DOI 10.31891/2307-5732-2024-341-5-14
УДК 004.412:519.237.5

FARIONOVA TETYANA
Admiral Makarov National University of Shipbuilding
https://orcid.org/0000-0003-3384-4712
e-mail: tetyana.farionova@nuos.edu.ua
PUKHALEVYCH ANDRII
Admiral Makarov National University of Shipbuilding
https://orcid.org/0000-0002-8827-3251
e-mail: andrii.pukhalevych@nuos.edu.ua
VORONA MYKHAILO
Admiral Makarov National University of Shipbuilding
https://orcid.org/0000-0003-4288-0096
e-mail: mykhailo.vorona@nuos.edu.ua

THE NON-LINEAR REGRESSION MODEL TO ESTIMATE THE DEVELOPMENT DURATION OF JAVA APPLICATIONS FOR THE MIDRANGE PLATFORM

Data from the ISBSG company shows that the midrange computer platform occupies a software share of approximately 25%. At the same time, the Java language is mainly used for development on this platform. Java applications for the midrange platform have such characteristics as a large size, a significant set of used components, and special requirements for stability. For this reason, performing a reliable assessment of the duration of the development of such applications is an important task, the solution of which has scientific and practical interest.

Analysis of modern models for the duration estimation of development of software applications was done. The most used models for estimating the duration of software development are COCOMO and ISBSG: nonlinear regression equations for estimating the duration of software development depending on development effort. Only the ISBSG model takes into account the features of the platform for which the software is being created. However, this model does not take into account the programming language used to develop the application.

A non-linear regression model was built for duration estimation of applications development written in Java for the midrange platform depending on effort, by creating non-linear regression equation, bounds of the confidence interval and of the prediction interval. The specified model was built using statistical data of the ISBSG repository from 129 projects and the appropriate method based on normalizing transformations. This method was used because empirical data on the duration and effort of software development have a distribution law that differs from the Normal distribution. The data normalization required by this method is done using a decimal logarithm. In the process of examining the data for outliers, 29 of the 129 applications were removed. Better values for the characteristics of the coefficient of determination, MMRE and the percentage of prediction PRED(0.25) were obtained for created model compared to the ISBSG model. The construction of the specified model made it possible to improve the reliability of the obtained duration estimates of the applications development written in Java for the midrange platform.

Keywords: duration of Java application development, nonlinear regression, midrange platform.

ФАРІОНОВА ТЕТЯНА
Національний університет кораблебудування імені адмірала Макарова
ПУХАЛЕВИЧ АНДРІЙ
Національний університет кораблебудування імені адмірала Макарова
ВОРОНА МИХАЙЛО
Національний університет кораблебудування імені адмірала Макарова

НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ ТРИВАЛОСТІ РОЗРОБКИ ЗАСТОСУНКІВ НА МОВІ JAVA ДЛЯ ПЛАТФОРМИ MIDRANGE

Статистичні дані компанії ISBSG показують, що комп'ютерна платформа midrange займає частку ПЗ приблизно 25%. При цьому, в основному, для розробки для цієї платформи використовується мова Java. Застосунки на мові Java для платформи midrange мають такі характеристики як великий розмір, значна кількість компонентів та відзначаються спеціальними вимогами щодо надійності. Тому виконання достовірного оцінювання тривалості розробки таких застосунків – це актуальна задача, вирішення якої матиме науковий та практичний інтерес.

В статті проведено аналіз сучасних моделей, які дозволяють оцінювати тривалість розробки програмних застосунків. Була побудована нелінійна регресійна модель для оцінювання тривалості розробки застосунків на мові Java для платформи midrange в залежності від трудомісткості, шляхом побудови рівнянь нелінійної регресії, границі довірчого інтервалу та границі інтервалу прогнозування. Побудова вказаної моделі дозволила підвищити достовірність отриманих оцінок тривалості розробки застосунків на мові Java для платформи midrange.

Ключові слова: тривалість розробки Java-застосунків, нелінійна регресія, платформа midrange.

Introduction

Midrange is a class of computer systems between mainframes and microcomputers/personal computers [1]. These are mid-level systems, primarily high-performance network servers and other types of servers that perform large-scale data handling for different business purposes. Midrange platform software is characterized by its large size, significant number of components, and extended reliability requirements. Specified reasons lead to the problem of reliable duration estimation of software applications development for the midrange platform.

According to PMBOK and the ISO/IEC 12207 Software Life Cycle Process Standard, the process of estimating the software development duration is an integral part of software measurement engineering [2]. According to statistics, about 30% of development projects have complications with completion (they are not fit into initial cost, they violates of the duration limits, or have inadequate quality). At the same time, regarding to the statistical data of the ISBSG (International Software Benchmarking Standards Group), midrange computer platform occupies a software share of approximately. At the same time, the Java language is mainly used for development on this platform. For this reason, performing a reliable assessment of the duration of the development of such applications is an important task, the solution of which has scientific and practical interest.

Analysis of recent sources

The most used models for estimating the duration of software projects are COCOMO [3] and ISBSG [4]: nonlinear regression equations for estimating the duration of software projects depending on development effort. The empirical data on the duration and effort of software development have a distribution law that differs from the Normal distribution [3–5]. Therefore, the normalization of the empirical data on the duration and effort of the development of software projects was carried out using a decimal logarithm for the construction of such models. At the same time, only the ISBSG model takes into account the features of the platform for which the software is being created. For the midrange platform, this model is set as $D = 0.548\ E^{0.360}$, where D is the duration of software application development (months), E is the effort of software application development for the midrange platform (man-hours) [4]. However, this model does not take into account the programming language used to develop the application.

The aim of the research is to improve the reliability of estimating the duration of applications development written in Java for the midrange platform due to the construction of a non-linear regression model.

Main material

It is possible to build a non-linear regression model of the duration of development of Java applications for the midrange platform using the appropriate development method based on normalizing transformations [6]. Following steps should be performed to get the model:

- normalize empirical data using a normalizing transformation (decimal logarithm);
- build a linear regression based on the normalized data;
- build a nonlinear regression model based on the linear regression equation, using the inverse normalizing transformation; build the equation of the lower and upper bounds of the confidence interval of the nonlinear regression, the equation of the lower and upper bounds of the prediction interval.

The process of building a nonlinear regression model includes removing outliers from empirical data. Before constructing a linear regression, the squared Mahalanobis distance is calculated for each pair of values. If the squared Mahalanobis distance exceeds a critical value for any pair, then the data row with the largest Mahalanobis distance is removed. After constructing a linear regression model, the law of the distribution of residuals (random error) is checked. If the distribution is not Normal, then the data row with the largest absolute residual is removed. If values fall outside the linear regression prediction interval, they are also removed. The whole process is repeated until there are no outliers left.

A linear regression model in general form can be represented by the equation

$$y = \hat{y} + \varepsilon,$$

where y is the dependent random variable; $\hat{y}$ is a function that determines the type of regression model; x is an independent random variable; and ε is a random error.

Most often, the random variables included in the regression model have no Gaussian distribution law. It leads to the construction of nonlinear regression. In order to avoid a brute-force search method when constructing nonlinear regression equations, methods based on normalizing transformations are used [6]. The method of normalizing transformations makes it possible to switch from the original non-Gaussian random variables to Gaussian-distributed (normalized) variables.

For the normalized random variables, a linear regression equation is constructed, which is then transformed into a nonlinear one using inverse transformations:

$$z_{\mathrm{y}} = b_1 z_x + b_0 + \varepsilon, \tag{1}$$

where z_x, z_{y} are normalized random variables x, y; ε is a random error distributed according to the normal law $\varepsilon \sim N(0,1)$; b_0, b_1 are linear regression coefficients calculated by the method of least squares.

In the research, data normalization is done using a decimal logarithm, since this is the transformation was used in the development of ISBSG models:

$$z_x = \lg x,\ z_y = \lg y. \tag{2}$$

The next step in constructing the regression equation is to get the confidence interval and the prediction interval. When the law of the distribution of random variables is Normal, these intervals for the linear regression equation can be constructed using the Student's t- distribution with a significance level α/2 and the number of degrees of freedom n-2:

$$\hat{y} \pm t_{\left(\frac{\alpha}{2},n-2\right)} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}; \quad \hat{y} \pm t_{\left(\frac{\alpha}{2},n-2\right)} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \tag{3}$$

where $\hat{y}$ is the value of y estimated by the regression equation; $t_{(\alpha/2,n\text{-}2)}$ – is the quantile of Student's t-distribution; α – level of significance; n – is the number of values in the dataset; $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

This approach was used to build a model for duration estimation of development applications written in Java for the midrange platform. Let D be a random variable, empirical values of the duration of software projects (in months), which depends on a random variable E – empirical values of development effort (in man-hours).

Using the normalizing transformation (2), we obtain the normalized random variables $z_{\mathrm{E}} = \lg\mathrm{E}, z_D = \lg D$. The linear regression equation for normalized values has the form $\hat{z}_D(z_E) = b_1 z_E + b_0 + \varepsilon$.

To assess the reliability of the estimation using the regression model, can be used the coefficient of determination R^2, or the mean magnitude of relative error (MMRE) and $\mathrm{PRED}_{0.25}$ – the percentage of prediction with the magnitude of relative error less than 0.25.

The coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{SS_E}{SS_T},$$

where SSE is the sum of squares of the residuals, $SS_E = \sum_{i=1}^{n}\left(D_i - \widehat{D_\iota}\right)^2$; SST is the total sum of squares, $SS_T = \sum_{i=1}^{n}(D_i - \overline{D})^2$; $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$.

If the R^2 value is $R^2 \geq 0.5$ it is considered that this model is acceptable. A model with a determination index $R^2 \geq 0.8$ is considered sufficiently effective and efficient.

MMRE is calculated by the formula

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i,$$

where $MRE_i = \left|\frac{D_i - \widehat{D_\iota}}{D_i}\right|$.

The prediction level $\mathrm{PRED}_{0.25}$ is calculated by the formula

$$\mathrm{PRED}_{0.25} = \frac{1}{n}\sum_{i=1}^{n}\begin{Bmatrix} 1, if\ MRE_i \leq 0.25 \\ 0 \qquad\qquad else \end{Bmatrix},$$

where $MRE_i = \left|\frac{D_i - \widehat{D_\iota}}{D_i}\right|$.

It is usually considered acceptable estimation accuracy if $\mathrm{PRED}_{0,25} \geq 0.75$.

To create a non-linear regression model for estimating the development duration of applications written in Java for the midrange platform, we will use the statistical data of the ISBSG repository from 129 projects. Fig. 1 shows ISBSG data: development duration of applications written in Java for the midrange platform (D, months) and development effort (E) of these projects according to the ISBSG repository.

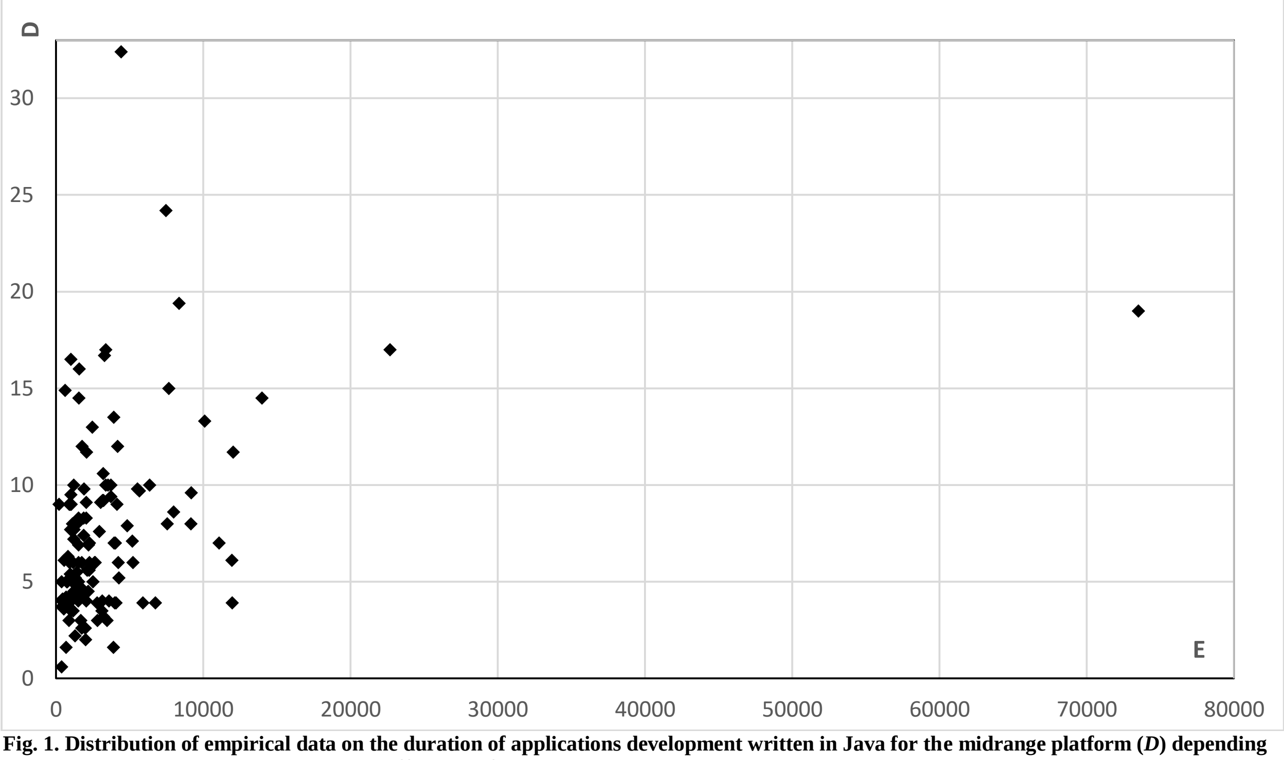

Fig. 1. Distribution of empirical data on the duration of applications development written in Java for the midrange platform (*D*) depending on the effort (*E*) of these projects according to the ISBSG repository

In the process of examining the data for outliers, 29 of the 129 applications were removed.

The final set contains 100 data pairs of the duration and effort of developing Java applications for the midrange platform, for which the linear regression model has the form

$$z_D = 0.270 z_E - 0.070 + \varepsilon.$$

To switch from a linear regression model to a non-linear one, the inverse normalization transformation to (2) was applied: $\mathrm{E} = 10^{z_{\mathrm{E}}}, \mathrm{D} = 10^{z_{\mathrm{D}}}$. The constructed non-linear regression model for estimating the duration of applications development written in Java for the midrange platform is the following:

$$D = 0.85E^{0.27+\varepsilon}, \qquad (4)$$

where D is the duration of application development written in Java for the midrange platform (in months); E is the effort of application development written in Java for the midrange platform (in man-hours).

Fig. 2 presents the nonlinear regression equation with the corresponding intervals (confidence and prediction interval) of the regression of the development time of Java applications of the midrange platform.

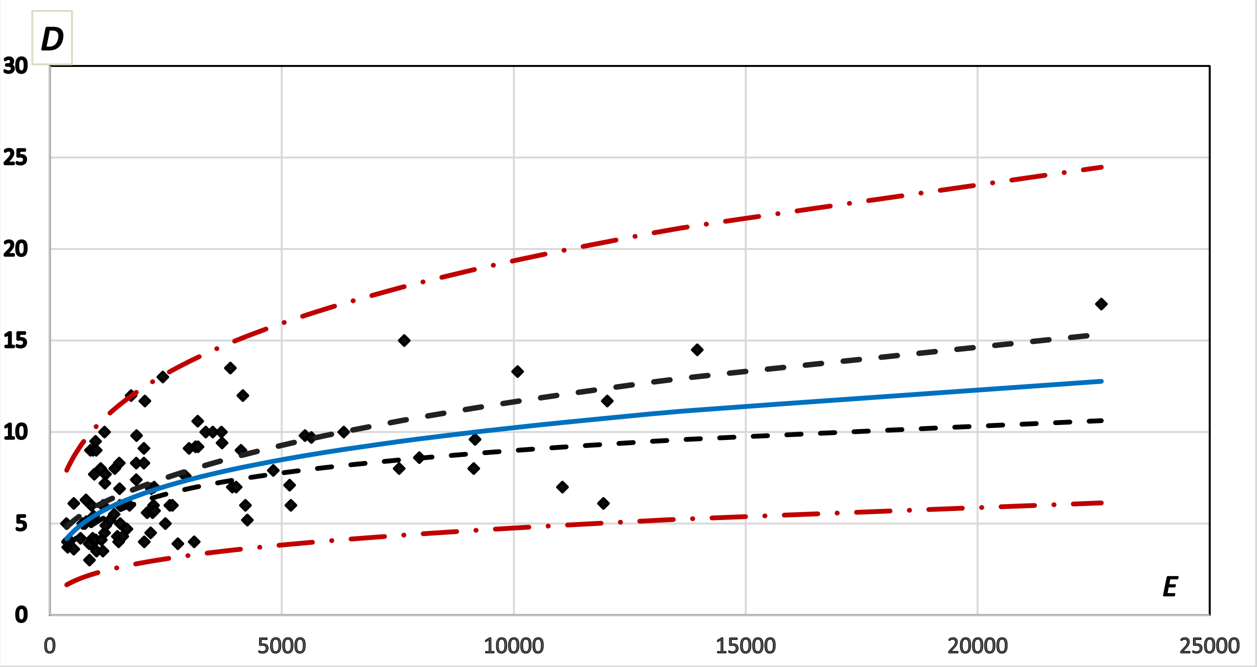

Fig. 2. Initial empirical data and constructed non-linear regression equation, confidence interval and prediction interval of the development duration of midrange Java applications (n=100)

The constructed non-linear regression model for estimation the development duration of applications written in Java for midrange has the following characteristics: $\text{PRED}_{0.25} = 0.57$; $R^2 = 0.365$; MMRE=0.268. Thus, better values for the characteristics of R^2, MMRE and $\text{PRED}_{0.25}$, were obtained comparing to the ISBSG model, for which R^2 =0.253; MMRE=0.708; $\text{PRED}_{0.25} = 0.301$. It should be noted that the obtained model has $0.25 < \text{PRED}_{0.25} < 0.75$ and additional research is required to improve the constructed model for estimating the of Java application development for the midrange platform by applying other normalizing transformation

Conclusions

A non-linear regression model was built to estimate the duration of Java application development for the midrange platform depending on effort, by constructing non-linear regression equations, confidence interval bounds and prediction interval bounds. The construction of the specified model made it possible to improve the reliability of the obtained estimates of the application development duration written in Java for the midrange platform.

References

1. Midrange. URL: https://www.devx.com/terms/midrange/.
2. ISO/IEC/IEEE International Standard - Systems and software engineering - Software life cycle processes/. URL: https://standards.ieee.org/ieee/12207/5672/
3. Boehm B.W. Software engineering economics. Englewood Cliffs, NJ: Prentice Hall, 1981.
4. Oligny S., Bourque P., Abran A, Fournie B. Exploring the relation between effort and duration in software engineering projects. In proc. of the World Computer Congress. 2000. P. 175-178.
5. Prykhodko S.B.,Prykhodko K.S., Makarova L.M., Pukhalevych A.V. Nonlinear regression models for estimating the duration of software development in Java for PC based on the 2021 ISBSG data. Radio Electronics, Computer Science, Control. No. 3 (62). 2022. DOI: https://doi.org/10.15588/1607-3274-2022-3-14
6. Prykhodko S.B. The method of constructing nonlinear regression equations based on normalizing transformations. Abstracts of interstate reports. Science and Methodology conf. "Problems of mathematical modeling". 2012. 31-33.