

ОЦІНЮВАННЯ РОЗМІРУ JAVA-ЗАСТОСУНКІВ ЗА ТРЬОХФАКТОРНОЮ НЕЛІНІЙНОЮ РЕГРЕСІЙНОЮ МОДЕЛЛЮ

Орехов О.С.¹; Фаріонова Т.А., к.т.н., доц.²

^{1,2} Національний університет кораблебудування імені адмірала Макарова

^{1,2} Україна, Миколаїв

¹ORCID 0000-0002-0001-0140; ²ORCID 0000-0003-3384-4712

Анотація. Робота присвячена побудові трьохфакторної нелінійної регресійної моделі для оцінювання кількості рядків коду Java-застосунків на базі взаємозворотного нормалізуючого перетворення Джонсона сім'ї SB в залежності від значення метрик загальної кількості класів, загальної кількості унікальних викликів методів в класах та середнього значення кількості *public* та *protected* методів на клас.

Ключові слова: нелінійна регресійна модель; нормалізуюче перетворення; java; оцінювання розміру програмних проєктів; *cross-validation*.

Вступна частина. Оцінювання кількості рядків коду JAVA-застосунків є важливою науково-практичною задачею, оскільки така оцінка розміру широко застосовуються для прогнозування трудомісткості розробки програмного забезпечення, яка є одним з ключових показників для бізнесу при формуванні бюджету та плануванні часу розробки. Ця задача є важливою для ІТ-компаній, оскільки дозволяє отримати прогноз кількісних показників ресурсів, які необхідні для створення програмного проєкту, допоможе враховувати ризики та дозволить підвищити ефективність процесу розробки. Оцінки трудомісткості розробки програмного забезпечення дозволяють отримати такі широковідомі параметричні моделі як СОСОМО, СОСОМО II, ISBSG, тощо, основним параметром яких є значення кількості рядків коду майбутнього програмного проєкту. Найбільш достовірні оцінки трудомісткості досягаються з урахуванням таких факторів середовища як мова програмування та категорія програмного застосунку [1].

Для оцінювання розміру JAVA-застосунків з відкритим кодом побудовано як лінійні так і нелінійні регресійні рівняння та моделі в залежності від різної кількості та варіації метрик з концептуальної моделі даних у вигляді UML діаграм, зокрема діаграм класів. Так, для порівняннн математичних моделей, були зібрані дані за метриками програмного коду 571 JAVA-застосунків, розташованих на платформі GitHub (<https://github.com>) за допомогою інструменту СК (<https://github.com/mauricioaniche/ck>) [3,4]. Були отримані наступні метрики: кількість рядків коду (KLOC), загальна кількість класів (CLASS), загальна кількість унікальних викликів методів в класі (RFC), загальна кількість видимих методів (VMQ), зв'язність між класами (CBO), тощо. Наведені 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 [3,4]. Оцінки кількості рядків коду за існуючими лінійними регресійними рівняннями та нелінійними регресійними моделями було порівняно за такими критеріями якості прогнозування регресійних моделей, як коефіцієнт детермінації R^2 , значення середньої величини відносної похибки *MMRE* [2] та значення відсотка передбачення для рівня відносної похибки 0,25 - *PRED*(0,25) [2]. Результат перевірки показав їх незадовільний рівень достовірності для оцінювання параметру розміру KLOC навчальної і тестової вибірок. Таким чином виникає необхідність вдосконалення існуючих моделей для оцінювання кількості рядків коду (KLOC) JAVA-застосунків.

Метою дослідження є підвищення достовірності оцінювання розміру JAVA-застосунків на ранніх стадіях розробки на основі наборів даних інформації з метрик коду застосунків з відкритим кодом, що отримані з діаграми класів, із застосування трьохфакторної нелінійної регресійної моделі.

Основна частина. Для побудови трьохфакторної нелінійної регресійної моделі для оцінювання кількості рядків коду JAVA-застосунків було обрано методи, запропоновані у роботах [5, 6], які ґрунтуються на застосуванні взаємозворотних нормалізуючих перетворень з ітеративним видаленням викидів. У випадку виявлення викиду, метод повертається до першого етапу, використовуючи модифіковану вибірку без виявленого викиду з попередньої ітерації. Застосування нормалізуючих перетворень дозволяє перейти до побудови лінійної регресійної моделі на основі нормалізованих даних із подальшим її перетворенням в нелінійну регресійну модель.

Зазвичай дані з метрик програмного коду не розподілені за нормальним законом, що робить обмежену можливість використання лінійних регресійних моделей для оцінювання розміру рядків коду ПЗ. Теоретичною умовою застосування лінійних регресійних моделей є нормальний розподіл багатовимірних даних або нормальний розподіл залишків регресії ϵ . Тому у випадку оцінювання розміру JAVA-застосунків переходять до застосування нелінійних регресійних моделей. В роботах [3, 4] було виявлено, що багатовимірні дані з метрик JAVA-застосунків не розподілені за нормальним законом, а отже перехід до побудови нелінійної регресійної моделі є виправданим.

Для досягнення мети підвищення достовірності оцінювання розміру KLOC (Y) JAVA-застосунків, побудуємо трьохфакторну нелінійну регресійну модель на основі навчальної вибірки [3, 4] за параметрами кількості класів (CLASS) X_1 , загальною кількістю унікальних викликів методів в класах (RFC) X_2 та середнім значенням кількості public та protected методів на клас (aVMQ) X_3 .

Вибір наведених факторів регресійної моделі зроблений з урахуванням мультиколінеарності. Відсутність мультиколінеарності визначено за коефіцієнтом впливу дисперсії (VIFs) для факторів регресії. Для багатфакторної регресійної моделі з k -факторами X_i , $i = 1, 2, \dots, k$, коефіцієнти VIFs представлені діагональними елементами оберненої коваріаційної $k \times k$ матриці. Якщо значення коефіцієнту VIF перевищує 10, то це свідчить про наявність проблем із мультиколінеарністю [7]. Для факторів X_1 , X_2 та X_3 значення коефіцієнтів VIFs дорівнюють 6,0889, 6,0890 та 1,0 відповідно, що свідчить про відсутність мультиколінеарності між факторами регресійної моделі.

Згідно з [5, 6] побудова нелінійної регресійної моделі складається з шести етапів.

На першому етапі проведено нормалізацію негаусівської багатовимірної випадкової величини (BB) з метрик JAVA-застосунків.

Для нормалізації багатовимірних негаусівських даних навчальної вибірки було обрано взаємозворотнє нормалізуюче перетворення Джонсона сімейства SB, яке задається як

$$\psi(X) = Z = \hat{\gamma} + \hat{\eta} \ln \left(\frac{X - \hat{\phi}}{\hat{\phi} + \hat{\lambda} - X} \right), \quad (1)$$

де $\psi(X)$ - функція нормалізуючого перетворення негаусівської BB X у гаусівську BB Z ; γ , η , ϕ та λ – параметри перетворення Джонсона, $\phi < X < \phi + \lambda$, $\eta > 0$, $\lambda > 0$.

Оцінки параметрів перетворення (1) були знайдені за допомогою методу максимальної правдоподібності. Для останньої ітерації значення оцінок для Y становлять:

$\hat{\gamma}_Y = -4,061130$, $\hat{\eta}_Y = -0,501291$, $\hat{\phi}_Y = -0,608187$, $\hat{\lambda}_Y = 82315,076609$; для X_1 : $\hat{\gamma}_{X_1} = -45,701128$, $\hat{\eta}_{X_1} = 48,673051$, $\hat{\phi}_{X_1} = -32,735283$, $\hat{\lambda}_{X_1} = 163,504695 \cdot 10^{12}$; для X_2 : $\hat{\gamma}_{X_2} = -41,803051$, $\hat{\eta}_{X_2} = 47,042155$, $\hat{\phi}_{X_2} = -110,321073$, $\hat{\lambda}_{X_2} = 627,451423 \cdot 10^{14}$; для X_3 : $\hat{\gamma}_{X_3} = -36,520911$, $\hat{\eta}_{X_3} = 37,582881$, $\hat{\phi}_{X_3} = -0,841250$, $\hat{\lambda}_{X_3} = 952,137743 \cdot 10^9$.

На другому етапі багатовимірні дані перевірені на відповідність нормального розподілу. На цьому етапі було вилучено 16 рядків багатовимірних викидів за методом квадрату відстані Махаланобіса.

На третьому етапі побудовано лінійну регресійну модель на основі нормалізованої багатовимірної вибірки, яка має вигляд:

$$Z_y = \hat{Z}_y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon, \quad (2)$$

де $\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3$ - оцінки параметрів лінійної регресійної моделі, які знаходяться за методом найменших квадратів, та мають значення $\hat{b}_0 = -23,784722$, $\hat{b}_1 = -0,005724$, $\hat{b}_2 = -0,005062$, $\hat{b}_3 = -0,008718$.

На четвертому етапі залишки регресії ε побудованої лінійної регресійної моделі (2) були перевірені на відповідність нормальному закону розподілу. На цьому етапі було знайдено та вилучено 1 багатовимірний викид.

На п'ятому кроці, застосувавши зворотнє нормалізуюче перетворення (1) до моделі лінійної регресії (2), нелінійна регресійна модель представлена як

$$Y = \psi_Y^{-1}(\hat{Z}_y + \varepsilon) = \psi_Y^{-1}(\hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon), \quad (3)$$

На шостому кроці визначено верхню та нижні границі інтервалі прогнозування. Було знайдена та вилучена 1 багатовимірна точка, яка знаходиться за межами інтервалу прогнозування.

Отриману трьохфакторну нелінійну регресійну модель (3) для оцінювання кількості рядків коду JAVA-застосунків перевірено за допомогою критеріїв якості регресійних моделей R^2 , $MMRE$ та $PRED(0, 25)$, значення яких 0,9073, 0,1645 та 0,7692 відповідно для навчальної вибірки та 0,9016, 0,1617 та 0,8175 відповідно для тестової вибірки, що свідчить про прийнятний рівень точності оцінки розміру JAVA-застосунків за допомогою нелінійної регресійної моделі (3).

Висновки. Побудовано трьохфакторну нелінійну регресійну математичну модель для оцінювання розміру JAVA-застосунків на базі взаємозворотнього нормалізуючого перетворення Джонсона сім'ї SB на основі метрик програмного коду CLASS, RFC, aVMQ. **Наукова новизна** отриманих результатів полягає у тому, що трьохфакторну нелінійну регресійну модель удосконалено у порівнянні з існуючими моделями та рівнянням для ранньої оцінки KLOC JAVA-застосунків із використанням багатовимірного нормалізуючого перетворення Джонсона сім'ї SB. Застосування методу перехресного затвердження підвищує впевненість в стійкості та надійності отриманої моделі. **Практичне значення** отриманих результатів дозволяє рекомендувати побудовану модель для використання на практиці. **Перспективи подальших досліджень** можуть включати розширення та розділення набору незалежних факторів, застосування інших взаємозворотних багатовимірних нормалізуючих

перетворень та збільшення розміру вибірки даних для побудови нелінійної регресійної моделі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Фаріонова, Т. А., & Орехов, О. С. (2024). Аналіз сучасного стану методів оцінювання трудомісткості розробки програмного забезпечення. *Збірник наукових праць НУК ім. адм. Макарова*, №1 (494), 102-111. DOI:[https://doi.org/10.15589/znp2024.1\(494\).15](https://doi.org/10.15589/znp2024.1(494).15)
- [2] Port, D., & Korte, M. (2008). Comparative studies of the model evaluation criteria 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
- [3] Орехов, О. С., & Фаріонова, Т. А. (2024), Математичні моделі для оцінювання розміру JAVA-застосунків. *Вісник Херсонського національного технічного університету*, № 2 (89), Херсон, 196-203, DOI: <https://doi.org/10.35546/kntu2078-4481.2024.2.28>
- [4] Oriekhov, O., Farionova, T., & Chernova, L. S. (2024). Three-factor nonlinear regression model of estimating the size of Java-software. *Proceedings of the 12th International Conference Information Control Systems & Technologies (ICST 2024)*, Odesa, Ukraine, September 23–25, 2024. URL: <https://ceur-ws.org/Vol-3790/paper44.pdf>
- [5] 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
- [6] 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.
- [7] Olkin, I., & Sampson, A. R. (2001). Multivariate Analysis: Overview. in N. J. Smelser & P. B. Baltes, *International encyclopedia of social & behavioral sciences (eds.) 1st edn.*, Elsevier, Pergamon, 10240–10247.

Oriekhov O. S., Farionova T. A.

JAVA-applications size estimation using the three-factor nonlinear regression model

Abstract. This work is devoted to constructing a three-factor nonlinear regression model for estimating the number of code lines of JAVA-applications using the reciprocal normalizing Johnson transform of the SB family, depending on the value of the metrics of the total quantity of classes, the total quantity of unique method invocations in classes, and the average number of public and protected methods per class.

Keywords: nonlinear regression model; normalizing transformation; java; software development cost estimation; cross-validation.