Autocorrelation - Prevalence of identification of collinearity cause

Merce, Emilian and Merce, Cristian Calin and Pocol, Cristina Bianca

USAMV Cluj-Napoca, USAMV Cluj-Napoca, USAMV Cluj-Napoca

16 November 2017

A U T O C O R R E L A T I O N
- PREVALENCE OF IDENTIFICATION OF COLLINEARITY CAUSE-

Merce E., C. C. Merce, Cristina B. Pocol*
Corresponding author

Abstract: *The paper demonstrates that autocorrelation is an accidental statistical phenomenon, whose origin is the incomplete data base. It also shows that the attempts to redistribute factors interactions have focused on the development of methods of solving the effect rather than identifying the cause that generates collinearity. Three possible methods for collinearity removal are analysed comparatively. The premise for two of these methods is autocorrelation redistribution, and the third reveals the cause of collinearity and, implicitly, its cancellation. The three methods are named as follows:*

1. *Classic method [1,7];*
2. *Method of Merce E., Merce C.C.[6];*
3. *Method of Merce E., Merce C.C.[5];*

It is demonstrated that the first two methods are conventional approximations on the distribution of factors' interaction, with possible subjective consequences.

The ideal solution is the use of a complete data base. If this is not possible, as is often the case with databases of economic or sociological research, solving can be the completion of information with theoretical values, obtained by adjusting the causal relationship, in the hypothesis of a certain regression model, a procedure that represents, in fact and implicitly, a way of redistributing the interaction on the influence factors included in the causal model.

INTRODUCTION

Collinearity is an objective reality in the research of complex causal relationships, which externalises, as it will be demonstrated, 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 studied effect, on the other. The phenomenon of collinearity cannot, however, always be avoided. This is primarily about research in economics, sociology, psychology. Therefore, it seems natural to evaluate the collinearity and then correct the determination relationship between factors and effect. For this purpose, methods of individualizing the influence of each factor have been outlined, respectively by calculating the partial correlation coefficients [1,6,7]. It will be emphasized that such attempts, although rigorous from a methodological point of view, are working conventions and that neither of these methods leads to the actual numerical determination ratios between factors and effects, ratios which can only be obtained in the case in which there are complete information on the causal complex. If the specificity of the researches necessary requires the use of an incomplete data base, then, in our opinion, the way of redistribution of the factors' interaction must be solved through the integration of the data base.

MATERIAL AND METHOD

We appreciate that, in the construction of methods of separating factors' influence, in cases of incomplete information, a principle mistake was made by which enthusiasm pushes us to combat or adjust the effects and not to explain the causes that produce them. This explains the presence in specialty literature of many methods which, with a higher or lower dose of conventionalism, offer the possibility of deciphering collinearity and collinearity redistribution by factors. All these methods, however, fall under the scope of conventional or, even more severely, of approximation of research results in violation of scientific rigor. That is, in the search for causality, according to these methods, the cause of collinearity was not identified. What is, therefore, the cause that generates collinearity (interdependence) among factors?

Studies, observations and concrete processing are the grounds that lead us to the conclusion that the source of collinearity is the incomplete information on the way of the exteriorization of the effect under the influence of the investigated factors. In such a case, the effect

of collinearity, respectively autocorrelation does not occur if all the states of the resultant variable (y_{ij}) are known for all possible combinations of states comprised of the factorial variables (x_{1j}, x_{2j}). Any deviation from this imperative generates collinearity. Compliance with this requirement means complete experimental plans, including all possible combinations of predetermined factors variants.

In the case of socio-economic phenomena, where the experiment is often impossible, the only alternative is to fill in the information with data adjusted in the hypothesis of a certain regression model, based on incomplete data in the experiment.

And in the case of agricultural experiments, it happens often to encounter situations that only contain some of the possible combinations of influence factors variants. In this regard, it was assumed the following experimental plan for corn crops, which is aimed at the evolution of average production according to NP doses (Table 1).

Table 1The evolution of average corn production according to NP doses (conventional data)

Dose	Kg/ha	Dose	Kg/ha	Dose	Kg/ha	Dose	Kg/ha
N_0P_0	4600	$N_{50}P_{80}$	5865	$N_{100}P_{120}$	7725	$N_{150}P_{160}$	7920
N_0P_{40}	4945	$N_{100}P_{40}$	6095	$N_{150}P_{80}$	7820	$N_{200}P_{120}$	8050
$N_{50}P_{40}$	5980	$N_{100}P_{80}$	7590	$N_{150}P_{120}$	7935	$N_{200}P_{160}$	7915

RESULTS AND DISCUSSIONS

The picture of the possible combinations, respectively the corresponding average production, is shown in Table 2.

Table 2 The range of possible combinations of the five variants of each factor

X_1 / X_2	0	50	100	150	200
0	4600	?	?	?	?
40	4945	5980	6095	?	?
80	?	5865	7590	7820	?
120	?	?	7725	7935	8050
160	?	?	?	7920	7915

This is a typical example of incomplete information, which generates collinearity and all shortcomings related to redistribution. Correspondences between the levels of factors allocated and the average production obtained for data processing are presented in Table 3.

Table 3

X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y
0	0	4600	100	40	6095	150	120	7935
0	40	4945	100	80	7590	150	160	7920
50	40	5980	100	120	7725	200	120	8050
50	80	5865	150	80	7820	200	160	7915

In the case of the first two methods, from those mentioned, for autocorrelation redistribution, it is necessary to determine the correlation coefficients in the hypothesis of a certain theoretical regression model. To express the causal relation between the two factors and the average production, a linear bifactorial model was used. The bifactorial model is, at the same time, the starting point for calculating the adjusted values for completing the baseline data for the third method.Based on the hypothesis that the link could be expressed by a bifactorial, respectively a mono-factorial linear model, by processing the database, the following concrete forms of the models mentioned were obtained:

$$\bar{y}(x_1x_2) = 4985{,}1 + 11{,}83x_1 + 8{,}56x_2\ ;\ \ R_{yx_1x_2} = 0{,}934;\ D_{yx_1x_2} = 87{,}2\,\%$$

$$\bar{y}(x_1) = 5097{,}1 + 17{,}02x_1\ ;\ \ r_{yx_1} = 0{,}914;\ \ \ ;$$

$$\bar{y}(x_2) = 4978{,}9 + 21{,}82x_2\,; \quad r_{yx_2} = 0{,}862\,;$$
$$\bar{x}_1(x_2) = 7{,}02 + 1{,}12x_2\,; \quad r_{x_1x_2} = 0{,}824\,;$$

Taking into consideration the concrete form of the calculated regression models, it is possible to emphasize the methodological content of the three methods.

Method 1:

According to this method, the individualization of the influence of the two factors implies the redistribution of the interaction between them. For this purpose, it is recommended to calculate the coefficients of partial correlation, according to practices established in specialty literature [1,7].

$$r_{yx_1 \bullet x_2} = \frac{R_{yx_1x_2} \cdot r_{yx_1}}{\sqrt{r_{yx_1}^2 + r_{yx_2}^2}} = \frac{0{,}934 * 0.914}{\sqrt{0{,}914^2 + 0{,}862^2}} = 0{,}680$$

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

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

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

Method 2:

Another method of distributing collinearity by influence factors recommended in specialty literature [6], namely the calculation of partial correlation coefficients and partial determinations is based on the illustration of how determinations are calculated in a specific causal complex in all successions possible. The illustration of the judgments underlying this method of calculation is shown in Figure 1.

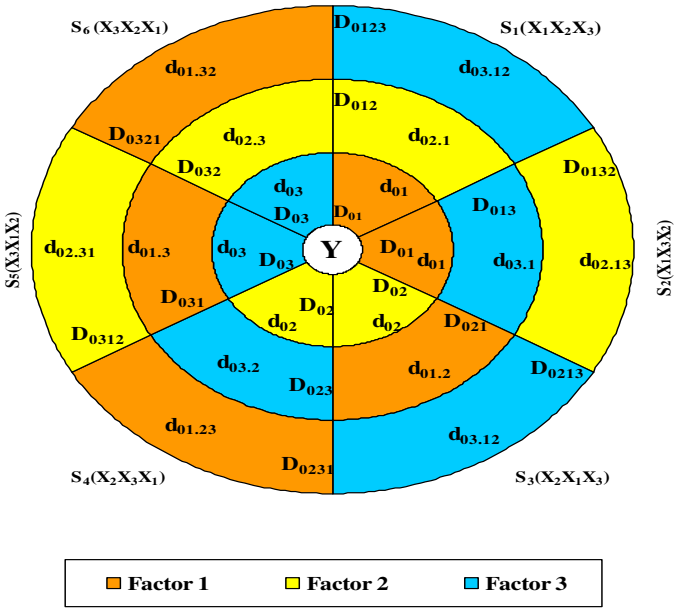

Fig.1 – Determinations in a causal complex of three partially autocorrelated factors

The calculation relations, respectively the calculations made according to the judgments presented in Figure 1, are as follows:

a. The general case:

The coefficient of partial correlation represents the square root of the average of determinations average explained step by step (iterative) in the context of a certain causal complex, in all possible successions, calculating according to the relationship:

$$r_{01\circ 23\ldots n} = \sqrt{\frac{A}{n}}$$

$$A = R^2_{01} + \frac{\left(R^2_{012} - R^2_{02}\right) + \ldots + \left(R^2_{01n} - R^0_{0n}\right)}{C^1_{n-1}} +$$
$$+ \frac{\left(R^2_{0123} - R^2_{023}\right) + \ldots + \left[R^2_{01(n-1)n} - R^2_{0(n-1)n}\right]}{C^2_{n-1}} +$$
$$+ \frac{\left(R^2_{01234} - R^2_{0234}\right) + \ldots + \left[R^2_{01(n-2)(n-1)n} - R^2_{0(n-2)(n-1)n}\right]}{C^3_{n-1}} + \ldots$$
$$\ldots + \frac{\left[R^2_{012\cdots(n-3)} - R^2_{02\cdots(n-3)}\right] + \ldots + \left[R^2_{015\ldots n} - R^2_{05\ldots n}\right]}{C^{n-4}_{n-1}} +$$
$$+ \frac{\left[R^2_{012\cdots(n-2)} - R^2_{02\cdots(n-2)}\right] + \ldots + \left[R^2_{014\ldots n} - R^2_{04\ldots n}\right]}{C^{n-3}_{n-1}} +$$
$$+ \frac{\left[R^2_{012\cdots(n-1)} - R^2_{02\cdots(n-1)}\right] + \ldots + \left[R^2_{013\ldots n} - R^2_{03\ldots n}\right]}{C^{n-2}_{n-1}} +$$
$$+ \left[R^2_{012\cdots n)} - R^2_{02\cdots n}\right]$$

b. The three-factors case:

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

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

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

c. The two-factors case and related processing:

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

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

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

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

Method 3

As in many other areas, scientists remain stuck in efforts to combat the effects, neglecting the decipherment of causes that produce unwanted effects. This is the case with collinearity. As a result of many applications and statistical processing by authors, there was a suspicion that autocorrelation could be caused by the incomplete data base. Remaining in the field of scientific speculation, it has been shown that interaction distribution could be done by filling in the missing information with the adjusted values of the regression model used. By generating the adjusted values, using the elaborated bifactorial model, the complete database is as shown in Table 4.

Table 4

X_2 \ X_1	0	50	100	150	200
0	**4600**	5487	6078	6670	7261
40	**4945**	**5919**	**6095**	7012	7604
80	5580	**6261**	**7590**	**7820**	7946
120	5922	6514	**7725**	**7935**	**8050**
160	6265	6856	7448	**7920**	**7915**

Through data processing, the following concrete forms of the bifactorial model and of the mono-factorial models were obtained:

$$\bar{y}(x_1 x_2) = 4919{,}3 + 11{,}70x_1 + 8{,}59x_2\,;\ \mathrm{R}_{\mathrm{yx}_1x_2} = 0{,}955;\ \mathrm{D}_{\mathrm{yx}_1x_2} = 91{,}20\,\%$$

$$\bar{y}(x_1) = 5606{,}8 + 11{,}70x_1\,;\ \mathrm{r}_{\mathrm{yx}_1} = 0{,}823;\ \ \mathrm{d}_{\mathrm{yx}_1} = 67{,}73\,\%$$

$$\bar{y}(x_2) = 6089{,}1 + 8{,}59x_2\,;\ \mathrm{r}_{\mathrm{yx}_2} = 0{,}4845;\ \ \mathrm{d}_{\mathrm{yx}_2} = 23{,}47\,\%$$

$$\bar{x}_1(x_2) = 100 + 0x_2\,;\ \mathrm{r}_{\mathrm{yx}_2} = 0;\ \ \mathrm{d}_{\mathrm{yx}_2} = 0\%$$

It can be noticed that, for the third method, the interaction does not operate, and the coefficients of the simple correlation are at the same time coefficients of the partial correlation, respectively reflecting the pure influence of each factor.

Synthetically, the aggregate influence and the separate influences of the two factors for the three methods are presented in Table 5.

Table 5 The comparative situation of total determination and by factors (%)

Factor's influence	Incomplete data base		Complete data base
	Method 1	Method 2	Method 3
X_1	46,2	48,2	67,73
X_2	41,0	39,0	23,47
X_1, X_2	87,2	87,2	91,20

For all three methods, the assignment of the total determination by factors is complete, but not unique. Moreover, the total determination is the same for the first two methods, but different for the third.

Method three confirms the truth that autocorrelation is generated by incomplete data bases, but, even in this case, it is assumed that total determination and true partial determinations can only be obtained in the case of the complete data base, obtained through the experimental plan.

B I B L I O G R A P H Y

1. Merce E. – Statistică – aplicaţii practice, Babeş-Bolyai University, Faculty of Economics, Cluj-Napoca, 1986, p.92
2. Merce E. – Cu privire la calculul coeficientului corelaţiei parţiale (I). Definire, conţinut. Studia 1/1989, Babeş-Bolyai University, Cluj-Napoca, pag. 51-60.
3. Merce E., V. Pârv – O nouă metodă de determinare a coeficientului corelaţiei parţiale. Revista de Statistică nr. 3/1991, pag. 32-39.
4. Merce E., B. Pârv, Flavia Laun - Calculul coeficienţilor determinaţiei parţiale. Metodă şi program. Romanian Journal of Statistics no. 6, 1992, pag. 29-37.
5. Merce E., C. C. Merce (2003) – Eliminarea autocorelaţiei dintre variabilele cauzale prin întregirea bazei de date, in the volume Specializare, dezvoltare şi integrare; ISBN; 973-86547-4-2; pag,275-281, Babeş-Bolyai University, Faculty of Economics, Cluj-Napoca
6. Merce E.. C. C. Merce – Statistică – paradigme consacrate şi paradigme întregitoare, AcademicPres Publishing House, Cluj-Napoca, 2009, p.311;
7. Moineagu C. - Modelarea corelaţiilor în economie, Scientific Publishing House, Bucharest, 1974.