

МАТЕМАТИЧНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ З ВІДКРИТИМ КОДОМ ТИПУ MP3PLAYERS НА JAVA

Кулаковська Інесса Василівна, кандидат фізико-математичних наук,
Чорноморський національний університет імені Петра Могили,
Україна, м. Миколаїв
inessa.kulakovska@chmnu.edu.ua

Анотація. У статті описуються основні види метрик ПЗ і можливість їх застосування до аналізу трудомісткості розробки ПЗв залежності від розміру проекту. Модель, що побудована, в порівнянні з іншими регресійними моделями (як лінійними, так і нелінійними), має більший множинний коефіцієнт детермінації, менше значення середньої величини відносної похибки. Для отриманого рівняння досліджено нормальне розподілення остач, обчислені ширини довірчого інтервалу для нелінійної регресії.

Ключові слова: нелінійна регресійна модель, довірчий інтервал, оцінювання розміру програмного забезпечення, нормалізуюче перетворення, управління проектами, оцінка трудомісткості, розробка ПЗ, метод функціональних точок.

Kulakovska I.V. The non-linear regression equations for software size assessment with java mp3players type.

Abstract. The article describes the main types of software metrics and the possibility of applying them to the analysis of the complexity of software development depending on the size of the project. The model constructed, compared to other regression models (both linear and nonlinear), has a larger multiple coefficient of determination, less than the mean value of the relative error. For the obtained equation, the normal distribution of residuals was investigated, and the confidence interval widths for nonlinear regression were calculated.

Keywords: nonlinear regression model, confidence interval, software size estimation, normalization transformation, project management, complexity estimation, software development, functional point method.

Оцінювання розміру проекту, особливо на ранньому етапі розробки, відіграє важливу роль у практичних завданнях з його управління, розробки та впровадження. Як правило, кількість рядків коду програми є негаусівською випадковою величиною, яка залежить від ряду факторів, у тому числі й метрик ПЗ, які впливають на кінцевий результат. Метрики ПЗ кількісно визначають різні властивості програмних продуктів та програмних процесів у вигляді чисельного відображення. Мета полягає у виведенні одного або декількох факторів особливостей ПЗ, що дає змогу порівнювати ці значення з подібними проектами, зі специфічними стандартами, які притаманні застосункам mp3players на JAVA. З отриманих результатів можна прийти до висновку щодо якості ПЗ та всього програмного процесу, а також, якщо необхідно, подальших заходів.



Рис. 1. Залежність зростання обсягу робіт від збільшення розміру проекту

Оскільки для різних категорій методів оцінки існують принципово різні способи беручи до уваги фактор розміру проекту, виникає потреба у виборі метрики розміру проекту. Серед чинників, що впливають на оцінку, розмір проекту є найбільш важливим показником. Хоча оцінки розміру недостатньо для розуміння цілому розробляється продукту, існує явна залежність між розміром проекту і його трудомісткістю. На рис. 1 показана залежність зростання обсягу робіт від збільшення розміру проекту, розрахована за моделлю СОСОМО. Модель СОСОМО II - це модель регресії, заснована на кількості рядків коду (LOC). [2]

Метою роботи є побудова нелінійного регресійного рівняння для оцінювання розміру ПЗ інформаційних систем з відкритим кодом для програм mp3players на JAVA на основі рівняння множинної регресії, та дослідження множинного коефіцієнту детермінації, значення середньої величини відносної похибки і менші ширини довірчого інтервалу та інтервалу передбачення регресії.

Нормалізуючі перетворення є надійним способом побудови нелінійних регресійних моделей та рівнянь [1, 3]. Виникає потреба порівнянні моделей та у побудові відповідного рівняння для різних видів ПЗ інформаційних систем з відкритим кодом, наприклад для програм mp3 players на JAVA.

При використанні точної та надійної методології для оцінювання розміру можна зробити висновок про якість ПЗ та навіть усього програмного процесу. Перевірка адекватності математичної моделі для оцінювання розміру програмного забезпечення. Для перевірки адекватності лінійного рівняння регресії використовуємо коефіцієнт детермінації:

$$R^2 = 1 - (\sum_{i=1}^n (y_i - \hat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2), \quad (1)$$

де y_i - емпіричне значення y ; $\hat{y}_i$ - розрахункове значення y ; $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ середнє значення випадкової величини y . R^2 характеризує частку дисперсії, яка обумовлена регресією, в загальній дисперсії показника y . Достатньо ефективною та результативною можна вважати модель з показником детермінації $R^2 > 0,8$. Якщо $R^2 = 1$, тоді лінія регресії точно відповідає усім спостереженням та вимогам, а модель можна вважати адекватною та достовірною.

Величина коефіцієнта детермінації виступає важливим критерієм оцінки якості лінійних і нелінійних моделей. Для перевірки якості знайденого рівняння регресії, окрім критерію R^2 , використовується також сума квадратів відхилень S_v : $S_y = \sum_{i=1}^n [y_i - f(x_i)]^2$, де y_i - фактичне значення випадкової величини y ; $f(x_i)$ - розрахункове значення згідно рівняння регресії. Сума квадратів відхилень S_v використовується для перевірки якості як нелінійного, так і лінійного рівняння регресії. Цей параметр також використовується для порівняння різних моделей.

Вихідні дані для оцінювання розміру ПЗ, як правило, не мають нормального розподілу, що в результаті призводить до помилок у розрахунках та негативно впливає на достовірність отриманих результатів, у нашому випадку, на оцінювання розміру ПЗ. Щоб уникнути цієї проблеми, перед тим, як будувати математичну модель, потрібно нормалізувати вихідні дані. Для побудови нелінійного регресійного рівняння використовуємо вибірку шестивимірних даних у концептуальній моделі даних з 50 програм на відкритим кодом програм mp3 players на базі мови програмування Java взятих з репозиторію GitHub.

Більшість соціально-економічних показників формується під впливом не одного, а багатьох факторів. Метод побудови моделі такого зв'язку має назву багатофакторного кореляційно-регресійного аналізу. В цьому випадку результативна ознака (Y) пов'язується з допомогою рівняння множинної регресії з двома або більше факторними ознаками ($x_1, \dots, x_m$). Для побудови цього рівняння були використані дані всього з 32 інформаційних систем, кількість яких фактично межує з малою вибіркою. Рівняння регресії

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_6 x_6 \quad (2)$$

Позначення Y - LOC - число рядків коду програми; x_1 - NC - загальна кількість класів; x_2 - ANA - середня кількість атрибутів; x_3 - ANM - середня кількість методів; x_4 - ANSM -

середня кількість set методів; x_5 - ANGM - середня кількість get методів; x_6 - ANCM - середня кількість методів-конструкторів; x_7 - N_{inh} - кількість відносин між класами. Для побудови нелінійного регресійного рівняння використовуємо вибірку шестивимірних даних з: фактичний розмір ПЗ в тисячах рядків коду (LOC) Y , загальна кількість класів x_1 , загальна кількість зв'язків x_2 та інші у концептуальній моделі даних з 54 інформаційних систем з відкритим кодом програм mp3 players на базі мови програмування Java взятих з репозиторію GitHub.

Вибірка програм на відкритим кодом програм mp3 players на базі мови програмування Java перевірена на виброси методом Евклідової відстані та обмежена 32 програмами. При дослідженні також розглядалися різні види рівняння регресії. Коефіцієнти детермінації для різних видів такі: лінійна регресія - 0,697, параболічна регресія - 0,636, **експоненціальна регресія - 0.714**, степенева регресія - 0,626, гіперболічна регресія - 0,111; логарифмічна регресія - 0,615, показникова регресія - 0,521. За ними обрано експоненціальну регресію і надалі досліджувалося рівняння:

$$\ln Y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_6 x_6 \quad (3)$$

Далі за даними будуємо лінійне регресійне рівняння

$$\ln Y = 5.3318 + 0.022X_1 + 0.01X_2 + 0.08061X_3 - 0.07548X_4 - 0.7223X_5 + 1.2149X_6 \quad (4)$$

Значення множинного коефіцієнта детермінації (5) R^2 і середньої величини відносної похибки $MMRE$ дорівнюють 0,7146 та 0,042 відповідно. Значенні відсотка прогнозування $PRED(0.1) = 0.968$.

Перевіримо гіпотезу про те, що X розподілено по нормальному закону з допомогою критерію згоди Пірсона, p_i - ймовірність попадання в i -й інтервал випадкової величини, розподіленої по теоретичному закону. Значення суми із статистики Пірсона, **Кнабл = 2,6782**. Так як статистика Пірсона вимірює різницю між емпіричним і теоретичним розподілами, то чим більше її бачимо значення Кнабл, тим сильніше аргумент проти основної гіпотези. Тому критична область для цієї статистики завжди правобічна: $[K_{кр}; +\infty)$. Її кордон $K_{кр} = \chi^2(k-r-1; \alpha)$ знаходимо за таблицями розподілу χ^2 і заданим значенням s , k (число інтервалів), $r = 2$ (параметри $x_{ср}$ і s оцінені за вибіркою). **$K_{кр} = \chi^2(6-2-1; 0.05) = 7.81473$** ; Кнабл = 2.68 < 7.81.

Спостережуване значення статистики Пірсона не влучає у критичну область: Кнабл < $K_{кр}$, тому немає підстав відкидати основну гіпотезу. Справедливо припущення про те, що дані вибірки мають **нормальний розподіл**.



Рис. 2. Діаграма розподілу частот

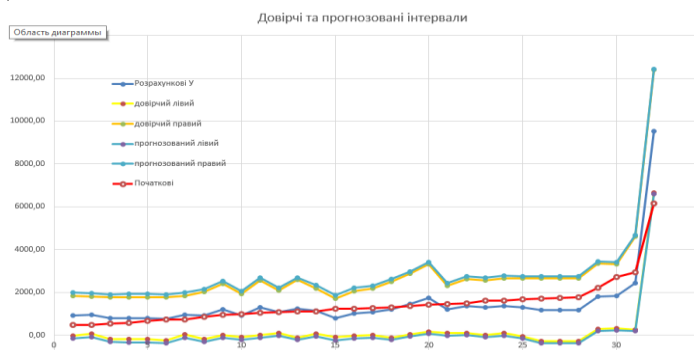


Рис.3. Довірчі та прогнозовані інтервали для Y .

Індивідуальні довірчі інтервали для Y при даному значенні вектора $Z(x_1, \dots, x_6)$. $(y \pm \varepsilon)$, де

$$\varepsilon = t_{кр} * S * \sqrt{\frac{1}{n} + (\bar{Z}_i) * R_{11}^{-1} * (\bar{Z}_i)^T} = 2.042 * 0.32 * \sqrt{\frac{1}{32} + (\bar{Z}_i) * R_{11}^{-1} * (\bar{Z}_i)^T};$$

де $t_{крит}(n-m-1; \alpha/2) = (30; 0.025) = 2.042$, S – оцінка середньоквадратичного відхилення (стандартна похибка Y) = 0,324629232.

Індивідуальні інтервали прогнозування для Y при даному значенні вектора $Z(x_1, \dots, x_6)$.
($y \pm \varepsilon$),

$$\varepsilon = t_{kr} * S_{Z_y} * \sqrt{1 + \frac{1}{n} + (\bar{Z}_i) * R_{11}^{-1} * (\bar{Z}_i)^T} = 2.042 * 267,55 * \sqrt{1 + \frac{1}{32} + (\bar{Z}_i) * R_{11}^{-1} * (\bar{Z}_i)^T}$$

де $t_{\text{крит}}(n-m-1; \alpha/2) = (30; 0.025) = 2.042$, R_{11} – матриця міжфакторної кореляції,

$$S_{Z_y}^2 = \frac{1}{n - m - 1} \sum_{s=1}^n (\bar{Z}_i - \bar{Z}_{\text{ср}})^2;$$

З імовірністю 95% можна гарантувати, що значення Y при необмежено великому числі спостережень не вийде за межі знайдених інтервалів.

Висновки. В результаті розрахунків було отримано рівняння множинної регресії:

$$\ln Y = 5.3318 + 0.022X_1 + 0.01X_2 + 0.08061X_3 - 0.07548X_4 - 0.7223X_5 + 1.2149X_6.$$

$$Y = 206,8099 * e^{0.022x_1 + 0.01x_2 + 0.08061x_3 - 0.07548x_4 - 0.7223x_5 + 1.2149x_6}.$$

Статистична значимість рівняння перевірена за допомогою коефіцієнта детермінації. Встановлено, що в досліджуваній ситуації 71.46% загальної варіабельності Y пояснюється зміною факторів x_j . Справедливо припущення про те, що дані вибірки мають нормальний розподіл. Таким чином, модель адекватна по нормальності розподілу залишкової компоненти. Удосконалено нелінійне регресійне рівняння для оцінювання розміру ПЗ інформаційних систем з відкритим кодом типу mp3players реалізованих мовою Java.

Отримані при аналізі метрики ПЗ кількісно визначають різні властивості програмних продуктів у вигляді чисельного відображення. Поставлена ціль дослідження полягала у виведенні шістьох немультіколінеарних факторів особливостей ПЗ, що дало змогу порівнювати ці значення з подібними проектами.

Список літератури

1. Prykhodko S.B. (2018) Constructing the non-linear regression equation to estimate the software size of open source PHPbased information systems / *Проблеми інформаційних технологій*. № 1 (023). – С.118-125. – ISSN 1998-7005
2. Boehm B. (2000) Software Cost Estimation with Cocomo II. *New Jersey, Prentice-Hall*. 544 p
3. Приходько Н.В., Приходько С.Б. (2018) Нелинейное регрессионное уравнение для оценки размера программного обеспечения информационных систем с открытым кодом на PHP. *Вісник східноукраїнського національного університету імені Володимира Даля*. № 6 (247).