Mathematical Modeling of Effort of Mobile Application Development in a Planning Phase

Sergiy Prykhodko[1[0000-0002-2325-018X]], Natalia Prykhodko[1[0000-0002-3554-7183]], Kateryna Knyrik[1[0000-0001-8434-4035]] and Andrii Pukhalevych[1[0000-0002-8827-3251]]

[1] Admiral Makarov National University of Shipbuilding, Mykolaiv, 54025, Ukraine
sergiy.prykhodko@nuos.edu.ua

Abstract. Mathematical modeling of effort of development of mobile applications (apps) by non-linear regression model using multivariate normalizing transformation is performed. A three-factor non-linear regression model to estimate the effort (in man-hours) of developing the mobile apps in a planning phase is constructed on the basis of the Johnson four-variate transformation for S_B family. This model is constructed around the Requirement Analysis Document (RAD) variables: number of screens, number of functions, and number of files. Comparison of the constructed model with the linear regression model and non-linear regression models based on the univariate normalizing transformations is performed. This model, in comparison with other regression models, has a larger multiple coefficient of determination, a smaller value of the mean magnitude of relative error, a larger value of percentage of prediction, and smaller widths of the confidence and prediction intervals of regression. Such a good result for the constructed model may be explained best multivariate normalization of the non-Gaussian data set, which used to build the three-factor non-linear regression model based on the Johnson four-variate transformation for S_B family.

Keywords: Mathematical Modeling, Effort Estimation, Mobile Application, Non-linear Regression Model, Prediction Interval.

1 Introduction

The problem of estimating software development effort is one of the important ones in the planning phase, which is the first of the five phases of the software development lifecycle [1]. Today, the solution of this problem is carried out, including using mathematical modeling. One of the more well-known mathematical models for estimating software development effort is COCOMO II. But its use for mobile apps has some difficulties. First, the main factor for this model is the size of the software, which is still unknown in the planning phase. Second, COCOMO II is a non-linear regression equation built on a univariate transformation in the form of a decimal logarithm, which does not always allow for proper normalization of the data. In addition, the regression equation does not include random variables [2-4] as and a effort estimation model based on Function Points Analysis method [5]. And, as you know, the effort is a random variable. Third, while mobile app development is similar to web app development and has its roots in more traditional software development, however, one significant difference is that mobile apps are often written specifically to take advantage of the unique features that a particular mobile device offers [6].

Therefore, over the last decade, the various models for forecasting the effort of developing the mobile apps in a planning phase, including regression ones [7, 8], were constructed. It is the regression models that describe an effort as a random variable. And since the effort distribution is not Gaussian, it is necessary to use non-linear regression models, and their construction should be based on multivariate normalizing transformations [9].

2 Model construction

At first, the three-factor linear regression model to estimate the effort Y (in man-hours) of developing the mobile apps in a planning phase is constructed for the four-dimensional data set from Table 1. This model is constructed around the Requirement Analysis Document (RAD) variables: number of screens X_1, number of functions X_2, and number of files X_3.

Table 1. The data set and MD^2 values.

No	Y	X_1	X_2	X_3	MD^2	No	Y	X_1	X_2	X_3	MD^2
1	192	5	4	3	0.66	20	198	6	5	4	0.50
2	272	5	4	3	2.31	21	146	4	3	2	1.18
3	288	3	2	2	6.43	22	191	6	6	5	0.96
4	116	6	6	4	0.95	23	99	3	3	2	1.47
5	372	5	5	4	6.82	24	382	11	12	9	8.35
6	504	9	8	6	10.55	25	270	9	10	8	4.84
7	28	6	7	2	7.11	26	282	12	7	3	7.16
8	176	6	7	3	4.53	27	213	10	5	2	6.14
9	364	10	11	9	6.90	28	322	11	7	5	4.32
10	120	10	10	5	6.76	29	290	10	6	4	3.67
11	22	6	5	4	6.72	30	223	7	7	6	1.69
12	224	11	6	2	7.08	31	241	5	5	6	4.95
13	24	2	2	1	3.05	32	87	5	5	2	1.53
14	200	11	7	4	4.88	33	36	3	3	1	2.24
15	160	6	6	7	9.41	34	216	8	7	5	0.54
16	120	2	2	1	2.86	35	67	5	6	2	4.26
17	96	4	4	1	2.60	36	115	7	7	3	2.59
18	202	6	5	4	0.49	37	36	2	2	1	2.84
19	145	4	3	2	1.17	38	98	3	3	2	1.47

The data set from Table 1 was obtained by combining two data sets for 17 mobile apps from [5] and for 21 mobile apps (rows 18 to 38). Also, Table 1 contains the values of squared Mahalanobis distance (MD^2). We use the technique based on the squared Mahalanobis distance [10] for detecting the outliers in the data from Table 1. There are no outliers in the data from Table 1 for 0.005 significance level, since for all data rows, the MD^2 values are smaller than the value of the quantile of the Chi-Square distribution, which equals to 14.86.

Following [2-4] the three-factor linear regression model has the form

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \varepsilon_x , \quad (1)$$

where ε_x is a Gaussian random variable which defines residuals, $\varepsilon_x \sim N(0, \sigma_x)$; the estimators for parameters of the model (1) are: $\hat{b}_0 = 0.26513$, $\hat{b}_1 = 0.23116$, $\hat{b}_2 = -0.00082$, $\hat{b}_3 = 0.08374$. Parameters of the model (1) were estimated by the least square method.

To judge the prediction accuracy of linear regression model (1) we first used the well-known standard metrics of prediction accuracy, i.e., a multiple coefficient of determination R^2, a mean magnitude of relative error MMRE and percentage of prediction at the level of magnitude of relative error (MRE), which equals 0.25, PRED(0.25) [11, 12]. The values of R^2, MMRE, and PRED(0.25) equal respectively 0.5449, 0.5713, and 0.5789 for the linear regression model (1). These values show us bad prediction results of the regression model (1).

Besides, the null hypothesis that the observed frequency distribution of residuals for the linear regression model (1) is the same as the normal distribution was tested by Pearson's chi-squared test. There is a reason to reject the null hypothesis that the distribution of residuals for the model (1) is the same as the normal distribution, since the chi-squared test statistic value equals to 13.33 is higher than the critical value of the chi-square, which equals to 7.81 for 3 degrees of freedom and 0.05 significance level. Also, for the distribution of residuals in linear regression model (1), estimators of skewness and kurtosis equal to 0.78 and 5.69, respectively. Although for the Gaussian distribution, the values of skewness and kurtosis equal to 0 and 3, respectively.

It is known [2], one of the underlying assumptions that justify the use of linear regression models is the normality of the distribution of residuals. But this assumption is not valid for the linear regression model (1). What leads to the need to construct a multiple non-linear regression model to estimate the effort of developing the mobile apps in a planning phase.

The three-factor non-linear regression model to estimate the effort of developing the mobile apps in a planning phase was constructed based on the Johnson four-variate transformation for S_B family according [9]. The three-factor non-linear regression model has the form [9]

$$Y = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y\right)/\hat{\eta}_Y} \right]^{-1} , \quad (2)$$

where ε is a Gaussian random variable which defines residuals, $\varepsilon \sim N(0,1)$; $\hat{Z}_Y$ is a prediction result by linear regression equation for normalized data, which were transformed using the Johnson four-variate transformation for S_B family, $\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3$; $Z_j = \gamma_j + \eta_j \ln \frac{X_j - \varphi_j}{\varphi_j + \lambda_j - X_j}$, $\varphi_j < X_j < \varphi_j + \lambda_j$, $j = 1,2,3$; the estimators for parameters of the Johnson four-variate transformation for S_B family are: $\hat{\gamma}_Y = 5.69898$, $\hat{\gamma}_1 = 0.524119$, $\hat{\gamma}_2 = 0.776179$, $\hat{\gamma}_3 = 0.540973$, $\hat{\eta}_Y = 2.40219$, $\hat{\eta}_1 = 0.743879$, $\hat{\eta}_2 = 0.79545$, $\hat{\eta}_3 = 0.534447$, $\hat{\varphi}_Y = \text{-}114.5452$, $\hat{\varphi}_1 = 1.7242$, $\hat{\varphi}_2 = 1.6885$, $\hat{\varphi}_3 = 0.90$, $\hat{\lambda}_Y = 3328.564$, $\hat{\lambda}_1 = 12.3743$, $\hat{\lambda}_2 = 12.091$, $\hat{\lambda}_3 = 8.30648$; the estimators for parameters of the linear regression equation for normalized data are: $\hat{b}_0 = 0$, $\hat{b}_1 = 0.808152$, $\hat{b}_2 = \text{-}0.928296$, $\hat{b}_3 = 0.854262$. Parameters of the linear regression equation for normalized data were estimated by the least square method.

The values of R^2, MMRE, and PRED(0.25) equal respectively 0.5789, 0.4933 and 0.5263 for non-linear regression model (2). These values show us bad prediction results of the non-linear regression model (2) approximately also as for the linear regression model (1).

Because of this, the method [13] for improving non-linear regression models was further used to construct a non-linear regression model to estimate the effort of developing the mobile apps in a planning phase. The method [13] consists of four stages. In the first stage, a set of multivariate non-Gaussian data is normalized using a multivariate normalizing transformation. After that, normalized data are checked for outliers, and, if ones are detected, outliers are cut off. The method based on the squared Mahalanobis distance [14] is used for outlier detection. In the second stage, the non-linear regression model is constructed based on the multivariate normalizing transformation [9]. In the third stage, the prediction intervals of non-linear regression is built according [9]. And finally, at the fourth stage, it is checked whether among the data for which the non-linear regression model was built, those that go beyond the found boundaries of the prediction interval. And, if the outliers are detected, they are cut off, and we repeat all the stages, starting with the first, for new data.

For the non-linear regression model (2) with the parameter estimators obtained from the data in Table 1 of the 38 mobile apps, it turned out that *Y* values for the three apps (5, 6, and 11) go beyond the prediction interval. In Table 2, the lower bound of the prediction interval obtained in the first iteration is denoted as LB_1, and the upper bound is denoted as UB_1. In the second iteration, data from three mobile apps (5, 6, and 11) were cut off, and data from the remaining 35 apps were used for model construction. For the model (2) with the parameter estimators obtained from the data in Table 1 of the 35 mobile apps, it turned out that the value of *Y* for app 17 goes beyond the prediction interval. There were four such iterations, after which 30 mobile apps remained (1, 3, 4, 7, 9, 10, 12-14, 18-38). At the fifth iteration, there were no outliers; the repeat of the stages was completed, the nonlinear regression model (2) was constructed using data from 30 apps. In Table 2, the lower bound of the prediction interval obtained in the fifth iteration is denoted as LB_5, and the upper bound is denoted as UB_5. The row numbers (i.e., mobile apps) with the outliers in data are highlighted in bold. A dash (-) depicts the exclusion of the corresponding numbers of data in the relevant iteration (i.e., iteration 5).

Table 2. Lower and upper bounds nonlinear regression before and after outlier cutoff.

No	*Y*	LB_1	UB_1	LB_5	UB_5	No	*Y*	LB_1	UB_1	LB_5	UB_5
1	192	60.5	377.3	131.4	228.5	20	198	70.8	402.2	148.4	248.2
2	**272**	**60.5**	**377.3**	-	-	21	146	49.2	353.3	115.3	209.6
3	288	88.6	524.1	218.1	332.6	22	191	66.2	392.5	139.8	238.7
4	116	51.1	352.9	105.7	195.7	23	99	24.7	290.0	73.0	149.9
5	**372**	**54.5**	**362.4**	-	-	24	382	140.1	624.9	317.2	402.2
6	**504**	**90.1**	**453.3**	-	-	25	270	93.4	477.2	202.6	309.0
7	28	-0.7	232.5	25.1	70.9	26	282	104.6	532.5	223.9	332.1
8	**176**	**18.9**	**277.8**	-	-	27	213	78.5	452.7	158.6	265.3
9	364	157.4	665.2	331.6	411.5	28	322	126.8	560.3	257.5	355.1
10	120	48.7	363.8	97.1	188.5	29	290	109.1	513.1	219.5	322.5
11	**22**	**70.8**	**402.2**	-	-	30	223	78.6	425.2	164.1	266.7
12	224	73.5	447.0	148.5	255.8	31	241	84.9	449.3	194.4	299.7
13	24	-23.9	170.9	15.3	51.1	32	87	17.1	267.3	49.2	111.2
14	200	106.5	511.6	214.3	318.8	33	36	-29.0	153.6	15.2	49.6
15	**160**	**100.6**	**490.0**	-	-	34	216	77.1	418.6	153.2	253.9
16	**120**	**-23.9**	**170.9**	-	-	35	67	1.4	233.2	29.0	77.0
17	**96**	**-33.4**	**149.2**	-	-	36	115	31.0	306.4	64.9	137.9
18	202	70.8	402.2	148.4	248.2	37	36	-23.9	170.9	15.3	51.1
19	145	49.2	353.3	115.3	209.6	38	98	24.7	290.0	73.0	149.9

In the fifth iteration, for the data from 30 mobile apps, the estimators of parameters of the Johnson four-variate transformation for S_B family are: $\hat{\gamma}_Y = 0.58590$, $\hat{\gamma}_1 = 0.316749$, $\hat{\gamma}_2 = 0.86299$, $\hat{\gamma}_3 = 0.48606$, $\hat{\eta}_Y = 1.01714$, $\hat{\eta}_1 = 0.63606$, $\hat{\eta}_2 = 0.86557$, $\hat{\eta}_3 = 0.612856$, $\hat{\varphi}_Y = -12.7422$, $\hat{\varphi}_1 = 1.84255$, $\hat{\varphi}_2 = 1.5560$, $\hat{\varphi}_3 = 0.73913$, $\hat{\lambda}_Y = 500.266$, $\hat{\lambda}_1 = 11.3796$, $\hat{\lambda}_2 = 13.2488$, $\hat{\lambda}_3 = 8.52637$; the estimators for parameters of the linear regression equation for normalized data are: $\hat{b}_0 = 0$, $\hat{b}_1 = 1.1190$, $\hat{b}_2 = -1.3765$, $\hat{b}_3 = 1.2027$.

The values of R^2, MMRE, and PRED(0.25) equal respectively 0.965, 0.117 and 0.867 for non-linear regression model (2). These values show us good prediction results of the non-linear regression model (2) with parameter estimators obtained from the data in Table 1 of the 30 mobile apps.

Following [9], appropriate equations were constructed to determine the lower and upper bounds of the non-linear regression prediction intervals

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{1+\frac{1}{N}+\left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right), \quad (3)$$

where ψ_Y is a first component of a vector of normalizing transformation, $\boldsymbol{\psi} = \{\psi_Y, \psi_1, \psi_2, \ldots, \psi_k\}^T$; k is a number of factors (regressors or independent variables); $t_{\alpha/2,\nu}$ is a quantile of student's t-distribution with $\alpha/2$ significance level and ν degrees of freedom; $\mathbf{Z}_X^+$ is a matrix of centered regressors that contains the values of normalized data $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$; $\mathbf{z}_X^+$ is a vector with components $Z_{1_i} - \overline{Z}_1$, $Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$ for i-row; $S_{Z_Y}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$, $\nu = N - k - 1$; $\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+$ is $k \times k$ matrix

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

where $S_{Z_qZ_r} = \sum_{i=1}^{N}\left[Z_{q_i} - \overline{Z}_q\right]\left[Z_{r_i} - \overline{Z}_r\right]$, $q, r = 1, 2, \ldots, k$. In our case, k=3.

In the fifth iteration, for the data which normalized by the Johnson four-variate transformation for S_B family from 30 mobile apps, 3×3 matrix

$$\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+ = \begin{pmatrix} 29.8 & 25.5 & 19.1 \\ 25.5 & 30.2 & 24.5 \\ 19.1 & 24.5 & 29.7 \end{pmatrix}.$$

3 Comparison of models

Also, for comparison of the model (2) with other models, a linear regression model and nonlinear regression models on the basis of the univariate decimal logarithm transformation (Log10) and the Johnson univariate transformation for the S_B family were constructed for data from Table 1 of the 30 mobile apps. The three-factor linear regression model for data from Table 1 of the 30 apps has the form

$$\hat{Y} = 40{,}250 + 28{,}973\, X_1 - 41{,}798\, X_2 + 50{,}665\, X_3 + \varepsilon_x . \quad (4)$$

The three-factor non-linear regression model is constructed based on the decimal logarithm transformation for data from Table 1 of the 30 apps

$$Y = 10^{\varepsilon_x + \hat{b}_0} X_1^{\hat{b}_1} X_2^{\hat{b}_2} X_3^{\hat{b}_3}, \qquad (5)$$

where the estimators for parameters are: $\hat{b}_0 = 1.73898$, $\hat{b}_1 = 1.6687$, $\hat{b}_2 = \text{-}2.1116$, $\hat{b}_3 = 1.30125$.

The three-factor non-linear regression model based on the Johnson univariate transformation for the S_B family has the form (2) with only the following parameter estimators: $\hat{b}_3 = 1.1148$ $\hat{\gamma}_Y = 0.25204$, $\hat{\gamma}_1 = 0.10255$, $\hat{\gamma}_2 = 0.49345$, $\hat{\gamma}_3 = 0.61963$, $\hat{\eta}_Y = 0.58192$, $\hat{\eta}_1 = 0.51359$, $\hat{\eta}_2 = 0.63352$, $\hat{\eta}_3 = 0.58967$, $\hat{\varphi}_Y = 19.9286$, $\hat{\varphi}_1 = 1.90$, $\hat{\varphi}_2 = 1.81688$, $\hat{\varphi}_3 = 0.90$, $\hat{\lambda}_Y = 370.175$, $\hat{\lambda}_1 = 10.20$, $\hat{\lambda}_2 = 10.6468$, $\hat{\lambda}_3 = 8.6277$, $\hat{b}_0 = 0$, $\hat{b}_1 = 0.60292$, $\hat{b}_2 = \text{-}0.80179$, $\hat{b}_3 = 1{,}1148$. Parameters of the Johnson transformation for S_B family were estimated by the maximum likelihood method.

The values of R^2, MMRE and PRED(0.25) equal respectively 0.838, 0.237 and 0.733 for linear regression model (4), and equal respectively 0.789, 0.206 and 0.733 the model (5), and equal respectively 0.878, 0.190 and 0.767 for the model (2) with estimators of parameters for the Johnson univariate transformation. The values of R^2, MMRE, and PRED(0.25), which equal respectively 0.965, 0.117, and 0.867, is better for the model (2) with estimators of parameters for the Johnson four-variate transformation in comparison with all previous models.

The null hypothesis that the distribution of residuals for the linear regression model (4) is the same as the normal distribution was tested by Pearson's chi-squared test. There is a reason to reject the null hypothesis that the distribution of residuals for the linear regression model (4) is the same as the normal distribution, since the chi-squared test statistic value equals to 10.78 is higher than the critical value of the chi-square, which equals to 7.81 for 3 degrees of freedom and 0.05 significance level. Also, for the distribution of residuals in linear regression model (4), estimators of skewness and kurtosis equal respectively to 1.52 and 7.73. There is no reason to reject the null hypothesis that the distribution of residuals for nonlinear regression models (2) and (5) is the same as the normal distribution, since the chi-squared test statistic values are less than the critical value of the chi-square, which equals to 7.81. The chi-squared test statistic values equal to 4.78, 2.91, and 2.30 for the distribution of residuals in nonlinear regression models (5), (2) with estimators of parameters for the Johnson univariate transformation and (2) with estimators of parameters for the Johnson four-variate transformation respectively. For the distribution of residuals in nonlinear regression models (2) and (5), estimators of skewness and kurtosis are close to 0 and 3, respectively. Only the estimator of kurtosis equals to 5.39 for the distribution of residuals in the nonlinear regression model (2) with estimators of parameters for the Johnson univariate transformation for the S_B family.

The lower (LB) and upper (UB) bounds of the linear regression and non-linear regression prediction intervals were also determined by (3) based on the decimal logarithm transformation, Johnson's univariate and four-variate transformations for a significance level of 0.05. These bounds are shown in Table 3.

Table 3. Lower and upper bounds of prediction intervals for regressions.

No	*Y*	linear regression		Log10 univariate		Johnson univariate		Johnson four-variate	
		LB	UB	LB	UB	LB	UB	LB	UB
1	192	80.7	259.1	104.3	310.1	68.5	302.5	131.4	228.5
3	288	52.8	237.0	108.0	354.1	95.9	352.6	218.1	332.6
4	116	77.3	254.6	87.5	259.1	71.0	306.0	105.7	195.7
7	28	-75.2	120.8	24.4	79.7	28.3	155.6	25.1	70.9
9	364	229.5	422.9	159.5	499.1	269.4	383.7	331.6	411.5
10	120	69.9	260.7	91.9	280.6	70.5	312.2	97.1	188.5
12	224	113.5	305.5	93.0	303.6	60.0	299.9	148.5	255.8
13	24	-26.6	157.1	22.6	71.9	21.6	59.9	15.3	51.1
14	200	176.6	361.5	171.2	522.3	129.0	355.4	214.3	318.8
18	202	118.6	296.9	128.4	381.5	89.1	327.4	148.4	248.2
19	145	42.2	222.0	77.3	233.0	49.7	262.8	115.3	209.6
20	198	118.6	296.9	128.4	381.5	89.1	327.4	148.4	248.2
21	146	42.2	222.0	77.3	233.0	49.7	262.8	115.3	209.6
22	191	127.0	306.3	116.3	348.6	99.1	336.5	139.8	238.7
23	99	12.7	193.5	47.7	144.5	40.2	227.7	73.0	149.9
24	382	215.8	410.9	155.4	487.5	204.8	378.0	317.2	402.2
25	270	194.1	382.5	141.0	437.2	179.3	370.8	202.6	309.0
26	282	151.5	343.2	134.0	421.8	168.9	375.6	223.9	332.1
27	213	127.5	317.1	116.6	380.3	59.9	294.2	158.6	265.3
28	322	226.4	413.0	228.3	700.0	174.3	369.1	257.5	355.1
29	290	189.5	374.2	201.4	619.1	125.1	352.3	219.5	322.5
30	223	163.9	345.1	137.2	414.2	126.6	352.8	164.1	266.7
31	241	184.3	376.0	156.0	490.8	143.7	361.7	194.4	299.7
32	87	-13.1	168.0	38.1	115.2	33.9	191.9	49.2	111.2
33	36	-38.8	143.7	18.9	60.0	21.0	48.9	15.2	49.6
34	216	143.9	321.6	136.4	404.8	105.4	340.0	153.2	253.9
35	67	-58.8	130.2	25.3	80.4	29.4	162.2	29.0	77.0
36	115	10.6	194.3	55.7	168.0	45.5	249.8	64.9	137.9
37	36	-26.6	157.1	22.6	71.9	21.6	59.9	15.3	51.1
38	98	12.7	193.5	47.7	144.5	40.2	227.7	73.0	149.9

Note that the width of the non-linear regression prediction interval based on the Johnson four-variate transformation is less than after the Johnson univariate transformation for 29 from 30 data rows (except one with number 25), smaller than after decimal log transformation and less compared with the linear regression prediction interval width for all 30 data rows. Approximately the same results were obtained for the confidence intervals of regressions. Herewith a confidence interval of non-linear regression is defined as (3) with the only difference that in the sum in curly brackets, there will not be 1.

Such good prediction results for the constructed model may be explained best multivariate normalization of the non-Gaussian data set, which used to build the three-factor non-linear regression model based on the Johnson four-variate transformation for S_B family. The measures of multivariate skewness β_1 and kurtosis β_2 [15] allow one to test two hypotheses that are compatible with the assumption of multivariate normality. In our case for 30 apps $\beta_1 = 4$ and $\beta_2 = 24$. The estimators of multivariate skewness and kurtosis equal 8.42, 5.44, 12.86, 6.82, and 26.78, 23.08, 33.57, 25.71 for the data for 30 apps from Table 1, the normalized data on the basis of the decimal logarithm transformation, the Johnson univariate and four-variate transformations respectively. The values of these estimators indicate that the necessary condition for multivariate normality is approximately performed for the normalized data on the basis of the decimal logarithm and the Johnson four-variate transformation. Also, multivariate normality was tested by MD^2 [16]. A multivariate normality condition is only performed for the normalized data on the basis of the decimal logarithm and the Johnson four-variate transformation, since for all 30 rows of the normalized data, the MD^2 values are smaller than the value of the quantile of the Chi-Square distribution, which equals to 14.86 for 0.005 significance level.

4 Conclusions

Mathematical modeling of effort of development of mobile apps by non-linear regression model using multivariate normalizing transformation is performed. A three-factor non-linear regression model to estimate the effort of developing the mobile apps in a planning phase is firstly constructed on the basis of the Johnson four-variate transformation for S_B family. This model, in comparison with other regression models (both linear and non-linear), has a more significant multiple coefficient of determination, a smaller value of the mean magnitude of relative error, a more significant value of percentage of prediction, and smaller widths of the confidence and prediction intervals of regression. An example of the construction of the three-factor non-linear regression model confirms the efficiency of the method for improving non-linear regression models on the basis of multivariate normalizing transformations, the squared Mahalanobis distance, and prediction intervals. Prospects for further research may include the application of other data sets to construct the multiple non-linear regression models for estimating the effort of developing the mobile apps in a planning phase.

References

1. Zhu, H.: Software design methodology: From principles to architectural styles. Butterworth-Heinemann, Elsevier, Oxford (2005).
2. Ryan T.P.: Modern regression methods. 2nd edn. John Wiley & Sons, New York (2008).
3. Chatterjee, S., Simonoff, J.S.: Handbook of Regression Analysis. John Wiley & Sons, New York (2013).
4. Drapper, N.R., Smith, H.: Applied Regression Analysis. John Wiley & Sons, New York (1998).
5. Arnuphaptrairong, T., Suksawasd, W.: An empirical validation of mobile application effort estimation models. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2017), pp. 697-701. Newswood Limited, Hong Kong (2017).
6. Rouse, M.: Mobile application development, https://searchmicroservices.techtarget.com/definition/mobile-application-development, last accessed 2019/10/12.
7. Francese, R., Gravino, C., Risi, M., Scanniello, G., Tortora, G.: On the use of requirements measures to predict software project and product measures in the context of Android mobile apps: A preliminary study. In: Proceedings of the 41st Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2015), pp. 357-364. IEEE Computer Society, Funchal (2015). doi: 10.1109/SEAA.2015.22
8. Shahwaiz, S.A., Malik, A.A., Sabahat N.: A parametric effort estimation model for mobile apps. In: Proceedings of the 19th International Multi-Topic Conference (INMIC 2016), pp. 1-6. IEEE, Islamabad (2016). doi: 10.1109/INMIC.2016.7840114
9. Prykhodko, N.V., Prykhodko, S.B.: Constructing the non-linear regression models on the basis of multivariate normalizing transformations. Electronic modeling 6(40), 101-110 (2018). doi: 10.15407/emodel.40.06.101
10. Johnson, R.A., Wichern, D.W.: Applied multivariate statistical analysis. Pearson Prentice Hall (2007).
11. Foss, T., Stensrud, E., Kitchenham, B., Myrtveit, I.: A simulation study of the model evaluation criterion MMRE. IEEE Transactions on software engineering 11(29), 985–995 (2003).
12. Port, D., Korte, M.: 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, pp. 51–60. ACM, New York (2008).
13. Prykhodko, S.B., Prykhodko, N.V.: A method for improving non-linear regression models based on multivariate normalizing transformations. In: Proceedings of the 3d International Conference on Applied Scientific and Technical Research, pp. 20. Symfoniya fortu, Ivano-Frankivsk (2019). (in Ukrainian).
14. Prykhodko, S., Prykhodko, N., Makarova, L., Pukhalevych, A.: 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 Radioelectronics, Telecommunications and Computer Engineering (TCSET), pp. 962-965. IEEE, Lviv-Slavske (2018). doi: 10.1109/TCSET.2018.8336353
15. Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika 3(57), 519–530 (1970). doi: 10.1093/biomet/57.3.519
16. Olkin, I., Sampson, A.R.: Multivariate Analysis: Overview. In: Smelser, N.J., Baltes, P.B. (eds.) International encyclopedia of social & behavioral sciences. 1st edn. Elsevier, Pergamon (2001).