СИСТЕМНИЙ АНАЛІЗ І ТЕОРІЯ ПРИЙНЯТТЯ РІШЕНЬ

SYSTEM ANALYSIS AND DECISION-MAKING THEORY

UDC 004:519.24:681.3.06 **DOI: 10.20998/2079-0023.2022.02.01**

A. A. PAVLOV, M. N. HOLOVCHENKO

MODIFIED METHOD OF CONSTRUCTING A MULTIVARIATE LINEAR REGRESSION GIVEN BY A REDUNDANT DESCRIPTION

A number of scientific works of Prof. O. A. Pavlov and his disciples is devoted to the development of an original method of efficient estimation of coefficients at nonlinear terms of multivariate polynomial regression given by a redundant description under the conditions of an active experiment. The solution of the formulated problem is reduced to the sequential construction of univariate polynomial regressions (finding efficient estimates for the coefficients at nonlinear terms) and solving the corresponding systems of linear nondegenerate equations, the variables of which are the estimates for coefficients at nonlinear terms of the multivariate polynomial regression given by the redundant description. Thus, the problem was reduced to the estimation of the coefficients at linear terms of a multivariate linear regression given by a redundant description in the conditions of an active experiment. We have proposed an original method of its solution that uses a cluster analysis algorithm. The algorithm's implementation significantly reduces the enumeration of partial descriptions of multivariate linear regression followed by the finding of the residual sum of squares for each of them. This allows using the chi-squared criterion to build a linguistic variable which value gives a qualitative assessment (high reliability, acceptable reliability, low reliability, unreliability) to the obtained result. The analysis of the computational experiments made it possible to modify the proposed method, which significantly increased its efficiency, first of all, of finding a reliable structure of the sought multivariate linear regression given by the redundant description. The method modification, in particular, has reduced the enumeration of partial descriptions and has led to a more efficient use of the general procedure of the least squares method.

Keywords: multivariate linear regression, least squares method, redundant description, cluster analysis, active experiment, linguistic variable.

О. А. ПАВЛОВ, М. М. ГОЛОВЧЕНКО

МОДИФІКОВАНИЙ МЕТОД ПОБУДОВИ БАГАТОВИМІРНОЇ ЛІНІЙНОЇ РЕГРЕСІЇ, ЗАДАНОЇ НАДЛИШКОВИМ ОПИСОМ

Низка наукових робіт проф. Павлова О. А. та його учнів присвячена розробці оригінального метода ефективної оцінки коефіцієнтів при нелінійних членах багатовимірної поліноміальної регресії, заданої надлишковим описом в умовах активного експерименту. Розв'язання сформульованої задачі зводиться до послідовної побудови одновимірних поліноміальних регресій (знаходження ефективних оцінок коефіцієнтів при нелінійних членах) та розв'язання відповідних систем лінійних невироджених рівнянь, змінними яких є оцінки коефіцієнтів при нелінійних членах багатовимірної поліноміальної регресії, заданої надлишковим описом. Таким чином, задача звелася до оцінки коефіцієнтів при лінійних членах багатовимірної лінійної регресії, заданої надлишковим описом в умовах активного експерименту. Був запропонований оригінальний метод її розв'язання, що використовує алгоритм кластерного аналізу, реалізація якого суттєво зменшує перебір варіантів часткового опису лінійної багатовимірної регресії з наступним находженням для кожної з них залишкової суми квадратів, що дозволяє з використанням критерія хі-квадрат побудувати лінгвістичну змінну, значення якої дає якісну оцінку (висока достовірність, допустима достовірність, мала достовірність, недостовірність) отриманого результату. Аналіз проведених обчислювальних експериментів дозволив модифікувати запропонований метод, що суттєво підвищило його ефективність, в першу чергу знаходження достовірної структури шуканої лінійної багатовимірної регресії, заданої надлишковим описом. Модифікація методу, зокрема, зменшила перебір варіантів часткових описів та привела до більш ефективного використання загальної процедури методу найменших квадратів.

Ключові слова: багатовимірна лінійна регресія, метод найменших квадратів, надлишковий опис, кластерний аналіз, активний експеримент, лінгвістична змінна.

Introduction. Models of regression analysis are widely used in various spheres of human activity, such as science, engineering, economics, medicine [1–9], etc. Forecasting models are created using multivariate (in particular, linear) regression. As the analysis of the scientific literature shows, the theoretical and practical aspects of creating efficient universal methods for constructing regression models with the simultaneous finding of input deterministic variables that affect the value of the output variable are still an actual problem today. The method of constructing a multidimensional polynomial regression given by a redundant description based on the results of an active experiment described in scientific works [10, 11] reduces this problem to the problem of constructing a multidimensional linear regression (MLR) given by a redundant description. An original method of its solution is given in [12] based on:

- using the algorithm of cluster analysis which reduces the selection of partial descriptions of the searched MLR;
- construction of a linguistic variable based on the simultaneous fulfillment of two criteria: minimization of the residual sum of squares (RSS) and the χ^2 criterion. The value of the linguistic variable characterizes the reliability of the final result.

In this work, we increase the efficiency of the new method, in particular, due to the modification of the cluster analysis algorithm; the use of a test sequence (results of a repeated active experiment) to find the RSS; refinement of the estimates of the coefficients of the found MLR structure for the entire data ensemble using only the previously constructed inverse matrix.

The problem statement. A MLR given by a redundant description has the following form:

$$Y(\bar{\boldsymbol{x}}) = b_0 + \sum_{j=1}^{m} b_j x_j + E\,, \qquad (1)$$

where $\bar{\boldsymbol{x}} = \left(x_1, \ldots, x_m\right)^{\mathrm{T}}$ is the redundant vector of deterministic input variables;

E is a random variable, its mathematical expectation $ME = 0$, its variance $DE < \infty$;

$\bar{x}_i = \left(x_{1i}, \ldots, x_{mi}\right)^{\mathrm{T}}$ is the vector of values of input deterministic variables in i -th test, $i = \overline{1,n}$;

The result of an active experiment is the data of $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$ where

$$y_i = b_0 + \sum_{j=1}^{m} b_j x_{ji} + \varepsilon_i\,, \qquad (2)$$

where ε_i is the realization of the random variable E in the i -th experiment.

We consider y_i as an implementation of a virtual random variable Y_i, $i = \overline{1,n}$; random variables Y_i, $i = \overline{1,n}$, are independent

$$Y_i = b_0 + \sum_{j=1}^{m} b_j x_{ji} + E_i, \, i = \overline{1,n}\,, \qquad (3)$$

where the random variables E, $i = \overline{1,n}$, are virtual independent copies of the random variable E (ε_i are considered realizations of the random variable E_i, $i = \overline{1,n}$). Let $\boldsymbol{y} = \left(y_1, \ldots, y_n\right)^{\mathrm{T}}$, $\boldsymbol{Y} = \left(Y_1, \ldots, Y_n\right)^{\mathrm{T}}$. The values of the coefficients b_j, $j = \overline{0,m}$, are unknown. A redundant description (or representation) means that some input variables may not affect the value of the output variable. We need to find the true structure of the MLR using the general procedure of the least squares method (LSM) and estimate the values of its coefficients.

Modified procedures of the construction method for the MLR given by a redundant description [12]. *1. Design of the active experiment.* As computer experiments have shown, it is desirable not to set the values of the input variables equal to zero, i.e. $\forall x_{ij} \neq 0$. The explanation of the recommendation follows from the fact that, based on the results of the active experiment, all input variables that do not affect the output variable should be excluded.

2. Modified cluster analysis algorithm. The cluster analysis algorithm [12] consists of the following routines.

Finding estimates of the coefficients b_j, $j = \overline{0,m}$, and ranking of the modules of their values according to the results of an active experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$ using the redundant description (1) and the general scheme of LSM:

$$\left|\hat{b}_{j_1}\right| \geq \left|\hat{b}_{j_2}\right| \geq \ldots \geq \left|\hat{b}_{j_{m+1}}\right|.$$

We use a sequential procedure to split all coefficients b_j, $j = \overline{0,m}$, into two classes: M_1 and M_2 .

The first step: $b_{j_1} \in M_1, b_{j_{m+1}} \in M_2$. If $\left|\hat{b}_{j_1}\right| - \left|\hat{b}_{j_2}\right| < \left|\hat{b}_{j_2}\right| - \left|\hat{b}_{j_{m+1}}\right|$, then $b_{j2} \in M_1$, else $b_{j2}, b_{j3}, \ldots, b_{jm+1} \in M_2$. The partitioning is complete.

The second step: $b_{j_2} \in M_1$. If $\frac{1}{2}\left(\left|\hat{b}_{j_1}\right| + \left|\hat{b}_{j_2}\right|\right) - \left|\hat{b}_{j_3}\right| < \left|\hat{b}_{j_3}\right| - \left|\hat{b}_{j_{m+1}}\right|$, then $b_{j_3} \in M_1$, else the partitioning is complete, $M_1 = \left\{b_{j_1}, b_{j_2}\right\}$, $M_2 = \left\{b_{j_3}, \ldots, b_{j_{m+1}}\right\}$.

Step l : $b_{j_l} \in M_1$. If $\frac{1}{l}\sum_{i=1}^{l}\left|\hat{b}_{j_i}\right| - \left|\hat{b}_{j_{l+1}}\right| < \left|\hat{b}_{j_{l+1}}\right| - \left|\hat{b}_{j_{m+1}}\right|$, then $b_{j_{l+1}} \in M_1$, otherwise the partitioning is complete, $M_1 = \left\{b_{j_1}, \ldots, b_{j_l}\right\}$, $M_2 = \left\{b_{j_{l+1}}, \ldots, b_{j_{m+1}}\right\}$.

It is obvious that in a limited number of steps the algorithm completes the partitioning of coefficients b_j, $j = \overline{0,m}$, into two classes M_1 and M_2 .

The partitioning of the coefficients of the MLR given by a redundant description into two classes [12] allows to find estimates of the coefficients of partial descriptions of the desired MLR based on the results of an active experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$. We form the partial descriptions according to the following rule: each of them includes all terms whose coefficients are in the class M_1 as well as all possible different combinations of the members whose coefficients belong to the set M_2 .

Remark. If $\hat{b} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{y}$ in the LSM formula, and the matrix $\boldsymbol{A}$ was built to estimate the MLR coefficients given by the redundant description based on the results of an active experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$, then in order to build the corresponding matrix to estimate the coefficients of a partial description of the MLR based on the results of the active experiment $\left(\bar{x}_i \to y_i, i = \overline{1,n}\right)$, we need to keep only those columns in the matrix $\boldsymbol{A}$ corresponding to the coefficients included in the partial description. The vector $\boldsymbol{y}$ does not change.

In the general case, having a sufficient number m of input variables and a significant scatter of the absolute non-zero coefficient values of the sought MLR, the set M_2 can include non-zero coefficients, and its cardinality $|M_2|$ can be a sufficiently large number, which can lead to the construction of an unacceptably large number of partial descriptions of the MLR. Therefore, we propose the following modification of the cluster analysis algorithm: the algorithm is supplemented with the following procedure.

Exclude from the set M_2 the members $b_{j_l}, b_{j_{l+1}}, \ldots, b_{j_{m+1}}$ that satisfy the conditions

$$\left|\hat{b}_{j_l}\right| - \left|\hat{b}_{j_{m+1}}\right| \le \Delta_{DE,n}, \; \left|\hat{b}_{j_{l-1}}\right| - \left|\hat{b}_{j_{m+1}}\right| > \Delta_{DE,n}, \qquad (5)$$

where $\Delta_{DE,n} > 0$ is an expert bound, its value is found according to the results of experiments and depends on the value of DE and the number of experiments n. The bound statistically significantly guarantees that $b_{j_k} = 0$, $k = \overline{l, m+1}$. That is $\Delta_{DE,n} > 0$ must be a sufficiently small number. This means that non-zero coefficients may remain in M_2 and the need to enumerate the set M_2 remains, but the number of redundant partial descriptions of the MLR significantly decreases. The desired regression is found [12] using simultaneously two criteria: the minimum of RSS and the value of χ^2 that checks the hypothesis that the estimates of the random variable E realizations correspond to the given distribution.

3. Modification of RSS construction algorithm for partial descriptions of a MLR. In [12], the RSS for each partial description is given by formula (22).

Remark. For the MLR problem, we set to zero $f(\bar{x}_i)$, $i = \overline{1,n}$, in formulas (22), (23) [12]. In [12], the RSS is found according to the results of an active experiment $\left(\bar{x}_i \to y_i,\, i = \overline{1,n}\right)$. This leads to such a value of the RSS of the partial description that has a correct structure and is practically minimal but never minimal. We proposed there to use the main idea of the group method of data handling by O. G. Ivakhnenko, namely, the RSS of each partial description is found by data that are not the components of the vector $\boldsymbol{y}$ (2). To find such data, we implement a repeated active experiment $\left(\bar{x}_i \to y_{n+i},\, i = \overline{1,n}\right)$. Next we will show that in this case we use only the inverse matrices found by the results of the active experiment $\left(\bar{x}_i \to y_{n+i},\, i = \overline{1,n}\right)$ to find the coefficient estimates of the MLR partial description on the whole dataset. Thus, the numbers y_i, $i = \overline{1,n}$, are replaced with the numbers y_{n+i}, $i = \overline{1,n}$, in the formula (23) [12] to find the RSS of each partial description.

That is

$$\mathrm{RSS}\left(M_1, M_2^j\right) = \sum_{i=1}^{n}\Big[y_{n+i} - \sum\nolimits_{\forall \hat{b}_l \in M_1}\left(\hat{b}_l \cdot x_{li}\right) \vee \hat{b}_0 - \\ - \sum\nolimits_{\forall \hat{b}_l \in M_2^j}\left(\hat{b}_l \cdot x_{li}\right) \vee \hat{b}_0 \Big]^2$$

where $M_2^j \subset M_2$; M_1, M_2^j unambiguously define a partial description of the MLR.

Remark. The expression $\vee \hat{b}_0$ included in the formulas (12) and (22), (23) [12] means that the coefficient b_0 can either be included in the set M_1 or M_2^j, or not belong to M_1 or M_2^j.

4. Modification of the algorithm for finding realizations of the random variable E. 4.1. Rank and renumber the partial descriptions of the MLR by the increasing values of their RSSs, and then keep the first t of them that meet the conditions

$$\mathrm{RSS}\left(M_1, M_2^1\right) \lesssim \mathrm{RSS}\left(M_1, M_2^2\right) \lesssim \ldots \lesssim \mathrm{RSS}\left(M_1, M_2^t\right), \quad (6)$$

$$\mathrm{RSS}\left(M_1, M_2^t\right) - \mathrm{RSS}\left(M_1, M_2^1\right) < \\ < \mathrm{RSS}\left(M_1, M_2^{t+1}\right) - \mathrm{RSS}\left(M_1, M_2^t\right).$$

Inequality (6) means that the values of RSSs of the first t partial descriptions are practically the same, and the values of RSSs of other partial descriptions are sufficiently different from them.

Remark. t can be equal to one. The practical equality of RSS values can be set in the form of fractions of the value of the minimal $\mathrm{RSS}\left(M_1, M_2^1\right)$ and found experimentally. On average, it states 2÷3 % of the $\mathrm{RSS}\left(M_1, M_2^1\right)$ value.

Let us find again for each of the partial descriptions $\left(M_1, M_2^l\right)$, $l = \overline{1,t}$, estimates of their coefficients using the whole experimental dataset: $\left(\bar{x}_i \to y_i,\, i = \overline{1,n}\right)$ and $\left(\bar{x}_i \to y_{n+i},\, i = \overline{1,n}\right)$. Let us show that for this it is not necessary to find a new matrix $\boldsymbol{A}$ and the corresponding inverse matrix $\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}$ in the general formula of the LSM when using the vector of initial data with the components y_i, $i = \overline{1, 2n}$.

Let us consider the basic formula of LSM in the general case, i.e. for an arbitrary partial description of a MLR.

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}, \qquad (7)$$

$$\hat{\boldsymbol{\theta}}_{\mathbf{RV}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}, \qquad (8)$$

where $\boldsymbol{y} = \left(y_1, \ldots, y_n\right)^{\mathrm{T}}$ is the realization of the random vector $\boldsymbol{Y}$;

$\hat{\boldsymbol{\theta}}_{\mathbf{RV}}$ is a random vector, the vector $\hat{\boldsymbol{\theta}}$ is its realization, $M\hat{\boldsymbol{\theta}}_{\mathbf{RV}} = 0$;

$DY_i = DE$;

$\boldsymbol{\theta}$ is a vector of exact coefficients of an arbitrary partial description $\left(M_1, M_2^j\right)$.

The matrix $\boldsymbol{A}$ was built based on the results of an active experiment $\left(\bar{x}_i \to y_i,\, i = \overline{1,n}\right)$. Let us conduct the repeated active experiment $\left(\bar{x}_i \to y_{n+i},\, i = \overline{1,n}\right)$, let $\boldsymbol{y}^1 =$

$= \left(y_{n+i},\, i=\overline{1,n}\right)^{\mathrm{T}}$ be the implementation of a virtual random vector $\boldsymbol{Y}_1$. Then,

$$M\hat{\boldsymbol{\theta}}_{\mathbf{RV}} = M\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}_1 = \boldsymbol{\theta}\,, \tag{9}$$

where $DY_{1i} = DE,\, i=\overline{1,n}$;

$Y_i,\, i=\overline{1,n},\, Y_{1i},\, i=\overline{1,n}$ are independent random variables. The matrix $\boldsymbol{A}$ in formulas (7)–(9) is the same.

Statement. The estimates obtained by formulas

$$\hat{\boldsymbol{\theta}}^{*} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{2}\left(y+y^{1}\right), \tag{10}$$

$$\hat{\boldsymbol{\theta}}^{*}_{\mathbf{RV}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{2}\left(\boldsymbol{Y}+\boldsymbol{Y}_1\right) \tag{11}$$

are unbiased, that is, $\forall j\; M\hat{\theta}^{*}_{j\mathrm{RV}} = \theta_j$ and

$$\forall j\; D\hat{\theta}^{*}_{j\mathrm{RV}} = \frac{1}{2}D\hat{\theta}_{j\mathrm{RV}}\,,$$

where $\hat{\theta}_{j\mathrm{RV}}$, $\hat{\theta}^{*}_{j\mathrm{RV}}$ are the j-th components of the vectors $\hat{\boldsymbol{\theta}}_{\mathbf{RV}}$, $\hat{\boldsymbol{\theta}}^{*}_{\mathbf{RV}}$, respectively.

Proof.

$$M\hat{\boldsymbol{\theta}}^{*}_{\mathbf{RV}} = M\left[\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{2}\left(\boldsymbol{Y}+\boldsymbol{Y}_1\right)\right] =$$

$$= \frac{1}{2}\left[M\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y} + M\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}_1\right] = \frac{1}{2}\left(\boldsymbol{\theta}+\boldsymbol{\theta}\right) = \boldsymbol{\theta}.$$

Vectors $\boldsymbol{Y}$ and $\boldsymbol{Y}_1$ are independent.

$$D\hat{\boldsymbol{\theta}}^{*}_{\mathbf{BB}} = D\left[\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{2}\left(\boldsymbol{Y}+\boldsymbol{Y}_1\right)\right] =$$

$$= \frac{1}{4}\left[D\left(\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}\right) + D\left(\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}_1\right)\right] = \frac{1}{2}D\hat{\boldsymbol{\theta}}_{\mathbf{RV}}\,.$$

Remark. Formulas (10), (11) follow from the equality

$$\left(\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)\begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{A}\end{pmatrix}\right)^{-1}\left(\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{Y}_1\end{pmatrix}\right) = \frac{1}{2}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{2}\left(\boldsymbol{Y}+\boldsymbol{Y}_1\right).$$

Corollary. Suppose that the experiment $\left(\bar{x}_i \to y_i,\; i=\overline{1,n}\right)$ is repeated k times. Then the estimates obtained by the formula

$$\hat{\boldsymbol{\theta}}^{*}_{\mathbf{RV}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\frac{1}{k}\left(\boldsymbol{Y}+\sum_{j=1}^{k-1}\boldsymbol{Y}_j\right)$$

are unbiased, and the variances of the components of the random vector $\hat{\boldsymbol{\theta}}^{*}_{\mathbf{RV}}$ are k times less than the variances of the components of the random vector $\hat{\boldsymbol{\theta}}_{\mathbf{RV}}$ (8).

Remark. An active experiment $\left(\bar{x}_i \to y_i,\, i=\overline{1,n}\right)$ can consist of k repeated experiments, and then the whole dataset can be split in two parts: the first one for finding estimates of the coefficients of partial descriptions, the second one for finding their RSSs. Inverse matrices can be built for the input data of a single experiment.

4.2. We find estimates E_i^j of realizations of the random variable E for a partial description $\left(M_1, M_2^j\right),\; j=\overline{1,t}$ of the MLR by the formulas

$$\begin{aligned} E_i^j &= y_i - \sum\nolimits_{\forall\hat{b}_l\in M_1}\left(\hat{b}_l\cdot x_{li}\right)\vee\hat{b}_0 - \\ &\quad -\sum\nolimits_{\forall\hat{b}_l\in M_2^j}\left(\hat{b}_l\cdot x_{li}\right)\vee\hat{b}_0,\; i=\overline{1,n}, \\ E_{n+i}^j &= y_{n+i} - \sum\nolimits_{\forall\hat{b}_l\in M_1}\left(\hat{b}_l\cdot x_{li}\right)\vee\hat{b}_0 - \\ &\quad -\sum\nolimits_{\forall\hat{b}_l\in M_2^j}\left(\hat{b}_l\cdot x_{li}\right)\vee\hat{b}_0,\; i=\overline{1,n}. \end{aligned} \tag{12}$$

5. A modified algorithm for a linguistic variable construction. From the set of the MLR's partial descriptions $\left(M_1, M_2^j\right),\;\; j=\overline{1,t}$, we choose the one containing the minimum number of terms.

Remark. In most cases, it corresponds to the minimum RSS.

We propose to consider this partial description as an effective approximation of the sought MLR that contains only the input variables significantly affecting the output variable's value. For this description we build a linguistic variable. Let us denote this MLR's description $\left(M_1, M_2^{j_1}\right)$.

If we know the density function of the random variable E or aware of its analytical expression with accuracy up to the values of its numerical parameters, then we check the hypothesis that these estimates are its realizations. We do this based on the estimates of the realization of the random variable E (12) for the partial description $\left(M_1, M_2^{j_1}\right)$ of the MLR using χ^2 criterion. Let $\chi^2\left(M_1, M_2^{j_1}\right)$ be the realization of χ^2 criterion for the partial description of the MLR $\left(M_1, M_2^{j_1}\right)$ that has $r \ge 3$ degrees of freedom if the checked hypothesis is true. Suppose expert numbers q_1, q_2 are given satisfying the conditions

$$P\left(\chi^2 \ge q_1\right) = 0.05\,,\; P\left(\chi^2 \ge q_2\right) = 0.4\,, \tag{13}$$

the random variable χ^2 has $r \ge 3$ degrees of freedom, and $r-2$ is the argument value at which the density function of the random variable χ^2 reaches its unique maximum. Then, if the condition

$$r-2 \le \chi^2\left(M_1, M_2^{j_1}\right) \le q_2 \tag{14}$$

is met, then, with a high probability, the partial description $\left(M_1, M_2^{j_1}\right)$ of the MLR is the sought MLR that contains all variables, each of which essentially affects the original variable. If

$$q_2 \le \chi^2\left(M_1, M_2^{j_1}\right) < q_1\,, \tag{15}$$

then the obtained result has a sufficient reliability degree. If

$$\chi^2\left(M_1, M_2^{j_1}\right) \ge q_1\,,$$

then the obtained result is unreliable, primarily due to insufficiently accurate estimates of the partial description $\left(M_1, M_2^{j_1}\right)$ coefficients of the MLR.

Remark 1. If there are several descriptions with a minimum number of members $\left(M_1, M_2^{j_1}\right)$, ..., $\left(M_1, M_2^{j_p}\right)$, then we study all of them, but the first contender for the solving is the one that achieves

$$\min_{i=\overline{1,p}} \left[r - 2 - \chi^2 \left(M_1, M_2^{j_i}\right) \right].$$

Remark 2. As a possible additional analysis, specialists in the subject field may consider other partial descriptions from the set $\left(M_1, M_2^{j}\right)$, $j = \overline{1,t}$, which have been checked by the χ^2 criterion for belonging of the numbers (12) to the distribution of the random variable E.

Conclusions. The content of this scientific article is the presentation of modified algorithmic procedures for the method of constructing a multivariate linear regression given by a redundant description [12]. The modifications follow from the analysis of statistical experiments results.

1. We have given recommendations for an active experiment design.

2. We have modified the cluster analysis algorithm for splitting the coefficients of a multivariate linear regression given by a redundant description into two classes.

3. We have modified the algorithm for finding the residual sum of squares for partial descriptions of a multivariate linear regression.

4. We have modified the algorithm for finding realizations of the random variable E.

5. We have presented a new theoretical property of the least squares method.

6. We have modified the algorithm for the linguistic variable construction.

References

1. Yu L. Using negative binomial regression analysis to predict software faults: a study of Apache Ant. *International Journal of Information Technology and Computer Science (IJITCS)*. 2012. Vol. 4, no. 8. P. 63–70. doi: 10.5815/ijitcs.2012.08.08
2. Shahrel M.Z., Mutalib S., Abdul-Rahman S. PriceCop – price monitor and prediction using linear regression and LSVM-ABC methods for e-commerce platform. *International Journal of Information Engineering and Electronic Business (IJIEEB)*. 2021. Vol. 13, no. 1. P. 1–14. doi: 10.5815/ijieeb.2021.01.01
3. Satter A., Ibtehaz N. A regression based sensor data prediction technique to analyze data trustworthiness in cyber-physical system. *International Journal of Information Engineering and Electronic Business (IJIEEB).* 2018. Vol. 10, no. 3. P. 15–22. doi: 10.5815/ijieeb.2018.03.03
4. Isabona J., Ojuh D. O. Machine learning based on kernel function controlled gaussian process regression method for in-depth extrapolative analysis of Covid-19 daily cases drift rates. *International Journal of Mathematical Sciences and Computing (IJMSC)*. 2021. Vol. 7, no. 2. P. 14–23. doi: 10.5815/ijmsc.2021.02.02
5. Sinha P. Multivariate polynomial regression in data mining: methodology, problems and solutions. *International Journal of Scientific & Engineering Research*. 2013. vol. 4, iss. 12. P. 962–965
6. Kalivas J. H. Interrelationships of multivariate regression methods using eigenvector basis sets. *Journal of Chemometrics*. 1999. Vol. 13 (2). P. 111–132. doi: 10.1002/(SICI)1099-128X(199903/04)13:2<111::AID-CEM532>3.0.CO;2-N
7. Ortiz-Herrero L., Maguregui M. I., Bartolomé L. Multivariate (O)PLS regression methods in forensic dating. *TrAC Trends in Analytical Chemistry*. 2021. Vol. 141. 116278. doi: 10.1016/j.trac.2021.116278
8. Guo G., Niu G., Shi Q., Lin Q., Tian D., Duan Y. Multi-element quantitative analysis of soils by laser induced breakdown spectroscopy (LIBS) coupled with univariate and multivariate regression methods. *Analytical Methods*. 2019. Vol. 11, iss. 23. P. 3006–3013, doi: 10.1039/C9AY00890J
9. Babatunde G., Emmanuel A. A., Oluwaseun O. R., Bunmi O. B., Precious A. E. Impact of climatic change on agricultural product yield using *k*-means and multiple linear regressions. *International Journal of Education and Management Engineering (IJEME).* 2019. Vol. 9, no. 3. P. 16–26. doi: 10.5815/ijeme.2019.03.02
10. Pavlov A. A. Holovchenko M. N., Drozd V. V. Construction of a multivariate polynomial given by a redundant description in stochastic and deterministic formulations using an active experiment. *Вісник Нац. техн. ун-ту «ХПІ»: зб. наук. пр. Темат. вип.: Системний аналіз, управління та інформаційні технології*. Харків: НТУ «ХПІ», 2022. № 1 (7). С. 3–8. doi: 10.20998/2079-0023.2022.01.01
11. Pavlov A., Holovchenko M., Mukha I. et al. Mathematics and software for building nonlinear polynomial regressions using estimates for univariate polynomial regressions coefficients with a given (small) variance. *Lecture Notes on Data Engineering and Communications Technologies*. 2022. Vol. 134. P. 288–303. doi: 10.1007/ 978-3-031-04812-8_25
12. Павлов О. А., Головченко М. М., Ревич М.М. Метод оцінки коефіцієнтів при лінійних членах багатовимірної поліноміальної регресії, заданої надлишковим описом. *Адаптивні системи автоматичного управління: міжвідомчий наук.-техн. збірник*. Київ: НТУУ «КПІ», 2022. Том 1, № 40. С. 110–117. doi: 10.20535/1560-8956.40.2022.261665

References (transliterated)

1. Yu L. Using negative binomial regression analysis to predict software faults: a study of Apache Ant. *International Journal of Information Technology and Computer Science (IJITCS)*. 2012, vol. 4, no. 8, pp. 63–70. doi: 10.5815/ijitcs.2012.08.08
2. Shahrel M.Z., Mutalib S., Abdul-Rahman S. PriceCop – price monitor and prediction using linear regression and LSVM-ABC methods for e-commerce platform. *International Journal of Information Engineering and Electronic Business (IJIEEB)*. 2021, vol. 13, no. 1, pp. 1–14. doi: 10.5815/ijieeb.2021.01.01
3. Satter A., Ibtehaz N. A regression based sensor data prediction technique to analyze data trustworthiness in cyber-physical system. *International Journal of Information Engineering and Electronic Business (IJIEEB).* 2018, vol. 10, no. 3, pp. 15–22. doi: 10.5815/ijieeb.2018.03.03
4. Isabona J., Ojuh D. O. Machine learning based on kernel function controlled gaussian process regression method for in-depth extrapolative analysis of Covid-19 daily cases drift rates. *International Journal of Mathematical Sciences and Computing (IJMSC)*. 2021, vol. 7, No. 2, pp. 14–23. doi: 10.5815/ijmsc.2021.02.02
5. Sinha P. Multivariate polynomial regression in data mining: methodology, problems and solutions. *International Journal of Scientific & Engineering Research*. 2013, vol. 4, iss. 12, pp. 962–965
6. Kalivas J. H. Interrelationships of multivariate regression methods using eigenvector basis sets. *Journal of Chemometrics*. 1999, vol. 13 (2), pp. 111–132. doi: 10.1002/(SICI)1099-128X(199903/04)13:2<111::AID-CEM532>3.0.CO;2-N
7. Ortiz-Herrero L., Maguregui M. I., Bartolomé L. Multivariate (O)PLS regression methods in forensic dating. *TrAC Trends in Analytical Chemistry*. 2021, vol. 141, 116278. doi: 10.1016/j.trac.2021.116278
8. Guo G., Niu G., Shi Q. et al. Multi-element quantitative analysis of soils by laser induced breakdown spectroscopy (LIBS) coupled with univariate and multivariate regression methods. *Analytical Methods*. 2019, vol. 11, iss. 23, pp. 3006–3013, doi: 10.1039/C9AY00890J.
9. Babatunde G., Emmanuel A. A., Oluwaseun O. R., Bunmi O. B., Precious A. E. Impact of climatic change on agricultural product yield using *k*-means and multiple linear regressions. *International Journal of Education and Management Engineering (IJEME).* 2019, vol. 9, no. 3, pp. 16–26. doi: 10.5815/ijeme.2019.03.02
10. Pavlov A. A. Holovchenko M. N., Drozd V. V. Construction of a multivariate polynomial given by a redundant description in stochastic and deterministic formulations using an active experiment. *Visnyk Nats.*

tekhn. un-tu "KhPI": zb. nauk. pr. Temat. vyp.: Systemnyy analiz, upravlinnya ta informatsiyni tekhnologiyi [Bulletin of the National Technical University "KhPI": a collection of scientific papers. Thematic issue: System analysis, management and information technology]. Kharkov, NTU "KhPI" Publ., 2022, no. 1 (7), pp. 3–8. doi: 10.20998/2079-0023.2022.01.01

11. Pavlov A., Holovchenko M., Mukha I., Lishchuk K. Mathematics and software for building nonlinear polynomial regressions using estimates for univariate polynomial regressions coefficients with a given (small) variance. *Lecture Notes on Data Engineering and Communications Technologies*. 2022, vol. 134, pp. 288–303. doi: 10.1007/978-3-031-04812-8_25
12. Pavlov O. A., Holovchenko M. M., Revych M. M. Metod otsinky koefitsiyentiv pry liniynykh chlenakh bahatovymirnoyi polinomial'noyi rehresiyi, zadanoyi nadlyshkovym opysom. [Method for estimating coefficients for linear terms of multidimensional polynomial regression given by redundant description] *Adaptyvni systemy avtomatychnoho upravlinnya: mizhvidomchyy nauk.-tekhn. zbirnyk* [Adaptive systems of automatic control: interdepartmental scientific and technical. collection]. Kyiv, NTUU "KPI" Publ., 2022, vol. 1, no. 40, pp. 110–117. doi: 10.20535/1560-8956.40.2022.261665

Received 26.09.2022

Відомості про авторів / About the Authors

Павлов Олександр Анатолійович – доктор технічних наук, професор каф. інформатики та програмної інженерії Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»; м. Київ, Україна; ORCID: https://orcid.org/0000-0002-6524-6410; e-mail: pavlov.fiot@gmail.com

Головченко Максим Миколайович – старший викладач каф. інформатики та програмної інженерії Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»; м. Київ, Україна; ORCID: https://orcid.org/0000-0002-9575-8046; e-mail: ma4ete25@ukr.net

Pavlov Alexander Anatolievich – Doctor of Technical Sciences, Full Professor of Informatics and Software Engineering Department of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"; Kyiv, Ukraine; ORCID: https://orcid.org/0000-0002-6524-6410; e-mail: pavlov.fiot@gmail.com

Holovchenko Maxim Nikolaevich – Senior Lecturer of Informatics and Software Engineering Department of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"; Kyiv, Ukraine; ORCID: https://orcid.org/0000-0002-6524-6410; e-mail: ma4ete25@ukr.net