DOI https://doi.org/10.15407/csc.2022.04.047
UDC 519.816

O.V. BABAK, PhD (Eng.), Senior Researcher, Department of Ecological Digital Systems, International Research and Training Center for Technologies and Systems of the NAS and MES of Ukraine, Acad. Glushkov ave., 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0002-7451-3314, dep115@irtc.org.ua, babak@irtc.org.ua

O.E. TATARINOV, Researcher of the Ecological Digital Systems Department, International Research and Training Center for Information Technologies and Systems of the National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, 40, Academician Glushkov av., Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0001-7206-6859, E-mail: dep115@irtc.org.ua, al.ed.tatarinov@gmail.com

A.K. SIERIEBRIAKOV, PhD Student, Researcher of Intellectual Control Department, International Research and Training Center for Information Technologies and Systems of the National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, 40, Academician Glushkov av., Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0003-3189-7968, E-mail: sier.artem1002@outlook.com

I.M. YAKOVENKO, Researcher of Intellectual Automatic Systems Department, International Research and Training Center for Information Technologies and Systems of the National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, 40, Academician Glushkov av., Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0002-4477-3254, E-mail: yakvan@ukr.net

THE QUASI-ORTHOGONALIZATION APPROACH TO SOLVING THE MULTICOLLINEARITY PROBLEM OF EMPIRICAL DATA

This article proposes an approach to solving the problem of regressors multicollinearity using the procedure of quasiorthogonalization of data. The specified approach is based on the transformation of factors during their coding according to the rules of a full factorial experiment. It is shown that the proposed coding of factors leads to a reduction of multicollinearity of the data. This approach can be used both for building models based on short samples and for batch processing of Big Data.

Keywords*: multicollinearity problem, quasi-orthogonalization procedure, coding of input data, full factorial experiment.*

Introduction

One of the conditions for the correct application of regression analysis when identifying regularities in empirical data is the absence of multicollinear regressors. The latter are called multicollinear if they are correlated almost linearly [1,2]. Let's briefly consider the features and consequences of the multicollinearity problem.

A classic multivariate regression model or multiple regression model looks like this

$$y = a_0 \cdot x_0 + \sum_{j=1}^{n} a_j \cdot x_j + \varepsilon, \quad x_0 = 1, \qquad (1)$$

where y is the dependent variable, x is an independent variable, a_0, a_j are estimates of regression coefficients, n is their number, ε is a random (stochastic) component. At the same time, linear independence of the columns of the regressor matrix X is assumed, which is the matrix $(X^T \cdot X)$. If this condition is violated, that is, when one of the columns of the matrix is a linear combination of the

others, full collinearity occurs and it is impossible to solve the normal equations. In practice, it rarely occurs and most often we have to face a situation when there is a high level of correlation between regressors and $(X^T \cdot X)$ is close to being singular, i.e. $\det(X^T \cdot X) \approx 0$. In this case it is said that multicollinearity is present, which can appear for various reasons, for example, from a time series x_j. Let us list some of the most important common causes of multicollinearity:

– factors that duplicate each other in terms of content are included in the modeling function;

– total value of indicators, which model contains as factors, is constant;

– factors characterizing the same side of the phenomenon are included in the model;

– factors that are constituent elements of each other are used in the model;

– spurious correlation between several independent variables due to the fact that they are all expressed through some other independent variable (this problem is not only typical for Big Data, but also for any amount of data).

The problem of multicollinearity is common in time series regression, that is, the data consists of a number of observations over a period of time.

Let us highlight the most typical consequences of multicollinearity:

– a small change in the input data (for example, the addition of new observations) leads to a significant change in the coefficients of the model;

– estimates have large standard errors, low significance, although the model is generally significant (high value of the coefficient of determination R^2 and the corresponding F-statistic);

– the interpretation of coefficient estimates from the point of view of significance becomes very problematic and their estimates have incorrect signs from the point of view of theory or unreasonably large values.

There are various approaches to overcome multicollinearity, for example, excluding the factor(s) from the model. However, it is not clear which factors are redundant and their removal may affect the substantive meaning of the model. Sometimes a possible solution to this problem is to increase the size of the sample, which leads to a decrease in the variance of the coefficient estimates and, therefore, the pairwise covariance values.

In connection with the mentioned consequences of multicollinearity, if their influence is not reduced or eliminated, the main goal of machine data processing - machine learning (ML), which consists in the automatic detection of hidden patterns, becomes impossible. These regularities and the knowledge extracted from them allow to understand the essence of the researched process and, relying on the available data, to predict new facts [3,4]. The above-mentioned task becomes important in connection with the expectation of an unprecedented increase in the amount of information coming from the Internet of Things (IoT) and the Industrial Internet of Things (IIoT). According to analysts, the entire world volume of information will reach 500 zettabytes by 2025 (1 zettabyte is 10^{12} gigabytes). Thus, we are talking about the socio-economic phenomenon of Big Data [5].

In [6], the two most effective methods of coping with multicollinearity in data are given.

The most radical technique for solving the multicollinearity problem is a linear transformation of variables, which leads to new orthogonal variables. Usually we are talking about conversion to main components. However, they, being linear combinations of all independent variables x, are poorly subjected to meaningful interpretation, and therefore the values of the regression coefficients for the main components tell the researcher little about the influence of the input variables on the dependent variable y. A common method of coping with multicollinearity is the use of ridge regression for regression estimation. However, in this case, the estimates are biased and this method should be used with caution. In [6,7], a method of significantly reducing multicollinearity with incomplete orthogonalization of input variables by introducing new variables is considered, which makes it possible to effectively obtain predictive estimates of the dependent variable. Therefore, the use of this method, apparently, does not always contribute to meaningful interpretation of regression estimates.

Current article proposes an approach to solving the multicollinearity problem in the tasks of identifying regularities in empirical data with qua-

si-orthogonalization of input variables. This approach to a certain extent guarantees the solution of the specified problems and always contributes to meaningful interpretation of regression estimates.

Problem Statement

Let there be n natural factors of passive experiment x and response with the number of their observations l represented by matrices X and Y. Let's write down the system of linear equations

$$Y = A \cdot X,$$

where A is the matrix of unknown coefficients. Let $\det(X^T \cdot X) << 1$, and respectively, there is multicollinearity of the input data X. The multiple regression model (1) is found by determining coefficient estimates

$$A = (X^T \cdot X)^{-1} \cdot X^T \cdot Y.$$

It is necessary to find an approach to solving the problem of multicollinearity using quasi-orthogonalization of input variables based on a hypothetical (anticipated) full factorial experiment (FFE) [8]. At the same time, compare the determinants of correlation matrices for normalization and coding according to the rules of the FFE of the matrix X, i.e., respectively, X_n and X_c (indices "***n***" and "***c***" in the further explanation, accordingly, denote normalization and coding operations).

Solution of the Problem

The optimal solution to the stated task of solving the multicollinearity problem with a regularity hidden in the passive experiment data is orthogonalization of the matrix according to the rules of FFE [8]. However, the very nature of the appearance of certain data (especially Big Data) usually precludes conducting FFE, and the quality of identifying regularities can only be discussed at the hypothetical level. Meanwhile, the quality of possible results with FFE is impressive. Thus, with orthogonal planning, if $N = 2^n$ (N is the number of experiments), the matrix of factors is obtained and, accordingly, a symmetric matrix of normal equations $M = X^T \cdot X \cdot s_y^2$, where always $\det M \neq 0$ (s_y^2 is an estimate of the variance of reproducibility). Its inverse covariance matrix $M^{-1} = (X^T \cdot X \cdot s_y^2)^{-1}$ is also symmetric, on the main diagonal of which there are estimates of variances of regression coefficients, and outside it are estimates of covariances between all pairs of regression coefficients, which are equal to zero. This remarkable property, due to the complete elimination of multicollinearity, determines the independence of the estimates of the regression coefficients. It allows to draw a conclusion about the possibility of removing non-informative independent variables x, which are determined by the significance of the corresponding estimates of the regression coefficients α without harm to the meaning of the model interpretation. An important circumstance should be noted, that knowing M^{-1}, it is also possible to find the correlation matrix R, that is, the matrix of pairwise correlation coefficients ρ.

The solution to the given problem is based on two statements derived from the theory of applied mathematical statistics [9].

Statement 1. Let there be a matrix of natural values of a passive volume experiment l and its regressors are known in the columns x_j, $x_{j\max}$ and $x_{j\min}$ respectively, $j = \overline{1,n}$ and coding of regressors is carried out according to the rules of FFE [8], i.e.

$$x_{j\mathbf{c}} = \frac{x_j - 0{,}5 \cdot (x_{j\max} + x_{j\min})}{0{,}5 \cdot (x_{j\max} - x_{j\min})}, \qquad (2)$$

then there is always a matrix of hypothetical FFE $N=2^n$.

Proof. Based on the reconstructed regression (1) of data X of the volume l coded using the ratio (2), it is possible to establish the results of observations of the hypothetical orthogonal planning matrix by calculating $N=2^n$. Thus, the specified circumstance confirms the validity of the statement.

Statement 2. Let there be some independent random sample of natural values of independent variables (factors) x and the results of observations of dependent variables y

$$\{x_{ji},\ y_i\},\ j = \overline{1,n},\ i = \overline{1,l}, \qquad (3)$$

that is, the matrices X and Y are given. In (3) n is the number of factors, l is the size of the sample.

Let multiple regression (1) be reconstructed on the basis of data (3) using normalization procedures, for example $x_{j\mathbf{n}} = \dfrac{x_j}{x_{j\max}}$, and coding using the ratio (2).

Table 1 . **Sample V_1** .

x_1	x_2	x_3	x_4
3,21	5,39	1,74	8,86
3,50	5,65	1,79	8,48
3,45	5,38	1,29	8,21
2,95	5,93	1,75	8,96
3,25	4,98	1,99	8,65

Table 2 . **Sample V_2**.

x_1	x_2	x_3	y
3,21	5,39	1,74	8,86
3,50	5,65	1,79	8,48
3,45	5,38	1,29	8,21
2,95	5,93	1,75	8,96
3,25	4,98	1,99	8,65
3,35	5,56	1,82	8,53
2,97	5,32	1,67	8,39
3,34	5,54	1,62	8,69
3,06	5,40	1,80	8,25
3,07	5,35	1,94	8,29
3,41	5,54	1,48	8,62
2,97	5,82	1,43	8,79
3,02	5,60	1,56	8,27
3,48	5,37	1,73	8,91
2,96	5,15	1,93	8,24
3,31	5,78	1,41	8,44
3,33	5.70	1,69	8,57
3,07	5,61	1,72	8,85
3,17	5,04	1,87	8,48
3,25	5,61	1,66	8,33

At the same time, the determinants of the matrices $M_{\text{н}} = X_{\text{н}}^T \cdot X_{\text{н}}$ and $M_{\text{с}} = X_{\text{с}}^T \cdot X_{\text{с}}$:

$$\det(M_{\text{н}}) \neq 0, \quad \det(M_{\text{с}}) \neq 0. \tag{4}$$

That is, there are inverse variance-covariance matrices $M_{\text{н}}^{-1}$ and $M_{\text{с}}^{-1}$.

Then, under the condition of the existence of partial multicollinearity in the data, the inequality is always valid

$$\det(R_{\text{с}}) > \det(R_{\text{н}}), \tag{5}$$

where R_n and R_c are correlation matrices.

Proof. Since, taking into account available in sample (3) values $x_{j\max}$ and $x_{j\min}$, $j=\overline{1,n}$, the existence of a hypothetical FFE of the type $N=2^n$ is possible. In this case M_N is a scalar matrix whose determinant is equal to $\det(M_N) = (2^n)^n$ and has the following property.

Property. If in data $x_{j\text{н}}$, $j=\overline{1,n}$ there is a partial multicollinearity, then $\det(M_{\text{н}}) < 1$ and its value approaches zero, $\det(M_{\text{с}}) > 1$ and its value in the limiting case, with a hypothetical FFE, approaches $(2^n)^n$.

Thus, we have

$$\det(M_{\text{н}}) < \det(M_{\text{с}}). \tag{6}$$

The validity of the specified property is explained by the fact that the coded values $x_{j\text{с}}$, $j=\overline{1,n}$ are located in the FFE area, which is limited by the hypersphere lying on the vertices of the hypercube, the value of which is equal to ± 1 [8].

Given (4), there exist inverse matrices $M_{\text{н}}^{-1} = (X_{\text{н}}^T \cdot X_{\text{н}})^{-1}$ and $M_{\text{с}}^{-1} = (X_{\text{с}}^T \cdot X_{\text{с}})^{-1}$. If $\det(M^{-1}) = \dfrac{1}{\det(M)}$, we have $M_{\text{н}}^{-1} > 1$, while $M_{\text{с}}^{-1} < 1$. Note that in the limiting case, the FFE M_N^{-1} is a scalar matrix whose determinant is equal to

$$\det(M_N^{-1}) = \left(\frac{1}{2^n}\right)^n. \tag{7}$$

Therefore, the value $\det(M_{\text{с}}^{-1})$ tends to the limit value $\det(M_N^{-1})$, i.e. almost to zero. Given the above property, the inequality naturally holds

$$\det(M_{\text{с}}^{-1}) < \det(M_{\text{н}}^{-1}). \tag{8}$$

Taking into account $M_{\text{н}}^{-1}$ and $M_{\text{с}}^{-1}$ respectively, the correlation matrices R_c and R_n can be obtained. At the same time, in the limiting case (hypothetical FFE) $\det(R_{\text{с}}) = 1$ and under full multicollinearity $\det(R_{\text{н}}) = 0$. Note that at the same time

$$\det(R_{\text{с}}) \to 1, \quad \det(R_{\text{н}}) \to 0. \tag{9}$$

So, $\det(R_{\text{с}}) > \det(R_{\text{н}})$ and the statement is proved.

Statements 1 and 2 form the basis of the proposed approach to solving the problem of multicollinearity in the input natural data of a passive experiment given by a rectangular matrix of size $L = l \times n$.

The essence of the mentioned approach is to encode the input data according to formula (2),

immersing them in the region of the hypothetical FFE and thus performing the quasi-orthogonalization procedure.

It is advisable to consider the results of the mentioned approach on the following simple example.

Let samples 1 and 2 of some passive experiment be known, represented by artificially generated matrices of input natural values x $L_1 = 5 \times 3$ and $L_2 = 20 \times 3$ and feedback y within bounds: $x_1 = [2{,}95;\ 3{,}5], x_2 = [4{,}98;\ 5{,}93], x_3 = [1{,}29;\ 1{,}99]$; $y = [8{,}21;\ 8{,}96]$.

$\hat{y}_{\mathbf{n}} = 10{,}09 - 3{,}24 \cdot x_{1\mathbf{n}} + 0{,}81 \cdot x_{2\mathbf{n}} + 0{,}96 \cdot x_{3\mathbf{n}}$.

$\hat{y}_{\mathbf{c}} = 8{,}64 - 0{,}25 \cdot x_{1\mathbf{c}} + 0{,}07 \cdot x_{2\mathbf{c}} + 0{,}17 \cdot x_{3\mathbf{c}}$.

$\hat{y}_{\mathbf{n}} = 4{,}80 + 0{,}63 \cdot x_{1\mathbf{n}} + 2{,}69 \cdot x_{2\mathbf{n}} + 0{,}80 \cdot x_{3\mathbf{n}}$.

$\hat{y}_{\mathbf{c}} = 8{,}51 + 0{,}05 \cdot x_{1\mathbf{c}} + 0{,}22 \cdot x_{2\mathbf{c}} + 0{,}14 \cdot x_{3\mathbf{c}}$.

Processing of input data was carried out by Matlab programs. Normalization was carried out according to ratios $x_{j\mathbf{n}} = \dfrac{x_j}{x_{j\max}}$, and coding according to (2). The obtained results are summarized in Table 3. In it $\rho_{\max}$ is the maximum value of the pairwise correlation coefficient in the matrices R_n and R_c.

Assuming that the matrix R contains $\rho > 0{,}7$, then multicollinearity exists in this multiple regression model. Also, it is possible to assume that in each case (Tab. 3) the connection is strong during normalization, and when coding, the connection is moderate $\rho = (0{,}3 - 0{,}5)$ and multicollinearity is almost absent [9].

The obtained practical results of the computational experiment presented in Tab. 3, confirm the theoretical side of the considered approach to solving the multicollinearity problem without active FFE. Comparing the results of input data processing (Tab. 3), it is possible to draw conclusion about the validity of Statement 1, since for matrices of dimensions L_1 and L_2 there is an inequality $\det(M_{\mathbf{n}}) < \det(M_{\mathbf{c}})$. At the same time $\det(M_{\mathbf{n}}) < 1$, and $\det(M_{\mathbf{c}}) > 1$, which indicates the presence of multicollinearity of the data of the normalized matrices L_1 and L_2 and its absence in the case of their coding. The validity of Proposition 2 is determined by the fact that for the given matrices L_1 and L_2 (Tab. 3) there is an inequality $\det(R_{\mathbf{c}}) > \det(R_{\mathbf{n}})$. At the same time $\det(R_{\mathbf{c}}) \to 1$, and $\det(R_{\mathbf{n}}) \to 0$, which indicates the elimination of multicollinearity in the case of input data encoding.

Table 3. **Processing of input data.**

M \ L	$L_1 = 5 \times 3$		$L_2 = 20 \times 3$	
$\det(M)$	**n**	**c**	**n**	**c**
$\det(M^{-1})$	0,015	10,82	0,362	199,07
$\det(R)$	67,44	0,09	2,745	0,005
$\rho_{\max}$	0,008	0,7341	0,0044	0,7011

The disadvantage of the stated approach, not in the case when samples are small ($\dfrac{l}{n+1} < 10$), but the construction of a technology for eliminating multicollinearity during batch processing of the database is considered, there is necessity to identify the maximum and minimum regressors value. However, it can be overcomed in the case of homogeneity of data by explicitly setting limiting values for regressors, which are determined by practical, theoretical or heuristic means. Nevertheless, this issue, as well as the issue of the quality of the resulting models after factors coding, especially after removal of uninformative regressors, requires additional study, because it is the subject of a separate study.

In conclusion, it should be noted that the mathematical transition from the coded values of the factors to their natural values is carried out in the usual way and is not a challenging task.

Conclusions

An approach to solving the multicollinearity problem of regressors using the quasi-orthogonalization procedure of data is proposed. The specified approach is based on the transformation of factors during their coding ac-

cording to the rules of FFE. A significant distinguishing feature of the procedure is that the results of real FFE are absent and exist only hypothetically. However, it is shown that the proposed coding of factors leads to a reduction of multicollinearity in the data.

The specified effect is substantiated by known theoretical basics of correlation analysis. With the help of a computer experiment, a comparison of correlation matrices obtained during normalization and coding of artificially generated data of a passive experiment was carried out. The results of the specified comparison showed the presence of the effect of significantly reducing the multicollinearity of the data.

The proposed approach to solving the multicollinearity problem can be used both for building models based on short samples and for batch processing of databases.

REFERENCES

1. *Draper, N., Smith, H.*, 2007. Applied Regression Analysis. Transl. from English, M.: Publishing house "Williams", 912 p. (In Russian).
2. *Magnus, Ya.R., Katyshev, P.K., Peresetskiy, A.A.*, 2004. Econometrics, Initial course: Textbook. 6th ed., Rev. and add, Moscow: Delo, 576 p. ISBN 5-7749-0055-X (In Russian).
3. *Zagoruyko, N.G.*, 1999. Prikladnyye metody analiza dannykh i znaniy ["Applied methods of data and knowledge analysis"], Novosibirsk: IM SO RAN, 270 p. (In Russian).
4. *Zagoruyko, N.G.* et al., 2013. "Obnaruzheniye zakonomernostey v massivakh eksperimental'nykh dannykh", Vychislitel'nyye tekhnologii, Volume 18, Spetsial'nyy vypusk, pp. 12-20. (In Russian).
5. *Gritsenko, V.I., Surovtsev, I.V., Babak, O.V.,* 2021. "Peculiarities of interconnection 5G, 6G networks with Big Data, Internet of Things and Artificial Intelligence", Cyb. And Comp. Eng., No. 2 (204), pp. 5-19. DOI : https://doi . org /10.15407/ kvt 204.02.005 (In Ukrainian).
6. *Orlova, I.V.*, 2019. "The approach to solving the problem of multicollinearity by using the transformation of variables", Fundamental research, No. 5. [online]. Available at: <https://s.fundamental-research.ru/pdf/2019/5/42464.pdf> (Last accessed: 23 December 2021). (In Russian).
7. *Orlova, I.V.*, 2019. " Korrektirovka spetsifikatsii modeli mnozhestvennoy regressii pri nalichii mul'tikollinearnosti iskhodnykh regressorov ", V knige: Upravleniye razvitiyem krupnomasshtabnykh sistem (MLSD'2019) , Materialy Dvenadtsatoy mezhdunar. Conference, Oct. 13, 2019, Moskva , Nauchnoye elektronnoye izdaniye, M.: IPU RAN, pp. 993- 995 (In Russian).
8. *Adler, Yu.V., Markova, Ye.V., Granovskiy, Yu.V.*, 1976. Planirovaniye eksperimenta pri poiske optimalnykh usloviy, Monograph, M.: Nauka, 280 p. (In Russian).
9. *Kobzar, A.I.,* 2006. Applied mathematical statistics, For engineers and scientists, Moscow: FIZMATLIT Publ., 816 p. (In Russian).

Received 19.01.2022

ЛІТЕРАТУРА

1. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Пер. с англ. М.: Издательский дом «Вильямс». 2007. 912 с.
2. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс: Учеб. 6-е изд., перераб. и доп. М.: Дело. 2004. 576 с. ISBN 5-7749-0055-X.
3. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН. 1999. 270 с.
4. *Загоруйко Н.Г.* и др. Обнаружение закономерностей в массивах экспериментальных данных. Вычислительные технологии. Том 18, Специальный выпуск, 2013. С. 12 -20.
5. *Гриценко В.І., Бабак О.В., Суровцев І.В.* Особливості взаємозв'язку мереж 5G, 6G з великими даними, інтернетом речей та штучним інтелектом. Кібернетика та обчислювана техніка. 2021. № 2 (204). С. 5 19. DOI: https://doi.org/10.15407/kvt204.02.005.
6. *Орлова И.В.* Подход к решению проблемы мультиколлинеарности с помощью преобразования переменных. Фундаментальные исследования. 2019. № 5. С. 78-84.

7. *Орлова И.В.* Корректировка спецификации модели множественной регрессии при наличии мультиколлинеарности исходных регрессоров. В книге: Управление развитием крупномасштабных систем (MLSD'2019). Материалы Двенадцатой междунар. конфер., 1 3 окт. 2019 г., Москва. Научное электронное издание. М.: ИПУ РАН. 2019. С. 993-995.
8. *Адлер Ю.В., Маркова Е.В., Грановский Ю.В.* Планирование эксперимента при поиске оптимальных условий. Монография. М.: Наука. 1976. 280 с.
9. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2005. 816 с.

Надійшла 19.01.2022

О.В. Бабак, кандидат технічних наук, старший науковий співробітник, Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України, 03187, м. Київ, просп. Академіка Глушкова, 40, Україна, ORCID: https://orcid.org/0000-0002-7451-3314, dep115@irtc.org.ua, babak@irtc.org.ua
О.Е. Татарінов, науковий співробітник, Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України, 03187, м. Київ, просп. Академіка Глушкова, 40, Україна, ORCID: https://orcid.org/0000-0001-7206-6859, dep115@irtc.org.ua, al.ed.tatarinov@gmail.com
А.К. Сєребряков, аспірант, молодший науковий співробітник, відділ інтелектуального управління, Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України, 03187, м. Київ, просп. Академіка Глушкова, 40, Україна, ORCID: https://orcid.org/0000-0003-3189-7968, sier.artem1002@outlook.com
І.М. Яковенко, науковий співробітник, відділ інтелектуальних автоматичних систем, Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України, 03187, м. Київ, просп. Академіка Глушкова, 40, Україна, ORCID: orcid.org/0000-0002-4477-3254, yakvan@ukr.net

ПІДХІД ДО ВИРІШЕННЯ ПРОБЛЕМИ МУЛЬТИКОЛІНЕАРНОСТІ ЗА ДОПОМОГОЮ КВАЗІОРТОГОНАЛІЗАЦІЇ ЕМПІРИЧНИХ ДАНИХ

Вступ. Однією з умов коректного застосування регресійного аналізу при виявленні закономірностей в емпіричних даних є відсутність мультиколінеарних регресорів. Існують різні підходи подолання мультиколінеарності, проте незрозуміло, які фактори виявляються зайвими та їх видалення може позначитися на змістовному сенсі моделі. Якщо не послабити або не позбутися впливу мультиколінеарності, то стає неможливою головна мета машинної обробки даних – машинне навчання. Приховані в даних закономірності й знання, що витягуються з прихованих закономірностей, дозволяють зрозуміти сутність досліджуваного процесу і на основі наявних даних передбачати нові факти. Вищезазначена задача набуває важливого значення у зв'язку з очікуванням небаченого зростання обсягу інформації, що надходить від Інтернету Речей та Промислового Інтернету Речей. Поширеним прийомом боротьби з мультиколінеарністю є застосування рідж-регресії для оцінки регресії. Однак при цьому оцінки виходять зміщеними і користуватися цим методом потрібно обережно.

Мета статті – створення способу вирішення проблеми мультиколінеарності в задачах виявлення закономірностей в емпіричних даних, що сприяє змістовній інтерпретації оцінок регресії.

Методи. Для реалізації вирішення проблеми мультиколінеарності використовувався метод квазіортогоналізації вхідних змінних на основі гіпотетичного повного факторного експерименту (ПФЕ).

Результат. Запропоновано підхід до вирішення проблеми мультиколінеарності регресорів за допомогою процедури квазіортогоналізації даних, що базується на перетворенні факторів при їх кодуванні за правилами ПФЕ. Показано, що запропоноване кодування факторів призводить до зменшення мультиколінеарності даних. Зазначений ефект обґрунтовано відомими теоретичними положеннями кореляційного аналізу.

Висновки. Підсумки дослідження, представлені в цій статті, показують можливість побудови інформаційної технології усунення мультиколінеарності як при побудові моделей за короткими вибірками, так і при пакетній обробці Великих Даних.

Ключові слова: *проблема мультиколінеарності, процедура квазіортогоналізації, кодування вхідних даних, повний факторний експеримент.*