the two variables X3 * X2

It is worth saying here also that the partial correlation (and simple correlation) between any variable and itself equal to one as given in the previous table above.

4. Multiple Correlation

Multiple Correlation is used to study the relationship between random variables when there are several variables and we want to calculate the correlation between a specific variable or characteristic X1 with the rest of the X2, X3, ... Xn variables. So if we have several X1 variables, X2,....Xn We wanted to calculate the correlation between a specific variable or characteristic X1 with the rest of the variables Xn....,3; Then we call the desired association the multiple association, and to calculate the multiple association we use the following general relation:

$$r_{1,23} = \sqrt{r_{12}^{\ 2} + r_{13}^{\ 2}} - 2 * r_{12} * r_{13} * r_{23} / 1 - r^2_{23}$$

Example

We use the previous example, which is to study the relationship between the characteristics of the number of grains in one ear, the number of ears per plant, and the production characteristic of one of the crops, which is high-yielding wheat, where we have the following:

Number of spike in a plant	The number of grains in the spikes	Production
X2	X1	X3
3	9	5
2	7	13
6	5	16
5	3	23
4	1	33

We apply the above-mentioned relationship to calculate the multiple correlation coefficient (r1.23), and we get the results presented in the following table:

Correlation Statistics

Multiple R	0.998213
R Square	0.996429
Adjusted R Square	0.992857
Standard Error	0.894427
Observations	5

As it appears from the previous table above, we find that:

0.998213 = r1,23 the multiple correlation between the variable X1 and the two variables X2,3

And this indicates the existence of a very strong positive multiple regression.

5. Determination factor

The determination factor is one of the very important statistical indicators, as it provides us with the percentage of the influence of one of the factors (variable or characteristic) or several factors on one factor when there is multiple correlation, and the determination factor is mathematically a square of the correlation factor, meaning that:

Where $$B = r^2$$

B is the determination operator

r. is the correlation operator

Example

We also use the previous example, which is to study the relationship between the characteristic of the number of grains per spike, the number of spikes per plant, and the production characteristic of one of the crops, which is high-yielding wheat. After analyzing the statistical data, we obtained the following results:

Correlation Statistics

Multiple R	0.998213
R Square	0.996429
Standard Error	0.894427
Observations	5

As it appears from the previous table above, we find that the determination factor is 99.64% = B, and this means that both the factor of the number of grains / spike and the number of spikes / plant affect the production by 99.64%, while the remaining percentage, which is 0.36% of the effect, is due to other factors Unknown.

In Part 18, time series will be explained, how to use them, and their laws. Wait for this interesting part

Mohamed abulibdah

LinkedIn

Statistician and Data Analyst
at the NCI - Cairo University -
Egypt