УДК 004.412:519.237.5
DOI https://doi.org/10.32838/2663-5941/2020.2-1/25

Приходько С.Б.
Національний університет кораблебудування імені адмірала Макарова

Приходько Н.В.
Національний університет кораблебудування імені адмірала Макарова

Смикодуб Т.Г.
Національний університет кораблебудування імені адмірала Макарова

ЧОТИРЬОХФАКТОРНА НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ JAVA-ЗАСТОСУНКІВ З ВІДКРИТИМ КОДОМ

Метою роботи є створення множинної нелінійної регресійної моделі для оцінювання розміру Java-застосунків з відкритим кодом на основі багатовимірного нормалізуючого перетворення за значеннями змінних, що можуть бути визначені за діаграмою класів. Чотирьохфакторну нелінійну регресійну модель для оцінювання розміру Java-застосунків з відкритим кодом побудовано на основі нормалізації за допомогою п'ятивимірного перетворення Джонсона для сімейства S_B негаусівського набору даних: кількості строк коду (LOC); кількості класів (Classes); кількості статичних методів (NOSM); метрики, що характеризує відсутність згуртованості методів (Lack of Cohesion of Methods, LCOM), та кількості викликів унікального методу в класі (the Response for Class, RFC) з 38 застосунків, розташованих на сайті GitHub (https://github.com) за допомогою інструменту CK (https://github.com/mauricioaniche/ck). Також нормалізацію цього набору даних було здійснено і за допомогою двох одновимірних перетворень: у вигляді десяткового логарифму та перетворення Джонсона для сімейства S_B. Використання п'ятивимірного перетворення порівняно з одновимірними дозволяє врахувати кореляцію між змінними, що призводить до покращення нормалізації даних, яка пов'язана з виконанням статистичної гіпотези щодо відповідності їх розподілу п'ятивимірному розподілу Гаусу, з подальшим підвищенням достовірності відповідного оцінювання. Виконано порівняння побудованої нелінійної моделі з лінійною регресійною моделлю і нелінійними регресійними моделями на основі десяткового логарифму і одновимірного перетворення Джонсона. Нелінійна модель, що побудована, порівняно з іншими регресійними моделями (як лінійними, так і нелінійними) має більші значення множинного коефіцієнту детермінації та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, менші значення середньої величини відносної похибки та ширини інтервалу передбачення нелінійної регресії. Цей результат може бути пояснений найкращою багатовимірною нормалізацією і тим, що немає підстав відкидати нульову гіпотезу про те, що п'ятивимірний розподіл для нормалізованих даних, який нормалізується за допомогою п'ятивимірного перетворення Джонсона для сімейства S_B, є таким самим, як і п'ятивимірний нормальний розподіл.

Ключові слова: *нелінійна регресійна модель, інтервал передбачення, оцінювання розміру програми, Java-застосунок, нормалізуюче перетворення, негаусівські дані.*

Постановка проблеми. Сьогодні найпопулярнішою мовою програмування у світі [1] та серед програмістів в Україні [2] є Java, вперше випущена Sun Microsystems у 1995 році (https://www.java.com). Зараз Java використовується практично скрізь – від ноутбуків до центрів обробки даних, ігрових консолей до суперкомп'ютерів, мобільних телефонів до Інтернету.

Задача оцінювання розміру Java-застосунків з відкритим кодом, як і іншого програмного забезпечення (ПЗ) на ранній стадії розробки, є важливою, оскільки ця інформація використовується для прогнозування трудомісткості створення ПЗ за допомогою такої відомої моделі, як COCOMO II [3]. Це потребує відповідних моделей для оцінювання розміру ПЗ, включаючи Java-застосунки з відкритим кодом.

Аналіз останніх досліджень і публікацій. Натепер для оцінювання кількості строк коду інформаційних Java-систем з відкритим кодом існують як лінійні, так і нелінійні регресійні рівняння та моделі залежно від трьох метрик концептуальної моделі даних у вигляді діаграми класів [4–7]. В [4; 5] відповідне лінійне рівняння побудовано на основі методів множинного лінійного регресійного аналізу. Але, як відомо, під час

побудови лінійних регресійних моделей необхідно виконання певних умов, зокрема, похибки повинні бути розподілені за нормальним законом, що має місце лише в поодиноких випадках. А це веде до необхідності побудови нелінійних регресійних моделей, у тому числі і для оцінювання розміру ПЗ, та застосування певних методів множинного нелінійного регресійного аналізу [8].

Тому для оцінювання розміру інформаційних Java-систем з відкритим кодом в [7] було запропоновано рівняння нелінійної регресії, а в [6] – нелінійна регресійна модель. Запропоновані нелінійні регресійні рівняння та модель побудовано за допомогою множинного нелінійного регресійного аналізу із застосуванням чотиримірного перетворення Джонсона сім'ї S_B на основі таких же трьох метрик діаграми класів, що і в [4; 5]: загальна кількість класів, загальна кількість зв'язків та середня кількість атрибутів на клас. Але для Java-застосунків з відкритим кодом, що не є інформаційними системами, таких як різноманітні інструменти і фреймворки, регресійні моделі можуть залежати від інших метрик.

Зазвичай для побудови моделей нелінійної регресії використовують одновимірні нормалізуючі перетворення [9–13]. Але їх застосування для побудови нелінійних регресійних моделей не завжди призводить до задовільних результатів прогнозування, насамперед за такими стандартними показниками, як середня величина відносної похибки та відсоток передбачення [6–8]. Також нелінійні регресійні моделі, що побудовані за допомогою одновимірних нормалізуючих перетворень, зазвичай мають більші ширини довірчих інтервалів та інтервалів передбачення. Це призводить до необхідності використання багатовимірних нормалізуючих перетворень під час побудови нелінійних регресійних моделей, у тому числі і для оцінювання розміру Java-застосунків з відкритим кодом.

Постановка завдання. Метою статті є побудова чотирьохфакторної моделі нелінійної регресії та визначення нижньої і верхньої границь її інтервалів передбачення для оцінювання розміру Java-застосунків з відкритим кодом залежно від: кількості класів (Classes); кількості статичних методів (NOSM); метрики, що характеризує відсутність згуртованості методів (LCOM), та кількості викликів унікального методу в класі (RFC) на основі п'ятивимірного нормалізуючого перетворення Джонсона. Це дозволе підвищити достовірність оцінювання залежної змінної нелінійної регресії порівняно з лінійними моделями та нелінійними моделями з використанням одновимірних нормалізуючих перетворень.

Виклад основного матеріалу дослідження. Для досягнення цілі статті, що сформульована вище, ми скористалися методами, наведеними в [8, с. 100–102]. Згідно з [8, с. 100] спочатку виконується нормалізація багатовимірних негаусових даних за багатовимірним нормалізуючим перетворенням. Для побудови нелінійної регресійної моделі для оцінювання розміру Java-застосунків з відкритим кодом були зібрані дані з метрик 38 програм, розташованих на сайті GitHub (https://github.com): фактична кількість строк коду (в тисячах рядків коду) Y; кількість класів (Classes) X_1; кількість статичних методів (NOSM) X_2; метрика, що характеризує відсутність згуртованості методів (LCOM) X_3, та кількість викликів унікального методу в класі (RFC) X_4. Ці дані були отримані за допомогою інструменту CK (https://github.com/mauricioaniche/ck) та наведені у табл. 1. Вибір саме цих метрик був зумовлений практичною відсутністю мультиколінеарності між ними. Наявність мультиколінеарності свідчить про те, що в множинній лінійній регресійній моделі два або більше факторів пов'язані між собою або мають високий ступінь кореляції [14].

Наявність або відсутність мультиколінеарності ми визначали за коефіцієнтами впливу дисперсії (VIFs) серед майбутніх факторів у моделі множинної лінійної регресії. Для множинної лінійної регресійної моделі з k-факторами X_i, $i = 1, 2, \ldots, k$, VIFs – це діагональні елементи оберненої коваріаційної $k \times k$ матриці [15]. Значення VIFs більше за 10 часто сприймаються як сигнал, що дані мають проблеми з мультиколінеарністю. У разі, якщо значення VIFs знаходяться у межах від 1 до 5, то мультиколінеарності немає. Для X_1, X_2, X_3 та X_4 значення VIFs відповідно дорівнюють 4,79, 3,10, 1,12 та 5,63, що свідчить про практичну відсутність мультиколінеарності між цими факторами.

Згідно з [16], п'ятивимірні дані для змінних Y, X_1, X_2, X_3 та X_4, що наведені в табл. 1, мають негаусівський розподіл, оскільки для п'яти застосунків (2, 10, 12, 37 та 39) значення квадрату відстані Махаланобіса MD^2, які, відповідно, дорівнюють 30,17, 20,21, 34,02, 28,28 та 21,47, є більшими, ніж величина квантіля розподілу χ^2, що становить 16,75 для рівня значущості 0,005. Значення MD^2, що є більшими за 16,75, в табл. 1 виділені напівжирним шрифтом. Також про негаусівський розподіл п'ятивимірних даних для змінних Y, X_1, X_2, X_3 та X_4 з табл. 1 свідчить оцінка багатовимірного ексцесу β_2, яка визначалася за [16]. Відомо, що для m-вимірного нормального розподілу $\beta_2 = m(m+2)$. У нашому випадку $\beta_2 = 35$. Для цих п'ятивимірних даних оцінка β_2 дорівнює

100,77, що майже у 3 рази перевищує теоретичне значення.

Також у табл. 1 наведені значення нормалізованих змінних з метрик Java-застосунків з відкритим кодом, які були отримані за допомогою п'ятивимірного перетворення Джонсона сімейства S_B, компоненти якого визначаються, як і в [8]:

$$Z_j = \gamma_j + \eta_j \ln \frac{X_j - \phi_j}{\phi_j + \lambda_j - X_j}, \qquad (1)$$

де γ_j, η_j, ϕ_j та λ_j – параметри перетворення Джонсона, $\varphi_j<X_j<,\varphi_j<\lambda_j$, $j = 1,2,3,4$.

Значення нормалізованої залежної змінної Z_Y також визначається за (1) з тою різницею, що в (1) замість Z_j, X_j, γ_j, η_j, ϕ_j та λ_j потрібно підставити, відповідно, Z_Y, Y, γ_Y, η_Y, λ_Y та λ_Y.

Для даних з табл. 1 оцінки параметрів п'ятивимірного перетворення Джонсона сімейства S_B такі: $\hat{\gamma}_Y = 1{,}39056$, $\hat{\gamma}_1 = 1{,}0380$, $\hat{\gamma}_2 = 1{,}18671$, $\hat{\gamma}_3 = 2{,}44690$, $\hat{\gamma}_4 = 1{,}82065$, $\hat{\eta}_Y = 0{,}455264$, $\hat{\eta}_1 = 0{,}373337$, $\hat{\eta}_2 = 0{,}388760$, $\hat{\eta}_3 = 0{,}352541$, $\hat{\eta}_4 = 0{,}519554$, $\hat{\phi}_Y = 0{,}875$, $\hat{\phi}_1 = 13{,}500$, $\hat{\phi}_3 = 2{,}500$, $\hat{\phi}_3 = 137{,}252$, $\hat{\phi}_4 = 77{,}9181$, $\hat{\lambda}_Y = 868{,}212$, $\hat{\lambda}_1 = 4276{,}548$, $\hat{\lambda}_2 = 4627{,}30$, $\hat{\lambda}_3 = 26483781{,}6$

Таблиця 1

Дані з метрик Java-застосунків з відкритим кодом

№	Y	X_1	X_2	X_3	X_4	MD^2	Z_Y	Z_1	Z_2	Z_3	Z_4
1	1,075	22	5	994	217	0,77	-2,42	-1,28	-1,74	-1,20	-1,84
2	114,784	239	1306	9743493	6321	**30,17**	0,53	-0,04	0,82	2,26	0,16
3	29,468	298	529	15129	2589	0,70	-0,15	0,05	0,39	-0,19	-0,33
4	149,332	1243	1050	273258	19129	0,86	0,67	0,70	0,71	0,84	0,78
5	349,794	1528	1033	1237536	36496	16,05	1,21	0,81	0,70	1,38	1,18
6	75,862	845	361	33312	8862	0,88	0,32	0,51	0,22	0,09	0,34
7	133,992	553	472	272629	17434	1,49	0,61	0,32	0,34	0,84	0,73
8	1,771	14	9	564	156	0,76	-1,74	-2,34	-1,37	-1,44	-2,14
9	9,802	36	90	11729	1342	0,70	-0,69	-0,92	-0,35	-0,28	-0,69
10	262,21	1710	1276	118134	12494	**20,21**	1,01	0,88	0,81	0,54	0,53
11	111,224	752	848	62203	15890	0,33	0,51	0,45	0,60	0,31	0,67
12	817,049	4256	4553	562576	136233	**34,02**	2,64	2,84	2,77	1,10	2,72
13	29,148	198	182	37993	2882	0,39	-0,15	-0,12	-0,06	0,14	-0,27
14	196,028	1690	1523	104556	22694	2,99	0,83	0,87	0,91	0,50	0,88
15	52,132	499	286	13735	5423	0,27	0,13	0,27	0,13	-0,22	0,07
16	29,033	201	89	22545	2510	0,45	-0,16	-0,11	-0,35	-0,05	-0,35
17	1,949	15	26	175	292	0,76	-1,66	-1,93	-0,87	-2,30	-1,62
18	11,385	40	6	4388	1729	0,78	-0,61	-0,86	-1,61	-0,63	-0,55
19	3,266	17	16	1001	785	0,81	-1,29	-1,62	-1,08	-1,20	-0,99
20	3,117	19	4	887	436	0,77	-1,32	-1,45	-1,94	-1,25	-1,35
21	1,744	26	3	263	236	0,76	-1,75	-1,14	-2,36	-1,87	-1,77
22	39,118	171	155	8404	2796	0,62	-0,01	-0,18	-0,13	-0,40	-0,29
23	44,566	345	340	20297	7664	0,60	0,05	0,11	0,20	-0,08	0,26
24	30,496	256	157	5590	3320	0,38	-0,13	-0,01	-0,12	-0,55	-0,20
25	198,999	1891	542	279133	28402	4,95	0,84	0,95	0,40	0,85	1,02
26	202,452	1339	310	98554	26446	2,98	0,85	0,74	0,16	0,48	0,98
27	3,441	24	21	556	339	0,73	-1,26	-1,20	-0,96	-1,45	-1,51
28	147,937	748	716	81283	18112	0,90	0,67	0,45	0,53	0,41	0,75
29	242,821	1870	1480	4341969	27904	7,66	0,96	0,94	0,89	1,87	1,01
30	79,187	662	439	25388	8376	0,20	0,34	0,40	0,31	0,00	0,31
31	19,884	234	85	2618	4241	0,86	-0,34	-0,05	-0,37	-0,82	-0,06
32	27,83	357	100	7703	3399	0,57	-0,18	0,13	-0,31	-0,43	-0,18
33	12,079	72	37	12157	1867	0,71	-0,58	-0,56	-0,71	-0,27	-0,51
34	22,274	223	94	9883	3627	0,63	-0,28	-0,07	-0,33	-0,34	-0,15
35	63,609	388	513	75303	7228	0,25	0,23	0,16	0,38	0,38	0,23
36	102,473	247	1270	2042353	9178	3,28	0,47	-0,03	0,81	1,57	0,36
37	131,096	834	3403	189931	10820	**28,28**	0,60	0,50	1,58	0,71	0,45
38	377,943	3829	1759	195632	48231	**21,47**	1,27	1,83	1,00	0,72	1,38

і $\hat{\lambda}_4 = 160289{,}7$.

Далі для нормалізованих даних будуємо лінійну регресійну модель

$$Z_Y = \hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \hat{b}_4 Z_4 + \varepsilon , \qquad (2)$$

де ε – випадкова величина з розподілом Гаусу, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$; оцінки параметрів для даних з табл. 1 є такими: $\hat{b}_0 = 0$, $\hat{b}_1 = 0{,}094587$, $\hat{b}_2 = 0{,}178654$, $\hat{b}_3 = 0{,}090668$, $\hat{b}_4 = 0{,}654896$. Оцінки параметрів моделі (2) визначалися за методом найменших квадратів. За даними з табл. 1, сума квадратів відхилень для моделі (2) склала 1,0378.

Потім будуємо нелінійну регресійну модель за [8, с. 101]

$$Y = \hat{\phi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y\right)/\hat{\eta}_Y} \right]^{-1}, \qquad (3)$$

де відповідні складники визначаються за (2). За даними з табл. 1, сума квадратів відхилень для моделі (3) склала 25 414,8.

Побудована модель (3) була перевірена за множинним коефіцієнтом детермінації R^2, середньою величиною відносної помилки MMRE і відсотком прогнозованих результатів, для яких величини відносної помилки MRE менші за 0,25, PRED(0,25). Ці показники зазвичай використовуються для оцінювання якості прогнозування за допомогою регресійних моделей і в інженерії програмного забезпечення [17; 18]. Допустимі значення MMRE і PRED(0,25) складають не більше 0,25 і не менше 0,75 відповідно. Допустиме значення R^2 приблизно таке ж, як для PRED(0,25).

Для моделі (3), що була побудована за даними з табл. 1 на основі п'ятивимірного перетворення Джонсона сімейства S_B, значення R^2, MMRE та PRED(0,25) мають, відповідно, такі значення: 0,971, 0,173 та 0,789, що вказує на добру її якість стосовно оцінювання розміру Java-застосунків з відкритим кодом.

Для порівняння моделі (3) для оцінювання розміру Java-застосунків з відкритим кодом в подальшому було також побудовано лінійну та нелінійну регресійні моделі на основі одновимірного перетворення Джонсона. Для цього ми також використовуємо значення змінних Y, X_1, X_2, X_3 та X_4, що наведені в табл. 1.

Лінійна регресійна модель для оцінювання розміру Java-застосунків з відкритим кодом має вигляд

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \hat{b}_4 X_4 + \varepsilon , \qquad (4)$$

де оцінки параметрів такі: $\hat{b}_0 = 5{,}9139$, $\hat{b}_2 = 0{,}051232$, $\hat{b}_2 = 0{,}015127$, $\hat{b}_3 = 6{,}0916 \cdot 10^{-6}$, $\hat{b}_4 = 0{,}0039244$.

Сума квадратів відхилень для моделі (4) склала 26 204,7, що всього на 3,1% більше за цю суму для моделі (3). Також невелика різниця між значеннями R^2 для моделей (3) і (4): всього 0,1%. Значення двох інших показників лінійної моделі (4) – MMRE і PRED(0,25), що дорівнюють 0,744 і 0,526 відповідно, вказують на незадовільну її якість з оцінками параметрів, що були отримані за значеннями змінних Y, X_1, X_2, X_3 та X_4 з табл. 1.

Перевірку нульової гіпотези про нормальність закону розподілу випадкової величини ε для моделі (4) здійснюємо за критерієм Пірсона. Для вибірки значень випадкової величини ε значення χ^2 , яке дорівнює 81,32, більше за $\chi^2_{кр}$, що становить 7,81 для 3 ступенів вільності та 0,05 рівня значущості. Тобто цю гіпотезу про нормальність розподілу випадкової величини ε потрібно відкинути. Це свідчить взагалі про відсутність теоретичного обґрунтування використання моделі лінійної регресії (4) і призводить до необхідності застосування нелінійної регресійної моделі для оцінювання розміру Java-застосунків з відкритим кодом.

Нелінійна регресійна модель для оцінювання розміру Java-застосунків з відкритим кодом на основі одновимірного перетворення Джонсона сімейства S_B також має вигляд (3), але оцінки параметрів, що отримані за значеннями змінних Y, X_1, X_2, X_3 та X_4 з табл. 1, дещо інші: $\hat{\gamma}_Y = 1{,}39056$, $\hat{\gamma}_1 = 1{,}0380$, $\hat{\gamma}_2 = 1{,}18671$, $\hat{\gamma}_3 = 1{,}17761$, $\hat{\gamma}_4 = 1{,}66557$, $\hat{\eta}_Y = 0{,}455264$, $\hat{\eta}_1 = 0{,}373337$, $\hat{\eta}_2 = 0{,}388760$, $\hat{\eta}_3 = 0{,}209433$, $\hat{\eta}_4 = 0{,}468450$, $\hat{\phi}_Y = 0{,}875$, $\hat{\phi}_1 = 13{,}500$, $\hat{\phi}_2 = 2{,}500$, $\hat{\phi}_3 = 174{,}50$, $\hat{\phi}_4 = 139{,}989$, $\hat{\lambda}_Y = 868{,}212$, $\hat{\lambda}_1 = 4276{,}548$, $\hat{\lambda}_2 = 4627{,}30$, $\hat{\lambda}_3$=9743319,0 і $\hat{\lambda}_4 = 153563{,}5$. Також інші і оцінки параметрів моделі (2): $\hat{b}_0 = 0$, $\hat{b}_1 = 0{,}153713$, $\hat{b}_2 = 0{,}228632$, $\hat{b}_3 = 0{,}0670969$, $\hat{b}_4 = 0{,}586594$.

Сума квадратів відхилень для моделі (3) з оцінками параметрів, що отримані на основі одновимірного перетворення Джонсона сімейства S_B, склала 28398,6, що на 11,7% більше за цю суму для моделі (3) з оцінками параметрів, що отримані на основі п'ятивимірного перетворення Джонсона сімейства S_B. Також гірші значення для моделі (3) з оцінками параметрів, що отримані на основі одновимірного перетворення Джонсона сімейства S_B, мають три інші показники – R^2, MMRE і PRED(0,25), що дорівнюють 0,968, 0,176 і 0,711 відповідно. Хоча слід зазначити, що значення MMRE і PRED(0,25) значно кращі, ніж у лінійної моделі (4).

Таблиця 2

Межі інтервалів передбачення нелінійних регресій

№	Одновимірне		П'ятивимірне		№	Одновимірне		П'ятивимірне	
	LB	UB	LB	UB		LB	UB	LB	UB
1	1,20	3,12	1,24	2,99	20	1,52	5,08	1,55	4,57
2	45,99	316,74	41,88	212,74	21	1,13	2,78	1,15	2,57
3	14,75	79,99	13,12	63,11	22	11,42	59,01	10,96	49,27
4	78,51	319,04	91,23	322,20	23	27,24	135,89	28,78	125,71
5	126,11	441,94	162,79	480,73	24	13,18	68,27	12,43	56,43
6	34,58	167,74	36,16	153,74	25	93,53	366,15	114,60	384,17
7	54,64	250,17	68,06	267,52	26	72,33	311,51	88,90	329,42
8	0,98	1,62	1,09	2,09	27	1,78	6,49	1,67	5,17
9	5,39	27,67	5,39	23,72	28	61,97	271,63	71,82	274,32
10	64,20	283,07	66,32	260,91	29	125,62	436,82	154,86	471,30
11	59,02	260,37	66,05	255,55	30	33,39	161,86	34,43	146,74
12	743,90	849,30	748,40	847,83	31	13,04	68,27	12,65	58,63
13	12,75	65,93	12,74	57,90	32	12,83	67,57	12,35	56,64
14	97,62	371,75	107,86	363,15	33	6,23	31,73	6,42	28,24
15	22,21	112,77	21,81	97,25	34	12,59	65,26	12,67	57,00
16	10,12	52,56	10,04	45,71	35	30,23	148,45	32,53	139,25
17	1,34	4,39	1,35	3,91	36	43,17	212,42	51,75	226,62
18	3,52	18,84	3,89	18,04	37	72,18	328,61	72,31	296,22
19	2,34	10,93	2,37	9,49	38	208,17	587,82	232,28	581,96

Для визначення нижньої і верхньої границь інтервалів передбачення нелінійних регресій ми використовували відповідний метод, запропонований у [8, с. 101]. Нижні (LB) і верхні границі (UB) інтервалів передбачення нелінійних регресій для моделі (3), як для одновимірного, так і для п'ятивимірного перетворення Джонсона сімейства S_B, наведені в табл. 2.

Для моделі (3) з оцінками параметрів, що були отримані за даними з табл. 1 з 38 Java-застосунків на основі п'ятивимірного перетворення Джонсона сімейства S_B, ширини інтервалу передбачення нелінійної регресії менші для 32 рядків даних (всі, окрім рядків 5, 7, 8, 26, 29 та 36) порівняно з одновимірним перетворенням Джонсона. Причому різниця у ширині інтервалу передбачення для даних з рядка 2 складає 58%.

Кращі показники оцінювання розміру Java-застосунків з відкритим кодом за моделлю нелінійної регресії на основі п'ятивимірного нормалізуючого перетворення Джонсона сімейства S_B можна передусім пояснити кращою багатовимірною нормалізацією, яка перевірялася за відомими критеріями [15]. Так, якщо за критерієм на основі квадрата відстані Махаланобіса гіпотеза про нормальність багатовимірного закону розподілу нормалізованих за допомогою п'ятивимірного нормалізуючого перетворення Джонсона сімейства S_B даних для 38 застосунків з табл. 1 приймається для рівня значущості 0,0001, то у разі застосування одновимірного перетворення та без нього – відкидається.

Висновки. Удосконалено чотирьохфакторну модель нелінійної регресії та визначення нижньої і верхньої границь її інтервалів передбачення для оцінювання розміру Java-застосунків з відкритим кодом залежно від кількості класів; кількості статичних методів; метрики, що характеризує відсутність згуртованості методів, та кількості викликів унікального методу в класі на основі п'ятивимірного нормалізуючого перетворення Джонсона сімейства S_B. Це дозволяє підвищити достовірність оцінювання залежної змінної нелінійної регресії порівняно з використанням одновимірних нормалізуючих перетворень. Модель, що побудовано, порівняно з іншими регресійними моделями має більший відсоток прогнозування, менші середні величини відносної похибки та ширини інтервалу передбачення нелінійної регресії. В майбутньому планується використання інших наборів даних для побудови нелінійної регресійної моделі для оцінювання розміру Java-застосунків з відкритим кодом.

Список літератури:

1. TIOBE Index for April 2020. URL: https://www.tiobe.com/tiobe-index/ (дата звернення: 03.04.2020)
2. Рейтинг мов програмування 2019: JavaScript майже зрівнялася з Java, популярність Go знижується URL: https://dou.ua/lenta/articles/language-rating-jan-2019/ (дата звернення: 03.04.2020)
3. Boehm B.W., Abts C., Brown A.W., Chulani S., Clark B.K., Horowitz E., Madachy R., Reifer D.J., Steece B. Software Cost Estimation with COCOMO II. Upper Saddle River, NJ : Prentice Hall PTR, 2000. 544 p.
4. Tan H.B.K., Zhao Y., Zhang H. Estimating LOC for information systems from their conceptual data models. *Proceedings of the 28th International Conference on Software Engineering (ICSE '06)*. (May 20-28, 2006, Shanghai, China). Shanghai, 2006. P. 321–330.
5. Tan H.B.K., Zhao Y., Zhang H. Conceptual data model-based software size estimation for information systems. *Transactions on Software Engineering and Methodology*. 2009. Vol. 19. Issue 2. October 2009. Article No. 4.
6. Prykhodko N.V., Prykhodko S.B. The non-linear regression model to estimate the software size of open source Java-based systems. *Radio Electronics, Computer Science, Control*. 2018. No. 3 (46). P. 158–166. DOI: 10.15588/1607-3274-2018-3-17.
7. Prykhodko S.B., Prykhodko N.V., Mandra A.V. Building the nonlinear regression equations to estimate the software size of Java-based information systems. *Materials of the VII International scientific-practical conference on Information Control Systems and Technologies*. (17th-18th September, 2018, Odessa). Odessa, Astroprint, 2018. P. 222–224.
8 Prykhodko N.V., Prykhodko S.B. Constructing the non-linear regression models on the basis of multivariate normalizing transformations. *Electronic modeling*. 2018. Vol. 40. No. 6. P. 101-110. DOI: 10.15407/emodel.40.06.101.
9. Bates D.M., Watts D. G. Nonlinear regression analysis and its applications. New York: John Wiley & Sons, 1988. 384 p.
10. Seber G.A.F., Wild C.J. Nonlinear regression. New York : John Wiley & Sons, 1989. 768 p.
11. Ryan T.P. Modern regression methods. 2nd Edition. New York : John Wiley & Sons, 2008. 672 p.
12. Drapper N.R., Smith H. Applied regression analysis. New York : John Wiley & Sons, 1998. 736 p.
13. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. – Pearson Prentice Hall, 2007. 800 p.
14. Chatterjee S., Price B. Regression analysis by example. New York: John Wiley & Son, 1977. 228 p.
15. Olkin I., Sampson A.R. Multivariate Analysis: Overview. *International encyclopedia of social & behavioral sciences* / N. J. Smelser, P. B. Baltes (eds.) 1st edn. Elsevier, Pergamon, 2001. P. 10240–10247.
16. Mardia K.V. Measures of multivariate skewness and kurtosis with applications. *Biometrika*. 1970. Vol. 57. P. 519–530. DOI: 10.1093/biomet/57.3.519.
17. Foss T., Stensrud E., Kitchenham B., Myrtveit I. A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on software engineering*. 2003. 11(29). P. 985–995.
18. 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*. ACM, New York, 2008. P. 51–60.

Prykhodko S.B., Prykhodko N.V., Smykodub T.G. FOUR-FACTOR NON-LINEAR REGRESSION MODEL TO ESTIMATE THE SIZE OF OPEN SOURCE JAVA-BASED APPLICATIONS

The goal of the work is the creation of the multiple non-linear regression model for estimating the size of open source Java-based applications based on the multivariate normalizing transformation. A four-factor non-linear regression model to estimate the size of open source Java-based applications is constructed on the basis of the Johnson five-variate normalizing transformation for S_B family of the non-Gaussian data set from 38 applications hosted on GitHub (https://github.com). The data set was obtained using the CK tool (https://github.com/mauricioaniche/ck). The model is built around the metrics (variables) of class diagram: number of classes (Classes), number of static methods (NOSM), a measure of the number of response abilities of classes (Lack of Cohesion of Methods, LCOM), number of unique method invocations in a classes (the Response for Class, RFC). Comparison of the constructed model with the linear model and non-linear regression model based on the Johnson univariate transformation has been performed. In comparison with other linear regression models both linear and non-linear models based on the univariate normalizing transformations, constructed model has larger values of multiple coefficient of determination and the percentage of prediction at the level of magnitude of relative error, which equals 0.25, smaller values of the mean magnitude of relative error and width of the prediction intervals of non-linear regression. This may be explained best multivariate normalization and the fact that there is no reason to reject the null hypothesis that the four-variate distribution for normalized data, which normalized by the Johnson five-variate transformation for S_B family, is the same as the four-variate normal distribution. The practical significance of obtained results is that the software realizing the constructed model is developed in the sci-language for Scilab. The experimental results allow to recommend the constructed model for use in practice. Prospects for further research may include the application of other multivariate normalizing transformations and data sets to construct the multiple non-linear regression model for estimating the size of open source Java-based applications.

Key words: *nonlinear regression model, prediction interval, software size estimation, Java application, normalizing transformation, non-Gaussian data.*