ADAPTIVE REGRESSION ANALYSIS: THEORY AND APPLICATIONS IN ECONOMETRICS

G. Perez, * Y.V. Chebrakov and V.V. Shmagin
Universidad de Almeria
* St-Petersburg Technical University,
Apt.22, Bldg.11/1, Polustrovsky pr., 195221, S.-Petersburg, Russia
e-mail: gchebra@mail.ru

ABSTRACT

Some new methods adaptable to regression analysing experimental dependences of the heterogeneous systems are described. These methods efficiency for analysing econometric heterogeneous systems is demonstrated.

KEY WORDS

adaptive regression analysis, heterogeneous econometric systems.

1. INTRODUCTION

It is well-known, many computational problems, arising in the investigation of the multivariate statistical systems (for instance, econometric systems), can hardly be solved without using some kind of fitting techniques. Such problems refer to the construction of mathematical models, description of the investigated systems behaviour, and/or finding the system parameters values by solving approximation type of problems [4]. It has been shown in [1– 3], that the application of available fitting techniques leads to some theoretical and computational difficulties (paradoxes). They are mainly explained by the fact that the modern statements of data analysis problems are too far from the real experimental situations. In particular, these statements make no provision for facts that:

1. Results of calculations may depend on the way the investigated data are obtained;
2. The given accuracy of estimated values cannot be achieved for the strongly contaminated data;
3. A single solution of the data analysis problems for contaminated data arrays is incomplete one;
4. Every investigated system may contain some heterogeneity, which leads to (i) an adequate, (ii) a removable (local) inadequate or (iii) an irremovable (global) inadequate postulated fitting model, describing the behaviour of the system.

The main goals of this paper are:

a) Discussing some theoretical and computational difficulties of regression analysing experimental dependences, describing the behaviour of the heterogeneous econometric systems;
b) Offering some new methods adaptable to regression analysing experimental dependences of the heterogeneous econometric systems.
c) Demonstrating the new methods efficiency for analysing econometric heterogeneous systems.

2. COMPUTATIONAL DIFFICULTIES OF REGRESSION ANALYSING HETEROGENEOUS DEPENDENCES

As generally known [1, 2, 4, 5], for found experimental dependence $\{y_n, \mathbf{X}_n\}$ ($n = 1, 2, \ldots, N$) and given fitting function $F(\mathbf{A}, \mathbf{X})$, the main problems of regression analysis theory are finding estimates of $\mathbf{A}'$ and y', where $\mathbf{A}'$ is an estimate of vector parameter $\mathbf{A}$ of the function $F(\mathbf{A}, \mathbf{X})$ and $\{y_n'\} = \{F(\mathbf{A}', \mathbf{X}_n)\}$. In particular, if $F(\mathbf{A}, \mathbf{X}) = \sum_{l=1}^{L} a_l h_l(\mathbf{X})$ $\{F(\mathbf{A}, \mathbf{X})$ is a linear model$\}$, where $h_l(\mathbf{X})$ are some functions on $\mathbf{X}$, then least squares (LS) method gives classical regression analysis solution:

$$\mathbf{A}'_{\mathrm{LS}} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\,\mathbf{H}^{\mathrm{T}}\mathbf{Y}, \qquad (1)$$

where $\mathbf{H}$ is a matrix $L \times N$ in size with n-th row $(h_1(\mathbf{X}_n), h_2(\mathbf{X}_n), \ldots, h_L(\mathbf{X}_n))$; $\mathbf{H}^{\mathrm{T}}$ is the transposed matrix $\mathbf{H}$; $\mathbf{Y} = \{y_n\}$.

We found [1, 2] that three various types of heterogeneity affection may be revealed when analysing experimental dependences:

a) The heterogeneity has no effect on the results of solving fitting problems.

b) The heterogeneity leads to a removable (local) of fitting function $F(\mathbf{A}, \mathbf{X})$.

c) The heterogeneity leads to a irremovable (global) of fitting function $F(\mathbf{A}, \mathbf{X})$.

In this section some computational difficulties arising in the cases (b) and (c) are discussed.

A. Pseudo-multicollinearity. Let experimental dependence $\{y_n, \mathbf{X}_n\}$ ($n = 1, 2, \ldots, N$) and the linear fitting

function $F(\mathbf{A}, \mathbf{X}) = \sum_{l=1}^{L} a_l h_l(\mathbf{X})$ be given. Assume that rank $\mathbf{H} < L$, or, in other words, there is a linear dependence between columns of matrix $\mathbf{H}$:

$$c_1 h_1 + c_2 h_2 + \ldots + c_L h_L = 0, \qquad (2)$$

where at least one coefficient $c_l \neq 0$. In this case matrix $(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}$ does not exist and, consequently, the value of $\mathbf{A}'$ cannot be found from the equation (1). Such situation is known as strict multicollinearity [4].

It should be noted that the values of the independent variable $\mathbf{X}$ are always determined with a certain round-off error, although this error may be very small. Thus, if even strict multicollinearity is present, the equation (2) is satisfied only approximately and therefore rank $\mathbf{H} = L$. In such situation the application of the equation (1) for calculation of the estimate of vector parameter $\mathbf{A}$ gives $\mathbf{A}'_{\mathrm{LS}}$ values drastically deviating from true coefficients values. To correct the discussed situation, in regression on characteristics roots [6] it is suggested to exclude from the linear model $\sum_{l=1}^{L} a_l h_l(\mathbf{X})$ such l-components, whose eigennumbers λ_l and elements V_{0l} are small (the vector $\{V_{0l}\}_{l=1,2,\ldots,L}$ is a measure of a l-th vector contribution to rise of the accuracy of regression model predictions). Following values are recommended to use as critical ones: $\lambda_{\mathrm{cr}} = 0.05$ and $V_{\mathrm{cr}} = 0.1$.

Below we demonstrate that in some cases the difference between $\mathbf{A}'_{\mathrm{LS}}$ (calculated by the formula (1)) and $\mathbf{A}'_{\mathrm{CHR}}$ (calculated in regression on characteristics roots) may be explained not only by the effect of multicollinearity but also by the outliers presence in the data array $\{y_n, \mathbf{X}_n\}$.

Indeed, assume that the data array is following:

$$\{y_n, \mathbf{X}_n\} = \{1 + 0.5n + 0.05n^2 + 0.005n^3;\ n, n^2, n^3\}, \quad (3)$$

$n = 1, 2, \ldots, 11$. Introduce two outliers into (3) by means of increasing values y_3 and y_8 on 0.5. The following results are obtained for the fitting function $F(\mathbf{A}, \mathbf{X}) = F(\mathbf{A}, x) = \sum_{l=0}^{3} a_l x^l$ and the heterogeneous experimental dependence under the consideration:

$$\mathbf{A}'_{\mathrm{LS}} = (1.017;\ 0.542;\ 0.0462;\ 0.0050), \qquad (4)$$
$$\mathbf{A}'_{\mathrm{CHR}} = (1.046;\ 0.516;\ 0.0516;\ 0.0047).$$

B. Nonequivalence of computational procedures. It may be concluded from (3) and (4) that the presence of outliers in some analysing data arrays can lead to a distortion of estimation problems solution results. To obtain better solution it is sufficient sometimes to clear the initial heterogeneous data arrays from all outliers. For revealing outliers it is suggested in [7] to use one of two equivalent combined statistical procedures, in which parameter estimates, minimising the median (MED) of the array $\{(y_n - y_n')^2\}$ or the sum (SUM) of K first elements of the same array, are considered as the best ones. Let us demonstrate that in some cases revealing outliers problems solutions may depend on a type of the combined statistical procedures.

Indeed, assume that the data array is following

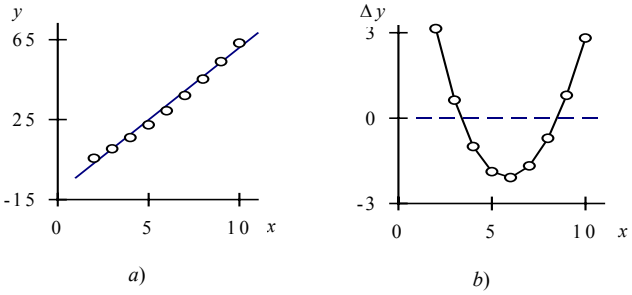
$$\{y_n\} = (2.48;\ 0.73;\ -0.04;\ -1.44;\ -1.32;\ 0), \qquad (5)$$
$$\{x_n\} = (-4;\ -3;\ -2;\ -1;\ 0;\ 10),$$

where, on simulation conditions [8], the record with number 6 is the outlier (such reading that contrasts sharply from others); the fitting function $F(\mathbf{A}, \mathbf{X}) = a_0 + a_1 x$ and $\mathbf{A}_{\mathrm{true}} = (-2;\ -1)$. The following results are obtained for the data array (5): for the first combined procedure MED the element y_6 is the outlier, but for the second combined procedure SUM none of records of the data array (5) is outlier.

C. LS-estimations of inadequate fitting functions parameters. Let the data array $\{y_n, x_n\}$ be following:

$$\{y_n\} = (5.66;\ 10.39;\ 16.00;\ 22.36;\ 29.39;\ 37.04; \qquad (6)$$
$$45.25;\ 54.00;\ 63.25),\ \{x_n\} = (2;\ 3;\ \ldots;\ 10).$$

As it may be calculated by (1), if the fitting function $F(\mathbf{A}, \mathbf{X}) = a_0 + a_1 x$, for the data array (6) $\mathbf{A}'_{\mathrm{LS}} = (-11.95;\ 7.24)$. Plots of the dependence (6) (circles) and the fitting function $F(\mathbf{A}, \mathbf{X}) = -11.95 + 7.24\,x$ (continuous line) are shown in Figure 1(*a*); the plot of residues $\{y_n + 11.95 - 7.24x_n;\ x_n\}$ is shown in Figure 1(*b*).



a) *b*)

Figure 1. (*a*) — Plots of the dependence (6) (circles) and the function $-11.95 + 7.24\,x$ (continuous line);
(*b*) — The plot of residues $\{y_n + 11.95 - 7.24x_n;\ x_n\}$

There are three criteria that the fitting function $a_0 + a_1 x$ is inadequate for the data array (6):

a) the plot of residues, shown in Figure 1(*b*), has the form of a parabola, and, consequently, for the dependence (6) a second degree multinomial must have the better fitting properties then the linear model;

b) the data array (6) was generated by the function $f(x) = 2x^{1,5}$:

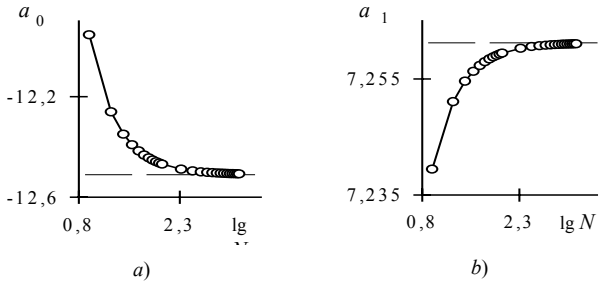
$$\{y_n\} = \{[k\,f(x_n)] / k\}, \qquad (7)$$

where square brackets mean the integer part, a factor $k = 100$ and its presence in (7) is necessary for calculating all values of y_n within error $\varepsilon = 0.01$;

c) the maximal value of Δy in Figure 1(*b*) exceeds appreciably the value of the computational error ε for the values of y ($\varepsilon = 0.01$).

Dependences of LS-estimations of linear model parameters on data array size N of the function $2x^{1,5}$,

given uniformly on the interval [2; 10], are shown in Figure 2. Limiting values of parameters a_1 and a_0 (when $N \to \infty$) are marked by the dotted lines.



a) *b*)

Figure 2. Dependences of LS-estimations of linear model parameters on data array size N of the function $2x^{1,5}$, given uniformly on the interval [2; 10].

Analysis of dependences shown in Figure 2 allows to conclude

a) inadequacy of the fitting function a_0+a_1x leads to the distortion of LS-estimations of its parameters;

b) the distortion value of LS-estimations may be decreased by increasing the records amount N of the function $2x^{1,5}$, given uniformly on the interval [2; 10].

3. THEORETICAL ANALYSIS OF COMPUTATIONAL DIFFICULTIES EMERGENCE

For a homogeneous system, let a connection between the characteristics y and $\mathbf{X}$ be described by a functional: $y = F(\mathbf{A}, \mathbf{X})$. As discussed in the Section 2, when some heterogeneity is observed in the system, three various situations can be met in the discussed experiment.

1. The heterogeneity has no effect on the experimental dependence $\{y_n, \mathbf{X}_n\}$. In this case, it is impossible to distinguish the homogeneous system from heterogeneous ones based on the dependence $\{y_n, \mathbf{X}_n\}$;

2. The heterogeneity leads to a distortion of the dependence $\{y_n, \mathbf{X}_n\}$ *in some small region* $\{\mathbf{X}_{n*}\} \subset \{\mathbf{X}_n\}$. It is said that the fitting model $F(\mathbf{A}, \mathbf{X})$ has removable (local) inadequacy. In this case, for extracting the effects connected with the presence of a heterogeneity in the investigated system, one may use the following strategy [1, 2]:

A. Solve the problem to reveal the outliers $\{y_{n*}, \mathbf{X}_{n*}\}$;

B. Determine the value $\mathbf{A}'$ on the data array $\{y_n, \mathbf{X}_n\} \setminus \{y_{n*}, \mathbf{X}_{n*}\}$, which is well-fitted by the model $F(\mathbf{A}, \mathbf{X})$;

C. Detect the type and degree of the effects connected with the presence of a heterogeneity in the investigated objects on the data array $\{y_{n*} - F(\mathbf{A}', \mathbf{X}_{n*}), \mathbf{X}_{n*}\}$.

It is evident that solving problem (A) some difficulties, which are unsurmountable in the framework of modern regression analysis theory, will arise.

3. The heterogeneity leads to a distortion of the dependence $\{y_n, \mathbf{X}_n\}$ *in a big region* $\{\mathbf{X}_{n*}\} \subseteq \{\mathbf{X}_n\}$: {the fitting model $F(\mathbf{A}, \mathbf{X})$ has irremovable (global) inadequacy}.

In this case it is impossible to find a reliable solution of the problem in the framework of modern regression analysis theory (see Section 6).

In summary we may conclude that since the methods under the consideration are not effective for bypassing the computational difficulties (paradoxes) or even absent from modern regression analysis theory, it is not easy to obtain reliable solutions for the problems of quantitative processing of experimental dependences describing the heterogeneous systems behaviour.

4. POSSIBLE WAY FOR BYPASSING COMPUTATIONAL DIFFICULTIES

The authors opinion is that modern regression analysis theory incapable to cope worthily with the computational difficulties described in previous Sections without engaging a set of concepts of experiment design theory [9] and approximation theory [10]. For instance, such concepts are:

a) measurement error of dependent variable values;

b) best even-approximative function;

c) problem on estimating parameters of the inadequate regression model;

d) active and passive regression experiments.

One of possible ways for bypassing the computational paradoxes in the regression analysis theory, is the further development of this theory by means of a set of methods, concepts and results of approximation theory and experiment design theory [1, 2].

From Section 3, it can be seen that for heterogeneous systems it is possible to obtain reliable solutions for the estimation problems when a more advanced approach is considered in order to include:

a) Objective criteria, allowing on the found dependence $\{y_n, \mathbf{X}_n\}$ to determine whether the heterogeneity has effect on the adequacy of the fitting model and how many outliers are present in the data array $\{y_n, \mathbf{X}_n\}$;

b) Method of regression analysing of heterogeneous experimental dependences for which the fitting model $F(\mathbf{A}, \mathbf{X})$ is inadequate.

5. NEW METHODS ADAPTABLE TO ANALYSING HETEROGENEOUS DEPENDENCES

As mentioned in Sections 2 and 3, three various situations can be met in the regression analysing of experimental dependences describing the heterogeneous systems

behaviour. The postulated fitting model $F(\mathbf{A}, \mathbf{X})$ may be (a) adequate, (b) removable (local) inadequate and (c) irremovable (global) inadequate. The situation (b) has the most wide occurrence in practice. From the point of view of the modern data analysis theory the situation (b) means that the data array contains some number of outliers and consequently for finding the correct value of vector parameter **A** it is necessary to solve a problem of revealing the outliers in the data array. As it is proved in [1, 2], in the general case, one may obtain the correct solution of this problem by virtual device method (VD-method). The VD-method is the following iterative procedure:

Determine the minimum value $\alpha = \alpha_{\min}$ for a fixed value n_0 such that for all the experiment realizations of $\{\mathbf{X}_U\}$ containing U possible subsamples from N, $N-1$, …, $N-n_0$ records, the inequality

$$|y_n - g_\alpha(y_n')| \le \alpha_{\min} \qquad (8)$$

is fulfilled, where N is the dimension of the initial data array $\{y_n, \mathbf{X}_n\}$, y_n' is an estimate of n-th value of the dependent variable, n_0 is a given integer number, which assigns a maximum level of truncating the initial data array $\{y_n, \mathbf{X}_n\}$ in the discussed computation procedure, g_α is a function, allowing to describe the way of dependent variable measurement, α is the vector, characterizing the contamination level of the dependent variable in the data array $\{y_n, \mathbf{X}_n\}$.

Remark 1. If the form of measurement function g_α is unknown then one may assume that g_α is the truncation function:

$$g_\alpha(y) = 2\alpha[y/(2\alpha)] + 2\alpha \text{ if } |y - 2\alpha\,[y/(2\alpha)]| \ge \alpha, \quad (9)$$
$$\text{else } g_\alpha(y) = 2\alpha\,[y/(2\alpha)].$$

Remark 2. In practice one may determine the value of VD-method vector parameter $\alpha_{\min}$ as the solution of the following extremal problem

$$\min_{\alpha} \max_{U} \max_{n} |y_n - g_\alpha(y_n')|, \qquad (10)$$

where the maximum with respect to U means finding the solution for all U sets of $\{\mathbf{X}_U\}$, composed of N, $N-1$, … , $N-n_0$ readings;

Remark 3. The analysing data array contains n_0 outliers if

a) For all experiment realizations of $\{\mathbf{X}_U\}$, containing all possible U subsamples, formed by N, $N-1$, … , $N-n_0-1$ records, the inequality $\alpha_{\min} > \varepsilon$ is satisfied.

b) There exists such n_0-truncation of the initial data array $\{y_n, \mathbf{X}_n\}$ containing $N-n_0$ records such that the inequality $\alpha_{\min} \le \varepsilon$ is satisfied, where ε is a vector characterizing the experimental contamination level of the dependent variable in the data array $\{y_n, \mathbf{X}_n\}$;

Remark 4. To construct a multiple solution of the data analysis problems for the contaminated data arrays (see Section 1) one may use the following method of equivalent analytical formulae (EAF-method) [1, 2]:

Replace the vector parameters **A** of functions $g_\alpha(F(\mathbf{A}, \mathbf{X}))$ by $\mathbf{A} = \{P_{m_i}(\mathbf{C}_i, \xi)\}$ and then determine the minimum values of the degrees, the estimates of coefficients and the values of ξ in the multinomials $P_{m_i}(\mathbf{C}_i, \xi)$.

Example 1. As stated in [11, 12], for the data array $\{y_n, \mathbf{X}_n\}$ shown in the Table 1, the multivariate model $Y_3 = a_0 + a_1x_1 + a_2x_2$ demonstrates well some effects of multicollinearity and so fitting models $Y_1 = b_0 + b_1x_1$ and $Y_2 = d_0 + d_1x_2$ are more preferable.

Table 1. The econometric data $\{y_n, \mathbf{X}_n\}$ on investigating the import turnover y (billion dollars) on gross national product x_1 (billion dollars) and consumer price index x_2 in USA.

Years	y	x_1	x_2
1964	28.4	635.7	92.9
1965	32.0	688.1	94.5
1966	37.7	753.0	97.2
1967	40.6	796.3	100.0
1968	47.7	868.5	104.2
1969	52.9	935.5	109.8
1970	58.5	982.4	116.3
197l	64.0	1063.4	121.3
1972	75.9	1171.1	125.3
1973	94.4	1306.6	133.1
1974	131.9	1412.9	147.7
1975	126.9	1528.8	161.2
1976	155.4	1702.2	170.5
1977	185.8	1899.5	181.5
1978	217.5	2127.6	195.4
1979	260.9	2368.5	217.4

Table 2. Computation results.

a) The function $g_\alpha(Y_i)$ (all readings)

$i=1$		$i=2$		$i=3$	
n_0	$\alpha_{\min}$	n_0	$\alpha_{\min}$	n_0	$\alpha_{\min}$
0	37.7	0	40.6	0	47.7
1	45.6	1	55.6	1	56.6

b) The function $g_\alpha(Y_i)$ (the first ten readings)

$i=1$		$i=2$		$i=3$	
n_0	$\alpha_{\min}$	n_0	$\alpha_{\min}$	n_0	$\alpha_{\min}$
0	2.7	0	6.4	0	2.7
1	5.7	1	11.1	1	5.7

c) The function $g_\alpha(Y_i)$ (the last six readings)

$i=1$		$i=2$		$i=3$	
n_0	$\alpha_{\min}$	N_0	$\alpha_{\min}$	n_0	$\alpha_{\min}$
0	9.0	0	16.0	0	8.0
1	15.5	1	22.2	1	29.0

Let us find the best fitting model $g_\alpha(Y_i)$ ($i = 1, 2, 3$) for readings of Table 1. Our calculations give the results presented in points (a – c) of Table 2. Analysing these results we may conclude that

α) for case (a) of Table 2

i) the model $g_\alpha(Y_1)$ has the minimal value $\alpha_{\min}$ and so it is the best fitting model among all tested models $g_\alpha(Y_i)$;

ii) the econometric system under analysis is heterogeneous for all tested models since even the minimal value $\alpha_{\min}$ exceeds appreciably the (measurement) error of the dependent variable y ($\varepsilon_{\max} = \varepsilon = 1$);

β) for cases (b) and (c) of Table 2

i) in order to reduce fitting errors of models $g_\alpha(Y_i)$, the econometric data of Table 1 are to divide into two data arrays, contained respectively the first ten (A1) and the last six (A2) readings of Table 1;

ii) for arrays A1 and A2, the model $g_\alpha(Y_1)$ is the best fitting model and so the value of y may be calculated by the formula

$$y = g_{2.7}(-34.7 + 0.09553\, x_1) \qquad (11)$$

for readings of array A1 and by the formula

$$y = g_{9.0}(-81.8 + 0.14212\, x_1) \qquad (12)$$

for readings of array A2.

6. CONCLUSION

In this paper the emergence of computational difficulties of regression analysing heterogeneous experimental dependences is explained by inadequacy of the fitting function $F(\mathbf{A}, \mathbf{X})$. It is stated that, for bypassing the computational difficulties, the following criteria and methods to be constructed:

a) Objective criteria allowing on the given dependence $\{y_n, \mathbf{X}_n\}$ to determine whether the heterogeneity has effect on the adequacy of the fitting model.

b) Methods of regression analysing of heterogeneous experimental dependences for which the fitting model is inadequate.

In Section 5 the problem (a) is solved completely. Besides some new methods adaptable to regression analysing experimental dependences of the heterogeneous systems with the removable (local) inadequate fitting model are suggested. Thus to carry the theory of adaptive estimation of parameters to completion it is necessary to elaborate advanced estimation method for the case, when the fitting model is irremovable (global) inadequate. This problem is theme of authors further studies. Some preliminary results on solving this problem may be found in [1, 13].

REFERENCES

[1] Y. V. Chebrakov, *Parameters Estimation Theory in Measuring Experiments* (St.–Petersburg State Univ. Press, 1997).

[2] Y. V. Chebrakov & V. V. Shmagin, *Regression Data Analysis for Physicists and Chemists* (St.–Petersburg State Univ. Press, 1998).

[3] Perez G., Chebrakov Y.V. y Shmaguin V.V. *Analisis de datos: dificultades y metodos alternativos* (Universidad de Almeria, 1999).

[4] K. R. Draper & H. Smith, *Applied Regression Analysis* (Wiley & Sons, 1981).

[5] C. P. Rao, *Linear statistical inference and its applications* (Wiley & Sons, 1973).

[6] J. Webster, R. Gunst & R. Mason, Latent root regression analysis, *Technometrics*, *16*, 1974, 513-522

[7] P. J. Rousseeuw & A. M. Leroy, *Robust Regression and Outlier Detection* (Wiley & Sons, 1987).

[8] P. J. Huber, *Robust Statistics* (Wiley & Sons, 1981).

[9] S. M. Ermakov & A. A. Zhiglyavskyi, *Mathematical Theory of Optimal Designing experiments*; (Nauka, 1987).

[10] P. J. Laurent, *Approximation et Optimisation* (Hermann, 1972).

[11] D. Salvatore, *Statistics and Econometrics* (McGraw-Hill, 1982).

[12] P. K. Katyshev, Y. R. Magnus & A. A. Peresectskiy *Econometrics for beginners: Problems and solutions* (Delo, 2002).

[13] J. Sacks & D. Ylvisaker, Linear estimation for approximately linear models, *Annals. Statist*., 6(5), 1978, 1122–1137.