ANALYSIS OF PEARSON'S LINEAR CORRELATION COEFFICIENT WITH THE USE OF NUMERICAL EXAMPLES

Katarzyna Brożek – Justyna Kogut

Abstract

This paper is a kind of repetition and consolidation of the knowledge of a certain segment of descriptive statistics, namely it concerns the Pearson linear correlation coefficient. The general objective of the study was to investigate and to determine the direction and strength of dependence between the analyzed features. The two selected numerical examples were used. In addition, the work looks at the results obtained from own calculations. The theoretical nature of the article is negligible, therefore the most important issues that are the subject of this work were briefly discussed. In contrast, much more attention has been paid to the empirical part, which is the culmination of the article. Considerations are addressed especially to students of economics but not strictly statistical, because the analysis it is not comprehensive, and at the same time thematic scope is largely limited. Therefore, the presented article should be considered as auxiliary material to study this issue. However, acquired knowledge that comes from the analysis should be strengthened by the information contained in the basic books positions or further reading relating to the descriptive statistics, and more precisely, to the Pearson correlation coefficient.

Key words: analysis, descriptive statistics, Pearson linear correlation coefficient

JEL Code: C1, C10

Introduction

In today's world all kinds of phenomena, for example economic ones (more Krištofík et al., 2015, pp. 189-197), social or natural are almost always conditioned by the action of other phenomena. Therefore the existence of relationships between the phenomena is often the subject of scientific inquiry. Then it turns out to be indispensable to use right the correlation coefficient (see Bedrick, 1991, pp. 369-378).

Assuming that the interdependence of the examined random variables X and Y is statistically significant, there are four basic types of measures the strength of these variables correlation (Sobczyk, 2007, p. 232):

1. Czuprow's convergence coefficient,
2. Pearson's correlation relations (indicators),
3. The coefficient of Pearson's linear correlation,
4. Spearman's rank correlation coefficient.

Pearson's linear correlation coefficient is generally smaller than the correlation relations. Equality between the correlation coefficient and the correlation relationships occurs only when the regression is linear (Sobczyk, 2007, p. 239).

Pearson's linear correlation coefficient (symbol r_{xy}) was created by Karl Pearson, it is used to study the strength of the rectilinear relationship between the two measurable characteristics (see Borroni, 2009, pp. 81-95). Rectilinear relationship is the kind of dependence in which the individual increments of one variable - the reason, is accompanied by, on average, steady growth of the second variable - effect (Sobczyk, 2007, p. 237). It can therefore be assumed that it is a factor determining the level of the linear dependence between random variables (Maesono, 2010, p. 1344).

It is worth noting that in contrast to the ratio of the correlation, the coefficient of correlation is a symmetric measure, i.e. it measures simultaneously the strength of Y dependence with respect to variable X, and vice versa - the strength of the variable X dependence with respect to variable Y, that is why $r_{xy} = r_{yx}$ (Buga, 1999, pp. 84 -85).

In general, the linear correlation coefficient of two variables is the ratio of the covariance and the product of the standard deviations of these variables (Klonecki, 1999, p. 117):

$$r_{x,y} = \frac{\operatorname{cov}(x, y)}{S_x S_y} \tag{1}$$

In contrast, to calculate the covariance, the following formula is helpful:

$$\operatorname{cov}_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N} \tag{2}$$

The correlation coefficient value is between the closed interval <-1.1>. It should be noted that the higher the absolute value is, the stronger the linear relationship between the variables. It is worth noting that the square of the correlation coefficient is the coefficient of determination, which informs about the extent to which changes of one variable are explained by changes in other one.

In the literature there are two basic types of correlation, namely, positive and negative. The first occurs when both variables increase in the same direction or when both variables decrease $r > 0$. In contrast, when one variable increases, and the other one decreases $r < 0$, or vice versa (the first one decreases and the other one increases), then we can talk about negative correlation. Thus, if the correlation coefficient is positive, the correlation is positive and by analogy, vice versa. Correlation relationship may be rectilinear (in short - linear) or curvilinear; it should be borne in mind that the coefficient r measures only the correlation of a rectilinear type.

Correlations can be interpreted also as strong or weak. Nevertheless, it is worth noting that such an interpretation is arbitrary and should not be taken too strictly and clearly, as this might lead to errors of measurement (Nielsen, 2016, pp. 169-200). Well, for individual representatives of the divergent sciences, it can mean something completely different. It is worth recalling an example of sociologists and physicists, and so even a factor of 0.9 for the first group will mean a very strong correlation, and for people who use high-quality measurements (physicists) it can mean a weak correlation (see Nelsen, 1998, pp. 343 -345).

Interpretation of the results is as follows (Aczel, 2000, pp. 479-480):

- $r_{xy} = 0$ –means there is no linear relationship between variables; lack of linear correlation (when x increases, then y sometimes increases and sometimes decreases), but it can then be curvilinear correlation,
- $r_{xy} = 1$ – means an exact positive linear relationship between variables,
- $r_{xy} = -1$ – means an exact negative linear relationship between the variables (i.e. if the variable x increases, then y decreases, and vice versa).

In order to provide details concerning the correct interpretation of the resulting coefficient of correlation, the following ranges of correlation strength might also help to solve the problem:

- $r_{xy} < 0,2$ – no linear relationship,
- 0,2 - 0,4 – linear relationship clear but low,
- 0,4 – 0,7 – moderate,
- 0,7 – 0,9 – significant (strong),
- > 0,9 – very strong.

1 Pearson's correlation study

To analyze, it was decided to recall the sample numerical task here. The following are the contents of a hypothetical task.

Examine the relationship between the annual meat consumption per person and the age of the consumer. Using the correlation coefficient determine if there is a relationship between these two variables. If so, designate the direction and strength of the relationship (other examples Rabiej, 2012, pp. 219-221).

Based on the data presented in Table 1, the correlation of both analyzed variables should be estimated.

Tab. 1: Consumption of meat and the age of consumers

$\mathbf{y_i}$	75	70	62	68	64	52	38	31
$\mathbf{x_i}$	25	32	47	54	58	53	67	68

where:

y_i - the number of meat in kg

x_i – age of consumers (years)

To calculate the correlation coefficient, y, x, Sy, Sx and the value of covariance should be calculated in the first place. In Table 2, the auxiliary parameters are presented and calculated.

Tab. 2: Support table

No of meat in kg	**Age of consumers**	$\mathbf{y_i - \bar{y}}$	$\mathbf{x_i - \bar{x}}$	$\mathbf{(y_i-\bar{y})(x_i-\bar{x})}$	$\mathbf{(x_i - \bar{x})^2}$	$\mathbf{(y_i-\bar{y})^2}$
$\mathbf{y_i}$	$\mathbf{x_i}$					
75	25	17,5	-28	-490	784	306,25
70	32	12,5	-21	-262,5	441	156,25
62	47	4,5	-6	-27	36	20,25
68	54	10,5	1	10,5	1	110,25
64	58	6,5	5	32,5	25	42,25
52	63	-5,5	10	-55	100	30,25
38	67	-19,5	14	-273	196	380,25
31	78	-26,5	25	-662,5	625	702,25
∑ **460**	∑ **424**	**X**	**X**	**-1727**	**2208**	**1748**

Source: own calculations

Average values

$$\bar{y} = \frac{\sum y_i}{N} = \frac{460}{80} = 57{,}5 \tag{3}$$

$$\bar{x} = \frac{\sum x_i}{N} = \frac{424}{8} = 53 \tag{4}$$

Standard deviations

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N}} = \sqrt{\frac{1748}{8}} = \sqrt{218{,}5} = 14{,}78 \tag{5}$$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} = \sqrt{\frac{2208}{8}} = \sqrt{276} = 16{,}61 \tag{6}$$

Covariance is regarded as a measure of compatibility of two random variables (more Stanisz, 2007, pp. 479-487). When the covariance is large and positive, x and y have a tendency to take both the large and small values at the same time. When the covariance is large in an absolute value, but negative, then x has a tendency to take large values when y is small and vice versa (Gajek, Kaluszka, 2000, p. 64).

$$\operatorname{cov}(x, y) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{N} = \frac{-1727}{8} = -215{,}875 \tag{7}$$

Correlation coefficient

$$r_{yx} = \frac{\operatorname{cov}(y, x)}{S_y * S_x} = \frac{-215{,}875}{14{,}78 * 16{,}61} = -0{,}88 \tag{8}$$

Results of the analysis

In the presented example the relationships between the annual consumption of meat for 1 person and the age of the consumer were shown. Correlation coefficient was used, which showed a negative linear relationship between the two tested variables. The correlation coefficient was -0.88, which means that if the number of consumed meat increases, the age of the consumers decreases and vice versa, when the variable x decreases then y increases.

1.1 Analysis of the correlation coefficient - the actual data

After the analysis of the correlation index for imaginary numerical example, it is worth repeating the analysis, but now on actual data concerning the two selected indicators of the

Polish economy. It was decided to examine the strength of the relationship between the level of GDP and export in the years 2007-2014 (Table 3).

Tab. 3: GDP and export in Poland in ECU / EUR billion in the years 2007-2014

	2007	2008	2009	2010	2011	2012	2013	2014
y_i	313,7	363,7	314,7	361,7	380,2	389,3	394,6	410,8
x_i	121,8	139,4	118,2	144,8	161,7	172,9	182,7	194,9

Source: AMECO Database, 2016.

where:

y_i - GDP at current prices in ECU / EUR billion

x_i - Export of goods and services at current prices (national accounts) in ECU / EUR billion

Again, in order to calculate the correlation index, Table 4, which contains all the necessary calculation, was created below.

Tab. 4: Support table

GDP ECU/EUR billion Y_i	Export ECU/EUR billion X_i	$Y_i - \bar{Y}$	$X_i - \bar{X}_i$	$(y_i-\bar{y})(x_i-\bar{x})$	$(X_i - \bar{X})^2$	$(y_i-\bar{y})^2$
313,7	121,8	-52,3875	-32,75	1715,690625	1072,5625	2744,45016
363,7	139,4	-2,3875	-15,15	36,170625	229,5225	5,70015625
314,7	118,2	-51,3875	-36,35	1867,935625	1321,3225	2640,67516
361,7	144,8	-4,3875	-9,75	42,778125	95,0625	19,2501562
380,2	161,7	14,1125	7,15	100,904375	51,1225	199,162656
389,3	172,9	23,2125	18,35	425,949375	336,7225	538,820156
394,6	182,7	28,5125	28,15	802,626875	792,4225	812,962656
410,8	194,9	44,7125	40,35	1804,149375	1628,1225	1999,20766
∑ **2928,7**	∑ **1236,4**	**X**	**X**	**6796,205**	**5526,86**	**8960,229**

Source: own calculations

Average values

$$\bar{y} = \frac{\sum y_i}{N} = \frac{2928,7}{8} = 366,0875 \tag{9}$$

$$\bar{x} = \frac{\sum x_i}{N} = \frac{1236,4}{8} = 154,55 \tag{10}$$

Standard deviations

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N}} = \sqrt{\frac{8960,229}{8}} = \sqrt{1120,03} = 33,47 \quad (11)$$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} = \sqrt{\frac{5526,86}{8}} = \sqrt{690,86} = 26,28 \quad (12)$$

Covariance

$$\operatorname{cov}(x, y) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{N} = \frac{6796,205}{8} = 849,526 \quad (13)$$

Correlation coefficient

$$r_{yx} = \frac{\operatorname{cov}(y, x)}{S_y * S_x} = \frac{849,526}{33,47 * 26,28} = -0,97 \quad (14)$$

Conclusions

In the presented example we showed actual relations that take place in Poland between the GDP level and the size of export in the years 2007-2014. Correlation coefficient was used, which showed the strength of the relationship between the above two variables. The correlation coefficient is 0.97 and so $r_{yx} > 0.9$, therefore, there is a very strong correlation. That is to say, with GDP growth the level of export rises, and vice versa. In view of the fact that both variables are increasing / decreasing in the same direction, it is a positive correlation.

Conclusion

Descriptive statistics in its scope, available methods and tools, tries to solve many afflicting problems. One of them is even the answer to the question about the relationships between one variable and another. Undoubtedly, an essential tool for the calculation is Pearson's linear correlation coefficient. It is worth noting that on the basis of the observation of dispersion period, we can roughly determined the nature of the relationship and its strength. The strength of the interdependence of two variables can be quantified using a wide variety of gauges. However, the choice of the gauge depends, for example, on the type of variables, among which the relationship is tested (measurable, immeasurable, mixed), the number of observations (correlation table, correlation ranks), and finally the shape of dependence (rectilinear or curvilinear regression).

To sum up, undoubtedly, this factor is extremely useful in many fields of science, so it is widely used. However, it is important to take into consideration the correct interpretation, since it is the key to any discussion.

References

Aczel, A. D. (2000). *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warsaw, pp. 479-480. ISBN 978-83-01-15338-0.

AMECO Database (2016). Retrieved January 20, 2016, from http://www.ec.europa.eu.

Bedrick, E. J. (1999). *Approximate confidence-intervals for the correlation from data in 2-by-2 tables.* British Journal of Mathematical & Statistical Psychology, Vol. 44, Publisher British Psychological SOC, England, pp. 369-378. ISSN 0007-1102.

Borroni, C. G. (2009). *Understanding Karl Pearson's Influence on Italian Statistics in the Early 20th Century*. International Statistical Review, Vol. 77, Issue 1, Publisher INT Statistical INST, Netherlands, pp. 81-95. ISSN 0306-7734.

Buga, J. ed. (1999). *Statystyka opisowa w przykładach*, Copyright by Politechnika Radomska, Radom: 1999, 84-85pp. ISSN 0860-9241.

Gajek, L., Kałuszka, M. (2000). *Wnioskowanie statystyczne. Modele i metody*. Wydawnictwa Naukowo-Techniczne, Warsaw, pp. 63-65. ISBN 83-204-2489-5.

Klonecki, W. (1999). *Statystyka dla inżynierów*. Wydawnictwo Naukowe PWN, Warsaw, p. 117. ISBN 83-01-12754-6.

Kristofik, P., Lament, M., Musa, H., Wolak-Tuzimek, A. (2015). *Financial tools in management of small and medium-sized enterprises*, Sciemcee Publishing, London, pp. 189-197. ISBN 978-0-9928772-7-9.

MAesono, Y. (2010). *Edgeworth Expansion and Normalizing Transformation of Ratio Statistics and Their Application*. Communications in Statistics-Theory and Methods, Vol. 39, Issue 8-9, Publisher Taylor & Francis INC, Philadelphia, p. 1344. ISSN 0361-0926.

Nelsen, R. B. (1998). *Correlation, regression lines, and moments of inertia*. American Statistician, Vol. 52, Issue 4, Publisher Amer Statistical ASSOC, USA, pp. 343-345. ISSN 0003-1305.

Nielsen, H. B. (2016). *The Co-Integrated Vector Autoregression with Errors-in-Variables.* Econometric Reviews, Vol. 35, Issue 2, Publisher Taylor & Francis INC, Philadelphia, pp. 169-200. ISSN 0747-4938.

Rabiej, M. (2012). *Statystyka z programem Statistica*, Wydawnictwo Helion, Gliwice, pp. 219-221. ISBN 978-83-246-4110-9.

Sobczyk, M. (2007). *Statystyka. Nowe wydanie,* Wydawnictwo Naukowe PWN, Warszawa, p. 237. ISBN 978-83-01-15199-7.

Stanisz, A. (2007). *Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny. Volume 2. Modele liniowe i nieliniowe,* Copyright by StatSoft Polska Sp. z o.o., Cracow, pp. 479-487. ISBN 978-83-88724-30-5.

Contact

Katarzyna Brożek
Kazimierz Pulaski University of Technology and Humanities in Radom
Malczewskiego 29, 26-600 Radom, Poland
kania6669@wp.pl

Justyna Kogut
Kazimierz Pulaski University of Technology and Humanities in Radom
Malczewskiego 29, 26-600 Radom, Poland
justynakogut5@wp.pl