UDC 004.412:519.237

S.B.PRYKHODKO, N.V.PRYKHODKO, L.M.MAKAROVA, O.O.KUDIN, T.G.SMYKODUB
Admiral Makarov National University of Shipbuilding

CONSTRUCTING NON-LINEAR REGRESSION EQUATIONS ON THE BASIS OF BIVARIATE NORMALIZING TRANSFORMATIONS

The techniques for constructing equations, confidence and prediction intervals of non-linear regressions on the basis of bivariate normalizing transformations for non-Gaussian data are proposed. Application of the techniques is considered for the bivariate non-Gaussian data set: actual effort (hours) and size (adjusted function points) from 133 maintenance and development software projects.

Keywords: non-linear regression equation, confidence interval, prediction interval, normalizing transformation, bivariate non-Gaussian data

С.Б.ПРИХОДЬКО, Н.В.ПРИХОДЬКО, Л.М.МАКАРОВА, О.О.КУДІН, Т.Г.СМИКОДУБ
Національний університет кораблебудування імені адмірала Макарова

ПОБУДОВА НЕЛІНІЙНИХ РЕГРЕСІЙНИХ РІВНЯНЬ НА ОСНОВІ ДВОМІРНИХ НОРМАЛІЗУЮЧИХ ПЕРЕТВОРЕНЬ

Запропоновано методи побудови рівнянь, довірчих інтервалів та інтервалів передбачення нелінійних регресій на основі двомірних нормалізуючих перетворень для негаусовських даних. Застосування методів розглядається для одного набору двомірних негаусовських даних: для фактичної трудомісткості (години) і розміру (скориговані функціональні точки) зі 133 проектів з підтримки та розробки програмного забезпечення.

Ключові слова: нелінійне рівняння регресії, довірчий інтервал, інтервал передбачення, нормалізуюче перетворення, двовимірні негаусовські дані

С.Б.ПРИХОДЬКО, Н.В.ПРИХОДЬКО, Л.Н.МАКАРОВА, О.А.КУДИН, Т.Г.СМЫКОДУБ
Национальный университет кораблестроения имени адмирала Макарова

ПОСТРОЕНИЕ НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ УРАВНЕНИЙ НА ОСНОВЕ ДВУМЕРНЫХ НОРМАЛИЗУЮЩИХ ПРЕОБРАЗОВАНИЙ

Предложены методы построения уравнений, доверительных интервалов и интервалов предсказания нелинейных регрессий на основе двумерных нормализующих преобразований для негауссовских данных. Применение методов рассматривается для одного набора двумерных негауссовских данных: для фактической трудоемкости (часы) и размера (скорректированные функциональные точки) из 133 проектов по поддержке и разработке программного обеспечения.

Ключевые слова: нелинейное уравнение регрессии, доверительный интервал, интервал предсказания, нормализующее преобразование, двумерные негауссовские данные

Problem formulation

A normalizing transformation is often a good way to construct equations, confidence and prediction intervals of non-linear regressions [1-5], and it is often used for that purpose in empirical software engineering, information technology, biometry, ecology, finance, etc. However, well-known techniques for constructing equations, confidence and prediction intervals of non-linear regressions are based on univariate normalizing transformations, which do not take into account the correlation between random variables in the case of normalization of bivariate non-Gaussian data. This leads to the need to use the bivariate normalizing transformations, which take into account that correlation to construct equations, confidence and prediction intervals of non-linear regressions.

Analysis of recent research and publications

Transformations are an extremely important part of regression analysis, but the use of transformations can be somewhat tricky [2]. According [2] transformations are made for essentially four purposes, two of which are: first, to obtain approximate normality for the distribution of the error term (residuals) or the dependent random variable, 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 batter than the linear relationship between dependent and independent random variables. Now well-known normalizing transformations are used to construct the equations, confidence and prediction intervals of non-linear regressions. For that purpose, for example, it is known the application of such normalizing transformations as the decimal logarithm transformation [1-6], the Box-Cox transformation [2, 4], the Johnson translation system [7, 8]. However, known techniques for constructing equations, confidence and prediction intervals of 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 bivariate non-Gaussian data.

Purpose of the study

The purpose of the study is to propose the techniques for constructing the equations, confidence and prediction intervals of non-linear regressions for bivariate non-Gaussian data in general case, when it necessary to take into account the correlation between the response and the predictor (dependent and independent random variables) in the case of normalization of that variables.

Presentation of the main research material

We propose the techniques for constructing the equations, confidence and prediction intervals of non-linear regressions for bivariate non-Gaussian data. As and in [9, 10] the techniques consist of three steps. In the first step, a set of bivariate non-Gaussian data is normalized using a bijective bivariate normalizing transformation. In the second step, the equation, confidence and prediction intervals of non-linear regression for the normalized data are built. In the third step, the equations, confidence and prediction intervals of non-linear regressions for bivariate non-Gaussian data are constructed on the basis of the equation, confidence and prediction intervals of non-linear regression for the normalized data and the normalizing transformation.

The techniques. Consider bijective multivariate normalizing transformation of non-Gaussian random vector $\mathbf{P} = \{Y, X_1\}^T$ to Gaussian random vector $\mathbf{T} = \{Z_Y, Z_1\}^T$ is given by

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

and the inverse transformation for (1)

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

The linear regression equation for normalized data according to (1) will have the form [2]

$$\hat{Z}_Y = \overline{Z}_Y + \hat{b}_1\left(Z_1 - \overline{Z}_1\right), \tag{3}$$

where $\hat{Z}_Y$ is prediction linear regression equation result for values of Z_1; $\hat{b}_1$ is estimator for linear regression equation parameter b_1.

The non-linear regression equation will have the form

$$Y = \psi_1^{-1}\left[\overline{Z}_Y + \hat{b}_1\left(Z_1 - \overline{Z}_1\right)\right]. \tag{4}$$

The technique for constructing of confidence interval is based on transformation (1) and equation

$$Z_{Y_{CI}} = \hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_{Z_Y} \sqrt{\frac{1}{N} + \frac{\left(Z_{1_i} - \overline{Z}_1\right)^2}{S_{Z_1 Z_1}}}\,. \tag{5}$$

where $t_{\alpha/2,N-2}$ is a quantile of student's *t*-distribution with ν degrees of freedom and $\alpha/2$ significance level; $S_{Z_Y}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Z_{Y_i} - \overline{Z}_Y\right)^2$; $\overline{Z}_Y = \frac{1}{N}\sum_{i=1}^{N} Z_{Y_i}$; $S_{Z_1 Z_1} = \sum_{i=1}^{N}\left(Z_{1_i} - \overline{Z}_1\right)^2$.

The technique consists of three steps. In the first step, non-Gaussian data is normalized using a bijective normalizing transformation (1), and linear regression equation (3) is built on the basis of the normalized data. In the second step, the confidence interval for linear regression is detected. In the third step, the confidence interval for nonlinear regression is built on the basis of the confidence interval for linear regression and the normalizing transformation. The confidence interval for non-linear regression will have the form

$$Y_{CI} = \psi_1^{-1}\left(\hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_{Z_Y} \sqrt{\frac{1}{N} + \frac{\left(Z_{1_i} - \overline{Z}_1\right)^2}{S_{Z_1 Z_1}}}\right). \tag{6}$$

The technique for constructing a prediction interval is based on the transformation (1) and equation [2]

$$Z_{Y_{PI}} = \hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_{Z_Y} \sqrt{1 + \frac{1}{N} + \frac{\left(Z_{1_i} - \bar{Z}_1\right)^2}{S_{Z_1 Z_1}}}\,. \tag{7}$$

Like previously the technique consists of three steps, with the difference that instead of the confidence intervals, we define the prediction intervals. The prediction interval for non-linear regression will have the form

$$Y_{PI} = \psi_1^{-1}\left(\hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_{Z_Y} \sqrt{1 + \frac{1}{N} + \frac{\left(Z_{1_i} - \bar{Z}_1\right)^2}{S_{Z_1 Z_1}}} \right). \tag{8}$$

The equations (4), (6) and (8) are used for constructing the equations, confidence and prediction intervals of non-linear regressions for bivariate non-Gaussian data. The lines of equations, confidence and prediction intervals of non-linear regressions can also be built by the inverse transformation (2) of the values of variables $\hat{Z}_Y$, $Z_{Y_{CI}}$ and $Z_{Y_{PI}}$ from equation (3), (5) and (7) respectively.

Bivariate normalizing transformations. Some normalizing transformations have been proposed for normalizing bivariate non-Gaussian data, such as, transformation on the basis of the Box-Cox transformation, the Johnson translation system and others. However, only a few normalizing transformations are bijective. Such bijective transformation is the transformation of S_U family of the Johnson translation system. The Johnson normalizing translation is given by [11]

$$\mathbf{Z} = \boldsymbol{\gamma} + \boldsymbol{\eta}\mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{X} - \varphi)\right] \sim N_m(\mathbf{0}_m, \mathbf{S}), \tag{9}$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, φ and $\boldsymbol{\lambda}$ are parameters of the Johnson normalizing translation; $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$; $\boldsymbol{\eta} = diag(\eta_1, \eta_2)$; $\varphi = (\varphi_1, \varphi_2)^T$; $\boldsymbol{\lambda} = diag(\lambda_1, \lambda_2)$; $\mathbf{h}[(y_1, y_2)] = \{h_1(y_1), h_2(y_2)\}^T$; $h_i(.)$ is one of the translation functions

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

S is the covariance matrix

$$\mathbf{S} = \begin{pmatrix} S_{Z_1}^2 & S_{Z_1 Z_2} \\ S_{Z_1 Z_2} & S_{Z_2}^2 \end{pmatrix}.$$

Here $y = (x - \varphi)/\lambda$; $\operatorname{Arsh}(y) = \ln\left(y + \sqrt{y^2 + 1}\right)$.

The inverse transformation for the Johnson normalizing translation (9) is given by [11]

$$\mathbf{x} = \varphi + \boldsymbol{\lambda}\mathbf{h}^{-1}\left[\boldsymbol{\eta}^{-1}(\mathbf{z} - \boldsymbol{\gamma})\right]. \tag{10}$$

Example. We consider the example of constructing the equation, confidence and prediction intervals of non-linear regression for the bivariate non-Gaussian data set: actual effort (hours) and size (adjusted function points) from 133 maintenance and development projects [12] after the cutoff of12 outliers by the technique for detecting bivariate outliers on the basis of the normalizing transformations for non-Gaussian data [13]. On Fig. 1 the linear regression (solid line), the borders of confidence (dot-dash lines) and prediction (dotted lines) intervals ($\alpha = 0.05$) of linear regression for normalized data (points in the form of circles) from 133 projects are presented.

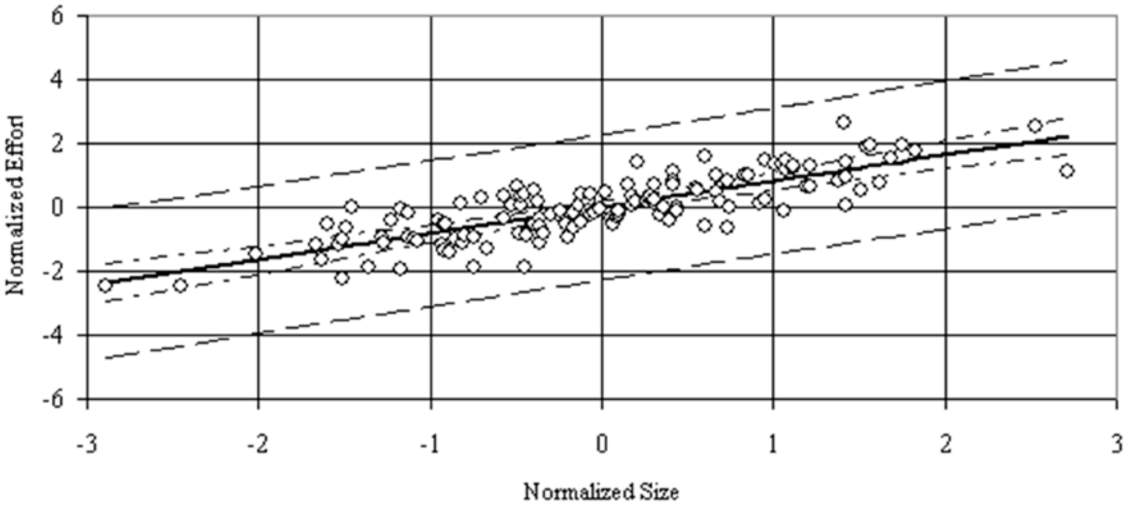

Fig. 1. Equation, confidence and prediction intervals of linear regression for normalized data from 133 projects

These data is normalized by S_B family of the transformation (9). In these case the point estimates of parameters are such: $\gamma_1 = 1.881055$, $\gamma_2 = 2.731134$, $\eta_1 = 0.793776$, $\eta_2 = 0.954031$, $\varphi_1 = 23.1890$, $\varphi_2 = 96.5557$, $\lambda_1 = 2768.509$ and $\lambda_2 = 28270.72$. The sample covariance matrix of the $\mathbf{Z}$ is used as the approximate moment-matching estimator of covariance matrix **S**

$$\mathbf{S}_N = \begin{pmatrix} 0.99248 & 0.81428 \\ 0.81428 & 0.99248 \end{pmatrix}.$$

On Fig. 2 the non-linear regression (solid line), the borders of confidence (dot-dash lines) and prediction (dotted lines) intervals ($\alpha = 0.05$) of non-linear regression for non-Gaussian data (points in the form of circles) from 133 projects are presented. That non-linear regression, the confidence and prediction intervals were built on the bases of transformations (9) and (10). Also the non-linear regression, the confidence and prediction intervals were built on the bases of the decimal logarithm transformation. For that transformation on Fig. 2 the borders of prediction interval (dotted lines with short dashes) are also presented. We note, in this case (the decimal logarithm transformation) at the maximum value of the independent variable the width of prediction interval is higher by 160 percent compared to prediction interval, which constructed on the bases of transformation (9).

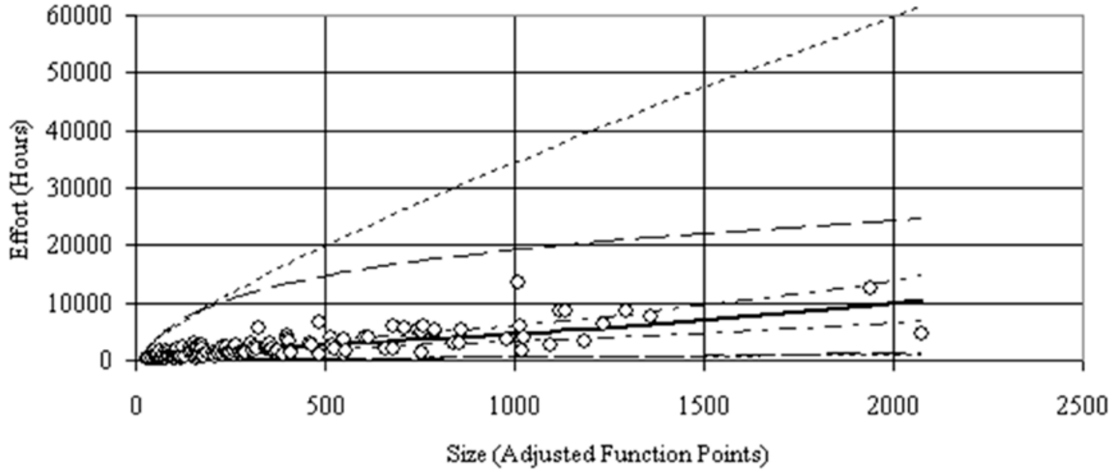

Fig. 2. Equation, confidence and prediction intervals of non-linear regression for non-Gaussian data from 133 projects

In our opinion, such a big difference is due to poor bivariate data normalization by the decimal logarithm transformation. We note, Mardia's multivariate kurtosis [14] β_2 equals 8 under bivariate normality for our case. The values of point estimate of kurtosis β_2 equal respectively 8.002 and 6.923 for the normalized data on Fig. 1 and the data, which normalized by the decimal logarithm transformation. These values indicate that the necessary condition for bivariate normality is practically performed for the normalized data by transformation (9) only.

At the same time the non-linear regressions, which were built on the bases of transformations (9) and the decimal logarithm transformation, are approximately similar: the values of coefficient of determination R^2 equal 0.5664 and 0.5759 respectively.

Conclusions

From the examples we conclude that the proposed techniques for constructing the equations, confidence and prediction intervals of non-linear regressions for bivariate non-Gaussian data are promising. Application of the techniques is considered for the bivariate non-Gaussian data set: actual effort (hours) and size (adjusted function points) from 133 maintenance and development software projects. Accounting the correlation between random variables in the case of normalization of that bivariate non-Gaussian data leads to reduction of the width of confidence and prediction intervals of the non-linear regression compared to the same intervals, which constructed on the bases of the decimal logarithm transformation. In the future, we intend to try other bivariate non-Gaussian data sets.

References

1. D.M. Bates and D.G. Watts. Nonlinear Regression Analysis and Its Applications. Wiley, 1988, 384 p.

2. T.P. Ryan. Modern regression methods. Wiley, 1997, 529 p.

3. G.A.F. Seber and C.J. Wild. Nonlinear Regression. John Wiley & Sons, Inc., 2003, 792 p.

4. R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, 800 p.

5. Iain Pardoe. Applied regression modelling. Wiley, 2012, 325 p.

6. S Chatterjee and J.S. Simonoff. Handbook of Regression Analysis. John Wiley & Sons, Inc., 2013, 236 p.

7. S.B. Prykhodko and A.V. Pukhalevich, “Developing PC Software Project Duration Model based on Johnson transformation”, in “Modern Problems of Radio Engineering, Telecommunications and Computer Science”, Proceedings of the 12th International Conference, Lviv-Slavske, Ukraine, 25 February - 1 March, 2014, pp. 114-116.

8. S.B. Prykhodko and A.V. Pukhalevich, “Confidence interval estimation of PC software project duration regression based on Johnson transformation” Radioelectronic and Computer Systems, No 2, Vol. 66, pp. 104-107, 2014.

9. S.B. Prykhodko, “Statistical anomaly detection techniques based on normalizing transformations for non-Gaussian data”, in “Computational Intelligence (Results, Problems and Perspectives)”, Proceedings of the International Conference, Kyiv-Cherkasy, Ukraine, May 12-15, 2015, pp. 286-287.

10. 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” (PTTAS-2016), Abstracts of the Research and Practice Seminar, Poltava, Ukraine, May 25-26, 2016, pp. 6-7.

11. P.M. Stanfield, J.R. Wilson, G.A. Mirka, N.F. Glasscock, J.P. Psihogios, 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.

12. B. Kitchenham, S.L. Pfleeger, B. McColl, and S. Eagan, “An empirical study of maintenance and development estimation accuracy”, The Journal of Systems and Software, 64, pp.57-77, 2002.

13. S. Prykhodko, N. Prykhodko, L. Makarova, O. Kudin, T. Smykodub and A. Prykhodko, “Detecting bivariate outliers on the basis of normalizing transformations for non-Gaussian data” in “Advanced Information Systems and Technologies”, Proceedings of the V International Scientific Conference, Sumy, Ukraine, May 17-19, pp. 95-97, 2017.

14. K.V. Mardia, “Measures of multivariate skewness and kurtosis with applications”, Biometrika, 57, pp. 519–530, 1970.