UDC 004.412:519.237.5

A NONLINEAR REGRESSION MODEL TO ESTIMATE THE SIZE OF WEB APPS CREATED USING THE CAKEPHP FRAMEWORK

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

Shutko I. S. – Post-graduate student of the Department of Software of Automated Systems, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine.

Prykhodko A. S. – Student 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 because a software size estimate is used for predicting the software development effort, including Web apps created using the CakePHP framework. The object of the study is the process of estimating the size of Web apps created using the CakePHP framework. The subject of the study is the nonlinear regression models to estimate the size of Web apps created using the CakePHP framework.

Objective. The goal of the work is the building the nonlinear regression model with three predictors for estimating the size of Web apps created using the CakePHP framework on the basis of the Box-Cox four-variate normalizing transformation to increase the confidence in early size estimation of these apps.

Method. The model, confidence and prediction intervals of multiply nonlinear regression to estimate the size of Web apps created using the CakePHP framework are constructed based on the Box-Cox multivariate normalizing transformation for non-Gaussian data with the help of appropriate techniques. The techniques to build the models, confidence, and prediction intervals of nonlinear regressions are based on the multiple nonlinear regression analysis using the multivariate normalizing transformations. The techniques allow taking into account the correlation between dependent and independent 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 nonlinear models constructed using univariate normalizing transformations.

Results. Comparison of the constructed model with the nonlinear regression models based on the decimal logarithm and the Box-Cox univariate transformation has been performed.

Conclusions. The nonlinear regression model with three predictors to estimate the size of Web apps created using the CakePHP framework is constructed on the basis of the Box-Cox four-variate transformation. This model, in comparison with other nonlinear regression models, has a larger multiple coefficient of determination, a smaller value of the mean magnitude of relative error and smaller widths of the confidence and prediction intervals. The prospects for further research may include the application of other multivariate normalizing transformations and data sets to construct the nonlinear regression model to estimate the size of Web apps created using the other frameworks.

KEYWORDS: software size estimation, Web app, nonlinear regression model, normalizing transformation, non-Gaussian data.

ABBREVIATIONS

DIT is a depth of inheritance tree;
KLOC is a thousand lines of code;
LB is a lower bound;
MMRE is a mean magnitude of relative error;
MRE is a magnitude of relative error;
PHP is a hypertext preprocessor;
PRED is a percentage of prediction;
SMD is a squared Mahalanobis distance;
UB is an upper bound.

NOMENCLATURE

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

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

k is a number of predictors (independent variables);

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, $\mathbf{S}_N = \left[S_{ij}\right]$;

$\mathbf{T}$ 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 ν degrees of freedom and $\alpha/2$ significance level;

X_1 is a number of classes;

X_2 is a average number of methods per class;

X_3 is a DIT mean value per class;

Y is an actual software size in KLOC;

Z_j is a *j*-th standard Gaussian variable that is obtained by transforming variable X_j, $Z_j \sim N(0,1)$, $j = 1,2,\ldots,k$;

Z_Y is a standard Gaussian variable that is obtained by transforming variable Y, $Z_Y \sim N(0,1)$;

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

DOI 10.15588/1607-3274-2021-4-12

$\hat{Z}_Y$ is a prediction result by linear regression equation for normalized data;

α is a significance level;

β_1 is a multivariate skewness;

β_2 is a multivariate kurtosis;

ε is a Gaussian random variable which defines residuals, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$;

ν is a number of degrees of freedom;

σ_ε is a standard deviation of ε;

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

$\mathbf{z}_X^+$ is a vector with components $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$ for *i*-row.

INTRODUCTION

Early software size estimation is one of the project managers' significant problems in evaluating software development efforts using mathematical models like COCOMO II [1]. Now many Web apps are created using the CakePHP framework making app development faster. However, today some software size estimation models that base metrics that can be measured from the class diagram are known [2–10]. There are only some regression equations and models, both linear [3, 4] and nonlinear [7, 8] ones, for estimating the software size of information open-source PHP-based systems. This demands the construction of the models for early size estimation of Web apps created using the CakePHP framework.

The object of study is the process of estimating the size of Web apps created using the CakePHP framework.

The subject of study is the regression models to estimate the size of Web apps created using the CakePHP framework.

The purpose of the work is to increase the confidence in early size estimation of Web apps created using the CakePHP framework.

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 average number of methods per class X_2, the average of Depth of Inheritance Tree (DIT) per class X_3 in class diagram from *N* Web apps. Suppose that there are bijective five-variate normalizing transformation of non-Gaussian random vector $\mathbf{P} = \{Y, X_1, X_2, X_3\}^T$ to Gaussian random vector $\mathbf{T} = \{Z_Y, Z_1, Z_2, Z_3\}^T$ is given by

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

and the inverse transformation for (1)

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

It is required to build the nonlinear regression model in the form $Y = Y(X_1, X_2, X_3, \varepsilon)$ based on the transformations (1) and (2).

2 REVIEW OF THE LITERATURE

In paper [3] the linear regression equations were proposed for estimating the software size of open-source PHP- and Java-based information systems. These equations are constructed on the basis of three metrics that can be measured from conceptual data model based a class diagram: a total number of classes, a total number of relationships, and an average number of attributes per class. However, there are four basic assumptions that justify the use of linear regression models, one of which is normality of the error distribution [11–13]. But this assumption is valid only in particular cases. Therefore, in papers [7] and [8], the nonlinear regression models were constructed using the same above metrics for estimating the software size of PHP- and Java-based information systems, respectively. But the size of Web apps may depend on other metrics. That is why in [9] the nonlinear regression model was constructed for estimating the size of Web apps created using the Laravel framework. This model depends on three factors (predictors), namely the total number of classes, the average number of methods per class, and the sum of average afferent coupling and average efferent coupling per class. However, the size of Web apps created using the CakePHP framework may depend on other metrics, and the model might have other parameters. This leads to the need of building the nonlinear regression model to estimate the size of Web apps created using the CakePHP framework.

A normalizing transformation is often a good way to construct nonlinear regression models [14–19]. According to [16], transformations are made for essentially four purposes, two of which are: firstly, to obtain approximate normality for the distribution of the error term (residuals), 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 initial dependent and independent variables.

Well-known techniques to construct the nonlinear regression models are based on the univariate normalizing transformations (such as, the decimal logarithm, Box-Cox transformation), and do not take into account the correlation between dependent and independent variables. The use of such univariate normalizing transformations for constructing the nonlinear regression models does not always lead to good normality and linear relationship between normalized variables. This leads to the need to apply multivariate normalizing transformations.

In [7] the techniques to build the models, confidence, and prediction intervals of nonlinear regressions based on the bijective multivariate normalizing transformations were proposed. However, according to [20], there may be

DOI 10.15588/1607-3274-2021-4-12

data sets for which the results of creating nonlinear regression models depend on, firstly, which normalizing transformation is used, univariate, or multivariate, and, secondly, are there any outliers in the data set. That is why, in [20] the technique to build nonlinear regression models based on the multivariate normalizing transformations and prediction intervals was considered. In this technique the prediction intervals of nonlinear regressions are used to detect the outliers in the process of constructing the nonlinear regression models. We apply the above technique for building the nonlinear regression model with three predictors to estimate the size of Web apps created using the CakePHP framework.

3 MATERIALS AND METHODS

According to [20], the technique to build nonlinear regression models based on the multivariate normalizing transformations and prediction intervals consist of four steps. In the first step, multivariate non-Gaussian data are normalized using a multivariate normalizing transformation (1). To do this, we use the Box-Cox multivariate transformation.

In the second step, the nonlinear regression model is constructed based on the multivariate normalizing transformation (1) as in [7]. Before that, we first determine whether one data point of a multivariate non-Gaussian data set is a multidimensional outlier. To do this, we apply the statistical technique based on the normalizing transformations and the Mahalanobis squared distance (MSD) as in [7, 8, 20]. If there is a multidimensional outlier in a multivariate non-Gaussian data set, then we discard the one, and return to step 1, else build the linear regression model for normalized data based on the transformation (1) in the form

$$Z_Y = \hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \ldots + \hat{b}_k Z_k + \varepsilon, \qquad (3)$$

ε is a Gaussian random variable which defines residuals, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$.

After that the nonlinear regression model is built on the basis of the linear regression model (3) and the transformations (1) and (2) as

$$Y = \psi_Y^{-1}\left(\hat{Z}_Y + \varepsilon\right). \qquad (4)$$

In the third step, the prediction interval of nonlinear regression is defined [20]

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{1 + \frac{1}{N} + \left(\mathbf{z}_X^+\right)^T \mathbf{S}_Z^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right), \qquad (5)$$

where $t_{\alpha/2,\nu}$ is a student's *t*-distribution quantile with $\alpha/2$ significance level and ν degrees of freedom; $\nu = N - k - 1$; k is a number of independent variables (in our case, k is 3); $\mathbf{z}_X^+$ is a vector with components $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$ for *i*-row; $\overline{Z}_j = \frac{1}{N}\sum_{i=1}^{N} Z_{j_i}$, $j = 1,2,\ldots,k$; $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$, $\nu = N - k - 1$; $\mathbf{S}_Z$ is a $k \times k$ matrix

$$\mathbf{S}_Z = \begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \ldots & S_{Z_1Z_k} \\ S_{Z_1Z_2} & S_{Z_2Z_2} & \ldots & S_{Z_2Z_k} \\ \ldots & \ldots & \ldots & \ldots \\ S_{Z_1Z_k} & S_{Z_2Z_k} & \ldots & S_{Z_kZ_k} \end{pmatrix}. \qquad (6)$$

In (6) $S_{Z_qZ_r} = \sum_{i=1}^{N}\left[Z_{q_i} - \overline{Z}_q\right]\left[Z_{r_i} - \overline{Z}_r\right]$, $q, r = 1,2,\ldots,k$.

In the fourth step, we check if there are data that exit the bounds of the prediction interval. And if we detect the outliers, we discard them and repeat all the steps starting with the first for new data without outliers, else nonlinear regression model construction is completed.

We constructed a nonlinear regression model to estimate the size of Web apps created using the CakePHP framework by the above technique from 38 apps hosted on GitHub (https://github.com). The data set was obtained using the PhpMetrics tool (https://phpmetrics.org/) around following variables: actual software size in the thousand lines of code (KLOC) *Y*, the total number of classes X_1, the average number of methods per class X_2, and the average of Depth of Inheritance Tree (DIT) per class X_3. Table 1 contains that data set. We chose the above predictors X_1, X_2, and X_3 for two reasons. Firstly, these predictors can be obtained from the class diagram, and, secondly, there is no multicollinearity between these predictors according to [21, 22] since variance inflation factors for predictors X_1, X_2, and X_3 are equal to 1.08, 1.03, and 1.11, respectively.

We checked the four-dimensional data from Table 1 for multivariate outliers. But before that, we tested the normality of multivariate data from Table I because well-known statistical methods (for example, multivariate outlier detection based on the squared Mahalanobis distance (SMD)) are used to detect outliers in multivariate data under the assumption that the data is described by a Gaussian distribution [18, 23]. We applied a multivariate normality test proposed by Mardia and based on measures of the multivariate skewness β_1 and kurtosis β_2 [24, 25]. According to this test, the distribution of four-dimensional data from Table I is not Gaussian since the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals to 161.59, is greater than the quantile of the Chi-Square distribution, which is 40.00 for 20 degrees of freedom and 0.005 significance level.

DOI 10.15588/1607-3274-2021-4-12

Similarly, the test statistic for multivariate kurtosis β_2, which equals to 44.38, is greater than the value of the Gaussian distribution quantile, which is 29.79 for 24 mean, 5.05 variance, and 0.005 significance level. Because we used the statistical technique [26] to detect multivariate outliers in the four-dimensional non-Gaussian data from Table I based on the multivariate normalizing transformations and the SMD for normalized data. To normalize the data from Table 1, we applied the four-variate Box-Cox transformation with components [18]

$$Z_j = x(\lambda_j) = \begin{cases} \left(X_j^{\lambda_j} - 1\right) \Big/ \lambda_j, & \text{if} \quad \lambda_j \neq 0; \\ \ln(X_j), & \text{if} \quad \lambda_j = 0. \end{cases} \tag{7}$$

Here Z_j is a Gaussian variable; λ_j is a parameter of the Box-Cox transformation, $j = 1,2,3$. The variable Z_Y is defined analogously (7) with the only difference that instead of Z_j, X_j, and λ_j should be put respectively Z_Y, Y, and λ_Y.

Table 1 – The data set and SMD values

No	Y	X_1	X_2	X_3	SMD	SMD_Z
1	0.448	4	4.75	1.40	2.07	4.02
2	7.846	90	3.86	1.71	0.08	1.01
3	4.345	42	4.19	1.81	0.11	0.63
4	2.717	60	2.20	2.11	3.84	6.24
5	2.954	45	3.13	2.13	2.34	3.42
6	1.717	26	2.54	1.44	2.09	5.11
7	0.212	1	7.00	2.00	4.31	6.07
8	1.149	14	3.86	1.73	0.30	0.56
9	0.477	8	2.63	2.20	3.84	3.55
10	61.269	487	4.93	1.79	19.41	5.77
11	0.124	2	2.50	2.00	2.60	4.23
12	0.358	2	3.50	2.00	1.22	**16.19**
13	0.349	8	2.25	1.38	3.15	5.86
14	3.486	29	3.62	2.00	1.05	4.67
15	1.538	11	5.82	1.80	1.51	1.71
16	0.365	4	4.00	1.50	1.08	2.53
17	2.332	23	4.00	1.79	0.21	0.26
18	24.347	328	3.59	1.24	7.78	4.59
19	12.433	66	6.52	1.79	3.08	3.82
20	0.948	10	4.30	1.78	0.28	0.84
21	1.826	20	4.15	1.60	0.47	0.28
22	0.882	9	3.56	1.60	0.72	1.52
23	3.567	23	6.83	1.24	7.30	6.55
24	3.735	49	3.57	1.70	0.22	0.53
25	47.61	289	5.33	1.33	18.86	7.55
26	4.029	58	3.38	1.54	0.61	0.72
27	3.5	40	4.18	1.63	0.23	0.55
28	0.131	3	2.00	1.67	2.73	6.61
29	2.227	23	4.22	1.46	1.11	0.83
30	0.521	2	8.50	3.00	**20.51**	12.06
31	1.494	22	3.41	1.35	1.94	2.30
32	0.906	5	7.80	1.60	7.59	5.31
33	0.471	5	3.80	2.00	0.97	1.26
34	24.04	303	3.67	1.75	5.34	3.74
35	2.681	23	4.78	1.71	0.39	0.39
36	20.327	308	3.94	1.20	9.35	6.58
37	24.66	367	3.36	1.64	11.55	4.01
38	0.142	3	3.00	2.00	1.76	10.13

The parameter estimates of the four-variate Box-Cox transformation for the data from Table 1 are calculated by the maximum likelihood method according to [18] and are $\hat{\lambda}_Y = -0.004265$, $\hat{\lambda}_1 = 0.012715$, $\hat{\lambda}_2 = -0.267106$, $\hat{\lambda}_3 = -0.610377$.

Table 1 contains the SMD for normalized data (SMD_Z), which is transformed using the four-variate BCT. The SMD_Z values from Table 1 indicate there is one multivariate outlier in four-dimensional non-Gaussian data since the SMD_Z value for row 12 is greater than the quantile of the Chi-Square distribution, which equals to 14.86 for the 0.005 significance level. In Table 1, the row numbers with the outliers are highlighted in bold. Also, Table 1 contains the SMD values for data without normalization. Note, for data without normalization, row 30 is the multivariate outlier since the SMD value for row 30 is greater than the quantile of the Chi-Square distribution for the 0.005 significance level.

Then, we built the linear regression model for data from Table 1 in the form

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \varepsilon, \tag{8}$$

where $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_3$ are parameter estimates, $\hat{b}_0 = -4.4155$, $\hat{b}_1 = 0.09974$, $\hat{b}_2 = 1.02683$, $\hat{b}_3 = -0.05258$, ε is a error term, which must be a Gaussian random variable to describe residuals, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, σ_ε is a standard deviation with the estimate $\hat{\sigma}_\varepsilon$ of 4.764.

According to [14], for error term in a linear regression model "the assumption of normality may be checked by examining the residuals." The null hypothesis H_0 that the observed frequency distribution of the ε values in (8) is the same as the normal distribution (there is no difference between the distributions) was checked by the Pearson Chi-Squared test. We rejected the null hypothesis H_0 with the 0.05 significance level since the χ^2 test statistic, which equals to 30.06, surpasses the critical value from the Chi-Squared distribution that is 7.81 for 0.05 significance level and 3 degrees of freedom. That is why there is no justification for the use of a linear regression model for mathematical modeling of the size of Web apps created using the CakePHP framework because the error distribution is not the Gaussian one. This creates a need to apply the nonlinear regression models in our case.

The nonlinear regression model with three predictors for estimating the size of Web apps created using the CakePHP framework is constructed based on the four-variate Box-Cox transformation for 37 data rows from Table 1 (without row 12) according to [20] and has the form [27]

$$Y = \left[\hat{\lambda}_Y \left(\hat{Z}_Y + \varepsilon\right) + 1\right]^{1/\hat{\lambda}_Y}, \tag{9}$$

DOI 10.15588/1607-3274-2021-4-12

where ε is a Gaussian random variable, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$, with the estimate $\hat{\sigma}_\varepsilon$ of 0.1286; $\hat{Z}_Y$ is a prediction result by the linear regression equation $\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3$ for normalized data, which are transformed by the four-variate Box-Cox transformation with components (7); $\hat{b}_0 = -4.27326$, $\hat{b}_1 = 1.06624$, $\hat{b}_2 = 1.17959$, $\hat{b}_3 = 0.37559$, $\hat{\lambda}_Y = -0.02762$, $\hat{\lambda}_1 = -0.02795$, $\hat{\lambda}_2 = 0.020293$, $\hat{\lambda}_3 = -0.67526$.

According to [20], after constructing a model (9), we have to find the nonlinear regression prediction interval by (5).

In the second iteration, for the data normalized by the four-variate Box-Cox transformation from 37 Web apps (without row 12), the matrix (6) is following

$$\mathbf{S}_Z = \begin{pmatrix} 79.91 & -1.12 & -2.60 \\ -1.12 & 4.50 & 0.049 \\ -2.60 & 0.049 & 0.594 \end{pmatrix}.$$

As we observe, there are two values of *Y* for Web applications 14 and 38 that are out of the prediction intervals computed by (5) for a significance level of 0.05. In Table 2, we marked the prediction interval lower (LB) and upper (UB) bounds calculated in the second iteration as LB_2, and UB_2.

Next, we erased data in Web applications 14 and 38. After that, we used data from the remaining 35 apps to construct the model. In nonlinear regression model (9) with parameters' estimates acquired from 35 rows of data, it appeared that two values of *Y* for rows 6 and 22 exceed the prediction interval. After five iterations, we saved 32 Web applications from Table 1 (excluding rows 6, 12, 14, 22, 25, and 38). There were no outliers in the fifth iteration. We completed the stages' iterations, and constructed nonlinear regression model (9) with 32 Web applications data. In Table 2, we marked the prediction interval LB and UB calculated in the third and fifth iterations as LB_3 and UB_3, and as LB_5 and UB_5, respectively. We highlighted the row numbers with the data outliers in bold, and a dash (–) shows the exception of the relevant numbers of data at the corresponding iteration.

In the fifth iteration, the parameter estimates for the model (9) constructed by the four-variate Box-Cox transformation from 32 Web apps are $\hat{\lambda}_Y = -0.05722$, $\hat{\lambda}_1 = -0.03731$, $\hat{\lambda}_2 = -0.10586$, $\hat{\lambda}_3 = -0.74223$, $\hat{b}_0 = -4.43109$, $\hat{b}_1 = 1.05416$, $\hat{b}_2 = 1.39398$, $\hat{b}_3 = 0.51635$, the estimate $\hat{\sigma}_\varepsilon$ is 0.05174.

In this case, the matrix (6) is following

$$\mathbf{S}_Z = \begin{pmatrix} 66.35 & -1.72 & -1.88 \\ -1.72 & 2.86 & 0.0605 \\ -1.88 & 0.0605 & 0.485 \end{pmatrix}.$$

We checked the multivariate normality of 32 rows of normalized data from Table I in the fifth iteration with a test proposed by Mardia [24].

Table 2 – LB and UB of nonlinear regression prediction intervals in various iterations

No	LB_2	UB_2	LB_3	UB_3	LB_5	UB_5
1	0.327	0.578	0.370	0.542	0.389	0.488
2	5.943	10.764	6.270	9.713	6.711	8.686
3	3.153	5.635	3.344	5.083	3.593	4.602
4	2.054	3.783	2.127	3.287	2.267	2.936
5	2.421	4.386	2.550	3.902	2.768	3.560
6	0.993	1.766	1.062	1.582	–	–
7	0.140	0.247	0.161	0.234	0.177	0.221
8	0.951	1.661	1.030	1.516	1.096	1.379
9	0.363	0.640	0.406	0.596	0.435	0.547
10	43.538	84.434	49.177	83.162	53.748	73.141
11	0.086	0.150	0.106	0.151	0.112	0.139
13	0.268	0.476	0.305	0.447	0.309	0.390
14	1.864	3.320	–	–	–	–
15	1.240	2.199	1.336	1.992	1.435	1.819
16	0.274	0.478	0.313	0.453	0.330	0.411
17	1.637	2.886	1.750	2.610	1.873	2.373
18	17.263	32.698	18.271	29.783	18.985	25.330
19	8.599	15.940	9.276	14.712	9.962	13.073
20	0.781	1.363	0.852	1.250	0.912	1.145
21	1.448	2.546	1.553	2.308	1.644	2.079
22	0.545	0.948	0.602	0.878	–	–
23	2.797	5.212	2.988	4.675	3.085	4.022
24	2.958	5.274	3.120	4.729	3.320	4.245
25	26.354	50.163	29.090	48.006	–	–
26	3.170	5.667	3.332	5.063	3.504	4.488
27	2.917	5.189	3.094	4.681	3.286	4.198
28	0.095	0.167	0.115	0.166	0.118	0.148
29	1.649	2.921	1.766	2.641	1.852	2.352
30	0.377	0.705	0.412	0.627	0.465	0.598
31	1.182	2.103	1.271	1.898	1.316	1.671
32	0.768	1.387	0.830	1.245	0.882	1.122
33	0.347	0.603	0.391	0.566	0.422	0.526
34	18.389	34.579	19.637	31.872	21.267	28.262
35	2.014	3.567	2.152	3.229	2.299	2.923
36	18.081	34.434	19.288	31.626	19.989	26.765
37	19.411	36.531	20.576	33.428	22.101	29.389
38	0.159	0.275	–	–	–	–

According to Mardia's test, the distribution of 32 rows of normalized data from Table I (excluding rows 6, 12, 14, 22, 25, and 38) is Gaussian since the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals to 14.05, is less than the quantile of the Chi-Square distribution, which is 40.00 for 20 degrees of freedom and 0.005 significance level. Similarly, the test statistic for multivariate kurtosis β_2, which equals to 23.80, is less than the quantile of the Gaussian distribution, which is 30.31 for 24 mean, 6.0 variance, and 0.005 significance level.

DOI 10.15588/1607-3274-2021-4-12

4 EXPERIMENTS

For comparison of the model (9) with other nonlinear regression models with three predictors, two nonlinear regression models are built based on 32 data rows from Table 1 (without rows 6, 12, 14, 22, 25, and 38) using the Box-Cox univariate transformation and the decimal logarithm univariate one.

The nonlinear regression model based on the linear regression model (3) for the normalized data and the decimal logarithm univariate transformation has the form

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

where the estimators for parameters are: $\hat{b}_0 = -1.80484$, $\hat{b}_1 = 0.981865$, $\hat{b}_2 = 1.186738$, $\hat{b}_3 = 0.319789$. The estimate $\hat{\sigma}_\varepsilon$ is 0.029761.

The nonlinear regression model based on the Box-Cox univariate transformation is analogously (9) with the only difference that the data for variables are normalized by the Box-Cox univariate transformation using the maximum likelihood method [18]. The estimators for parameters of the Box-Cox univariate transformation for each from variables Y, X_1, X_2, and X_3 are $\hat{\lambda}_Y = -0.054396$, $\hat{\lambda}_1 = -0.04861$, $\hat{\lambda}_2 = -0.15995$, $\hat{\lambda}_3 = -0.70493$. The parameter estimators of the linear regression model for normalized data by the Box-Cox univariate transformation are $\hat{b}_0 = -4.55626$, $\hat{b}_1 = 1.09337$, $\hat{b}_2 = 1.50882$, $\hat{b}_3 = 0.529486$. The estimate $\hat{\sigma}_\varepsilon$ is 0.055974.

The computer program implementing the constructed models (9) and (10) 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 [28].

5 RESULTS

The prediction results $\hat{Y}$ of nonlinear regression models (9) and (10) for values of predictors from Table 1 (without rows 6, 12, 14, 22, 25, and 38) and values of MRE are shown in the Table 3. The prediction results by model (9) and values of MRE are shown in the Table 3 for two cases: Box-Cox univariate and four-variate normalizing transformations.

Table 3 – The prediction results and confidence intervals of multiple regressions

No	the four-variate Box-Cox transformation				Univariate transformations							
					the decimal logarithm transformation				the Box-Cox transformation			
	$\hat{Y}$	MRE	LB	UB	$\hat{Y}$	MRE	LB	UB	$\hat{Y}$	MRE	LB	UB
1	0.436	0.0274	0.419	0.454	0.433	0.0344	0.409	0.457	0.435	0.0285	0.417	0.455
2	7.631	0.0274	7.401	7.869	7.667	0.0229	7.399	7.943	7.701	0.0185	7.451	7.960
3	4.065	0.0645	3.958	4.174	4.072	0.0629	3.947	4.201	4.149	0.0450	4.032	4.270
4	2.578	0.0510	2.445	2.720	2.826	0.0400	2.643	3.021	2.592	0.0461	2.446	2.746
5	3.138	0.0622	3.012	3.269	3.247	0.0992	3.085	3.418	3.201	0.0836	3.063	3.346
7	0.197	0.0686	0.188	0.207	0.197	0.0710	0.184	0.211	0.189	0.1077	0.180	0.199
8	1.229	0.0699	1.204	1.255	1.238	0.0775	1.204	1.273	1.254	0.0918	1.226	1.283
9	0.488	0.0230	0.468	0.508	0.490	0.0262	0.462	0.519	0.493	0.0328	0.471	0.515
10	62.657	0.0227	58.937	66.626	54.581	0.1092	51.153	58.238	59.540	0.0282	55.807	63.537
11	0.125	0.0041	0.119	0.130	0.115	0.0757	0.108	0.122	0.121	0.0213	0.116	0.127
13	0.347	0.0053	0.331	0.364	0.350	0.0040	0.328	0.374	0.346	0.0097	0.328	0.364
15	1.615	0.0500	1.566	1.665	1.611	0.0474	1.549	1.675	1.642	0.0676	1.589	1.697
16	0.368	0.0083	0.356	0.381	0.361	0.0119	0.344	0.378	0.368	0.0078	0.355	0.382
17	2.107	0.0964	2.061	2.154	2.126	0.0884	2.069	2.185	2.156	0.0755	2.105	2.208
18	21.916	0.0998	20.734	23.170	22.596	0.0719	21.294	23.978	21.199	0.1293	19.984	22.493
19	11.406	0.0826	10.903	11.934	10.687	0.1404	10.152	11.250	11.475	0.0771	10.933	12.045
20	1.021	0.0774	0.999	1.045	1.021	0.0766	0.991	1.051	1.040	0.0972	1.015	1.066
21	1.848	0.0119	1.809	1.888	1.868	0.0229	1.816	1.921	1.887	0.0333	1.844	1.931
23	3.521	0.0130	3.326	3.727	3.567	0.0000	3.332	3.819	3.553	0.0040	3.343	3.777
24	3.753	0.0047	3.660	3.848	3.839	0.0280	3.725	3.958	3.819	0.0225	3.717	3.924
26	3.964	0.0161	3.858	4.073	4.114	0.0211	3.979	4.253	4.018	0.0028	3.901	4.137
27	3.712	0.0607	3.628	3.799	3.743	0.0694	3.639	3.849	3.785	0.0814	3.692	3.880
28	0.132	0.0098	0.126	0.139	0.124	0.0563	0.115	0.133	0.129	0.0174	0.122	0.136
29	2.086	0.0631	2.031	2.144	2.122	0.0470	2.050	2.197	2.128	0.0445	2.067	2.191
30	0.527	0.0115	0.494	0.563	0.558	0.0702	0.505	0.615	0.517	0.0078	0.481	0.555
31	1.482	0.0080	1.431	1.536	1.539	0.0299	1.472	1.609	1.507	0.0087	1.451	1.566
32	0.994	0.0971	0.950	1.040	1.013	0.1176	0.952	1.077	0.990	0.0926	0.943	1.040
33	0.471	0.0002	0.458	0.485	0.463	0.0166	0.445	0.482	0.475	0.0086	0.460	0.490
34	24.502	0.0192	23.361	25.702	23.957	0.0034	22.720	25.261	23.885	0.0065	22.702	25.132
35	2.591	0.0336	2.531	2.653	2.588	0.0346	2.514	2.665	2.648	0.0124	2.581	2.716
36	23.116	0.1372	21.792	24.526	23.475	0.1549	22.067	24.973	22.397	0.1018	21.034	23.854
37	25.471	0.0329	24.281	26.723	25.506	0.0343	24.174	26.912	24.616	0.0018	23.395	25.904

DOI 10.15588/1607-3274-2021-4-12

The MRE values for the model (9) based on the Box-Cox four-variate transformation are smaller than for the model (9) based on the Box-Cox univariate transformation for 17 from 32 rows of data (rows 1, 5, 7–11, 13, 15, 18, 20, 21, 24, 27, 28, 31, 33). Also, the MRE values for the model (9) based on the Box-Cox four-variate transformation are less than for the model (10) based on the decimal logarithm univariate transformation for 21 from 32 rows of data (rows 1, 5, 7–11, 16, 19, 21, 24, 26–28, 30–33, 35–37).

To evaluate the prediction accuracy of the nonlinear regression models we applied the standard metrics R^2, MMRE, and PRED(0.25). MMRE and PRED(0.25) are accepted as standard evaluations of prediction results by regression models. These metrics are applied in software engineering too [29, 30]. The acceptable values of MMRE and PRED(0.25) are not more than 0.25 and not less than 0.75 respectively. The values of R^2, MMRE and PRED(0.25) equal respectively 0.9963, 0.0425, and 1.0 for model (9) based on the Box-Cox four-variate transformation, and equal respectively 0.9961, 0.0442 and 1.0 for the model (9) based on the Box-Cox univariate transformation, and equal respectively 0.9871, 0.0552 and 1.0 for the model (10) for the decimal logarithm univariate transformation. The MMRE and R^2 values are better for the model (9) based on the Box-Cox four-variate transformation.

The confidence and prediction intervals of nonlinear regression are defined for the data from Table 1 (without rows 6, 12, 14, 22, 25, and 38). Table 3 contains the lower (LB) and upper (UB) bounds of the confidence intervals of nonlinear regressions based on the univariate and four-variate transformations respectively for 0.05 significance level. We defined the confidence intervals of the sample mean of size using (5) with the only difference that in the sum in curly brackets, there is not 1. The widths of the confidence interval of nonlinear regression based on the Box-Cox four-variate transformation are less than for nonlinear regression based on the Box-Cox univariate transformation for all 32 rows of data. Also, the widths of the confidence interval of nonlinear regression based on the Box-Cox four-variate transformation are less than for nonlinear regression based on the decimal logarithm univariate transformation for 31 from 32 rows of data (except row 10). Approximately the same results are obtained for the prediction intervals of nonlinear regressions.

Table 4 contains the lower (LB) and upper (UB) bounds of the prediction intervals of non-linear regressions based on the univariate and multivariate transformations respectively for 0.05 significance level. The widths of the prediction interval of nonlinear regression based on the Box-Cox four-variate transformation are less than for nonlinear regression based on the Box-Cox univariate transformation for all 32 rows of data. Also, the widths of the prediction interval of nonlinear regression based on the Box-Cox four-variate transformation are less for nonlinear regression based on the decimal logarithm univariate transformation for 31 from 32 rows of data (except row 10).

Note, a more significant advantage of the model (9) constructed by the four-variate Box-Cox transformation compared with the two above models based on the univariate transformations is the smaller widths of the confidence and prediction intervals. Such, the width of the nonlinear regression prediction interval for the four-variate Box-Cox transformation is less than after the univariate Box-Cox transformation for all 32 data rows (with the difference up to 11%) and less than after decimal logarithm univariate transformation for 31 (with the difference up to 49%) from 32 data rows (except row 10 with the difference of 9%).

Table 4 – The bounds of prediction interval

No	Y	univariate				four-variate	
		decimal logarithm		Box-Cox		Box-Cox	
		LB	UB	LB	UB	LB	UB
1	0.448	0.369	0.507	0.385	0.493	0.389	0.488
2	7.846	6.586	8.924	6.707	8.853	6.711	8.686
3	4.345	3.501	4.735	3.633	4.744	3.593	4.602
4	2.717	2.403	3.323	2.255	2.982	2.267	2.936
5	2.954	2.777	3.797	2.796	3.669	2.768	3.560
7	0.212	0.167	0.232	0.168	0.214	0.177	0.221
8	1.149	1.065	1.439	1.108	1.421	1.096	1.379
9	0.477	0.418	0.574	0.435	0.558	0.435	0.547
10	61.269	46.450	64.135	50.568	70.205	53.748	73.141
11	0.124	0.098	0.135	0.108	0.137	0.112	0.139
13	0.349	0.298	0.412	0.305	0.392	0.309	0.390
15	1.538	1.383	1.877	1.445	1.867	1.435	1.819
16	0.365	0.309	0.421	0.326	0.415	0.330	0.411
17	2.332	1.829	2.470	1.898	2.451	1.873	2.373
18	24.347	19.271	26.495	18.181	24.751	18.985	25.330
19	12.433	9.140	12.496	9.922	13.287	9.962	13.073
20	0.948	0.878	1.187	0.920	1.177	0.912	1.145
21	1.826	1.607	2.171	1.662	2.143	1.644	2.079
23	3.567	3.031	4.197	3.081	4.101	3.085	4.022
24	3.735	3.302	4.464	3.346	4.362	3.320	4.245
26	4.029	3.536	4.787	3.518	4.593	3.504	4.488
27	3.500	3.220	4.350	3.318	4.321	3.286	4.198
28	0.131	0.105	0.146	0.114	0.145	0.118	0.148
29	2.227	1.824	2.470	1.871	2.422	1.852	2.352
30	0.521	0.467	0.666	0.451	0.593	0.465	0.598
31	1.494	1.319	1.795	1.325	1.716	1.316	1.671
32	0.906	0.863	1.188	0.869	1.128	0.882	1.122
33	0.471	0.397	0.540	0.422	0.536	0.422	0.526
34	24.04	20.478	28.028	20.527	27.827	21.267	28.262
35	2.681	2.226	3.009	2.327	3.016	2.299	2.923
36	20.327	20.001	27.551	19.173	26.199	19.989	26.765
37	24.66	21.797	29.846	21.150	28.685	22.101	29.389

Also, the width of the nonlinear regression prediction interval for the four-variate Box-Cox transformation is less than after the univariate Box-Cox transformation for all 32 data rows (with the difference up to 11%) and less than after decimal logarithm univariate transformation for 31 (with the difference up to 59%) from 32 data rows (except row 10 with the difference of 8%).

6 DISCUSSION

We apply four-variate normalizing transformations to build the nonlinear regression model for estimating the size of Web apps created using the CakePHP framework

DOI 10.15588/1607-3274-2021-4-12

by appropriate techniques [20] since the error distribution of the linear regression model is not Gaussian what the chi-squared test result indicates. Also, there are outliers in the data from Table 1. Moreover, the four-variate distribution of the data from Table 1 is not Gaussian what the Mardia multivariate normality test based on measures of the multivariate skewness and kurtosis indicates. Because we use the statistical technique [26] to detect multivariate outliers in the four-dimensional non-Gaussian data from Table I based on the multivariate normalizing transformations and the SMD for normalized data. Note, we have other four-variate outliers for the data from Table 1 without applying normalization compared to outlier detection results using the above technique [26].

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 multiple nonlinear regression for a larger number of data rows compared to the Box-Cox four-variate transformation.

The widths of the confidence and prediction intervals of multiple nonlinear regression based on the Box-Cox four-variate transformation are smaller for more data rows than for multiple nonlinear regressions following the univariate transformations, both the decimal logarithm and the Box-Cox ones. Also the MMRE value is smaller for the model (9) for the Box-Cox four-variate transformation in comparison with all other nonlinear models based on univariate transformations. This may be explained best four-variate normalization of non-Gaussian data from Table 1 using the Box-Cox four-variate transformation.

The obtained results and results from [9] indicate that constructing a multiple nonlinear regression model to estimate the size (in KLOC) of Web apps using the specific framework (CakePHP in our case and Laravel in [9]) leads to an increase of confidence in estimating.

CONCLUSIONS

The important problem of increase of confidence in estimating the size of Web apps created using the CakePHP framework is solved.

The scientific novelty of obtained results is that three-factors nonlinear regression model to estimate the size of Web apps created using the CakePHP framework is firstly constructed on the basis of the Box-Cox four-variate transformation. This model, in comparison with other nonlinear regression models, has a smaller value of the mean magnitude of relative error, smaller widths of the confidence and prediction intervals of three-factors nonlinear 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 multiple nonlinear regression models for estimating the size of Web apps created using the specific frameworks.

REFERENCES

1. Boehm B. W., Abts C., Brown A. W. et al. Software cost estimation with COCOMO II. Upper Saddle River, NJ, Prentice Hall PTR, 2000, 506 p.
2. 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
3. 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
4. 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
5. Zifen Y. An improved software size estimation method based on object-oriented approach, *Electrical & Electronics Engineering : IEEE Symposium EEESYM*'12, Kuala Lumpur, Malaysia, 24–27 June, 2012, proceedings, *IEEE*, 2012, pp. 615–617. DOI: 10.1109/EEESym.2012.6258733.
6. 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, IEEE, 2016, pp. 1–6. DOI: 10.1109/JCSSE.2016.7748880
7. Prykhodko N. V., Prykhodko S. B. Constructing the nonlinear regression models on the basis of multivariate normalizing transformations, *Electronic modeling*, 2018, Vol. 40, No. 6, pp. 101–110. DOI: 10.15407/emodel.40.06.101
8. Prykhodko N. V., Prykhodko S. B. The non-linear regression model to estimate the software size of open source Java-based systems, *Radio Electronics, Computer Science, Control*, 2018, No. 3 (46), pp. 158–166. DOI: 10.15588/1607-3274-2018-3-17
9. Prykhodko S. B., Prykhodko N. V., Vorona M. V. et al. Nonlinear regression model for estimating the size of web applications created using the Laravel framework, *Information technology and computer engineering*, 2021, Vol. 50, No. 1, pp. 115–121. DOI: 10.31649/1999-9941-2021-50-1-115-121 [Published in Ukrainian]
10. Nassif A. B., AbuTalib M., Capretz L. F. Software effort estimation from use case diagrams using nonlinear regression analysis, *Electrical and Computer Engineering : IEEE Canadian Conference CCECE'20.*, London, ON, Canada, 30 Aug.–2 Sept., 2020, proceedings, IEEE, 2020, pp. 1–4. DOI: 10.1109/CCECE47787.2020.9255712.
11. Montgomery D. C., Peck D. C., Introduction to linear regression analysis / D. C. Montgomery, // 2nd edn. – New York: John Wiley & Sons, 1992. – 544 p.
12. Seber G.A.F., Lee A. J. Linear regression analysis, 2nd edn. New York, John Wiley & Sons, 2003, 582 p.
13. Weisberg, S. Applied linear regression, 4th edn. New York, John Wiley & Sons, 2013, 368 p.
14. Bates D. M., Watts D. G. Nonlinear regression analysis and its applications. New York, John Wiley & Sons, 1988, 384 p. DOI:10.1002/9780470316757
15. Seber G.A.F., Wild C. J. Nonlinear regression. New York: John Wiley & Sons, 1989, 768 p. DOI: 10.1002/0471725315

DOI 10.15588/1607-3274-2021-4-12

16. Ryan T.P. Modern regression methods. New York, John Wiley & Sons, 1997, 529 p. DOI: 10.1002/9780470382806
17. Drapper N. R., Smith H. Applied regression analysis. New York, John Wiley & Sons, 1998, 736 p.
18. Johnson R. A., Wichern D. W. Applied multivariate statistical analysis. Pearson Prentice Hall, 2007, 800 p.
19. Chatterjee S., Simonoff J. S. Handbook of regression analysis. New York, John Wiley & Sons, 2013, 236 p. DOI: 10.1002/9781118532843
20. Prykhodko S., Prykhodko N. Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations, *Mathematical Modeling and Simulation of Systems : 15th International Scientific-practical Conference MODS'2020, Chernihiv, Ukraine, June 29 – July 01, 2020 : selected papers. Springer*, Cham., 2021, pp. 166–174. (Advances in Intelligent Systems and Computing, Vol. 1265). DOI: 10.1007/978-3-030-58124-4_16
21. Belsley D. A., Kuh E., Welsch R. E. Regression diagnostics: Identifying influential data and sources of collinearity / D. A. Belsley,. – New York: John Wiley, 1980. – 300 p. DOI:10.1002/0471725153
22. Chatterjee S., Price B. Regression analysis by example. New York, John Wiley & Son, 2012, 424 p.
23. Olkin I., Sampson A. R., Smelser N. J., Baltes P. B. Eds. Multivariate analysis: Overview, *International encyclopedia of social & behavioral sciences. 1st edn.* Elsevier, Pergamon, 2001, pp. 10240–10247.
24. Mardia K. V. Measures of multivariate skewness and kurtosis with applications, *Biometrika*, 1970, Vol. 57, pp. 519–530. DOI: 10.1093/biomet/57.3.519
25. Mardia K. V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhya: The Indian Journal of Statistics, Series B (1960–2002)*, 1974, Vol. 36, Issue 2, pp. 115–128.
26. 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
27. Prykhodko S., Prykhodko N., Knyrik K. Estimating the efforts of mobile application development in the planning phase using nonlinear regression analysis, *Applied Computer Systems*, 2020, Vol. 25, No. 2, pp. 172–179. DOI: 10.2478/acss-2020-0019
28. Campbell S. L., Chancelier J.-P., Nikoukhah R. Modeling and simulation in Scilab/Scicos. Springer, 2005, 313 p.
29. Foss T., Stensrud E., Kitchenham B., Myrtveit I. A simulation study of the model evaluation criterion MMRE, *IEEE Transactions on software engineering*, 2003, Vol. 29, Issue 11, pp. 985–995. DOI: 10.1109/TSE.2003.1245300
30. Port D., Korte M. Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research, *Empirical Software Engineering and Measurement, the 2nd ACM-IEEE International Symposium ESEM, Kaiserslautern*. Germany, October, 2008, proceedings. New York, ACM, 2008, pp. 51–60.

Received 04.09.2021.
Accepted 25.10.2021.

УДК 004.412:519.237.5

НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ ВЕБ-ЗАСТОСУНКІВ, ЩО СТВОРЮЮТЬСЯ З ВИКОРИСТАННЯМ ФРЕЙМВОРКУ CAKEPHP

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

Шутко І. С. – аспірант кафедри програмного забезпечення автоматизованих систем Національного університету кораблебудування імені адмірала Макарова, Миколаїв, Україна.

Приходько А. С. – студент 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 оцінювання розміру програмного забезпечення використовується для прогнозування трудомісткості розробки програмного забезпечення, включаючи веб-застосунки з відкритим кодом на PHP, що створені із використанням фреймворку CakePHP. Об'єктом дослідження є процес оцінювання розміру веб-застосунків з відкритим кодом на PHP, що створені із використанням фреймворку CakePHP. Предметом дослідження є нелінійні регресійні моделі для оцінювання розміру вебзастосунків з відкритим кодом на PHP, що створені із використанням фреймворку CakePHP.

Мета. Метою роботи є побудова нелінійної регресійної моделі з трьома предикторами для оцінювання розміру вебзастосунків, шо створюються із використанням фреймворку CakePHP на основі чотиривимірного нормалізуючого перетворення Бокса-Кокса, щоб підвищити достовірність раннього оцінювання розміру цих застосунків.

Метод. Модель, довірчі інтервали та інтервали передбачення багатовимірної нелінійної регресії для оцінювання розміру веб-застосунків з відкритим кодом на PHP, створених із використанням фреймворку CakePHP, побудовані на основі багатовимірного нормалізуючого перетворення Бокса-Кокса для негаусівських даних за допомогою відповідних методів. Методи побудови моделей, рівнянь, довірчих інтервалів і інтервалів передбачення нелінійних регресій засновані на множинному нелінійному регресійному аналізі з використанням багатовимірних нормалізуючих перетворень. Ці методи дозволяють враховувати кореляцію між залежними та незалежними змінними у разі нормалізації багатовимірних негаусівських даних. Загалом, це призводить до зменшення середньої величини відносної похибки, ширини довірчих інтервалів і інтервалів передбачення в порівнянні нелінійними моделями, побудованими з використанням одновимірних нормалізуючих перетворень.

Результати. Проведено порівняння побудованої моделі з нелінійними регресійними моделями на основі десяткового логарифму та одновимірного перетворення Бокса-Кокса.

DOI 10.15588/1607-3274-2021-4-12

Висновки. Модель нелінійної регресії з трьома предикторами для оцінювання розміру веб-застосунків, створених за допомогою фреймворку CakePHP, побудована на основі чотиривимірного перетворення Бокса-Кокса. Ця модель, у порівнянні з іншими нелінійними регресійними моделями, має більший множинний коефіцієнт детермінації, менше значення середньої величини відносної похибки та менші ширини довірчих інтервалів та інтервалів передбачення. Перспективи подальших досліджень можуть включати застосування інших багатовимірних нормалізуючих перетворень та наборів даних для побудови нелінійних регресійних моделей для оцінювання розміру веб-додатків, створених за допомогою інших фреймворків.

КЛЮЧОВІ СЛОВА: оцінка розміру програмного забезпечення, веб-додаток, нелінійна регресійна модель, нормалізуюче перетворення, негаусівські дані.

УДК 004.412:519.237.5

НЕЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ РАЗМЕРА ВЕБ-ПРИЛОЖЕНИЙ, СОЗДАВАЕМЫХ С ИСПОЛЬЗОВАНИЕМ ФРЕЙМВОРКА CAKEPHP

Приходько С. Б. – д-р техн. наук, профессор, заведующий кафедрой программного обеспечения автоматизированных систем Национального университету кораблестроения имени адмирала Макарова, Николаев, Украина.

Шутко И. С. – аспирант кафедры программного обеспечения автоматизированных систем Национального университету кораблестроения имени адмирала Макарова, Николаев, Украина.

Приходько А. С. – студент кафедры программного обеспечения автоматизированных систем Национального университету кораблестроения имени адмирала Макарова, Николаев, Украина.

АННОТАЦИЯ

Актуальность. Проблема оценки размера программного обеспечения на ранней стадии программного проекта является важной, поскольку оценки размера программного обеспечения используется для прогнозирования трудоемкости разработки программного обеспечения, включая веб-приложения с открытым кодом на PHP, созданных с использованием фреймворка CakePHP. Объектом исследования является процесс оценки размера веб-приложений с открытым кодом на PHP, созданных с использованием фреймворка CakePHP. Предметом исследования является нелинейные регрессионные модели для оценки размера веб-приложений с открытым кодом на PHP, созданных с использованием фреймворка CakePHP.

Цель. Целью работы является построение нелинейной регрессионной модели с тремя предикторами для оценки размера веб-приложений, создаваемых с использованием фреймворка CakePHP на основе четырехмерного нормализующего преобразования Бокса-Кокса, чтобы повысить достоверность раннего оценивания размера этих приложений.

Метод. Модель, доверительные интервалы и интервалы предсказания многомерной нелинейной регрессии для оценки размера веб-приложений с открытым кодом на PHP, созданных с использованием фреймворка CakePHP, построены на основе многомерного нормализующего преобразования Бокса-Кокса для негауссовских данных с помощью соответствующих методов. Методы построения моделей, уравнений, доверительных интервалов и интервалов предсказания нелинейных регрессий основаны на множественном нелинейном регрессионном анализе с использованием многомерных нормализующих преобразований. Эти методы позволяют учитывать корреляцию между зависимыми и независимыми переменными в случае нормализации многомерных негауссовских данных. В общем, это приводит к уменьшению средней величины относительной погрешности, ширины доверительных интервалов и интервалов предсказания по сравнению нелинейными моделями, построенными с использованием одномерных нормализующих преобразований.

Результаты. Проведено сравнение построенной модели с нелинейными регрессионными моделями на основе десятичного логарифма и одномерного преобразования Бокса-Кокса.

Выводы. Модель нелинейной регрессии с тремя предикторами для оценки размера веб-приложений, созданных с помощью фреймворка CakePHP, построена на основе четырехмерного преобразования Бокса-Кокса. Эта модель, по сравнению с другими нелинейными регрессионными моделями, имеет больший множественный коэффициент детерминации, меньшее значение средней величины относительной погрешности и меньшие ширины доверительных интервалов и интервалов предсказания. Перспективы дальнейших исследований могут включать применение других многомерных нормализующих преобразований и наборов данных для построения нелинейных регрессионных моделей для оценки размера веб-приложений, созданных с помощью других фреймворков.

КЛЮЧЕВЫЕ СЛОВА: оценка размера программного обеспечения, веб-приложение, нелинейная регрессионная модель, нормализующее преобразование, негауссовские данные.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Boehm B.W. Software cost estimation with COCOMO II / [B. W. Boehm, C. Abts, A. W. Brown et al.]. – Upper Saddle River, NJ : Prentice Hall PTR, 2000. – 506 p.
2. 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
3. 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
4. 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
5. Zifen Y. An improved software size estimation method based on object-oriented approach / Y. Zifen // Electrical & Electronics Engineering : IEEE Symposium EEESYM’12, Kuala Lumpur, Malaysia, 24–27 June, 2012 : proceedings. – IEEE, 2012. – P. 615–617. DOI: 10.1109/EEESym.2012.6258733.

DOI 10.15588/1607-3274-2021-4-12

6. 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. – IEEE, 2016. – P. 1–6. DOI: 10.1109/JCSSE.2016.7748880
7. Prykhodko N.V. Constructing the non-linear regression models on the basis of multivariate normalizing transformations / N. V. Prykhodko, S. B. Prykhodko // Electronic modeling. – 2018. – Vol. 40, No. 6. – P. 101–110. DOI: 10.15407/emodel.40.06.101
8. Prykhodko N.V. The non-linear regression model to estimate the software size of open source Java-based systems / N. V. Prykhodko, S. B. Prykhodko // Radio Electronics, Computer Science, Control. – 2018. – No. 3 (46). – P. 158–166. DOI: 10.15588/1607-3274-2018-3-17
9. Приходько С.Б. Нелінійна регресійна модель для оцінювання розміру Web-застосунків, що створюються з використанням фреймворку Laravel / [С. Б. Приходько, Н. В. Приходько, М. В. Ворона та ін.] // Інформаційні технології та комп'ютерна інженерія. – 2021. – Том 50, № 1. – С. 115–121. DOI: 10.31649/1999-9941-2021-50-1-115-121
10. Nassif A. B. Software effort estimation from use case diagrams using nonlinear regression analysis / A. B. Nassif, M. AbuTalib, L. F. Capretz // Electrical and Computer Engineering : IEEE Canadian Conference CCECE'20, London, ON, Canada, 30 Aug.–2 Sept., 2020 : proceedings. – IEEE, 2020. – P. 1–4. DOI: 10.1109/CCECE47787.2020.9255712.
11. Montgomery D.C. Introduction to linear regression analysis / D. C. Montgomery, D. C. Peck, // 2nd edn. – New York: John Wiley & Sons, 1992. – 544 p.
12. Seber G.A.F. Linear regression analysis / G. A. F. Seber, A. J. Lee // 2nd edn. – New York: John Wiley & Sons, 2003. – 582 p.
13. Weisberg, S. Applied linear regression / S. Weisberg // 4th edn. – New York: John Wiley & Sons, 2013. – 368 p.
14. 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
15. 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
16. Ryan T.P. Modern regression methods / T. P. Ryan. – New York: John Wiley & Sons, 1997. – 529 p. DOI: 10.1002/9780470382806
17. Drapper N.R. Applied regression analysis / N. R. Drapper, H. Smith. – New York: John Wiley & Sons, 1998. – 736 p.
18. Johnson R.A. Applied multivariate statistical analysis / R. A. Johnson, D. W. Wichern. – Pearson Prentice Hall, 2007. – 800 p.
19. Chatterjee S. Handbook of regression analysis / S. Chatterjee, J. S. Simonoff. – New York : John Wiley & Sons, 2013. – 236 p. DOI: 10.1002/9781118532843
20. Prykhodko S. Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations / S. Prykhodko, N. Prykhodko // Mathematical Modeling and Simulation of Systems : 15th International Scientific-practical Conference MODS'2020, Chernihiv, Ukraine, June 29 – July 01, 2020 : selected papers. – Springer, Cham., 2021. – P. 166–174. – (Advances in Intelligent Systems and Computing, Vol. 1265). DOI: 10.1007/978-3-030-58124-4_16
21. Belsley D. A. Regression diagnostics: Identifying influential data and sources of collinearity / D. A. Belsley, E. Kuh, R. E. Welsch. – New York : John Wiley, 1980. – 300 p. DOI:10.1002/0471725153
22. Chatterjee S. Regression analysis by example / S. Chatterjee, B. Price. – New York : John Wiley & Son, 2012. – 424 p.
23. Olkin I. Multivariate analysis: Overview / I. Olkin, A. R. Sampson // N. J. Smelser, P. B. Baltes, Eds. International encyclopedia of social & behavioral sciences. 1st edn. – Elsevier, Pergamon, 2001. – P. 10240–10247.
24. Mardia K. V. Measures of multivariate skewness and kurtosis with applications / K. V. Mardia // Biometrika. – 1970. – Vol. 57. – P. 519–530. DOI: 10.1093/biomet/57.3.519
25. Mardia K.V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies / K. V. Mardia // Sankhya: The Indian Journal of Statistics, Series B (1960–2002). – 1974. – Vol. 36, Issue 2. – P. 115–128.
26. 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
27. Prykhodko S. Estimating the efforts of mobile application development in the planning phase using nonlinear regression analysis / S. Prykhodko, N. Prykhodko, K. Knyrik // Applied Computer Systems. – 2020. – Vol. 25, No. 2. – P. 172–179. DOI: 10.2478/acss-2020-0019
28. Campbell S.L. Modeling and simulation in Scilab/Scicos / S. L. Campbell, J.-P. Chancelier, R. Nikoukhah. – Springer, 2005. – 313 p.
29. A simulation study of the model evaluation criterion MMRE / [T. Foss, E. Stensrud, B. Kitchenham, I. Myrtveit] // IEEE Transactions on software engineering. – 2003. – Vol. 29, Issue 11. – P. 985–995. DOI: 10.1109/TSE.2003.1245300
30. Port D. Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research / D. Port, M. Korte // Empirical Software Engineering and Measurement : the 2nd ACM-IEEE International Symposium ESEM, Kaiserslautern, Germany, October, 2008 : proceedings. – New York : ACM, 2008. – P. 51–60.

DOI 10.15588/1607-3274-2021-4-12