# A Modified Technique for Constructing Nonlinear Regression Models Based on the Multivariate Normalizing Transformations

Sergiy Prykhodko and Natalia Prykhodko

*Admiral Makarov National University of Shipbuilding, Heroes of Ukraine Ave., 9, Mykolaiv, 54007, Ukraine*

**Abstract**

The technique for constructing nonlinear regression models based on the multivariate normalizing transformations and prediction intervals is modified by testing the normality of error distribution in the linear regression model for normalized data, which is applied to construct the nonlinear one. We have demonstrated that there may be multidimensional data sets (for example, software metrics) for which the constructed nonlinear regression model has poor prediction accuracy even in the cases of, first, multivariate normality of the normalized data, second, outlier cutoff using the technique based on the multivariate normalizing transformations and prediction intervals of nonlinear regression. As a rule, this is because the error distribution in the linear regression model for normalized data becomes non-Gaussian. Therefore, in this case, we propose to discard the multidimensional data point for which the value of the error modulus in the model is maximum. We have considered the application of a modified technique for constructing a nonlinear regression model with three predictors (software metrics) to estimate the size of open source PHP-based apps. This model has the better values of well-known prediction accuracy metrics compared with the model, which is only constructed based on the multivariate normalizing transformation and prediction intervals without testing the normality of error distribution in the linear regression model for normalized data and discarding the multidimensional data point for which the value of the error modulus in the linear model is maximum if the distribution of residuals in the linear one is not Gaussian.

**Keywords** [1]

Nonlinear regression model, normalizing transformation, prediction interval, outlier, software size, estimation, PHP

## 1. Introduction

Regression models are one of the general models in mathematical modeling of various dependent random variables and other applications in many fields, including information technologies. As known, the use of linear regression models has some limitations. For example, linear regression models have significant limitations as practical techniques for pattern recognition, particularly for problems involving input spaces of high dimensionality [1]. And the application of linear regression models demands the assumption fulfillment of error distribution normality [2] that is valid in particular cases only. This leads to the need to apply nonlinear regression models.

As we know [2-7], normalizing transformations are used to build nonlinear regression models. In this case, methods for outlier detection in nonlinear regression models based on the normalizing transformations can be applied [8, 9]. Like the method proposed in [10], which combines a new robust nonlinear regression method with a new method for identifying outliers, the technique to construct nonlinear regression models considered in [9] includes a new technique for detecting outliers, but multivariate ones in contrast to [10]. Notice that nonlinear regression models built by the above technique [9] based on the multivariate normalizing transformations and prediction intervals usually lead to better results compared to the models that are constructed using univariate transformations and without taking into account the presence of outliers. However, there may be multidimensional data sets for which the

*Information Technology and Implementation (IT&I-2021), December 01–03, 2021, Kyiv, Ukraine*
EMAIL: sergiy.prykhodko@nuos.edu.ua (A. 1); natalia.prykhodko@nuos.edu.ua (A. 2)
ORCID: 0000-0002-2325-018X (A. 1); 0000-0002-3554-7183 (A. 2)


constructed nonlinear regression model has unsatisfactory prediction accuracy even after all outliers are removed and ending the construction of the nonlinear regression model using the technique [9]. This result, as we have found out, may be explained by the fact that the error distribution in the linear regression model for normalized data becomes non-Gaussian at the final step of constructing the nonlinear regression model. And as we know [2-7, 11-14], the error distribution in the linear regression model should be Gaussian. For this reason, the technique [9] requires some modification.

According to [2], for a disturbance (error) term in a linear regression model "the assumption of normality may be checked by examining the residuals." And the existing methods for outlier detection in regressions by analysis of residuals (including the standardized, studentized, and studentized deleted residuals) are based on the assumption of normality of residuals. Also, as noted in [12], "procedures for the detection of outliers rely almost exclusively on the detection of extreme residuals, so much so that the two are used interchangeably by some authors and researchers." Therefore, we have proposed to modify the technique [9] using testing the normality of error distribution in the linear regression model for normalized data, which is applied to construct the nonlinear model and discarding the multidimensional data point for which the value of the residual modulus in the linear regression model is maximum in the case if this error distribution is not Gaussian.

Next, we describe the proposed modification of the technique [9] and demonstrate its viability by the example of constructing the nonlinear regression model with three predictors (software metrics) to estimate the size of open source PHP-based apps.

## 2. A modified technique

We modify the technique to construct nonlinear regression models based on the multivariate normalizing transformations and prediction intervals [9]. We combined the technique [9] with outlier detection using residuals in the linear regression model for normalized data, which is applied to construct the nonlinear one. A modified technique follows six steps.

Step 1. Normalize a multivariate non-Gaussian data set by a normalizing transformation. We apply a multivariate normalizing transformation.

Step 2. Determine whether one multidimensional data point of a multivariate non-Gaussian data set is a multidimensional outlier. If there is a multidimensional outlier in a multivariate non-Gaussian data set, then discard the one and go to step 1, else go to step 3.

Step 3. Build the linear regression model for normalized data, which is applied to construct the nonlinear one.

Step 4. Test the normality of distribution of residuals in the linear regression model for normalized data. If the distribution of residuals in the linear regression model for normalized data is not Gaussian, then discard the multidimensional data point for which the value of the modulus of residual in the model is maximum and go to step 1, else go to step 5.

Step 5. Construct the nonlinear regression model based on the multivariate normalizing transformation and the linear regression model for normalized data.

Step 6. Build the prediction interval of nonlinear regression, and determine whether one or more values of the response (dependent random variable) are outliers. If there are outliers in the nonlinear regression model, then discard the ones and go to step 1, else complete constructing the nonlinear regression model.

Notice that outlier detection in a multivariate non-Gaussian data set is included in both the technique [9] and the modified one because after going to step 1 from step 6, an outlier may again appear in the reduced data set. To detect outliers in step 2, we apply the statistical technique [15, 16] based on the normalizing transformations and the Mahalanobis squared distance (MSD). As in [9, 15, 16], we use the Johnson multivariate transformation to normalize a multivariate non-Gaussian data set at step 1.

In step 3, we build the linear regression model for normalized data by the least squared method. Next, we calculate the *i*-value of residual in the linear regression model for normalized data by the formula

$$\varepsilon_i = Z_{Y_i} - \hat{Z}_{Y_i} = Z_{Y_i} - \hat{b}_0 - \hat{b}_1 Z_{1_i} - \hat{b}_2 Z_{2_i} - \ldots - \hat{b}_k Z_{k_i}, \quad (1)$$

where $Z_{Y_i}$ is the normalized *i*-value of the non-Gaussian dependent random variable *Y*; $\hat{Z}_{Y_i}$ is a prediction result by linear regression equation for normalized data for *i*-values of normalized predictors $Z_1$, $Z_2$, …, $Z_k$, which are transformed from independent variables $X_1$, $X_2$, …, $X_k$, by a bijective multivariate normalizing transformation $\mathbf{T}=\boldsymbol{\psi}(\mathbf{P})$ of non-Gaussian random vector $\mathbf{P}=\{Y,X_1,X_2,\ldots,X_k\}^T$ to Gaussian one $\mathbf{T}=\{Z_Y,Z_1,Z_2,\ldots,Z_k\}^T$, $\hat{Z}_{Y_i}=\hat{b}_0+\hat{b}_1Z_{1_i}+\hat{b}_2Z_{2_i}+\ldots+\hat{b}_kZ_{k_i}$; *k* is a number of independent variables (predictors or regressors); $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, …, $\hat{b}_k$ are estimates of parameters in linear regression equation for normalized data.

To test the normality of distribution of residuals in the linear regression model for normalized data, we apply the Kolmogorov-Smirnov test if the number of values of the random variable is less than 30, and the Pearson chi-squared test in the opposite case. As we note above, if the distribution of residuals in the linear regression model for normalized data is not Gaussian, then we discard the multidimensional data point for which the value of the modulus of residual in the model is maximum and go to step 1. In this case, the residual value is determined by the formula (1). The detection of extreme residual is the same as that and the detection of an outlier. But here, we use such outlier detection because the distribution of residuals in the linear regression model for normalized data is not Gaussian. Also, notice that we go to step 1 if the distribution of residuals in the linear regression model for normalized data is not Gaussian because there is no justification for the use of a linear regression model in this case.

We construct the nonlinear regression model based on the multivariate normalizing transformation and the linear regression model for normalized data if the null hypothesis $H_0$ that the observed frequency distribution of residuals is the same as the normal distribution is accepted. To do this, we apply the technique [9]. Also, we build the prediction interval of nonlinear regression according to [9].

Modification of the technique [9] consists of adding step 4 and highlighting step 3 to build the linear regression model for normalized data, which is applied to construct the nonlinear one. Next, we consider the example of constructing the nonlinear regression model with three predictors by a modified technique to demonstrate its viability for estimating the size of open source PHP-based apps.

## 3. Model construction example

We consider the example of constructing a nonlinear regression model with three predictors to estimate the size of open source PHP-based apps. We construct a model for the non-Gaussian data set from 44 apps hosted on GitHub (https://github.com) by a modified technique. The data set was obtained using the PhpMetrics tool (https://phpmetrics.org/). The model is constructed around the following metrics (variables) of the app: the size (in KLOC, thousand lines of code) *Y*, the number of classes $X_1$, the average number of methods per class $X_2$, a sum of average afferent coupling, and average efferent coupling per class $X_3$. Table 1 contains the data set of the above metrics from 44 rows. Moreover, *Y* is the dependent variable, and $X_1$, $X_2$, and $X_3$ are independent variables (predictors).

We tested the normality of four-dimensional data from Table 1 by the Mardia test based on the measures of the multivariate skewness $\beta_1$ and kurtosis $\beta_2$ [17]. According to the Mardia test, the four-variate distribution of four-dimensional data from Table 1 is not Gaussian since the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals 237.49, is greater than the value of the Chi-Square distribution quantile, which is 45.31 for 20 degrees of freedom and 0.001 significance level. Similarly, the test statistic for multivariate kurtosis $\beta_2$, which equals 59.21, is greater than the value of the Gaussian distribution quantile, which is 30.46 for 24 mean, 4.36 variance, and 0.001 significance level. We apply the above test statistics because the number of rows of multi-dimensional data *N* is greater than 20. Notice that otherwise, other test corrected statistics should be used according to [18].

The above results indicate to us the need for the application of the method to determine multidimensional outliers in multivariate non-Gaussian data further. To do this we apply the statistical technique [15, 16] based on the normalizing transformations and the MSD. In the beginning, we normalize the four-dimensional non-Gaussian data set from Table 1 by the Johnson four-variate transformation for the $S_B$ family, for which the components are defined using

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

where $Z_j$ is a standard Gaussian variable, $Z_j \sim N(0,1)$; $\gamma_j$, $\eta_j$, $\varphi_j$, and $\lambda_j$ are parameters of the Johnson transformation for $S_B$ family, $\eta_j > 0$, $\lambda_j > 0$, $j = 1,2,3$. The component $Z_Y$ is defined analogously (2) with the only difference that instead of $Z_j$, $X_j$, $\gamma_j$, $\eta_j$, $\varphi_j$, $\lambda_j$ should be put respectively $Z_Y$, $Y$, $\gamma_Y$, $\eta_Y$, $\varphi_Y$, $\lambda_Y$. The estimates of parameters of the Johnson four-variate transformation for the $S_B$ family for the data from Table 1 are calculated by the maximum likelihood method

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\mathbf{P}, \boldsymbol{\theta}), \tag{3}$$

where the log-likelihood function is

$$l(\mathbf{P}, \boldsymbol{\theta}) = N \ln(\eta_Y \lambda_Y) - N \sum_{j=1}^{k} \ln(\eta_j \lambda_j) - \frac{N(k+1)\ln(2\pi)}{2} - \frac{N}{2} \ln[\det(\mathbf{S_T})] -$$
$$- \sum_{i=1}^{N} \ln[(Y_i - \varphi_Y)(\varphi_Y + \lambda_Y - Y_i)] - \sum_{j=1}^{k} \sum_{i=1}^{N} \ln[(X_{j_i} - \varphi_j)(\varphi_j + \lambda_j - X_{j_i})] - \frac{1}{2} \sum_{i=1}^{N} \mathbf{T}_i^T \mathbf{S_T}^{-1} \mathbf{T}_i. \tag{4}$$

Here $\mathbf{S_T}$ is the sample covariance matrix for the components of $\mathbf{T}$. Notice in (4), we take into account that for the Johnson multivariate translation system, the components of the mean vector are equal to zero.

**Table 1**

The data set from 44 open-source PHP-based apps

| No | $Y$ | $X_1$ | $X_2$ | $X_3$ | No | $Y$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 174.927 | 2075 | 4.809 | 6.425 | 23 | 10.044 | 314 | 2.226 | 4.242 |
| 2 | 112.048 | 445 | 2.591 | 3.054 | 24 | 15.477 | 280 | 3.400 | 3.793 |
| 3 | 82.551 | 411 | 7.856 | 6.358 | 25 | 15.595 | 115 | 7.965 | 5.096 |
| 4 | 12.022 | 132 | 5.811 | 5.053 | 26 | 2.323 | 15 | 10.067 | 4.933 |
| 5 | 5.347 | 5 | 32.600 | 2.400 | 27 | 7.101 | 25 | 9.400 | 3.440 |
| 6 | 0.601 | 25 | 1.080 | 1.760 | 28 | 1.431 | 22 | 3.409 | 2.591 |
| 7 | 1.561 | 25 | 2.240 | 4.040 | 29 | 37.081 | 278 | 5.888 | 4.385 |
| 8 | 33.276 | 216 | 7.671 | 10.102 | 30 | 32.826 | 235 | 8.191 | 6.430 |
| 9 | 36.028 | 126 | 10.849 | 2.587 | 31 | 12.219 | 58 | 13.379 | 9.155 |
| 10 | 100.245 | 448 | 9.569 | 6.315 | 32 | 59.618 | 568 | 5.974 | 6.285 |
| 11 | 4.458 | 73 | 4.452 | 5.219 | 33 | 24.864 | 363 | 5.595 | 11.474 |
| 12 | 2.988 | 18 | 9.611 | 7.000 | 34 | 2.362 | 28 | 5.357 | 6.250 |
| 13 | 4.047 | 31 | 6.323 | 5.355 | 35 | 0.381 | 8 | 3.125 | 2.750 |
| 14 | 6.688 | 125 | 2.824 | 3.816 | 36 | 4.308 | 52 | 6.519 | 4.923 |
| 15 | 1.247 | 2 | 18.000 | 2.000 | 37 | 3.412 | 52 | 3.327 | 7.654 |
| 16 | 5.966 | 74 | 6.176 | 4.851 | 38 | 15.785 | 126 | 8.968 | 4.190 |
| 17 | 38.996 | 269 | 7.684 | 6.394 | 39 | 0.535 | 3 | 9.333 | 3.333 |
| 18 | 3.269 | 37 | 5.108 | 4.946 | 40 | 31.676 | 76 | 14.105 | 7.474 |
| 19 | 35.548 | 335 | 5.818 | 11.096 | 41 | 13.94 | 251 | 3.410 | 3.378 |
| 20 | 8.910 | 117 | 4.641 | 3.735 | 42 | 3.334 | 16 | 8.063 | 8.813 |
| 21 | 14.019 | 209 | 4.536 | 8.713 | 43 | 24.298 | 794 | 0.487 | 2.242 |
| 22 | 2.920 | 19 | 8.158 | 5.737 | 44 | 42.941 | 282 | 2.872 | 3.915 |

The estimates of parameters of the Johnson four-variate transformation for $S_B$ family for the data from Table 1 are $\hat{\gamma}_Y = 1.61708$, $\hat{\gamma}_1 = 2.17927$, $\hat{\gamma}_2 = 13.84357$, $\hat{\gamma}_3 = 0.860885$, $\hat{\eta}_Y = 0.538609$, $\hat{\eta}_1 = 0.612224$, $\hat{\eta}_2 = 1.87447$, $\hat{\eta}_3 = 0.90490$, $\hat{\varphi}_Y = 0.2810$, $\hat{\varphi}_1 = 0.124604$, $\hat{\varphi}_2 = -1.29688$, $\hat{\varphi}_3 = 1.50142$, $\hat{\lambda}_Y = 201.7538$, $\hat{\lambda}_1 = 3128.508$, $\hat{\lambda}_2 = 11606.794$, $\hat{\lambda}_3 = 12.0339$. The sample covariance matrix $\mathbf{S_T}$ is following

$$\mathbf{S_T} = \begin{pmatrix} 1.00000 & 0.87083 & 0.05023 & 0.37632 \\ 0.87083 & 1.00000 & -0.38093 & 0.33310 \\ 0.05023 & -0.38093 & 1.00000 & 0.21822 \\ 0.37632 & 0.33310 & 0.21822 & 1.00000 \end{pmatrix}.$$

We tested the normality of normalized four-dimensional data from Table 1 by the Mardia test based on the measures of the multivariate skewness $\beta_1$ and kurtosis $\beta_2$ [17]. According to the Mardia test, the four-variate distribution of normalized four-dimensional data from Table 1 is not Gaussian since the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals 61.97, is greater than the value of the Chi-Square distribution quantile, which is 45.31 for 20 degrees of freedom and 0.001 significance level. Similarly, the test statistic for multivariate kurtosis $\beta_2$, which equals 31.35, is greater than the value of the Gaussian distribution quantile, which is 30.46 for 24 mean, 4.364 variance, and 0.001 significance level.

According to [15, 16], row 2 from Table 1 is the multivariate outlier in four-dimensional non-Gaussian data since the MSD value for normalized data of row 2, which equals 19.71, is greater than the value of the Chi-Square distribution quantile, which equals 14.86 for the 0.005 significance level. After that, we discard the outlier and go to step 1. The first iteration is completed.

Next, we normalize the four-dimensional non-Gaussian data set from 43 rows of Table 1 (without row 2) by the Johnson four-variate transformation for the $S_B$ family using (2). In this case, the estimates of parameters of the Johnson four-variate transformation for $S_B$ family for the data from Table 1 (without row 2) are $\hat{\gamma}_Y = 1.75572$, $\hat{\gamma}_1 = 2.19844$, $\hat{\gamma}_2 = 15{,}3097$, $\hat{\gamma}_3 = 0.798411$, $\hat{\eta}_Y = 0.559115$, $\hat{\eta}_1 = 0.610394$, $\hat{\eta}_2 = 2.10956$, $\hat{\eta}_3 = 0.893772$, $\hat{\varphi}_Y = 0.2810$, $\hat{\varphi}_1 = 0.239008$, $\hat{\varphi}_2 = -2.07165$, $\hat{\varphi}_3 = 1.49364$, $\hat{\lambda}_Y = 212.2741$, $\hat{\lambda}_1 = 3116{,}897$, $\hat{\lambda}_2 = 11606.864$, $\hat{\lambda}_3 = 11.8220$. The sample covariance matrix $\mathbf{S_T}$ is following

$$\mathbf{S_T} = \begin{pmatrix} 1.00000 & 0.86822 & 0.09732 & 0.43126 \\ 0.86822 & 1.00000 & -0.37186 & 0.36348 \\ 0.09732 & -0.37186 & 1.00000 & 0.17306 \\ 0.43126 & 0.36348 & 0.17306 & 1.00000 \end{pmatrix}.$$

We tested the normality of normalized four-dimensional data from Table 1 (without rows 2) by the Mardia test based on the measures of the multivariate skewness $\beta_1$ and kurtosis $\beta_2$ [17]. According to the Mardia test, the four-variate distribution of normalized four-dimensional data from Table 1 (without rows 2) is Gaussian since, first, the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals 38.79, is less than the value of the Chi-Square distribution quantile, which is 45.31 for 20 degrees of freedom and 0.001 significance level, second, the test statistic for multivariate kurtosis $\beta_2$, which equals 28.73, is less than the value of the Gaussian distribution quantile, which is 30.53 for 24 mean, 4.465 variance, and 0.001 significance level.

According to [15, 16], there is no multivariate outlier in the four-dimensional non-Gaussian data set from 43 rows of Table 1 (without row 2) at the second iteration since the MSD values for normalized data of 43 rows are less than the value of the Chi-Square distribution quantile, which equals 14.86 for the 0.005 significance level. Therefore, we go to step 3.

We build the linear regression model for normalized data in the form

$$Z_Y = \hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon, \qquad (5)$$

where $\hat{b}_0 = 0$, $\hat{b}_1 = 1.07222$, $\hat{b}_2 = 0.503937$, $\hat{b}_3 = -0.045683$, $\varepsilon$ is the error term that is the Gaussian random variable to describe residuals, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$, $\sigma_\varepsilon$ is the standard deviation of $\varepsilon$ with the estimate $\hat{\sigma}_\varepsilon$ of 0.2017.

To test the normality of the distribution of residuals in the linear regression model (5), we apply the Pearson chi-squared test. According to this test, the null hypothesis $H_0$ that the observed frequency distribution of residuals is the same as the normal distribution is accepted because the $\chi^2$ test statistic, which equals 1.92, does not surpass the critical value from the Chi-Squared distribution which is 7.82 for 3 degrees of freedom and 0.05 significance level. Therefore, we go to step 5.

According to [9], we construct the nonlinear regression model based on the Johnson four-variate transformation for the $S_B$ family and the linear regression model (5) for normalized data from 43 rows of Table 1 (without row 2). This nonlinear regression model has the form [9]

$$Y = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y\right)/\hat{\eta}_Y}\right]^{-1} \tag{6}$$

with estimates of parameters in the second iteration.

Next, according to [9], we build the prediction interval of nonlinear regression as

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

where $\psi_Y$ is the transformation (1) for $Y$, $\psi_Y^{-1} = \varphi_Y + \lambda_Y / \left(1 + e^{-(Z_Y - \gamma_Y)/\eta_Y}\right)$; $t_{\alpha/2,\nu}$ is a student's $t$-distribution quantile with $\alpha/2$ significance level and $\nu$ 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 $Z_{1_i} - \bar{Z}_1$, $Z_{2_i} - \bar{Z}_2$, …, $Z_{k_i} - \bar{Z}_k$ for $i$-row; $\bar{Z}_j = \frac{1}{N}\sum_{i=1}^{N} Z_{j_i}$, $j = 1,2,\ldots,k$; $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$; $\mathbf{S}_Z$ is the $k \times k$ matrix

$$\mathbf{S}_Z = \begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \ldots & S_{Z_1Z_k} \\ S_{Z_1Z_2} & S_{Z_2Z_2} & \ldots & S_{Z_2Z_k} \\ \ldots & \ldots & \ldots & \ldots \\ S_{Z_1Z_k} & S_{Z_2Z_k} & \ldots & S_{Z_kZ_k} \end{pmatrix}, \tag{8}$$

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

In the second iteration, for the data normalized by the Johnson four-variate transformation for $S_B$ family from 43 rows of Table 1 (without row 2), the matrix (8) is following

$$\mathbf{S}_Z = \begin{pmatrix} 43.00 & -15.99 & 15.63 \\ -15.99 & 43.00 & 7.44 \\ 15.63 & 7.44 & 43.00 \end{pmatrix}.$$

It turned out that only two $Y$ values for rows 35 and 44 are outside the prediction interval, which was calculated by (7) for a significance level of 0.05. In Table 2, the lower and upper bounds of the prediction interval obtained in the second iteration are denoted as $LB_2$, and $UB_2$, respectively. In Table 2, as in [9], the row numbers with the outliers in data are highlighted in bold at the relevant iteration, and a dash (-) depicts the exception of the corresponding numbers of data at the relevant iteration.

Next, we discard two outliers (rows 35 and 44). The second iteration is completed. We start the third iteration and go to step 1. In step 1, we normalize the four-dimensional non-Gaussian data set from 41 rows of Table 1 (without rows 2, 25, and 44) by the Johnson four-variate transformation for the $S_B$ family using (2). In this case the estimates of parameters of the Johnson four-variate transformation for $S_B$ family for the data from Table 1 (without rows 2, 25, and 44) are $\hat{\gamma}_Y = 3.01458$, $\hat{\gamma}_1 = 2.87815$, $\hat{\gamma}_2 = 16.5902$, $\hat{\gamma}_3 = 0.710304$, $\hat{\eta}_Y = 0.698176$, $\hat{\eta}_1 = 0.652161$, $\hat{\eta}_2 = 2.322625$, $\hat{\eta}_3 = 0.878073$,

$\hat{\varphi}_Y = 0.086857$, $\hat{\varphi}_1 = -0.063309$, $\hat{\varphi}_2 = -2.840555$, $\hat{\varphi}_3 = 1.48156$, $\hat{\lambda}_Y = 707.4161$, $\hat{\lambda}_1 = 6959.670$, $\hat{\lambda}_2 = 11672.25$, $\hat{\lambda}_3 = 11.5940$. The sample covariance matrix $\mathbf{S_T}$ is following

$$\mathbf{S_T} = \begin{pmatrix} 1.00000 & 0.86577 & 0.06900 & 0.42189 \\ 0.86577 & 1.00000 & -0.41849 & 0.35461 \\ 0.06900 & -0.41849 & 1.00000 & 0.12036 \\ 0.42189 & 0.35461 & 0.12036 & 1.00000 \end{pmatrix}.$$

We tested the normality of normalized four-dimensional data from Table 1 (without rows 2, 25, and 44) by the Mardia test based on the measures of the multivariate skewness $\beta_1$ and kurtosis $\beta_2$ [17]. According to the Mardia test, the four-variate distribution of normalized four-dimensional data from Table 1 (without rows 2, 25, and 44) is Gaussian since, first, the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals 28.66, is less than the value of the Chi-Square distribution quantile, which is 45.31 for 20 degrees of freedom and 0.001 significance level, second, the test statistic for multivariate kurtosis $\beta_2$, which equals 26.06, is less than the value of the Gaussian distribution quantile, which is 30.69 for 24 mean, 4.683 variance, and 0.001 significance level.

**Table 2**

Lower and upper bounds of the prediction interval for nonlinear regression

| No | $LB_2$ | $UB_2$ | $LB_4$ | $UB_4$ | No | $LB_2$ | $UB_2$ | $LB_4$ | $UB_4$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 141.219 | 193.978 | 137.430 | 221.703 | 23 | 6.670 | 27.764 | 8.062 | 20.813 |
| **2** | - | - | - | - | 24 | 9.091 | 36.423 | 10.679 | 27.061 |
| 3 | 36.445 | 105.427 | 40.098 | 89.814 | 25 | 9.176 | 36.212 | 11.029 | 27.574 |
| 4 | 7.023 | 28.295 | 8.508 | 21.425 | 26 | 1.433 | 5.636 | 1.844 | 4.699 |
| 5 | 2.569 | 12.792 | 3.733 | 10.881 | 27 | 2.285 | 9.420 | 2.918 | 7.490 |
| 6 | 0.470 | 1.301 | 0.458 | 1.106 | 28 | 0.722 | 2.396 | 0.817 | 2.008 |
| **7** | 0.580 | 1.754 | - | - | 29 | 17.071 | 61.480 | 19.367 | 47.085 |
| 8 | 15.512 | 58.221 | 18.403 | 45.611 | 30 | 20.645 | 71.044 | 23.650 | 56.401 |
| 9 | 16.803 | 63.803 | 19.485 | 49.153 | 31 | 8.111 | 33.723 | 10.298 | 26.543 |
| 10 | 50.521 | 127.947 | 55.464 | 117.672 | 32 | 37.982 | 108.120 | 40.874 | 91.235 |
| 11 | 2.607 | 10.705 | 3.298 | 8.440 | 33 | 17.585 | 66.332 | 20.438 | 51.732 |
| 12 | 1.533 | 6.159 | 1.995 | 5.132 | 34 | 1.199 | 4.582 | 1.502 | 3.815 |
| 13 | 1.629 | 6.446 | 2.069 | 5.257 | **35** | **0.396** | **0.858** | - | - |
| 14 | 2.963 | 12.392 | 3.723 | 9.633 | 36 | 2.922 | 11.992 | 3.697 | 9.415 |
| 15 | 0.555 | 1.758 | 0.675 | 1.718 | 37 | 1.275 | 5.081 | 1.620 | 4.217 |
| 16 | 4.003 | 16.475 | 5.000 | 12.711 | 38 | 12.274 | 47.116 | 14.481 | 35.991 |
| 17 | 22.098 | 74.708 | 25.110 | 59.510 | 39 | 0.430 | 1.020 | 0.416 | 0.925 |
| 18 | 1.533 | 6.015 | 1.931 | 4.902 | 40 | 12.426 | 48.388 | 15.259 | 38.290 |
| 19 | 17.166 | 64.386 | 20.036 | 50.228 | 41 | 8.126 | 33.053 | 9.605 | 24.503 |
| 20 | 4.797 | 19.821 | 5.881 | 15.004 | 42 | 1.061 | 4.112 | 1.350 | 3.499 |
| 21 | 7.879 | 32.283 | 9.622 | 24.565 | 43 | 9.401 | 41.759 | 11.033 | 30.214 |
| 22 | 1.361 | 5.296 | 1.735 | 4.408 | **44** | **7.620** | **31.153** | - | - |

According to [15, 16], there is no multivariate outlier in the four-dimensional non-Gaussian data set from 41 rows of Table 1 (without rows 2, 25, and 44) at the third iteration since the MSD values for normalized data of 41 rows are less than the value of the Chi-Square distribution quantile, which equals 14.86 for the 0.005 significance level. Therefore, we go to step 3.

In step 3, we build the linear regression model (5) whose parameter estimates $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$ are equal to 0, 1.100603, 0.533514, and -0.032605, respectively. The estimate $\hat{\sigma}_\varepsilon$ is 0.1571.

Next, we test the normality of the distribution of residuals in the linear regression model (5) by the Pearson chi-squared test. According to this test, the null hypothesis $H_0$ that the observed frequency

distribution of residuals is the same as the normal distribution is rejected because the $\chi^2$ test statistic, which equals 11.35, surpasses the critical value from the Chi-Squared distribution which is 7.82 for 3 degrees of freedom and 0.05 significance level. Therefore, we discard the raw 7 for which the value of the modulus of residual in the model is maximum and equal to 0.2999. The third iteration is completed and we go to step 1.

In step 1 of the fourth iteration, we normalize the four-dimensional non-Gaussian data set from 40 rows of Table 1 (without rows 2, 7, 25, and 44) by the Johnson four-variate transformation for the $S_B$ family using (2). In this case, the estimates of parameters of the Johnson four-variate transformation for $S_B$ family for the data from Table 1 (without rows 2, 7, 25, and 44) are $\hat{\gamma}_Y = 2.331506$, $\hat{\gamma}_1 = 2.502955$, $\hat{\gamma}_2 = 16.23225$, $\hat{\gamma}_3 = 0.628931$, $\hat{\eta}_Y = 0.670065$, $\hat{\eta}_1 = 0.638907$, $\hat{\eta}_2 = 2.263781$, $\hat{\eta}_3 = 0.814678$, $\hat{\varphi}_Y = 0.131886$, $\hat{\varphi}_1 = -0.091614$, $\hat{\varphi}_2 = -2.51524$, $\hat{\varphi}_3 = 1.559526$, $\hat{\lambda}_Y = 330.6481$, $\hat{\lambda}_1 = 4464.000$, $\hat{\lambda}_2 = 11714.039$, $\hat{\lambda}_3 = 11.1327$. The sample covariance matrix $\mathbf{S_T}$ is following

$$\mathbf{S_T} = \begin{pmatrix} 1.00000 & 0.86660 & 0.02741 & 0.41858 \\ 0.86660 & 1.00000 & -0.45561 & 0.34855 \\ 0.02741 & -0.45561 & 1.00000 & 0.12038 \\ 0.41858 & 0.34855 & 0.12038 & 1.00000 \end{pmatrix}.$$

We tested the normality of normalized four-dimensional data from Table 1 (without rows 2, 7, 25, and 44) by the Mardia test based on the measures of the multivariate skewness $\beta_1$ and kurtosis $\beta_2$ [17]. According to the Mardia test, the four-variate distribution of normalized four-dimensional data from Table 1 (without rows 2, 7, 25, and 44) is Gaussian since, first, the test statistic for multivariate skewness $N\beta_1/6$ of this data, which equals 28.95, is less than the value of the Chi-Square distribution quantile, which is 45.31 for 20 degrees of freedom and 0.001 significance level, second, the test statistic for multivariate kurtosis $\beta_2$, which equals 26.22, is less than the value of the Gaussian distribution quantile, which is 30.77 for 24 mean, 4.8 variance, and 0.001 significance level. We apply the above test statistics because the number of rows of multi-dimensional data $N$ is greater than 20. Notice that otherwise, other test corrected statistics should be used according to [18].

According to [15, 16], there is no multivariate outlier in the four-dimensional non-Gaussian data set from 40 rows of Table 1 (without rows 2, 7, 25, and 44) at the fourth iteration since the MSD values for normalized data of 40 rows are less than the value of the Chi-Square distribution quantile, which equals 14.86 for the 0.005 significance level. Therefore, we go to step 3.

In step 3 we build the linear regression model (5) whose parameter estimates $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$ are equal to 0, 1.130412, 0.547413, and -0.041326, respectively. The estimate $\hat{\sigma}_\varepsilon$ is 0.1525.

Next, we test the normality of the distribution of residuals in the linear regression model (5) by the Pearson chi-squared test. According to this test, the null hypothesis $H_0$ that the observed frequency distribution of residuals is the same as the normal distribution is accepted because the $\chi^2$ test statistic, which equals 2.69, does not surpass the critical value from the Chi-Squared distribution which is 7.82 for 3 degrees of freedom and 0.05 significance level. Therefore, we go to step 5. We construct the nonlinear regression model (4) with estimates of parameters in the fourth iteration.

After that, we build the prediction interval of nonlinear regression by (7). In the fourth iteration, for the data normalized by the Johnson four-variate transformation for $S_B$ family from 40 rows of Table 1 (without rows 2, 7, 25, and 44), the matrix (8) is following

$$\mathbf{S}_Z = \begin{pmatrix} 40.00 & -18.22 & 13.94 \\ -18.22 & 40.00 & 4.82 \\ 13.94 & 4.82 & 40.00 \end{pmatrix}.$$

In Table 2, the lower and upper bounds of the prediction interval obtained in the fourth iteration are denoted as $LB_4$ and $UB_4$, respectively. At the fourth iteration, there are no outliers; the repeat of the iterations is completed, and the nonlinear regression model (6) is constructed using data from 40 apps.

Note, that the constructed nonlinear regression model (6) allows performing early size estimation (in KLOC) of open source apps in PHP based on the above three metrics (independent variables $X_1$, $X_2$, and $X_3$) that may be derived from a class diagram. In model (6) size $Y$ is the non-Gaussian dependent random variable and $\varepsilon$ is the Gaussian random variable with zero mean and a standard deviation estimate $\hat{\sigma}_\varepsilon$ of 0.1525.

To judge the prediction accuracy of the nonlinear regression model (6) we used the well-known prediction accuracy metrics such as a multiple coefficient of determination $R^2$, a mean magnitude of relative error MMRE, and prediction percentage at the level of magnitude of relative error of 0.25, PRED(0.25) [19, 20]. The $R^2$, MMRE, and PRED(0.25) values equal respectively 0.9812, 0.1753, and 0.750 for the nonlinear regression model (6) with the estimators of parameters which are calculated for the 40 data rows from Table 1 (without rows 2, 7, 25 and 44). These values indicate to us good prediction results of the nonlinear regression model (6), which is constructed by a modified technique for parameter estimators calculated from the 40 data rows.

Notice that if the nonlinear regression model (6) is built based on the data set from Table 1 by the technique [9] then we have three iterations. At the third iteration, there are no outliers in step 5, the repeat of the iterations is completed, and the nonlinear regression model (6) is constructed using data from 41 rows of Table 1 (without rows 2, 25, and 44). In this case, the $R^2$, MMRE, and PRED(0.25) values equal respectively 0.9776, 0.1795, and 0.6829 for the nonlinear regression model (6) with the estimates of parameters which are calculated for the 41 data rows from Table 1 (without rows 2, 25 and 44). It is not hard to see the values of all prediction accuracy metrics for the nonlinear regression model (6) which is constructed by the technique [9] are worse than in the case of using a modified technique we propose. This is especially true for PRED(0.25). Its value indicates unsatisfactory prediction accuracy since the recommended minimum value of PRED(0.25) should not be less than 0.75 [20].

Note, a more significant advantage of the model (6) constructed by a modified technique compared to the same model constructed using a nonmodified technique is the smaller widths of the confidence and prediction intervals. The width of the nonlinear regression prediction interval after modification is less than before modification by 67% for app 1 (row 1). Also, the width of the nonlinear regression confidence interval after modification is less than before modification by 57% for app 1. App 1 is Symfony, a PHP framework for web and console apps.

Such better prediction results for the nonlinear regression model (6), which is constructed by a modified technique for parameter estimates calculated from the 40 data rows by the maximum likelihood method (3) with the log-likelihood function (4) might be explained by, firstly, the normality of four-variate distribution of normalized four-dimensional data, secondly, the bigger number of data that has been discarded as outliers, and, thirdly, the normality of distribution of residuals in the linear regression model for normalized data.

## 4. Discussion and interpretations

The four-variate distribution of the software metrics from Table 1 is not Gaussian which the Mardia test for multivariate normality based on measures of the multivariate skewness and kurtosis indicates. Because we apply the statistical technique [15] to detect multivariate outliers in the four-dimensional non-Gaussian data from Table I based on the multivariate normalizing transformations and the MSD for normalized data. According to [15], there is a multivariate outlier in four-dimensional non-Gaussian data from Table 1 (data row 2) based on the Johnson four-variate transformation for the $S_B$ family. Note, that we have the multivariate outlier in the data from Table 1 (data row 1) without using normalization. This may be explained by the four-variate distribution of the data from Table 1 distribution differs significantly from Gaussian. That is why we rejected row 2 as a multivariate outlier.

We apply the Johnson four-variate transformation for the $S_B$ family to build the nonlinear regression model for estimating the size of open source PHP-based apps based on the modified technique [9] since there are outliers in the data from Table 1, which are detected in the model construction process by the MSD values for normalized data, nonlinear regression prediction intervals, and residuals (see Table 2).

We detected four outliers in the model construction process (rows 2, 7, 35, and 44). These data outliers are derived for the following four various apps: Faker, cphalcon, country-master, and PHP_CodeSniffer. Faker is a PHP library that generates fake data for users [21]. Phalcon is an open-

source web framework delivered as a C language extension for the PHP language providing high performance and lower resource consumption [22]. Country-master is an Android library project to create a login and registration form using phone numbers (uses Google's libphonenumber) [23]. PHP_CodeSniffer is a set of two PHP scripts that ensure your code remains clean and consistent [24]. The data of these so different apps are united by the fact that the average number of methods to the number of classes (predictor $X_2$) for them is in a fairly narrow range from 2.240 to 3.125. Therefore, we propose to divide the range of change of the predictor $X_2$ into two ranges: the first, from 0.48 to 2.22, and the second, from 3.14 to 32.60.

We propose to apply the nonlinear regression model (6) for estimating the size (in KLOC) of open source PHP-based apps such as frameworks, libs, and scripts around the following predictor ranges: from 2 to 2075 for $X_1$ (number of classes), from 0.48 to 2.22 or from 3.14 to 32.60 for $X_2$, and from 1.76 to 11.48 for $X_3$ (sum of average afferent and efferent coupling per class).

Also, we additionally calculated MRE (magnitude of relative error) values for prediction regression results by the model (6) for well-known PHP frameworks Symfony, Yii2, Laravel, and CakePHP (respectively data rows 1, 3, 6, and 10), which equal 0.0322, 0.2595, 0.1581, and 0.1748. Note, MRE values for prediction regression results by the model (6) before using the modified technique for the same PHP frameworks are 0.0775, 0.2748, 0.2594, and 0.1753, respectively. The above calculations indicate better prediction results of the size of open source PHP-based apps by the nonlinear regression model (6), which was constructed using the modified technique.

## 5. Conclusions

The technique for constructing nonlinear regression models based on the multivariate normalizing transformations and prediction intervals is modified by testing the normality of error distribution in the linear regression model for normalized data, which is applied to construct the nonlinear one. Using the example of building the nonlinear regression model with three predictors based on the Johnson four-variate transformation for the $S_B$ family to estimate the size (in KLOC) of open source PHP-based apps (such as frameworks, libs, and scripts), we have demonstrated that there may be multidimensional data sets (software metrics) for which the constructed nonlinear regression model has an unsatisfactory prediction accuracy of software size even after all outliers are removed as in the technique [9]. From the example, we have concluded that a modified technique is promising to apply, including its application for the construction of nonlinear regression models to estimate a software size. The nonlinear regression model, which is constructed by a modified technique, has better values of well-known prediction accuracy metrics such as $R^2$, MMRE, and PRED(0.25) compared to the model, which is only constructed based on the multivariate normalizing transformation and prediction intervals (without testing the normality of error distribution in the linear regression model for normalized data and discarding the multidimensional data point for which the value of the error modulus in the linear model is maximum if the distribution of residuals in the linear one is not Gaussian). Also, calculated MRE values for well-known PHP frameworks (Symfony, Yii2, Laravel, and CakePHP) indicate better prediction results of the size of open source PHP-based apps by the nonlinear regression model, which was constructed using the modified technique. Further research may be directed to search other multidimensional data sets to confirm that the modified technique works we have proposed, including the construction of nonlinear regression models for estimating the size of other types of apps and applying other multivariate normalizing transformations.

## 6. References


[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.

[2] D. M. Bates, D. G. Watts, Nonlinear Regression Analysis and Its Applications, 2nd ed., John Wiley & Sons, New York, 1988. doi:10.1002/9780470316757.

[3] G. A. F. Seber, C. J. Wild, Nonlinear Regression, John Wiley & Sons, New York, 1989. doi:10.1002/0471725315.

[4] T. P. Ryan, Modern Regression Methods, John Wiley & Sons, New York, 1997. doi:10.1002/9780470382806.

[5] N. R. Drapper, H. Smith, Applied Regression Analysis, 3rd ed., John Wiley & Sons, New York 1998.
[6] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis, Pearson Prentice Hall, 2007.
[7] S. Chatterjee, J. S. Simonoff, Handbook of Regression Analysis, John Wiley & Sons, New York, 2013.
[8] S. Prykhodko, N. Prykhodko, L. Makarova, A. Pukhalevych, Outlier detection in non-linear regression analysis based on the normalizing transformations, in: Proceedings of the 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), IEEE, 2020, pp. 407–410. doi:10.1109/TCSET49122.2020.235464.
[9] S. Prykhodko, N. Prykhodko, Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations, in: S. Shkarlet, A. Morozov, A. Palagin (eds.) Mathematical Modeling and Simulation of Systems (MODS'2020), MODS 2020, volume 1265 of Advances in Intelligent Systems and Computing, Springer, Cham., 2021, pp. 166–174. doi:10.1007/978-3-030-58124-4_16.
[10] H. J. Motulsky, R. E. Brown, Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics 7, 123 (2006) 1–20. doi:10.1186/1471-2105-7-123.
[11] D. C. Montgomery, E. A. Peck, Introduction to Linear Regression Analysis, 2nd ed., John Wiley & Sons, 1992.
[12] E. J. Pedhazur, Multiple Regression in Behavioral Research: Explanation and Prediction, 3rd ed., Fort Worth, Harcourt Brace College Publishers, 1997.
[13] G. A. F. Seber, A. J. Lee, Linear Regression Analysis, 2nd ed., John Wiley & Sons, 2003.
[14] S. Weisberg, Applied Linear Regression, 3rd ed., John Wiley & Sons, 2005.
[15] S. Prykhodko, N. Prykhodko, L. Makarova, K. Pugachenko, Detecting outliers in multivariate non-Gaussian data on the basis of normalizing transformations, in: Proceedings of the First Ukraine Conference on Electrical and Computer Engineering (UKRCON), IEEE, 2017, pp. 846–849. doi:10.1109/UKRCON.2017.8100366.
[16] S. Prykhodko, N. Prykhodko, L. Makarova, A. Pukhalevych, Application of the squared Mahalanobis distance for detecting outliers in multivariate non-Gaussian data, in: Proceedings of the 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), IEEE, 2018, pp. 962–965, doi: 10.1109/TCSET.2018.8336353.
[17] K. V. Mardia, Measures of multivariate skewness and kurtosis with applications, Biometrika 3(57), (1970) 519–530. doi:10.1093/biomet/57.3.519.
[18] K. V. Mardia, 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), 36(2) (1974) 115–128.
[19] L. C. Briand, I. Wieczorek, Resource modeling in software engineering, in: J. Marciniak (Ed.), Encyclopedia of Software Engineering, 2nd ed. John Wiley & Son, New York, 2002, pp. 1160–1196. doi:10.1002/0471028959.sof282.
[20] D. Port, M. Korte, Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research, in: Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, New York, 2008, pp. 51–60. doi:10.1145/1414004.1414015.
[21] fzaninotto/Faker, 2020. URL: https://github.com/fzaninotto/Faker.
[22] phalcon/cphalcon, 2021. URL: https://github.com/phalcon/cphalcon.
[23] uglymittens/country-master, 2015. URL: https://github.com/uglymittens/country-master.
[24] squizlabs/PHP_CodeSniffer, 2021. URL: https://github.com/squizlabs/PHP_CodeSniffer.