UDC 519.237.5

Constructing the nonlinear regression equations based on multivariate normalizing transformations

N.V. Prykhodko, S.B. Prykhodko
Admiral Makarov National University of Shipbuilding, Heroes of Ukraine Ave., 9, Nikolayev, 54025, Ukraine
e-mail: sergiy.prykhodko@nuos.edu.ua

In the paper we consider the techniques to construct the equations, confidence and prediction intervals of nonlinear regressions on the basis of multivariate normalizing transformations for non-Gaussian data. We demonstrate that the poor normalization of multivariate non-Gaussian data using the univariate transformations leads to an expansion of the confidence and prediction intervals of non-linear regression for a larger number of data rows compared to the multivariate normalizing transformation.

Keywords: *non-linear regression equation, confidence interval, prediction interval, normalizing transformation, multivariate non-Gaussian data.*

В статті розглядаються методи побудови рівнянь, довірчих інтервалів та інтервалів передбачення нелінійних регресій на основі багатовимірних нормалізуючих перетворень для негаусовських даних. У якості прикладу побудовано нелінійне регресійне рівняння для оцінювання розміру програмного забезпечення інформаційних систем з відкритим кодом на PHP із застосуванням багатовимірного нормалізуючого перетворення Джонсона для сімейства S_B. Це рівняння отримано за вибіркою чотиривимірних негаусовських даних: фактичний розмір програмного забезпечення у тисячах рядків коду, загальна кількість класів, загальна кількість зв'язків та середня кількість атрибутів на клас у концептуальній моделі даних з 32 інформаційних систем, розроблених з використанням мови програмування PHP. Попередньо зазначені дані були перевірені на наявність викидів із використанням квадрату відстані Махаланобиса (Mahalanobis): для рівня значимості, що дорівнює 0,005, викиди відсутні. Гіпотезу про багатовимірну нормальність було перевірено за критерієм квадрату відстані Махаланобиса. Побудоване нелінійне рівняння у порівнянні з іншими регресійними рівняннями (як лінійними, так і нелінійними, які отримані за допомогою одновимірних нормалізуючих перетворень Джонсона та десяткового логарифму) має більший множинний коефіцієнт детермінації і менше значення середньої величини відносної похибки. Продемонстровано, що погана нормалізація багатовимірних негаусовських даних за допомогою одновимірних перетворень або її відсутність призводить до збільшення ширини довірчих інтервалів та інтервалів передбачення як нелінійної так і лінійної регресії для більшої кількості рядків даних у порівнянні з багатовимірним нормалізуючим перетворенням.

Ключові слова: *нелінійне рівняння регресії, довірчий інтервал, інтервал передбачення, нормалізуюче перетворення, багатовимірні негаусовські дані.*

1 Introduction

A normalizing transformation is a good way to construct equations, confidence and prediction intervals of non-linear regressions [1-5], and it is often used in information technology, software engineering, biometry, ecology, finance, etc. According to [3] the transformations are mainly used for four purposes, two of which are: the first – to obtain approximate normality for the distribution of the residuals, the second – to transform the dependent and independent random variables in such a way that the strength of the linear relationship between new variables (normalized variables) is better than the linear relationship between the response and the predictor (or predictors) without transformation. Well-known techniques to construct the equations, confidence and prediction intervals of nonlinear regressions are based on the univariate normalizing transformations, which do not take into account the correlation between dependent and independent variables when multivariate non-Gaussian data is normalized. Therefore the multivariate normalizing transformations needs to be applied.

2 Unsolved problems and objectives of the paper

Well-known techniques for constructing the non-linear regression equations are based on the univariate normalizing transformations (such as, the decimal logarithm, the Box-Cox transformation), which do not take into account the correlation between the dependent and independent random variables in the case of normalization of multivariate non-Gaussian data. Application of such univariate normalizing transformations for building the nonlinear regression equations does not always lead to good multivariate normality and linear relationship between normalized variables. This demands the usage of the multivariate normalizing transformations. The objective of the paper is to consider techniques for constructing the equations, confidence and prediction intervals of multivariate nonlinear regressions on the basis of multivariate normalizing transformations. The nonlinear regression prediction results obtained by constructing the equations should be better in comparison with other nonlinear regression equations based on univariate normalizing transformations, primarily on such

standard evaluations as the multiple coefficient of determination and mean magnitude of relative error. Application of multivariate normalizing transformations should lead to a narrowing of confidence and prediction intervals of nonlinear regressions for a larger number of data rows compared to the univariate normalizing transformations.

3 Problem statement

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 $\mathbf{T} = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$

$$\mathbf{T} = \boldsymbol{\psi}(\mathbf{P}) \tag{3.1}$$

and the inverse transformation for (3.1)

$$\mathbf{P} = \boldsymbol{\psi}^{-1}(\mathbf{T}). \tag{3.2}$$

Here $\boldsymbol{\psi}$ is the vector of normalizing transformation, $\boldsymbol{\psi} = \{\psi_Y, \psi_1, \psi_2, \ldots, \psi_k\}^T$. It is required to build the nonlinear regression equation in the form $Y = Y(X_1, X_2, \ldots, X_k)$ on the basis of the transformations (3.1) and (3.2).

4 The techniques

The techniques to construct the equations, confidence and prediction intervals of nonlinear regressions are based on the nonlinear regression analysis using the multivariate normalizing transformations and they consist of three steps [5]. For the first step, a set of multivariate non-Gaussian data is normalized using a bijective multivariate normalizing transformation (3.1). 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 nonlinear 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 transformations (3.1) and (3.2).

The linear regression equation for normalized data will have the form [3]

$$\hat{Z}_Y = \overline{Z}_Y + \left(\mathbf{Z}_X^+\right)\hat{\mathbf{b}}, \tag{4.1}$$

where $\hat{Z}_Y$ is a prediction result obtained by linear regression equation 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 $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$; $\hat{\mathbf{b}}$ is estimator for vector of parameters of equation (4.1), $\mathbf{b} = \{b_1, b_2, \ldots, b_k\}^T$.

The nonlinear regression equation will be

$$Y = \psi_Y^{-1}\left[\overline{Z}_Y + \left(\mathbf{Z}_X^+\right)\hat{\mathbf{b}}\right]. \tag{4.2}$$

The technique to construct a confidence interval of nonlinear regression is based on transformations (3.1) and (3.2), linear regression equation (4.1) and a confidence interval for normalized data. The confidence interval of nonlinear regression is

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{\frac{1}{N} + \left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right), \tag{4.3}$$

where $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$, $\nu = N - k - 1$; $t_{\alpha/2,\nu}$ is a quantile of the Student *t*-distribution with ν degrees of freedom and $\alpha/2$ significance level; $\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+$ is the $k \times k$ matrix

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

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

The technique to construct a prediction interval of nonlinear regression is based on transformations (3.1) and (3.2), linear regression equation (4.1) and a prediction interval for normalized data. The prediction interval of non-linear regression is

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{\frac{1}{N} + \left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right). \quad (4.4)$$

5 Examples

We consider the examples of constructing the equations, confidence and prediction intervals of nonlinear regressions for multivariate non-Gaussian data for two cases: univariate and multivariate normalizing transformations. Table 1 contains the data [6] on metrics of software for open-source PHP-based information systems.

Table 1. The data and prediction results by regression equations.

i	Y in KLOC	X_1	X_2	X_3	prediction results by regressions			
					linear	non-linear regressions		
						decimal logarithm	**Johnson univariate**	**Johnson multivariate**
1	3.038	5	2	10.6	3.237	4.707	4.675	4.283
2	22.599	17	7	7	24.142	22.681	19.965	21.048
3	32.243	21	13	4.524	37.524	32.351	32.098	34.906
4	16.164	13	11	7.077	25.916	20.232	23.171	23.191
5	83.862	35	24	6.571	74.624	69.290	80.265	76.393
6	24.22	13	9	8.077	23.224	19.275	20.524	20.495
7	63.929	35	19	8.029	67.215	65.909	65.913	67.297
8	2.543	5	3	9.4	4.127	5.297	5.789	5.029
9	6.697	5	5	7	5.906	6.028	7.353	6.223
10	55.537	25	14	8.64	46.843	43.089	42.098	45.209
11	55.752	39	10	9.077	57.814	60.290	67.070	56.644
12	62.602	30	17	7	56.995	53.494	53.497	56.727
13	67.111	23	22	14.957	61.856	49.720	65.500	63.792
14	2.552	3	1	8.333	-2.395	2.179	2.202	2.324
15	12.17	10	5	3.7	9.959	10.977	9.693	9.659
16	12.757	13	9	5	21.218	18.042	18.682	19.002
17	5.695	7	3	8.429	5.976	7.285	7.083	6.520
18	7.744	9	6	9.222	13.991	11.914	12.911	11.988
19	7.514	4	1	8	-1.371	2.882	2.496	2.820
20	11.054	9	9	3.667	15.385	12.006	13.301	12.884
21	29.77	17	15	3.412	35.179	26.465	27.321	29.362
22	11.653	9	8	8.778	17.045	13.020	15.461	14.338
23	6.847	5	4	3.6	2.017	5.107	5.435	4.850
24	13.389	7	5	11.714	11.462	9.033	10.367	9.102
25	14.45	12	6	16.583	22.513	17.181	20.191	18.741
26	4.414	6	3	3.667	1.630	5.575	5.318	4.966
27	2.102	3	1	3.333	-5.655	1.921	2.142	2.168
28	42.819	20	18	3.5	43.975	33.150	37.967	40.170
29	4.077	4	2	9	0.953	3.688	3.892	3.508
30	57.408	33	14	9.242	57.164	57.273	53.121	55.910
31	7.428	7	3	7	5.044	7.101	6.861	6.359
32	8.947	15	5	4	16.360	16.585	12.934	13.808

Let us remind that Y implies the actual software size in the thousand lines of code (KLOC), X_1, X_2 and X_3 determine respectively the total number of classes, the total number of relationships and the average number of attributes per class, that is, $X_3 = A/X_1$, where A is the total number of attributes in conceptual data model.

For normalizing the multivariate non-Gaussian data from Table 1, we use the Johnson translation system

$$\mathbf{T} = \boldsymbol{\gamma} + \boldsymbol{\eta}\mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{P} - \varphi)\right] \sim N_m(\mathbf{0}_m, \boldsymbol{\Sigma}), \tag{5.1}$$

where $\mathbf{0}_m$ is the m-dimensional vector of means equal to zero; $\boldsymbol{\Sigma}$ is the $m \times m$ covariance matrix with variances equal to one; $\mathbf{h}[(y_Y, y_1, \ldots, y_k)] = \{h_Y(y_Y), h_1(y_1), \ldots, h_k(y_k)\}^T$; $h_i(\cdot)$ is one of the translation functions

$$h = \begin{cases} \ln(y), & \text{for } S_L \text{ (lognormal) family;} \\ \ln[y/(1-y)], & \text{for } S_B \text{ (bounded) family;} \\ \operatorname{Arsh}(y), & \text{for } S_U \text{ (unbounded) family;} \\ y & \text{for } S_N \text{ (normal) family.} \end{cases} \tag{5.2}$$

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.

We use the technique [7] based on multivariate normalizing transformations and the squared Mahalanobis distance (MD) to detect the outliers in the data from Table 1. There are no outliers in the data from Table 1 after their normalization by the Johnson multivariate transformation (5.1) for S_B family for 0.005 significance level. The same result has been obtained for the transformation (5.1) for S_U family. In [6] it is also assumed that the data contains no outliers. The values of squared MD for data normalized by the Johnson univariate transformation for S_B family from Table 1 indicate that the data of systems 11 and 19 are multivariate outliers, since for these data rows the values of squared MD equal to 18.29 and 17.16 respectively are greater than the value of the quantile of the Chi-Square distribution, which equals to 14.86 for 0.005 significance level. Without using normalization, the data of system 11 is multivariate outlier, since for this data row the squared MD equals to 15.44. It should be noted that there are no outliers in the data from Table 1 after their normalization by the decimal logarithm transformation.

Estimators for parameters of the multivariate transformation (5.1) for S_B family have been calculated by the maximum likelihood method and are: $\hat{\gamma}_Y = 0.96954$, $\hat{\gamma}_1 = 1.05143$, $\hat{\gamma}_2 = 0.89436$, $\hat{\gamma}_3 = 0.68406$, $\hat{\eta}_Y = 0.52252$, $\hat{\eta}_1 = 0.66953$, $\hat{\eta}_2 = 0.75874$, $\hat{\eta}_3 = 0.50162$, $\hat{\varphi}_Y = 1.9227$, $\hat{\varphi}_1 = 2.57124$, $\hat{\varphi}_2 = 0.219$, $\hat{\varphi}_3 = 3.319$, $\hat{\lambda}_Y = 90.756$, $\hat{\lambda}_1 = 47.425$, $\hat{\lambda}_2 = 28.507$ and $\hat{\lambda}_3 = 13.668$.

The sample covariance matrix S_N of the $\mathbf{T}$ is used as the approximate moment-matching estimator of covariance matrix $\boldsymbol{\Sigma}$

$$S_N = \begin{pmatrix} 1.0000 & 0.9328 & 0.9171 & 0.2258 \\ 0.9328 & 1.0000 & 0.8942 & 0.1903 \\ 0.9171 & 0.8942 & 1.0000 & 0.1024 \\ 0.2258 & 0.1903 & 0.1024 & 1.0000 \end{pmatrix}.$$

After normalizing the non-Gaussian data by the multivariate transformation (5.1) for S_B family the linear regression equation is built for normalized data

$$\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3. \tag{5.3}$$

Parameters of the equation (5.3) have been estimated by the least square method. The estimators for parameters of the equation (5.3) are: $\hat{b}_0 = 0$, $\hat{b}_1 = 0.522133$, $\hat{b}_2 = 0.441941$ and $\hat{b}_3 = 0.081219$.

After that the non-linear regression equation (4.2) is built

$$\hat{Y} = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y - \hat{\gamma}_Y\right)/\hat{\eta}_Y} \right]^{-1}, \tag{5.4}$$

where $\hat{Z}_Y$ is prediction result by the equation (5.3), $Z_j = \gamma_j + \eta_j \ln \frac{X_j - \varphi_j}{\varphi_j + \lambda_j - X_j}$, $\varphi_j < X_j < \varphi_j + \lambda_j$, $j = 1,2,3$.

The prediction results by nonlinear regression equation (5.4) for values of components of vector $\mathbf{X} = \{X_1, X_2, X_3\}$ from Table 1 are shown in the Table 1 for two cases: univariate and multivariate normalizing transformations. For univariate the Johnson normalizing transformations of S_B family (5.2) the estimators for parameters are: $\hat{\gamma}_Y = 0.77502$, $\hat{\gamma}_1 = 0.59473$, $\hat{\gamma}_2 = 0.57140$, $\hat{\gamma}_3 = 0.68734$, $\hat{\eta}_Y = 0.44395$, $\hat{\eta}_1 = 0.48171$, $\hat{\eta}_2 = 0.49553$, $\hat{\eta}_3 = 0.51970$, $\hat{\varphi}_Y = 2.063$, $\hat{\varphi}_1 = 2.900$, $\hat{\varphi}_2 = 0.900$, $\hat{\varphi}_3 = 3.304$, $\hat{\lambda}_Y = 83.059$, $\hat{\lambda}_1 = 36.695$, $\hat{\lambda}_2 = 23.525$ and $\hat{\lambda}_3 = 13.660$. In the case of univariate normalizing transformations the estimators for parameters of the equation (5.3) are: $\hat{b}_0 = 0$, $\hat{b}_1 = 0.43519$, $\hat{b}_2 = 0.52239$ and $\hat{b}_3 = 0.08546$.

Also the nonlinear regression equation (4.2) is built by the decimal logarithm transformation

$$\hat{Y} = 10^{\hat{b}_0} X_1^{\hat{b}_1} X_2^{\hat{b}_2} X_3^{\hat{b}_3}, \tag{5.5}$$

where the estimators for parameters of the equation (5.5) are: $\hat{b}_0 = -0.26161$, $\hat{b}_1 = 0.99151$, $\hat{b}_2 = 0.33232$ and $\hat{b}_3 = 0.13777$.

Table 1 also contains the prediction results by linear regression equation from [6] for values of components of vector $\mathbf{X} = \{X_1, X_2, X_3\}$ from Table 1. It should be noted that the prediction results obtained by the linear regression equation from [6] are negative for the three rows of data: 14, 19 and 27. All prediction results obtained by non-linear regression equations (5.4) and (5.5) are positive.

The values of multiple coefficient of determination R^2, Mean Magnitude of Relative Error (MMRE) and Percentage of Prediction (PRED(0.25)) for the regression equations are shown in the Table 2.

Table 2. Values of R^2, MMRE and PRED(0.25)

Coefficients	Linear regression equation	Non-linear regression equations		
		univariate transformations		multivariate transform.
		logarithm	Johnson	
R^2	0.9491	0.9375	0.9591	0.9730
MMRE	0.4919	0.2455	0.2535	0.2243
PRED(0.25)	0.5313	0.6250	0.7188	0.6875

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 are not more than 0.25 for nonlinear regression equation (5.4) on the basis of multivariate normalizing transformation and for non-linear regression equation (5.5) based on the decimal logarithm transformation. Although all values of PRED(0.25) in the Table 2 are less than 0.75 nevertheless the values are greater for nonlinear regression equation (5.4). All values of R^2 in the Table 2 are greater than 0.75 but the value R^2 is greater for nonlinear regression equation (5.4) on the basis of the Johnson multivariate transformation.

The confidence and prediction intervals of nonlinear regression are defined by (4.3) and (4.4) respectively for the data from Table 1. Table 3 contains the lower (LB) and upper (UB) bounds of the confidence intervals of linear and nonlinear regressions based on univariate and multivariate transformations respectively for 0.05 significance level. The values from Table 3 indicate that the lower bounds of the confidence interval of linear regression from [6] are negative for the seven rows of data: 1, 14, 19, 23, 26, 27 and 29. All the lower bounds of the confidence interval for nonlinear regression equations (5.4) and (5.5) are positive. For the fourteen rows of data the widths of the confidence

interval of linear regression are greater than for nonlinear regressions. The widths of the confidence interval of nonlinear regression on the basis of the Johnson multivariate transformation are less than following the decimal logarithm univariate transformation for the seventeen rows of data: 1, 3, 5, 7-14, 19, 23, 26, 27, 29 and 30. The widths of the confidence interval of nonlinear regression on the basis of the Johnson multivariate transformation are less than following the Johnson univariate transformation for the twenty-four rows of data: 1-4, 6-12, 15-18, 20-26, 28-31. Approximately the same results are obtained for the prediction interval of regressions.

Table 3. Bounds of the confidence intervals of regressions.

i	Bounds for linear regression		Bounds for nonlinear regressions					
			decimal logarithm transformation		Johnson univariate transformation		Johnson multivariate transformation	
	LB	UB	LB	UB	LB	UB	LB	UB
1	-0.402	6.877	3.725	5.947	3.673	6.267	3.532	5.370
2	21.413	26.871	18.933	27.172	15.473	25.455	16.565	26.460
3	34.344	40.704	26.415	39.621	24.791	40.266	28.688	41.670
4	23.172	28.660	16.855	24.285	17.982	29.365	18.876	28.212
5	69.187	80.062	55.076	87.173	74.107	83.078	67.843	82.455
6	21.015	25.433	16.557	22.438	16.129	25.819	17.128	24.385
7	62.690	71.740	52.309	83.045	56.434	72.961	59.139	74.114
8	1.013	7.241	4.288	6.544	4.484	7.748	4.147	6.243
9	3.084	8.728	4.569	7.951	5.456	10.203	4.848	8.199
10	43.863	49.824	35.573	52.195	33.947	50.366	38.167	52.354
11	49.560	66.068	41.448	87.698	42.891	79.359	39.199	71.614
12	53.265	60.725	43.512	65.766	44.275	61.787	48.759	64.130
13	54.146	69.566	35.747	69.156	49.897	75.572	48.637	75.635
14	-5.673	0.883	1.639	2.897	2.125	2.375	2.155	2.614
15	6.609	13.309	8.838	13.632	6.979	13.684	7.416	12.700
16	18.574	23.862	15.339	21.222	14.673	23.576	15.903	22.604
17	3.165	8.787	6.139	8.644	5.548	9.233	5.311	8.130
18	11.381	16.601	10.039	14.139	9.902	16.849	9.865	14.593
19	-4.587	1.845	2.106	3.945	2.253	3.046	2.471	3.389
20	11.684	19.085	9.137	15.776	9.186	19.255	9.685	17.159
21	30.767	39.591	20.246	34.593	16.796	41.072	20.369	40.407
22	14.250	19.840	10.430	16.253	11.581	20.525	11.374	18.048
23	-1.579	5.613	3.900	6.687	4.071	7.662	3.909	6.214
24	7.648	15.276	7.099	11.493	7.462	14.583	7.130	11.736
25	16.199	28.828	13.006	22.695	10.971	34.746	12.089	28.318
26	-1.967	5.227	4.421	7.029	4.083	7.261	4.032	6.293
27	-9.730	-1.580	1.360	2.712	2.092	2.281	2.040	2.436
28	38.873	49.077	25.212	43.588	25.181	51.940	29.184	52.086
29	-2.236	4.142	2.947	4.616	3.177	5.048	2.995	4.262
30	52.335	61.993	44.400	73.879	41.599	63.278	45.867	65.148
31	2.314	7.774	6.042	8.344	5.463	8.784	5.225	7.856
32	12.515	20.205	12.503	22.001	9.080	18.449	9.767	19.490

Following [8], multivariate kurtosis β_2 is estimated for the data from Table 1 and the normalized data on the basis of the Johnson univariate and multivariate transformations for S_B family. It is known that $\beta_2 = m(m+2)$ holds under 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 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 assumption of multivariate

normality for the data from Table 1 and for the data from Table 1 normalized by the Johnson univariate transformation of S_B family is rejected.

Squared MD is used for checking multivariate normality (MVN). According MD MVN test, the assumption of multivariate normality for the data from Table 1 normalized by the Johnson multivariate transformation (5.1) of S_B family and the decimal logarithm transformation is not rejected for 0.005 significance level. The assumption of multivariate normality for the data from Table 1 and for the data from Table 1 normalized by the Johnson univariate transformation of S_B family is rejected for 0.005 significance level.

It should be noted that the poor normalization of multivariate non-Gaussian data using the Johnson univariate transformation leads to an expansion of the confidence and prediction intervals of nonlinear regression for a larger number of data rows compared to both the Johnson multivariate transformation and the decimal logarithm transformation. The values of R^2 and MMRE are better for the equation (5.4) for the Johnson multivariate transformation in comparison with all previous regression equations, both linear and nonlinear, based on univariate transformations. This can be explained best by the 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.

6 Conclusions

To sum it up, when constructing the equations, confidence and prediction intervals of nonlinear regressions for multivariate non-Gaussian data multivariate normalizing transformations should be used.

From the examples we can make a conclusion that the considered techniques based on multivariate normalizing transformations are promising ones, since they lead to a narrowing of the confidence and prediction intervals of nonlinear regression for a larger number of data rows compared to the univariate normalizing transformations.

As a rule, poor normalization of multivariate non-Gaussian data using univariate transformations instead of multivariate ones can result in an expansion of the confidence and prediction intervals of nonlinear regression.

Prospects for the further research include the application of new multivariate normalizing transformations and data sets for constructing the equations, confidence and prediction intervals of nonlinear regressions for multivariate non-Gaussian data.

REFERENCIES

1. D. M. Bates, and D. G. Watts, Nonlinear Regression Analysis and Its Applications. New York: John Wiley & Sons, 1988. DOI:10.1002/9780470316757
2. G. A. F. Seber, and C. J. Wild, Nonlinear Regression. New York: John Wiley & Sons, 1989. DOI: 10.1002/0471725315
3. T. P. Ryan, Modern regression methods. New York: John Wiley & Sons, 1997. DOI: 10.1002/9780470382806
4. R. A. Johnson, and D. W. Wichern, Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.
5. S. B. Prykhodko, “Developing the software defect prediction models using regression analysis based on normalizing transformations” in “Modern problems in testing of the applied software” (PTASS-2016), Abstracts of the Research and Practice Seminar, Poltava, Ukraine, May 25-26, 2016, pp. 6-7.
6. 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. DOI: 10.1145/1134285.1134331
7. 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. DOI: 10.1109/UKRCON.2017.8100366
8. K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications", Biometrika, 57, 1970, pp. 519-530. DOI: 10.1093/biomet/57.3.519

ЛІТЕРАТУРА

1. 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
2. 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
3. Ryan T.P. Modern regression methods / T. P. Ryan. New York: John Wiley & Sons, 1997. 529 p. DOI: 10.1002/9780470382806
4. Johnson R.A. Applied Multivariate Statistical Analysis / R. A. Johnson, D. W. Wichern. Pearson Prentice Hall, 2007. 800 p.
5. Prykhodko S.B. Developing the software defect prediction models using regression analysis based on normalizing transformations / S. B. Prykhodko // Сучасні проблеми тестування прикладного програмного забезпечення: збірник тез доповідей науково-практичного семінару, Полтава, 25-26 травня 2016 р. Полтава, 2016. С. 6-7.
6. 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
7. 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
8. Mardia K.V. Measures of multivariate skewness and kurtosis with applications / K. V. Mardia // Biometrika, 57, 1970. P. 519-530. DOI: 10.1093/biomet/57.3.519

Приходько Наталія Василівна *– кандидат економічних наук, доцент; Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв, проспект Героїв України, 9, 54025; e-mail: natalia.prykhodko@nuos.edu.ua; ORCID: 0000-0002-3554-7183.*

Приходько Сергій Борисович *– доктор технічних наук, професор; Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв, проспект Героїв України, 9, 54025; e-mail: sergiy.prykhodko@nuos.edu.ua; ORCID: 0000-0002-2325-018X.*

Надійшла у першій редакції 21.06.2018, в останній - 20.10.2018.