УДК 004.412:519.237.5
DOI https://doi.org/10.32838/2663-5941/2020.1-1/23

Приходько С.Б.
Національний університет кораблебудування імені адмірала Макарова

Приходько Н.В.
Національний університет кораблебудування імені адмірала Макарова

Фаріонова Т.А.
Національний університет кораблебудування імені адмірала Макарова

Ворона М.В.
Національний університет кораблебудування імені адмірала Макарова

ТРЬОХФАКТОРНА НЕЛІНІЙНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ PHP-ЗАСТОСУНКІВ ІЗ ВІДКРИТИМ КОДОМ

Метою роботи є створення множинної нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим кодом на основі багатовимірного нормалізуючого перетворення за змінними, що визначаються за діаграмою класів. Трьохфакторну нелінійну регресійну модель для оцінювання розміру PHP-застосунків із відкритим кодом побудовано на основі нормалізації чотиривимірного негаусівського набору даних: кількість строк коду (LOC); кількість класів (Classes); сума кількості класів, на які впливає даний клас (Afferent Coupling), і кількості класів, із яких даний клас отримує ефекти (Efferent Coupling), та кількість методів (Methods) із 44 застосунків, розташованих на сайті GitHub (https://github.com) за допомогою інструменту PhpMetrics (https://phpmetrics.org/). Нормалізація цього набору даних здійснена за допомогою і двох одновимірних перетворень: у вигляді десяткового логарифму та перетворення Джонсона для сімейства S_B. Використання чотиривимірного перетворення в порівнянні з одновимірними дозволяє врахувати кореляцію між змінними, що призводить до покращення нормалізації даних, яка пов'язана з виконанням статистичної гіпотези щодо відповідності їх розподілу чотиривимірному 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, в порівнянні з іншими регресійними моделями (як лінійними, так і нелінійними) має більше значення множинного коефіцієнту детермінації, менше значення середньої величини відносної похибки та менші ширини інтервалу передбачення нелінійної регресії. Цей результат може бути пояснений найкращою багатовимірною нормалізацією і тим, що немає підстав відкидати нульову гіпотезу про те, що чотиривимірний розподіл для нормалізованих даних, який нормалізується за допомогою чотиривимірного перетворення Джонсона для сімейства S_B, є таким самим, як і чотиривимірний нормальний розподіл.

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

Постановка проблеми. PHP – це популярна мова сценаріїв загального призначення з відкритим вихідним кодом, яка особливо добре підходе для розроблення веб-застосунків на стороні сервера. У наш час PHP використовується більш ніж у 80% всіх веб-сайтів, наприклад, таких як Tesla, Wikipedia, WordPress.com [1]. Хоча основне призначення цієї мови полягає в тому, щоб дозволити веб-розробникам швидко писати веб-сторінки, що генеруються динамічно, але за допомогою PHP роблять набагато більше, в тому числі різноманітні фреймворки, конвертори та інші застосунки.

Задача оцінювання розміру PHP-застосунків із 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 [2]. Це потребує відповідних моделей для оцінювання розміру ПЗ, включаючи PHP-застосунки з відкритим кодом.

Аналіз останніх досліджень і публікацій. Для оцінювання кількості строк коду інформаційних PHP-систем із відкритим кодом відомо лінійне регресійне рівняння в залежності від трьох метрик концептуальної моделі даних у вигляді діаграми класів [3; 4]. Це рівняння побудовано на основі методів множинного лінійного регресійного аналізу. Але, як відомо, під час побудови лінійних регресійних моделей необхідно виконання певних умов, зокрема, залишки (residuals) повинні бути розподілені за нормальним законом, що має місце лише в поодиноких випадках. А це веде до необхідності побудови нелінійних регресійних моделей для оцінювання кількості строк ПЗ та застосування відповідних методів множинного нелінійного регресійного аналізу [5].

Тому для оцінювання розміру інформаційних PHP-систем із відкритим кодом в [6] було запропоновано рівняння нелінійної регресії, а в [5] – нелінійна регресійна модель. Запропоновані нелінійні регресійні рівняння та модель побудовано за допомогою множинного нелінійного регресійного аналізу із застосуванням чотиривимірного перетворення Джонсона сім'ї S_B на основі таких же трьох метрик діаграми класів, що і в [3; 4]: загальна кількість класів, загальна кількість зв'язків та середня кількість атрибутів на клас. Але для PHP-застосунків із відкритим кодом, що не є інформаційними системами, наприклад, таких як різноманітні фреймворки та конвертори, регресійні моделі можуть залежати в тому числі від інших метрик.

Зазвичай для побудови нелінійних регресійних рівнянь та моделей використовують одновимірні нормалізуючі перетворення [7–11]. Але їх застосування для побудови рівнянь і моделей нелінійної регресії не завжди призводить до задовільних результатів прогнозування, насамперед за такими стандартними оцінками, як середня величина відносної похибки, відсоток передбачення, ширина довірчого інтервалу та інтервалу передбачення [5; 6]. Це призводить до необхідності використання багатовимірних нормалізуючих перетворень.

Формулювання цілей статті. Метою статті є побудова трьохфакторної моделі нелінійної регресії та рівнянь нижньої і верхньої границь її інтервалів передбачення для оцінювання розміру PHP-застосунків із відкритим кодом в залежності від кількості класів (Classes); суми середньої кількості класів, на які впливає даний клас (Average Afferent Coupling) і середньої кількості класів, з яких даний клас отримує ефекти (Average Efferent Coupling), та середньої кількості методів (Average Methods) на основі чотиривимірного нормалізуючого перетворення, що дозволе підвищити достовірність оцінювання залежної змінної нелінійної регресії в порівнянні з використанням одновимірних нормалізуючих перетворень.

Виклад основного матеріалу дослідження. Для досягнення мети статті, що сформульована вище, ми скористалися методами наведеними в [5]. Згідно з [5] спочатку виконується нормалізація багатовимірних негаусових даних за багатовимірним нормалізуючим перетворенням. Для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим кодом були зібрані дані з метрик 44 програм, розташованих на сайті GitHub (https://github.com): фактична кількість строк коду *LOC*; кількість класів *Classes*; сума кількості класів, на які впливає даний клас, і кількості класів, із яких даний клас отримує ефекти, *AEC* та кількість методів *Methods*. Ці дані були отримані за допомогою інструменту PhpMetrics (https://phpmetrics.org/) та наведені в табл. 1.

Дані, що наведені в табл. 1, мають негаусівський розподіл, оскільки для трьох застосунків (1, 2 та 43) значення квадрату відстані Махаланобіса MD^2, які, відповідно, дорівнюють 33,67, 34,32 та 27,13, є більшими ніж величина квантіля розподілу χ^2, що становить 14,86 для рівня значущості 0,005. Також про негаусівський розподіл чотиривимірних даних із табл. 1 свідчить оцінка багатовимірного ексцесу β_2, яка визначалася за [12]. Відомо, що для m-вимірного нормального розподілу $\beta_2 = m(m+2)$. У нашому випадку $\beta_2 = 24$. Для чотиривимірних даних з табл. 1 оцінка β_2 дорівнює 83,11, що майже в 3,5 рази перевищує теоретичне значення.

Також майбутні фактори (*Classes*, *AEC* та *Methods*) були перевірені на наявність мультиколінеарності. Наявність мультиколінеарності свідчить про те, що в множинній регресійній моделі два або більше факторів пов'язані між собою або мають високий ступінь кореляції [13].

Наявність мультиколінеарності будемо визначати за коефіцієнтами впливу дисперсії (VIFs) серед майбутніх предикторів (факторів) у моделі множинної лінійної регресії. Для лінійної моделі множинної регресії з k-предикторами X_i, $i = 1, 2, \ldots, k$, VIFs – це діагональні елементи оберненої коваріаційної матриці $k \times k$ k-предикторів [13]. Значення VIFs більше за 10 часто сприймаються як сигнал, що дані мають проблеми з мультиколінеарністю. У разі, якщо значення VIFs знаходяться в межах від 1 до 5, то мультиколінеарності немає.

Коваріаційная матриця для трьох факторів (*Classes*, *AEC* та *Methods*) за даними табл. 1 має вигляд

$$\begin{pmatrix} 1,00000 & 0,96669 & 0,91919 \\ 0,96669 & 1,00000 & 0,96698 \\ 0,91919 & 0,96698 & 1,00000 \end{pmatrix}, \quad (1)$$

а обернена матриця до (1) така

$$\begin{pmatrix} 16,1865 & -19,4018 & 3,8827 \\ -19,4018 & 38,6533 & -19,5431 \\ 3,8827 & -19,5431 & 16,3289 \end{pmatrix}. \quad (2)$$

Елементи на головній діагоналі оберненої матриці (2) – значення VIFs, більше за 10. Це вказує на те, що дані мають проблеми з мультиколінеарністю. Для подолання проблеми мультиколінеарності дані з табл. 1 були перетворені в такі: фактичний розмір PHP-застосунків із відкритим кодом у тисячах рядків коду *Y*, загальна кількість класів X_1, сума середньої кількості класів, на які впливає даний клас (Average Afferent Coupling), і середньої кількості класів, з яких даний клас отримує ефекти (Average Efferent Coupling), X_2 та середня кількість методів на клас X_3. Ці дані наведені в табл. 2.

Коваріаційная матриця для факторів X_1, X_2 та X_3, за даними табл. 2, має такий вигляд:

$$\begin{pmatrix} 1,00000 & 0,12465 & -0,22070 \\ 0,12465 & 1,00000 & -0,00293 \\ -0,22070 & -0,00293 & 1,00000 \end{pmatrix}, \quad (3)$$

а обернена матриця до (3) така;

$$\begin{pmatrix} 1,068 & -0,133 & 0,235 \\ -0,133 & 1,016 & -0,026 \\ 0,235 & -0,026 & 1,052 \end{pmatrix}. \quad (4)$$

Елементи на головній діагоналі оберненої матриці (4) – значення VIFs, менше за 5. Це вказує на відсутність мультиколінеарності факторів X_1, X_2 та X_3.

Зазначимо, що розподіл даних із табл. 2 також є негаусівським тому, що для трьох застосунків (1, 2 та 5) значення MD^2, які, відповідно, дорівнюють 32,57, 17,75 та 26,38, є більшими ніж величина квантіля розподілу χ^2, що становить 14,86 для рівня значущості 0,005. Про негаусівський розподіл даних із табл. 2 свідчить оцінка багатовимірного ексцесу β_2, що дорівнює 59,21. Це значення більш ніж удвічі перевищує теоретичне, що в нашому випадку дорівнює 24.

У подальшому чотиривимірні негаусові дані, що наведені в табл. 2, використовуються для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим

Таблиця 1

Дані з метрик PHP-застосунків із відкритим кодом

№	*LOC*	*Classes*	*AEC*	*Methods*	№	*LOC*	*Classes*	*AEC*	*Methods*
1	174927	2075	13332	9979	23	10044	314	1332	699
2	112048	445	1359	1153	24	15477	280	1062	952
3	82551	411	2613	3229	25	15595	115	586	916
4	12022	132	667	767	26	2323	15	74	151
5	5347	5	12	163	27	7101	25	86	235
6	601	25	44	27	28	1431	22	57	75
7	1561	25	101	56	29	37081	278	1219	1637
8	33276	216	2182	1657	30	32826	235	1511	1925
9	36028	126	326	1367	31	12219	58	531	776
10	100245	448	2829	4287	32	59618	568	3570	3393
11	4458	73	381	325	33	24864	363	4165	2031
12	2988	18	126	173	34	2362	28	175	150
13	4047	31	166	196	35	381	8	22	25
14	6688	125	477	353	36	4308	52	256	339
15	1247	2	4	36	37	3412	52	398	173
16	5966	74	359	457	38	15785	126	528	1130
17	38996	269	1720	2067	39	535	3	10	28
18	3269	37	183	189	40	31676	76	568	1072
19	35548	335	3717	1949	41	13940	251	848	856
20	8910	117	437	543	42	3334	16	141	129
21	14019	209	1821	948	43	24298	794	1780	387
22	2920	19	109	155	44	42941	282	1104	810

кодом. Але спочатку для остаточного обґрунтування необхідності її побудови отримуємо лінійну регресійну модель для оцінювання розміру PHP-застосунків із відкритим кодом у вигляді

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \hat{b}_3 X_3 + \varepsilon, \qquad (5)$$

де оцінки параметрів такі:

$\hat{b}_0 = -3{,}1977$, $\hat{b}_1 = 0{,}0886$, $\hat{b}_2 = 0{,}7045$, $\hat{b}_3 = 0{,}6840$;

ε – випадкова величина з розподілом Гаусу, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Сума квадратів відхилень для моделі (5) склала 14465,3.

Перевірку нульової гіпотези про нормальність закону розподілу випадкової величини ε для моделі (5) здійснюємо за критерієм Пірсона. Для вибірки значень випадкової величини ε значення χ^2, яке дорівнює 52,04, 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 про відсутність теоретичного обґрунтування використання моделі лінійної регресії (5) і призводить до необхідності побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим кодом.

Спочатку для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим кодом негаусівські дані з табл. 2 ми нормалізуємо за одновимірним перетворенням у формі десяткового логарифму. Далі для нормалізованих даних будуємо лінійну регресійну модель [5]

$$Z_Y = \hat{Z}_Y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3 + \varepsilon, \qquad (6)$$

де параметри моделі (6) оцінювалися методом найменших квадратів, та їх оцінки є такими:

$\hat{b}_0 = -1{,}5110$, $\hat{b}_1 = 1{,}0610$, $\hat{b}_2 = -0{,}4459$, $\hat{b}_3 = 1{,}0151$.

Сума квадратів відхилень для моделі (6) склала 14,867.

Після чого за (6) та перетворенням у вигляді десяткового логарифму будуємо нелінійну регресійну модель для оцінювання розміру PHP-застосунків із відкритим кодом

$$Y = 10^{\varepsilon + \hat{b}_0} X_1^{\hat{b}_1} X_2^{\hat{b}_2} X_3^{\hat{b}_3}. \qquad (7)$$

Побудована модель (7) була перевірена за множинним коефіцієнтом детермінації R^2, середньою величиною відносної помилки MMRE і відсотком

Таблиця 2

Перетворені дані з метрик PHP-застосунків із відкритим кодом

№	Y	X_1	X_2	X_3	№	Y	X_1	X_2	X_3
1	174,927	2075	6,425	4,809	23	10,044	314	4,242	2,226
2	112,048	445	3,054	2,591	24	15,477	280	3,793	3,400
3	82,551	411	6,358	7,856	25	15,595	115	5,096	7,965
4	12,022	132	5,053	5,811	26	2,323	15	4,933	10,067
5	5,347	5	2,400	32,600	27	7,101	25	3,440	9,400
6	0,601	25	1,760	1,080	28	1,431	22	2,591	3,409
7	1,561	25	4,040	2,240	29	37,081	278	4,385	5,888
8	33,276	216	10,102	7,671	30	32,826	235	6,430	8,191
9	36,028	126	2,587	10,849	31	12,219	58	9,155	13,379
10	100,245	448	6,315	9,569	32	59,618	568	6,285	5,974
11	4,458	73	5,219	4,452	33	24,864	363	11,474	5,595
12	2,988	18	7,000	9,611	34	2,362	28	6,250	5,357
13	4,047	31	5,355	6,323	35	0,381	8	2,750	3,125
14	6,688	125	3,816	2,824	36	4,308	52	4,923	6,519
15	1,247	2	2,000	18,000	37	3,412	52	7,654	3,327
16	5,966	74	4,851	6,176	38	15,785	126	4,190	8,968
17	38,996	269	6,394	7,684	39	0,535	3	3,333	9,333
18	3,269	37	4,946	5,108	40	31,676	76	7,474	14,105
19	35,548	335	11,096	5,818	41	13,940	251	3,378	3,410
20	8,910	117	3,735	4,641	42	3,334	16	8,813	8,063
21	14,019	209	8,713	4,536	43	24,298	794	2,242	0,487
22	2,920	19	5,737	8,158	44	42,941	282	3,915	2,872

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 [14; 15]. 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 (7), що була побудована за даними з табл. 2, лише значення R^2, яке дорівнює 0,802, є задовільним. Значення двох інших показників – MMRE і PRED(0,25), що дорівнюють 0,284 і 0,500 відповідно, вказують на незадовільну якість моделі (7) з оцінками параметрів, що були отримані за даними з табл. 2.

Зважаючи на це, в подальшому для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків із відкритим кодом було застосовано метод покращення нелінійних регресійних моделей на основі нормалізуючих перетворень із застосуванням квадрату відстані Махаланобіса та інтервалів передбачення [16]. Суть цього методу [16] полягає в такому. Спочатку на першому етапі, як це зазвичай робиться, початкові негаусівські дані перевіряються на наявність викидів і, якщо останні знайдено, то вони відкидаються. Для цього використовується квадрат відстані Махаланобіса для нормалізованих даних. На першому етапі рівень значущості дорівнює 0,005. Далі на другому етапі будується нелінійна регресійна модель із застосуванням відповідного методу на основі нормалізуючих перетворень [5]. Після цього на третьому етапі для рівня значущості, що дорівнює 0,05, визначаються границі інтервалу передбачення нелінійної регресії за методом, наведеним в [5]. І на завершення, на четвертому етапі перевіряють, чи є серед даних, за якими будувалася нелінійна регресійна модель, такі, що виходять за визначені границі інтервалу передбачення. Та якщо останні знайдено, вони відкидаються, і ми повторюємо знову всі етапи, починаючи з першого, для нових даних. Якщо таких викидів не було, то повторення етапів завершується, відповідна нелінійна регресійна модель побудована.

Для визначення нижньої і верхньої границь інтервалів передбачення нелінійних регресій побудовано відповідні рівняння за [5]

$$\hat{Y}_{PI} = \psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{ 1 + \frac{1}{N} + \left(\mathbf{z}_X^+\right)^T \left[\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+ \right]^{-1} \left(\mathbf{z}_X^+\right) \right\}^{1/2} \right). \quad (8)$$

У (8) знаки мінус та плюс відповідають нижній та верхній границям інтервалів передбачення нелінійних регресій; ψ_Y – перша компонента нормалізуючого перетворення $\mathbf{T} = \psi(\mathbf{P})$ негаусівського випадкового вектору $\mathbf{P} = \{Y, X_1, X_2, \ldots, X_k\}^T$ в гаусівський випадковий вектор $\mathbf{T} = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$; $t_{\alpha/2,\nu}$ – квантіль *t*-розподілу Стьюдента з кількістю ступенів вільності ν та рівнем значущості $\alpha/2$; $\mathbf{Z}_X^+$ – матриця центрованих регресорів, яка містить значення $Z_{1_i} - \bar{Z}_1$, $Z_{2_i} - \bar{Z}_2$, $Z_{3_i} - \bar{Z}_3$; $\mathbf{z}_X^+ = \left\{ Z_{1_i} - \bar{Z}_1, Z_{2_i} - \bar{Z}_2, Z_{3_i} - \bar{Z}_3 \right\}^T$;

$S_{Z_Y}^2 = \frac{1}{\nu} \sum_{i=1}^{N} \left(Z_{Y_i} - \hat{Z}_{Y_i} \right)^2$, $\nu = N - k - 1$; $\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+$ – $k \times k$ матриця

$$\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+ = \begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \ldots & S_{Z_1Z_k} \\ S_{Z_1Z_2} & S_{Z_2Z_2} & \ldots & S_{Z_2Z_k} \\ \ldots & \ldots & \ldots & \ldots \\ S_{Z_1Z_k} & S_{Z_2Z_k} & \ldots & S_{Z_kZ_k} \end{pmatrix},$$

де $S_{Z_qZ_r} = \sum_{i=1}^{N} \left[Z_{q_i} - \bar{Z}_q \right] \left[Z_{r_i} - \bar{Z}_r \right]$, $q, r = 1, 2, \ldots, k$. В нашому випадку, що розглядається, k=3, а $\psi = \{\lg Y, \lg X_1, \lg X_2, \lg X_3\}^T$.

Нижні і верхні границі інтервалів передбачення нелінійних регресій для моделі (7) наведені в табл. 3. Для моделі (7) з оцінками параметрів, що були отримані за даними з табл. 2, з 44 PHP-застосунків виявилося: значення *Y* для одного застосунку 2 виходить за визначені межі інтервалу передбачення, що визначалися за (8). У табл. 3 ліва границя (межа) інтервалу передбачення на першій ітерації позначена як LB_1, а права – як UB_1.

Усього було 4 таких ітерації, після яких залишилося 40 застосунків (1, 3-6, 8-42). На четвертій ітерації викидів не було, повторення етапів завершується, нелінійна регресійна модель остаточно побудована за даними з 40 застосунків. У табл. 3 ліва границя інтервалу передбачення на четвертій ітерації позначена як LB_4, а права – як UB_4. Остаточно на четвертій ітерації для даних із 40 застосунків оцінки параметрів такі:

$$\hat{b}_0 = -1{,}7565,\ \hat{b}_1 = 1{,}0009,\ \hat{b}_2 = -0{,}2121,\ \hat{b}_3 = 1{,}2170.$$

Сума квадратів відхилень для моделі (7) у цьому випадку склала 0,326, що майже у 3,5 рази менше за відповідну суму на першій ітерації. На четвертій ітерації для нормалізованих за перетворенням у вигляді десяткового логарифму даних з 40 застосунків матриця $\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+$ для визначення нижньої і верхньої границь інтервалів передбачення нелінійної регресії за (8) є такою:

$$\left(\mathbf{Z}_X^+\right)^T \mathbf{Z}_X^+ = \begin{pmatrix} 17,116 & 2,419 & -1,954 \\ 2,419 & 1,539 & 0,267 \\ -1,954 & 0,267 & 2,655 \end{pmatrix}.$$

Після четвертої ітерації модель (7) виявилася кращою, ніж була після першої ітерації за трьома показниками: R^2, MMRE і PRED(0,25), значення яких для моделі (7) після четвертої ітерації складають 0,972, 0,172 і 0,775 відповідно, що краще за ці показники після першої ітерації, відповідно, на 17,5%, 65,2% і 35,5%.

Спробуємо побудувати кращу за (7) нелінійну регресійну модель для оцінювання розміру PHP-застосунків із відкритим кодом. Для цього спочатку дані 40 застосунків (1, 3-6, 8-42) із табл. 2 ми нормалізуємо за чотиривимірним перетворенням Джонсона сімейства S_B із такими оцінками параметрів:

$\hat{\gamma}_Y = 3,22402$, $\hat{\gamma}_1 = 3,0914$, $\hat{\gamma}_2 = 0,741344$, $\hat{\gamma}_3 = 18,3888$,

$\hat{\eta}_Y = 0,673845$, $\hat{\eta}_1 = 0,652695$, $\hat{\eta}_2 = 0,880739$, $\hat{\eta}_3 = 2,0838$

$\hat{\varphi}_Y = 0,054249$, $\hat{\varphi}_1 = -0,023277$, $\hat{\varphi}_2 = 1,52687$, $\hat{\varphi}_3 = -1,25874$,

$\hat{\lambda}_Y = 1057,484$, $\hat{\lambda}_1 = 8737,832$, $\hat{\lambda}_2 = 11,7836$ і $\hat{\lambda}_3 = 52797,72$.

Далі для нормалізованих даних будуємо лінійну модель (6) із такими оцінками параметрів:

$\hat{b}_0 = 0$, $\hat{b}_1 = 1,05746$, $\hat{b}_2 = -0,0428158$, $\hat{b}_3 = 0,504146$.

У цьому разі сума квадратів відхилень для модель (6) склала 0,7258.

Потім будуємо нелінійну регресійну модель [5]

$$Y = \hat{\varphi}_Y + \hat{\lambda}_Y \left[1 + e^{-\left(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y\right)/\hat{\eta}_Y}\right]^{-1}, \qquad (9)$$

де $Z_j = \gamma_j + \eta_j \ln \dfrac{X_j - \phi_j}{\phi_j + \lambda_j - X_j}$, $\phi_j < X_j < \phi_j + \lambda_j$, $j = 1, 2, 3$.

Зазначимо, що перед побудовою моделі (9) дані 40 застосунків (1, 3-6, 8-42) із табл. 2 було перевірено на наявність викидів за допомогою методу на основі багатовимірних нормалізуючих перетворень і квадрату відстані Махаланобіса MD^2 [17]. З'ясовано, що немає викидів у цих даних для рівня значущості 0,005 та чотиривимірного перетворення Джонсона сімейства S_B тому, що всі значення MD^2 менше, ніж величина квантіля розподілу χ^2, яка становить 14,86.

Модель (9) виявилася кращою за модель (7) за двома показниками: R^2 і MMRE. Значення R^2 і MMRE для моделі (9) складають 0,982 і 0,161 відповідно, що краще за ці показники для моделі (7), відповідно, на 1,0% і 7,1%. Для моделі (9) значення PRED(0,25) дорівнює 0,75, що на 3,3% гірше за цей показник для моделі (7). Але основна перевага моделі (9) в порівнянні з моделлю (7) полягає в менших ширинах інтервалу передбачення нелінійної

Таблиця 3

Межі інтервалів передбачення на першій і четвертій ітераціях

№	LB_1	UB_1	LB_4	UB_4	№	LB_1	UB_1	LB_4	UB_4
1	94,43	507,78	103,42	269,14	23	7,20	36,72	6,75	17,23
2	13,91	72,69	-	-	24	10,35	52,29	10,38	26,17
3	28,75	146,85	37,94	95,21	25	8,44	42,00	11,39	28,13
4	7,14	35,32	8,93	22,01	26	1,24	6,28	1,97	4,93
5	1,62	9,67	3,04	8,43	27	2,34	11,83	3,27	8,16
6	0,33	1,89	0,26	0,72	28	0,82	4,22	0,88	2,24
7	0,50	2,62	-	-	29	16,81	84,96	19,57	48,96
8	11,47	59,28	17,49	44,17	30	16,61	83,65	22,86	56,86
9	16,33	90,32	20,32	53,54	31	5,21	27,04	9,39	23,91
10	38,30	198,74	52,39	132,78	32	30,85	157,51	37,64	94,58
11	2,86	14,20	3,54	8,75	33	13,49	71,32	19,38	49,53
12	1,22	6,28	2,07	5,23	34	1,14	5,80	1,62	4,08
13	1,62	8,11	2,29	5,67	35	0,25	1,31	0,28	0,73
14	3,65	18,30	3,69	9,29	36	3,02	14,94	4,07	10,02
15	0,37	2,14	0,62	1,68	37	1,21	6,43	1,60	4,11
16	4,19	20,68	5,44	13,39	38	11,34	57,48	14,94	37,34
17	18,00	90,73	24,23	60,30	39	0,24	1,29	0,38	0,99
18	1,63	8,15	2,14	5,31	40	8,08	41,32	13,76	34,72
19	13,12	68,96	18,92	48,19	41	9,71	49,33	9,54	24,19
20	5,70	28,40	6,40	15,89	42	0,79	4,27	1,39	3,62
21	6,96	35,73	9,23	23,22	43	5,01	30,54	-	-
22	1,20	6,10	1,87	4,69	44	8,66	43,79	-	-

регресії розміру PHP-застосунків із відкритим кодом для більшої кількості даних. Межі інтервалів передбачення нелінійних регресій розміру 40 PHP-застосунків із відкритим кодом (1, 3-6, 8-42) із табл. 2 для моделі (9) наведені в табл. 4 для двох перетворень Джонсона сімейства S_B: одновимірного і чотиривимірного. Дані табл. 3 і табл. 4 вказують на те, що модель (9) із відповідними параметрами для чотиривимірного перетворення Джонсона сімейства S_B у порівнянні з моделлю (7) має менші ширини інтервалу передбачення для 36 PHP-застосунків (1-3, 6-28, 30-37, 39 і 40). Також з табл. 4 можна побачити, що ширини довірчого інтервалу нелінійної регресії на основі чотиривимірного перетворення Джонсона сім'ї S_B менші, ніж для одновимірного перетворення Джонсона для 34 з 40 рядків даних (2-4, 6, 7, 9-12, 14-32, 34-36, 38-40).

Кращі показники оцінювання розміру PHP-застосунків із відкритим кодом за моделлю нелінійної регресії на основі чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B можна, в першу чергу, пояснити кращою нормалізацією, яка перевірялася за відомими критеріями [18]. Так, якщо за критерієм на основі квадрата відстані Махаланобіса гіпотеза про нормальність багатовимірного закону розподілу нормалізованих за допомогою чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B даних для 40 застосунків із табл.2 приймається для рівня значущості 0,025, то у випадку застосування одновимірного перетворення та без нього – відкидається.

Висновки. Удосконалено трьохфакторну модель нелінійної регресії та рівняння нижньої і верхньої границь її інтервалу передбачення для оцінювання розміру PHP-застосунків із відкритим кодом у залежності від загальної кількості класів, суми середньої кількості класів, на які впливає даний клас, і середньої кількості класів, з яких даний клас отримує ефекти, та середньої кількості методів на клас на основі чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B, що дозволяє підвищити достовірність оцінювання залежної змінної нелінійної регресії в порівнянні з використанням одновимірних нормалізуючих перетворень. Модель, що побудовано, в порівнянні з іншими регресійними моделями має більші значення множинного коефіцієнту детермінації, менші середні величини відносної похибки та ширини інтервалу передбачення нелінійної регресії. На прикладі вдосконалення трьохфакторної нелінійної регресійної моделі підтверджено працездатність методу покращення нелінійних регресійних моделей на основі багатовимірних нормалізуючих перетворень із застосуванням квадрату відстані Махаланобіса та інтервалів передбачення. У майбутньому планується використання інших наборів даних для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків.

Таблиця 4

Межі інтервалів передбачення нелінійних регресій

№	Одновимірне		Чотиривимірне		№	Одновимірне		Чотиривимірне	
	LB	UB	LB	UB		LB	UB	LB	UB
1	137,623	188,547	125,591	268,898	23	5,830	19,347	7,273	17,605
3	40,093	97,026	40,135	92,247	24	9,000	28,760	10,261	24,516
4	7,806	24,536	8,764	20,551	25	10,601	32,674	11,429	26,788
5	3,030	11,250	3,657	9,684	26	1,688	5,224	1,992	4,719
6	0,413	0,830	0,368	0,898	27	2,735	8,764	3,160	7,485
8	17,186	51,333	18,567	44,132	28	0,744	1,967	0,828	1,951
9	20,028	59,892	20,119	49,065	29	18,694	54,189	19,375	45,476
10	55,794	120,683	56,409	128,228	30	23,274	64,438	23,916	55,612
11	2,794	8,883	3,433	8,092	31	9,539	30,801	10,531	25,436
12	1,789	5,636	2,145	5,129	32	40,258	97,406	39,786	91,584
13	1,848	5,710	2,225	5,233	33	18,430	56,066	20,162	49,046
14	2,855	9,281	3,648	8,728	34	1,319	3,955	1,599	3,783
15	0,580	1,467	0,676	1,672	35	0,399	0,725	0,299	0,667
16	4,535	14,461	5,273	12,370	36	3,342	10,629	3,943	9,254
17	24,684	67,512	25,280	58,707	37	1,281	3,935	1,642	3,967
18	1,689	5,171	2,055	4,832	38	14,340	43,233	14,954	35,234
19	18,154	54,792	19,803	47,763	39	0,451	0,916	0,402	0,920
20	5,184	16,687	6,050	14,301	40	14,665	44,649	15,478	36,875
21	8,191	26,272	9,563	22,788	41	8,092	26,159	9,299	22,307
22	1,575	4,823	1,875	4,432	42	1,201	3,689	1,446	3,519

Список літератури:

1. Hayden J. 80% of the web powered by PHP. URL : https://haydenjames.io/80-percent-web-powered-by-php/ (дата звернення: 19.09.2019).
2. 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.
3. 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.
4. 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.
5. 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
6. Prykhodko S.B., Prykhodko N.V., Smykodub T.G., Spinov A.V. Constructing the non-linear regression model to estimate the software size of open source PHP-based information systems. *Problems of information technologies*. 2018. № 1(023). P. 118–125.
7. Bates D.M., Watts D. G. Nonlinear regression analysis and its applications. New York : John Wiley & Sons, 1988. 384 p.
8. Seber G.A.F., Wild C.J. Nonlinear regression. New York : John Wiley & Sons, 1989. 768 p.
9. Ryan T.P. Modern regression methods. 2nd Edition. New York : John Wiley & Sons, 2008. 672 p.
10. Drapper N.R., Smith H. Applied regression analysis. New York : John Wiley & Sons, 1998. 736 p.
11. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. Pearson Prentice Hall, 2007. 800 p.
12. 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
13. Chatterjee S., Price B. Regression analysis by example. New York : John Wiley & Son, 1977. 228 p.
14. 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.
15. 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.
16. Приходько С.Б., Приходько Н.В. Метод покращення нелінійних регресійних моделей на основі багатовимірних нормалізуючих перетворень. *Прикладні науково-технічні дослідження: матеріали III міжнар. наук.-практ. конф*. (Івано-Франківськ, 3–5 квітня 2019 р.). Івано-Франківськ : Сімфоня Форте, 2019. С. 20.
17. Prykhodko S., Prykhodko S., Makarova L., Pugachenko K. Detecting Outliers in Multivariate Non-Gaussian Data on the basis of Normalizing Transformations. *Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) «Celebrating 25 Years of IEEE Ukraine Section»*. (May 29 – June 2, 2017, Kyiv, Ukraine). Kyiv, 2017. P. 846–849. DOI: 10.1109/UKRCON.2017.8100366.
18. 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.

Prykhodko S.B., Prykhodko N.V., Farionova T.A., Vorona M.V. THREE-FACTOR NON-LINEAR REGRESSION MODEL TO ESTIMATE THE SIZE OF OPEN SOURCE PHP-BASED APPLICATIONS

The goal of the work is the creation of the multiple non-linear regression model for estimating the size of open source PHP-based applications on the basis of the multivariate normalizing transformation. A three-factor non-linear regression model to estimate the size of open source PHP-based applications is constructed on the basis of the Johnson four-variate normalizing transformation for S_B family of the non-Gaussian data set from 44 applications hosted on GitHub (https://github.com). The data set was obtained using the PhpMetrics tool (https://phpmetrics.org/). The model is built around the metrics (variables) of class diagram: number of classes, sum of average afferent coupling and average efferent coupling, average number of methods. Comparison of the constructed model with the linear model and non-linear regression models based on the decimal logarithm and the Johnson univariate transformation has been performed. In comparison with other linear regression models and non-linear regression models based on the univariate normalizing transformations, constructed model has a larger multiple coefficient of determination, a smaller value of the mean magnitude of relative error and smaller widths 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 four-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 PHP-based applications.

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