Section 4. Mathematical and simulation modeling

УДК 004.412:519.25

ЧОТИРЬОХФАКТОРНА НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ JAVA-ЗАСТОСУНКІВ НА РАННІХ СТАДІЯХ РОЗРОБКИ

О. Орєхов ORCID: 0000-0002-0001-0140
Національний університет кораблебудування імені адмірала Макарова, Україна
E-mail: oleksandr.oriekhov@nuos.edu.ua

Анотація. *У роботі розглядається побудова чотирьохфакторної нелінійної регресійної моделі для оцінки кількості рядків коду KLOC JAVA-застосунків. Оцінювання розміру JAVA-застосунку є актуальною науково-практичною задачею, яка невід'ємно пов'язана з життєвим циклом розробки програмного забезпечення (ПЗ). Метою роботи є підвищення достовірності оцінювання кількості рядків коду JAVA-застосунків на ранніх стадіях розробки ПЗ за метриками діаграми класів шляхом побудови чотирьохфакторної нелінійної регресійної моделі. 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 навчальну та тестові вибірки інформації з метрик JAVA-застосунків із відкритим програмних кодом, проведено аналіз та порівняння існуючих математичних моделей і рівнянь для оцінювання розміру JAVA-застосунків. На базі навчальної вибірки, побудовано чотирьохфакторну нелінійну регресійну модель та її інтервали прогнозування для оцінювання розміру JAVA-застосунків на основі нормалізуючого перетворення Бокса-Кокса за метриками кількості класів, загальною кількістю унікальних викликів методів в класах, середнім значенням кількості зв'язків між класами та середнім значенням кількості видимих методів на клас. Отримана чотирьохфакторна нелінійна регресійна модель має меншу середню величину відносної похибки, вище значення відсотка передбачення для рівня відносної похибки 0,25 та вище значення коефіцієнту детермінації у порівнянні з існуючими моделями, що дозволяє підвищити достовірність оцінювання кількості рядків коду JAVA-застосунків.*

Ключові слова: *оцінювання кількості рядків коду, JAVA-застосунок, негаусівські дані, нормалізуюче перетворення Бокса-Кокса, нелінійна регресійна модель.*

1. Вступ

Оцінка трудомісткості розробки програмного забезпечення є одним з ключових показників для бізнесу при формуванні бюджету та плануванні часу

розробки. Ця задача є важливою для IT-компаній, оскільки дозволяє отримати прогноз кількісних показників ресурсів, які необхідні для створення програмного проєкту, допоможе враховувати ризики і ресурси та дозволить підвищити ефективність процесу розробки. Критерії достовірної оцінки ґрунтуються на здастності її підтримувати успішну реалізацію програмного проєкту. Заниження або завищення оцінки трудомісткості може призвести до певних проблем. Достовірна оцінка є джерелом інформації для формування реальних витрат по ресурсам проєкту, що дозволяє менеджерам проєкту прийняти найбільш ефективні рішення для досягнення поставлених цілей [1].

Статистика успішності програмних проєктів, яка представлена компанією The Standish Group за період 25 років [2], має помірну позитивну динаміку в бік збільшення частки успішно реалізованих проєктів. Так, в 1994 частка успішно реалізованих проєктів складала 16%, а в 2020 році відсоток таких проєктів (виконані вчасно, в межах бюджету, реалізовано 100% вимог) вже становив 35%. Статистика провальних проєктів зменшилась з 31% до 19% за період з 1994 по 2020 рік. А частка проєктів, які зазнали проблеми (не вклалися в терміни, або вийшли за рамки бюджету, або функціонал не був повністю реалізований у відповідності до вимог), має незначну тенденцію до зменшення, біля 4% , рис. 1.

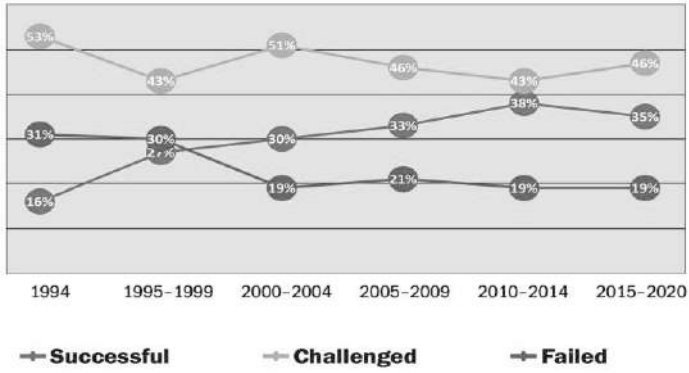

Рисунок 1. *Дослідження успішності реалізації IT проєктів за даними The Standish Group за період з 1994 року по 2020 рік [2].*

Де **successful** – успішні проєкти, **challenged** – проєкти з певними проблемами при реалізації, **failed** – проєкти, які провалились

Крім того дослідження The Standish group [3] доводить, що чим більший і складніший проєкт, тим вище імовірність появи труднощів при реалізації програмного проєкту відповідно до вимог, бюджету та часових рамок (рис. 2.), що були встановлені на початкових етапах планування.

	SUCCESSFUL	CHALLENGED	FAILED	TOTAL
Grand	6%	51%	43%	100%
Large	11%	59%	30%	100%
Medium	12%	62%	26%	100%
Moderate	24%	64%	12%	100%
Small	61%	32%	7%	100%

Рисунок 2. *Статистика успішності розробки ПЗ в залежності від його розміру [3]*

Мова програмування JAVA є однією з найбільш поширених мов програмування на сучасному ринку розробки ПЗ. JAVA - це високорівнева мова програмування, створена компанією Sun Microsystems (зараз Oracle) у 1995 році. Вона широко використовується в багатьох сферах, таких як розробка веб-застосунків, наукові обчислення, розробка штучного інтелекту, розробка мобільних застосунків та ігор [4]. Згідно з результатами досліджень та індексами популярності мов програмування, такими як IEEE, TIOBE, JAVA постійно утримується на передових позиціях [5, 6].

Але, не дивлячись на стрімкий розвиток галузі інформаційних технологій, дослідження успішності завершення проєктів ПЗ показали, що існують проблеми із достовірністю оцінювання трудомісткості та вартості розробки ПЗ. Аналіз останніх досліджень і публікацій показав, що збільшення достовірності оцінювання розміру ПЗ можливо досягти враховуючи такі особливості, як фактори середовища при оцінки трудомісткості розробки ПЗ за допомогою таких параметричних моделей як COCOMO II та інших [7].

Тому оцінка розміру програмного забезпечення на ранніх етапах розробки програмного проєкту, яка використовується в параметричних моделях оцінках трудомісткості та вартості є важливою науково-практичною задачею. Для досягнення достовірного рівня оцінювання розміру ПЗ потрібні відповідні математичні моделі, які враховують особливості мови програмування, такої як JAVA [7].

Метою і завданнями дослідження є підвищення достовірності оцінювання розміру JAVA-застосунків на ранніх стадіях розробки на основі наборів даних інформації з метрик коду проєктів з відкритим кодом за метриками діаграми класів із застосування чотирьохфакторної нелінійної регресійної моделі.

Для досягнення поставленої мети необхідно вирішити наступні завдання:

1. Провести аналіз існуючих моделей оцінювання розміру JAVA-застосунків. Дослідити їх переваги та недоліки. Порівняти існуючі моделі за критеріями якості регресійних моделей.

2. Удосконалити чотирьохфакторну нелінійні регресійну модель для оцінювання розміру програмних JAVA-застосунків на ранніх стадіях розробки за метриками діаграми класів із застосування багатовимірних нормалізуючих перетворень та визначити довірчий інтервал та інтервал прогнозування.

3. Побудувати довірчий інтервал і інтервал прогнозування нелінійної регресійної моделі оцінювання розміру JAVA-застосунків.

Об'єктом дослідження є процес оцінювання розміру кількості рядків коду JAVA-застосунків.

Предметом дослідження є нелінійні регресійні моделі для оцінювання розміру програмних JAVA-застосунків.

2. Аналіз існуючих математичних моделей для оцінки розміру JAVA-застосунків

Для оцінювання розміру JAVA-застосунків з відкритим кодом побудовано як лінійні [8, 9] так і нелінійні [10,11,12,13] регресійні рівняння та моделі в залежності від різної кількості та варіації метрик з концептуальної моделі даних у вигляді UML діаграм, зокрема діаграм класів. Так в [8, 9] запропоновані трьохфакторні лінійні регресійні рівняння для оцінювання JAVA-застосунків з відкритим кодом та інформаційних систем. Але, як відомо, при побудові лінійних регресійних моделей необхідно виконання певних умов, зокрема похибки повинні бути розподілені за нормальним законом, що можливо лише в поодиноких випадках для інформації з метрик програмних застосунків. Роботи [10,11,12,13] пропонують нелінійні регресійні моделі з різною кількістю факторів та наборів даних, які були використані для побудови цих моделей. Так [10] пропонує трьохфакторну нелінійну регресійну модель для оцінки розміру інформаційних систем на базі набору даних з робіт [8, 9], робота [11] пропонує однофакторну нелінійну регресійна модель для оцінювання JAVA веб-застосунків, ці регресійні моделі не беруть до уваги особливості оцінювання JAVA-застосунків які не є інформаційними системами або веб-застосунками. Робота [12] пропонує чотирьох факторну нелінійну регерсійну модель на основі перетворення Джонсона з сім'ї SB, але вона, як і [10], має певні обмеження границь допустимих значень, що накладене нормалізуючим перетворенням. Крім того, робота [12] доводить, що зі збільшенням факторів нелінійних регресійних моделей, зменшується ширина інтервалу прогнозування та довірчого інтервалу, це, в свою чергу, дозволяє врахувати оптимістичний та песимістичний сценарії. У [13, 14, 15] доведено що застосування вибірки більшого розміру та методів перехресного затвердження (Cross-Validation) дозволяє досягти більшої стійкості та надійності регресійних моделей у порівнянні з регресійними моделями, які

побудовані на вибірках з відносно невеликим розміром та без застосування методу перехресного затвердження.

Для порівнянні математичних моделей за критеріями якості, 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. 3.

Для оцінки достовірності прогнозування регресійних моделей та рівнянь використовуються такі критерії якості, як коефіцієнт детермінації R^2, значення середньої величини відносної похибки $MMRE$ [14] та значення відсотка передбачення для рівня відносної похибки 0,25 - $PRED\,(0{,}25)$ [14].

Коефіцієнт детермінації R^2 визначається як

$$R^2 = 1 - \frac{SSE}{SST}, \tag{1}$$

де SSE - сума квадратів залишків (the residual sum of squares) або сума квадратів помилок (the error sum of squares), $SSE = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$; SST – загальна сума квадратів (the total sum of squares), $SST = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

Критерій $MMRE$ визначається як

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i \tag{2}$$

де MRE_i - значення величини відносної похибки для і-го рядку даних випадкової величини

$$MRE_i = \left|\frac{(Y_i - \hat{Y}_i)}{Y_i}\right|. \tag{3}$$

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

$$PRED\,(0{,}25) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & if\ MRE_i \le 0{,}25\\ 0. & otherwise\end{cases}. \tag{4}$$

Достовірність оцінювання регресійними моделями вважається прийнятною, якщо значення $MMRE \le 0{,}25$, значення $PRED\,(0{,}25) \ge 0{,}75$. Значення R^2 вважається прийнятним, якщо воно більше за 0,75 [17].

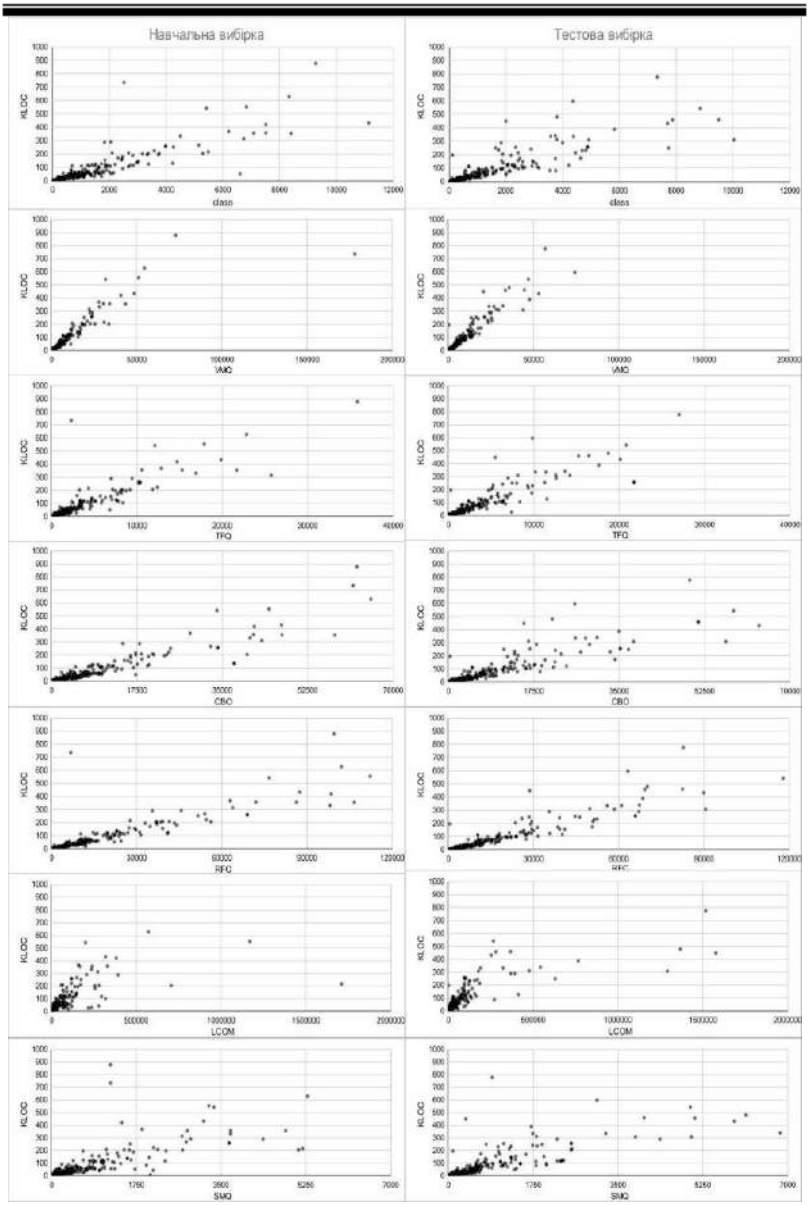

Рисунок 3. *Розподіл даних навчальної та тестової вибірок*

У роботі [13] існуючі регресійні моделі і рівняння оцінювання розміру рядків коду JAVA-застосунків [8, 9, 10, 11, 12] були перевірені за множинним коефіцієнтом детермінації R^2 (1), середньою величиною відносної помилки $MMRE$ (2) і відсотком прогнозованих результатів, для яких величини відносної помилки MRE (3) менші за 0,25, $PRED\,(0{,}25)$ (4). Результат перевірки показав їх незадовільний рівень достовірності як для навчальної так і для тестової вибірок.

Так оцінки R^2, $MMRE$ та $PRED\,(0{,}25)$ для лінійних регресійних рівнянь [8, 9] вийшли за межі допустимих значень, що свідчить про відсутність достовірних оцінок кількості рядків коду для JAVA-застосунків. Трьохфакторна нелінійна регресійна модель [10] не може бути застосована, оскільки через обмеження нормалізуючих перетворень значення оцінки кількості рядків коду не існують для 242 точок значень навчальної вибірки та для 236 точок значень тестової вибірки через обмеження нормалізуючого перетворення. В свою чергу чотирьохфакторна нелінійна регресійна модель [12] не може бути застосована для 22 точок значень навчальної вибірки та для 25 точок значень тестової вибірки через обмеження нормалізуючого перетворень. Незадовільні значення критеріїв можуть бути викликани різними причинами, зокрема, нормалізуючі перетворення існуючих моделей можуть мати певні обмеження, змінами в мові програмування JAVA та підходах розробки ПЗ, що впливають на структури та синтаксис програмних проєктів, недостатній рівень об'єктивності представлення генеральної сукупності вибірками, на основі яких були побудовані математичні моделі, тощо.

В роботі [13] було побудовано однофакторні нелінійні регресійні моделей для оцінювання кількості рядків коду (Y) в залежності від кількості класів (X) на базі нормалізуючих перетворень десяткового логарифму, Бокса-Кокса і Джонсона сім'ї SB та двофакторну нелінійну регресійну модель на базі нормалізуючого перетворення Джонсона сім'ї SB із застосуванням вище наведених навчальної і тестової вибірок (за метриками кількості класів X_1 та X_2 – загальною кількістю видимих методів класів). Результати порівнянн отриманих нелінійних регерсійних моделей за критеріями якості R^2, $MMRE$ та $PRED\,(0{,}25)$ наведені в Таблиці 1.

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 у порівнянні з побудованими в роботі [13] нелінійними регресійними моделей не дивлячись на застосування складних нормалізуючих перетворень. Це пояснюється тим, що введення додаткової незалежного фактору регресії підвищує достовірність оцінювання залежної змінної. Модель (23) має високі

значення критеріїв якості R^2 та $MMRE$ та задовільне значення критерію якості $PRED\,(0{,}25)$.

Таблиця1
Критерії якості однофакторних та двофакторних 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

Отже, існуючі регресійні моделі і рівняння [8, 9, 10, 11, 12, 13] були перевірені за множинним коефіцієнтом детермінації R^2, середньою величиною відносної помилки $MMRE$ (2) і відсотком прогнозованих результатів, для яких величини відносної помилки MRE (4) менші за 0,25, $PRED\,(0{,}25)$ (4). Результат перевірки показав їх незадовільний рівень достовірності для навчальної і тестової вибірок. Таким чином виникає необхідність вдосконалення існуючих моделей для оцінювання кількості рядків коду (KLOC) JAVA-застосунків.

3. Метод побудови нелінійної регресійної моделі

Зазвичай дані за метриками програмного коду не розподілені за нормальним законом, що робить обмежену можливість використання лінійних регерсійних моделей для оцінювання розміру рядків коду ПЗ. Теоретичною умовою застосування лінійних регресійних моделей є нормальний розподіл багатовимірних даних або нормальний розподіл залишків регресії є. Тому у випадку оцінювання розміру JAVA-застосунків переходять до застосування

нелінійних регресійних моделей. Для побудови нелінійних регресійних моделей для оцінювання кількості рядків коду JAVA-застосунків було обрано методи, застосовані у роботах [16, 17], які ґрунтуються на застосуванні взаємозворотних нормалізуючих перетворень з ітеративним видаленням викидів. У випадку виявлення викиду, метод повертається до першого кроку, використовуючи модифіковану вибірку без виявленого викиду з попередньої ітерації. Застосування нормалізуючих перетворень дозволяє перейти до побудови лінійних регерсійних моделей на основі нормалізованих даних із подальшим їх перетворенням в нелінійні регресійні моделі.

Нелінійна регресійна модель має такий загальний вигляд

$$Y = F(X_1, X_2, \dots, X_k, \varepsilon), \quad (5)$$

де ε – випадкова величина (ВВ) розподілена за нормальним законом, $\varepsilon = \mathcal{N}(0, \sigma_\varepsilon^2)$, яка визначає залишки, F - нелінійна регресійна модель на основі k - факторів, Y - залежна змінна KLOC.

На першому кроці проводиться нормалізації негаусівських багатовимірних даних.

Нехай взаємозворотнє багатовимірне нормалізуюче перетворення негауссівського випадкового вектора $P = \{Y, X_1, X_2, \dots, X_k\}^T$ у гаусівський випадковий вектор $T = \{Z_Y, Z_1, Z_2, \dots, Z_k\}^T$ задано як

$$T = \psi(P) \quad (6)$$

Обернене перетворення до (6) має вигляд

$$P = \psi^{-1}(T) \quad (7)$$

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

Для нормалізації багатовимірних негаусівських даних навчальної вибірки було обрано взаємозворотнє нормалізуюче перетворення Бокса-Кокса [18], яке задається як

$$Z = \begin{cases} (X^\lambda - 1)/\lambda, & if\ \lambda \neq 0 \\ ln(X), & if\ \lambda = 0 \end{cases} \quad (8)$$

де λ – параметри перетворення Бокса-Кокса.

Для оцінки параметрів нормалізуючого перетворення (6) було обрано метод метод максимальної правдоподібності

$$\hat{\theta} = \underset{\theta}{arg\ max}\ l(\theta), \quad (9)$$

де $l(\theta)$ - логарифмічна функція правдоподібності, θ - вектор параметрів перетворення $\theta = \{\lambda_1, \lambda_2, \dots, \lambda_k\}$, k - розмірність багатовимірних даних. Для перетворення Бокса-Кокса логарифмічна функція правдоподібності [19] для багатовимірних даних задано як

$$l(X, \theta) = \sum_{j=1}^{k} (\lambda_j - 1) \sum_{i=1}^{N} \ln(X_{ji}) - \frac{N}{2} ln[det(S_N)]. \quad (10)$$

де N - розмір вибірки, S_N – вибіркова коваріаційна матриця

$$S_N = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z})(Z_i - \bar{Z})^T \quad (11)$$

де Z_i - гаусівський випадковий вектор, $Z_i = (Z_{1_i}, Z_{2_i}, \dots, Z_{k_i})^T$; $\bar{Z}$- вектор вибіркових середніх, $\bar{Z} = (\overline{Z_1}, \overline{Z_2}, \dots, \overline{Z_k})^T$.

На другому кроці виконується перевірка нормального розподілу багатовимірних даних за допомогою критерію Мардія [21], який ґрунтується на відповідності нормальному закону розподілу параметрів багатовимірної асиметрії ($\beta_{1,k}$) та ексцесу ($\beta_{2,k}$)

$$\beta_{1,k} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[(X_i - \bar{X})^T S_N^{-1}(X_j - \bar{X})\right]^3, \quad (12)$$

$$\beta_{2,k} = \frac{1}{N}\sum_{i=1}^{N}[(X_i - \bar{X})^T S_N^{-1}(X_i - \bar{X})]^2, \quad (13)$$

де X - k-розмірний вектор випадкової величини, $X = (X_1, X_2, \dots, X_k)$, а S_N — зміщена вибіркова коваріаційна матриця багатовимірної ВВ X, визначена як

$$S_N = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T, \quad (14)$$

де $\bar{X}$ - це вектор вибіркових середніх, $\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)^T$.

За тестом Мардія очікувана асиметрія багатовимірних даних дорівнює 0, очікуваний ексцес становить $k(k+2)$ для багатовимірного нормального закону розподілу.

Для $\beta_{1,k}$ застосовується тестова статистика

$$\frac{N}{6}\beta_{1,k} \le \chi^2, \quad (15)$$

яка має наближений розподіл χ^2 з $k(k+1)(k+2)/6$ ступенями свободи та рівнем значущості α.

Для $\beta_{2,k}$ у якості тестової статистики використовується $1-\alpha$ квантиль $\mathcal{N}$ нормального закону розподілу з математичним сподіванням $\mu = k(k+2)$ та дисперсією $\sigma^2 = 8k(k+2)/N$

$$\beta_{2,k} \le \mathcal{N}_{1-\alpha}(\mu, \sigma^2). \quad (16)$$

На цьому кроці багатовимірні викиди із багатовимірних нормалізованих даних знаходяться за методом квадрату відстані Махаланобіса [22]. Визначається чи є одна багатовимірна точка із набору даних викидом, якщо в багатовимірному негаусівському наборі даних є багатовимірний викид, то таку точку відкидають і відбувається перехід до першого етапу, інакше до наступного етапу. Значення квадрату відстані Махаланобіса це елементи головної діагоналі матриці d^2, розміром $N \times N$.

$$d^2 = (Z_i - \bar{Z})^T S_N^{-1}(Z_i - \bar{Z}), \quad (17)$$

де S_N — зміщена вибіркова коваріаційна матриця (14), Z - ВВ з нормальним розподілом. Суть цього методу полягає в тому, що значення $d_i^2, i = 1,2,\dots N$, які перевищують значення квантилю розподілу χ^2 для рівня значущості α вважаються викидами і виключаються із вибірки. Зазвичай використовується значення $\alpha = 0{,}005$ для цього методу.

На третьому кроці відбувається побудова лінійної регресійної моделі на основі нормалізованої багатовимірної вибірки, яка має вигляд

$$Z_y = \hat{Z}_y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + .. + \hat{b}_k Z_k + \varepsilon, \quad (18)$$

де $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ - оцінки параметрів лінійної регресійної моделі, які знаходяться за методом найменших квадратів.

На четвертому кроці необхідно перевірити розподіл залишків регресії є побудованої лінійної регресійної моделі на відповідність нормальному закону розподілу. Якщо розподіл залишків лінійної регресійної моделі для нормалізованих даних не є нормальним, тоді необхідно відкинути точку даних, для якої модуль залишку є максимальним і повернутись до першого кроку.

На п'ятому кроці, після побудови моделі лінійної регресії (18) для нормалізованих даних, застосувавши зворотнє нормалізуюче перетворення (7), нелінійна регресійна модель буде мати наступний вигляд

$$Y = \psi_Y^{-1}(\hat{Z}_y + \varepsilon) = \psi_Y^{-1}(\hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \dots + \hat{b}_k Z_k + \varepsilon), \quad (19)$$

де ψ_Y^{-1} - функція зворотнього нормалізуючого перетворення перетворення оцінки лінійної регресії.

На шостому кроці побудови нелінійної регресійної моделі, будується верхня та нижня границі інтервалу прогнозування та проводиться пошук точок значень які знаходяться поза інтервалами прогнозування, якщо такі точки знайдені, то вони вважаються викидами і вилучаються із вибірки. Границі інтервалів прогнозування нелінійної регресії визначаються за наступною формулою

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,v} S_{Z_Y}\left\{1 + \frac{1}{N} + (Z_X^+)^T S_Z^{-1}(Z_X^+)\right\}^{1/2}\right), \quad (20)$$

де $t_{\alpha/2,v}$ - квантиль t-розподілу Стьюдента з $v = N - k - 1$ ступенями свободи та α/2 - рівнем значущості; $S_{Z_Y}^2 = \frac{1}{v}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$; Z_X^+ - вектор центральних факторів(регресорів), який містить значення $\{Z_{1_i} - \bar{Z}_1, Z_{2_i} - \bar{Z}_2, \dots, Z_{k_i} - \bar{Z}_k\}$; S_Z - $k \times k$ матриця

$$S_Z = \left[S_{Z_q} S_{Z_r}\right], \quad (21)$$

де $S_{Z_q} S_{Z_r} = \sum_{n-1}^{N}\left(Z_{q_i} - \bar{Z}_q\right)\left(Z_{r_i} - \bar{Z}_r\right)$, $q,r = 1,2, \dots,k$

4. Побудова чотирьохфакторної нелінійної регресійної моделі для оцінки розміру JAVA-застосунків

Для досягнення мети підвищення достовірності оцінювання розміру KLOC (Y) JAVA-застосунків, побудуємо чотирьохфакторну нелінійну регресійну модель на основі навчальної вибіркою даних (рис. 3) за параметрами кількості класів (CLASS) X_1, загальною кількістю унікальних викликів методів в класах (RFC) X_2, середнім значенням кількості зв'язків між класами (aCBO) X_3 та середнім значенням кількості public та protected методів на клас (aVMQ) X_4. Вибір наведених факторів регресійної моделі зроблений з урахуванням мультиколінеарності, оскільки високий рівень кореляції факторів регресії між собою підвищує чутливість моделі до випадкових змін у даних, знижує стабільність моделі та робить оцінку ваг параметрів моделі менш точною. Відсутність мультиколінеарності визначено за за коефіцієнтом впливу дисперсії (VIFs) для факторів регресії. Для багатофакторної регресійної моделі з *k*-факторами X_i , i=1 2,...,k, коефіцієнти VIFs представлені діагональними елементами оберненої кореляційної $k \times k$ матриці. Якщо значення коефіцієнту VIF перевищує 10, то це свідчить про наявність проблем із мультиколінеарністю [20]. Для факторі X_1, X_2 , X_3 та X_4 значення коефіцієнтів VIFs дорівнюють 6,3, 6,5, 1,3, та 1,2 відповідно, що свідчить про відсутність мультиколінеарності між факторами регресійної моделі. Багатовимірні дані були перевірені за критерієм Мардія на відповідність нормальному закону розподілу для рівня значущості $\alpha = 0{,}005$. Результати перевірки показали, що розподіл п'ятивимірних даних X_1 (CLASS), X_2 (RFC), X_3 (aCBO), X_4 (aVMQ) та Y (KLOC) є негаусівським, оскільки тестова статистика для багатовимірної асиметрії $(N \cdot \beta_1/6 = 12707{,}58)$ цих даних перевищує значення 60,27 квантилю розподілу χ^2, для 35 ступенів свободи, та значення багатовимірного ексцесу $\beta_2 = 399{,}58$, що є більшим за значення квантилю розподілу Гауса, яке становить 37,55, де $m = 35$ та $\sigma = 0{,}99$. Отже перехід до побудови нелінійної регресійної моделі є виправданим. Негаусівські дані навчальної вибірки були перетворені в гаусівські дані із застосування нормалізуючого перетворення Бокса-Кокса. Оцінки параметрів багатовимірного нормалізуючого перетворення (8) залежного фактора Y та незалежних факторів X_1, X_2, X_3 та X_4 мають наступні значення $\hat{\lambda}_Y = -0{,}050162$, $\hat{\lambda}_{X_1} = 0{,}014653$, $\hat{\lambda}_{X_2} = -0{,}028591$, $\hat{\lambda}_{X_3} = 0{,}650591$ та $\hat{\lambda}_{X_4} = -0{,}02$ відповідно для останньої ітерації. Оцінку параметрів перетворення було отримано за методом максимальної правдоподібності для багатовимірних даних [19] шляхом максимізації логарифмічної функції правдоподібності (9). На другому кроці було визначено і виключено із вибірки 25 багатовимірних викидів із нормалізованої вибірки. На останній ітерації багатовимірні нормалізовані дані були перевірені за критерієм Мардія на відповідність

нормальному закону розподілу для рівня значущості α=0,005. Результати перевірки показали, що розподіл п'ятивимірних даних Z_1, Z_2, Z_3, Z_4 та Z_Y є нормальним, оскільки тестова статистика для багатовимірної асиметрії $N \cdot \beta_1/6 = 56{,}36$ цих даних не перевищує значення 60,27 квантилю розподілу χ^2, для 35 ступенів свободи, та значення багатовимірного ексцесу $\beta_2 = 35{,}29$ менше за значення квантиля розподілу Гауса, яке становить 37,70, де m = 35 та σ = 1,05. Далі для нормалізованих даних побудовано лінійну регресійну модель

$$Z_y = \hat{Z}_y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \hat{b}_4 Z_4 + \varepsilon. \quad (22)$$

Оцінки параметрів чотирьохфакторної лінійної регресійної моделі на основі нормалізованих даних, отримані за методом найменших квадратів, становлять $\hat{b}_0 = -4{,}726846$, $\hat{b}_1 = 0{,}238926$, $b_2 = 0{,}756442$, $\hat{b}_3 = -0{,}089933$ та $\hat{b}_4 = 0{,}415064$.

Шляхом застосування зворотнього нормалізуючого перетворення (7) до лінійної регресії (22), чотирьохфакторна нелінійна регресійна модель має наступний вигляд

$$Y = \psi_Y^{-1}(\hat{Z}_y + \varepsilon) = \psi_Y^{-1}(\hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \hat{b}_4 Z_4 + \varepsilon). \quad (23)$$

На шостому кроці, для моделі (23) побудовано інтервал прогнозування та виявлено 6 викидів, які були ітеративно вилучені із нормалізованої навчальної вибірки. Для останньої ітерації значення нормалізованих вибіркових середніх $\bar{Z}_1$, $\bar{Z}_2$, $\bar{Z}_3$ та $\bar{Z}_4$ становлять $6{,}741137$, $7{,}747876$, $3{,}375148$ та $1{,}530621$ відповідно. Квантиль t-розподілу Стьюдента $t_{\alpha/2,\nu} = 2{,}594580$ для рівня значущості $\alpha = 0{,}01$ та 250 ступенів свободи; $S_{Z_Y} = 0{,}137762$. Обернена матриця (14) має вигляд

$$S_Z^{-1} = \begin{bmatrix} 0{,}0758, & -0{,}1071, & 0{,}0195, & 0{,}0430 \\ -0{,}1071, & 0{,}1556, & -0{,}0285, & -0{,}0645 \\ 0{,}0195, & -0{,}0285, & 0{,}0121, & 0{,}0058, \\ 0{,}0430, & -0{,}0645, & 0{,}0058, & 0{,}0706 \end{bmatrix}$$

Чотирьохфакторна нелінійна регресійна модель побудована за 32 ітерації, з багатовимірних нормалізованих даних було вилучено 31 багатовимірну точку як викид.

Для чотирьохфакторної нелінійної регресійної моделі знайдено довірчий інтервал, який має наступний вигляд

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{\frac{1}{N} + (Z_X^+)^T S_Z^{-1}(Z_X^+)\right\}^{1/2}\right). \quad (24)$$

5. Результати

Отриману чотирьохфакторну нелінійну регресійну модель (23) для оцінювання кількості рядків коду JAVA-застосунків перевірено за допомогою

критеріїв якості регресійних моделей R^2, $MMRE$ та $PRED\,(0{,}25)$, які мають значення 0,8242, 0,1621 та 0,8042 відповідно для навчальної вибірки та 0,8981, 0,1536 та 0,8211 відповідно для тестової вибірки, що свідчить про прийнятний рівень достовірності оцінки розміру JAVA-застосунків за допомогою отриманої нелінійної регресійної моделі у порівнянні з існуючими лінійними регресійними рівняннями та нелінійними регресійними моделями з роботи [13] (табл. 1).

Для порівняння достовірності прогнозування та якості отриманої регресійної моделі (23) побудовано інтервал прогнозування за навчальною вибіркою для однофакторної нелінійної регресійної моделі оцінки KLOC від параметру кількості класів (CLASS) *X* на основі нормалізуючого перетворення Бокса-Кокса з роботи [13].

Подібно до отриманої чотирьохфакторної нелінійної регресійної моделі (23), однофакторна нелінійна модель побудовано за тими ж методами [16, 17], що базуються на статистичному аналізі багатовимірних даних та багатовимірних взаємозворотних нормалізуючих перетвореннях.

Оцінки параметрів перетворення Бокса-Кокса однофакторної моделі мають наступні значення $\hat{\lambda}_Y = 0{,}002738,\ \hat{\lambda}_X = 0{,}030707$ а оцінки лінійної регресії мають вигляд $\hat{b}_0 = -2{,}697047$, $\hat{b}_1 = 0{,}857236$. Для отриманого інтервалу прогнозування значення нормованих вибіркових середніх $\bar{Z}_X = 7{,}045051$. Квантиль t-розподілу Стьюдента має значення $t_{\alpha/2,\nu} = 2{,}594025$ для рівня значущості $\alpha = 0{,}01$ та 272 ступенів свободи; $S_{z_Y} = 0{,}358707$.

Обернена матриця (14) складається з одного елемента $S_Z^{-1} = 1{,}529482 \cdot 10^{-3}$.

З навчальної вибірки випадковим чином було обрано 30 з 286 рядків даних для порівняння фактичних значень рядків коду (Y) та їх оцінок KLOC ($\hat{Y}$) за регресійними моделями, а також їх нижньої (LB) та верхньої межі (UB) інтервалів прогнозування Таблиця 2.

Порівняння ширини інтервалу прогнозування однофакторної та чотирьохфакторної нелінійних регресійних моделей оцінено за формулою

$$\mathrm{DIFF} = \left(1 - \frac{\sum_{i=1}^{N}\left|\mathrm{UB}_{4X_i} - \mathrm{LB}_{4X_i}\right|}{\sum_{i=1}^{N}\left|\mathrm{UB}_{1X_i} - \mathrm{LB}_{1X_i}\right|}\right) \cdot 100\%, \qquad (25)$$

де UB_{4X_i}, LB_{4X_i} - *i*-тий інтервал прогнозування чотирьохфакторної нелінійної регресійної моделі (25) та UB_{1X_i}, LB_{1X_i} - *i*-тий інтервал прогнозування однофакторної нелінійної регресійної моделі на базі перетворення Бокса-Кокса [13].

Таблиця 2.
Порівняння оцінок KLOC та інтервалів прогнозування нелінійних регресійних моделей

		Однофакторна нелінійна регресійна модель на основі Johnson SB			Чотирьохфакторна нелінійна регресійна модель на основі Johnson SB		
#	Y	$\hat{Y}$	LB	UB	$\hat{Y}$	LB	UB
1	1,649	1,687	0,520	5,453	1,643	0,926	2,965
2	1,769	1,733	0,534	5,600	1,891	1,065	3,415
3	1,806	3,108	0,963	9,992	1,944	1,088	3,535
4	1,848	1,641	0,506	5,306	1,596	0,902	2,870
5	2,013	1,641	0,506	5,306	1,658	0,937	2,986
…	…	…	…	…	…	…	…
6	7,522	9,564	2,987	30,509	8,293	4,484	15,641
7	7,609	11,621	3,634	37,030	8,314	4,492	15,691
8	7,628	8,727	2,724	27,853	6,407	3,492	11,983
9	8,056	9,285	2,900	29,623	9,562	5,155	18,093
10	8,056	9,285	2,900	29,623	9,562	5,155	18,093
…	…	…	…	…	…	…	…
11	16,056	14,210	4,448	45,236	18,360	9,700	35,489
12	16,098	9,704	3,031	30,952	10,948	5,867	20,846
13	16,202	10,311	3,222	32,876	14,232	7,566	27,331
14	16,398	10,124	3,163	32,284	15,359	8,148	29,559
15	16,583	14,683	4,596	46,735	7,354	3,982	13,850
…	…	…	…	.	…	…	…
16	28,724	32,742	10,280	103,900	30,421	15,835	59,751
17	29,139	45,379	14,259	143,888	35,407	18,323	69,977
18	29,296	23,981	7,522	76,173	25,235	13,211	49,264
19	29,575	8,448	2,637	26,968	34,412	17,487	69,350
20	29,591	11,903	3,722	37,923	21,758	11,412	42,389
…	…	…	…	.	…	…	…
21	64,504	60,178	18,919	190,722	47,373	24,287	94,568
22	65,525	45,884	14,418	145,486	65,229	33,014	132,023
23	66,253	100,858	31,721	319,514	50,521	25,726	101,586
24	66,455	44,723	14,053	141,812	59,534	30,028	120,939
25	66,695	37,155	11,670	117,866	55,095	28,057	110,778
…	…	…	…	…	…	…	…
26	205,51	289,349	90,906	917,617	235,629	113,912	500,996
27	208,68	137,768	43,327	436,468	178,114	87,140	373,880
28	208,96	106,094	33,368	336,100	150,519	73,989	314,346
29	216,61	300,958	94,545	954,518	220,525	106,233	470,675
30	222,65	188,859	59,379	598,493	248,491	120,069	528,630
N	…	…	…	…	…	…	…

Отримана оцінка вказує, що інтервал прогнозування чотирьохфакторної нелінійної регресійної моделі на $42{,}62\%$ вужчий за інтервал прогнозування

однофакторної нелінійної регресійної моделі на базі перетворення Бокса-Кокса [13].

6. Обговорення отриманих результатів

Отримано чотирьохфакторну нелінійну регресійну модель для ранньої оцінки KLOC JAVA-застосунків, використовуючи методи побудови нелінійних регресійних моделей на основі багатовимірного нормалізуючого перетворення Бокса-Кокса. Вибір методу зумовлений негаусівським розподілом п'ятивимірних даних інформації з метрик коду за критерієм Мардіа та залишків регресії за критерієм χ^2 Пірсона.

В порівнянні з існуючими моделями отримана чотирьохфакторна нелінійна регресійна модель (23) має кращі показники якості для оцінювання кількості рядків коду Java-застосунків за критеріями якості R^2, $MMRE$ та $PRED\,(0{,}25)$. Значення оцінок критеріїв якості знаходяться в допустимих межах $R^2 \geq 0{,}75$, $MMRE \leq 0{,}25$ та $PRED\,(0{,}25) \geq 0{,}75$ як для навчальної, так і для тестової вибірок, що свідчить про прийнятну точність отриманої моделі та мають наступні значення $R^2 = 0{,}8242$, $MMRE = 0{,}1621$, $PRED\,(0{,}25) = 0{,}8042$ для навчальної вибірки та $R^2 = 0{,}8981$, $MMRE = 0{,}1536$, $PRED\,(0{,}25) = 0{,}8211$ для тестової вибірки. Застосування методу перехресного затвердження підвищує впевненість в стійкості та надійності отриманої моделі. Крім того, ширина інтервалу прогнозування (20) моделі (23) на $42{,}62\%$ менша (Таблиця 2), ніж інтервал прогнозування однофакторної нелінійної регресійної моделі, що дозволяє підвищити точність оцінки параметру розміру, з урахуванням оптимістичного та песимістичного сценаріїв при плануванні розробки ПЗ.

До переваг запропонованої моделі (23) можна віднести можливість оцінки KLOC на ранніх етапах проєктування JAVA-застосунків з використанням чотирьох метрик коду, таких як загальна кількість класів (CLASS), загальна кількість унікальних викликів методів у класах (RFC), середнім значенням кількості зв'язків між класами (aCBO) та середнім значення видимих методів на клас (aVMQ), які можна отримати на ранніх етапах планування проєкту з діаграм класів UML. До недоліків запропонованої моделі (23) можна віднести наступні обмеження інтервалів значень незалежних змінних, де $CLASS \in [25\,;\,11147]$, $RFC \in [45\,;\,117847]$, $aCBO \in [0{,}12\,;\,24{,}57]$ та $aVMQ \in [1{,}59\,;\,85{,}92]$.

Незважаючи на використання великої вибірки метрик коду, запропонована модель побудована лише на основі проєктів з відкритим кодом JAVA-застосунків з платформи GitHub. Хоча модель досягла прийнятного рівня достовірності відповідно до критеріїв якості регресійних моделей [14], напрямки подальших досліджень можуть бути спрямовані у бік вдосконалення

моделі, шляхом розширення кількості та розділення незалежних факторів моделі, наприклад з урахуванням успадкування, або використання параметру кількості інтерфейсів окремо від кількості класів.

7. Висновки

Отримана чотирьохфакторна нелінійна регресійна модель вирішує важливу проблему оцінки розміру JAVA-застосунків на ранніх стадіях планування програмного проєкту з використанням метрик діаграми класів UML: загальної кількості класів (CLASS), загальної кількості унікальних викликів методів у класах (RFC), середнє значення кількості зв'язків між класами (aCBO) та середнього значення кількості public та protected методів на клас (aVMQ).

Наукова новизна отриманих результатів полягає в тому, що чотирьохфакторну нелінійну регресійну модель удосконалено у порівнянні з існуючими моделями та рівнянням для ранньої оцінки KLOC JAVA-застосунків із використанням багатовимірного нормалізуючого перетворення Бокса-Кокса; вперше побудовано чотирьохфакторну нелінійну регресійну модель з використанням вибірки, розміром більше 250 точок даних; достовірність, стійкість та надійність моделі перевірено на тестовій вибірці аналогічного розміру.

Отримана модель, у порівнянні з іншими нелінійними регресійними моделями, має вище значення коефіцієнта детермінації R^2, менше значення середньої відносної похибки $MMRE$ та більше значення відсотка передбачення рівня відносної похибки $PRED\,(0{,}25)$ як для навчальної, так і для тестової вибірок, а інтервал передбачення є меншим у порівнянні з однофакторною нелінійною регресійною моделлю на основі однакового нормалізуючого перетворення.

Оцінки критеріїв якості отриманої регресійної моделі для навчально та тестової вибірок суттєво не відрізняються, а отже це свідчить, що вибірки мають високий рівень репрезентативності генеральної сукупності.

Практичне значення отриманих результатів дозволяє рекомендувати побудовану модель для використання на практиці. Запропонована модель реалізована у вигляді програмного продукту, який може бути використаний менеджерами проєктів для оцінки трудомісткості розробки програмного забезпечення на мові JAVA на ранніх стадіях планування проєкту з метою зниження ризиків та оптимізації витрат.

Перспективи подальших досліджень можуть включати розширення та розділення набору незалежних факторів, використання інших багатовимірних нормалізуючих перетворень та збільшення розміру вибірки даних для побудови нелінійної регресійної моделі.

8. Література

[1] S. McConnel, Software Estimation: Demystifying the Black Art, Microsoft Press, Redmond, Washington, USA, 2006, 352 p.
[2] J. Johnson, H. Mulder, Endless Modernization: How Infinite Flow Keeps Software Fresh, The Standish Group, Hans Mulder's Lab, 2021. https://www.researchgate.net/publication/348849361_Endless_Modernization_How_Infinite_Flow_Keeps_Software_Fresh
[3] The Standish Group, Chaos report 2015, 2015. https://standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf
[4] Oracle, Java, 2024, URL: https://www.oracle.com/my/java/
[5] IEEE, "Top Programming Languages 2022", url: https://spectrum.ieee.org/top-programming-languages-2022
[6] TIOBE. "TIOBE Index of programming languages", url: https://www.tiobe.com/tiobe-index/
[7] Фаріонова Т. А., Орєхов О. С., Аналіз сучасного стану методів оцінювання трудомісткості розробки програмного забезпечення, Збірник наукових праць НУК ім. адм. Макарова, No. 1 (494), Миколаїв, Україна, 2024, С. 102-111. doi: https://doi.org/10.15589/znp2024.1(494).15
[8] H. B. K. Tan, Y. Zhao, H. Zhang, 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.
[9] H. B. K. Tan, Y. Zhao, H. Zhang, Conceptual Data Model-Based Software Size Estimation for Information Systems, ACM Transactions of Software Engineering and Methodology, No. 19, 2009. doi: 10.1145/1571629.1571630.
[10] Приходько Н. В., Приходько С. Б., Нелінійна регресійна модель для оцінювання розміру програмного забезпечення промислових інформаційних систем на Java, Моделювання та інформаційні технології, том 85, 2018, С. 81–88. url: http://nbuv.gov.ua/UJRN/Mtit_2018_85_14
[11] 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, № 2 (69), 2019, С. 145–153. url: http://eir.nuos.edu.ua/handle/123456789/4443
[12] 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 ч.1, 2020, С. 157–162. doi:10.32838/2663-5941/2020.2-1/25
[13] Орєхов О. С., Фаріонова Т. А., Математичні моделі для оцінювання розміру JAVA-застосунків, Вісник Херсонського національного технічного університету», № 2 (89), Херсон, 2024, С. 196-203. doi: https://doi.org/10.35546/kntu2078-4481.2024.2.28
[14] D. Port, M. Korte, 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

[15] J. Jia , W. Qiu, Research on an Ensemble Classification Algorithm Based on Differential Privacy, IEEE Access, 2020, P. 99. doi:10.1109/ACCESS.2020.2995058

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

[17] S. Prykhodko, N. Prykhodko, L. Makarova, A. Pukhalevych, 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.

[18] G. E. P. Box, An analysis of transformations, Journal of the Royal Statistical Society (B), No.2 (26), 1964, pp. 211-252.

[19] R. A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Pearson Prentice Hall, 2007, p. 800.

[20] I. Olkin, A. R. Sampson, Multivariate Analysis: Overview, in N. J. Smelser, P. B. Baltes, International encyclopedia of social & behavioral sciences (eds.) 1st edn., Elsevier, Pergamon, 2001, pp. 10240–10247.

[21] K. V. Mardia, Measures of multivariate skewness and kurtosis with applications, volume 57 of Biometrika, 1970, pp. 519–530, doi:10.1093/biomet/57.3.519.

[22] S. Chatterjee, B. Price, Regression analysis by example, New York: John Wiley & Son, 1977, 228 p.

THE FOUR-FACTOR NONLINEAR REGRESSION MODEL FOR EARLY JAVA-APPLICATIONS SIZE ESTIMATION

O. Oriekhov. ORCID: 0000-0002-0001-0140
Makarov National University of Shipbuilding, Ukraine.
E-mail: oleksandr.oriekhov@nuos.edu.ua

Abstract. *The paper is devoted to a four-factor nonlinear regression model building for early software lines of code (KLOC) estimation of JAVA applications. JAVA-application size estimation is an important scientific and practical task that is inextricably linked to the software development life cycle. The aim of the study is to increase the reliability and accuracy of JAVA-applications code lines estimation at the early stages of software development by building a four-factor nonlinear regression model using UML class diagram metrics. The object of the study is the process of size estimation for open-source Java-software. The subject of the study is the nonlinear regression models to estimate the software size. To achieve this goal, we collected training and test samples of JAVA-software code metrics information, analyzed and compared existing mathematical models and equations for JAVA-application size estimation. Using the training sample, the four-factor nonlinear regression model and its prediction intervals are built to estimate JAVA applications software size on a basis of the Box-Cox normalization transformation by the metrics of the total quantity of classes, the total quantity of unique method invocations in classes, the average value of classes relationships, and the average value of visible methods per class. The obtained four-factor nonlinear regression model has a lower estimate of mean magnitude of relative error, a higher estimate of percentage of prediction for magnitude of relative error of 0.25, and a higher estimate of determination coefficient in comparison to existing models, which confirm the reliability and accuracy increasing of early KLOC estimation of JAVA applications.*

Keywords: *number of lines of code, JAVA application, non-Gaussian data, Box-Cox normalizing transformation, nonlinear regression model.*