CONSTRUCTING THE NON-LINEAR REGRESSION EQUATION TO ESTIMATE THE SOFTWARE SIZE OF OPEN SOURCE PHP-BASED INFORMATION SYSTEMS

UDC 004.412:519.237.5

Sergiy PRYKHODKO
Dr.Sc., Professor at the Department of Software of Automated Systems,
the Head of Department, Admiral Makarov National University of Shipbuilding.
Scientific interests: mathematical modeling of random variables and processes in information technologies.

Natalia PRYKHODKO
PhD, Associate Professor at the Finance Department, Admiral Makarov National University of Shipbuilding.
Scientific interests: mathematical modeling of random variables and processes in information technologies.

Tatyana SMYKODUB
Senior Teacher at the Department of Software of Automated Systems, Admiral Makarov National University of Shipbuilding.
Scientific interests: mathematical modeling of random variables and processes in information technologies.

Alexander SPINOV
a student of master's degree program in specialty 121 "Software Engineering",
the Department of Software of Automated Systems, Admiral Makarov National University of Shipbuilding.
Scientific interests: mathematical modeling of random variables in information technologies.

INTRODUCTION

PHP is a free programming language used primarily in information systems and web applications. Software size is one of the most important internal metrics of software including software of PHP-based open-source information systems. The information obtained from estimating the software size is useful for predicting the software development effort by such models as COCOMO 81, COCOMO II and COCOMO 2000. The papers [1, 2] proposed the linear regression equations for estimating the software size of some programming languages, such as VBA, PHP, Java and C++. The proposed equations are constructed by multiple linear regression analysis on the basis of the metrics that can be measured from 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 equations including for estimating the software size of PHP-based open-source information systems.

A normalizing transformation is often a good way to build the equations, confidence and prediction intervals of multiply non-linear regressions [3-8]. According to [4] transformations are used for essentially four purposes, two of which are: first, to obtain approximate normality for the distribution of the error term (residuals), second, 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, which do not take into account the correlation between random variables in the case of normalization of multivariate non-Gaussian data. This leads to the need to use the multivariate normalizing transformations.

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

In this article, we build the equation, confidence and prediction intervals of multivariate non-linear regression for estimating the software size of open-source PHP-based systems on the basis of the Johnson multivariate normalizing transformation (the Johnson normalizing translation) with the help of appropriate techniques proposed in [8, 9].

The techniques. The techniques to build the equations, confidence and prediction intervals of non-linear regressions are based on the multiple non-linear regression analysis using the multivariate normalizing transformations. A 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

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

and the inverse transformation for (1)

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

The linear regression equation for normalized data according to (1) will have the form [4]

$$\hat{Z}_Y=\overline{Z}_Y+\left(\mathbf{Z}_X^+\right)\hat{\mathbf{b}}, \tag{3}$$

where $\hat{Z}_Y$ is prediction linear regression equation result for values of components of vector $\mathbf{z}_X=\{Z_1,Z_2,\ldots,Z_k\}$; $\mathbf{Z}_X^+$ is the matrix of centered regressors that contains the values $Z_{1_i}-\overline{Z}_1$, $Z_{2_i}-\overline{Z}_2$, $\ldots$, $Z_{k_i}-\overline{Z}_k$; $\hat{\mathbf{b}}$ is estimator for vector of linear regression equation parameters, $\mathbf{b}=\{b_1,b_2,\ldots,b_k\}^T$.

The non-linear regression equation will have the form

$$\hat{Y}=\psi_Y^{-1}\left[\overline{Z}_Y+\left(\mathbf{Z}_X^+\right)\hat{\mathbf{b}}\right]. \tag{4}$$

where ψ_Y is the first component of vector $\boldsymbol{\psi}=\{\psi_Y,\psi_1,\psi_2,\ldots,\psi_k\}^T$.

The technique to build a confidence interval of non-linear regression is based on transformations (1) and (2), equation (3) and a confidence interval of linear regression for normalized data

$$\hat{Z}_Y\pm t_{\alpha/2,\nu}S_{Z_Y}\left\{\frac{1}{N}+\left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}. \tag{5}$$

where $t_{\alpha/2,\nu}$ is a quantile of student's *t*-distribution with ν degrees of freedom and $\alpha/2$ significance level; $\left(\mathbf{z}_X^+\right)^T$ is one of the rows of $\mathbf{Z}_X^+$; $S_{Z_Y}^2=\frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i}-\hat{Z}_{Y_i}\right)^2$, $\nu=N-k-1$; $\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+$ is the $k\times k$ matrix

$$\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+=\begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \cdots & S_{Z_1Z_k}\\ S_{Z_1Z_2} & S_{Z_2Z_2} & \cdots & S_{Z_2Z_k}\\ \cdots & \cdots & \cdots & \cdots\\ S_{Z_1Z_k} & S_{Z_2Z_k} & \cdots & S_{Z_kZ_k}\end{pmatrix},$$

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

The confidence interval for non-linear regression is built on the basis of the interval (5) and inverse transformation (2)

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

The technique to build a prediction interval is based on multivariate transformation (1), the inverse transformation

(2), linear regression equation for normalized data (3) and a prediction interval for normalized data

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

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

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{ 1 + \frac{1}{N} + \left(\mathbf{z}_X^+\right)^T \left[\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+ \right]^{-1} \left(\mathbf{z}_X^+\right) \right\}^{1/2} \right) . \quad (8)$$

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

$$\mathbf{T} = \boldsymbol{\gamma} + \boldsymbol{\eta}\mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{P} - \varphi)\right] \sim N_m(\mathbf{0}_m, \boldsymbol{\Sigma}), \quad (9)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix; $m = k + 1$; $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, φ and $\boldsymbol{\lambda}$ are parameters of translation (9); $\boldsymbol{\gamma} = (\gamma_Y, \gamma_1, \gamma_2, \ldots, \gamma_k)^T$; $\boldsymbol{\eta} = \operatorname{diag}(\eta_Y, \eta_1, \eta_2, \ldots, \eta_k)$; $\boldsymbol{\lambda} = \operatorname{diag}(\lambda_Y, \lambda_1, \lambda_2, \ldots, \lambda_k)$; $\varphi = (\varphi_Y, \varphi_1, \varphi_2, \ldots, \varphi_k)^T$; $\mathbf{h}[(y_Y, y_1, \ldots, y_k)] = \{h_Y(y_Y), h_1(y_1), \ldots, h_k(y_k)\}^T$; $h_i(\cdot)$ is one of the translation functions

$$h = \begin{cases} \ln(y), & \text{for } S_L \text{ (log normal) family;} \\ \ln[y/(1-y)], & \text{for } S_B \text{ (bounded) family;} \\ \operatorname{Arsh}(y), & \text{for } S_U \text{ (unbounded) family;} \\ y & \text{for } S_N \text{ (normal) family.} \end{cases} \quad (10)$$

Here $y = (X - \varphi)/\lambda$; $\operatorname{Arsh}(y) = \ln\left(y + \sqrt{y^2 + 1}\right)$. 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 PHP-based systems. The equation, confidence and prediction intervals of non-linear regression to estimate the software size of open-source PHP-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 average number of attributes per class X_3, the total number of classes X_1 and the total number of relationships X_2 in conceptual data model from 32 information systems developed using the PHP programming language with HTML and SQL. Table I contains the data from [1] on four metrics of software for 32 open-source PHP-based systems.

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

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: $\hat{\gamma}_Y = 9.63091$, $\hat{\gamma}_1 = 15.5355$, $\hat{\gamma}_2 = 25.4294$, $\hat{\gamma}_3 = 0.72801$, $\hat{\eta}_Y = 1.05243$, $\hat{\eta}_1 = 1.58306$, $\hat{\eta}_2 = 2.54714$, $\hat{\eta}_3 = 0.54312$, $\hat{\varphi}_Y = \text{-}1.4568$, $\hat{\varphi}_1 = \text{-}1{,}8884$, $\hat{\varphi}_2 = \text{-}6{,}9746$, $\hat{\varphi}_3 = 3.2925$, $\hat{\lambda}_Y = 153102.605$, $\hat{\lambda}_1 = 243051.0$, $\hat{\lambda}_2 = 311229.5$ and $\hat{\lambda}_3 = 13.900$. The sample covariance matrix S_N of the $\mathbf{T}$ is used as the approximate moment-matching estimator of $\boldsymbol{\Sigma}$

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

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

$$\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 . \quad (11)$$

Parameters of the linear regression equation (11) were estimated by the least square method. Estimators for parameters of the equation (11) are such: $\hat{b}_0 = 1.02 \cdot 10^{-5}$, $\hat{b}_1 = 0.56085$, $\hat{b}_2 = 0.42491$, $\hat{b}_3 = 0.05846$.

Table I

The data and prediction result by regression equations for 32 open-source PHP-based systems

No	Y	X_1	X_2	X_3	Linear regression		Non-linear regression			
							univariate transformation		multivariate transformation	
					$\hat{Y}$	RME	$\hat{Y}$	RME	$\hat{Y}$	RME
1	3.038	5	2	10.6	3.237	0.0656	4.675	0.5388	4.550	0.4976
2	22.599	17	7	7	24.142	0.0683	19.965	0.1166	19.990	0.1154
3	32.243	21	13	4.524	37.524	0.1638	32.098	0.0045	33.535	0.0401
4	16.164	13	11	7.077	25.916	0.6033	23.171	0.4335	21.292	0.3173
5	83.862	35	24	6.571	74.624	0.1102	80.265	0.0429	83.618	0.0029
6	24.22	13	9	8.077	23.224	0.0411	20.524	0.1526	18.901	0.2196
7	63.929	35	19	8.029	67.215	0.0514	65.913	0.0310	70.647	0.1051
8	2.543	5	3	9.4	4.127	0.6228	5.789	1.2764	5.169	1.0328
9	6.697	5	5	7	5.906	0.1181	7.353	0.0980	6.356	0.0509
10	55.537	25	14	8.64	46.843	0.1565	42.098	0.2420	43.126	0.2235
11	55.752	39	10	9.077	57.814	0.0370	67.070	0.2030	49.823	0.1064
12	62.602	30	17	7	56.995	0.0896	53.497	0.1454	56.651	0.0951
13	67.111	23	22	14.957	61.856	0.0783	65.500	0.0240	60.617	0.0968
14	2.552	3	1	8.333	-2.395	1.9384	2.202	0.1370	2.447	0.0412
15	12.17	10	5	3.7	9.959	0.1816	9.693	0.2035	10.029	0.1759
16	12.757	13	9	5	21.218	0.6632	18.682	0.4644	18.105	0.4192
17	5.695	7	3	8.429	5.976	0.0493	7.083	0.2438	6.687	0.1743
18	7.744	9	6	9.222	13.991	0.8067	12.911	0.6673	11.301	0.4593
19	7.514	4	1	8	-1.371	1.1825	2.496	0.6678	3.096	0.5880
20	11.054	9	9	3.667	15.385	0.3918	13.301	0.2032	12.850	0.1625
21	29.77	17	15	3.412	35.179	0.1817	27.321	0.0823	29.061	0.0238
22	11.653	9	8	8.778	17.045	0.4627	15.461	0.3268	13.268	0.1386
23	6.847	5	4	3.6	2.017	0.7054	5.435	0.2062	5.112	0.2534
24	13.389	7	5	11.714	11.462	0.1440	10.367	0.2257	8.661	0.3531
25	14.45	12	6	16.583	22.513	0.5580	20.191	0.3973	15.888	0.0995
26	4.414	6	3	3.667	1.630	0.6307	5.318	0.2048	5.260	0.1916
27	2.102	3	1	3.333	-5.655	3.6902	2.142	0.0192	1.873	0.1090
28	42.819	20	18	3.5	43.975	0.0270	37.967	0.1133	38.631	0.0978
29	4.077	4	2	9	0.953	0.7662	3.892	0.0454	3.732	0.0846
30	57.408	33	14	9.242	57.164	0.0043	53.121	0.0747	54.381	0.0527
31	7.428	7	3	7	5.044	0.3209	6.861	0.0764	6.571	0.1154
32	8.947	15	5	4	16.360	0.8285	12.934	0.4456	14.258	0.5936

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

$$\hat{Y} = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y - \hat{\gamma}_Y\right)/\hat{\eta}_Y} \right]^{-1} . \quad (12)$$

where $\hat{Z}_Y$ is prediction result by the equation (11), $Z_j = \gamma_j + \eta_j \ln \frac{X_j - \varphi_j}{\varphi_j + \lambda_j - X_j}$, $\varphi_j < X_j < \varphi_j + \lambda_j$, $j = 1,2,3$.

The prediction results by equation (12) for values of components of vector $\mathbf{X} = \{X_1, X_2, X_3\}$ from Table I and values of magnitude of relative error MRE are shown in the Table I for two cases: the Johnson univariate and multivariate normalizing transformations. Table I also contains the prediction results by linear regression equation from [1] for values of components of vector **X** from Table I and MRE values. Note the prediction results by linear regression equation from [1] are negative for the three rows of data: 14, 19 and 27. All prediction results by non-linear regression equation (12) are positive.

For univariate normalizing transformations (10) of S_B family the estimators for parameters are such: $\hat{\gamma}_Y = 0.77502$, $\hat{\gamma}_1 = 0.59473$, $\hat{\gamma}_2 = 0.57140$, $\hat{\gamma}_3 = 0.68734$, $\hat{\eta}_Y = 0.44395$, $\hat{\eta}_1 = 0.48171$, $\hat{\eta}_2 = 0.49553$, $\hat{\eta}_3 = 0.51970$, $\hat{\varphi}_Y = 2.063$, $\hat{\varphi}_1 = 2.900$, $\hat{\varphi}_2 = 0.900$, $\hat{\varphi}_3 = 3.304$, $\hat{\lambda}_Y = 83.059$, $\hat{\lambda}_1 = 36.695$, $\hat{\lambda}_2 = 23.525$ and $\hat{\lambda}_3 = 13.660$. In the case of univariate normalizing transformations the estimators for parameters of the equation (11) are such: $\hat{b}_0 = 3.11 \cdot 10^{-7}$, $\hat{b}_1 = 0.43519$, $\hat{b}_2 = 0.52239$ and $\hat{b}_3 = 0.08546$.

Also the non-linear regression equation (4) is built by the decimal logarithm transformation

$$\hat{Y} = 10^{b_0} X_1^{b_1} X_2^{b_2} X_3^{b_3}, \tag{13}$$

where the estimators for parameters of the equation (13) are such: $\hat{b}_0 = \text{-}0.26161$, $\hat{b}_1 = 0.99151$, $\hat{b}_2 = 0.33232$ and $\hat{b}_3 = 0.13777$.

The values of multiple coefficient of determination R^2, mean magnitude of relative error MMRE and percentage of prediction PRED(0.25) equal respectively 0.9491, 0.4919 and 0.5313 for linear regression equation from [1], and equal respectively 0.9375, 0.2455 and 0.625 for the equation (13). The values of R^2, MMRE and PRED(0.25) are better for the equation (12), in comparison with both the equation from [1] and equation (13), and are 0.9692, 0.2199 and 0.7188 for the Johnson multivariate transformation and equal 0.9591, 0.2535 and 0.7188 for the Johnson univariate transformation respectively. 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 MMRE indicate that only the values for equation (12) on the basis of the Johnson multivariate normalizing transformation and the decimal logarithm transformation are less than 0.25. Although all values of PRED(0.25) are less than 0.75 nevertheless the values are greater for equation (12) with estimators of parameters for the Johnson transformations, both multivariate and univariate.

The confidence and prediction intervals of non-linear regression are defined by (6) and (8) respectively for the data from Table I. Table II 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 from [1] are negative for the seven rows of data: 1, 14, 19, 23, 26, 27 and 29. The upper bound for the data row 27 is negative too. All the lower and upper 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 from [1] for the twenty rows of data: 1, 6, 8, 9, 14-20, 22-27, 29, 31 and 32. 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 the twenty-seven rows of data: 1-4, 6-12, 15-26, 28, 30-32. And ones are less than following the Johnson univariate transformation for the twenty-five rows of data: 1-4, 6, 8-11, 15-18, 20-26, 28-32. Approximately the same results are obtained for the prediction intervals of regressions.

Table III 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 from [1] are negative for the thirteen rows of data: 1, 8, 9, 14, 15, 17, 19, 23, 24, 26, 27, 29, 31. 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 from [1] for the twenty rows of data: 1, 6, 8, 9, 14-20, 22-27, 29, 31 and 32. 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 the twenty-nine rows of data: 1-13, 15-18, 20-26, 28-32. And ones are less than following the Johnson univariate transformation for the twenty-three rows of data: 1-4, 6, 8-10, 15-18, 20-26, 28, 29, 31 and 32.

Table II

Bounds of the confidence intervals

No	Y	Bounds for linear regression		Bounds for non-linear regressions					
				univariate transformations				Johnson multivariate transformation	
				decimal logarithm		Johnson			
		LB	UB	LB	UB	LB	UB	LB	UB
1	3.038	-0.402	6.877	3.725	5.947	3.673	6.267	3.655	5.601
2	22.599	21.413	26.871	18.933	27.172	15.473	25.455	16.856	23.660
3	32.243	34.344	40.704	26.415	39.621	24.791	40.266	29.157	38.539
4	16.164	23.172	28.660	16.855	24.285	17.982	29.365	18.542	24.421
5	83.862	69.187	80.062	55.076	87.173	74.107	83.078	68.095	102.603
6	24.22	21.015	25.433	16.557	22.438	16.129	25.819	16.935	21.075
7	63.929	62.690	71.740	52.309	83.045	56.434	72.961	59.182	84.277
8	2.543	1.013	7.241	4.288	6.544	4.484	7.748	4.260	6.223
9	6.697	3.084	8.728	4.569	7.951	5.456	10.203	5.135	7.803
10	55.537	43.863	49.824	35.573	52.195	33.947	50.366	37.344	49.768
11	55.752	49.560	66.068	41.448	87.698	42.891	79.359	36.261	68.257
12	62.602	53.265	60.725	43.512	65.766	44.275	61.787	48.298	66.405
13	67.111	54.146	69.566	35.747	69.156	49.897	75.572	45.429	80.723
14	2.552	-5.673	0.883	1.639	2.897	2.125	2.375	1.806	3.214
15	12.17	6.609	13.309	8.838	13.632	6.979	13.684	8.320	12.037
16	12.757	18.574	23.862	15.339	21.222	14.673	23.576	16.253	20.151
17	5.695	3.165	8.787	6.139	8.644	5.548	9.233	5.655	7.870
18	7.744	11.381	16.601	10.039	14.139	9.902	16.849	9.960	12.799
19	7.514	-4.587	1.845	2.106	3.945	2.253	3.046	2.391	3.930
20	11.054	11.684	19.085	9.137	15.776	9.186	19.255	10.550	15.591
21	29.77	30.767	39.591	20.246	34.593	16.796	41.072	22.899	36.782
22	11.653	14.250	19.840	10.430	16.253	11.581	20.525	11.347	15.478
23	6.847	-1.579	5.613	3.900	6.687	4.071	7.662	4.063	6.360
24	13.389	7.648	15.276	7.099	11.493	7.462	14.583	7.262	10.286
25	14.45	16.199	28.828	13.006	22.695	10.971	34.746	12.197	20.576
26	4.414	-1.967	5.227	4.421	7.029	4.083	7.261	4.241	6.461
27	2.102	-9.730	-1.580	1.360	2.712	2.092	2.281	1.087	2.902
28	42.819	38.873	49.077	25.212	43.588	25.181	51.940	30.491	48.845
29	4.077	-2.236	4.142	2.947	4.616	3.177	5.048	2.959	4.640
30	57.408	52.335	61.993	44.400	73.879	41.599	63.278	44.841	65.885
31	7.428	2.314	7.774	6.042	8.344	5.463	8.784	5.585	7.695
32	8.947	12.515	20.205	12.503	22.001	9.080	18.449	11.119	18.180

Table III

Bounds of the prediction intervals

No	Y	Bounds for linear regression		Bounds for non-linear regressions					
				univariate transformations				Johnson multivariate transformation	
				decimal logarithm		Johnson			
		LB	UB	LB	UB	LB	UB	LB	UB
1	3.038	-8.886	15.361	2.264	9.787	2.507	15.664	2.053	8.822
2	22.599	12.260	36.024	11.075	46.451	5.800	53.204	11.088	35.207
3	32.243	25.530	49.517	15.704	66.644	9.341	65.987	19.149	57.962
4	16.164	14.031	37.802	9.874	41.455	6.642	57.342	11.955	37.129
5	83.862	61.845	87.403	33.367	143.886	59.920	84.392	47.603	146.045
6	24.22	11.451	34.998	9.475	39.211	5.956	53.906	10.617	32.866
7	63.929	54.797	79.633	31.724	136.931	31.210	81.247	40.528	122.355
8	2.543	-7.849	16.103	2.565	10.940	2.713	20.215	2.431	9.838
9	6.697	-5.998	17.810	2.856	12.722	2.996	26.099	3.097	11.949
10	55.537	34.901	58.785	20.980	88.499	13.397	72.304	24.761	74.346
11	55.752	43.606	72.022	27.405	132.638	26.251	82.571	26.759	91.726
12	62.602	44.844	69.146	25.940	110.319	19.861	77.358	32.563	97.782
13	67.111	47.957	75.755	23.063	107.190	28.542	81.562	33.153	109.857
14	2.552	-14.415	9.625	1.030	4.613	2.084	2.994	0.811	5.262
15	12.17	-2.080	21.999	5.307	22.704	3.441	33.425	5.255	18.197
16	12.757	9.355	33.081	8.849	36.788	5.492	51.258	10.150	31.513
17	5.695	-5.925	17.877	3.565	14.884	2.964	24.822	3.336	12.381
18	7.744	2.136	25.846	5.831	24.343	4.145	40.894	6.095	20.095
19	7.514	-13.374	10.632	1.346	6.172	2.127	4.916	1.198	6.351
20	11.054	3.243	27.527	5.697	25.303	4.154	42.480	6.867	23.133
21	29.77	22.801	47.556	12.581	55.670	7.324	63.400	15.978	51.960
22	11.653	5.148	28.943	6.285	26.973	4.693	46.152	7.200	23.590
23	6.847	-10.093	14.128	2.426	10.749	2.635	19.103	2.367	9.829
24	13.389	-0.715	23.638	4.334	18.826	3.576	35.238	4.477	15.796
25	14.45	9.337	35.689	8.136	36.280	5.323	56.560	8.396	29.076
26	4.414	-10.481	13.741	2.683	11.585	2.621	18.450	2.464	10.048
27	2.102	-17.916	6.606	0.885	4.168	2.073	2.648	0.410	4.484
28	42.819	31.335	56.615	15.725	69.883	10.895	70.978	21.432	68.748
29	4.077	-11.043	12.949	1.779	7.647	2.371	12.014	1.575	7.423
30	57.408	44.632	69.696	27.354	119.916	19.170	77.441	30.902	94.883
31	7.428	-6.838	16.926	3.483	14.475	2.926	23.959	3.273	12.168
32	8.947	4.173	28.547	7.842	35.078	4.090	41.560	7.530	26.021

Following [13] multivariate kurtosis β_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]

$$\hat{\beta}_2 = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)^T S_N^{-1}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\right\}^2 . \qquad (14)$$

In our case, in the formula (14), the vectors **Z** and $\overline{\mathbf{Z}}$ should be replaced by the vectors **P** and $\overline{\mathbf{P}}$ or **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 condition for multivariate normality. In our case $\beta_2 = 24$. The estimators of multivariate kurtosis equal 28.66, 23.87, 37.29 and 23.08 for the data from Table I, 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. Note that in our case, 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.

CONCLUSIONS

The non-linear regression equation to estimate the software size of open-source PHP-based information systems is improved on the basis of the Johnson multivariate transformation for S_B family. This equation, in comparison with other regression equations (both linear and nonlinear), has a larger multiple coefficient of determination and a smaller value of MMRE.

When building the equations, confidence and prediction intervals of non-linear regressions for multivariate non-Gaussian data to estimate the software size of open-source PHP-based information systems, one should use multivariate transformations.

Usually poor normalization of multivariate non-Gaussian data or application of univariate transformations instead of multivariate ones to normalize such data may lead to increase of width of the confidence and prediction intervals of regressions, both linear and non-linear, to estimate the software size of open-source PHP-based information systems.

In the future, we intend to try other multivariate normalizing transformations and non-Gaussian data sets.

REFERENCES

1. Hee Beng Kuan Tan, Yuan Zhao, and Hongyu Zhang, "Estimating LOC for information systems from their conceptual data models", in Proceedings of the 28th International Conference on Software Engineering (ICSE '06), May 20-28, 2006, Shanghai, China, pp. 321-330.
2. Matinee Kiewkanya, and Suttipong Surak, "Constructing C++ software size estimation model from class diagram", in 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), July 13-15, 2016, Khon Kaen, Thailand, pp. 1-6.
3. D.M. Bates and D.G. Watts. Nonlinear Regression Analysis and Its Applications. Wiley, 1988, 384 p.
4. T.P. Ryan. Modern regression methods. Wiley, 1997, 529 p.
5. G.A.F. Seber and C.J. Wild. Nonlinear Regression. John Wiley & Sons, Inc., 2003, 792 p.
6. R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, 800 p.
7. S Chatterjee and J.S. Simonoff. Handbook of Regression Analysis. John Wiley & Sons, Inc., 2013, 236 p.
8. S. B. Prykhodko, "Developing the software defect prediction models using regression analysis based on normalizing transformations", in Abstracts of the Research and Practice Seminar on Modern Problems in Testing of the Applied Software (PTTAS-2016), May 25-26, 2016, Poltava, Ukraine, pp. 6-7.
9. S. B. Prykhodko, N. V. Prykhodko, and K. S. Pugachenko, "Building the equations, confidence and prediction intervals of non-linear regressions on the basis of multivariate normalizing transformations", in Materials of the II International Scientific and Practical Conference "Applied Scientific and Technical Research", Ivano-Frankivsk, Ukraine, April 3-5, 2018, p. 16.
10. P.M. Stanfield, J.R. Wilson, G.A. Mirka, N.F. Glasscock, J.P. Psihogios, and J.R. Davis, "Multivariate input modeling with Johnson distributions", in Proceedings of the 28th Winter simulation conference WSC'96, December 8-11, 1996, Coronado, CA, USA, ed. S.Andradyttir, K.J.Healy, D.H.Withers, and B.L.Nelson, IEEE Computer Society Washington, DC, USA, 1996, pp. 1457-1464.
11. S. Prykhodko, N. Prykhodko, L. Makarova, and K. Pugachenko, "Detecting Outliers in Multivariate Non-Gaussian Data on the basis of Normalizing Transformations", in Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) «Celebrating 25 Years of IEEE Ukraine Section», May 29 – June 2, 2017, Kyiv, Ukraine, 2017, pp. 846-849.
12. S. Prykhodko, N. Prykhodko, L. Makarova, and A. Pukhalevych, "Application of the Squared Mahalanobis Distance for Detecting Outliers in Multivariate Non-Gaussian Data", in Proceedings of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, February 20–24, 2018, pp. 962-965.
13. K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications", Biometrika, 57, 1970, pp. 519–530.

Рецензент: д.т.н., проф. Ходаков В.Є.
Херсонський національний технічний університет