УДК 004.412:519.25 DOI https://doi.org/10.35546/kntu2078-4481.2024.2.28

О. С. ОРЄХОВ
аспірант кафедри програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
ORCID: 0000-0002-0001-0140

Т. А. ФАРІОНОВА
кандидат технічних наук,
доцент кафедри програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
ORCID: 0000-0003-3384-4712

МАТЕМАТИЧНІ МОДЕЛІ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ JAVA-ЗАСТОСУНКІВ

У статті розглядається застосування математичних моделей для оцінювання розміру Java-застосунків. Мова програмування Java є однією з найбільш поширених у світі та широко використовується в розробці різноманітних програмних проєктів. Оцінювання розміру Java-застосунку є актуальною задачею, яка невід'ємно пов'язана з життєвим циклом розробки програмного забезпечення на ранніх стадіях проєктування. Метою роботи є підвищення достовірності оцінювання кількості рядків коду Java-застосунків на ранніх стадіях розробки програмних проєктів за метриками діаграми класів шляхом побудови нелінійних регресійних моделей. Об'єктом дослідження є процес оцінювання розміру Java-застосунків з відкритим кодом. Предметом дослідження є математичні моделі для оцінювання розміру Java-застосунків. Для досягнення поставленої мети було зібрано 2 вибірки метрик Java-застосунків із відкритим програмних кодом – нвчальна, розміром 286, та тестова, розміром 285 точок даних, проведено аналіз та порівняння існуючих математичних моделей і рівнянь для оцінювання розміру Java-застосунків на тестовій вибірці. Доведено, що існуючі регресійні рівняння та моделі мають незадовільний рівень якості прогнозування розміру Java-застосунків або не можуть бути застосовані для наведеного набору даних через обмеження регресійних моделей. Із використанням навчальної вибірки, побудовано однофакторні нелінійні регресійні моделі для оцінювання розміру Java-застосунків на основі нормалізуючих перетворення десяткового логарифму, Бокса-Кокса та Джонсона сімейства SB за метрикою кількості класів (CLASS) та двофакторна нелінійна регресійна модель на основі нормалізуючого перетворення десяткового логарифму за метриками кількості класів (CLASS) та загальна кількість видимих методів (VMQ). Отримана двофакторна нелінійна регресійна модель на основі перетворення у вигляді десяткового логарифму має меншу середню величину відносної похибки, вище значення відсотка передбачення для рівня відносної похибки та вище значення коефіцієнту детермінації, що у порівнянні з існуючими моделями дозволяє підвищити достовірність оцінювання кількості рядків коду Java-застосунків.

Ключові слова: *розмір програмного забезпечення, кількість рядків коду, Java-застосунок, нелінійна регресійна модель, нормалізуюче перетворення, негаусівські дані.*

O. S. ORIEKHOV
Postgraduate Student at the Department of Automated Systems Software
Admiral Makarov National University of Shipbuilding
ORCID: 0000-0002-0001-0140

T. A. FARIONOVA
Candidate of Technical Sciences,
Associate Professor at the Automated Systems Software Department
Admiral Makarov National University of Shipbuilding
ORCID: 0000-0003-3384-4712

MATHEMATICAL MODELS FOR THE SIZE ESTIMATING OF JAVA APPLICATIONS

This paper introduces the usage of mathematical models for Java applications size estimation. The Java programming language is one of the most widely used in the world and is used in the development of various software projects. Size estimation of Java-applications is one of the key planning tasks at the early stages of software project planning. The aim of the study is to increase the accuracy of Java application code lines estimation at the early stages of software project development using class diagram metrics by building nonlinear regression models. The object of study is the Java applications size estimation process. The subject of the study is mathematical models for Java applications size estimation. To achieve this goal, we collected 2 samples of code metrics from open source Java applications – a training sample of 286 data points and a test sample of 285 data points. We analyzed and compared existing mathematical models

and equations of Java application size estimation using the test sample. Proven that the existing regression equations and models have an unsatisfactory level of accuracy for Java applications size estimation or cannot be applied to the given data set due to the limitations of regression models. For Java applications size estimation, using training sample we built one-factor nonlinear regression models based on the normalizing transformations of the decimal logarithm, Box-Cox and Johnson of the SB family by the number of classes (CLASS) metric and a two-factor nonlinear regression model based on the normalizing transformation of the decimal logarithm by the number of classes (CLASS) and the visible methods quantity (VMQ) metrics. The obtained two-factor nonlinear regression model based on the decimal logarithm normalizing transformation has a smaller mean magnitude of relative error, a higher value of the percentage of prediction of the relative error level and a higher value of the determination coefficient, which, in comparison with existing models, allows to increase the reliability and accuracy of source lines of code estimation of Java applications.

Key words: *software size, lines of code, Java-application, nonlinear regression model, normalizing transformation, non-Gaussian data.*

Постановка проблеми

Мова програмування Java є однією з найбільш поширених у світі [1] та широко застосовується в розробці програмних проєктів за різними напрямками, починаючи від веб-додатків та прикладного програмного забезпечення (ПЗ), до автомобільних та інформаційних систем. Оцінювання розміру програмного забезпечення тісно пов'язане з моделями оцінювання вартості програмного продукту та трудовитрат на його розробку. Інформація про розмір, а саме кількість рядків коду Java-застосунків, на ранніх етапах розробки дає змогу отримати прогноз трудомісткості розробки ПЗ за допомогою параметричних моделей COCOMO, COCOMO II, SLIM, SEER-SEM, тощо, які використовують параметр розміру у вигляді рядків коду для оцінки трудомісткості розробки програмних проєктів [2]. Тому саме оцінювання розміру Java-застосунків є актуальною задачею, яка невід'ємно пов'язан з життєвим циклом програмного проєкту. В якості міри розміру програмних застосунків в дослідження використовується метрика кількості рядків коду, оскільки вона дозволяє врахувати особливості мови програмування.

Аналіз останніх досліджень і публікацій

Для оцінювання розміру Java-застосунків з відкритим кодом побудовано як лінійні [3;4] так і нелінійні [5;6;7] регресійні рівняння та моделі в залежності від різної кількості метрик зібраних з концептуальної моделі даних у вигляді UML діаграм, зокрема діаграм класів. Крім того, розмір навчальних вибірок напряму впливає на достовірність прогнозування, оскільки вибірки можуть мати недостатній рівень репрезентативності генеральної сукупності [8; 9].

Так, в роботі [3] запропоновано трьохфакторне лінійне регресійне рівняння оцінювання кількості рядків коду Java-застосунків на основі методів множинного лінійного регресійного аналізу із використанням метрик загальної кількості класів (CLASS), загальної кількості зв'язків між класами (CBO) та загальної кількості атрибутів класів (TFQ) в вихідному коді. У роботі [4] удосконалені трьохфакторні лінійні регресійні рівняння оцінювання кількості рядків коду для великих промислових інформаційних Java-систем на основі 16-ти точок даних, Java-застосунків з відкритим кодом на основі 30 точок даних та узагальнене лінійне регресійне рівнянні на основі сукупного набору з 46 точок даних. В якості метрик використано значення загальної кількості класів (CLASS), загальної кількості зв'язків між класами (CBO) та середнє значення кількості атрибутів на клас (aTFQ).

Робота [5] присвячена удосконаленню оцінювання кількості рядків коду для промислових інформаційних Java-систем та запропонована трьохфакторна нелінійна регресійна модель із використанням багатовимірного перетворення Джонсона для сімейства S_B, за метриками програмного коду – кількості класів (CLASS) X_1, загальної кількість зв'язків між класами (CBO) X_2 та значення середньої кількості атрибутів на клас (aTFQ) X_3 із концептуальної моделі застосунку. Модель побудована із використанням 32-х точок даних.

В роботі [6] побудовано нелінійну регресійну модель для оцінювання розміру веб-застосунків, які створені мовою програмування Java. Запропонована математична модель побудована із застосуванням нормалізючого перетворення Джонсона сімества S_B і використовує метрику кількості класів (CLASS), як незалежний фактор X. Модель побудована із використанням 30 точок даних.

В роботі [7] авторами розроблена чотирьохфакторна нелінійна регресійна модель для оцінювання розміру Java-застосунків з відкритим кодом. Модель була побудована на основі багатовимірного нормалізуючого перетворення Джонсона сімества S_B. Для оцінювання кількості рядків коду використані метрики кількості класів (CLASS) X_1, кількості статичних методів (SMQ) X_2, значення згуртованості методів класу (LCOM – lack of cohesion of methods) X_3 та кількість унікальних викликів методу в класі (RFC) X_4. Ця модель побудована із використанням 38-ми точок даних.

Формулювання мети дослідження

Метою статті є підвищення достовірності оцінювання кількості рядків коду Java-застосунків на ранніх стадіях розробки програмних проєктів за метриками діаграми класів шляхом побудови нелінійних регресійних моделей. Об'єктом дослідження є процес оцінювання кількості рядків коду Java-застосунків з відкритим кодом. Предметом дослідження є математичні моделі для оцінювання розміру Java-застосунків

Викладення основного матеріалу дослідження

Авторами були зібрані дані за метриками програмного коду 571 Java-застосунків, розташованих на платформі GitHub (https://github.com). За допомогою інструменту CK (https://github.com/mauricioaniche/ck) отримані наступні метрики: кількість рядків коду (KLOC), загальна кількість класів (CLASS), загальна кількість унікальних викликів методу в класі (RFC), загальна кількість значень згуртованості методів класу (LCOM), загальна кількість статичних методів (SMQ), загальна кількість видимих методів (VMQ), загальна кількість атрибутів класів (TFQ), зв'язність між класами (CBO). Ці метрики можна отримати на ранній стадії проєктування з концептуальної моделі застосунку. Отриманий набір даних був розділений випадковим чином на навчальну і тестову вибірки з розмірами в 286 та 285 рядків даних відповідно. Розподіл метрик відносно KLOC наведені на рис. 1.

Для оцінки параметру кількості рядків коду існують різні підходи із застосуванням як лінійних так і нелінійних регресійних моделей. Зазвичай дані за метриками програмного коду не розподілені за нормальним законом, що робить обмежену можливість використання лінійних регерсійних моделей для оцінювання розміру рядків коду. Теоретичною умовою застосування лінійних регресійних моделей є нормальний розподіл залишків регресії ε або нормальний розподіл багатовимірних даних.

Для оцінки якості прогнозування регресійних моделей та рівнянь використовуються такі критерії якості, як коефіцієнт детермінації R^2, значення середньої величини відносної похибки *MMRE* та значення відсотка передбачення для рівня відносної похибки 0,25 – *PRED* (0,25) [8]. Критерій MMRE визначається як

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i, \quad (1)$$

де MRE_i – значення величини відносної похибки для *i*-го рядку даних випадкової величини

$$MRE_i = \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|. \quad (2)$$

Значення відсотка передбачення для рівня відносної похибки 0,25 PRED,

$$PRED(0{,}25) = \frac{1}{N}\sum_{i=1}^{N} \begin{cases} 1 & if\ MRE_i \le 0{,}25 \\ 0 & otherwise \end{cases}. \quad (3)$$

Прийнятними вважаються значення $MMRE \le 0{,}25$ для оцінки точності моделі та значення $PRED\ (0{,}25) \ge 0{,}75$ для прийнятної точності регресійних моделей. Значення R^2 вважається прийнятним, якщо воно більше за 0,75 [8].

Регресійні моделі і рівняння [3; 4; 5; 6; 7] були перевірені за множинним коефіцієнтом детермінації R^2, середньою величиною відносної помилки MMRE (1) і відсотком прогнозованих результатів, для яких величини відносної помилки MRE (2) менші за 0,25, PRED(0,25) (3). Результат перевірки показав їх незадовільну якість як для навчальної так і для тестової вибірок.

Таким чином виникає необхідність вдосконалення існуючих моделей для оцінювання кількості рядків коду (KLOC) Java-застосунків. Авторами за даними начальної вибірки було побудовані однофакторні нелінійні регресійні моделі на основі параметру кількості класів (CLASS) X ($KLOC = f(CLASS)$) із використанням перетворення десяткового логарифму, Бокса-Кокса і Джонсона сімейства SB, та двофакторну нелінійну регресійну модель для оцінювання кількості рядків коду за параметрами кількості класів (CLASS) X_1 та загальної кількості видимих методів класів (VMQ) X_2 ($KLOC = f(CLASS, VMQ)$), з метою виявлення найкращої моделі. Запропоновані варіанти моделей грунтуються на параметрі кількості класів (*CLASS*), оскільки класи є фундаментальними одиницями побудови будь-яких програм із використанням об'єктно-орієнтованих мов програмування.

Перевірка нульової гіпотези про можливість побудови однофакторної $y = f(X) + \varepsilon$ та двофакторної $y = f(X_1, X_2) + \varepsilon$ лінійних регресійних моделей була виконана за допомогою критерія Мардія [10] для рівня значущості $\alpha = 0{,}005$, який заснований на вимірюванні багатовимірного асиметрії β_1 і ексцесу β_2. Відповідно до цього тесту:

– розподіл двовимірних даних X (*CLASS*) та Y (*KLOC*) є негаусівським, оскільки тестова статистика для багатовимірної асиметрії $(N \cdot \beta_1/6 = 1809{,}77)$ цих даних перевищує значення 14,86 квантилю розподілу χ^2, для 4 ступенів свободи, та значення багатовимірного ексцесу β_2=75,09, що є більшим за значення квантиля розподілу Гауса, яке становить 9,22, де $m = 8$ та $\sigma = 0{,}47$;

– розподіл тривимірних даних X_1 *(CLASS),* X_2 *(VMQ) та* Y *(KLOC)* є негаусівським, оскільки тестова статистика для багатовимірної асиметрії $N \cdot \beta_1/6 = 12998{,}09$ цих даних перевищує значення 25,19 квантилю розподілу χ^2, для 10 ступенів свободи, та значення багатовимірного ексцесу $\beta_2 = 314{,}65$, що є більшим за значення квантиля розподілу Гауса, яке становить 16,67, де $m = 15$ та $\sigma = 0{,}65$.

Для побудови нелінійних регресійних моделей було обрано методи, застосовані у роботах [11; 12], які грунтуються на застосуванні взаємозворотних нормалізуючих перетворень з ітеративним видаленням викидів. Застосування нормалізуючих перетворень дозволяє перейти до побудови лінійних регерсійних моделей на основі нормалізованих даних із подальшим їх перетворенням в нелінійні регресійні моделі.

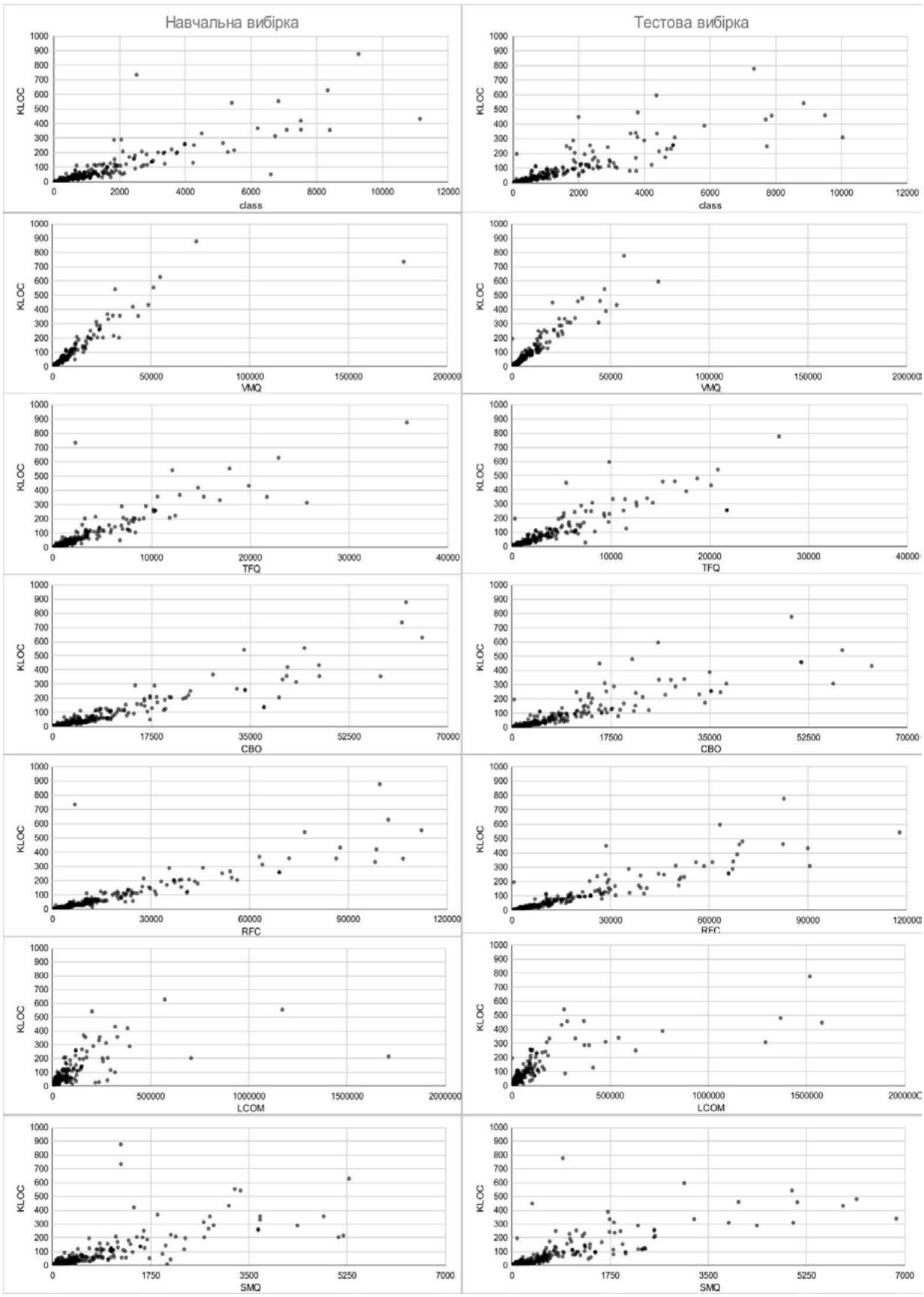

Рис. 1. Розподіл даних навчальної та тестової вибірок

Нормалізуюче перетворення негаусівського випадкового вектору $P = \{Y, X_1, X_2, \ldots, X_k\}^T$ у гаусівський випадковий вектор $T = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$ задається як

$$T = \psi(P), \tag{4}$$

де k – кількість факторів, тоді обернене перетворення до (4) задається як

$$P = \psi^{-1}(T), \tag{5}$$

де ψ – вектор взаємозворотніх функцій нормалізуючого перетворення, $\psi = \{\psi_Y, \psi_1, \psi_2, \ldots, \psi_k\}^T$.

Після нормалізації ненормальних даних вибірки за перетворенням (4), рівняння лінійної регресії буде мати наступний вигляд

$$Z_y = \hat{Z}_y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \ldots + \hat{b}_k Z_k + \varepsilon, \tag{6}$$

де ε – випадкова величина (ВВ) розподілена за нормальним законом, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$; k=1 для однофакторної регресійної моделі та k = 2 для двофакторної регресійної моделі; $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$ – оцінки параметрів лінійного регресійного рівняння (6), які знаходяться за методом найменших квадратів. Після побудови моделі лінійної регресії (6) для нормалізованих даних, застосувавши зворотнє перетворення (5), нелінійна регресійна модель буде мати наступний вигляд

$$Y = \psi_Y^{-1}(\hat{Z}_y + \varepsilon) = \psi_Y^{-1}(\hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \ldots + \hat{b}_k Z_k + \varepsilon) \tag{7}$$

Для нормалізації даних були обрані наступні нормалізуючі перетворення:

– у вигляді десяткового логарифму

$$Z = log_{10}(X), \tag{8}$$

– перетворення Бокса-Кокса

$$Z = \begin{cases} (X^\lambda - 1)/\lambda, & if\ \lambda \neq 0 \\ ln(X), & if\ \lambda = 0 \end{cases} \tag{9}$$

де λ – параметри перетворення Бокса-Кокса.

– перетворення Джонсона сімейства SB

$$Z = \gamma + \eta\ ln\left(\frac{X - \varphi}{\varphi + \lambda - X}\right), \tag{10}.$$

де γ, η, φ та λ – параметри перетворення Джонсона, $\varphi < X < \varphi + \lambda$, $\eta > 0$, $\lambda > 0$.

Для оцінки параметрів перетворень Бокса-Кокса і Джонсона використовувався метод максимальної правдоподібності.

При побудові регресійної моделі важливо щоб дані були однорідними та не містили аномальних значень – викидів. Для визначення викидів застосовані методи, запропоновані у роботі [12]. З навчальної вибірки були вилучені дані, для яких квадрат відстані Махаланобіса перевищував квантиль розподілу $\chi^2_{\text{крит}}$ для рівня значимості α = 0,005.

Також, відповідно до [12], були виключені дані, які знаходяться поза межами інтервалів передбачення. Границі інтервалів передбачення нелінійної регресії визначаються за наступною формулою:

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,v} S_{Z_Y}\left\{1 + \frac{1}{N} + (Z_X^+)^T S_{XX}^{-1}(Z_X^+)\right\}^{1/2}\right), \tag{13}$$

де ψ_Y^{-1} – функція оберненого перетворення оцінки лінійної регресії, $t_{\alpha/2,v}$ – квантиль t-розподілу Стьюдента з $v = N$ – k – 1 ступенями свободи та $\alpha/2$ – рівнем значущості; $S^2_{\hat{Z}_Y} = \frac{1}{N}\sum_{i=1}^{N} \left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$; Z_X^+ – вектор центральних факторів(регресорів), який містить значення $\left\{Z_{1_i} - \underline{Z}_1, Z_{2_i} - \underline{Z}_2, \ldots, Z_{k_i} - \underline{Z}_k\right\}$; S_{XX} k x k матриця

$$S_{XX} = \left[S_{X_q} S_{X_r}\right], \tag{14}$$

де $S_{X_q} S_{X_r} = \sum_{n-1}^{N} \left(X_{q_i} - \underline{X}_q\right)\left(X_{r_i} - \underline{X}_r\right)$, q, r = 1,2, …, k.

Згідно з [11; 12] процес виключення викидів є ітеративним.

Для навчальної вибірки застосовано нормалізуюче перетворення десяткового логарифму (8). Отримані нормалізовані дані було очищено від 12 точок викидів та перевірено на відповідність нормальному закону розподілу багатовимірних даних за критерієм Мардія [10]. Далі, за (6) побудовано лінійну регресійну модель. Оцінки параметрів моделі , які визначені за методом найменших квадратів, мають значення $\hat{b}_0 = -1{,}403401$, $\hat{b}_1 = 1{,}031958$.

Застосувавши обернену трансформацію (7) до перетворення десяткового логарифму (8), однофакторна нелінійна регресійна модель має вигляд

$$\hat{Y} = 10^{\varepsilon - 1{,}403401} X_1^{1{,}031958} \tag{15}$$

Наступним кроком, будуємо однофакторну нелінійну регресійну модель із застосуванням перетворення Бокса-Кокса. За методом максимальної правдоподібності оцінки параметрів перетворення (9) мають значення $\hat{\lambda}_Y = 0{,}002738$, $\hat{\lambda}_X = 0{,}030707$ для останньої ітерації. Нормалізовані дані було очищено від 12 точок викидів та перевірено на відповідність нормальному закону розподілу. Далі, за (6) було побудовано лінійну регресійну модель, яка має такі оцінки параметрів $\hat{b}_0 = -2{,}697047$, $\hat{b}_1 = 0{,}857236$.

Для однофакторної нелінійної регресійної моделі, яка побудована із застосуванням нормалізуючих перетворень Джонсона сімейства SB (10) маємо наступні значення оцінків параметрів перетворення $\hat{\gamma}_Y = 4311{,}992036$, $\hat{\gamma}_X = 3555{,}710081$, $\hat{\eta}_Y = 30479{,}194612$, $\hat{\eta}_X = 29712{,}602140$, $\hat{\varphi}_Y = -0{,}345327$, $\hat{\varphi}_X = -3{,}150988$, $\hat{\lambda}_Y = 2{,}471319$, $\hat{\lambda}_X = 22965{,}385684$ для останньої ітерації. Отримані нормалізовані дані було очищено від 14 точок викидів та перевірено на відповідність нормальному закону розподілу. Далі, за (6) було побудовано лінійну регресійну модель з відповідними оцінками параметрів $\hat{b}_0 = -799519{,}0395$, $\hat{b}_1 = 0{,}999920$.

Для побудови двофакторної нелінійної регресійної моделі із застосуванням перетворення десяткового логарифму, обрані метрики X_1 (CLASS) та X_2 (VMQ) були перевірені на наявність мультиколінеарності за коефіцієнтом впливу дисперсії VIFs [13]. Значення коефіцієнтів VIFs не перевищило порогове значення 10, що підтверджує можливість застосування цих факторів для побудови регресійної моделі. Нормалізовані дані було очищено від 13 точок викидів та перевірено на відповідність нормальному закону розподілу. Далі, за (6) було побудовано лінійну регресійну модель. Оцінку параметрів лінійної моделі було визначено за методом найменших квадратів $\hat{b}_0 = -1{,}925069$, $\hat{b}_1 = 0{,}074531$, $\hat{b}_2 = 0{,}920560$. Застосувавши обернену трансформацію (7) до перетворення десяткового логарифму (8), двофакторна нелінійна регресійна модель має вигляд

$$\hat{Y} = 10^{\varepsilon - 1{,}925069} X_1^{0{,}074531} X_2^{0{,}920560}. \tag{16}$$

Побудовані нелінійні регресійні моделі були перевірені на навчальній і тестовій вибірках за критеріями якості R^2, MMRE та PRED(0,25), значення яких наведені в таблиці 1.

Таблиця 1

Критерії якості нелінійних регресійних моделей

Математична модель	Навчальна вибірка			Тестова вибірка		
	R^2	MMRE	PRED (0,25)	R^2	MMRE	PRED (0,25)
Однофакторна нелінійна регресійна модель на базі перетворення десяткового логарифму	0,7266	0,3522	0,4720	**0,7170**	**0,3350**	0,4632
Однофакторна нелінійна регресійна модель на базі перетворення Бокса-Кокса	0,7291	0,3514	0,4790	0,7159	0,3354	**0,4780**
Однофакторна нелінійна регресійна модель на базі перетворення Джонсона сімейства SB	0,6825	0,3593	0,4790	0,6616	0,3417	0,4702
Двофакторна нелінійна регресійна модель на базі перетворення десяткового логарифму	0,8002	0,2332	0,6853	**0,8981**	**0,1964**	0,7158

Критерії якості оцінки побудованих нелінійних регресійних моделей для оцінювання кількості рядків коду Java-застосунків для навчально та тестової вибірок суттєво не відрізняються, а отже це свідчить що вибірки мають високий рівень репрезентативності генеральної сукупності. Серед побудованих однофакторних моделей, однофакторна нелінійна регресійна модель із застосування перетворення десяткового логарифму найкраще описує тестову вибірку за R^2 та має менше значення MMRE для тестової вибірки. Але краще значення PRED(0,25) має однофакторна нелінійна регресійна модель із застосування перетворення Бокса-Кокса. Слід зазначити що значення R^2, MMRE та PRED(0,25) суттєво не відрізняються для однофакторних нелінійних регресійних моделей між собою для обох вибірок. Побудована двофакторна нелінійна регресійна модель із застосування перетворення десяткового логарифма має найкращі показники серед побудованих нелінійних регресійних моделей. Це пояснюється тим, що введення додаткової незалежного фактору регресії підвищує достовірність оцінювання залежної змінної. Модель (16) має високі значення критеріїв якості R^2 та MMRE та задовільне значення критерію якості PRED(0,25), додаткова перевірка за критерієм PRED із збільшенням порогового значення до 0,30 показала що за таких умов модель досягає прийнятного рівня точності PRED(0,30) = 0,8, що може свідчити про задовільний рівень точності отриманої моделі.

Висновки

1. Вирішено задачу удосконалення оцінювання кількості рядків коду Java-застосунків на ранніх стадіях розробки проєкту за метриками діаграми класів шляхом побудови двофакторної нелінійної регресійної моделі.

2. Наукова новизна отриманих результатів полягає в тому, що побудована на основі перетворення у вигляді десяткового логарифму двофакторна нелінійна регресійна модель має меншу середню величину відносної похибки, вище значення відсотка передбачення для рівня відносної похибки та вище значення коефіцієнту детермінації, що у порівнянні з існуючими моделями дозволяє підвищити достовірність оцінювання кількості рядків коду Java-застосунків.

3. Практичне значення отриманих результатів полягає у розробці програмного забезпечення для автоматизації оцінювання розміру Java-застосунків, яке реалізує двофакторну нелінійну регресійну модель на базі нормалізуючого перетворення десяктвого логарифму.

4. Перспективи подальших досліджень пов'язані із вдосконаленням багатофакторної нелінійної регресійної моделі за рахунок застосування багатовимірних нормалізуючих перетворень, збільшення набору даних навчальної вибірки, розширенні кількості незалежних факторів регресії для досягнення більшої достовірності оцінювання та побудови інтервалів прогнозування регресійної моделі.

Список використаної літератури

1. TIOBE Index. URL: https://www.tiobe.com/tiobe-index/ (дата звернення 08.04.2024).

2. Munialo S.W. A Review of Agile Software Effort Estimation Methods. *International Journal of Computer Applications Technology and Research. Association of Technology and Science*. 2016. Vol. 5. pp. 612–618. DOI:10.7753/IJCATR0509.1009.

3. Tan H.B.K., Zhao Y., Zhang H. Estimating LOC for information systems from their conceptual data models. *Proceedings – International Conference on Software Engineering*. 2006. pp. 321-330. DOI:10.1145/1134285.1134331.

4. Tan H.B.K., Zhao Y., Zhang H. Conceptual Data Model-Based Software Size Estimation for Information Systems, *ACM Transactions of Software Engineering and Methodology*. 2009. Vol. 19. DOI:10.1145/1571629.1571630.

5. Приходько Н.В., Приходько С.Б. Нелінійна регресійна модель для оцінювання розміру програмного забезпечення промислових інформаційних систем на Java. *Моделювання та інформаційні технології.* 2018. Вип. 85. С. 81–88. URL: http://nbuv.gov.ua/UJRN/Mtit_2018_85_14

6. Макарова Л.М., Приходько Н.В., Кудін О.О. Побудова нелінійної регресійної моделі для оцінювання розміру веб-додатків, реалізованих мовою Java. *Вісник Херсонського національного технічного університету*. 2019. №2 (69). С. 145–153. URL: http://eir.nuos.edu.ua/handle/123456789/4443

7. Приходько С.Б., Приходько Н.В., Смикодуб Т.Г. Чотирьохфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків з відкритим кодом. *Вчені записки ТНУ імені В.І. Вернадського. Серія: технічні науки* Том 31 (70) №2 Частина 1. 2020. С. 157–162. DOI:10.32838/2663-5941/2020.2-1/25

8. Port D., Korte M. Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, New York, 2008. pp. 51–60. DOI:10.1145/1414004.1414015

9. Jia J., Qiu W. Research on an Ensemble Classification Algorithm Based on Differential Privacy. *IEEE Access*. 2020. P. 99. DOI:10.1109/ACCESS.2020.2995058

10. Mardia K. V., Measures of multivariate skewness and kurtosis with applications, *Biometrika*. 1970. Vol. 57. pp. 519–530. DOI:10.1093/biomet/57.3.519

11. Prykhodko S., Prykhodko N., Mathematical Modeling of Non-Gaussian Dependent Random Variables by Nonlinear Regression Models Based on the Multivariate Normalizing Transformations, *Mathematical Modeling and Simulation of Systems (MODS'2020). Advances in Intelligent Systems and Computing*. 2021. Vol. 1265. PP. 166–174. DOI:10.1007/978-3-030-58124-4_16

12. Prykhodko S., Prykhodko N., Makarova L., Pukhalevych A. Outlier Detection in Non-Linear Regression Analysis Based on the Normalizing Transformations, *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*. Lviv-Slavske, Ukraine, 2020. pp. 407–410, DOI:10.1109/TCSET49122.2020.235464.

13. Olkin I., Sampson A.R. Multivariate Analysis: Overview. *International encyclopedia of social & behavioral sciences (eds.) 1st edn.*, Elsevier, Pergamon, 2001. pp. 10240–10247.

References

1. TIOBE (2024). TIOBE Index. URL: https://www.tiobe.com/tiobe-index/ (Accessed 08.04.2024)

2. Munialo, S.W., & Muketha, G.M. (2016). A review of Agile Software effort estimation methods. *International Journal of Computer Applications Technology and Research*, 5 (9), 612–618. https://doi.org/10.7753/ijcatr0509.1009

3. Tan, H.B.K., Zhao, Y. & Zhang H. (2006). Estimating LOC for information systems from their conceptual data models, *Proceedings – International Conference on Software Engineering*, 321-330. doi:10.1145/1134285.1134331.

4. Tan, H.B.K., Zhao Y. & Zhang, H. (2009). Conceptual Data Model-Based Software Size Estimation for Information Systems. *ACM Transactions of Software Engineering and Methodology*, Vol. 9. doi:10.1145/1571629.1571630.

5. Prykhodko, N.V. & Prykhodko S.B. (2018). A nonlinear regression model for estimation of the size of Java enterprise information systems software. *Modeling and Information Technologies*, Vol. 85, 81–88. URL: http://nbuv.gov.ua/UJRN/Mtit_2018_85_14 [in Ukrainian]

6. Makarova L.M., Prykhodko N.V. & Kudin O.O. (2019). Constructing the non-linear regression model for size estimation of WEB-applications implemented in JAVA. *Bulletin of Kherson National Technical University*, Vol. 69, P. 145–153. URL: http://eir.nuos.edu.ua/handle/123456789/4443 [in Ukrainian]

7. Prykhodko, S.B., Prykhodko, N.V. & T. G. Smykodub. (2020). Four-factor non-linear regression model to estimate the size of open source Java-based applications, *Scientific Notes of Taurida National V.I. Vernadsky University. Series: Technical Sciences*, Vol. 70, 157-–162. https://doi.org/10.32838/2663-5941/2020.2-1/25 [in Ukrainian]

8. Port, D. & Korte, M. (2008). Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, New York, 51–60. doi:10.1145/1414004.1414015

9. Jia, J. & Qiu W. (2020). Research on an Ensemble Classification Algorithm Based on Differential Privacy. *IEEE Access*. 99. doi:10.1109/ACCESS.2020.2995058

10. Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika*, Vol. 57, 519–530. doi:10.1093/biomet/57.3.519.

11. Prykhodko, S. & Prykhodko, N. (2021). Mathematical Modeling of Non-Gaussian Dependent Random Variables by Nonlinear Regression Models Based on the Multivariate Normalizing Transformations, *Mathematical Modeling and Simulation of Systems (MODS'2020). Advances in Intelligent Systems and Computing*, Vol. 1265, 166–174. doi:10.1007/978-3-030-58124-4_16

12. Prykhodko, S. & Prykhodko, N., Makarova, L. & Pukhalevych, A. (2020). Outlier Detection in Non-Linear Regression Analysis Based on the Normalizing Transformations, *IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, Lviv-Slavske, Ukraine, 407–410. doi:10.1109/TCSET49122.2020.235464.

13. Olkin, I. & Sampson, A.R. (2001). Multivariate Analysis: Overview, *International encyclopedia of social & behavioral sciences (eds.) 1st edn*, Elsevier, Pergamon, 10240–10247.