DOI: 10.20998/2079-0023.2025.02.02
UDC 004:519.24:681.3.06

A. A. PAVLOV, Doctor of Technical Sciences, Full Professor, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, Professor of Computer Science and Software Engineering Department; e-mail: alexanderpavlov1944@gmail.com; ORCID: https://orcid.org/0000-0002-6524-6410
A. V. KUSHCH, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Postgraduate Student of Computer Science and Software Engineering Department, Kyiv, Ukraine, e-mail: antonkusch1@gmail.com, ORCID: https://orcid.org/0009-0003-0994-2985

ALGORITHMS FOR CONSTRUCTING A REGRESSION LINEAR WITH RESPECT TO UNKNOWN COEFFICIENTS ON A LIMITED AMOUNT OF EXPERIMENTAL DATA

This publication continues the series of scientific works of the authors on the creation of algorithms for constructing multivariate regressions which are linear with respect to unknown coefficients by using linear programming models. To simplify the simulation modeling of their efficiency, we present the algorithms for the multivariate linear regression problem. The use of linear programming models requires minimizing the sum of the absolute differences used in the general procedure of the least squares method. The estimates of the unknown coefficients obtained as a result of solving the linear programming problem are linear with respect to the vector of the values of the regression model in the statistical experiment. It is known that, by virtue of the Markov theorem, the estimates of the unknown coefficients obtained by the general procedure of the least squares method are efficient in the class of linear unbiased estimates. Thus, it would seem that the transition from the least squares method to the least absolute deviations used in the least squares method is a priori unproductive. But this is not so. From the proof of the Markov theorem, it follows that the linear estimation matrix must be constant and independent of the values of the regression model in the statistical experiment. The estimates obtained by the least absolute deviations method do not meet this condition. Indeed, the estimation matrix is the optimal basis for solving the linear programming problem by the simplex method and depends on the values of the regression model in the statistical experiment. Such a formulation of the problem allows introducing, into the optimization model, linear constraints that use the results of statistical tests and implement additional properties of the searched multivariate regression. The first studies of these algorithms have shown their efficiency, this allowed the authors to set the task of creating such algorithms that can not only compete with the general algorithmic procedure of the least squares method, but also be efficient for the case of a limited volume of experimental data, when the ratio of the average absolute value of the realizations of a random factor in the experiment to the average absolute value of the true regression on it is a sufficiently large value. In this case, it is incorrect to raise the problem of finding estimates of unknown coefficients that practically do not differ from the true ones, but, as experiments and, in particular, the examples given in this paper have shown, it is possible to find sufficiently good estimates of the average values of the true regression in the experiments conducted, which can be used, for example, in diagnosing the early stages of the onset of an epidemic of various diseases or in other recognition tasks.

Keywords: multivariate regression, least squares method, least absolute deviations method, linear programming model, simplex method, optimal basis.

1. Introduction. The problem of constructing multivariate regressions on a small volume of experimental data with a significant value of the variance of a random factor is still of interest to researchers both in theoretical and practical aspects [1–10]. In most cases, practical results for such problems are obtained using heuristic methods, in particular, the classical method of group consideration of arguments and its numerical modifications. This publication continues the series of papers by the authors [11, 12] on the creation of efficient algorithms for constructing multivariate regressions linear with respect to unknown coefficients, which formally have the form

$$\bar{\boldsymbol{Y}}(\bar{\boldsymbol{x}}) = \sum\nolimits_{j=0}^{r} \theta_j \psi_j(\bar{\boldsymbol{x}}) + E, \quad (1)$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_r)^{\mathrm{T}}$ is a vector of unknown coefficients; $\psi_j(\bar{\boldsymbol{x}}), j = \overline{0, r},$ are known basis functions; random variable (RV) E is a random factor, the distribution of this RV in this work is considered normal with known parameters $ME = 0, DE = \sigma^2 < \infty$.

Remark 1. The distribution of the RV is not essential for the algorithms proposed in this paper. The fundamental feature of these algorithms is that they use as the main formal model a linear programming model, the functionality of which minimizes the sum of the absolute differences used in the least squares method (LSM). As in publications [11, 12], the algorithm for constructing a multivariate regression is presented for a partial case of the model (1), namely, for a multivariate linear regression. This allows one to conduct simulated statistical modeling of the efficiency of algorithms for a sufficiently wide class of linear multivariate regressions, which is impossible for models presented in the form of (1).

2. Formal statement of the problem. A multivariate linear regression has the following matrix form:

$$\bar{\boldsymbol{Y}}(\bar{\boldsymbol{x}}) = \boldsymbol{\theta}^{\mathrm{T}} \bar{\boldsymbol{x}} + E, \quad (1)$$

where E is a RV that has a normal distribution with zero mathematical expectation and known variance σ^2. $\boldsymbol{\theta}$ is a vector of unknown coefficients $\boldsymbol{\theta}^{\mathrm{T}} = (\theta_0, \theta_1, \ldots, \theta_r)$. The vector $\bar{\boldsymbol{x}}^{\mathrm{T}} = (\boldsymbol{1}, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_r)$, $\bar{\boldsymbol{x}}_i, i = \overline{1, r},$ are deterministic input variables of the regression model. Let's write the results of statistical tests on model (2) in the form $(\bar{\boldsymbol{x}}_i \to y_i, i = \overline{1, n})$, that is,

$$y_i = \theta_0 + \theta_1 x_{1i} + \ldots + \theta_r x_{ri} + \varepsilon_i, i = \overline{1, n},$$

where ε_i is the realization of the RV E. With a sufficiently large amount of experimental data, the algorithms presented below can successfully compete with the general procedure of the LSM for finding efficient estimates of the values of the coefficients of a multivariate regression linear with respect to the unknown coefficients. However, this paper shows that the proposed algorithms can be successfully used in the case of a limited amount of experimental data to estimate scalar characteristics from the values of the true regression on the values of the input variables of the experiment conducted, when the average absolute value of the random factor realizations and the average absolute value of the true regression on the values of the input variables in the experiment are of the same order.

3. An algorithm that uses a single linear programming model. The first algorithm finds the optimal solution to the following linear programming problem (LPP):

$$\min \sum_{i=1}^{n} z_i \qquad (3)$$

$$-z_i \le y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i \le z_i,\ z_i \ge 0,\ i = \overline{1,n} \qquad (4)$$

$$0 \le (1 - t_2(n)) \cdot k_1 \le \sum_{i=1}^{n} z_i \le (1 + t_1(n)) \cdot k_1 \qquad (5)$$

$$0 \ge -(1 + t_4(n)) \cdot |k_2| \le \sum_{i=1}^{n} \left(y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i\right) \le (1 + t_3(n)) \cdot |k_2|. \qquad (6)$$

The variables of the LPP (3)–(6) are $z_1, \dots, z_n$, $\theta_0, \theta_1, \dots, \theta_r$. Constants $t_i(n) > 0,\ i = \overline{1,4}$, are chosen experimentally.

$$k_1 = \tfrac{1}{n} \sum_{i=1}^{n} |\varepsilon_i|,\ k_2 = \tfrac{1}{n} \sum_{i=1}^{n} \varepsilon_i > 0,$$

where ε_i are artificially generated realizations of the RV and do not coincide with the realizations of the RV in a statistical experiment.

Remark 2. The heuristic for choosing the regions of constraints on the LPP (3)–(6) variables is a modification of the heuristic presented in [12] and consists in the fact that with an appropriate choice of constants $t_i(n),\ i = \overline{1,4}$, the unknown values of

$$\tfrac{1}{n} \sum_{i=1}^{n} \left|y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i\right|,\ \tfrac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i\right)$$

belong to constraints (5), (6), respectively. The components of the vector $\boldsymbol{\theta}$ are the unknown values of the multivariate linear regression. The union of regions (5), (6) is significantly smaller than the region given by constraints (4).

Remark 3. At a qualitative level, it is clear that as the number of tests n increases, the number of errors $t_i(n)$ should decrease.

Remark 4. If $n \le r + 1$, then the general procedure of the LSM or the least absolute deviations method gives a degenerate solution, that is, if $\hat{\boldsymbol{\theta}}$ is the estimate of the values of $\theta_i,\ i = \overline{0,r}$, then these equalities hold:

$$y_i - \hat{\boldsymbol{\theta}}^{\mathrm{T}} \bar{\mathbf{x}}_i = 0,\ i = \overline{1,n}.$$

In other words, the estimates $\hat{\varepsilon}_i,\ i = \overline{1,n}$, of the realizations of the RV that it took in the tests are identically equal to zero. For the case $n \le r + 1$, the vector of estimates $\hat{\boldsymbol{\theta}}$ as a solution to the LLP (3)–(6) is not a degenerate vector of estimates, but the answer to how useful this estimate is in this case can only be given by careful statistical studies.

4. An iterative algorithm for constructing a multivariate linear regression. To find the efficient domain of the algorithm, we introduce the following definition.

Definition 1. A vector $\hat{\boldsymbol{\theta}}$ of estimates of the coefficients of a vector $\boldsymbol{\theta}$ is called a *consistent estimate* if the realization of the criterion χ^2, which is built on the estimates of the realizations of E, $\hat{\varepsilon}_i = y_i - \hat{\boldsymbol{\theta}}^{\mathrm{T}} x_i,\ i = \overline{1,n}$, does not contradict the hypothesis of a normal distribution with parameters $0, \sigma^2$ (parameters of the normal distribution of the RV E).

Remark 5. The definition of a consistent estimate does not depend on the distribution of the RV E since the criterion efficiently tests the hypothesis for any known distribution of the RV E.

From the definition of a consistent estimate it follows that the value of the number of tests in a statistical experiment and the value of $\sigma^2 = DE$ must be such that the following condition is fulfilled: the realization of the criterion χ^2 statistically significantly belongs to the feasible region only in the case when the numbers $\hat{\varepsilon}_i$ do not differ significantly from the realizations $\varepsilon_i,\ i = \overline{1,n}$, of the RV E. This is the condition that guarantees that the iterative algorithm presented below, the construction heuristic of which is aimed at finding consistent estimates $\hat{\theta}_j,\ j = \overline{0,r}$, of the components of the vector $\boldsymbol{\theta}$, is the most efficient algorithm. But, as the second example given in section 5 shows, it can be used in the case when the number of tests n does not satisfy the above restrictions.

The iterative algorithm consists of sequentially solving the following LPPs:

$$\min \sum_{i=1}^{n} z_i \qquad (7)$$

$$-z_i \le y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i \le z_i,\ z_i \ge 0,\ i = \overline{1,n} \qquad (8)$$

$$0 \ge -(1 + t_4(n)) \cdot |k_2| \le \sum_{i=1}^{n} \left(y_i - \boldsymbol{\theta}^{\mathrm{T}} \bar{\mathbf{x}}_i\right) \le (1 + t_3(n)) \cdot |k_2|. \qquad (9)$$

$$k_1 + (p-1)\Delta k_1 - j\Delta k_1 \le \sum_{i=1}^{n} z_i \le k_1 + p\Delta k_1 - j\Delta k_1,$$

$$j = \overline{0, p+l},\ \Delta k_1 > 0. \qquad (10)$$

The variables of the LPPs (7)–(10) are $\theta_i,\ i = \overline{0,r}$, $z_i,\ i = \overline{1,n}$. p, l are given natural numbers, $\Delta k_1 > 0$ is the given accuracy of the regions (10), one of which must contain the unknown number

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right|, (l+1)\Delta k_1<k_1.$$

Thus, the natural numbers p and l must statistically significantly guarantee that the unknown number $\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right|$ belongs to one of the regions (10), and the value of Δk_1 is chosen as a compromise between the number of LPPs and the absolute deviation from the number $\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right|$.

The macro-algorithm for obtaining the vector of estimates $\hat{\boldsymbol{\theta}}$ is as follows. The LPPs are solved sequentially in an arbitrary order. For each problem, starting from the first one, for the found vector of estimates $\hat{\boldsymbol{\theta}}$ we find estimates $\hat{\varepsilon}_i$ of the realizations ε_i of the RV $E, i=\overline{1,n},$ where

$$\hat{\varepsilon}_i=y_i-\hat{\boldsymbol{\theta}}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}},\ i=\overline{1,n}. \quad (11)$$

For them, the criterion χ^2 tests the hypothesis that the numbers $\hat{\varepsilon}_i, i=\overline{1,n},$ do not contradict the simple hypothesis of a normal distribution with parameters $0,\sigma^2$. All $p+l+1$ LPPs are solved. In general case, several consistent estimates $\hat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}$ can be obtained. In this case, the vector of estimates $\hat{\boldsymbol{\theta}}$ corresponding to the smallest value of the criterion χ^2 realization is selected. The logic of finding this vector of estimates is that on average, the realization of the criterion χ^2 is greater, the more the law of the distribution of numbers $\hat{\varepsilon}_i, i=\overline{1,n},$ differs from the simple hypothesis tested by the criterion χ^2. If no valid solution is found, then by the same reasoning, the vector of estimates $\hat{\boldsymbol{\theta}}$ corresponding to the smallest value of the criterion χ^2 realization is selected.

Remark 6. The presented iterative algorithm is easily modified for the case when the region (9) is represented as a union of subregions of the form (10). However, since the real value of a number k_2 can be both positive and negative, not one but two LPPs are solved for fixed values of $p, p_1, p_2, l, l_1, l_2, j, j_1, j_2$. In the first one, instead of (9), the following constraint is used:

$$|k_2|+(p_1-1)\Delta k_2-j_1\Delta k_2\le\sum_{i=1}^{n}\left(y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right)\le$$

$$\le|k_2|+p_1\Delta k_2-j_1\Delta k_2. \quad (12)$$

j_1 can take values from 0 to p_1+l_1, and p_1, l_1 are natural numbers, $\Delta k_2>0$. For l_1, the condition $l_1\Delta k_2<$ $<|k_2|+1$ is fulfilled.

In the second LPP, instead of the region (9), the following constraint is used:

$$-|k_2|+(p_2-1)\Delta k_2-j_2\Delta k_2\le\sum_{i=1}^{n}\left(y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right)\le$$

$$\le-|k_2|+p_2\Delta k_2-j_2\Delta k_2. \quad (13)$$

$j_2=\overline{0,p_2+l_2}$, the condition $-|k_2|+p_2\Delta k_2<0$ is imposed on p_2.

Thus, $(p+l+1)(p_1+l_1+1)(p_2+l_2+1)$ LPPs are solved in the modified iterative algorithm, since the intersection of the regions (12), (13) is an empty set.

5. Illustrative examples. *5.1. The first example.*

Remark 7. The following measure was chosen as an integral measure of comparisons of the components of two vectors $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$:

$$\left\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right\|=\left\|\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}-\frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|}\right\|, \quad (14)$$

where $\|\boldsymbol{\theta}\|=+\sqrt{\sum_{j=0}^{r}\theta_j^2}$.

The formula for the average deviation of experimental values from model values on input experimental data has the form

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right|, \quad (15)$$

where $\overline{\mathbf{x}}_{\boldsymbol{i}}=(1,x_{1i},\ldots,x_{mi})$, $\overline{\mathbf{x}}_{\boldsymbol{i}}\to y_i, i=\overline{1,n},$ $\boldsymbol{\theta}^{\mathrm{T}}=(\theta_0,$ $\theta_1,\ldots,\theta_m)$.

The formula for the average deviation of the average values of the estimated and ideal regression on the input experimental data has the form

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\boldsymbol{\theta}}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}-\boldsymbol{\theta}^{\mathrm{T}}\overline{\mathbf{x}}_{\boldsymbol{i}}\right|, \quad (16)$$

where $\hat{\boldsymbol{\theta}}^{\mathrm{T}}=\left(\hat{\theta}_0,\hat{\theta}_1,\ldots,\hat{\theta}_m\right)$ is the optimal solution to the LPP.

Below we give an illustrative example of using the first algorithm to estimate unknown coefficients of multivariate linear regression with finding the values of (14)–(16) by ideal multivariate linear regression, the true values of the 16 coefficients of which $\left(\theta_j, j=\overline{0,15}\right)$ are 1.59, 4.90, 3.58, –2.57, –2.25, 4.13, 1.45, 5.00, –1.47, 1.26, 4.49, –2.18, 4.78, –2.46, 4.98, –1.38. The modeling parameters used are as follows: $ME=0, DE=1200,$ the number of tests is 48, the ratio of the average absolute value of the ideal regression on the input experimental data to the average absolute value of the realizations of RV E is 69.42/26.19.

Table 1 shows the values of $y_i, \overline{\mathbf{x}}_{\boldsymbol{i}}, i=\overline{1,48}$.

The values of $k_1=\frac{1}{48}\sum_{i=1}^{48}|\varepsilon_i|$, $k_2=\frac{1}{48}\sum_{i=1}^{48}\varepsilon_i$, where ε_i are artificially generated realizations of the RV E, are equal to, 28.69 and –1.42, respectively.

Table 1 – The values of $y_i, \bar{x}_i, i = \overline{1, 48}$, for the first example.

y_i	$\bar{x}_i$															
71.20	1.00	1.65	4.48	4.80	1.53	1.22	2.47	3.71	2.35	1.15	3.08	1.68	4.58	2.43	2.87	3.45
46.44	1.00	4.37	1.81	2.22	4.65	3.98	2.76	2.27	2.88	2.66	1.60	3.17	2.04	1.27	1.71	2.67
36.19	1.00	2.59	4.35	4.35	4.18	4.38	3.73	4.81	3.95	4.87	3.65	4.57	3.12	4.58	1.14	4.97
53.45	1.00	1.93	3.05	4.65	2.03	3.71	2.57	1.14	3.99	1.98	2.25	1.09	4.13	1.52	3.40	3.99
51.00	1.00	4.85	1.65	2.80	1.65	2.61	4.51	4.68	3.66	1.68	2.62	1.92	2.56	3.92	4.53	2.18
81.15	1.00	3.63	1.02	2.11	3.18	2.26	1.94	1.10	3.77	2.62	3.10	3.02	4.35	2.37	2.68	2.77
18.35	1.00	4.21	1.03	1.31	4.44	3.12	2.19	2.27	3.91	3.97	4.25	2.53	4.62	4.09	2.25	4.12
21.17	1.00	2.59	3.43	2.67	4.73	4.07	1.40	1.52	3.76	3.93	1.82	3.76	1.97	3.33	2.06	1.04
39.63	1.00	4.92	3.96	2.82	3.05	2.37	1.97	3.70	1.64	2.30	3.02	1.27	3.33	1.37	1.14	2.94
60.66	1.00	2.93	1.60	1.50	1.84	2.94	4.13	1.53	1.78	1.70	3.42	4.98	1.28	3.87	1.93	3.53
7.23	1.00	2.00	1.97	4.80	1.92	2.17	3.91	4.12	3.20	4.35	1.71	2.33	2.70	3.36	1.93	3.71
98.11	1.00	2.99	4.72	4.89	3.46	3.13	1.14	4.94	1.78	1.76	1.40	1.45	4.26	1.66	4.92	1.36
53.66	1.00	4.07	4.79	3.39	3.06	3.99	1.69	4.93	3.54	3.94	4.30	3.74	2.78	3.10	2.07	2.81
50.27	1.00	1.85	4.84	3.48	3.62	2.50	3.09	4.72	1.67	1.83	3.41	4.48	1.44	1.43	4.84	4.92
92.97	1.00	1.76	2.61	4.59	3.66	1.26	1.75	4.41	3.86	1.17	3.08	3.51	2.97	2.24	4.99	4.22
104.99	1.00	3.23	2.34	2.48	3.91	3.59	4.56	4.95	1.72	2.96	3.41	4.14	2.36	4.18	3.96	1.25
40.65	1.00	1.34	2.39	3.27	2.53	2.60	2.04	3.67	2.44	3.24	1.03	3.76	2.85	4.78	1.91	3.91
32.24	1.00	3.73	1.19	4.23	4.64	2.33	3.86	1.39	4.08	2.43	4.95	1.29	1.13	1.06	1.68	2.64
110.91	1.00	4.30	3.92	4.36	1.40	4.99	1.12	2.83	3.59	4.92	1.45	3.24	3.03	2.86	2.39	3.82
107.73	1.00	1.80	1.60	2.62	1.74	1.41	4.93	1.74	2.50	3.76	1.98	1.49	4.86	3.42	3.40	4.04
89.53	1.00	3.81	1.96	3.01	3.18	2.69	4.21	2.23	4.41	2.54	3.26	1.18	3.39	2.06	3.27	3.68
110.72	1.00	3.44	1.71	1.42	2.36	4.49	4.63	4.84	1.18	1.39	3.34	3.23	4.60	1.12	3.22	3.95
69.43	1.00	1.06	4.36	4.01	2.58	4.43	3.95	3.85	4.70	2.46	2.04	4.37	4.71	2.83	4.22	4.19
24.80	1.00	2.78	4.27	3.62	1.57	2.37	1.67	2.57	1.33	3.08	4.47	2.20	3.33	1.37	2.05	1.46
93.57	1.00	2.48	1.89	4.57	1.07	2.57	1.35	1.53	2.74	4.61	4.47	3.99	1.60	2.74	1.67	3.00
106.43	1.00	3.68	4.10	2.50	3.70	4.73	1.72	2.66	4.44	4.78	3.22	3.28	3.12	4.41	3.15	3.48
31.05	1.00	3.92	4.43	4.13	1.73	3.10	3.83	1.14	4.51	4.24	2.08	3.55	3.13	4.04	3.22	2.61
59.66	1.00	4.78	1.39	1.86	4.37	1.23	2.50	1.42	1.54	2.65	1.66	4.81	4.33	3.86	4.13	3.59
72.63	1.00	3.93	3.34	4.88	1.12	4.43	2.68	2.26	1.97	1.72	1.51	3.37	1.05	4.72	3.79	2.30
120.61	1.00	3.26	3.62	3.44	4.68	3.72	3.82	3.33	4.61	3.83	2.09	4.27	4.79	3.87	2.31	1.33
57.61	1.00	3.10	1.81	4.48	4.00	2.89	2.18	3.02	1.33	4.11	3.82	3.64	2.87	1.32	1.58	3.80
70.50	1.00	2.90	4.28	2.93	3.30	3.41	3.58	3.38	1.44	1.89	2.85	4.93	4.27	2.87	1.76	3.93
70.05	1.00	2.11	3.76	3.36	4.82	3.55	4.95	3.80	4.64	2.84	1.09	1.44	1.35	2.82	1.32	3.81
58.06	1.00	2.95	3.37	1.92	2.77	1.32	2.80	3.04	1.12	4.27	1.31	2.66	1.12	1.28	4.21	3.92
72.03	1.00	2.96	3.95	4.79	3.05	3.79	4.74	2.14	4.35	3.58	4.76	4.79	2.03	1.20	4.33	2.65
64.63	1.00	1.65	4.94	1.16	1.19	4.10	2.15	2.49	2.72	3.35	1.33	4.87	1.88	1.85	2.26	4.44
30.52	1.00	1.01	1.26	2.72	1.66	4.81	2.86	1.02	4.75	3.87	2.81	1.47	2.43	4.52	4.47	1.32
151.75	1.00	3.79	2.12	2.23	2.68	3.43	1.84	4.68	2.91	1.12	3.92	3.59	2.32	1.15	1.88	2.45
136.47	1.00	4.34	4.52	1.41	2.34	4.29	4.68	1.37	2.16	3.62	4.99	4.88	2.26	3.40	4.15	4.07
77.82	1.00	2.43	3.20	1.02	4.08	1.76	1.19	4.28	4.05	1.50	4.87	1.90	3.39	2.88	4.38	3.53
98.32	1.00	3.56	2.45	3.53	4.44	4.41	4.62	1.99	4.95	3.69	1.99	2.30	2.53	3.57	1.56	2.12
105.21	1.00	4.78	4.64	3.99	2.64	3.17	4.61	2.73	3.34	4.35	4.05	1.71	3.96	3.07	3.58	3.72
8.03	1.00	1.58	1.13	1.29	4.43	1.75	3.25	2.05	1.90	2.26	2.82	3.11	4.63	1.05	4.26	1.59
61.37	1.00	3.34	3.03	1.15	4.37	3.84	1.81	2.42	3.01	4.84	1.24	4.98	3.41	4.62	4.55	2.98
30.32	1.00	2.95	1.44	1.79	2.24	1.22	4.33	4.79	3.76	4.03	2.33	1.00	4.63	4.36	4.14	2.62
104.27	1.00	2.18	2.95	2.71	3.09	3.79	1.69	3.00	4.27	3.84	2.67	1.02	4.97	4.39	3.41	2.28
74.71	1.00	3.05	3.24	3.24	4.07	4.24	4.51	1.80	4.66	2.37	4.64	1.01	3.31	2.30	4.82	2.71
175.05	1.00	4.20	3.78	2.86	3.18	1.96	3.02	4.71	1.57	3.47	4.28	3.48	4.76	2.73	3.79	3.69

The LPP has the form:

$$\min \sum_{i=1}^{48} z_i \,, \quad (17)$$

$$-z_i \le y_i - \sum_{j=0}^{12} \theta_j \bar{\boldsymbol{x}}_{\boldsymbol{i}} \le z_i \,,\ z_i \ge 0,\ i = \overline{1,48}\,, \quad (18)$$

$$k_1 - \tfrac{1}{3}k_1 \le \tfrac{1}{48}\sum_{i=1}^{48}\Big(y_i - \sum_{j=0}^{12}\theta_j \bar{\boldsymbol{x}}_{\boldsymbol{i}}\Big) \le k_1 + \tfrac{1}{3}k_1\,, \quad (19)$$

$$-1.3|k_2| \le \tfrac{1}{48}\sum_{i=1}^{48}\Big(y_i - \sum_{j=0}^{12}\theta_j \bar{\boldsymbol{x}}_{\boldsymbol{i}}\Big) \le 1.3|k_2|. \quad (20)$$

As a result of solving the LLP (17)–(20), the following estimates of the unknown coefficients were obtained: –23.06, 10.40, 1.56, –1.89, –9.22, 4.50, –0.21, 6.28, –0.38, 1.59, 5.69, –0.96, 5.24, –0.63, 4.58, 3.63. The value of (14) is equal to 1.07. The value of (15) is 26.19. The value of (16) is equal to 8.73. Thus, the scalar measure of deviation of the components of vectors $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ is 1.07, with ideal values of this measure being 0.02, but the filtering effect of the average value of the ideal regression line on the input experimental data decreased from 26.19 to 8.73, i.e. by 66.65 %, which allows the value (16) to be used in recognition systems for various purposes (for example, recognizing the beginning of a disease epidemic in a region from the list of diseases contained in the recognition system).

5.2. The second example. Below we give an illustrative example of using the second algorithm to estimate unknown coefficients of a multivariate linear regression with finding the values of (14)–(168) by ideal multivariate linear regression, the true values of 13 coefficients of which $\left(\theta_j,\ j=\overline{0,12}\right)$ are 1.05, –2.08, 2.02, –1.10, –1.55, 3.20, –3.70, –4.97, –2.61, 4.95, 3.49, 2.87, –3.54. The modeling parameters used are as follows: $ME = 0, DE = 3000$, the number of tests is 48, the ratio of the average absolute value of the ideal regression on the input experimental data to the average absolute value of the realizations of RV E is 14.15/42.03. It should be emphasized that the average absolute value of the realizations of RV E is 2.97 times greater than the average absolute value of the ideal regression on the input experimental data.

Since the size of the article does not allow us to present all the stages of the iterative algorithm, we will present only the LPP that corresponds to the consistent estimate $\hat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}$ (the χ^2 criterion has five degrees of freedom, the realization of the χ^2 criterion is 4.00, and the critical region for $\alpha = 0.05$ is given by the number 11.07). For this purpose, we give in Table 2 the values of $y_i, \bar{\boldsymbol{x}}_{\boldsymbol{i}}, i = \overline{1,48}$.

The value of $k_2 = \frac{1}{48}\sum_{i=1}^{48}\varepsilon_i$, where ε_i are the artificially generated realizations of the RV E, is –14.29.

The LPP has the form:

$$\min \sum_{i=1}^{48} z_i \,, \quad (21)$$

$$-z_i \le y_i - \sum_{j=0}^{12} \theta_j \bar{\boldsymbol{x}}_{\boldsymbol{i}} \le z_i \,,\ z_i \ge 0,\ i = \overline{1,48}\,, \quad (22)$$

$$41.61 \le \tfrac{1}{48}\sum_{i=1}^{48} z_i \le 42.45\,, \quad (23)$$

$$-1.3|k_2| \le \tfrac{1}{48}\sum_{i=1}^{48}\Big(y_i - \sum_{j=0}^{12}\theta_j \bar{\boldsymbol{x}}_{\boldsymbol{i}}\Big) \le 1.3|k_2|. \quad (24)$$

As a result of solving the LLP (21)–(24), we obtained the following estimates of the unknown coefficients: –20.36, –4.84, 1.41, 0.50, 0.46, 4.25, 4.87, –3.01, 0.03, –3.86, 6.53, 6.22, –8.32. The value of (14) is 1.28. The value of (15) is 42.03. The value of (16) is 9.94. Thus, compared to the first example, the scalar measure of deviation of the components of vectors $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ became worse, but the effect of filtering the average value of the ideal regression line on the input experimental data decreased from 42.03 to 9.94, i.e. by 76.35 %, which is better than in the first example.

6. Methodology of using the proposed algorithms. As shown by the two illustrative examples given in section 5, the proposed algorithms for estimating multivariate regression linear with respect to unknown coefficients are potentially efficient. Indeed, even with a fairly limited number of tests (48), a random factor variance of 1200 (the first example), 3000 (the second example), and a significant ratio of the average absolute value of the real regression on the values of the input variables in the statistical experiment tests to the average absolute value of the random factor realizations in these tests (69.42/26.19 for the first example, 14.15/42.03 for the second example), both algorithms demonstrated high filtering properties (66.65 % for the first example, 76.35 % for the second example).

For the correct use of the proposed algorithms in the general case, the following methodology of statistical simulation modeling is proposed.

1) set the parameters of the regression problem: the analytical expression of a multivariate regression linear with respect to unknown coefficients, the distribution of a random factor, the range of values of input arguments and unknown coefficients of the multivariate regression, the number of tests of the statistical experiment (and there may be several such values);

2) select two out of three or one out of three proposed algorithms and the algorithm with which its efficiency is compared (for example, LSM);

3) using a uniform distribution to generate the coefficients of the ideal regression (their absolute values and their signs), the values of the input variables, model a sufficient number of individual ideal regressions, for each of which simulate a statistical experiment $\left(\bar{\boldsymbol{x}}_{\boldsymbol{i}} \to y_i, i = \overline{1,n}\right)$. As the result of each statistical experiment, the estimates of unknown coefficients are found by the selected algorithms. Using the scalar measure (14), average comparative characteristics of their efficiency are found by them. According to the results of average comparative characteristics for a given class of multivariate regressions (item 1 of the methodology), the best algorithm is selected.

Table 2 – The values of $y_i, \bar{x}_i, i=\overline{1,48}$, for the second example.

y_i	$\bar{x}_i$												
–33.81	1.00	4.76	3.92	1.67	3.07	1.93	2.13	3.31	4.09	3.62	2.71	2.94	1.96
–78.63	1.00	2.61	2.28	2.94	2.20	1.25	2.00	3.55	1.71	4.76	4.50	3.61	3.11
10.00	1.00	3.49	4.74	3.18	4.08	1.20	3.48	1.76	4.74	2.83	2.96	4.82	2.18
–73.60	1.00	1.04	1.12	4.65	3.57	4.01	1.59	2.20	2.87	1.33	3.74	3.53	4.97
–1.08	1.00	1.49	2.01	2.28	1.40	1.62	3.48	1.87	1.94	4.86	2.74	2.40	1.26
39.96	1.00	2.39	2.61	4.14	2.75	1.47	2.75	2.45	1.72	1.24	1.29	3.98	3.91
–54.97	1.00	3.17	1.55	4.56	3.67	1.94	2.11	1.84	1.92	1.33	3.01	3.72	2.54
3.05	1.00	3.19	1.94	3.31	4.41	4.46	4.49	3.15	3.55	4.57	3.45	2.04	3.96
–25.91	1.00	2.38	2.90	3.41	1.25	4.22	1.40	1.95	4.18	2.12	1.57	3.91	2.04
–27.47	1.00	1.44	3.79	1.23	3.15	1.42	3.16	4.20	3.26	3.85	1.76	1.04	2.34
12.03	1.00	1.46	1.75	2.83	4.38	3.40	1.29	2.38	3.83	2.29	2.21	4.16	1.39
–0.54	1.00	3.86	3.25	1.60	1.39	2.70	1.11	4.21	3.15	1.89	3.21	3.88	1.16
10.28	1.00	3.74	3.19	1.22	3.13	3.63	4.52	1.67	4.08	3.43	4.91	1.20	1.99
–26.98	1.00	2.39	4.29	2.77	2.81	3.32	2.89	1.37	4.27	4.35	4.71	4.55	3.31
–24.92	1.00	3.31	1.27	4.26	3.41	3.10	4.70	2.62	1.68	2.49	3.31	4.45	3.67
102.55	1.00	1.94	1.91	1.20	1.54	1.70	1.88	4.20	3.67	1.52	1.40	2.50	4.44
–72.91	1.00	1.17	3.15	3.97	3.52	4.92	1.71	3.24	3.10	3.07	3.13	4.36	3.75
–39.34	1.00	1.04	3.65	1.70	4.67	4.12	3.59	2.84	1.72	4.75	3.42	3.30	1.76
–27.53	1.00	4.98	3.67	3.57	2.63	2.36	3.79	2.66	1.21	2.57	1.04	4.51	4.41
75.31	1.00	1.74	4.40	1.51	4.62	4.57	2.84	1.48	4.68	4.77	4.09	2.81	1.22
–129.01	1.00	2.35	1.54	3.97	4.18	1.62	1.31	1.41	4.60	1.17	4.73	2.67	3.88
–86.66	1.00	1.50	2.02	1.75	1.53	1.11	3.14	2.00	3.40	1.20	4.06	3.07	4.66
43.84	1.00	2.38	2.14	2.01	3.91	2.61	2.94	1.91	1.77	2.77	1.35	1.02	1.08
74.30	1.00	1.36	1.84	3.97	5.00	4.98	2.81	3.90	2.30	1.46	2.26	2.10	1.84
11.83	1.00	4.87	4.13	4.38	1.64	3.53	2.90	3.33	1.42	3.59	3.60	4.76	1.38
–72.21	1.00	3.79	2.64	2.69	1.46	2.10	3.91	3.40	1.74	1.54	1.83	4.64	1.99
–78.39	1.00	3.14	2.42	2.93	1.18	3.95	4.62	2.14	2.50	4.14	4.03	2.13	3.59
–15.96	1.00	3.10	2.38	4.46	4.66	3.75	4.39	4.80	4.05	1.31	3.75	1.48	4.80
–32.83	1.00	1.25	3.80	4.51	3.41	1.78	1.84	4.31	3.34	1.77	1.76	3.19	4.65
–52.88	1.00	1.31	1.19	4.49	1.72	4.04	3.25	1.55	2.94	2.88	1.66	1.85	2.41
11.88	1.00	1.62	1.61	2.53	4.80	1.45	1.15	2.83	1.52	1.30	2.30	4.30	3.59
92.08	1.00	3.04	4.56	1.46	3.92	3.56	1.14	1.10	1.19	3.22	1.71	4.86	2.24
–8.43	1.00	1.62	1.33	4.15	3.31	4.95	4.66	3.06	1.99	4.08	1.53	2.66	3.24
–37.11	1.00	4.55	1.59	1.40	4.77	3.02	4.87	3.92	2.45	2.45	1.02	1.19	3.53
–45.35	1.00	3.61	1.73	3.62	3.55	4.30	4.65	1.04	1.19	3.38	3.92	2.19	3.19
43.38	1.00	2.88	3.70	1.37	3.57	1.26	4.01	2.53	1.38	1.80	1.98	4.53	3.99
19.76	1.00	4.16	3.47	2.80	3.98	3.31	3.06	4.05	3.05	1.99	4.55	2.12	1.49
67.07	1.00	3.22	3.71	1.28	1.95	1.57	2.49	1.49	1.77	2.87	4.79	3.71	1.06
–8.05	1.00	4.60	1.98	1.93	4.26	3.95	1.64	4.49	1.64	2.87	4.05	1.67	1.31
–85.96	1.00	1.17	1.89	2.18	2.84	1.34	4.64	4.62	3.68	1.68	2.30	4.46	1.39
43.70	1.00	3.60	2.75	1.30	4.33	3.97	4.96	2.07	2.13	3.14	2.45	4.85	3.62
–6.41	1.00	3.18	2.52	4.65	1.74	4.60	2.23	2.14	4.72	3.37	2.90	1.90	2.31
25.51	1.00	3.47	4.14	4.11	1.86	3.04	4.44	3.53	4.15	1.30	1.52	1.19	2.12
85.34	1.00	4.98	4.00	4.36	2.50	4.65	2.51	4.72	4.28	1.86	3.60	2.35	1.94
10.16	1.00	2.14	1.82	3.32	2.02	3.39	3.17	2.64	4.48	3.48	3.50	1.59	3.99
68.99	1.00	2.04	2.05	1.14	3.41	1.52	2.97	4.07	2.66	3.65	1.70	4.68	3.96
–10.22	1.00	3.45	1.16	1.70	3.65	2.08	4.16	4.03	3.15	3.81	4.11	4.68	4.28
–6.35	1.00	3.83	1.65	4.74	4.03	4.77	1.14	3.07	1.76	4.74	4.27	2.76	2.09

Remark 7. The choice of a uniform distribution guarantees the generation of the most "rigorous" for estimating unknown parameters of individual regression problems. If desired, the uniform distribution can be replaced by other distributions in the simulation modeling system by the user.

Conclusions. 1. We substantiated the feasibility of using linear programming models to find estimates of a multivariate regression linear with respect to unknown coefficients.

2. We proposed a new algorithm for constructing estimates of unknown coefficients of a multivariate regression using the example of a linear multivariate regression, which uses a single linear programming model. The peculiarity of the algorithm is, in particular, that, unlike LSM, it does not give degenerate estimates for the case when the number of tests does not exceed the number of unknown coefficients.3. We proposed a new iterative algorithm for constructing estimates of a multivariate regression using the example of a multivariate linear regression, which finds consistent estimates of unknown coefficients implementing a special linear programming model at each iteration.

4. We give two illustrative examples confirming the efficiency of using the proposed algorithms with a small number of tests and a significant value of the variance of the random factor in comparison with the average value of the ideal regression on the values of the input variables in tests of a statistical experiment on the regression model.

5. We give the methodology for using the proposed algorithms in a statistical simulation modeling system.

References

1. Roustaei N. Application and interpretation of linear-regression analysis. *Medical hypothesis discovery and innovation in ophthalmology*. 2024. Vol. 13. No. 3. P. 151–159. DOI: doi.org/10.51329/mehdiophthal1506.
2. Liu J., Bellows B., Hu X. J., Wu J., Zhou Z., Soteros C., Wang L. A new time-varying coefficient regression approach for analyzing infectious disease data. *Scientific Reports.* 2023. Vol. 13. Art. 14687. DOI: doi.org/10.1038/s41598-023-41551-1.
3. Gao C. Robust regression via multivariate regression depth. *Bernoulli.* 2020. Vol. 26. No. 2. P. 1139–1170. DOI: doi.org/10.3150/19-BEJ1144.
4. Haghighat P., Gandara D., Kang L., Anahideh H. Fair multivariate adaptive regression splines for ensuring equity and transparency. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2024. Vol. 38. No. 20. P. 22076–22086. DOI: doi.org/10.1609/aaai.v38i20.30211.
5. Zhou Z., Ji J., Wang Y. Zhu Z., Chen J. Hybrid regression model via multi-variate adaptive regression spline and online sequential extreme learning machine and its application in vision servo system. *International Journal of Advanced Robotic Systems.* 2022. Vol. 19. No. 3. P. 1–13. DOI: doi.org/10.1177/17298806221108603.
6. Omer A., Sedeeq B., Ali T. A proposed hybrid method for multivariate linear regression model and multivariate wavelets (Simulation study). *Polytechnic Journal of Humanities and Social Sciences.* 2024. Vol. 5. No. 1. P. 112–124. URL: https://www.researchgate.net/publication/377558021 (дата звернення: 24.11.2025).
7. Abdulrahman A. T., Alshammari N. S. Factor analysis and regression analysis to find out the influencing factors that led to the countries' debt crisis. *Advances and Applications in Statistics.* 2022. Vol. 78. P. 1–16. DOI: doi.org/10.17654/0972361722047.
8. Hünermund P., Louw B. On the nuisance of control variables in causal regression analysis. *Organizational Research Methods.* 2023. Vol. 28. No. 1. P. 138–151. DOI: doi.org/10.1177/10944281231219274.
9. Szentpéteri Sz., Csáji B. C. Finite-sample identification of linear regression models with residual-permuted sums. *IEEE Control Systems Letters.* 2024. Vol. 8. P. 1523–1528. DOI: doi.org/10.1109/LCSYS.2024.3412856.
10. Nie B., Du Y., Du J., Rao Y., Zhang Y., Zheng X., Ye N., Jin H. A novel regression method: partial least distance square regression methodology. *Chemometrics and Intelligent Laboratory Systems.* 2023. Vol. 237. Art. 104827. DOI: doi.org/10.1016/j.chemolab.2023.104827.
11. Pavlov A. A., Holovchenko M. N., Drozd V. V. An adaptive method for building a multivariate regression. *Вісник Нац. техн. ун-ту «ХПІ»: зб. наук. пр. Темат. вип.: Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ». 2024. № 1 (11). С. 3–8. DOI: doi.org/10.20998/2079-0023.2024.01.01.
12. Павлов О. А., Кущ А. В. Задача лінійної регресії з лінійно залежними коефіцієнтами [Pavlov A.A., Kushch A.V. Formulation of a linear regression problem with linearly dependent coefficients]. *Адаптивні системи автоматичного управління.* Київ: КПІ ім. Ігоря Сікорського. 2025. № 2 (47). С. 147–157. DOI: doi.org/10.20535/1560-8956.47.2025.340201.

References (transliterated)

1. Roustaei N. Application and interpretation of linear-regression analysis. *Medical hypothesis discovery and innovation in ophthalmology*. 2024, vol. 13, no. 3, pp. 151–159. DOI: doi.org/10.51329/mehdiophthal1506.
2. Liu J., Bellows B., Hu X. J., Wu J., Zhou Z., Soteros C., Wang L. A new time-varying coefficient regression approach for analyzing infectious disease data. *Scientific Reports.* 2023, vol. 13, article 14687. DOI: doi.org/10.1038/s41598-023-41551-1.
3. Gao C. Robust regression via multivariate regression depth. *Bernoulli.* 2020, vol. 26, no. 2, pp. 1139–1170. DOI: doi.org/10.3150/19-BEJ1144.
4. Haghighat P., Gandara D., Kang L., Anahideh H. Fair multivariate adaptive regression splines for ensuring equity and transparency. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2024, vol. 38, no. 20, pp. 22076–22086. DOI: doi.org/10.1609/aaai.v38i20.30211.
5. Zhou Z., Ji J., Wang Y. Zhu Z., Chen J. Hybrid regression model via multi-variate adaptive regression spline and online sequential extreme learning machine and its application in vision servo system. *International Journal of Advanced Robotic Systems.* 2022, vol. 19, no. 3, pp. 1–13. DOI: doi.org/10.1177/17298806221108603.
6. Omer A., Sedeeq B., Ali T. A proposed hybrid method for multivariate linear regression model and multivariate wavelets (Simulation study). *Polytechnic Journal of Humanities and Social Sciences.* 2024, vol. 5, no. 1, pp. 112–124. Available at: https://www.researchgate.net/publication/377558021 (accessed 24.11.2025).
7. Abdulrahman A. T., Alshammari N. S. Factor analysis and regression analysis to find out the influencing factors that led to the countries' debt crisis. *Advances and Applications in Statistics.* 2022, vol. 78, pp. 1–16. DOI: doi.org/10.17654/0972361722047.
8. Hünermund P., Louw B. On the nuisance of control variables in causal regression analysis. *Organizational Research Methods.* 2023, vol. 28, no. 1, pp. 138–151. DOI: doi.org/10.1177/10944281231219274.
9. Szentpéteri Sz., Csáji B. C. Finite-sample identification of linear regression models with residual-permuted sums. *IEEE Control Systems Letters.* 2024, vol. 8, pp. 1523–1528. DOI: doi.org/10.1109/LCSYS.2024.3412856.
10. Nie B., Du Y., Du J., Rao Y., Zhang Y., Zheng X., Ye N., Jin H. A novel regression method: partial least distance square regression methodology. *Chemometrics and Intelligent Laboratory Systems.* 2023, vol. 237, article 104827. DOI: doi.org/10.1016/j.chemolab.2023.104827.
11. Pavlov A. A., Holovchenko M. N., Drozd V. V. An adaptive method for building a multivariate regression. *Vestnik Nats. tekhn. un-ta "KhPI": sb. nauch. tr. Temat. vyp.: Sistemnyy analiz, upravlenie i informatsionnye tekhnologii* [Bulletin of the National Technical University "KhPI": a collection of scientific papers. Thematic issue: System analysis, control and information technology]. Kharkov, NTU "KhPI" Publ., 2024. no. 1 (11). pp. 3–8. DOI: doi.org/10.20998/2079-0023.2024.01.01.

12. Pavlov A. A., Kushch A. V. Zadacha liniynoyi rehresiyi z liniyno zalezhnymy koefitsiyentamy [Formulation of a linear regression problem with linearly dependent coefficients]. *Adaptyvni systemy avtomatychnoho upravlinnya* [Adaptive Systems of Automatic Control]. Kyiv, *Igor Sykorsky KPI,* 2025. no. 2 (47). pp. 147–157. DOI: doi.org/10.20535/1560-8956.47.2025.340201. (In Ukr.).

Received 10.11.2025

УДК 004:519.24:681.3.06

О. А. ПАВЛОВ, доктор технічних наук, професор, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, професор кафедри інформатики та програмної інженерії; e-mail: alexanderpavlov1944@gmail.com; ORCID: https://orcid.org/0000-0002-6524-6410

А. В. КУЩ, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, аспірант кафедри інформатики та програмної інженерії; e-mail: antonkusch1@gmail.com, ORCID: https://orcid.org/0009-0003-0994-2985

АЛГОРИТМИ ПОБУДОВИ РЕГРЕСІЇ, ЛІНІЙНОЇ ВІДНОСНО НЕВІДОМИХ КОЕФІЦІЄНТІВ, НА ОБМЕЖЕНОМУ ОБ’ЄМІ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ

Дана публікація продовжує цикл наукових робіт авторів по створенню алгоритмів побудови багатовимірних регресій, лінійних відносно невідомих коефіцієнтів, з використанням моделей лінійного програмування. Для спрощення імітаційного моделювання їх ефективності, алгоритми наводяться для задачі багатовимірної лінійної регресії. Використання моделей лінійного програмування вимагає мінімізувати суму модулів різниць, що використовуються в загальній процедурі метода найменших квадратів. Оцінки невідомих коефіцієнтів, отриманих внаслідок розв’язання задачі лінійного програмування, є лінійними відносно вектору значень регресійної моделі в статистичному експерименту. Відомо, що в силу теореми Маркова оцінки невідомих коефіцієнтів, отриманих загальною процедурою метода найменших квадратів, є ефективними в класі лінійних незміщених оцінок. Таким чином, здавалось би, перехід від методу найменших квадратів до методу мінімізації суми модулів різниць, що використовується в методі найменших квадратів, є заздалегідь не продуктивним. Але це не так. З доведення теореми Маркова випливає, що матриця лінійної оцінки має бути сталою і не залежати від значень регресійної моделі в статистичному експерименту. Оцінки, отримані методом мінімізації суми модулів, цій умові не відповідають. Дійсно, матриця оцінок є оптимальним базисом для розв’язання задачі лінійного програмування симплекс-методом і залежить від значень регресійної моделі в статистичному експерименту. Така постановка задачі дозволяє в моделі оптимізації вводити лінійні обмеження, що використовують результати статистичних випробувань і реалізують додаткові властивості шуканої багатовимірної регресії. Перші дослідження цих алгоритмів показали їх ефективність, що дозволило авторам поставити задачу створення таких алгоритмів, які не тільки можуть конкурувати з загальною алгоритмічною процедурою метода найменших квадратів, але і для випадку обмеженого об’єму експериментальних даних, коли відношення середнього значення модуля реалізацій випадкового фактору в експерименті до середнього значення на ньому модуля істинної регресії є достатньо великою величиною. В цьому випадку ставити питання про знаходження оцінок невідомих коефіцієнтів, які практично не відрізняються від істинних, є не коректним, але, як показали експерименти і, зокрема, наведені в даній роботі приклади, можна знаходити достатньо хороші оцінки середніх значень істинної регресії на проведених експериментах, які можна використовувати, наприклад, при діагностуванні на ранній стадії початку епідемії різних захворювань чи в інших задачах розпізнавання.

Ключові слова: багатовимірна регресія, метод найменших квадратів, метод мінімізації суми модулів, модель лінійного програмування, симплекс-метод, оптимальний базис.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Павлов Олександр Анатолійович / Pavlov Alexander Anatolievich
Автор 2 / Author 2: Кущ Антон Віталійович / Kushch Anton Vitaliyovich