UDC 519.21:519.24

V.V. Krokhin

INVESTIGATION OF THE STABILITY OF METHODS OF SELECTING THE OPTIMAL MODEL OF MULTIPLE LINEAR REGRESSION IN THE CASE WHEN INDEPENDENT VARIABLES ARE OBSERVED WITH ERRORS

Annotation. The situation is considered when the input variables of the regression model contain errors. For such a case, the stability of the various criteria commonly used to select the regression model optimal from the point of view of the set of input variables is carried out. The study was carried out by simulation using a software specially developed in the MATLAB environment.
It is shown that the most stable to the presence of errors in input variables is an algorithm based on the method of sequential elimination.
Keywords: multiple linear regression, optimal model, optimality criterion, errors in input variables, simulation modeling.

Introduction. In the classical linear regression model, it is assumed that the input variables are observed without errors. It is clear, that in most cases this assumption is not fulfilled exactly, since the values of the input variables contain, at least, measurement errors. Therefore, it is of considerable interest to study the influence of errors in independent variables on various regression analysis procedures. From the practical point of view, such analysis is important, first of all, when using samples of limited volume. An effective means of research in this situation is simulation modeling.

Analysis of publications on the research topic. Effects associated with the presence of errors in the input variables of the regression model have been repeatedly studied. It is shown that the estimates of the coefficients of the multiple linear regression model obtained by the least squares method (MLS) become biased and unconsistent. (see, for example, [1,2,3]). You can obtain unbiased estimates if you have additional information about the errors inherent in the input variables, for example, if their variances are known [1]. In the work [4] the value of the root-mean-square error (RMS) of the MLS estimations (MLSEs) was

 ISSN 1562-9945

investigated by simulation in the presence of errors in the input variables with a limited sample size. It has been demonstrated that MLSEs have smaller RMSs compared to unbiased estimates obtained by the maximum likelihood method (MLM) at small sample sizes (up to 50), since they have significantly less selective variability.

One of the important practical problems of regression analysis is the selection of those input variables that significantly affect the output variable (system response). Such selection makes it possible to reduce the time and material costs of conducting experiments using the multiple regression model. Screening of input variables that do not significantly affect the response of the system, in this paper, we will call the optimal search for a set of input effects of the multiple linear regression model.

Earlier, we considered the problem of choosing the optimal multiple input regression in a set of input parameters in the case when the hypotheses of the classical Gauss-Markov model are satisfied. [3,5] The study was carried out using simulation. Five different algorithms for solving the problem of choosing the optimal regression model were analyzed. [5,6] Then, in work [7] the stability of these five algorithms was studied in the case when multicollinearity is present between some of the input variables.

Formulation of the problem. In this paper, we study the effectiveness of the same algorithms for choosing the optimal model of multiple linear regression, when the input variables are observed with errors.

Main part. Consider the multiple linear regression (MLR) model with errors in the input variables.

$$\mathbf{Y} = \xi\beta + \mathbf{U}, \qquad (1)$$

where $\mathbf{Y} - [n \times 1]$ vector of values of the dependent variable, $\xi - [\mathbf{n} \times \mathbf{k}]$ matrix of true values of independent variables, which, however, are inaccessible for observation, $\beta - [(k+1) \times 1$ vector of parameters to be evaluated and a perturbation vector (which is often called model errors). Instead of true values of independent variables ξ, variables available for observation are

$$\mathbf{X} = \xi + \varepsilon, \qquad (2)$$

where ε represents a matrix of errors in independent variables.

For this model the usual MLSEs of parameters β

$$\tilde{\beta} = (X'X)^{-1}X' \quad (3)$$

lose their optimal properties. They are even asymptotically biased and are unconsistent [1,3]. At the same time, they probably remain the best among all estimates, which are calculated only using the observed values of the input variables **X** and the output variable **Y**. Therefore, it is of interest to investigate the influence of errors in input variables on the choice of the optimal MLR in the sense indicated above. This study was carried out by simulation using a specially developed software in the MATLAB environment.

Simulation modeling was carried out as follows. It is assumed that **m** inputs are possible, of which only **k** inputs actually affect the response of the system. Therefore, an array of random numbers **Xfull [nxm]** was generated, the contents of which were considered as **m** possible inputs (here **n** is the sample size). From this array, the subarray was chosen as **k** input influences, that really affects the response of the Y system. But since we consider that input variables are observed with errors, then elements of array **Xk** are superimposed with perturbations simulating Ex errors. These perturbations were set as follows

$$Ex = Xk \times diag(mean(Xk)) \times sigmaX \quad (4)$$

The parameter allows you to control the error level, and the coefficient, where mean denotes the calculation of the mean value, and $diag$ uses the diagonal element of the matrix, allows to obtain perturbations proportional to the sample mean of each of the input parameters. Then the array was calculated

$$X = Xk + Ex\ , \quad (5)$$

which was used to form the output of the system **Y**, caused by input effects **X**

$$Yn = X \times \beta. \quad (6)$$

The exact values of the coefficients of the model β were also set as initial parameters. However, the net response of the system to the input effects in the classical linear regression model is unobservable, therefore, as the observed system response,

$$Y = Yn + U. \quad (7)$$

Errors of the model U were set so

$$U = mean(Yn) \times sigmaU \times randn(size(Y)). \quad (8)$$

 ISSN 1562-9945

Thus, it is possible to control the level of model errors **U** with the ***stgmaU*** parameter and set this level in proportion to the average level of the system's exact response to the input effects due to the ***mean*(Yn)** multiplier.

The initial data for searching optimal for the set of input parameters of the multiple linear regression model were **Xfull** and **Y** arrays. As well as in the works **[5,6]** to select the optimal MLR, 5 criteria were applied:

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

- the method of all possible regressions with the corrected coefficient of determination as an optimality criterion and an estimate of the significance of the MLR coefficients (based on t-statistics) (MR2t);

- a method of all possible regressions using the Mallows statistics as an optimality criterion (Mlz);

- method of sequential elimination of input variables (BWE);

- a method of step-by-step inclusion of input variables (SWP).

The algorithms of each of these criteria are presented in the works mentioned above **[5,6]**, therefore here are not given.

MLSEs **(3)**, calculated from the observed sampling values, are random variables. Therefore, any results, obtained on the basis of the MLSEs, also include an element of randomness. However, when analyzing the results of a set of numerical experiments carried out with unchanged initial data, it is possible to establish the regularities characterizing the regression analysis procedure under study.

In particular, since the true model of multiple linear regression is known in advance in simulation simulation performed in accordance with the above method, it is possible to estimate the percentage of identifications of the "correct" model using each of the above optimality criteria.

Numerical experiments were carried out at such values of the initial parameters:

- number of possible input variables **m** = 10;

- the number of input variables that affect the output response of the **system k** = 5;

- number of experiments **num** = 500.

Variable values were:

- the level of errors in the input variables ***sigmaX*** ;

- sample size **n**.

The input effects and errors were generated using the randn function of MATLAB, which produces pseudo-random sequences with a distribution law close to normal. In this case, it is possible for each new experiment to return the sensor of random numbers to the initial state. Therefore, identifications using different optimality criteria were carried out for the same sequences of sets of input variables, which is very important in a comparative analysis of simulation data.

The results obtained are presented in the following table 1.

This table shows the percentage of identification of the "correct" regression models when using 5 various criteria of optimality mentioned above. "Correct" is a model in which only input variables that really affect the output response of the system ("true" input variables) were included. Bold type identifies cases when the number of "correct" identifications exceeds 50%.

Incorrect identifications are generated by errors of two kinds:

- include "extra" input variables (identification error of the first kind);

- Some of the "true" input variables (identification error of the second kind) are not included.

From the point of view of further application of the obtained regression model, identification errors of the second kind lead to more severe consequences than errors of the first kind. In particular, estimating the regression coefficients for the "true" variables included in the model turns out to be biased. [5] Therefore, in the table 1, along with the percentage of "correct" identifications, the percentage of identified models is shown in which some (or even all) of the "true" input variables were not included (this value is indicated in the table as "% with a lack of "true" variab.").

The first column of the table 1 contains the results of identifications in the absence of any errors in the "true" input variables (***sigmaX*** =0).

 ISSN 1562-9945

Table 1

Optimization criteria	Volume of sample **n**	Error level in independent variables							
		***sigmaX*=0**		***sigmaX*=0,1**		***sigmaX*=0,2**		***sigmaX*=0,5**	
		% of correct identifications	% with a lack of "true" variab.	% of correct identifications	% with a lack of "true" variab.	% of correct identifications	% with a lack of "true" variab.	% of correct identifications	% with a lack of "true" variab.
MR2	20	11,8	0,6	7,4	1,4	10,6	6,5	9,0	30,0
	30	9,8	0	13,8	0,6	12,8	1,4	12,6	13,2
	50	16,2	0	15,4	0	13,4	0,2	13,6	3,4
MR2t	20	31,6	2,4	29,8	6,0	26,4	20,2	13,0	50,0
	30	37,6	0,4	40,6	1,4	40,0	4,8	23,8	61,8
	50	44,6	0	42,6	0	44,2	0,8	36,4	18,6
Mlz	20	**86,4**	12,5	**69,0**	30,0	40,2	58,6	9,0	81,0
	30	**90,4**	9,6	**78,6**	21,2	48,8	51,2	7,6	92,4
	50	**92,6**	7,4	**82,5**	17,5	**57,4**	42,6	12,0	88,0
BWE	20	**80,0**	4,0	**79,6**	7,0	**67,8**	22,4	31,0	64,0
	30	**91,8**	0	**86,6**	1,4	**82,8**	5,4	**52,4**	39,6
	50	**93,6**	0	**93,0**	0	**92,4**	0,6	**76,6**	17,4
SWP	20	31,2	65,6	29,2	67,4	25,6	72,4	8,8	91,0
	30	48,0	48,3	46,8	49	43,2	53,2	24,4	72,2
	50	**71,4**	25,6	**69,4**	25	**67,8**	36,0	**53,6**	40,4

Analysis of the results. The data given show that errors in the input effects of the level, which is of the order of 10% from the level of the input itself, lead to a slight decrease in the percentage of "correct" identifications (sigmaX = 0,1). The best stability in this case is the identification method, which uses the algorithm to sequentially exclude input variables (BWE). [2,5]. This algorithm allows you to get more 50% correct identifications, even when the level of errors in the input variables is about half of the level of the values of the input variable itself (sigmaX = 0,5), with a sample size of at least 30. It is also important that this algorithm, less often than the others analyzed, identifies models with a lack of "true" input variables. In the work [6] also shown that it is also the most rapid. The algorithm, based on the search of all possible regression models, using Mallows statistics as the criterion of optimality (Mlz) [5], also gives a high percentage of "correct" identifications in the absence of errors in the input variables. However, this percentage decreases faster than the BWE algorithm as the level of errors in the input variables grows. In addition, almost all "incorrect" identifications contain errors of the second kind, that is, they do not contain some of the "true" input variables. The remaining identification algorithms considered, with sample sizes up to 50 elements, give a low percentage of "correct" identifications, even if there are no errors in the input variables.

Conclusions. Errors in input variables, the level of which is of the order of 10%, do not lead to a significant reduction in the percentage of "correct" identification of multiple linear regression models. Consequently, errors in measurements, the level of which is usually lower 10%, will not have a significant influence on the choice of the optimal linear regression model for a set of input parameters. Among the considered algorithms of identification, the algorithm of successive exclusion of possible input variables (BWE) has the best stability. It allows you to get a high percentage of correct identifications at a level of errors up to 50% from the level of the values of the input variable itself, even for small sample sizes (from 30 before 50 values).

LITERATURE

1. Fuller W.A. " Measurement Error Models", New York, John Wiley & Sons, 1987. – 442 p.

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

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

4. Крохин В.В., Цыганков К. Е. Исследование эффективности оценивания коэффициентов множественной линейной регрессии при наличии ошибок в независимых переменных. // Системні технології. Регіональній міжвузівський збірник наукових праць. - Вип. 5 (64). Днепропетровск, 2009. - с. 44 – 56.

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

6. 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.

7. Крохин В.В. Исследование устойчивости методов выбора оптимальной модели множественной линейной регрессии в случае, когда независимые переменные являются квазиколлинеарными. // Системні технології. Регіональній міжвузівський збірник наукових праць. - Випуск 1 (108), Дніпропетровськ, 2017, с. 46-54.