UDC 005.8

S. B. PRYKHODKO, A. V. PUKHALEVYCH

National University of Shipbuilding named after admiral Makarov, Ukraine

CONFIDENCE INTERVAL ESTIMATION OF PC SOFTWARE PROJECT DURATION REGRESSION BASED ON JOHNSON TRANSFORMATION

The regression models of software project duration based on project effort are considered. The aim of the research is to derive equations of confidence interval of a non-linear regression model of software project duration based on Johnson transformation for the personal computer (PC) development platform. Non-linear regression models of software project duration based on project effort are widely used. However, existing models do not provide equations to estimate the confidence interval of non-linear regression. Therefore, it is required to build a confidence interval of non-linear regression. In this paper, equations of a confidence interval of a non-linear regression model of software project duration based on Johnson transformation from the S_B family are derived for the PC development platform. Mathematical statistics, regression and interval analysis methods are used to analyze data samples and derive equations of a confidence interval of a non-linear regression model.

Keywords: *non-linear regression model, software project duration, Johnson transformation, project management, time management, confidence interval.*

Introduction

Many parametric models based on project effort have been proposed in the literature to predict the duration of software development projects. Among these, COCOMO that was built by Boehm (1981) has received wide attention. COCOMO [1] is a non-linear regression model of software project duration based on project effort. COCOMO models were built for three development platforms. Oligny et al. (1997) derived another regression model of software project duration based on project effort from a set of historical data maintained by the *International Software Benchmarking Standards Group* (ISBSG) [*3*]. Oligny et al. (2000) built separate duration models for subsets of projects for personal computer (PC), mid-range (MR) and mainframe (MF) development platforms using the same ISBSG data set [4].

Distribution of software project duration and effort is not normal. Therefore, it is impossible to develop an adequate linear regression model and there is a need to develop non-linear regression models. This article describes one of the effective methods for non-linear regression model building that does not require brute force, based on the application of normalizing transformations.

Deriving a non-linear regression model based on normalizing transformation is performed in three steps:

1) empirical data is normalized using normalizing transformation;

2) a linear regression model is derived from the normalized data;

3) a non-linear regression model is built based on normalizing transformation.

Models such as COCOMO and ISBSG were developed using common logarithm transformation for normalization. Prykhodko and Pukhalevich (2012) showed that common logarithm transformation does not enable normalization of some sets of empirical data [8]. Therefore, it is required to use other normalizing transformations.

Kendal and Stuart (1963) suggested that Johnson transformation can be used as a normalizing transformation [*2*]. Prykhodko and Pukhalevich (2012) built a regression model using Johnson transformation for entire ISBSG dataset. Furthermore, Prykhodko and Pukhalevich (2014) showed that the non-linear regression model of software project duration based on Johnson transformation has better characteristics than the models based on a common logarithm transformation [*6*]. The regression model of software project duration based on Johnson transformation for PC development platform that was developed in [*6*] allows to estimate duration more accurately since this model was built for a specific platform. However, the specified regression model for PC development platform does not provide equations to calculate the confidence interval of estimates of the software project duration. **Therefore, the aim of the research is to derive equations of the confidence interval of the non-linear regression model of software project duration based on Johnson transformation for PC development platform.**

1. Analysis of the Data sample

Among the 789 projects of the ISBSG repository, projects showing the following characteristics were selected (as was suggested by Oligny et al. (1997) in [*3*]) to build the non-linear regression model of software project duration for PC development platform based on Johnson transformation in [*6*]:

– no reasonable doubt as to data validity; that is the ISBSG has not flagged this project as having uncertain data and has retained it for its own analyses;

– known effort;

– known duration;

– software was developed for PC platform.

52 projects satisfied all of these criteria.

Let's assume that D is empirical values of software project duration; E is empirical values of software project effort. Basic descriptive statistics of D and E are shown in table 1. A scatter plot of the empirical data is shown in fig. 1.

Table 1

Descriptive statistics of sample

	Duration (D)	Effort (E)
Units	calendar months	man-hours
Number of observations	52	52
Minimum value	2	170
Maximum value	30	14520
Mean value	9,25	2348,77
Standard deviation	5,58	3232,77
Skewness	1,41	2,47
Kurtosis	2,68	6,44
χ^2	28,08	156,66
χ^2_{cr} ($\nu = 4;\ \alpha = 0{,}05$)	9,49	9,49

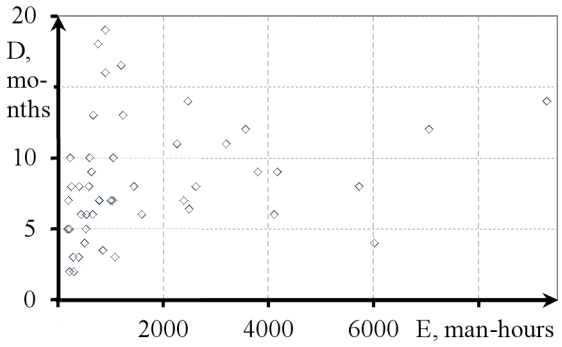

Fig. 1. Scatter plot of project effort vs. duration

Fig 1 indicates that the regression function of Duration vs. Effort is non-linear. Distributions of effort and duration data differ from normal distribution. Values of skewness and kurtosis differ from corresponding values for normal distribution. Pearson's chi-squared test rejects the hypothesis that variables D and E are normally distributed: χ^2 value for both variables is greater than critical value from the chi-squared distribution.

2. Deriving a confidence interval

The family of Johnson transformation for the software project duration and the effort dataset normalization were determined with the Johnson diagram [*7*] using values of skewness and kurtosis.

For the software project duration and effort, S_B family of Johnson transformation was chosen:

$$z = \gamma + \eta \ln\left(\frac{x-\varphi}{\lambda+\varphi-x}\right), \qquad (1)$$

x – non-gaussian random variable; z – normalized (gaussian) random variable; $\varphi < y < \varphi + \lambda$; $\eta > 0$; $\lambda > 0$; $-\infty < \gamma < \infty$; $-\infty < \varphi < \infty$.

To determine values of the Johnson transformation coefficients ($\{\gamma_D, \eta_D, \varphi_D, \lambda_D\}$ and $\{\gamma_E, \eta_E, \varphi_E, \lambda_E\}$) following expression was used [*5*]:

$$\hat{\theta} = \arg\min_{\theta}\left\{\bar{z}^2 + (s_z^2 - 1)^2 + b_{1z}^2 + b_{2z}{}^2\right\}, \qquad (2)$$

$\hat{\theta}$ – estimate of vector of unknown coefficients; θ – vector of unknown coefficients, $\theta = \{\gamma, \eta, \varphi, \lambda\}$; $\bar{z}$ – mean of normalized random variable z, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$; s_z^2 - unbiased variance of z, $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$; b_1 - unbiased skewness of z; b_2 - unbiased excess kurtosis of normalized random variable z.

Estimated transformation coefficient values for duration normalization were $\gamma_D = 2{,}390$; $\eta_D = 1{,}303$; $\varphi_D = 0{,}433$; $\lambda_D = 54{,}747$. Estimated transformation coefficient values for effort normalization are $\gamma_E = 1{,}612$; $\eta_E = 0{,}539$; $\varphi_E = 158{,}853$; $\lambda_E = 18248$.

After determining the Johnson transformation coefficients ($\{\gamma_D, \eta_D, \varphi_D, \lambda_D\}$ and $\{\gamma_E, \eta_E, \varphi_E, \lambda_E\}$), variables D and E were normalized by (1). Following two normalized variables were given: z_D – normalized values of the software project duration; z_E – normalized values of the effort. Both variables are normally distributed. Values of skewness and kurtosis meet with corresponding values for normal distribution. Pearson's chi-squared test accepts the hypothesis that the variables z_D and z_E are normally distributed: χ^2 value for both variables is smaller than the critical value

from the chi-squared distribution. Basic descriptive statistics of z_D and z_E are shown in table 2.

Table 2

Statistics of normalized sample

	z_D	z_E
Number of observations	52	52
Minimum value	-2,20	-2,38
Maximum value	2,60	2,32
Mean value	0	0
Standard deviation	1,01	1,01
Skewness	0	0
Kurtosis	0,12	0,12
χ^2	7,85	2,09
χ^2_{cr} $(\nu = 4,\ \alpha = 0{,}05)$	9,49	9,49

After Johnson transformation, linear regression has been performed on the normalized variables:

$$z_D = b_0 + b_1 z_E . \qquad (3)$$

3 projects were removed from the sample because they showed a high leverage or large studentized residuals (outliers) on the regression results. These projects were not included in the final regression analysis. For the linear regression, constant $b_0 = 0{,}492$, coefficient $b_1 = 0{,}235$; $R^2 = 14{,}450$; F=-0,051.

The $(1-\alpha)100\%$ of confidence interval on linear regression (3) is described by following equation [*9*]:

$$\left[\hat{z}_D(z_E)\right] = \hat{z}_D(z_E) \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2_{z_D}} \sqrt{\frac{1}{n} + \frac{(z_E - \overline{z}_E)^2}{S_{z_E}}} \qquad (4)$$

$$\hat{z}_D(z_E) = b_0 + b_1 z_E ;\ s^2_{z_D} = \frac{1}{n-2}\sum_{i=1}^{n}\left(z_{D_i} - \hat{z}_D(z_{E_i})\right)^2 ;$$

$$S_{z_E} = \sum_{i=1}^{n}(z_{E_i} - \overline{z}_E)^2 .$$

The linear regression model and lines of 95% confidence interval are shown in fig. 2.

The empirical model linking project effort and duration can be characterized by the following equation:

$$\hat{D}(E) = \frac{\varphi_D + (\lambda_D + \varphi_D)e^k}{1 + e^k}, \qquad (5)$$

$$k = \left(\hat{z}_D(z_E) - \gamma_D\right)/\eta_D ; \qquad \hat{z}_D(z_E) = b_0 + b_1 z_E ;$$

$$z_E = \gamma_E + \eta_E \ln\left(\frac{e - \varphi_E}{\lambda_E + \varphi_E - E}\right);$$ b_0 and b_1 –linear regression coefficients; $\{\gamma_D, \eta_D, \varphi_D, \lambda_D\}$ and $\{\gamma_E, \eta_E, \varphi_E, \lambda_E\}$ – Johnson transformation coefficients.

The equation of the confidence interval $[\hat{D}(E)]$ of the non-linear regression model is the same as (5), but $k = \left(\left[\hat{z}_D(z_E)\right] - \gamma_D\right)/\eta_D$.

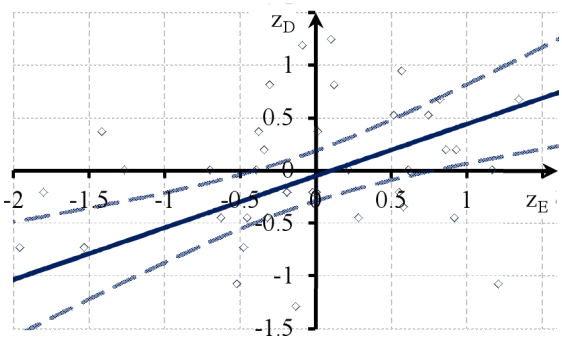

Fig. 2. Linear regression model and confidence interval lines for normalized values

The non-linear regression model built using (5) and lines of 95% confidence interval are shown in fig. 3.

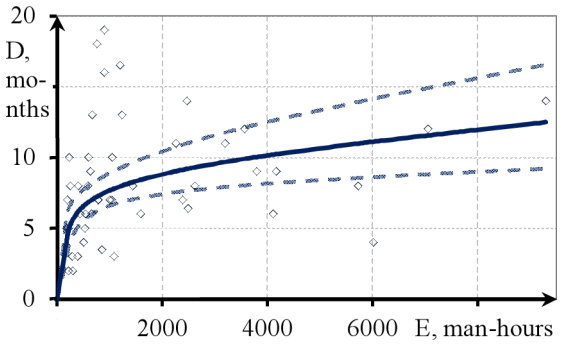

Fig. 3. Non-linear regression model and confidence interval lines for empirical values

Conclusions

In this research, equations of a confidence interval of a non-linear regression model of software project duration based on Johnson transformation from S_B family were derived for the PC development platform. In following research, the confidence interval of regression models of software project duration should also be derived for MR and MF platforms.

References

1. Boehm, B. W. Software engineering economics [Text] / B. W. Boehm. – Englewood Cliffs, NJ : Prentice Hall, 1981. – 768 p. – ISBN 0-13-822122-7.

2. Kendal, M. G. The Advanced Theory of Statistics Vol. 1 [Text] / M. G. Kendal, A. Stuart. – 2nd Edition. – London : Charles Griffin, 1963. – 433 p.

3. Oligny, S. An empirical assessment of project duration models in software engineering [Electronic resource] / S. Oligny, P. Bourque, A. Abran // In Proc. 8th European Software Control and Metrics Conference ESCOM. – Berlin, 1997. – Access mode: http://s3.amazonaws.com/publicationslist.org/data/p.bourque/ref-731/201.pdf. – 2.06.2014.

4. Exploring the relation between effort and

duration in software engineering projects [Text] / S. Oligny, P. Bourque, A. Abran, B. Fournier // In proc. of the World Computer Congress, Aug. 2000. – P. 175-178.

5. Prykhodko, S. B. Interval estimation of statistical moments non-gaussian random variables based on normalizing transformations [Text] / S. B. Prykhodko // Mathematical modeling. – Dniprodzerzhynsk, 2011. – № 1 (24). – P. 9-13 (in Ukrainian).

6. Prykhodko, S. B. Developing PC Software Project Duration Model based on Johnson transformation [Text] / S. B. Prykhodko, A. V. Pukhalevich // Proceedings of the 12th International Conference TCSET'2014, Lviv-Slavske, Ukraine. – Lviv : Polytechnic National University, 2014. – P. 114-116.

7. Prykhodko, S. B. Analytical dependence for selecting Johnson distribution of SL family [Text] / S. B. Prykhodko, L .N. Makarova // Vystnyk KhNTU. – Kherson : KhNTU, 2012. – № 2 (45). – P. 101–104 (in Ukrainian).

8. Prykhodko, S. B. Development of Non-linear Regression Model of Software Project Duration on the basis of Johnson Normalizing Transformation [Text] / S. B. Prykhodko, A. V. Pukhalevich // Radioelectronic and computer systems. – Kharkiv, 2012. – № 4 (56). – P. 90-93 (in Ukrainian).

9. Yan, X. Linear regression analysis : theory and computing [Text] / X. Yan, X. G. Su. – Singapore : World Scientific Publishing Co. Pte. Ltd., 2009. – 349 p. – ISBN 978-981-283-410-2.

Надійшла до редакції 02.04.2014, розглянута на редколегії 19.05.2014

Рецензент: д-р техн. наук, професор, Заслужений діяч науки і техніки України, директор інституту комп'ютерних і інженерно - технологічних наук, зав. каф. інформаційних управляючих систем та технологій К. В. Кошкін, Національний університет кораблебудування ім. адмірала Макарова.

ОЦІНЮВАННЯ ДОВІРЧИХ ІНТЕРВАЛІВ РЕГРЕСІЙНОЇ МОДЕЛІ ТРИВАЛОСТІ ПРОГРАМНИХ ПРОЕКТІВ ДЛЯ ПК НА ОСНОВІ ПЕРЕТВОРЕННЯ ДЖОНСОНА

С. Б. Приходько, А. В. Пухалевич

Розглянуто регресивні моделі тривалості програмних проектів залежно від зусиль на їх виконання. Мета дослідження – отримати рівняння довірчого інтервалу нелінійної регресійної моделі тривалості програмних проектів на основі перетворення Джонсона для платформи ПК. У даний час широко використовуються нелінійні регресійні моделі тривалості програмних проектів. Але існуючі моделі не дозволяють оцінювати довірчі інтервали регресії. Тому необхідно побудувати рівняння довірчого інтервалу нелінійної регресії. В даній роботі були отримані рівняння довірчого інтервалу нелінійної регресійної моделі тривалості програмних проектів на основі перетворення Джонсона сім'ї S_B для платформи ПК. Використовувалися методи математичної статистики, регресійного та інтервального аналізу.

Ключові слова: нелінійна регресійна модель, тривалість програмних проектів, перетворення Джонсона, управління часом, довірчий інтервал.

ОЦЕНИВАНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ РЕГРЕССИОНОЙ МОДЕЛИ ДЛИТЕЛЬНОСТИ ПРОГРАМНЫХ ПРОЕКТОВ ДЛЯ ПК НА ОСНОВЕ ПРЕОБРАЗОВАНИЯ ДЖОНСОНА

С. Б. Приходько, А. В. Пухалевич

Рассмотрены регрессионные модели длительности программных проектов в зависимости от усилий на их выполнение. Цель исследования – получить уравнения доверительного интервала нелинейной регрессионной модели длительности программных проектов на основе преобразования Джонсона для платформы ПК. В данное время широко используются нелинейные регрессионные модели длительности программных проектов. Но существующие модели не позволяют оценивать доверительные интервалы регрессии. Поэтому необходимо построить уравнения доверительного интервала нелинейной регрессии. В этой работе были получены уравнения доверительного интервала нелинейной регрессионной модели длительности программных проектов на основе преобразования Джонсона семьи S_B для платформы ПК. Использовались методы математической статистики, регрессионного и интервального анализа.

Ключевые слова: нелинейная регрессионная модель, длительность программных проектов, преобразование Джонсона, управление временем, доверительный интервал.

Приходько Сергій Борисович – д-р техн. наук, доц., зав. каф. програмного забезпечення автоматизованих систем, Національний університет кораблебудування ім. адмірала Макарова, Миколаїв, Україна, e-mail: sergiy.prykhodko@nuos.edu.ua.

Пухалевич Андрій Володимирович – ст. лаб. каф. програмного забезпечення автоматизованих систем, Національний університет кораблебудування ім. адмірала Макарова, Миколаїв, Україна, e-mail: a.puhalevich@gmail.com.