Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 37(3), 2010, Pages 35–51
ISSN: 1223-6934

Statistical Correlation Study

Dorel Săvulea and Nicolae Constantinescu

Abstract. Correlation is a statistical method used to determine relationships between two or more variables. The intensities of various factors influence change in time and space conditions such that the evolution of the dependent phenomena is changing comparing with the previous trends. This paper represents a study on the correlation between three very important economic factors: Government public debt as a percentage of gross domestic product, labor productivity and employment rate, Eur per Gigajoule for natural gas. The statistics contain data from 1998 to 2009 for various countries. In order to analyze the economic influences of each factor we will present statistics for countries that differ depending on geographic location and the citizens' standard of living. We also exemplify a situation where the correlation coefficient obtained has an aberrant value.

2010 Mathematics Subject Classification. Primary: 62P20 ; Secondary: 46N30, 62B15.
Key words and phrases. correlation coefficient, regression, economics, Government public debt as a percentage of gross domestic product, labor productivity, employment rate.

1. Introduction

Causal relationship between socio-economic phenomena can be quantified and analyzed using correlation. The information obtained is very useful, especially for the fact that the specific statistical methods provide the researcher knowledge of the following [3]:

- existence of causal relationships between phenomena;
- each factor contribution to overall variability phenomena effect;
- intensity of causal links between socio-economic phenomena and processes;
- evolutionary trends of correlation between events.

Despite the fact that determining the correlation specific factors is very difficult, the correlation analysis is often used mainly because it offers much more information.

The correlation can be defined as the synthetic expression of the causal links intensity between phenomena. The correlative tuple can contain two or more variables. One of these variables is called effect variable, while the others are cause variables. The effect variable is a result variable while the cause variables are factorial variables.

The correlation can be classified in [8]:

(1) depending on the number of variables from the correlative tuple
 (a) simple correlation – the correlative tuple contains two variables (one is a factor and the other is a result);
 (b) multiple correlation – the correlative tuple contains at least three variables (one is a result and the others are factors);
(2) depending on the factorial links
 (a) direct correlation – the variability of the result occurs in the same sense as the variability of the factor/s;

Received June 26, 2010. Revision received August 16, 2010.

   (b) inverse correlation – the variability of the result occurs in reverse with the variability of the factor/s;
(3) depending on the causal links
   (a) linear correlation – the result variable shows a linear trend due to the influence of determinants;
   (b) non-linear correlation – the result variable fits into a trend of non-linear type (parabolic, exponential, etc.).

Knowing the correlations shape are of particular interest in estimating the evolutionary trends of effect phenomena closely related to the determinant factors variability.

Clearly, a linear trend of the result variable, suggests that levels of this variable increases or decreases approximately in arithmetic progression; non-linear trend shows the changing of these levels in geometric progression, exponential, etc. [5]. This information is particularly useful in forecasting calculations and in modeling evolutionary trends.

This article uses the correlation between three important economic factors: total population, labor productivity and employment rate to emphasize and analyze the influences on the economies of countries which differ by geographical location and the standard of living of the citizens.

2. State of Art

Correlation coefficient is a quantitative value that describes the relationship between two or more variables. It varies between $(-1, 1)$, where the extreme values assume a perfect relationship between variables while 0 means a total lack of linear relationship. A proper interpretation of values is obtained by comparing results obtained with certain defaults on tables of correlations depending on the number of subjects, type of connection desired and materiality [9].

The two main conditions that must be met to use parametric tests are:

- Normal distribution of variable interest in a sample survey;
- Homogeneity of sample variance that is studied on the subject variable.

Parametric samples are preferred for the fulfillment of these conditions because they are stronger, which means increasing the chance of rejecting a false hypothesis. These conditions can be verified by locating the average in the normal distribution of data and calculating the indicators of homogeneity of the sample studied.

The most important types of parametric correlations are [7]:

(1) Simple correlation coefficient (Bravais-Pearson's)
(2) Eneahoric correlation coefficient;
(3) Coefficient of partial correlation;
(4) Biserial and triserial correlation coefficients.

In this paper we use the methodology of applying statistical-mathematical function in the study of simple regression and simple correlation intensity quantification methodology.

In the correlation analysis the two key issues that are taken in consideration are [6]:

- regression – which determines the determining factors contributions to variability effect phenomena, using and interpreting the regression coefficients of different mathematical functions statistic;

- correlation intensity – summarized by correlation coefficients. For simple correlation, the first part can be shown using functions: linear – for linear causal links; higher-order parabolic, hyperbolic, exponential, log, semi-logarithmic, logistic, etc. – for non-linear causal link.

The linear regression of order 1 is given by the expression [4]:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

y = dependent variable
x = independent variable
β_0 = the initial value of y
β_1 = the modification of y
caused by the changes of x
ϵ = variable's error

To compute the parameters β_0 and β_1, estimator parameters of the following linear function are used:

$$\widehat{y} = b_0 + b_1 x$$

where

$$b_1 = \frac{cov(x,y)}{s_x^2}$$
$$b_0 = \overline{y} - b_1\overline{x}$$

The elements used to calculate the two parameters are determined as follows:

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

In the analysis of socio-economic phenomena are very frequent the cases where empirical values y_i and theoretical values $\overline{y}$ differ; these elements are determined by random factors and are the residual value estimated, known as error [1].

The average error (or standard error) of forecasts is calculated from the formula:

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

where SSE is the sum of squared errors of estimation, and can be computed using the expression:

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_1})^2$$

To determine if b_1 estimates correctly β_1 there must be verified the null hypothesis:

$$H_0 : \beta_1 = 0$$

Its alternative is represented by:

$$H : \beta_1 \neq 0$$

In order to verify the null hypothesis we use the test t given by the relation:

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

where

$$S_{b_1} = \frac{S_e}{\sqrt{(n-1)s_x^2}}$$

If the variable's errors are normally distributed, the statistic test t, considered as a referential level, is the one that corresponds to $n-2$ degrees of freedom [2].

The rejection region for the null hypothesis is:

$$t > t_{\alpha/2,n-2} \quad or \quad t < -t_{\alpha/2,n-2}$$

To determine the contribution of one of the factors influencing the variability of the dependent phenomenon we use the determining coefficient given by:

$$R^2 = \frac{[cov(x,y)]^2}{s_x^2 s_y^2}$$

or

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y_i})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{SSE}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

Since the deviations $(y_i - \overline{y})$ show the total variance of variable y, given by all the factors that influenced it and knowing that a part from this variation is taken by the regression function and highlighted through the deviations $(\widehat{y_i} - \overline{y})$ and the other part – which is the residual variation of the forecasts' error – is evidenced by the deviations $(y_i - \widehat{y_i})$, means that:

$$(y_i - \overline{y}) = (y_i - \widehat{y_i}) + (\widehat{y_i} - \overline{y})$$

From the above relation we have:

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \widehat{y_i})^2 + \sum_{i=1}^n (\widehat{y_i} - \overline{y})^2$$

If we consider:

$$SS_y = \sum_{i=1}^n (y_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \widehat{y_i})^2$$

the determining coefficient can be computed using various methods:

$$R^2 = 1 - \frac{SSE}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

$$R^2 = \frac{SS_y - SSE}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

$$R^2 = \frac{SSR}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

The determining coefficient R^2 shows the proportion of the dependent variable's variation y determined by the influence of the independent variable's variation x.

Correlation coefficient r shows the intensity of the causal link between the two variables. Both coefficients are based on Pearson's relationship; the determining coefficient is the square of the correlation coefficient. So:

$$r = \sqrt{R^2} = \frac{cov(x,y)}{s_x s_y}$$

The correlation coefficient r takes values between -1 and 1 with the following meaning:

- values between 0 and 1 show a direct correlation of increasingly intense as they approach one;
- values between 0 and -1 show an inverse correlation of increasingly intense as close to -1;
- zero value points out that between the two variables there is no connection.

Usually, in practice the interval between -1 and 1 is refined as follows [10]:

- If $0 \leq r < 0.2$ there is no significant relation between variables;
- If $0.20 \leq r \leq 0.50$ the relationship between variables is low;
- If $0.50 \leq r < 0.75$ the relationship between variables is average;
- If $0.75 \leq r < 0.95$ the relationship between variables is strong;
- If $0.95 \leq r \leq 1$ there is a functional relation between the two variables.

To test the significance of the correlation coefficient we must verify the null hypothesis, which says that between two variables that there is no linear relation. In this case the test t is given by:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where r is the correlation coefficient and n is the number of the pairs (x, y).

The computed value of t is compared with the theoretical value obtained from table t, for $n-2$ degrees of freedom and the significance level set. The rejection region is set in the same way as for the regression function.

If H_0 is rejected there is assumed that the correlation coefficient values is different from 0, which means that between the two variables exists a significant relation. If H_0 is accepted between the two variables exists a random relation.

Correlation between two variables can sometimes mislead and can be difficult to interpret when there exists a third variable responsible for the dependence between the other two. In such a case we use the partial correlation coefficient. In order to give a formula for computing this coefficient we suppose to have three variables $1, 2, 3$. The correlation coefficients have the following values: r_{12}, r_{23} and r_{13}. To compute $r_{12.3}$ we use:

$$r_{12.3} = \frac{r_{12} - r_{13} * r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

3. Economic significance of correlation data

This paper represents a study on the correlation between important economic factors like: Government public debt as a percentage of gross domestic product, labor productivity and employment rate, Euro per Gigajoule for natural gas, homes with Internet locations etc.. We have computed numerical data using the formulas presented in the above section in order to interpret them in economic terms. The study was made on various countries which were grouped depending on their geographical location, life standards and political influences. In this article we chose the representative countries for each group.

Belgium is located in the northwest Europe, having one of the highest living standards. Belgium is a constitutional, popular monarchy and a parliamentary democracy, being one of the founding members of the European Union and hosting its headquarters, as well as those of other major international organizations, including NATO. The statistics for Belgium are described in table 3.

The formula used for the correlation between two parameters is:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

where $n = 12$, x_i and y_i are the two parameters.

Table 1. Belgium statistics

	Government debt % of GDP	Labor productivity	Employment rate
1998	117,4	136,7	57,4
1999	113,7	133,6	59,3
2000	107,9	136,3	60,5
2001	106,6	134,6	59,9
2002	103,5	131,8	59,9
2003	98,5	129,8	59,6
2004	94,2	128,5	60,3
2005	92,1	126,8	61,1
2006	88,1	125,5	61
2007	84,2	124,8	62
2008	89,8	126,1	62,4
2009	96,7	126	61,6

We note the correlation coefficient for government debt % of GDP and labor productivity with r_{gl}, for government debt % of GDP and employment rate with r_{ge}, and for labor productivity and employment rate with r_{le}. All the computation are presented in the following tables. For r_{gl} we have the table 3 and:

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 117,4 + 113,7 + 107,9 + 106,6 \\ &\quad +103,5 + 98,5 + 94,2 + 92,1 + 88,1 \\ &\quad +84,2 + 89,8 + 96,7 = 1192,7\end{aligned}$$

Table 2. Government debt % of GDP and labor productivity

i	x_i	y_i	x_i^2	y_i^2	xy
1	117,4	136,7	13782,76	18686,89	16048,58
2	113,7	133,6	12927,69	17848,96	15190,32
3	107,9	136,3	11642,41	18577,69	14706,77
4	106,6	134,6	11363,56	18117,16	14348,36
5	103,5	131,8	10712,25	17371,24	13641,3
6	98,5	129,8	9702,25	16848,04	12785,3
7	94,2	128,5	8873,64	16512,25	12104,7
8	92,1	126,8	8482,41	16078,24	11678,28
9	88,1	125,5	7761,61	15750,25	11056,55
10	84,2	124,8	7089,64	15575,04	10508,16
11	89,8	126,1	8064,04	15901,21	11323,78
12	96,7	126	9350,89	15876	12184,2

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 1192,7^2 = 1422533,29$$

$$\begin{aligned}\sum_{i=1}^{n} y_i &= 136,7+133,6+136,3+134,6\\ &\quad +131,8+129,8+128,5+126,8\\ &\quad +125,5+124,8+126,1+126\\ &= 1560,5\end{aligned}$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 = 1560,5^2 = 2435160,25$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &= 13782,76+12927,69+11642,41\\ &\quad +11363,56+10712,25+9702,25\\ &\quad +8873,64+8482,41+7761,61\\ &\quad +7089,64+8064,04+9350,89\\ &= 119753,15\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= 18686,89+17848,96+18577,69\\ &\quad +18117,16+17371,24+16848,04\\ &\quad +16512,25+16078,24+15750,25\\ &\quad +15575,04+15901,21+15876\\ &= 203142,97\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= 16048,58+15190,32+14706,77\\ &\quad +14348,36+13641,3+12785,3\\ &\quad +12104,7+11678,28+11056,55\\ &\quad +10508,16+11323,78+12184,2\\ &= 155576,3\end{aligned}$$

$$\begin{aligned}r_{gl} &= \frac{12*155576,3-1192,7*1560,5}{\sqrt{12*119753,15-1422533,29}}\\ &\quad *\frac{1}{\sqrt{12*203142,97-2435160,25}}\\ &= \frac{5707,25}{120,43*50,55} = \frac{5707,25}{6087,73} = 0,93\end{aligned} \tag{1}$$

The value obtained emphasizes a high positive correlation in the sense that if the government debt raises with 100% the labor productivity raises with 93%. This shows Belgium people is consciuos of the fact that they have to work harder in order to cover and reduce the government public debt. However, the government must be very carefull to what extreme it can "exploit" its own people. The table 3 contains the data used for computing r_{ge}.

Table 3. Government debt % of GDP and employment rate

i	x_i	y_i	x_i^2	y_i^2	xy
1	117,4	57,4	13782,76	3294,76	6738,76
2	113,7	59,3	12927,69	3516,49	6742,41
3	107,9	60,5	11642,41	3660,25	6527,95
4	106,6	59,9	11363,56	3588,01	6385,34
5	103,5	59,9	10712,25	3588,01	6199,65
6	98,5	59,6	9702,25	3552,16	5870,6
7	94,2	60,3	8873,64	3636,09	5680,26
8	92,1	61,1	8482,41	3733,21	5627,31
9	88,1	61	7761,61	3721	5374,1
10	84,2	62	7089,64	3844	5220,4
11	89,8	62,4	8064,04	3893,76	5603,52
12	96,7	61,6	9350,89	3794,56	5956,72

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 117,4+113,7+107,9+106,6\\ &\quad +103,5+98,5+94,2+92,1+88,1\\ &\quad +84,2+89,8+96,7=1192,7\end{aligned}$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 1192,7^2 = 1422533,29$$

$$\begin{aligned}\sum_{i=1}^{n} y_i &= 57.4+59.3+60.5+59.9+59.9+59.6\\ &\quad +60.3+61.1+61+62+62.4+61.6\\ &= 725\end{aligned}$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 \approx 6270,62^2 = 525625$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &= 13782,76 + 12927,69 + 11642,41 \\ &\quad +11363,56 + 10712,25 + 9702,25 \\ &\quad +8873,64 + 8482,41 + 7761,61 \\ &\quad +7089,64 + 8064,04 + 9350,89 \\ &= 119753,15\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= 3294,76 + 3516.49 + 3660.25 \\ &\quad +3588.01 + 3588.01 + 3552.16 \\ &\quad +3636.09 + 3733.21 + 3721 \\ &\quad +3844 + 3893.76 + 3794.56 \\ &= 43822,3\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= 6738,76 + 6742,41 + 6527,95 \\ &\quad +6385,34 + 6199,65 + 5870,6 \\ &\quad +5680,26 + 5627,31 + 5374,1 \\ &\quad +5220,4 + 5603,52 + 5956,72 \\ &= 71927,02\end{aligned}$$

Then, we have:

$$\begin{aligned} r_{ge} &= \frac{12 * 71927,02 - 1192,7 * 725}{\sqrt{12 * 119753,15 - 1422533,29}} \\ &\quad * \frac{1}{\sqrt{12 * 43822,3 - 525625}} \\ &= \frac{-1583,26}{\sqrt{14504,6 * 242,6}} = \frac{-1583,26}{1875,85} = -0,84 \end{aligned} \tag{2}$$

The value of this correlation shows that the government is forced to borrow more money in order to maintain the same life standards for its unemployed population. So, if the employment rate decreases the public debt will grow higher and higher.

Table 3 contains the data used for computing the correlation coefficient for labor productivity and employment rate.

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 136,7 + 133,6 + 136,3 + 134,6 \\ &\quad +131,8 + 129,8 + 128,5 + 126,8 \\ &\quad +125,5 + 124,8 + 126,1 + 126 \\ &= 1560,5\end{aligned}$$

TABLE 4. Labor productivity and employment rate

i	x_i	y_i	x_i^2	y_i^2	xy
1	136,7	57,4	18686,89	3294,76	7846,58
2	133,6	59,3	17848,96	3516,49	7922,48
3	136,3	60,5	18577,69	3660,25	8246,15
4	134,6	59,9	18117,16	3588,01	8062,54
5	131,8	59,9	17371,24	3588,01	7894,82
6	129,8	59,6	16848,04	3552,16	7736,08
7	128,5	60,3	16512,25	3636,09	7748,55
8	126,8	61,1	16078,24	3733,21	7747,48
9	125,5	61	15750,25	3721	7655,5
10	124,8	62	15575,04	3844	7737,6
11	126,1	62,4	15901,21	3893,76	7868,64
12	126	61,6	15876	3794,56	7761,6

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 1560,5^2 = 2435160,25$$

$$\begin{aligned}\sum_{i=1}^{n} y_i &= 57.4 + 59.3 + 60.5 + 59.9 + 59.9 + 59.6 \\ &\quad +60.3 + 61.1 + 61 + 62 + 62.4 + 61.6 \\ &= 725\end{aligned}$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 \approx 6270,62^2 = 525625$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &= 18686,89 + 17848,96 + 18577,69 \\ &\quad +18117,16 + 17371,24 + 16848,04 \\ &\quad +16512,25 + 16078,24 + 15750,25 \\ &\quad +15575,04 + 15901,21 + 15876 \\ &= 203142,97\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= 3294,76 + 3516.49 + 3660.25 \\ &\quad +3588.01 + 3588.01 + 3552.16 \\ &\quad +3636.09 + 3733.21 + 3721 \\ &\quad +3844 + 3893.76 + 3794.56 \\ &= 43822,3\end{aligned}$$

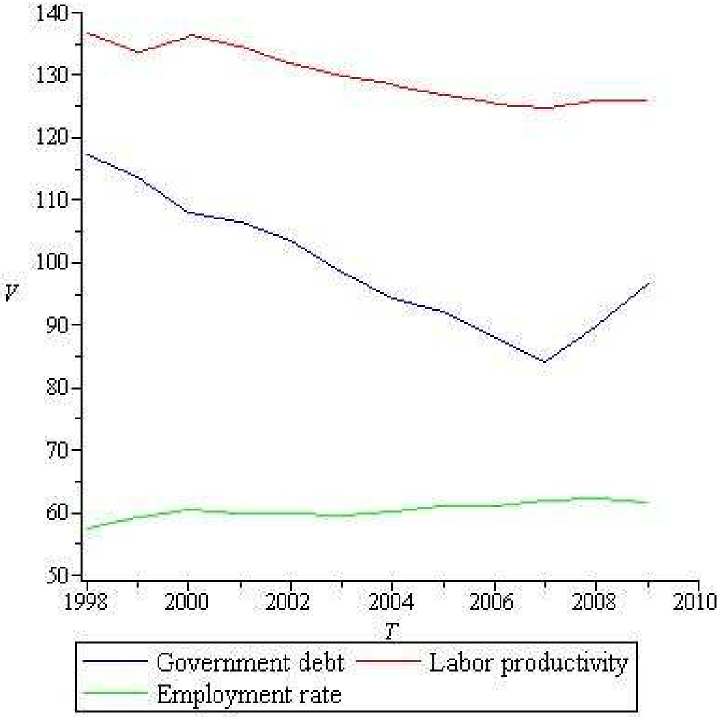

Figure 1. Belgium parameters

$$
\begin{aligned}
\sum_{i=1}^{n} x_i y_i &= 7846,58+7922,48+8246,15+8062,54 \\
&\quad +7894,82+7736,08+7748,55 \\
&\quad +7747,48+7655,5+7737,6+7868,64 \\
&\quad +7761,6=94228,02
\end{aligned}
$$

$$
\begin{aligned}
r_{le} &= \frac{12*94228,02-1560,5*725}{\sqrt{12*203142,97-2435160,25}} \\
&\quad \frac{1}{\sqrt{12*43822,3-525625}} \\
&= \frac{-626,26}{\sqrt{2555,39*242,6}}=\frac{-626,26}{787,36} \\
&= -0,79
\end{aligned} \tag{3}
$$

The value of r_{le} emphasizes a bad aspect of this country since the labor productivity decreases with the growth of the employment rate. It shows that Belgium productivity grows with a lower rate than the employment one.

Figure 1 is the graph for the three parameters described. As we can see from the numeric values obtained for the correlation coefficient the government debt and the labor productivity are in the same sense while the employment rate has an inverse sense.

In order to present a partial correlation we have chosen three factors: annual inflation rate, percent of homes with Internet connection and Eur per Gigajoule for natural gas for Slovenia. Our purpose is to compute the correlation coefficient for homes with Internet connection and Eur per Gigajoule for natural gas. Since these two factors do not have much in common and using a simple correlation factor can be difficult to interpret we chose the third factor, the annual inflation rate, as a link between these two. We note r_{ie} the correlation between homes with Internet connection and Eur per Gigajoule for natural gas, r_{ii} the correlation between homes with Internet connection and annual inflation rate, and r_{ei} the correlation between Eur per Gigajoule for natural and annual rate inflation. The partial correlation coefficient will be:

$$r_{ie.i} = \frac{r_{ie} - r_{ii} * r_{ei}}{\sqrt{1 - r_{ii}^2}\sqrt{1 - r_{ei}^2}}$$

Inflation, with all its meanings and intensities, is a macroeconomic problem, and because of that, it has multiple determinations, interdependencies and impacts on the country's economy. Given the very harmful effects of inflation on the economy of a country, we can say categorically that inflation is "the worst disease" of the economy.

The transition from inflationary economic evolution and non-inflationary growth to rational concepts is made through the concept of deflation and disinflation. Deflation is a monetary-rolling process characterized by low long-term sustainable price level resulted from a set of measures to restrict nominal demand for reducing tensions on the dynamics of increasing prices.

If inflation usually expresses an inverse relation between very strong price growth and dynamics of economic growth (slow growth or even a decrease in domestic production), deflation, by contrast, expresses a direct relationship (with the same sense) between price dynamics and production dynamics (both variables decrease).

Slovenia is a country in Central Europe, its form of government being parliamentary republic. The Slovenian head of state is the president, who is elected by popular vote every five years, and has mainly advisory and ceremonial duties. Slovenia has a medium life standard mainly because the number of pensioners is large. The data for Slovenia is presented in 3. First we compute the correlation factor r_{ie}.

Table 5. Slovenia statistic data

	Homes with internet connection	Euro per GJ for natural gas	Annual inflation rate
2004	47	7,2	3,7
2005	48	7,8	2,5
2006	54	10,0	2,5
2007	58	10,8	3,8
2008	59	12,1	8,5
2009	64	14,4	0,9

Table 6. Percent of homes with Internet connection and Eur per GJ for natural gas

i	x_i	y_i	x_i^2	y_i^2	xy
1	47	7,2	2209	51,84	338,4
2	48	7,8	2304	60,84	374,4
3	54	10,0	2916	100,00	540,0
4	58	10,8	3364	116,64	626,4
5	59	12,1	3481	146,41	713,9
6	64	14,4	4096	207,36	921,6

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 47+48+54+58+59+64\\ &= 330\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i &= 7,2+7,8+10,0+10,8\\ &\quad +12,1+14,4=62,3\end{aligned}$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 330^2 = 108900$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 = 62,3^2 = 3881,29$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &= 2209+2304+2916\\ &\quad +3364+3481+4096\\ &= 18370\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= 51.84+60.84+100.00\\ &\quad +116.64+146.41+207.36\\ &= 683,09\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= 338.4+374.4+540.0\\ &\quad +626.4+713.9+921.6\\ &= 3514,7\end{aligned}$$

The correlation coefficient is given by:

$$\begin{aligned} r_{ie} &= \frac{6*3514,7-330*62,3}{\sqrt{6*18370-108900}} \\ &\quad *\frac{1}{\sqrt{6*683,09-3881,29}} \\ &= \frac{529,2}{\sqrt{1320*217,25}}=\frac{529,2}{535,50}=0,98 \end{aligned} \tag{4}$$

Next we compute r_{ei} for Eur per GJ for natural gas and annual inflation rate.

Table 7. Eur per GJ for natural gas and annual inflation rate

i	x_i	y_i	x_i^2	y_i^2	xy
1	7,2	3,7	51,84	13,69	26,64
2	7,8	2,5	60,84	6,25	19,5
3	10,0	2,5	100,00	6,25	25
4	10,8	3,8	116,64	14,44	41,04
5	12,1	8,5	146,41	72,25	102,85
6	14,4	0,9	207,36	0,81	12,96

$$\begin{aligned} \sum_{i=1}^{n} x_i &= 7,2+7,8+10,0+10,8 \\ &\quad +12,1+14,4=62,3 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n} y_i &= 3,7+2,5+2,5+3,8+8,5+0,9 \\ &= 21,9 \end{aligned}$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 62,3^2 = 3881,29$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 = 21,9^2 = 479,61$$

$$\begin{aligned} \sum_{i=1}^{n} x_i^2 &= 51.84+60.84+100.00 \\ &\quad +116.64+146.41+207.36 \\ &= 683,09 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n} y_i^2 &= 13,69+6,25+6,25+14,44 \\ &\quad +72,25+0,81=113,69 \end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= 26,64+19,5+25+41,04 \\ &\quad +102,85+12,96=227,99\end{aligned}$$

$$\begin{aligned}r_{ei} &= \frac{6*227,99-62,3*21,9}{\sqrt{6*683,09-3881,29}} \\ &\quad *\frac{1}{\sqrt{6*113,69-479,61}} \\ &= \frac{3,57}{\sqrt{217,25*202,53}}=\frac{3,57}{209,76}=0,017\end{aligned} \tag{5}$$

Even if these two factors seem to be highly related the small values of the correlation coefficient is due to the fact that Slovenia has only foreign suppliers of natural gas and does not produce anything for its own use. So the factors that influenced the cost of Gigajoule are entirely independent of its annual inflation rate.

The data for computing r_{ii} is described in the next table.

TABLE 8. Percent of homes with Internet connection and annual inflation rate

i	x_i	y_i	x_i^2	y_i^2	xy
1	47	3,7	2209	13,69	173,9
2	48	2,5	2304	6,25	120
3	54	2,5	2916	6,25	135
4	58	3,8	3364	14,44	220,4
5	59	8,5	3481	72,25	501,5
6	64	0,9	4096	0,81	57,6

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 47+48+54+58+59+64 \\ &= 330\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i &= 3,7+2,5+2,5+3,8+8,5+0,9 \\ &= 21,9\end{aligned}$$

$$\left(\sum_{i=1}^{n} x_i\right)^2 = 330^2 = 108900$$

$$\left(\sum_{i=1}^{n} y_i\right)^2 = 21,9^2 = 479,61$$

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &= 2209 + 2304 + 2916 \\ &\quad +3364 + 3481 + 4096 \\ &= 18370\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= 13,69 + 6,25 + 6,25 + 14,44 \\ &\quad +72,25 + 0,81 = 113,69\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= 173,9 + 120 + 135 + 220,4 \\ &\quad +501,5 + 57,6 = 1208,4\end{aligned}$$

$$\begin{aligned}r_{ii} &= \frac{6 * 1208,4 - 330 * 21,9}{\sqrt{6 * 18370 - 108900}} \\ &\quad * \frac{1}{\sqrt{6 * 113,69 - 479,61}} \\ &= \frac{23,4}{\sqrt{1320 * 202,53}} = \frac{23,4}{517.04} = 0,045\end{aligned} \tag{6}$$

From (4), (5) and (6) we obtain the partial correlation coefficient:

$$\begin{aligned}r_{ie.i} &= \frac{r_{ie} - r_{ii} * r_{ei}}{\sqrt{1 - r_{ii}^2}\sqrt{1 - r_{ei}^2}} \\ &= \frac{0,98 - 0,045 * 0,017}{\sqrt{1 - 0,002}\sqrt{1 - 0,0002}} \\ &= \frac{0,97}{\sqrt{0,998 * 0,9998}} = \frac{0,97}{0,998} = 0,97\end{aligned}$$

Considering that the annual inflation rate and the cost of Gigajoule are entirely independent (which is a special case), we say that the correlation coefficient has an aberrant value. This means that these two factors cannot be statistically connected.

4. Conclusions

The two examples show that the correlation coefficient emphasizes whether there is a connection between two factors and, if there is, how strong it is. It can also be used for two factors which apparently are not related in any way. We can find a "linking" factor between them in order to compute the partial correlation coefficient. However, as presented in the Slovenia example, we can obtain no relation between two factors that apparently are strongly linked (Eur per Gigajoule for natural gas and the annual inflation rate), which means that other external factors influence them. The correlation coefficient is very important for both experts or beginners in a certain domain. For example, if an economist reads this article he is probably most interested in the value of the correlation coefficient because it shows to what extend the change of a certain factor (increase or decrease) leads to the modification of the other factor. But, if a beginner reads this paper he can find out new things about connection between economic factors which he probably had not know before.

There is no specific reason for choosing the two countries except that they have a different level of development influenced by various factors. Belgium is considered to be one of the countries with the highest life standard level, while Slovenia is a developing country. Even if its labor productivity is good, Belgium also has a high government debt because it has to borrow money in order to maintain the much higher life standard. However, Slovenia, even if it is less developed, manages to maintain a rather low annual inflation rate excepting 2008 when the financial crises stroke strong and we are witnesses to an explosion of the inflation.

References

[1] J. Bartko, On various intraclass correlation reliability coefficients, *Psychological Bulletin* (1976), 762–765.

[2] L. Crocker and J. Algina, *Introduction to classical and modern test theory*, Fort Worth: Holt, Rinehart and Winston, 1986.

[3] A.L. Edwards, The Correlation Coefficient, In *An Introduction to Linear Regression and Correlation* (1976), San Francisco, CA: W. H. Freeman, 33–46.

[4] M.G. Kendall, *Rank Correlation Methods*, Charles Griffin & Co., 1955.

[5] J.F. Kenney and E.S. Keeping, Linear Regression and Correlation, In *Mathematics of Statistics* **1** (1962), Princeton, NJ: Van Nostrand, 252–285.

[6] E.L. Lehmann, *Testing statistical hypotheses*, (2nd ed.), New York: Wiley, 1986.

[7] J.L. Rodgers and W.A. Nicewander, Thirteen ways to look at the correlation coefficient, *The American Statistician* **42** (1988), no. 1, 59-66.

[8] G.W. Snedecor and W.G. Cochran, The Sample Correlation Coefficient r and Properties of r, In *Statistical Methods* (1980), 7th ed. Ames, IA: Iowa State Press, 175–178.

[9] M.R. Spiegel, Correlation Theory, In *Theory and Problems of Probability and Statistics* (1992), 2nd ed. New York: McGraw-Hill, 294–323.

[10] B.J. Winer, *Statistical principles in experimental design* (2nd ed.), New York: McGraw-Hill, 1971.

(Dorel Săvulea and Nicolae Constantinescu) Faculty of Mathematics and Computer Science, University of Craiova, Al.I. Cuza Street, No. 13, Craiova RO-200585, Romania, Tel. & Fax: 40-251412673
E-mail address: savulea@central.ucv.ro, nikyc@central.ucv.ro