UDC 519.6
DOI: 10.15587/2312-8372.2019.175020

Geche F.,
Mulesa O.,
Hrynenko V.,
Smolanka V.

SEARCH FOR IMPACT FACTOR CHARACTERISTICS IN CONSTRUCTION OF LINEAR REGRESSION MODELS

Об'єктом дослідження є задача побудови лінійної регресійної моделі, яка виникає в процесі вирішення проблеми прогнозування значень залежної змінної від сукупності незалежних факторних ознак. Ця задача часто виникає в процесі аналізу показників економічної діяльності підприємств. Процес побудови рівняння регресії, яке адекватно відображає залежність між факторними ознаками та досліджуваними результуючими ознаками, є багатоетапною і трудомісткою процедурою. Важливим при цьому є етап вибору найвпливовіших факторних ознак. Від ефективності проведення такого етапу та правильності вибору системи ознак залежить адекватність регресійної моделі та ефективність аналізу діяльності підприємств. В наукових джерелах пропонується ряд методів та алгоритмів для вибору найвпливовіших факторних ознак. Деякі з них базуються на кореляційно-регресійному аналізі, проте є ряд евристичних методів. В дослідженнях показано, що використання різних методів відбору найвпливовіших факторних ознак для розв'язання конкретних задач, в загальному випадку призводить до отримання різних результатів. При цьому особливістю більшості методів є їх обчислювальна складність або нестійкість щодо умов застосування. Основним критерієм ефективності алгоритмів вибору факторних ознак є адекватність побудованої регресійної моделі.

В дослідженні проведено аналіз процесу побудови множинних лінійний регресійних моделей. Визначено основні його етапи та наведено базові поняття і розрахункові формули. Авторами пропонується алгоритм вибору найвпливовіших факторних ознак при побудові лінійних регресійних моделей. Особливістю запропонованого підходу є те, що він базується на властивостях частинних коефіцієнтів кореляції. Застосування розробленого алгоритму дозволяє зменшувати обчислювальну складність процесу вибору факторних ознак в порівнянні з відомими алгоритмами.

Виконана експериментальна верифікація розробленого алгоритму для задачі побудови залежностей між різними показниками діяльності двох підприємств у вигляді множинної лінійної регресії. В результаті проведених обчислень з системи 17 факторних ознак для кожного досліджуваного показника було відібрано одну або дві впливові ознаки. Побудовані при цьому рівняння множинної лінійної регресії мали достовірність, яка перевищує 90 %.

Ключові слова: *множинна лінійна регресія, частинні коефіцієнти кореляції, факторні ознаки, адекватність моделі.*

Received date: 26.11.2018
Accepted date: 21.12.2018
Published date: 30.06.2019

1. Introduction

One of the most pressing problems of the functioning and development of enterprises is to identify ways to improve the efficiency of their activities. Promising research in this direction is the identification and analysis of factors affecting the resulting indicators of enterprises, from which the efficiency of the enterprise is determined.

Most of the resulting economic indicators of the enterprise are formed under the influence of many factors. The identification of the main of these factors and the establishment of relations between them allows to determine the main efforts necessary for their effective use and, as a result, to ensure the efficient operation of the enterprise.

This is also true for economic problems that can be described using the correlation model, which determines the correlation (regression) relationship between factor attributes and resulting indicators.

An important stage in the construction of the correlation model is the stage of selecting influential factor attributes among the set of measured indicators. The adequacy of the constructed regression model depends on the quality of making such a choice. In the general case, the process of determining the degree of influence of a sign on the resulting indicator is laborious and requires large computational costs. In this regard, the development of new effective algorithms for finding the most influential factor attributes is an urgent and practically important task.

2. The object of research and its technological audit

The object of research is the task of constructing a linear regression model that arises in the process of solving the problem of predicting the values of a dependent variable on a set of independent factor characteristics.

A feature of the problem of predicting the economic indicators of enterprises is that in the process of analyzing their activities, it is necessary to decide on the advisability of including many factors in the model. The solution to this problem encourages the use of the apparatus of correlation analysis. To apply the known methods

and algorithms for solving problems of determining the influence degree of a sign on the indicator under study, it is necessary to use the mathematical apparatus for determining the distribution laws of some model parameters. Such an approach leads to a significant increase in the computational complexity of the methods and algorithms for determining the influence of a sign with an increase in the number of factor factors studied. This indicates the possibility of potential problems with the duration of the search for a solution to the problem in a study of the economic performance of real enterprises.

This implies the need to develop new effective algorithms for determining the most influential factor characteristics for the resulting indicator, the use of which in practice will avoid the problems described above. Important in this case is the achievement of the adequacy of the constructed models and the prediction of the possibility of identifying a case in which not a single set of the considered features can provide the necessary level of adequacy.

3. The aim and objectives of research

The aim of research is investigation of the problem of constructing the multiple linear regression equation and to propose new approaches to finding the most influential factor attributes for constructing linear regression models. To achieve this aim, the following objectives are set:

1. To study the relationship between the resulting and factor attributes in the process of constructing linear regression models.

2. To develop an effective algorithm for the selection of factor attributes, the use of which can provide a high level of adequacy of the constructed regression models.

3. To perform experimental verification of the developed algorithm.

4. Research of existing solutions of the problem

The tasks of quantifying the interdependencies between economic indicators are investigated in many scientific sources. So, for example, [1, 2] are devoted to the study of the effectiveness of the use of correlation and regression analysis in the process of solving problems associated with establishing relationships between factor attributes of different types. In this case, the construction of the multiple linear regression equation [3] is gaining wide application; therefore, a large number of scientific papers are devoted to this issue. So, in [4, 5] the mathematical apparatus used in the process of constructing linear regression models is given. Here are the basic concepts, definitions and statements that are key in the correlation and regression analysis. In [6], the procedure for constructing a model of correlation analysis for studying multifactorial processes and phenomena is described. When constructing equations of multiple linear regression to assess the degree of connection between factors, this study proposes to use higher-order partial correlation coefficients, which makes the proposed approach difficult to apply. In [7], a genetic algorithm for the selection of factor attributes for constructing regression models is proposed. However, as shown in the study, such an approach in the general case allows one to achieve or even improve the results of the work of classical algorithms for selecting factor attributes. Studies have been devoted to the development of heuristic approaches to reducing the set of factor attributes [8, 9]. A feature of the proposed approaches is that when they are used, it becomes necessary to work with large-dimensional matrices; in the case of their poor conditioning, the calculation process is very complicated. Research [10] devoted to the use of the apparatus of the theory of choice and the theory of fuzzy sets when choosing the most influential features. Such an approach leads to the fact that the expert's opinion, on the basis of which the construction of fuzzy sets occurs, is the main when calculating the influence level of an attribute on the resulting indicator. Thus, with the involvement of various experts to solve the same problem, with high probability, different results will be obtained.

A study of scientific sources indicates that in the analysis of the results of economic activity of enterprises, correlation-regression analysis is widely used in terms of constructing equations of multiple linear regression. An important problem in this case is the choice of the most influential factor attributes that will be included in the model. Algorithms are known or have great computational complexity, or depend on the conditions of their application. Thus, the issue of developing new approaches to solving this problem remains promising.

5. Methods of research

The construction of econometric models using the mathematical apparatus of correlation analysis consists of the following steps:

1. Choice of an independent variable (factor argument).
2. Processing of statistical information.
3. Establishment of the degree of dependence between the resulting variables (attributes) and factor variables (indicators).
4. Determination of fuel-factoring features for the resulting indicator.
5. Construction of the regression equation for the most influential factor attributes and checking it for adequacy.
6. Analysis of the correlation model.

The selection of factor attributes (influential factors) is made on the basis of a logical analysis: from a set of technical, technological, organizational and socio-economic conditions for the functioning of enterprises.

When constructing linear regression models, as a rule, the following requirements apply to factor attributes:

– actor variables should have a quantitative measurement (gross profit, net profit, material costs, depreciation, etc.);
– factor attributes should be linearly independent;
– their values are determined according to the current and operational reporting data (quarterly, annual reports, dispatch documents, etc.).

After determining the dependent variables (resulting indicators) and factor attributes (independent variables) from their statistical data are built:

– corresponding variational (interval variational) series;
– numerical characteristics of the variation series are calculated;
– removal of certain values of the resulting and factor attributes that are «significantly» different from the bulk of the observations;
– sample is checked for representativeness.

ISSN 2226-3780

6. Research results

6.1. Analysis of dependencies between resulting and factor attributes. Mathematical models of many economic problems contain a whole group of factor attributes. If the number of factor attributes is equal to unity, then such models belong to the class of paired, and otherwise multidimensional correlation models. In the study of odd correlation models, the mathematical apparatus of multivariate correlation analysis is used.

Let's suppose that there is a multidimensional set of features, among which one resulting y and m factor features $x_1, x_2, ..., x_m$.

Let's suppose there is a sample of volume n, that is, there are n points:

$$(y_i, x_{1i}, x_{2i}, ..., x_{mi}),\ i = 1,2,...,n,$$

in m+1-dimensional vector space.

Between each pair of features, it is possible to set a sample pair correlation coefficient. For example, $\tilde{r}_{yx_j}$ – the sample pair coefficient between the resulting indicator y and the factor variable x_j $(j \in \{1,2,...,m\})$, $\tilde{r}_{x_i x_j}$ – the sample coefficient between the factor attributes x_i and x_j.

The relationship between the attributes can be set using the correlation matrix:

$$Q_{m+1} = Q(y, x_1, ..., x_m),$$

the elements of which are selective paired correlation coefficients:

$$Q_{m+1} = \begin{pmatrix} 1 & \tilde{r}_{yx_1} & ... & \tilde{r}_{yx_m} \\ \tilde{r}_{x_1 y} & 1 & ... & \tilde{r}_{x_1 x_m} \\ ... & ... & ... & ... \\ \tilde{r}_{x_m y} & \tilde{r}_{x_m x_1} & ... & 1 \end{pmatrix}. \quad (1)$$

In order to find the influence of only one factor characteristic x_j on the effective indicator y, it is necessary to fix the values of the attributes $x_1, ..., x_{j-1}, x_{j+1}, ..., x_m$ and get a sample of volume n for different values of the attribute x_j. It is clear that in this case the variation y is explained by the variable x_j. The correlation coefficient obtained on the basis of this sample is called the partial correlation coefficient and denoted as $\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}}$. Partial correlation coefficients can be found using the correlation matrix Q_{m+1} according to the following formula [5]:

$$\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}} = \frac{-A_{1j+1}}{\sqrt{A_{11} \cdot A_{j+1j+1}}}, \quad (2)$$

where A_{sq} – algebraic complement to the element a_{sq} of the correlation matrix located at the intersection of the s-th row and q-th column of the correlation matrix Q_{m+1}.

To assess the significance of the partial correlation coefficients $\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}}$ $(j = 1,2,...,m)$, the following parameter values t_j are found:

$$t_j = \frac{\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}} \sqrt{k}}{\sqrt{1 - \tilde{r}^2_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}}}},$$

where t_j – random variables distributed according to Student's law with a degree of freedom $k = n - m - 1$, and are compared with a critical value $t_{k,\alpha}$ at a significance level α.

If $|t_j| > t_{k,\alpha}$, then the partial coefficient $\tilde{r}_{yx_j \cdot \{x_1,...,x_m\} \setminus \{x_j\}}$ is significant.

The multiple correlation between the resulting indicator y and factor attributes $x_1, x_2, ..., x_m$ can be determined through the multiple correlation coefficient $R_{y \cdot x_1, x_2, ..., x_m}$ [4, 5], which is found by the following formula:

$$R_{y \cdot x_1, x_2, ..., x_m} = \sqrt{1 - \frac{|Q_{m+1}|}{A_{11}}}, \quad (3)$$

where $|Q_{m+1}|$ – determinant of the correlation matrix Q_{m+1}.

After the detection of influential factor characteristics, the equation of multiple linear regression is constructed:

$$\tilde{y}(x_1, x_2, ..., x_m) = m_y(x_1, x_2, ..., x_m) = \\ = a_0 + a_1 x_1 + ... + a_m x_m, \quad (4)$$

where $m_y(x_1, x_2, ..., x_m)$ – conditional expectation.

Let $y_i = \tilde{y}_i + \tilde{u}_i$ $(i = 1,2,...,n)$, then the coefficient of determination [5] is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{u}_i^2}{\sum_{i=1}^{n} (y_i - \tilde{m}_y)^2},$$

where $\tilde{m}_y$ – sample expectation y.

The adequacy of the linear regression model (4) is checked by the parameter:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}.$$

The calculated value of the parameter F is compared with the critical value $F_{cr}(m, n - m - 1)$ [5] and, according to the Fisher test, the linear regression model under study will be adequate if $F > F_{cr}(m, n - m - 1)$.

6.2. Algorithm for determining the most influential factor attributes in constructing a multiple linear regression model. To solve the problem of choosing the most influential factor attributes, the following algorithm is proposed.

Step 1. With respect to factor characteristics (independent variables) $x_1, x_2, ..., x_q$, based on the empirical table (sample), let's find sample correlation coefficients:

$$\tilde{r}_{yx_i} = \frac{\tilde{\text{cov}}(y, x_i)}{\tilde{\sigma}_y \cdot \tilde{\sigma}_{x_i}}\ (i = 1,2,...,q),$$

where y – resulting feature (dependent variable); $\tilde{\text{cov}}(y, x_i)$ – sample covariance of the studied indicators y and x_i; $\tilde{\sigma}_y$, $\tilde{\sigma}_{x_i}$ – sample standard deviations of the studied indicators.

With the help of t-Student statisticians, let's define a subset of factor attributes $\{x_{j_1}, x_{j_2}, ..., x_{j_k}\} \subset \{x_1, x_2, ..., x_n\}$ for which:

$$t_{j_s} = \frac{|\tilde{r}_{yx_{j_s}}| \cdot \sqrt{n-2}}{\sqrt{1 - r^2_{yx_{j_s}}}} > t_{n-2,\alpha}, \quad (5)$$

where n – the number of observations; $t_{n-2,\alpha}$ – the critical value of the parameter t_{j_s} at a significance level α with a degree of freedom $n - 2$.

The fulfillment of inequality (5) indicates the significance of the sample pair correlation coefficient $\tilde{r}_{yx_{j_s}}$.

Step 2. Based on the sample data for the factor features $x_{j_1}, x_{j_2}, ..., x_{j_k}$ and the resulting indicator y, let's construct a correlation matrix $Q_{k+1}(y, x_{j_1}, ..., x_{j_k})$:

$$Q_{k+1}(y, x_{j_1}, ..., x_{j_k}) = \begin{pmatrix} 1 & \tilde{r}_{yx_{j_1}} & ... & \tilde{r}_{yx_{j_k}} \\ \tilde{r}_{x_{j_1}y} & 1 & ... & \tilde{r}_{x_{j_1}x_{j_k}} \\ ... & ... & ... & ... \\ \tilde{r}_{x_{j_k}y} & \tilde{r}_{x_{j_k}x_{j_1}} & ... & 1 \end{pmatrix}$$

and find the partial selective correlation coefficients:

$$\tilde{r}_{yx_{j_s}.\{x_{j_1},...,x_{j_k}\}\setminus\{x_{j_s}\}} = \frac{-A_{1s+1}}{\sqrt{A_{11} \cdot A_{s+1s+1}}} \ (s = 1,2,...,k).$$

Using t-Student statistics, let's define a subset $\{z_1, z_2, ..., z_m\} \subset \{x_{j_1}, x_{j_2}, ..., x_{j_k}\}$ for which:

$$t_{yz_i.\{x_{j_1},...,x_{j_k}\}\setminus\{z_i\}} = \frac{\left|\tilde{r}_{yz_i.\{x_{j_1},...,x_{j_k}\}\setminus\{z_i\}}\right|\sqrt{n-k-1}}{\sqrt{1-\tilde{r}^2_{yz_i.\{x_{j_1},...,x_{j_k}\}\setminus\{z_i\}}}} > t_{n-k-1,\alpha}, \quad (6)$$

where $t_{n-k-1,\alpha}$ – critical value of the parameter $t_{yz_i.\{x_{j_1},...,x_{j_k}\}\setminus\{z_i\}}$ at the level of significance and degree of freedom $n-k-1$.

Step 3. Find the multiple coefficient of determination:

$$R^2_{y.z_1,...,z_m} = 1 - \frac{|Q_{m+1}(y, z_1, ..., z_m)|}{A_{11}},$$

where $|Q_{m+1}(y, z_1, ..., z_m)|$ – matrix determinant $Q_{m+1}(y, z_1, ..., z_m)$.

If the coefficient value $R^2_{y.z_1,...,z_m}$ shows that the variation y is sufficiently explained by the variations of factor attributes $z_1, ..., z_m$, then go to step 4. Otherwise, it is possible to conclude that there are factor variables that significantly affect the resulting attribute y and are not taken into account. Therefore, when constructing a multiple linear regression model, it is necessary to conduct additional research to identify new significant factor attributes $x_{q+1}, x_{q+2}, ...$, and can be included in the linear regression model. Or increase the number of observations to clarify the relationship between the resulting attribute y and factor attributes $z_1, ..., z_m$, if possible.

Step 4. Concerning factor attributes $z_1, z_2, ..., z_m$, let's construct the equation of multiple linear regression:

$$\tilde{y} = a_0 + a_1 z_1 + ... + a_m z_m. \quad (7)$$

Step 5. Based on the sample values of factor attributes $z_1, z_2, ..., z_m$, let's construct a matrix:

$$Z(y, z_1, ..., z_m) = \begin{pmatrix} 1 & z_{11} & ... & z_{1m} \\ 1 & z_{21} & ... & z_{2m} \\ ... & ... & ... & ... \\ 1 & z_{n1} & ... & z_{nm} \end{pmatrix},$$

where $i+1$ – column of the matrix consists of the corresponding values of the factor variable z_i (z_{ji} – values of the factor variable z_i during its j-th observation).

Let's estimate the statistical significance of the coefficients ($i = 0,1,...,m$) of multiple linear regression (7).

To do this, let's find the estimated covariance matrix [5]:

$$\tilde{\sigma}^2_u (Z' \cdot Z)^{-1} = \begin{pmatrix} \tilde{\sigma}^2_{a_0} & & * & * \\ & \tilde{\sigma}^2_{a_1} & & * \\ * & & \ddots & \\ * & * & & \tilde{\sigma}^2_{a_m} \end{pmatrix}, \quad (8)$$

on the diagonal of which are estimates $\tilde{\sigma}^2_{a_0}, \tilde{\sigma}^2_{a_1}, ..., \tilde{\sigma}^2_{a_m}$ of the variances of the parameters:

$$\sigma_u = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-m-1}};$$

Z' – transposed matrix matrix.

Calculating the square roots of these estimates $\tilde{\sigma}^2_{a_0}$, $\tilde{\sigma}^2_{a_1}, ..., \tilde{\sigma}^2_{a_m}$, let's find the standard errors $\tilde{\sigma}_{a_0}, \tilde{\sigma}_{a_1}, ..., \tilde{\sigma}_{a_m}$ of the coefficients $a_0, a_1, \ldots, a_m$ and through them let's respectively determine the parameters $t_{a_0}, t_{a_1}, ..., t_{a_m}$:

$$t_{a_0} = \frac{|a_0|}{\tilde{\sigma}_{a_0}}, t_{a_1} = \frac{|a_1|}{\tilde{\sigma}_{a_1}}, ..., t_{a_m} = \frac{|a_m|}{\tilde{\sigma}_{a_m}}.$$

With the found parameters t_{a_i} ($i = 1,2,...,m$), let's use t-Student statistics. If $t_{a_i} > t_{n-m-1,\alpha}$, then the coefficient a_i is significant and leave it in equation (7), otherwise, let's believe that the coefficient a_i is insignificant and will remove it from equation (7). After such transformations, equations (7) can be written as follows:

– if $t_{a_0} > t_{n-m-1,\alpha}$ then:

$$\tilde{y} = a_0 + a_{j_1} z_{j_1} + ... + a_{j_h} z_{j_h}, \quad (9)$$

– or otherwise:

$$\tilde{y} = a_{j_1} z_{j_1} + ... + a_{j_h} z_{j_h}, \quad (10)$$

where $\{z_{j_1}, ..., z_{j_h}\} \subset \{z_1, ..., z_m\}$ and go to step 6.

Step 6. With respect to factor attributes $z_{j_1}, z_{j_2}, ..., z_{j_h}$, let's find the multiple coefficient of determination $R^2 = R^2_{y.z_{j_1},z_{j_2},...,z_{j_h}}$:

$$R^2_{y.z_{j_1},z_{j_2},...,z_{j_h}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \tilde{m}_y)^2}, \quad (11)$$

where $\tilde{m}_y$ – sample expectation y.

Let's check equations (9) or (10) for significance in general. To do this, let's find the value:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-h-1}{h}, \quad (12)$$

and compare it with the critical value $F_{cr}(h, n-s-1, \alpha)$.

According to Fisher's F statistics, equations (9) or (10) will be significant with a significance level if $F > F_{cr}(h, n-h-1, \alpha)$ and algorithm is complete.

Note. If no factor attribute satisfies conditions (5) and (6), then let's conclude that factor attributes are not significant and for these factor attributes the equations of multiple linear regression are not made.

ISSN 2226-3780

6.3. Analysis of the results of the algorithm for selecting the most influential factor attributes. To conduct experimental verification of the developed algorithm for the selection of the most influential factor attributes, the problem of constructing multiple linear regression based on the performance of enterprises was considered. The data for the problem were obtained from [11]. The activities of two enterprises during 2003–2012 were reviewed. The system of attributes included 17 elements, among which were inventories, cash and cash equivalents, accounts payable for goods, work and services, and others.

The aim of research was to determine the effect of these attributes on such resulting indicators as:

– $y1$ – capital productivity of fixed assets;
– $y2$ – profitability of fixed assets;
– $y3$ – return on equity for fixed values of the parameters ST;
– ST – significance level of factor attributes according to student criterion;
– F – adequacy level of the multiple linear regression equation according to the Fisher criterion;
– PM – «multicollinearity threshold».

The results of the algorithm at (ST=0.1; F=0.1; PM=0.8) are given in Tables 1, 2.

Table 1

Multiple linear regression models for resulting indicators $y1$, $y2$, $y3$ in the analysis of the first enterprise

Resulting indicators	Multiple linear regression equations	R^2
$y1$	$y1=0.807086+0.000191x13-0.000123x6$	0.936022
$y2$	$y2=-0.024015+0.000125x11+0.000656x2$	0.934932
$y3$	$y3=+4.7e-0.5x11+0.000253x2$	0.90165

Table 2

Multiple linear regression models for resulting indicators $y1$, $y2$, $y3$ in the analysis of the second enterprise

Resulting indicators	Multiple linear regression equations	R^2
$y1$	$y1=12.329052-0.005155x6$	0.939498
$y2$	$y2=-0.005758+0.000411x11$	0.977696
$y3$	$y3=-0.00275+0.000258x17$	0.992574

As can be seen from the Tables 1 and 2, all the constructed equations of multiple linear regression of the resulting indicators $Y1$, $Y2$, $Y3$ are adequate and have a high coefficient of determination for the studied enterprises.

If to study the relationship between $y2$ and the factor attributes $x1$, $x2$,..., $x17$, then it is possible to see that for each enterprise under study the reliability of the constructed equations is high and in percent they respectively amount to:

– 93.4932 % (first enterprise);
– 97.7696 % (second enterprise).

For comparison, $Y2$ (primary data) and $y2$ (data obtained by the regression equation) at each enterprise let's use their graphic image (Fig. 1, 2).

The obtained numerical results for model problems indicate the effectiveness of the application of the developed algorithm at the stage of determining the influential factor characteristics.

Fig. 1. Comparison of indicators of profitability of fixed assets ($Y2$) and ($y2$) for the first enterprise

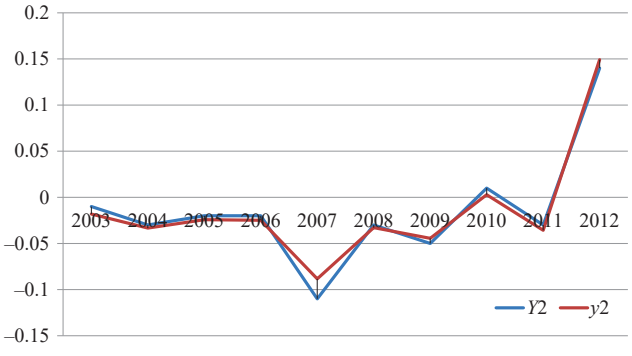

Fig. 2. Comparison of indicators of profitability of fixed assets ($Y2$) and ($y2$) for the second enterprise

7. SWOT analysis of research results

Strengths. The developed algorithm for choosing the most influential factor attributes in the task of constructing equations of multiple linear regression is based on the properties of particular correlation coefficients and provides a qualitative choice of factor attributes. This approach allows to reduce the computational complexity of the process of selecting factor features in comparison with known algorithms.

Weaknesses. The proposed approach to the selection of the most influential factor attributes allows multicollinearity between factor attributes in the constructed model with a given threshold.

Opportunities. The constructed multiple linear regression models can be successfully used to solve problems associated with forecasting the values of economic indicators of enterprises.

Threats. When solving real problems of choosing the most influential factor attributes, there may be cases when a given system of attributes can't provide an adequate linear regression model. The developed algorithm detects such cases, but does not answer the question about the methodology for expanding the system of factor attributes in order to ensure the possibility of constructing an adequate linear regression model.

8. Conclusions

1. The process of analyzing the dependencies between the resulting and factor attributes used in the construction of linear regression models is investigated. Working calculation formulas are given for calculating the corresponding

parameters, which determine the significance of factor attributes included in the linear regression model.

2. A new effective algorithm for the selection of factor attributes based on the properties of particular correlation coefficients is developed. The application of the developed algorithm allows to reduce the computational complexity of the process of selecting factor attributes in comparison with known algorithms.

3. The results of the developed algorithm are demonstrated using model examples. It is shown how only influential features are selected from the set of factor factors proposed, which ensured the reliability of the constructed equation of multiple linear regression at a level exceeding 90 %.

References

1. Smeekes, S., Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting, 34 (3),* 408–430. doi: https://doi.org/10.1016/j.ijforecast.2018.01.001
2. Alvarez-Diaz, M., Alvarez, A. (2010). Forecasting exchange rates using local regression. *Applied Economics Letters, 17 (5),* 509–514. doi: https://doi.org/10.1080/13504850801987217
3. Cleland, A. C., Earle, M. D., Boag, I. F. (2007). Application of multiple linear regression to analysis of data from factory energy surveys. *International Journal of Food Science & Technology, 16 (5),* 481–492. doi: https://doi.org/10.1111/j.1365-2621.1981.tb01841.x
4. Heche, F. E. (2019). *Teoriya ymovirnostei i matematychna statystyka*. Uzhhorod: AUTDOR-ShARK, 235.
5. Baltagi, B. (2008). *Econometric analysis of panel data*. John Wiley & Sons, 388.
6. Shojima, K., Usami, S., Hashimoto, T., Todo, N., Takano, K. (2018). Understanding Differences in Statistical Models. *The Annual Report of Educational Psychology in Japan, 57,* 302–308. doi: https://doi.org/10.5926/arepj.57.302
7. Depczynski, U., Frost, V. J., Molt, K. (2000). Genetic algorithms applied to the selection of factors in principal component regression. *Analytica Chimica Acta, 420 (2),* 217–227. doi: https://doi.org/10.1016/s0003-2670(00)00893-x
8. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58 (1),* 267–288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
9. Mulesa, O. (2016). Development of evolutionary methods of the structural and parametric identification for tabular dependencies. *Technology audit and production reserves, 4 (2 (30)),* 13–19. doi: https://doi.org/10.15587/2312-8372.2016.74482
10. Azadeh, A., Ziaei, B., Moghaddam, M. (2012). A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. *Expert Systems with Applications, 39 (1),* 298–315. doi: https://doi.org/10.1016/j.eswa.2011.07.020
11. *Ahentstvo z rozvytku infrastruktury fondovoho rynku Ukrainy*. Available at: https://smida.gov.ua/

Geche Fedir, *Doctor of Technical Sciences, Professor, Head of Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Ukraine, e-mail: fgeche@hotmail.com, ORCID: http://orcid.org/0000-0002-4757-9828*

Mulesa Oksana, *PhD, Associate Professor, Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Ukraine, e-mail: mulesa.oksana@gmail.com, ORCID: http://orcid.org/0000-0002-6117-5846*

Hrynenko Viktor, *Postgraduate Student, Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Ukraine, ORCID: http://orcid.org/0000-0002-4606-0792*

Smolanka Veronika, *Postgraduate Student, Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Ukraine, e-mail: veronika.smolanka@uzhnu.edu.ua, ORCID: http://orcid.org/0000-0002-8380-1967*