UDC 004.412:519.237.5

# THE NON-LINEAR REGRESSION MODEL TO ESTIMATE THE SOFTWARE SIZE OF OPEN SOURCE JAVA-BASED SYSTEMS

**Prykhodko N. V.** – PhD, Associate Professor, Associate Professor of the Finance Department, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine.

**Prykhodko S. B.** – Dr. Sc., Professor, Head of the Department of Software of Automated Systems, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine.

## ABSTRACT

**Context.** The problem of estimating the software size in the early stage of a software project is important, since the information obtained from estimating the software size is used for predicting the software development effort, including open-source Java-based information systems. The object of the study is the process of estimating the software size of open-source Java-based information systems. The subject of the study is the regression models for estimating the software size of open-source Java-based information systems.

**Objective.** The goal of the work is the creation of the non-linear regression model for estimating the software size of open-source Java-based information systems on the basis of the Johnson multivariate normalizing transformation.

**Method.** The model, confidence and prediction intervals of multiply non-linear regression for estimating the software size of open-source Java-based information systems are constructed on the basis of the Johnson multivariate normalizing transformation for non-Gaussian data with the help of appropriate techniques. The techniques to build the models, equations, confidence and prediction intervals of non-linear regressions are based on the multiple non-linear regression analysis using the multivariate normalizing transformations. The appropriate techniques are considered. The techniques allow to take into account the correlation between random variables in the case of normalization of multivariate non-Gaussian data. In general, this leads to a reduction of the mean magnitude of relative error, the widths of the confidence and prediction intervals in comparison with the linear models or nonlinear models constructed using univariate normalizing transformations.

**Results.** Comparison of the constructed model with the linear model and non-linear regression models based on the decimal logarithm and the Johnson univariate transformation has been performed.

**Conclusions.** The non-linear regression model to estimate the software size of open-source Java-based information systems is constructed on the basis of the Johnson multivariate transformation for $S_B$ family. This model, in comparison with other regression models (both linear and non-linear), has a larger multiple coefficient of determination, a larger value of percentage of prediction and a smaller value of the mean magnitude of relative error. The prospects for further research may include the application of other multivariate normalizing transformations and data sets to construct the non-linear regression model for estimating the software size of open-source Java-based information systems.

**KEYWORDS:** software size estimation, Java-based information system, non-linear regression model, univariate normalizing transformation, non-Gaussian data.

## ABBREVIATIONS

HTML is hypertext markup language;
JSP is Java server pages;
KLOC is the thousand lines of code;
LB is lower bound;
MD is Mahalanobis distance;
MMR is a magnitude of relative error;
MMRE is a mean magnitude of relative error;
PHP is hypertext preprocessor;
PRED is percentage of prediction;
SQL is structured query language;
UB is upper bound;
VBA is visual Basic for application.

## NOMENCLATURE

$\hat{\mathbf{b}}$ is estimator for vector of linear regression equation parameters, $\mathbf{b} = \{b_1, b_2, \ldots, b_k\}^T$;

$\hat{b}_i$ is estimator for the $i$-th parameter of linear regression equation;

$k$ is a number of independent variables (regressors);

$N$ is a number of data points;

$N(0,1)$ is a Gaussian distribution with zero mathematical expectation and unit variance;

$\mathbf{P}$ is a non-Gaussian random vector, $\mathbf{P} = \{Y, X_1, X_2, \ldots, X_k\}^T$;

$R^2$ is a multiple coefficient of determination;

$\mathbf{S}_N$ is a sample covariance matrix, $\mathbf{S}_N = [S_{ij}]$;

$\mathbf{T}$ is a Gaussian random vector, $\mathbf{T} = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$;

$t_{\alpha/2,\nu}$ is a quantile of student's $t$-distribution with $\nu$ degrees of freedom and $\alpha/2$ significance level;

$X_1$ is a total number of classes;

$X_2$ is a total number of relationships;

$X_3$ is an average number of attributes per class,

$Y$ is an actual software size in KLOC;

$\mathbf{Z}_X^+$ is a matrix of centered regressors that contains the values $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$;

$\mathbf{z}_X$ is a vector with components $Z_i$;


DOI 10.15588/1607-3274-2018-3-17

$(\mathbf{z}_X)^T$ is a transpose of $\mathbf{z}_X$;

$\overline{Z}_Y$ is a sample mean of the values of the variable $Z_Y$;

$\hat{Z}_Y$ is a prediction linear regression equation result;

$\alpha$ is a significance level;

$\beta_2$ is a multivariate kurtosis;

$\boldsymbol{\gamma}$ is a vector of parameters of the Johnson multivariate translation, $\boldsymbol{\gamma} = (\gamma_Y, \gamma_1, \gamma_2, \ldots, \gamma_k)^T$;

$\varepsilon$ is a Gaussian random variable which defines residuals, $\varepsilon \sim N(0,1)$;

$\boldsymbol{\eta}$ is a vector of parameters of the Johnson multivariate translation, $\boldsymbol{\eta} = diag(\eta_Y, \eta_1, \ldots, \eta_k)$;

$\boldsymbol{\lambda}$ is a vector of parameters of the Johnson multivariate translation, $\boldsymbol{\lambda} = diag(\lambda_Y, \lambda_1, \ldots, \lambda_k)$;

$\nu$ is a number of degrees of freedom;

$\boldsymbol{\Sigma}$ is a covariance matrix, $\boldsymbol{\Sigma} = [\Sigma_{ij}]$;

$\varphi$ is a vector of parameters of the Johnson multivariate translation, $\varphi = (\varphi_Y, \varphi_1, \varphi_2, \ldots, \varphi_k)^T$;

$\boldsymbol{\psi}$ is a vector of multivariate normalizing transformation, $\boldsymbol{\psi} = \{\psi_Y, \psi_1, \psi_2, \ldots, \psi_k\}^T$.

## INTRODUCTION

Java is a programming language and computing platform first released by Sun Microsystems in 1995 (https://www.java.com). Now Java is used practically everywhere from laptops to datacenters, game consoles to supercomputers, cell phones to the Internet, including information systems. Software size is one of the most important internal metrics of software including software of open-source Java-based information systems.

The information obtained from estimating the software size are useful for predicting the software development effort by such well-known model as COCOMO II. This leads to the need to develop appropriate models to estimate the software size [1–4].

The paper [2] proposed the linear regression equations for estimating the software size of some programming languages including Java. The proposed equation is constructed by multiple linear regression analysis on the basis of the metrics that can be measured from conceptual data model based a class diagram. However, there are four basic assumptions that justify the use of linear regression models, one of which is normality of the error distribution. But this assumption is valid only in particular cases. This leads to the need to use the non-linear regression models including for estimating the software size of Java-based open-source information systems.

**The object of study** is the process of estimating the software size of open-source Java-based information systems.

**The subject of study** is the non-linear regression models to estimate the software size of open-source Java-based information systems.

The known regression equation for estimating the software size of open-source Java-based information systems [2] is linear and generally have large widths of confidence and prediction intervals.

**The purpose of the work** is to construct the non-linear regression model for estimating the software size of open-source Java-based information systems. The software size prediction results by constructed model should be better in comparison with other regression models, both linear and nonlinear, primarily on such standard evaluations as the multiple coefficient of determination and mean magnitude of relative error.

## 1 PROBLEM STATEMENT

Suppose given the original sample as the four-dimensional non-Gaussian data set: actual software size in the thousand lines of code (KLOC) $Y$, the total number of classes $X_1$, the total number of relationships $X_2$ and the average number of attributes per class $X_3$ in conceptual data model from $N$ information systems developed using the Java programming language with JSP, HTML and SQL. Suppose that there are bijective multivariate normalizing transformation of non-Gaussian random vector $\mathbf{P} = \{Y, X_1, X_2, \ldots, X_k\}^T$ to Gaussian random vector $\mathbf{T} = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$ is given by

$$\mathbf{T} = \boldsymbol{\psi}(\mathbf{P}) \quad (1)$$

and the inverse transformation for (1)

$$\mathbf{P} = \boldsymbol{\psi}^{-1}(\mathbf{T}). \quad (2)$$

It is required to build the non-linear regression model in the form $Y = Y(X_1, X_2, X_3, \varepsilon)$ on the basis of the transformations (1) and (2).

## 2 REVIEW OF THE LITERATURE

In paper [2] the linear regression equation for estimating the software size of open-source Java-based information systems was proposed in the form

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3, \quad (3)$$

where $\hat{b}_0 = -10.121$, $\hat{b}_1 = 1.201$, $\hat{b}_2 = 1.439$ and $\hat{b}_3 = 0.726$.

A normalizing transformation is often a good way to build the models, equations, confidence and prediction intervals of non-linear regressions [5–8]. According to [7] transformations are made for essentially four purposes, two of which are: firstly, to obtain approximate normality for the distribution of the error term (residuals) or the


DOI 10.15588/1607-3274-2018-3-17

dependent random variable, secondly, to transform the response and/or the predictor in such a way that the strength of the linear relationship between new variables (normalized variables) is better than the linear relationship between dependent and independent random variables.

Well-known techniques for building the equations, confidence and prediction intervals of multivariate non-linear regressions are based on the univariate normalizing transformations (such as, the decimal logarithm, Box-Cox transformation), which do not take into account the correlation between random variables in the case of normalization of multivariate non-Gaussian data. Application of such univariate normalizing transformations for building the non-linear regression models does not always lead to good normality and linear relationship between normalized variables. This leads to the need to use the multivariate normalizing transformations.

In [9] the techniques to build the equations, confidence and prediction intervals of non-linear regressions for multivariate non-Gaussian data on the basis of the bijective multivariate normalizing transformations were proposed. The techniques consist of three steps. In the first step, a set of multivariate non-Gaussian data is normalized using a bijective multivariate normalizing transformation. In the second step, the equation, confidence and prediction intervals of linear regression for the normalized data are built. In the third step, the equations, confidence and prediction intervals of non-linear regressions for multivariate non-Gaussian data are constructed on the basis of the equation, confidence and prediction intervals of linear regression for the normalized data and the multivariate normalizing transformation. Note there is no the error term in non-linear regression equation. The absence of the error term in non-linear regression equation does not allow modeling the random dependent variable for its prediction. This leads to the need to develop the non-linear regression model for estimating the software size of open-source Java-based information systems.

## 3 MATERIALS AND METHODS

After normalizing the non-Gaussian data by the transformation (1) the linear regression model is built for normalized data. The linear regression model for normalized data according to (1) will have the form

$$Z_Y = \hat{Z}_Y + \varepsilon = \overline{Z}_Y + \left(\mathbf{Z}_X^{+}\right)\hat{\mathbf{b}} + \varepsilon . \quad (4)$$

After that the non-linear regression model is built on the basis of the linear regression model (4) for the normalized data and the transformations (1) and (2). The non-linear regression model will have the form

$$Y = \psi_Y^{-1}\left[\overline{Z}_Y + \left(\mathbf{Z}_X^{+}\right)\hat{\mathbf{b}} + \varepsilon\right] . \quad (5)$$

The technique to build a confidence interval of non-linear regression is based on transformations (1) and (2), and a confidence interval of linear regression for normalized data

$$\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{\frac{1}{N} + \left(\mathbf{z}_X^{+}\right)^T \left[\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+}\right]^{-1} \left(\mathbf{z}_X^{+}\right)\right\}^{1/2} , \quad (6)$$

where $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$, $\nu = N - k - 1$; $\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+}$ is the $k \times k$ matrix

$$\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+} = \begin{pmatrix} S_{Z_1 Z_1} & S_{Z_1 Z_2} & \dots & S_{Z_1 Z_k} \\ S_{Z_1 Z_2} & S_{Z_2 Z_2} & \dots & S_{Z_2 Z_k} \\ \dots & \dots & \dots & \dots \\ S_{Z_1 Z_k} & S_{Z_2 Z_k} & \dots & S_{Z_k Z_k} \end{pmatrix},$$

where $S_{Z_q Z_r} = \sum_{i=1}^{N}\left[Z_{q_i} - \overline{Z}_q\right]\left[Z_{r_i} - \overline{Z}_r\right]$, $q, r = 1,2,\dots,k$ .

The confidence interval for non-linear regression is built on the basis of the interval (6) and inverse transformation (2)

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{\frac{1}{N} + \left(\mathbf{z}_X^{+}\right)^T \left[\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+}\right]^{-1} \left(\mathbf{z}_X^{+}\right)\right\}^{1/2}\right) .$$

The technique to build a prediction interval is based on multivariate transformation (1), the inverse transformation (2) and a prediction interval for normalized data

$$\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{1 + \frac{1}{N} + \left(\mathbf{z}_X^{+}\right)^T \left[\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+}\right]^{-1} \left(\mathbf{z}_X^{+}\right)\right\}^{1/} \quad (7)$$

The prediction interval for non-linear regression is built on the basis of the interval (7) and inverse transformation (2)

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{1 + \frac{1}{N} + \left(\mathbf{z}_X^{+}\right)^T \left[\left(\mathbf{Z}_X^{+}\right)^T \mathbf{Z}_X^{+}\right]^{-1} \left(\mathbf{z}_X^{+}\right)\right\}^{1/2}\right).$$

For normalizing the multivariate non-Gaussian data, we use the Johnson translation system. In our case the Johnson normalizing translation is given by [10]

$$\mathbf{T} = \boldsymbol{\gamma} + \boldsymbol{\eta}\mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{P} - \boldsymbol{\varphi})\right] \sim N_m(\mathbf{0}_m, \boldsymbol{\Sigma}), \quad (8)$$


DOI 10.15588/1607-3274-2018-3-17

where $\mathbf{h}[(y_Y, y_1, \ldots, y_k)] = \{h_Y(y_Y), h_1(y_1), \ldots, h_k(y_k)\}^T$; $h_i(\cdot)$ is one of the translation functions

$$h = \begin{cases} \ln(y), & \text{for } S_L \text{ (log normal) family;} \\ \ln[y/(1-y)], & \text{for } S_B \text{ (bounded) family;} \\ \mathrm{Arsh}(y), & \text{for } S_U \text{ (unbounded) family;} \\ y & \text{for } S_N \text{ (normal) family.} \end{cases} \quad (9)$$

Here $y = (X - \varphi)/\lambda$; $\mathrm{Arsh}(y) = \ln\left(y + \sqrt{y^2 + 1}\right)$. In our case $X$ equals $Y$, $X_1$, $X_2$ or $X_3$ respectively.

The equation, confidence and prediction intervals of non-linear regression to estimate the software size of open-source Java-based systems are constructed on the basis of the Johnson multivariate normalizing transformation for the four-dimensional non-Gaussian data set: actual software size in the thousand lines of code (KLOC) $Y$, the total number of classes $X_1$, the total number of relationships $X_2$ and the average number of attributes per class $X_3$ in conceptual data model from 30 information systems developed using the Java programming language with JSP, HTML and SQL. Table 1 contains the data from [2] on four metrics of software for 30 open-source Java-based systems.

Table 1 – The data set and squared MDs

| No | $Y$ | $X_1$ | $X_2$ | $X_3$ | Squared MD | |
|---|---|---|---|---|---|---|
| | | | | | univariate | multivariate |
| 1 | 11.717 | 8 | 6 | 4.25 | 0.99 | 0.99 |
| 2 | 47.52 | 23 | 19 | 9.565 | 1.65 | 1.50 |
| 3 | 84.01 | 26 | 40 | 11.462 | 10.13 | 6.55 |
| 4 | 26.999 | 15 | 14 | 8.933 | 1.33 | 1.89 |
| 5 | 41.72 | 20 | 15 | 5.9 | 0.25 | 0.50 |
| 6 | 13.015 | 5 | 6 | 12.4 | 5.89 | 6.64 |
| 7 | 30.402 | 18 | 7 | 6.611 | 1.59 | 3.10 |
| 8 | 29.159 | 23 | 10 | 6.957 | 2.26 | 3.13 |
| 9 | 53.443 | 28 | 25 | 4.179 | 3.53 | 4.02 |
| 10 | 18.694 | 13 | 9 | 6.615 | 0.39 | 0.89 |
| 11 | 26.384 | 16 | 6 | 5.125 | 2.23 | 4.06 |
| 12 | 38.721 | 19 | 16 | 6.579 | 0.19 | 0.19 |
| 13 | 75.643 | 26 | 30 | 6.154 | 1.87 | 2.74 |
| 14 | 46.72 | 21 | 24 | 6.048 | 1.31 | 1.66 |
| 15 | 6.413 | 7 | 5 | 4.143 | 2.41 | 6.07 |
| 16 | 79.534 | 20 | 37 | 4.85 | 7.06 | 8.20 |
| 17 | 36.343 | 18 | 17 | 5.333 | 0.35 | 0.47 |
| 18 | 59.684 | 22 | 31 | 6.182 | 2.49 | 2.62 |
| 19 | 50.454 | 15 | 20 | 11.6 | 2.51 | 3.35 |
| 20 | 3.055 | 4 | 1 | 7 | 10.83 | 7.10 |
| 21 | 63.257 | 34 | 17 | 3.971 | 9.16 | 8.29 |
| 22 | 91.28 | 35 | 28 | 13.571 | **17.73** | 11.22 |
| 23 | 32.707 | 11 | 17 | 7.545 | 0.98 | 1.54 |
| 24 | 11 | 5 | 5 | 3.6 | 6.15 | 6.36 |
| 25 | 5.543 | 6 | 4 | 3.833 | 2.54 | 5.16 |
| 26 | 22.686 | 12 | 11 | 6.667 | 0.11 | 0.22 |
| 27 | 3.911 | 3 | 2 | 6.667 | 7.26 | 5.52 |
| 28 | 20.841 | 14 | 7 | 3 | 8.17 | 7.21 |
| 29 | 9.269 | 6 | 5 | 3.5 | 3.23 | 2.82 |
| 30 | 7.732 | 7 | 2 | 11.143 | 5.42 | 5.98 |

For detecting the outliers in the data from Table 1 we use the technique based on multivariate normalizing transformations and the squared Mahalanobis distance (MD) [11]. There are no outliers in the data from Table 1 for 0.005 significance level and the Johnson multivariate transformation (8) for $S_B$ family. The same result was obtained in [12] for the transformation (8) for $S_U$ family. In [2] it was also assumed that the data contains no outliers. The values of squared MD for normalized data by the Johnson univariate transformation (9) for $S_B$ family from Table 1 indicate the data of system 22 is multivariate outlier, since for this data row the squared MD equals to 17.73 is greater than the value of the quantile of the Chi-Square distribution, which equals to 14.86 for 0.005 significance level. Although note that without using normalization, the data of system 11 is multivariate outlier, since for this data row the squared MD equals to 15.44.

Parameters of the multivariate transformation (9) for $S_B$ family were estimated by the maximum likelihood method. Estimators for parameters of the transformation (9) are: $\hat{\gamma}_Y = 9.63091$, $\hat{\gamma}_1 = 15.5355$, $\hat{\gamma}_2 = 25.4294$, $\hat{\gamma}_3 = 0.72801$, $\hat{\eta}_Y = 1.05243$, $\hat{\eta}_1 = 1.58306$, $\hat{\eta}_2 = 2.54714$, $\hat{\eta}_3 = 0.54312$, $\hat{\varphi}_Y = -1.4568$, $\hat{\varphi}_1 = -1.8884$, $\hat{\varphi}_2 = -6.9746$, $\hat{\varphi}_3 = 3.2925$, $\hat{\lambda}_Y = 153102.605$, $\hat{\lambda}_1 = 243051.0$, $\hat{\lambda}_2 = 311229.5$ and $\hat{\lambda}_3 = 13.90$. The sample covariance matrix $\mathbf{S}_N$ of the $\mathbf{T}$ is used as the approximate moment-matching estimator of $\mathbf{\Sigma}$

$$\mathbf{S}_N = \begin{pmatrix} 1.0000 & 0.9514 & 0.9333 & 0.1574 \\ 0.9514 & 1.0000 & 0.9006 & 0.1345 \\ 0.9333 & 0.9006 & 1.0000 & 0.0554 \\ 0.1574 & 0.1345 & 0.0554 & 1.0000 \end{pmatrix}.$$

After normalizing the non-Gaussian data by the multivariate transformation (9) for $S_B$ family the linear regression model (3) is built for normalized data

$$Z_Y = \hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon. \quad (10)$$

Parameters of the linear regression model (10) were estimated by the least square method. Estimators for parameters of the equation (10) are such: $\hat{b}_0 = 1.02 \cdot 10^{-5}$, $\hat{b}_1 = 0.56085$, $\hat{b}_2 = 0.42491$, $\hat{b}_3 = 0.05846$.

After that the non-linear regression model (4) is built

$$Y = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y\right)/\hat{\eta}_Y}\right]^{-1}. \quad (11)$$

where $Z_j = \gamma_j + \eta_j \ln \dfrac{X_j - \varphi_j}{\varphi_j + \lambda_j - X_j}$, $\varphi_j < X_j < \varphi_j + \lambda_j$, $j = 1,2,3$.


DOI 10.15588/1607-3274-2018-3-17

The model (11) is the non-linear regression model to estimate the software size of open-source Java-based information systems.

## 4 EXPERIMENTS

For comparison of the model (11) with other models two non-linear regression models are built on the basis of the data from Table 1 and two univariate normalizing transformations: the decimal logarithm transformation and the Johnson transformation.

The non-linear regression model is constructed on the basis of the linear regression model (4) for the normalized data and the decimal logarithm transformation

$$Y = 10^{\varepsilon + \hat{b}_0} X_1^{\hat{b}_1} X_2^{\hat{b}_2} X_3^{\hat{b}_3} . \qquad (12)$$

where the estimators for parameters of the model (12) are: $\hat{b}_0 = -0.04536$, $\hat{b}_1 = 0.64235$, $\hat{b}_2 = 0.56305$ and $\hat{b}_3 = 0.18045$.

The non-linear regression model is constructed on the basis of the linear regression model (4) for the normalized data and the Johnson univariate transformation for $S_B$ family. In this case the estimators for parameters of the model (11) are: $\hat{\gamma}_Y = 0.46387$, $\hat{\gamma}_1 = 0.38093$, $\hat{\gamma}_2 = 0.60545$, $\hat{\gamma}_3 = 0.65592$, $\hat{\eta}_Y = 0.50326$, $\hat{\eta}_1 = 0.62689$, $\hat{\eta}_2 = 0.62215$, $\hat{\eta}_3 = 0.72789$, $\hat{\varphi}_Y = 2.817$, $\hat{\varphi}_1 = 2.634$, $\hat{\varphi}_2 = 0.700$, $\hat{\varphi}_3 = 2.839$, $\hat{\lambda}_Y = 89.930$, $\hat{\lambda}_1 = 33.711$, $\hat{\lambda}_2 = 41.428$, $\hat{\lambda}_3 = 11.780$, $\hat{b}_0 = 0$, $\hat{b}_1 = 0.46976$, $\hat{b}_2 = 0.53539$ and $\hat{b}_3 = 0.11397$.

The computer program implementing the constructed models (11) and (12) was developed to conduct experiments. The program was written in the sci-language for the Scilab system. Scilab (http://www.scilab.org) is the free and open source software, the alternative to commercial packages for system modeling and simulation packages such as MATLAB and MATRIXx.

## 5 RESULTS

If the Gaussian random variable $\varepsilon$ equals zero the regression models (11) and (12) are the non-linear regression equations for which the prediction results for values of components of vector $\mathbf{X} = \{X_1, X_2, X_3\}$ from Table 1 and values of MRE are shown in the Table 2. The prediction results by model (11) and values of MRE are shown in the Table 2 for two cases: the Johnson univariate and multivariate normalizing transformations. Table 2 also contains the prediction results by linear regression equation (3) from [2] for values of components of vector **X** from Table 1 and MRE values. Note, all prediction results by linear regression equation (3), non-linear regression models (11) and (12) are positive.

Table 2 – The prediction results and confidence intervals of regressions for 30 open-source Java-based systems

| No | Linear regression | | | | Non-linear regression | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | univariate normalizing transformation | | | | | | | | the Johnson multivariate normalizing transformation | | | |
| | | | | | the decimal logarithm | | | | the Johnson transformation | | | | | | | |
| | $\hat{Y}$ | RME | LB | UB | $\hat{Y}$ | RME | LB | UB | $\hat{Y}$ | RME | LB | UB | $\hat{Y}$ | RME | LB | UB |
| 1 | 11.205 | 0.0437 | 8.301 | 14.109 | 12.197 | 0.0410 | 10.923 | 13.620 | 10.671 | 0.0893 | 8.989 | 12.756 | 11.978 | 0.0223 | 10.322 | 13.263 |
| 2 | 51.784 | 0.0897 | 48.396 | 55.171 | 53.248 | 0.1205 | 47.404 | 59.812 | 55.919 | 0.1768 | 49.992 | 61.568 | 52.316 | 0.1009 | 48.849 | 57.132 |
| 3 | 86.989 | 0.0355 | 79.427 | 94.551 | 90.515 | 0.0774 | 77.073 | 106.302 | 87.586 | 0.0426 | 83.866 | 89.805 | 87.262 | 0.0387 | 82.880 | 89.624 |
| 4 | 34.523 | 0.2787 | 31.823 | 37.223 | 33.653 | 0.2465 | 30.525 | 37.102 | 35.152 | 0.3020 | 30.502 | 40.115 | 33.719 | 0.2489 | 31.087 | 37.648 |
| 5 | 39.765 | 0.0469 | 36.726 | 42.804 | 39.052 | 0.0639 | 35.787 | 42.615 | 40.333 | 0.0333 | 35.838 | 44.986 | 38.969 | 0.0659 | 35.830 | 42.245 |
| 6 | 13.516 | 0.0385 | 6.843 | 20.188 | 10.941 | 0.1593 | 8.918 | 13.424 | 10.900 | 0.1625 | 7.968 | 15.255 | 11.602 | 0.1086 | 9.430 | 15.320 |
| 7 | 26.365 | 0.1328 | 21.326 | 31.404 | 24.255 | 0.2022 | 21.099 | 27.884 | 24.630 | 0.1898 | 20.414 | 29.485 | 23.956 | 0.2120 | 20.880 | 26.937 |
| 8 | 36.937 | 0.2668 | 30.570 | 43.305 | 35.027 | 0.2012 | 30.424 | 40.326 | 37.853 | 0.2982 | 31.806 | 44.302 | 35.358 | 0.2126 | 31.382 | 39.866 |
| 9 | 62.516 | 0.1698 | 57.570 | 67.462 | 60.729 | 0.1363 | 52.927 | 69.681 | 64.288 | 0.2029 | 57.204 | 70.537 | 62.131 | 0.1626 | 56.682 | 67.124 |
| 10 | 23.243 | 0.2433 | 20.890 | 25.595 | 22.674 | 0.2129 | 21.041 | 24.433 | 22.251 | 0.1903 | 19.396 | 25.446 | 22.451 | 0.2009 | 20.399 | 24.563 |
| 11 | 21.446 | 0.1872 | 16.885 | 26.006 | 19.692 | 0.2536 | 17.110 | 22.665 | 18.897 | 0.2838 | 15.560 | 22.881 | 19.239 | 0.2708 | 16.405 | 21.492 |
| 12 | 40.496 | 0.0459 | 38.185 | 42.808 | 39.963 | 0.0321 | 36.836 | 43.355 | 41.354 | 0.0680 | 37.021 | 45.812 | 39.831 | 0.0287 | 36.978 | 43.139 |
| 13 | 68.744 | 0.0912 | 64.647 | 72.842 | 68.808 | 0.0904 | 61.557 | 76.912 | 70.832 | 0.0636 | 65.599 | 75.335 | 68.708 | 0.0917 | 64.545 | 72.542 |
| 14 | 54.028 | 0.1564 | 50.795 | 57.260 | 52.739 | 0.1288 | 47.735 | 58.266 | 55.262 | 0.1828 | 49.875 | 60.436 | 53.278 | 0.1404 | 49.565 | 57.466 |
| 15 | 8.487 | 0.3234 | 5.441 | 11.534 | 10.056 | 0.5681 | 8.926 | 11.329 | 8.703 | 0.3571 | 7.338 | 10.443 | 9.863 | 0.5379 | 8.468 | 10.966 |
| 16 | 70.670 | 0.1115 | 61.411 | 79.928 | 62.671 | 0.2120 | 53.424 | 73.519 | 73.437 | 0.0767 | 64.730 | 79.996 | 70.101 | 0.1186 | 64.881 | 77.099 |
| 17 | 39.831 | 0.0960 | 37.362 | 42.300 | 38.454 | 0.0581 | 35.184 | 42.028 | 39.331 | 0.0822 | 34.828 | 44.018 | 38.677 | 0.0642 | 35.425 | 41.986 |
| 18 | 65.401 | 0.0958 | 59.939 | 70.864 | 63.010 | 0.0557 | 55.949 | 70.961 | 66.982 | 0.1223 | 60.723 | 72.447 | 64.714 | 0.0843 | 60.331 | 69.576 |
| 19 | 45.094 | 0.1062 | 39.943 | 50.244 | 43.124 | 0.1453 | 37.198 | 49.994 | 47.605 | 0.0565 | 39.805 | 55.417 | 44.662 | 0.1148 | 40.487 | 51.875 |
| 20 | 1.200 | 0.6071 | –2.531 | 4.932 | 3.118 | 0.0206 | 2.525 | 3.850 | 3.430 | 0.1227 | 3.180 | 3.849 | 3.173 | 0.0388 | 2.947 | 3.480 |
| 21 | 58.055 | 0.0822 | 47.956 | 68.154 | 54.860 | 0.1328 | 45.880 | 65.597 | 70.088 | 0.1080 | 58.879 | 78.528 | 62.201 | 0.0167 | 56.926 | 73.899 |
| 22 | 82.053 | 0.1011 | 75.043 | 89.064 | 92.398 | 0.0123 | 77.034 | 110.827 | 88.089 | 0.0350 | 84.129 | 90.284 | 88.102 | 0.0348 | 84.265 | 91.745 |
| 23 | 33.031 | 0.0099 | 28.529 | 37.532 | 29.837 | 0.0878 | 26.096 | 34.113 | 31.165 | 0.0471 | 26.305 | 36.522 | 31.086 | 0.0496 | 27.987 | 35.574 |
| 24 | 5.692 | 0.4826 | 1.975 | 9.408 | 7.899 | 0.2819 | 6.666 | 9.359 | 6.502 | 0.4089 | 5.391 | 8.063 | 7.677 | 0.3021 | 6.421 | 8.778 |
| 25 | 5.622 | 0.0143 | 2.367 | 8.878 | 7.921 | 0.4290 | 6.917 | 9.071 | 6.795 | 0.2259 | 5.745 | 8.199 | 7.745 | 0.3972 | 6.621 | 8.678 |
| 26 | 24.958 | 0.1002 | 22.712 | 27.204 | 24.148 | 0.0644 | 22.338 | 26.104 | 23.874 | 0.0524 | 20.862 | 27.219 | 24.192 | 0.0664 | 22.075 | 26.610 |
| 27 | 1.197 | 0.6938 | -2.722 | 5.117 | 3.796 | 0.0295 | 3.169 | 4.547 | 3.548 | 0.0929 | 3.253 | 4.038 | 3.763 | 0.0378 | 3.404 | 4.254 |
| 28 | 18.942 | 0.0911 | 14.883 | 23.001 | 17.897 | 0.1413 | 15.224 | 21.039 | 13.423 | 0.3559 | 9.477 | 19.245 | 15.903 | 0.2369 | 11.765 | 18.810 |
| 29 | 6.820 | 0.2642 | 3.356 | 10.284 | 8.835 | 0.0468 | 7.597 | 10.274 | 7.279 | 0.2147 | 5.991 | 9.053 | 8.532 | 0.0795 | 7.117 | 9.686 |
| 30 | 9.248 | 0.1960 | 3.644 | 14.851 | 7.176 | 0.0719 | 4.599 | 11.199 | 7.022 | 0.0918 | 5.480 | 9.390 | 7.094 | 0.0825 | 5.826 | 8.605 |


DOI 10.15588/1607-3274-2018-3-17

MMRE and PRED(0.25) are accepted as standard evaluations of prediction results by regression models and equations. The acceptable values of MMRE and PRED(0.25) are not more than 0.25 and not less than 0.75 respectively. The acceptable value of $R^2$ is approximately the same as for PRED(0.25). The values of $R^2$, MMRE and PRED(0.25) equal respectively 0.9621, 0.1734 and 0.7667 for linear regression equation (3), and equal respectively 0.9541, 0.1441 and 0.8667 for the model (12), and equal respectively 0.9574, 0.1579 and 0.8000 for the model (11) for the Johnson univariate transformation. The values of $R^2$, MMRE and PRED(0.25) are better for the model (11) for the Johnson multivariate transformation in comparison with all previous equations, and are 0.9672, 0.1389 and 0.8667 respectively.

The confidence and prediction intervals of non-linear regression are defined for the data from Table 1. Table 2 contains the lower (LB) and upper (UB) bounds of the confidence intervals of linear and non-linear regressions on the basis of univariate and multivariate transformations respectively for 0.05 significance level. Note the lower bounds of the confidence interval of linear regression (3) from [2] are negative for the two rows of data: 20 and 27. All the lower bounds of the confidence interval of non-linear regressions are positive. The widths of the confidence interval of non-linear regression on the basis of the Johnson multivariate transformation are less than for linear regression (3) from [2] for 21 rows of data: 1, 3, 6–8, 10, 11, 13, 15, 16, 18, 20–25, 27–30. Also the widths of the confidence interval of non-linear regression on the basis of the Johnson multivariate transformation are less for more data rows than for non-linear regressions following the univariate transformations, both decimal logarithm and the Johnson. The widths of the confidence interval of non-linear regression on the basis of the Johnson multivariate transformation are less than following the decimal logarithm univariate transformation for 24 rows of data: 2–5, 7–9, 11–14, 16–25, 27, 29 and 30. And ones are less than following the Johnson univariate transformation for 27 rows of data: 1, 2, 4–21, 23–26, 28–30. Approximately the same results are obtained for the prediction intervals of regressions.

Table 3 contains the lower (LB) and upper (UB) bounds of the prediction intervals of linear and non-linear regressions on the basis of univariate and multivariate transformations respectively for 0.05 significance level. Note the lower bounds of the prediction interval of linear regression (3) are negative for the eight rows of data: 1, 15, 20, 24, 25, 27, 29 and 30. All the lower bounds of the prediction interval of non-linear regressions are positive. The widths of the prediction interval of non-linear regression on the basis of the Johnson multivariate transformation are less than for linear regression (3) from [2] for 16 rows of data: 1, 3, 6, 7, 10, 11, 15, 20, 22, 24–30. Also the widths of the prediction interval of non-linear regression on the basis of the Johnson multivariate transformation are less for more data rows than for non-linear regressions following the univariate transformations, both decimal logarithm and the Johnson. The widths of the prediction interval of non-linear regression on the basis of the Johnson multivariate transformation are less than following the decimal logarithm univariate transformation for 17 rows of data: 2–5, 8, 9, 12–14, 16–22 and 27. And ones are less than following the Johnson univariate transformation for 26 rows of data: 1, 2, 4–19, 21, 23–26, 28–30.

Table 3 – The bounds of the prediction intervals

| No | Bounds for linear regression | | Bounds for non–linear regression | | | |
|---|---|---|---|---|---|---|
| | | | Johnson univariate transformation | | Johnson multivariate transformation | |
| | LB | UB | LB | UB | LB | UB |
| 1 | –0.004 | 22.413 | 5.679 | 22.412 | 7.536 | 19.277 |
| 2 | 40.441 | 63.127 | 32.573 | 75.467 | 36.600 | 68.504 |
| 3 | 73.784 | 100.194 | 77.736 | 91.114 | 73.863 | 96.264 |
| 4 | 23.365 | 45.680 | 17.436 | 58.472 | 21.814 | 48.847 |
| 5 | 28.521 | 51.009 | 20.726 | 63.360 | 25.763 | 54.772 |
| 6 | 0.799 | 26.232 | 5.559 | 24.105 | 7.061 | 19.398 |
| 7 | 14.424 | 38.305 | 11.701 | 46.283 | 14.924 | 36.873 |
| 8 | 24.378 | 49.497 | 18.875 | 61.455 | 22.764 | 51.145 |
| 9 | 50.614 | 74.418 | 40.607 | 80.655 | 45.398 | 77.523 |
| 10 | 12.164 | 34.321 | 10.693 | 42.556 | 14.102 | 34.502 |
| 11 | 9.699 | 33.192 | 9.052 | 37.791 | 11.900 | 30.337 |
| 12 | 29.427 | 51.566 | 21.431 | 64.236 | 26.458 | 55.673 |
| 13 | 57.169 | 80.319 | 49.028 | 83.854 | 52.398 | 82.637 |
| 14 | 42.730 | 65.325 | 32.094 | 74.958 | 37.508 | 69.345 |
| 15 | –2.759 | 19.733 | 4.920 | 18.108 | 6.290 | 15.899 |
| 16 | 56.425 | 84.914 | 51.026 | 85.596 | 52.840 | 84.458 |
| 17 | 28.727 | 50.935 | 20.068 | 62.454 | 25.536 | 54.455 |
| 18 | 53.275 | 77.527 | 43.924 | 81.998 | 48.094 | 79.560 |
| 19 | 33.105 | 57.082 | 25.246 | 70.052 | 29.796 | 61.411 |
| 20 | –10.250 | 12.651 | 3.009 | 4.748 | 2.516 | 4.443 |
| 21 | 43.250 | 72.859 | 45.506 | 84.384 | 44.137 | 78.697 |
| 22 | 69.156 | 94.951 | 78.586 | 91.334 | 74.127 | 97.180 |
| 23 | 21.307 | 44.755 | 15.088 | 54.334 | 19.782 | 45.949 |
| 24 | –5.754 | 17.138 | 4.077 | 13.054 | 4.988 | 12.419 |
| 25 | –5.682 | 16.927 | 4.203 | 13.643 | 5.062 | 12.421 |
| 26 | 13.902 | 36.014 | 11.471 | 44.858 | 15.245 | 36.852 |
| 27 | –10.316 | 12.711 | 3.048 | 5.105 | 2.834 | 5.520 |
| 28 | 7.380 | 30.503 | 6.446 | 29.640 | 9.577 | 26.119 |
| 29 | –4.546 | 18.186 | 4.360 | 14.963 | 5.483 | 13.826 |
| 30 | –2.942 | 21.438 | 4.206 | 14.773 | 4.601 | 11.617 |

Following [13] multivariate kurtosis $\beta_2$ is estimated for the data on metrics of software from Table I and the normalized data on the basis of the decimal logarithm transformation, the Johnson univariate and multivariate transformations for $S_B$ family. The estimator of multivariate kurtosis given by [13]

$$\hat{\beta}_2 = \frac{1}{N}\sum_{i-1}^{N}\left\{\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)^T \mathbf{S}_N^{-1}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\right\}^2 . \qquad (13)$$

In our case, in the formula (13), the vectors $\mathbf{Z}$ and $\overline{\mathbf{Z}}$ should be replaced by the vectors $\mathbf{P}$ and $\overline{\mathbf{P}}$ or $\mathbf{T}$ and $\overline{\mathbf{T}}$, respectively, for the initial (non-Gaussian) or normalized data. It is known that $\beta_2 = m(m+2)$ holds under multivariate normality. The given equality is a necessary


DOI 10.15588/1607-3274-2018-3-17

condition for multivariate normality. In our case $\beta_2 = 24$. The estimators of multivariate kurtosis equal 27.17, 22.38, 32.05 and 24.02 for the data from Table 1, the normalized data on the basis of the decimal logarithm transformation, the Johnson univariate and multivariate transformations respectively. The values of these estimators indicate that the necessary condition for multivariate normality is practically performed for the normalized data on the basis of the decimal logarithm transformation and the Johnson multivariate transformation, it does not hold for other data.

## 6 DISCUSSION

As it evident from the Table 2 and Table 3, the values of lower bounds of the confidence and prediction intervals of linear regression (3) from [2] for estimating the software size of open-source Java-based information systems are negative for some data rows. In our opinion, the presence of negative values may be explained by two reasons. Firstly, for the initial data from Table 1, four basic assumptions that justify the use of linear regression model, one of which is normality of the error distribution, are not valid. Secondly, there is reason to reject the hypothesis that the sample of data from Table 1 comes from a multivariate normal distribution. Note all the lower bounds of the confidence and prediction intervals of non-linear regressions are positive.

Also note that in our case for the data from Table 1, the poor normalization of multivariate non-Gaussian data using the Johnson univariate transformation leads to an increase in the widths of the confidence and prediction intervals of non-linear regression for a larger number of data rows compared to both the Johnson multivariate transformation and the decimal logarithm transformation.

The widths of the confidence and prediction intervals of non-linear regression on the basis of the Johnson multivariate transformation are less for more data rows than for linear regression and non-linear regressions following the univariate transformations, both decimal logarithm and the Johnson. Also the values of $R^2$, MMRE and PRED(0.25) are better for the model (11) for the Johnson multivariate transformation in comparison with all previous equations and models, both linear and non-linear, based on univariate transformations. This may be explained best multivariate normalization and the fact that there is no reason to reject the hypothesis that the sample of data, which normalized by the Johnson multivariate transformation for $S_B$ family, comes from a multivariate normal distribution.

## CONCLUSIONS

The important problem of increase of confidence of software size estimation for open-source Java-based information systems is solved.

**The scientific novelty** of obtained results is that the techniques to build the non-linear regression model for multivariate non-Gaussian data on the basis of the multivariate normalizing transformations is firstly proposed. The non-linear regression model to estimate the software size of open-source Java-based information systems is constructed on the basis of the Johnson multivariate transformation for $S_B$ family. This model, in comparison with other regression models (both linear and non-linear), has a larger multiple coefficient of determination, a smaller value of the mean magnitude of relative error, a larger value of percentage of prediction and smaller widths of the confidence and prediction intervals of non-linear regression.

**The practical significance** of obtained results is that the software realizing the constructed model is developed in the sci-language for Scilab. The experimental results allow to recommend the constructed model for use in practice.

**Prospects for further research** may include the application of other multivariate normalizing transformations and data sets to construct the non-linear regression model for estimating the software size of open-source Java-based information systems.

## REFERENCES

1. Kaczmarek J., Kucharski M. Size and effort estimation for applications written in Java, *Information and Software Technology,* 2004, Vol. 46, Issue 9, pp. 589–601. DOI: 10.1016/j.infsof.2003.11.001
2. Tan H. B. K., Zhao Y., Zhang H. Estimating LOC for information systems from their conceptual data models, *Software Engineering : the 28th International Conference (ICSE '06), Shanghai*. China, May 20–28, 2006 : proceedings, pp. 321–330. DOI: 10.1145/1134285.1134331
3. Tan H. B. K., Zhao Y., Zhang H. Conceptual data model-based software size estimation for information systems, *Transactions on Software Engineering and Methodology,* 2009, Vol. 19, Issue 2, October 2009, Article No. 4. DOI: 10.1145/1571629.1571630
4. Kiewkanya M., Surak S. Constructing C++ software size estimation model from class diagram, *Computer Science and Software Engineering : 13th International Joint Conference, Khon Kaen, Thailand, July 13–15, 2016 : proceedings*, pp. 1–6. DOI: 10.1109/JCSSE.2016.7748880
5. Bates D. M., Watts D. G. Nonlinear Regression Analysis and Its Applications. New York, John Wiley & Sons, 1988, 384 p. DOI:10.1002/9780470316757
6. Seber G.A.F., Wild C. J. Nonlinear Regression. New York, John Wiley & Sons, 1989, 768 p. DOI: 10.1002/0471725315
7. Ryan T.P. Modern regression methods. New York, John Wiley & Sons, 1997, 529 p. DOI: 10.1002/9780470382806
8. Johnson R. A., Wichern D. W. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, 800 p.
9. Prykhodko S. B. Developing the software defect prediction models using regression analysis based on normalizing transformations, *Modern Problems in Testing of the Applied Software : the Research and Practice Seminar (PTTAS-2016),* Poltava, Ukraine, May 25–26, 2016 : abstracts, pp. 6–7.
10. Stanfield P. M., Wilson J. R., Mirka G. A., Glasscock N. F., Psihogios J. P., Davis J. R. Multivariate input modeling with Johnson distributions, The 28th Winter simulation conference WSC'96, Coronado, CA, USA, December 8–11, 1996 : proceedings, ed. S. Andradyttir, K. J. Healy, D.H.Withers, and B. L. Nelson, *IEEE Computer Society Washington, DC, USA,* 1996, pp. 1457–1464.


DOI 10.15588/1607-3274-2018-3-17

11. Prykhodko S., Prykhodko N., Makarova L., Pugachenko K. Detecting Outliers in Multivariate Non-Gaussian Data on the basis of Normalizing Transformations, *Electrical and Computer Engineering : the 2017 IEEE First Ukraine Conference (UKRCON) «Celebrating 25 Years of IEEE Ukraine Section», Kyiv, Ukraine, May 29 – June 2, 2017 : proceedings,* pp. 846–849. DOI: 10.1109/UKRCON.2017.8100366
12. Prykhodko S., Prykhodko N., Makarova L., Pukhalevych A. Application of the Squared Mahalanobis Distance for Detecting Outliers in Multivariate Non-Gaussian Data, *Radioelectronics, Telecommunications and Computer Engineering : 14th International Conference on Advanced Trends (TCSET), Lviv-Slavske, Ukraine, February 20–24, 2018 : proceedings,* pp. 962–965. DOI: 10.1109/TCSET.2018.8336353
13. Mardia K. V. Measures of multivariate skewness and kurtosis with applications, *Biometrika,* 1970, 57, pp. 519–530. DOI: 10.1093/biomet/57.3.519

Received 18.05.2018.
Accepted 05.06.2018.

УДК 004.412:519.237.5

## НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ СИСТЕМ З ВІДКРИТИМ КОДОМ НА JAVA

**Приходько Н. В.** – канд. екон. наук, доцент, доцент кафедри фінансів Національного університету кораблебудування імені адмірала Макарова, Миколаїв, Україна.

**Приходько С. Б.** – д-р техн. наук, професор, завідувач кафедри програмного забезпечення автоматизованих систем Національного університету кораблебудування імені адмірала Макарова, Миколаїв, Україна.

### АНОТАЦІЯ

**Актуальність.** Проблема оцінювання розміру програмного забезпечення на ранній стадії програмного проекту є важливою, оскільки інформація, отримана при оцінюванні розміру програмного забезпечення, використовується для прогнозування трудомісткості по розробці програмного забезпечення, включаючи інформаційні системи на базі Java з відкритим вихідним кодом. Об’єктом дослідження є процес оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java. Предметом дослідження є моделі регресії для оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java. Мета роботи – створення моделі нелінійної регресії для оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java на основі багатовимірного нормалізуючого перетворення Джонсона.

**Метод.** Моделі, довірчі інтервали та інтервали передбачення багатовимірної нелінійної регресії для оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java побудовані на основі багатовимірного нормалізуючого перетворення Джонсона для негаусівських даних за допомогою відповідних методів. Методи побудови моделей, рівнянь, довірчих інтервалів і інтервалів передбачення нелінійних регресій засновані на багатовимірному нелінійному регресійному аналізі з використанням багатовимірних нормалізуючих перетворень. Розглянуто відповідні методи. Ці методи дозволяють враховувати кореляцію між випадковими величинами в разі нормалізації багатовимірних негаусівських даних. Загалом, це призводить до зменшення середньої величини відносної похибки, ширини довірчих інтервалів і інтервалів передбачення в порівнянні з лінійними моделями або нелінійними моделями, побудованими з використанням одновимірних нормалізуючих перетворень.

**Результати.** Здійснено порівняння побудованої моделі з моделями лінійної регресії та нелінійними регресіями на основі десяткового логарифму та одновимірного перетворення Джонсона.

**Висновки.** Модель нелінійної регресії для оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java побудована на основі багатовимірного перетворення Джонсона для сімейства $S_B$. Ця модель в порівнянні з іншими регресійній моделі (як лінійними, так і нелінійними) має більший множинний коефіцієнт детермінації і менше значення середньої величини відносної похибки. Перспективи подальших досліджень можуть включати застосування інших багатовимірних нормалізують перетворень і наборів даних для побудови моделі нелінійної регресії для оцінювання розміру програмного забезпечення інформаційних систем з відкритим вихідним кодом на Java.

**КЛЮЧОВІ СЛОВА:** оцінювання розміру програмного забезпечення, інформаційна система на основі Java, модель нелінійної регресії, одновимірне нормалізуюче перетворення, негаусівські дані.

УДК 004.412:519.237.5

## НЕЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ РАЗМЕРА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ СИСТЕМ С ОТКРЫТЫМ КОДОМ НА JAVA

**Приходько Н. В.** – канд. екон. наук, доцент, доцент кафедри фінансів Національного університету кораблебудування імені адмірала Макарова, Миколаїв, Україна.

**Приходько С. Б.** – д-р техн. наук, професор, завідувач кафедри програмного забезпечення автоматизованих систем Національного університету кораблебудування імені адмірала Макарова, Миколаїв, Україна.

### АННОТАЦИЯ

**Актуальность.** Проблема оценки размера программного обеспечения на ранней стадии программного проекта важна, поскольку информация, полученная при оценке размера программного обеспечения, используется для прогнозирования трудоемкости по разработке программного обеспечения, включая информационные системы на базе Java с открытым исходным кодом. Объект исследования – процесс оценки размера программного обеспечения информационных систем с


DOI 10.15588/1607-3274-2018-3-17

открытым исходным кодом на Java. Предмет исследования – модели регрессии для оценки размера программного обеспечения информационных систем с открытым исходным кодом на Java. Цель работы – создание модели нелинейной регрессии для оценки размера программного обеспечения информационных систем с открытым исходным кодом на Java на основе многомерного нормализирующего преобразования Джонсона.

**Метод.** Модели, доверительные интервалы и интервалы прогнозирования многомерной нелинейной регрессии для оценки размера программного обеспечения информационных систем с открытым исходным кодом на Java построены на основе многомерного нормализирующего преобразования Джонсона для негауссовских данных с помощью соответствующих методов. Методы построения моделей, уравнений, интервалов доверия и прогнозирования нелинейных регрессий основаны на многократном нелинейном регрессионном анализе с использованием многомерных нормализирующих преобразований. Рассмотрены соответствующие методы. Методы позволяют учитывать корреляцию между случайными величинами в случае нормализации многомерных негауссовских данных. В общем, это приводит к уменьшению средней величины относительной погрешности, ширины доверительных интервалов и интервалов предсказания по сравнению с линейными моделями или нелинейными моделями, построенными с использованием одномерных нормализирующих преобразований.

**Результаты.** Проведено сравнение построенной модели с линейной моделью и нелинейными регрессионными моделями на основе десятичного логарифма и одномерного преобразования Джонсона.

**Выводы.** Модель нелинейной регрессии для оценки размера программного обеспечения информационных систем с открытым исходным кодом на Java построена на основе многомерного преобразования Джонсона для семейства $S_B$. Эта модель по сравнению с другими регрессионными моделями (как линейными, так и нелинейными) имеет больший множественный коэффициент детерминации и меньшее значение средней величины относительной погрешности. Перспективы дальнейших исследований могут включать применение других многомерных нормализующих преобразований и наборов данных для построения модели нелинейной регрессии для оценки размера программного обеспечения информационных систем с открытым исходным кодом на Java.

**КЛЮЧЕВЫЕ СЛОВА:** оценка размера программного обеспечения, информационная система на основе Java, модель нелинейной регрессии, одномерное нормализующее преобразование, негауссовские данные.

## ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Kaczmarek J. Size and effort estimation for applications written in Java / J. Kaczmarek, M. Kucharski // Information and Software Technology. – 2004. – Vol. 46, Issue 9. – P. 589–601. DOI: 10.1016/j.infsof.2003.11.001
2. Tan H. B. K. Estimating LOC for information systems from their conceptual data models / H. B. K. Tan, Y. Zhao, H. Zhang // Software Engineering : the 28th International Conference (ICSE '06), Shanghai, China, May 20–28, 2006 : proceedings. – P. 321–330. DOI: 10.1145/1134285.1134331
3. Tan H. B. K. Conceptual data model-based software size estimation for information systems / H. B. K. Tan, Y. Zhao, H. Zhang // Transactions on Software Engineering and Methodology. – 2009. – Vol. 19, Issue 2. – October 2009. – Article No. 4. DOI: 10.1145/1571629.1571630
4. Kiewkanya M. Constructing C++ software size estimation model from class diagram / M. Kiewkanya, S. Surak // Computer Science and Software Engineering : 13th International Joint Conference, Khon Kaen, Thailand, July 13–15, 2016 : proceedings. – P. 1–6. DOI: 10.1109/JCSSE.2016.7748880
5. Bates D. M. Nonlinear Regression Analysis and Its Applications / D. M. Bates, D. G. Watts. – New York : John Wiley & Sons, 1988. – 384 p. DOI:10.1002/9780470316757
6. Seber G. A. F. Nonlinear Regression / G. A. F. Seber, C. J. Wild. – New York : John Wiley & Sons, 1989. – 768 p. DOI: 10.1002/0471725315
7. Ryan T. P. Modern regression methods / T. P. Ryan. – New York : John Wiley & Sons, 1997. – 529 p. DOI: 10.1002/9780470382806
8. Johnson R. A. Applied Multivariate Statistical Analysis / R. A. Johnson, D. W. Wichern. – Pearson Prentice Hall, 2007. – 800 p.
9. Prykhodko S.B. Developing the software defect prediction models using regression analysis based on normalizing transformations / S. B. Prykhodko // Modern Problems in Testing of the Applied Software : the Research and Practice Seminar (PTTAS-2016), Poltava, Ukraine, May 25–26, 2016 : abstracts. – P. 6–7.
10. Stanfield P. M. Multivariate input modeling with Johnson distributions / [P. M. Stanfield, J. R. Wilson, G. A. Mirka, et al.] // the 28th Winter simulation conference WSC'96, Coronado, CA, USA, December 8–11, 1996 : proceedings, ed. S. Andradyttir, K. J. Healy, D. H. Withers, and B. L. Nelson. – IEEE Computer Society Washington, DC, USA, 1996. – P. 1457–1464.
11. Prykhodko S. Detecting Outliers in Multivariate Non-Gaussian Data on the basis of Normalizing Transformations / S. Prykhodko, N. Prykhodko, L. Makarova, K. Pugachenko // Electrical and Computer Engineering : the 2017 IEEE First Ukraine Conference (UKRCON) «Celebrating 25 Years of IEEE Ukraine Section», Kyiv, Ukraine, May 29 – June 2, 2017 : proceedings. – P. 846–849. DOI: 10.1109/UKRCON.2017.8100366
12. Application of the Squared Mahalanobis Distance for Detecting Outliers in Multivariate Non-Gaussian Data / [S.Prykhodko, N. Prykhodko, L. Makarova, A. Pukhalevych] // Radioelectronics, Telecommunications and Computer Engineering : 14th International Conference on Advanced Trends (TCSET), Lviv-Slavske, Ukraine, February 20–24, 2018 : proceedings. – P. 962–965. DOI: 10.1109/TCSET.2018.8336353
13. Mardia K. V. Measures of multivariate skewness and kurtosis with applications / K. V. Mardia // Biometrika. – 1970. – 57. – P. 519–530. DOI: 10.1093/biomet/57.3.519


DOI 10.15588/1607-3274-2018-3-17