УДК 519.21 : 519.24

V.V. KROKHIN

STABILITY INVESTIGATION OF THE METHODS OF SELECTION OF OPTIMUM MODEL OF MULTIPLE LINEAR REGRESSION IN THE CASE WHEN THE INDEPENDENT VARIABLES ARE QUASICOLLINEAR.

Annotation. *There has been undertaken an analysis of the effectiveness of choice of optimal model of multiple linear regression with a strong linear relationship between the input variables. The study was conducted using a simulation method. It was conducted by analyzing five methods for selecting the optimal MLR (method of all possible regressions with the corrected coefficient of determination as an optimality criterion, the method of all possible regressions with the corrected coefficient of determination as benchmarks, and assessment of the significance of MDR ratios (based on t-statistics) method of all possible regressions using statistics of Mallouza as an optimality criterion, successive elimination method, incremental method) in terms of their stability under conditions of multi-collinearity. The study was conducted with the help of specially designed tion in the environment MATLAB software.*

Keywords: multiple linear regression, optimal model, optimality criterion, multicollinearity, simulation.

Introduction. It is known that a serious problem in estimating coefficients in the multiple regression model is the linear relationship between the presence of certain input variables. [1,4,5]. Moreover, the presence of such a relationship is not evident if the values of one of the variables is the combination of the values of several other variables. [5]. In practice, strict (exact) multicollinearity is unlikely, if only because that there are any data errors that are random. Therefore it is necessary to talk about the possible presence kvazimultikollinearnosti between input variables. Significant interest is the study of the effect of multicollinearity on the quality of estimation of coefficients model of multiple linear regression. In particular, it is important to understand how multicollinearity affects the stability of various algorithms for selecting an optimal, in the sense of a set of input actions, the linear regression model.

Analysis of publications on the subject of research. Consider the model of multiple linear regression (MLR):

$$\mathbf{Y_i} = \beta_0 + \beta_1 \mathbf{X_{1i}} + ... + \beta_k \mathbf{X_{ki}} + \mathbf{U_i}, \quad \mathbf{i} = 1,2,...,\mathbf{n} \tag{1}$$

here i indicates the number of observations X_j, $j = 1,2,...,k$ - independent (input, predictor) variables Y - dependent (output) variable, U_i - disturbance (error patterns), $\beta_j, j = 1,2,...,k$ - coefficients of model.

The columns X_j, $j = 1,2,...,k$ of the matrix X of the input variables are called linearly independent if the equation

$$\lambda_1 \mathbf{X}_1 + \lambda_2 \mathbf{X}_2 + ... + \lambda_\mathbf{k} \mathbf{X}_\mathbf{k} = \vec{0},$$

(2)

is executed if and only if $\lambda_\mathbf{j} = 0$ for all $\mathbf{j} = 1,2,..., \mathbf{k}$. In other words, if the columns of the matrix are linearly independent, or one column can be represented as a linear combination of the other columns of the matrix. If in (2) not all $\lambda_\mathbf{j}$ are equal to zero, then the columns of the matrix are linearly dependent or multikollinear.

In practice the presence multikollinearity means that the determinant of the matrix $(\mathbf{X}'\mathbf{X})$ equals zero and, as a consequence, matrix $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist, in other words evaluation of the least squares method (GLS) ratios RLL (1)

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

(3)

is impossible to calculate. This conclusion shows how serious is the problem of multicollinearity.

In practice, however the exact (or strict) collinearity between the data is rare. Even if the actual data is strictly collinear, measurement errors always having a random component lead to some violation of the exact collinearity. In other words, it is expected that between the actual data used as the independent variables, can exist dependence, expressed as follows:

$$\lambda_1 \mathbf{X}_1 + \lambda_2 \mathbf{X}_2 + ... + \lambda_\mathbf{k} \mathbf{X}_\mathbf{k} + \nu = \vec{0},$$

(4)

where not all $\lambda_\mathbf{j}$ equal zero, and ν is a random variable.

If the columns of the matrix are related (4), then we say that there is an almost linear relationship, or approximately collinear (or quasicollinear) between the independent variables.

Note that collinearity is easily noticed when there are only two independent variables, as in this case the values of the variables are proportional to (or nearly proportional to) each other. However, when the number of independent variables is large, "visually" noted the presence of multicollinearity is not possible. It should be noted also that the problem of multicollinearity is a problem precisely of a strict linear relationship between the variables. If there is a strong, but non-linear dependence, evaluation of the least squares method is possible, but the standard errors of estimates of the coefficients are large. [1,5]

Note that the presence of an almost linear relationship between the independent variables does not violate the optimal properties of GLS. That is, these estimates, in the case when other hypotheses of the classical linear regression model, are the best linear unbiased estimates (NLNO). [1] This is an important conclusion, which shows that there is no other method of linear estimation, allowing to obtain the best results in terms approximate multicollinearity than the OLS.

Availability quasi multicollinearity leads to the fact that some of the standard errors of estimates of linear regression coefficients become very large. [1] Moreover, as we approach the strict collinearity, these errors tend to infinity. As a result, there is too much variability in sample estimates (estimates for the same model vary greatly from sample to sample, making them extremely unreliable). n practice this means that small changes in the input data leads to unnecessarily large changes estimates of regression coefficients. On the other hand, larger values standard error (standard deviations) lead to small values of the statistic t for MLR coefficients estimates. As a result they are not statistically significant in the case of testing the null hypothesis. That is the true value of the regression coefficients, which estimates have a large standard error, are tested as zero (in other words, the relevant independent variable is tested as not to affect the behavior of the dependent variable). At the same time it should be noted that along with the insignificant estimates for some of the coefficients, the value of the coefficient of determination $\mathbf{R}^2$ can be sufficiently large ($\mathbf{R}^2 > 0.9$). [2,5] This shows that in general, the model agrees well with the available data. We emphasize once again that all of the above applies to the case quasicollinearity, given that in the case of strict collinearity estimation of linear regression coefficients simply cannot be calculated. Bearing in mind the serious consequences that has multicollinearity in terms of quality estimation of a regression model, it is important, firstly, to be able to assess its presence in the experimental data, and secondly to have methods of dealing with the consequences of multicollinearity. The solution to these problems was considered in detail Jobs [1, 4, 5]. This article discusses only quasicollinear impact on the stability of the various methods of selecting the optimal MLR.

By choice of the optimal model we mean the following. Suppose that there are **m** variables that are possibly input actions. In reality, only **k** of them are really significant effect on the behavior of the output variable RLL (1), that is, are "true" input variables MLR. However, we don’t know “apriori”, which of **m** variables really affect the output variable. Therefore, using certain criteria need to select from all possible subset of input variables which best "explains" the behavior of the dependent variable linear regression through the mechanism (1).

To select the optimal in the sense described above used 5 MLR criteria described in detail in [2,3]:

- Method of all possible regressions with the corrected coefficient of determination as an optimality criterion (MR2);

- Method of all possible regressions with the corrected coefficient of determination as benchmarks, and assessment of the significance of MDR ratios (based on t-statistics) (MR2t);
- Method of all possible regressions using Mallouza statistics as optimality criterion (Mlz);
- Successive elimination method (BWE);
- A step by step method (SWP).

Formuating the problem. Assess the impact of the degree of linear relationship between the input variables on the stability of different algorithms for choosing the optimal model of linear regression (MLR) by simulation.

The main part. The effectiveness of these methods in the presence of qiuasi multicollinearity among input variables was investigated using simulation.

A computer simulated **m** random sequences, which are used as the possible values of the independent variables. Of these sequences **k** were chosen as variables involved in the construction of RLL (1). Then it was assumed that it is not known what and how much of **m** independent variables included in the model is true. Thus, the search was performed using models of various possible combinations of **m** of independent variables on the basis of the criteria listed above.

To determine the quality of each of the above methods of optimization carried out numerical experiments in which the search for the optimal model was made repeatedly by statistically independent data. Namely, the generated **N** (number of experiments) sets of **m** sequences each of which comprises at **n** (volume of sample) data. In each of the **N** experiments to determine the best model on the basis of the above five criteria listed. Since the true MDR known beforehand described method makes it possible to estimate the percentage of correct identification of each of the test methods. This estimate will be enough reliable, if **N** is sufficiently large. For the numerical experiments we have developed in Matlab environment, the appropriate software.

Using the described technique, was studied the influence of multicollinearity level of input (predictor) variables on the quality of the identification of the "right" or optimum model in the sense described above. Quasicollinearity between predictor variables $\mathbf{X}_i$ and $\mathbf{X}_j$ was modelled the following way:

$$\mathbf{X}_i = \mathbf{X}_j + 1/\mathbf{mu} * \xi, \quad (5)$$

where ξ is random value.

Parameter **mu** sets the level of deviation from strict collinearity. The value of this parameter is proportional to the degree of relationship between the input variables $\mathbf{X}_i$ and $\mathbf{X}_j$.

Of course, this is only one possible way to control the degree of multicollinearity, but it is enough to analyze its impact on the selection of the optimal model.

 ISSN 1562-9945

In order to bind the **mu** level to the classical criterion level of the relationship between the two sequences, was calculated sample correlation coefficient between $\mathbf{X_i}$ and $\mathbf{X_j}$:

$$\mathbf{r} = \frac{\sum_{\mathbf{k}=1}^{\mathbf{N}} [(\mathbf{X_i(k)} - \overline{\mathbf{X}}_\mathbf{j}) * (\mathbf{X_j(k)} - \overline{\mathbf{X}}_\mathbf{j})]}{\sqrt{\sum_{\mathbf{k}=1}^{\mathbf{N}} (\mathbf{X_i(k)} - \overline{\mathbf{X}}_\mathbf{i})^2 \sum_{\mathbf{k}=1}^{\mathbf{N}} (\mathbf{X_j(k)} - \overline{\mathbf{X}}_\mathbf{j})^2}}$$

(6)

The following Table 1 shows the results of numerical experiments on the effect of multicollinearity on the level of the percentage of correctly identified MLR using each of the five methods of selection of the optimal model listed earlier.

The experiments were conducted with the following values of simulation parameters:

- The number of possible input variables **m** = 10;
- The number of input variables used to construct the MLR **k** = 5;
- The volume of the sample **n** = 50;
- Number of experiments **N** = 500.

Table 1

Optimality criterion	The level of multicollinearity (**mu**)	Correlation coefficient (**r**)	% correct identifications	% ident. mod. with a lack significant variables	Calculation time (sec.)
MR2	10	0,99999	7	30,8 (13,8)	88,90
	1	0,99995	14,40	0,4 (0)	88,50
	0,1	0,995	15,5	0	87,40
	0,01	0,699	15	0	87,20
	0	0	14,4	0	87,30
MR2t	10	0,99999	23,8	29,4 (8,8)	88,60
	1	0,99995	46,4	0,4 (0)	91,56
	0,1	0,995	47,2	0	90,33
	0,01	0,699	47,2	0	90,25
	0	0	44,6	0	90,08
Mlz	10	0,99999	4,6	95,4(21,2)	39,40
	1	0,99995	53,4	46,6(0,2)	38,90
	0,1	0,995	100	0	38,70
	0,01	0,699	100	0	38,70
	0	0	100	0	39,00
BWE	10	0,99999	65,8	28,6(8)	1,38
	1	0,99995	92,2	0	1,32
	0,1	0,995	92,2	0	1,29
	0,01	0,699	92,4	0	1,29
	0	0	92,0	0	1,26
SWP	10	0,99999	87,8	2,4(1)	8,74
	1	0,99995	83,2	7,6(3,8)	8,50
	0,1	0,995	45,4	52(20)	8,42
	0,01	0,699	72,0	19,2 (0,8)	9,41

	0	0	82,6	8,6 (0,2)	8,34

The second column of the table shows the value of **mu** parameter defining the level of multicollinearity (Formula 5), and in the third column - corresponding to this **mu** selective correlation coefficient (Formula 6). Under proper identification is implied choice MLR, which as predictor variables included in full only those input variables that were used in the modeling regression dependence (1) (t. E. The "true" input variables). [1,2]

Errors in identifying MDR can be of two types [4]:

- Not to include some of the "real" input variables;
- Include some "extra" input variables.

It is known that the non-inclusion of the "true" input variables much more significant error than inclusion in the identification model "extra" variables [2,4]. Therefore, in the fourth column of Table 1 shows the number of the identified models are not fully set "true" input variables. However, in the presence of a strong multicollinearity between the two input variables, one of which bears almost the full information on the other. Therefore, not only in the identification model including these two variables results in a serious error identification. The number of such cases is shown in the fourth column of Table 1 in parentheses.

Analysis of the data in Table 1 indicates that the step by step method (the SWP) selection of the optimal MLR gives the highest percentage of correct identifications during a very strong linear relationship between the two input variables (ie. E., When the correlation between them is practically equal to unity). At the same time, this method is not dependent on the level and Multicollinearity identification errors often associated with non-inclusion of the "true" input variables. If the level of correlation is less than 0.995 one hundred percent correct identification provides a method of all possible regressions using Mallouza statistics as optimality criterion (Mlz). The disadvantage of this method of identification is that the computation time increases rapidly with increasing numbers of possible input variables.

Significantly faster is the method of successive elimination (BWE). This method provides a high percentage of correct identifications of MLR (over 90 percent), where the correlation between input variables is less than 0,99995 and there do not appear the errors associated with the non-inclusion of the "true" input variables.

Conclusions. Significant influence of quasicollinearity manifests itself only at a very high degree of correlation between input variables (more than 0,999). In the case where among the independent variables there is a strong linear relationship (quasicollinear), in order to select an optimal method MLR it is better to use of all the possible regression statistics Mallouza as the optimality criterion, if the number of possible independent variables is not too great. In practice, the authentication method can be used if the number of possible input variable does not exceed 15. Otherwise it is preferable to use the method of successive elimination.

LITERATURE

1. Дрейпер Н., Смит Г. – Прикладной регрессионный анализ: В 2-х кн. 2/Пер. с англ. – 3-е изд., перераб. и доп. – М.: Диалектика. 2016. – 912с.

2. Крохін В. В. Алгоритми ідентифікації багатопараметричних регресійних моделей / / Навчальний посібник до вивчення курсу «Цифрова обробка сигналів». – Дніпропетровськ.: РВВ ДНУ, 2012. – 32 с.

3. Крохин В. В., Кузьменко Н. О. Автоматизация выбора оптимальной модели линейной регрессии // Системні технології. Регіональній міжвузівський збірник наукових праць. - - Випуск 1 (78). Днепропетровск, 2012 - с. 73 – 83.

4. Gujarati D. N. Econometría. Tercera edición. McGraw-Hill - Santafe de Bogotá, 1997. - 824 p.

5. Rawlings J.O., Pantula S.G., Dickey D.A. Applied Regression Analysis. A Research Tool. Second edition - New York: Springer, 2001. - 671p.

ESSAY

UDC 519.21: 519.24

V. V. Krokhin. **Stability analysis method of choosing the optimal model of multiple linear regression when the independent variables are quasicollinear.** // System technologies. Regional Interuniversity collection of scientific papers. - Issue 1 (..) - Dnepropetrovsk, 2017 - ... with.

Simulation method is used to study the influence of quasicollinearity between input variables on the effectiveness of different methods to select the optimal model of multiple linear regression (MLR). The study was conducted using an original program developed in Matlab environment. Analyzed five different methods of selecting the optimal MLR.

Based on these results it is concluded that significant influence of quasicollinearity on choosing optimal MLR is manifested only at a very high degree of correlation between input variables (more than 0,999). If there is a strong linear relationship between the input variables, it is recommended to select the MLR method of all possible regressions using Mallouza statistics as the optimality criterion, where the number of possible independent variables is less than 15. Otherwise, preference should be given to the method of successive elimination.

Bible. 5, table 1.