COMPARATIVE STUDY ON THE USE OF STATISTICAL METHODS FOR THE DISTRIBUTION OF AUTOCORRELATION TO ANY INFLUENCE FACTOR

CRISTIAN MERCE[1], MANEA DRĂGHICI[2], EMILIAN MERCE[3]
RALUCA-ALEXANDRA NECULA[4]

Abstract*: Three methods reported in the literature are subject to comparative analysis in the present paper:*

1. *Classic method [1,5];*
2. *Merce E., Merce C.C. Method[2,3];*
3. *Merce E., at al Method[4];*

It is shown that in the case of the first two methods mentioned above, the attempts to distribute interactions on influence factors have as a prerequisite the determination of the simple correlation coefficients and of the partial correlation coefficients, the methods being of this particularly laborious nature. With obvious computational facilities, compared to the first two methods, the authors propose the use of a new method based on the principle of proportional distribution of autocorrelation with the coefficients of simple determination, and the following five steps are being performed: 1) Calculation of multiple correlation coefficient and simple correlation coefficients using the Regression function of the Microsoft Excel Data Analysis component; 2) The recording of the multiple correlation coefficient and the simple correlation coefficients in the Excel table used for this purpose; 3) Calculating the coefficients of the simple determination and the multiplication factor; 4) Sum of coefficients of simple determination; 5) Calculating the proportions of simple determinations, considering their sum equal to 100; 6) Determination of the influence of each factor as a product between multiple determinations and the proportion of simple determinations. Note that the last four steps in the Excel work table are generated instantly after the first two steps.

Keywords*: autocorrelation, comparative analysis methods, distribution of autocorrelation on each method, method and program.*

JEL Classification*: C36*

INTRODUCTION

Collinearity is an objective reality in the investigation of complex causal relationships, which is outlined, as demonstrated in the literature (Merce E., et al, 2004; 2017), whenever information about the causal complex is incomplete. The presence of collinearity alters the accuracy of numerical determinations between factors, on the one hand, and the effect studied, on the other. The phenomenon of collinearity cannot, however, always be avoided. This is primarily about economics, sociology, psychology, but also about complex agro-biological experiments.

It is, therefore, natural to be concerned about assessing collinearity and then correcting the relationship between determining factors and effect. For this purpose, a method for individualizing the influence of each factor has been outlined, based on the calculation of the coefficients of the simple correlation and the partial correlation (Merce E.,1986); Moineagu C.,1974).

Another method of distributing collinearity on the influence factors recommended in the literature is based on the calculation of the influence of factors in a certain causal complex as the average of simple and partial determinations in all possible successions (Merce E., et al; 2017).

In these two working hypotheses, the researcher must evaluate the collinearity numerically and then proceed to correct the relationship between the factors studied and the effect. However, the use of the two mentioned methods is difficult, requiring extremely laborious calculations to determine the coefficients of partial correlation, especially in the case of causal relationships with more than two factors.

[1] Conferenţiar dr., UASVM Cluj Napoca, merceccristian@hotmail.com
[2] Professor dr., UASVM Bucharest, dmprofesor@hotmail.com
[3] Professor dr., UASVM Cluj-Napoca, emerce@usamvcluj.ro
[4] Sef lucrari dr, USAMV Bucureşti, raluca_nec@yahoo.com

We propose and offer in this sense a new calculation method based on the distribution of the autocorrelation on the factors of influence, using the principle of the proportionality of the determinations with the simple correlation coefficients. To individualize the influence of each factor, a working method has been imagined that harnesses the benefits offered by Microsoft Excel as a workbook. With obvious computing facilities, compared to the first two methods, the authors suggest using this original method, following the next six steps, the first being mandatory, the next four being resolved instantly:

1. Calculation of the multiple correlation coefficient and simple correlation coefficients using the Regression function of the Microsoft Excel Data Analysis component;
2. The recording of the multiple correlation coefficient and the simple correlation coefficients in the Excel table;
3. Calculating the coefficients of the simple determination and the multiple correlation coefficient;
4. The sum of the coefficients of the simple determination;
5. Calculating the proportions of simple determinations, considering their sum equal to 100;
6. Determination of the influence of each factor, as a product of multiple determinations and the proportion of simple determinations.

MATERIALS AND METHODS

The complexity of causal relationships in different areas of activity, as well as the set of variables investigated, often make it impossible to obtain complete databases. Studies, observations, and concrete processing are grounds that have led us to conclude that the source of collinearity is incomplete information about all possible combinations of variants of influence factors.

And in the case of agricultural experiments there are often encountered situations that comprise only a few of the possible combinations of variants of influence factors. We assume, in this respect, an experience with the evolution of average maize production depending on nitrogen and phosphorus fertilizers doses (Table 1).

Table 1

The evolution of average maize production according to NP quantities (conventional data)

Dose	Kg/ha	Dose	Kg/ha	Dose	Kg/ha	Dose	Kg/ha
N_0P_0	5072	$N_{50}P_{80}$	6466	$N_{100}P_{120}$	8517	$N_{150}P_{160}$	8732
N_0P_{40}	5452	$N_{100}P_{40}$	6720	$N_{150}P_{80}$	8622	$N_{200}P_{120}$	8875
$N_{50}P_{40}$	6593	$N_{100}P_{80}$	8368	N_{150}P120	8748	$N_{200}P_{160}$	8726

RESULTS AND DISCUSSIONS

The picture of the possible combinations of NP variants and the corresponding average outputs is shown in Table 2.

Table 2

The range of possible combinations of the five variants of each factor

X_2 \ X_1	0	50	100	150	200
0	5072	?	?	?	?
40	5452	6593	6720	?	?
80	?	6466	8368	8622	?
120	?	?	8517	8748	8875
160	?	?	?	8732	8726

This is a typical example of incomplete information that generates collinearity and all attempts and achievements on how to redistribute it.

Correspondences between the levels of the factors allocated and the average outputs obtained as incomplete data are centralized in Table 3.

Table 3

Correspondence between NP combinations and average outputs on incomplete data base

N	P	Kg/ha
0	0	5072
0	40	5452
50	40	6593
50	80	6466
100	40	6720
100	80	8368
100	120	8517
150	80	8622
150	120	8748
150	160	8732
200	120	8875
200	160	8726

The three mentioned methods are presented comparatively, illustrating the distribution of autocorrelation on the factors of influence.

For all three methods, for the distribution of autocorrelation, it is necessary to determine the correlation coefficients in the hypothesis of a certain theoretical regression model. In order to express the causal relationship between the average production versus two factors it was hypothesized that the link could be expressed by a bifactorial linear model and by mono-factorial models respectively.

Through the processing of the database, the following concrete forms of the models were obtained:

$$\bar{y}(x_1x_2) = 5396.9 + 13.05x_1 + 9.44x_2 \ ; \quad \mathrm{R}_{yx_1x_2} = 0,934 \ ; \ \mathrm{D}_{yx_1x_2} = 87,2\ \%$$

$$\bar{y}(x_1) = 5619.5 + 18.76x_1 \ ; \quad \mathrm{r}_{yx_1} = 0,914;$$

$$\bar{y}(x_2) = 5489.2 + 24.06x_2 \ ; \quad \mathrm{r}_{yx_2} = 0,862;$$

$$\bar{x}_1(x_2) = 7,08 + 1,12x_2 \ ; \quad \mathrm{r}_{x_1x_2} = 0,824;$$

Taking into account the concrete form of calculated regression models, the methodological content of the three methods can be emphasized. It is specified that for the first two methods it is also necessary to calculate the partial correlation coefficients.

Method 1 (Moineagu C, 1974):

According to this method, the individualization of the influence of the two factors implies the redistribution of the interaction between them. To this end, it is mandatory to determine the partial correlation coefficients by using specific calculation relationships (Moineagu C, 1974).

$$r_{yx_1 \bullet x_2} = \frac{R_{yx_1x_2} \cdot r_{yx_1}}{\sqrt{r_{yx_1}^2 + r_{yx_2}^2}} = \frac{0,934 * 0.914}{\sqrt{0,914^2 + 0,862^2}} = 0,6795$$

$$d_{yx_1 \bullet x_2} = (0,680)^2 \cdot 100 = 46,2\,\%;$$

$$r_{yx_2 \bullet x_1} = \frac{R_{yx_1x_2} \cdot r_{yx_2}}{\sqrt{r_{yx_1}^2 + r_{yx_2}^2}} = \frac{0,934 * 0,862}{\sqrt{0,914^2 + 0,862^2}} = 0,641;$$

$$d_{yx_2 \bullet x_1} = (0,641)^2 \cdot 100 = 41,0\,\%$$

Method 2 (Merce E. et al., 2009; 2017):

And this method of distributing collinearity by factors is recommended in the literature [Merce E. At, al., 2009; 2017]. It includes the calculation of the determination of each factor as an average of the averages of all simple and partial determinations in all possible sequences for a particular causal complex. The judgments are graphically illustrated in Figure 1.

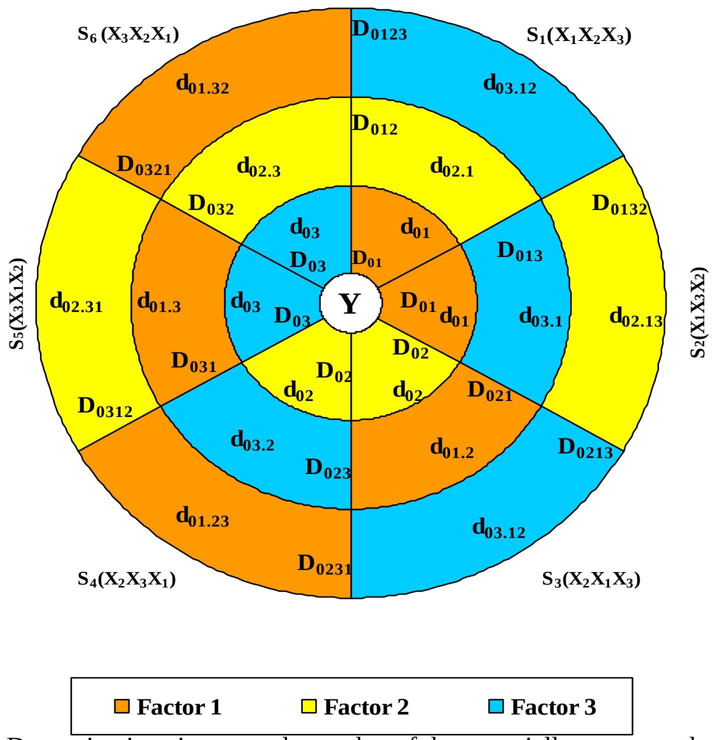

Fig.1 - Determinations in a causal complex of three partially auto-correlated factors

The calculation relations, respectively the calculations made according to the judgments of Method 2, are as follows:

a. The three-factor case:

$$r_{01\bullet 23} = \sqrt{\frac{R_{01}^2 + \frac{+\left(R_{012}^2 - R_{01}^2\right) + \left(R_{013}^2 - R_{03}^2\right)}{2} + \left(R_{0123}^2 - R_{023}^2\right)}{3}}$$

$$r_{02\bullet 13} = \sqrt{\frac{R_{02}^2 + \frac{+\left(R_{012}^2 - R_{01}^2\right) + \left(R_{023}^2 - R_{03}^2\right)}{2} + \left(R_{0123}^2 - R_{013}^2\right)}{3}}$$

$$r_{03\bullet 12} = \sqrt{\frac{R_{03}^2 + \frac{+\left(R_{013}^2 - R_{01}^2\right) + \left(R_{023}^2 - R_{02}^2\right)}{2} + \left(R_{0123}^2 - R_{012}^2\right)}{3}}$$

b. The two-factor case and the related calculations:

$$r_{yx_1 \bullet x_2} = \sqrt{\frac{r_{01}^2 + (R_{012}^2 - r_{02}^2)}{2}} = \sqrt{\frac{(0,914)^2 + [(0,934)^2 - (0,862^2]}{2}} = 0,694$$

$$d_{yx_1 \bullet x_2} = (0,694)^2 \cdot 100 = 48,2\,\%$$

$$r_{yx_2 \bullet x_1} = \sqrt{\frac{r_{02}^2 + (R_{012}^2 - r_{01}^2)}{2}} = \sqrt{\frac{(0,862)^2 + [(0,934)^2 - (0,914)^2]}{2}} = 0,624$$

$$d_{yx_2 \bullet x_1} = (0,624) \cdot 100 = 39,00\,\%$$

Method 3 (Merce E., et al., 2018)

Each method is based on a certain hypothesis, the differences in the operability of the calculations may be substantial. The method that we present has as a hypothesis the distribution of the autocorrelation on factors according to the principle of proportionality with the coefficients of the simple determination. The method is characterized by a high degree of promptness, with substantial facilities in integrating calculations.

The distribution of autocorrelation on the factors of influence implies the preliminary determination of the simple correlation coefficients and of the multiple correlation coefficient in the hypothesis of a certain theoretical regression model. Considering the database presented in Table 1, a linear bifactorial model was used to express the causal relationship between the two factors and the average production.

All calculations were performed using the Regression function of the Microsoft Excel Data Analysis component.

And this method assumes the determination of the bi-factorial model and of the mono-factorial models, respectively the coefficient of multiple correlation and of the simple correlation coefficients, the results being emphasized in the preamble of the three methods.

The introduction of the multiple correlation coefficient and the simple correlation coefficients in the centralized Excel table, which synthesizes the calculation steps of the pure determinative factor, automatically leads to the individualization of the influence of each factor (Table 4).

Table 4

Case of a linear multifactor model **)

Correlation and determination		Correlation coefficients	Determination Coefficients (%)	Percentage to one hundred (%)	Total and Factor Determination (%)
Sum of simple determinations		*	157.84	100.00	*
Simple correlation	Factor 1	0.914	83.54	52.93	46.17
	Factor 2	0.862	74.30	47.07	41.07
	Factor 3	0.000	0.00	0.00	0.00
	Factor 4	0.000	0.00	0.00	0.00
	Factor 5	0.000	0.00	0.00	0.00
Multiple correlation		0.934	87.24	*	87.24

**) The results presented in Table 2 as well as possible additional simulations can be checked by activating the table designed in Method 3 based on Microsoft Excel.

By comparison, the total determination and determinations of the two factors for the three methods are illustrated in Table 5.

Table 5

Comparative situation of total and factor determinations (%)

Factor	Method 1	Method 2	Method 3
X_1	46.2	48,2	46,2
X_2	41,0	39,0	41,0
X_1, X_2	87,2	87,2	87,2

For all three methods, the assignment of the total factor determination is complete. Moreover, factor determinations are identical for Methods 1 and 3. Method 3, however, has the great advantage of promptness and convenience of calculations. These features may be preferable to the processing of statistical data by specialists for attributing self-correlation to influence factors in incomplete databases

CONCLUSIONS

Colinarity is not a fiction. This is manifested in the context of the concrete realities caused by the impossibility of incorporating in experiments all the combinations of the many variants of the influence factors on the effect they are in a causal relationship. In such situations, the only way to individualise the pure influence of the factors is to distribute the collinearity according to working hypotheses with reasonable scientific support.

Concerns about the distribution of collinearity over factors of influence are numerous and have a substantial historical background (Moineagu C, 1974, Merce E. et al., 2009, 2017).

They are based on working hypotheses with appropriate scientific support, but the workload is impressive, making them even inapplicable in case of complex causal relationships with 3; 4; 5 or more factors of influence. The third method (Merce E., at al., 2018) is remarkable in terms of operability even in the case of particularly complex causal relationships, with only two steps to be taken. The first step is to run data using the Regression function of the Microsoft Excel Data Analysis component, assuming a specific regression model is used. The second step, recording the results obtained, in the first step, in the second column of the Excel table elaborated by the authors.

BIBLIOGRAPHY

1. Merce E. - Statistics - practical applications (1986), Babeş-Bolyai University, Faculty of Economic Sciences, Cluj-Napoca, p.92;
2. Merce E., C. C. Merce (2004) Elimination of autocorrelation in the case of a causal complex with incomplete data base, UASMV Scientific Bulletin, 61/2004, pp. 171-176
3. Merce E .. C. C. Merce - Statistics - Established paradigms and new paradigms (2009), AcademicPres Publishing House, Cluj-Napoca, p.311;
4. Merce E., C. C. Merce, Cristina B. Pocol (2017) - Autocorrelation - the prevalence of the identification of the cause of collinearity, Vol. Agrarian Economy and Rural Development - Realities and Perspectives for Romania, 2017", Editura ASE; ISSN-L 2247 – 7187; 7 pagini;
5. Merce E., M. Drăghici, M. Berca, C. Merce (2018) - metodă şi program pentru distribuirea autocorelaţiei pe factori de influenţă, Vol. Agrarian Economy and Rural Development - Realities and Perspectives for Romania, 2018", Editura ASE; 4 pagini;
6. Moineagu C. - Modeling Correlations in Economics, Scientific Publishing House, Bucharest, 1974.