DOI: 10.36910/6775-2524-0560-2019-37-11
УДК: 004.891
Кулаковська І. В.
Чорноморський національний університет імені Петра Могили

НЕЛІНІЙНЕ МНОЖИННЕ РЕГРЕСІЙНЕ РІВНЯННЯ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ З ВІДКРИТИМ КОДОМ ТИПУ MP3PLAYERS НА JAVA

Кулаковська І.В. Нелінійне множинне регресійне рівняння для оцінювання розміру програмного забезпечення з відкритим кодом типу mp3players на JAVA. У статті розглядаються різні підходи до оцінки трудомісткості розробки програмного забезпечення (ПЗ). Аналізується залежність оцінки трудомісткості розробки ПЗ від розміру проекту. Описуються основні види існуючих метрик і можливість їх застосування. Пропонуються показники для порівняння метрик. Нелінійна регресійна модель для оцінювання розміру програмного забезпечення з відкритим кодом типу mp3players на JAVA побудована на основі нормалізації шестивимірного негаусового набору даних (фактичний розмір програми в тисячах рядків коду, загальна кількість класів, загальна кількість зв'язків та інші) в концептуальній моделі даних з 32 програм за допомогою нелінійного регресійного рівняння. 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.

Кулаковская И. В. Нелинейное множественное регрессионное уравнение для оценки размера программного обеспечения с открытым кодом типа mp3players на JAVA. В статье рассматриваются различные подходы к оценке трудоемкости разработки программного обеспечения (ПО). Анализируется зависимость оценки трудоемкости разработки ПО от размера проекта. Описываются основные виды существующих метрик и возможность их применения. Предлагаются показатели для сравнения метрик. Нелинейная регрессионная модель для оценки размера программного обеспечения с открытым кодом типа mp3players на JAVA построена на основе нормализации шестимерного негауссового набора данных (фактический размер программы в тысячах строк кода, общее количество классов, общее количество связей и другие) в концептуальной модели данных из 32 программ с помощью нелинейного регрессионного уравнения. Модель, построенная, по сравнению с другими регрессионными моделями (как линейными, так и нелинейными), имеет больший множественный коэффициент детерминации, меньшее значение средней величины относительной погрешности. Для полученного уравнения исследовано нормальное распределения остатка, вычисленные ширины доверительного интервала для нелинейной регрессии.
Ключевые слова: нелинейная регрессионная модель, доверительный интервал, оценка размера программного обеспечения, нормализующее преобразование, управление проектами, оценка трудоемкости, разработка ПО, метод функциональных точек.

Kulakovska I.V. The non-linear regression equations for software size assessment with java mp3players type. The article discusses different approaches to assessing the complexity of software development (software). The dependence of the complexity of the software development on the size of the project is analyzed. The basic types of existing metrics and the possibility of their application are described. Metrics are offered to compare metrics. A nonlinear regression model for estimating the size of open source MP3players software on JAVA is based on the normalization of a six-dimensional non-Gaussian dataset (actual program size in thousands of lines of code, total classes, total links, etc.) in the conceptual data model of 32 programs using a nonlinear regression equation. 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.

Постановка наукової проблеми.

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 даній компанії. 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 COCOMO.

Задача оцінювання розміру програмного забезпечення (ПЗ) на ранній стадії його розробки є важливою, оскільки ця інформація використовується для прогнозування трудомісткості розробки ПЗ за допомогою таких моделей як COCOMO II. Це призводить до необхідності розробки відповідних моделей для оцінювання розміру ПЗ, включаючи ПЗ інформаційних систем з відкритим кодом для програм mp3 players.

Існують різні методи для оцінювання розміру програмного 'забезпечення, які використовуються сьогодні. Більшість з них походять від методу аналізу функціональних точок (FPA - Function Point Analysis). Інший підхід полягає в тому, щоб провести функціональне вимірювання, щоб «разити функціональність у кількості, що представляє розмір. Інші методи зазначення розміру програмного забезпечення включають оцінювання на генові варіантів використання (Use Case). Але історично найпоширенішою та найбільш вживаною методологією визначення розміру програмного забезпечення є підрахунок кількості рядків коду, написаних у вихідному коді програми. Крім того, всесвітньо відомою є модель COCOMO II - це модель регресії, заснована на кількість рядків коду (LOC).

Аналіз останніх досліджень і публікацій. Нормализуючі перетворення дуже часто є надійним способом побудови нелінійних регресійних моделей та рівнянь [3-4]. Однак добре відомі методи їх побудови, що засновані на одновимірних нормалізуючих перетвореннях (таких як логарифмічне і Бокса-Кокса), не враховують кореляції між випадковими змінними у разі нормалізації багатовимірних негаусових даних. Це призводить до необхідності використання багатовимірних нормализуючих перетворень [3-4], які враховують цю кореляцію, для побудови нелінійних регресійних моделей та рівнянь для оцінювання розміру ПЗ інформаційних систем, в тому числі з відкритим кодом на PHP. В [1-3] було побудовано нелінійне регресійне рівняння для оцінювання розміру ПЗ інформаційних систем з відкритим кодом на PHP на основі багатовимірного перетворення Джонсона для сімейства SB. Виникає потреба порівнянні моделей та у побудові відповідного рівняння для інших видів ПЗ інформаційних систем з відкритим кодом, наприклад для програм mp3 players на JAVA.

Мета статті – 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.

Існуючі моделі оцінювання розміру програмного забезпечення. Моделі оцінювання розміру ПЗ поділяються на п'ять категорій: *аналогові; регресійні; моделі на основі експертних оцінок; моделі, які базуються на функціональних точках; параметричні моделі.*

Існують різні методи для оцінювання розміру програмного забезпечення, які використовуються сьогодні. Більшість з них походять від методу аналізу функціональних точок (FPA - Function Point Analysis). Інший підхід полягає в тому, щоб провести функціональне вимірювання, щоб виразити функціональність у кількості, що представляє розмір. Інші методи визначення розміру програмного забезпечення включають оцінювання на основі варіантів використання (Use Case). Але історично найпоширенішою та найбільш вживаною методологією визначення розміру програмного забезпечення є

підрахунок кількості рядків коду, написаних у вихідному коді програми.

Серед методик підрахунку кількості рядків коду є дві основні:

- по числу фізичних рядків (LOC) - визначається як загальне число рядків вихідного коду, включаючи коментарі і порожні рядки;
- по числу логічних рядків коду (LLOC) - визначається як загальна кількість команд і залежить від використовуваної мови програмування. Якщо мова підтримує розміщення кількох команд в одному рядку, то один фізичний рядок повинен бути врахований як кілька логічних, якщо він містить більше однієї команди мови.

Також є похідні від основних методик, які в залежності від завдання можуть містити додаткову інформацію за такими показниками: число порожніх рядків; число рядків, що містять коментарі; відсоток коментарів (відношення рядків коду до рядків з коментарями, похідна метрика стилістики); середнє число рядків для функцій (класів, файлів); середня кількість рядків, що містять вихідний код для функцій (класів, файлів); середнє число рядків для модулів і т.д. Метрики, засновані на аналізі кількості рядків і синтаксичних елементів вихідного коду програми, були запропоновані багатьма відомими вченими, наприклад М. Холстедом в 1977р.

Аналіз функціональних точок (Function points) - це метод вимірювання розміру програмного забезпечення з точки зору користувачів системи. Метод був розроблений Аланом Альбрехтом ще в середині 1970-х років, вперше опублікований в 1979 році.

Метод UCP (Use Case Points) являє собою оцінку розміру проектів на основі діаграм UML (Unified Modeling Language) і методології RUP (Rational Unified Process). Як, і багато інших сучасні методів оцінки, UCP базується приблизно на тих же принципах, що і метод функціональних точок. Головна відмінність полягає в заміні одиниць вимірювання з функціональних точок на варіанти використання (Use Cases).

Крім того, всесвітньо відомою є модель COCOMO II - це модель регресії, заснована на кількість рядків коду (LOC). Ця процедурна модель оцінювання витрат для програмних проектів часто використовується для надійного прогнозування різних параметрів, пов'язаних з проектом, таких, як розмір, зусилля, витрати, час та якість, які необхідні для впровадження програмного забезпечення.

При використанні точної та надійної методології для оцінювання розміру можна зробити висновок про якість ПЗ та навіть усього програмного процесу і, при необхідності, вжити подальших заходів.

Перевірка адекватності математичної моделі для оцінювання розміру програмного забезпечення. Для перевірки адекватності лінійного рівняння регресії використаємо коефіцієнт детермінації:

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

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. Коефіцієнт детермінації R^2 приймає значення від 0 до 1, Чим ближче значення коефіцієнта за модулем до 1, тим тісніше зв'язок результативної ознаки з досліджуваними факторами. При значенні $R^2>0,5$ можна вважати, що дана модель є прийнятною. 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. Чим вагоміша частка пояснюваної варіації, тим менше роль інших факторів, а отже, модель регресії краще апроксимує вихідні дані і такою регресійної моделлю можна скористатися для прогнозу значень результативного показника. 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 Sv використовується для перевірки якості як нелінійного, так і лінійного рівняння регресії. Цей параметр також використовується для порівняння різних моделей.

Способи удосконалення математичної моделі для оцінювання розміру програмного забезпечення. 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 негаусових 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 50 програм на відкритим кодом програм mp3 players на базі мови програмування Java взятих з репозиторію GitHub.

Використання нормалізуючих перетворень дозволяє перейти до лінійної регресії для нормалізованих даних, для неї побудувати довірчий інтервал та інтервал прогнозування традиційним способом, і, нарешті, шляхом застосування відповідного зворотного перетворення перейти до нелінійної регресії. Тим самим ми отримаємо більш точну математичну модель для оцінки розміру mp3players реалізованих мовою Java.

Рівняння множинної регресії. 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,. . . , x_m). Найважливішими умовами побудови багатофакторної моделі зв'язку є достатня кількість одиниць у сукупності (як мінімум у 5 разів більше, ніж число факторів) та відсутність мультиколінеарності факторів (близького до функціонального зв'язку між ними). В тому випадку, якщо два факторних показники мультиколінеарні, один з них повинен бути виключений з моделі. Для побудови цього рівняння були використані дані всього з 32 інформаційних систем, кількість яких фактично межує з малою вибіркою. Рівняння регресії

$$Y=a_0+a_1x_1+a_2x_2+\ldots.+a_6x_6 \qquad (2)$$

Позначення У- 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} - кількіість відносин між класами.

Попередній аналіз даних виявив мультиколінеарність x_1 та x_7, тому для побудови рівняння обрані перші шість факторів. 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, данні яких наведені в таблиці 2. При дослідженні також розглядалися різні види рівняння регресії. Коефіцієнти детермінації для різних видів наведені в таблиці 1. За ними обрано експоненціальну регресію і надалі досліджувалося рівняння:

$$\ln Y=a_0+a_1x_1+a_2x_2+\ldots.+a_6x_6. \qquad (3)$$

Таблиця 1. Коефіцієнти детермінаціїї для різних рівняннь регресії

Показатели	Коэффициент детерминации
Лінійна регресія	0,697019
Параболічна регресія	0,636005
Експоненціальна регресія	**0.7146035**
Степенева регресія	0,626321
Гіперболічна регресія	0,110816
Логарифмічна регресія	0,615861
Показникова регресія	0,521657

Таблиця 2. Вибірка шести факторів

Name	LOC	У	NC	ANA	ANM	ANSM	ANGM	ANCM	Ninh
gabbiernacka_MP3Player	484	6.182	5	7.2	7.2	0.8	0.6	1	0,52
joshuba_JavaFX-MP3Player	484	6.182	8	2.75	3.5	1.25	0.13	1	1,74
pawelslupski_java-fx-repository	552	6.314	9	3	5.56	1.89	0.56	1	1,99
adrianDyj_mp3Player-javaFx	577	6.358	9	3	5.78	2.11	0.56	1	2,23
Kwarku_Mp3Player	653	6.482	9	3.11	5.78	2.11	0.56	1	2,33
cipek_simple-MP3Player-Android	723	6.583	10	3.3	5.5	1.1	0.9	1.1	2,33
szxv587_mp3Player	739	6.605	8	4.38	4.5	1.38	0.25	1	2,33
joseavel_Mp3Player	853	6.749	11	3.55	7.27	0.82	0.73	1	2,66
saquibhafiz_MP3Player	956	6.863	19	4.95	3.42	0.89	0.37	1.11	2,66

Name	LOC	У	NC	ANA	ANM	ANSM	ANGM	ANCM	Ninh
paczarny_mp3Player	972	6.879	10	4.3	5.7	2.1	0.4	1	2,66
kystudio_MP3Player	1041	6.948	22	1.91	3	0.41	0.23	1.05	2,66
ivartanian_MP3Player	1066	6.972	7	8	7.14	0.71	0.43	1	2,73
pandboy_Mp3Player	1104	7.007	25	2.68	2.6	0.44	0.32	1.04	2,73
dammyson_mp3player	1113	7.015	10	4.3	6.9	0.6	0.4	1	2,73
suryarajasekaran_Jamify	1221	7.107	9	4.44	3.33	0.67	0.44	1	2,73
dtgianggithub_mp3player	1242	7.124	12	6.33	4.67	1.08	0.33	1	2,73
tevop_Mp3Player	1278	7.153	11	5.27	7.18	1	0.64	1.09	2,73
jxnu707_MP3player	1299	7.169	23	3.65	2.87	0.43	0.35	1.04	2,73
kystudio_MP3Player_Tab	1365	7.219	27	2.04	3.07	0.59	0.19	1.04	2,79
XFHNever_MP3Player	1435	7.269	35	2.69	2.26	0.26	0.17	1.06	2,79
vbilenko78_MP3Player	1468	7.292	12	7.33	6.67	0.5	0.42	1	2,79
carloscj6_SampleMusicPlayer	1482	7.301	19	5.74	5.47	0.79	0.26	1	2,85
CreativityCz_Mp3Player	1596	7.375	21	2.1	3.67	0.67	0.24	1.05	2,85
Inzulus_MP3Player	1608	7.383	20	4.2	5.7	2.7	0.25	1.1	2,96
david-loman_MP3player	1682	7.428	24	9.17	3.38	0.46	0.33	1	3,63
karolminuth_Mp3Player-ProjectDoneWithJavaStart	1704	7.441	27	3	5.56	1.89	0.56	1	3,80
StanislawRusnak_Mp3Player	1722	7.451	27	3	5.78	2.11	0.56	1	4,72
polewa_Mp3Player	1779	7.484	27	3	5.78	2.11	0.56	1	5,21
Azzello_MP3Player	2223	7.707	32	3.44	3.25	0.34	0.03	1	0,52
ProjectMusicPlayer_MP3Player_frontend	2704	7.902	30	3.2	3.37	0.07	0	1	1,74
nicoliuli_Mp3Player	2923	7.98	66	2.3	3.45	0.79	0.82	1.12	1,99
yuanwofei_mp3player	6152	8.725	117	6.33	3.34	0.66	0.37	1.02	2,23

Далі за даними будуємо лінійне регресійне рівняння

$$\mathbf{Y = 5.3318 + 0.022X1 + 0.01X2 + 0.08061X3-0.07548X4-0.7223X5 + 1.2149X6} \quad (4)$$

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.*

Аналіз мультиколінеарності. $det(X^TX)=788527187,3 \geqq 0$. Вид мультиколінеарності, при якому факторні змінні пов'язані деякою стохастичною залежністю, називається частковою. Якщо між факторними змінними є високий ступінь кореляції, то матриця (X^TX) близька до виродженої, тобто, чим ближче до 0 визначник матриці межфакторной кореляції, тим сильніше мультиколінеарності факторів і ненадійніше результати множинної регресії).

Таблиця 3. Матриця парних коефіцієнтів кореляції R:

с Y	LOC	NC	ANA	ANM	ANSM	ANGM	ANCM
LOC	1	0,808853774	0,071438	-0,3458135	-0,28553	-0,2193	0,154278
NC	0,808853774	1	-0,03447	-0,4325747	-0,2136	-0,06362	0,219349
ANA	0,071438363	-0,03447114	1	0,33820412	-0,18537	0,030388	-0,24698
ANM	-0,34581353	-0,43257466	0,338204	1	0,448041	0,557441	-0,24972
ANSM	-0,28552624	-0,21360163	-0,18537	0,4480408	1	0,351865	-0,07159
ANGM	-0,21930037	-0,06362117	0,030388	0,55744137	0,351865	1	0,214305
ANCM	0,154278212	0,219348944	-0,24698	-0,2497173	-0,07159	0,214305	1

Таблиця 4. Матриця межфакторної кореляції R_{11}

	NC	ANA	ANM	ANSM	ANGM	ANCM
NC	1	-0,03447	-0,4325747	-0,2136	-0,06362	0,219349
ANA	-0,03447114	1	0,33820412	-0,18537	0,030388	-0,24698
ANM	-0,43257466	0,338204	1	0,448041	0,557441	-0,24972
ANSM	-0,21360163	-0,18537	0,4480408	1	0,351865	-0,07159
ANGM	-0,06362117	0,030388	0,55744137	0,351865	1	0,214305
ANCM	0,219348944	-0,24698	-0,2497173	-0,07159	0,214305	1

Для перевірки рівня значущості використовують F-критерій Фішера.

При цьому обчислюють фактичну (що спостерігається) значення F-критерію, через коефіцієнт детермінації R2, розрахований за даними безпосереднього спостереження. За таблицями розподілу Фішера-Снедоккора знаходять критичне значення F-критерію (Fкр). Для цього задаються рівнем значущості α (зазвичай його беруть рівним 0,05) і двома числами ступенів свободи k1 = m і k2 = n-m-1.Маємо: k1 = 6 і k2 = n-m-1 = 32 - 6 - 1 = 25. Табличне значення при ступенях свободи, Fкр (6; 25) = 2.51

$$R^2 = 0.7146 = 1 - \frac{detR}{detR_{11}} \quad (5)$$

де *detR= 0.0615* – визначник матриці парної кореляції; $detR_{11} = 0.215$ - визначник матриці міжфакторної кореляції. Уточнене значення коефіцієнту детермінації $\overline{R^2}$.

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} \quad (6)$$

$$\overline{R}^2 = 1 - (1 - 0.7146)\frac{32-1}{32-6-1} = 0.646$$

$$F = \frac{R^2}{1-R^2}\cdot\frac{(n-m-1)}{m} = \frac{0.7146}{1-0.7146}\cdot\frac{32-6-1}{6} = 10.431 \quad (7)$$

F = 0.71461 - 0.7146 • 32-6-16 = 10.431>2.51

Оскільки фактичне значення F> Fкр, то коефіцієнт детермінації статистично значимий і рівняння регресії статистично надійно (тобто коефіцієнти спільно значимі).

Перевірка нормальності розподілу залишкової компоненти.

1. Перевірка гіпотез про вид розподілу по Пирсону.

Перевіримо гіпотезу про те, що X розподілено по нормальному закону з допомогою критерію згоди Пірсона.

Error!, (8)

де рі - ймовірність попадання в і-й інтервал випадкової величини, розподіленої по гіпотетичному(теоретичному) закону.

Таблиця 5. Розподіл ймовірностей для перевірки залишкової компоненти

$x_i \div x_{i+1}$		f_i	$x_1 = (x_i - x_{cp})/s$	$x_2 = (x_{i+1} - x_{cp})/s$	$\Phi(x_1)$	$\Phi(x_2)$	$p_i=\Phi(x_2)-\Phi(x_1)$	Ожидаемая частота, $32p_i$	Слагаемые статистики Пирсона, K_i
-0,418	-0,202	1	-2.3159	-1.5275	-0.4898	-0.437	0.0528	1.6896	0.2814
-0,202	0,014	8.	-1.5275	-0.7391	-0.437	-0.2703	0.1667	5.3344	1.332
0,014	0,230	7.	-0.7391	0.04927	-0.2703	0.0199	0.2902	9.2864	0.5629
0,230	0,446	10	0.04927	0.8377	0.0199	0.2995	0.2796	8.9472	0.1238
0,446	0,662	4	0.8377	1.626	0.2995	0.4484	0.1489	4.7648	0.1227
0,662	0,878	2	1.626	2.4144	0.4484	0.4922	0.0438	1.4016	0.2554
		32							2.6782

Для обчислення ймовірностей рі застосуємо формулу і таблицю функції Лапласа

Error!, де s = 2.6782, xcp = 0.217 (9)

Теоретична (очікувана) частота дорівнює fi = f_{pi}, де f = 32

Ймовірність влучення в і-й інтервал: pi = Ф (x_2) - Ф (x_1)

Визначимо кордон критичної області. 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: [Ккр; + ∞). Її кордон Ккр = χ2 (k-r-1; α) знаходимо за таблицями розподілу χ2 і заданим значенням s, k (число інтервалів), r = 2 (параметри xcp і s оцінені за вибіркою).

Ккр = χ2 (6-2-1; 0.05) = 7.81473; Кнабл = 2.68 (10)

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

Рис. 2. Діаграма розподілу частот для остач

Індивідуальні довірчі інтервали для Y при даному значенні вектора Z(x_1,…x_6). **(y±ε), де**

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

де $t_{крит}$ (n-m-1;α/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,…x_6). **(y±ε),**

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

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

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

Таблиця 6. Розрахункова таблиця довірчих інтервалів та прогнозованих

LOC	ln(Y)	NC	ANA	ANM	ANSM	ANGM	ANCM	Y_solv	$E^{(Y\_solv)}$	Довірчий		Прогнозований	
										лівий	правий	лівий	правий
484	6,18	5	7,2	7,2	0,8	0,6	1	6,82	911,80	-20,85	1844,45	-169,08	1992,68
484	6,18	8	2,75	3,5	1,25	0,13	1	6,84	938,28	69,25	1807,32	-88,21	1964,78
552	6,31	9	3	5,56	1,89	0,56	1	6,68	792,93	-181,04	1766,89	-323,80	1909,66
577	6,36	9	3	5,78	2,11	0,56	1	6,68	793,82	-188,71	1776,35	-330,39	1918,03
653	6,48	9	3,11	5,78	2,11	0,56	1	6,68	794,70	-187,46	1776,85	-329,18	1918,57
723	6,58	10	3,3	5,5	1,1	0,9	1,1	6,63	758,58	-238,18	1755,34	-378,09	1895,25
739	6,61	8	4,38	4,5	1,38	0,25	1	6,84	938,71	32,43	1844,99	-119,50	1996,92
853	6,75	11	3,55	7,27	0,82	0,73	1	6,82	917,01	-176,84	2010,86	-305,69	2139,71
956	6,86	19	4,95	3,42	0,89	0,37	1,11	7,09	1198,83	-9,47	2407,14	-127,24	2524,91
972	6,88	10	4,3	5,7	2,1	0,4	1	6,82	917,59	-92,86	1928,04	-231,10	2066,28
1041	6,95	22	1,91	3	0,41	0,23	1,05	7,16	1280,86	-8,56	2570,28	-119,52	2681,25
1066	6,97	7	8	7,14	0,71	0,43	1	6,99	1088,10	99,13	2077,06	-41,74	2217,93
1104	7,01	25	2,68	2,6	0,44	0,32	1,04	7,12	1233,15	-122,07	2588,37	-228,04	2694,35
1113	7,01	10	4,3	6,9	0,6	0,4	1	7,03	1132,03	76,02	2188,04	-56,93	2321,00
1221	7,11	9	4,44	3,33	0,67	0,44	1	6,69	803,64	-90,06	1697,34	-243,83	1851,10
1242	7,12	12	6,33	4,67	1,08	0,33	1	6,93	1023,12	-17,47	2063,71	-152,17	2198,41
1278	7,15	11	5,27	7,18	1	0,64	1,09	6,99	1087,72	1,06	2174,37	-128,55	2303,98
1299	7,17	23	3,65	2,87	0,43	0,35	1,04	7,08	1192,60	-113,42	2498,63	-223,09	2608,29
1365	7,22	27	2,04	3,07	0,59	0,19	1,04	7,28	1444,32	26,94	2861,70	-74,71	2963,35
1435	7,27	35	2,69	2,26	0,26	0,17	1,06	7,46	1730,67	145,49	3315,85	53,98	3407,36

LOC	ln(Y)	NC	ANA	ANM	ANSM	ANGM	ANCM	Y_solv	$E^{(Y\_solv)}$	Довірчий		Прогнозований	
										лівий	правий	лівий	правий
1468	7,29	12	7,33	6,67	0,5	0,42	1	7,08	1188,76	87,07	2290,46	-40,95	2418,48
1482	7,30	19	5,74	5,47	0,79	0,26	1	7,22	1360,60	104,01	2617,18	-9,61	2730,81
1596	7,38	21	2,1	3,67	0,67	0,24	1,05	7,16	1289,94	13,14	2566,75	-98,84	2678,72
1608	7,38	20	4,2	5,7	2,7	0,25	1,1	7,23	1373,73	81,71	2665,74	-29,05	2776,50
1682	7,43	24	9,17	3,38	0,46	0,33	1	7,17	1294,48	-49,46	2638,43	-156,26	2745,23
1704	7,44	27	3	5,56	1,89	0,56	1	7,07	1178,22	-281,12	2637,55	-380,03	2736,46
1722	7,45	27	3	5,78	2,11	0,56	1	7,07	1179,54	-284,27	2643,36	-382,90	2741,99
1779	7,48	27	3	5,78	2,11	0,56	1	7,07	1179,54	-284,27	2643,36	-382,90	2741,99
2223	7,71	32	3,44	3,25	0,34	0,03	1	7,50	1807,65	273,80	3341,50	179,41	3435,90
2704	7,90	30	3,2	3,37	0,07	0	1	7,51	1817,31	325,62	3308,99	228,72	3405,89
2923	7,98	66	2,3	3,45	0,79	0,82	1,12	7,79	2425,31	255,29	4595,33	187,57	4663,05
6152	8,72	117	6,33	3,34	0,66	0,37	1,02	9,16	9515,13	6650,30	12379,96	6598,67	12431,59

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

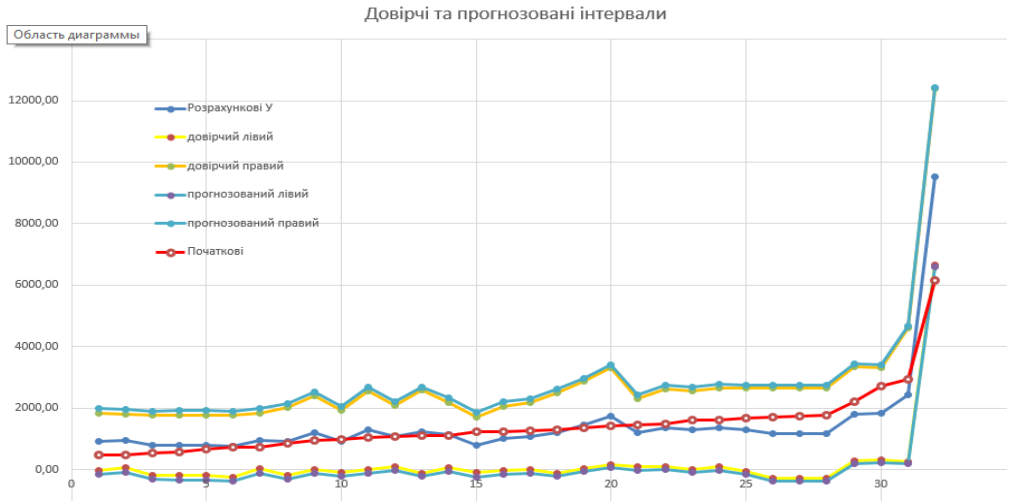

Рис.3. Довірчі та прогнозовані інтервали для У.

Висновки.

В результаті розрахунків було отримано рівняння множинної регресії:

lnY = 5.3318 + 0.022X1 + 0.01X2 + 0.08061X3-0.07548X4-0.7223X5 + 1.2149X6.

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

Статистична значимість рівняння перевірена за допомогою коефіцієнта детермінації і критерію Фішера. 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. Надалі планується застосування інших даних для побудови різних видів регресійних рівнянь для оцінювання розміру ПЗ інформаційних систем з відкритим кодом довільного типу.

Отримані при аналізі метрики ПЗ кількісно визначають різні властивості програмних продуктів у вигляді чисельного відображення. Поставлена ціль дослідження полягала у виведенні шістьох немультиколінеарних факторів особливостей ПЗ, що дало змогу порівнювати ці значення з подібними проектами, зі специфічними стандартами З отриманих результатів можна прийти до висновку щодо прогнозування розміру ПЗ, його якості та всього програмного процесу, а також, якщо необхідно,

планування необхідного часу для подальшої розробки.

Список бібліографічного опису.

1. Тан Х.Б.К. Оцінка LOC для інформаційних систем з їх концептуальних моделей даних. HBK Tan, Y. Zhao, H. Zhang. Матеріали 28-ї Міжнародної конференції з інженерії програмного забезпечення (ICSE 06), Шанхай, Китай, 20-28 травня 2006 р. - P . 321-330.
2. Тан Х.Б.К. Концептуальна оцінка розміру програмного забезпечення на основі моделей даних для інформаційних систем. Х. Б. К. Тан, Ю. Чжао, Х. Чжан . Операції з інженерії та методології програмного забезпечення. - 2009. - Вип. 19. - Випуск 2. - жовтень 2009. - Стаття № 4. DOI: 10.1145 / 1134285.1134331
3. Приходько С.Б. Побудова рівняння нелінійної регресії для оцінки розміру програмного забезпечення відкритих джерел інформаційних систем, що базуються на PHP.С. Б. Приходько, Н. В. Приходько, Т. Г. Смикодуб, А. В. Спінов - 2018. - № 1 (023). - С.118-125. - ISSN 1998-7005
4. Приходько С. Застосування відстані квадратичного махаланобіса для виявлення людей, що переживають багатофакторні дані, що не містять газів. С. Приходько, Н. Приходько, Л. Макарова, А. Пухалевич . Матеріали 14-ї міжнародної конференції з передових тенденцій радіоелектроніки, телекомунікацій та комп'ютерів Техніка (TCSET), Львів-Славське, Україна, 20–24 лютого 2018 р. - С. 962–965. DOI: 10.1109. TCSET.2018.8336353
5. Боем Б. Економіка програмного забезпечення. Б. Боем. - Нью-Джерсі: Прентіс-Холл, 1981. - 42 с.
6. Boehm V. Оцінка витрат на програмне забезпечення з Cocomo II. Нью-Джерсі, Прентіс-Холл. 2000. 544 с
7. Приходько Н.В., Приходько С.Б. Нелінійне рівняння регресії для оцінки розміру програмного забезпечення інформаційних систем з відкритим кодом у PHP. Приходько Н. В., Приходько С. Б. БІЗНЕС ЛІКУ СХІДНОГО УКРАЇНСЬКОГО НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ имени Володимира Даля. - 2018. - № 6 (247).

References

1. Tan H.B.K. Estimating LOC for information systems from their conceptual data models. H. B. K. Tan, Y. Zhao, H. Zhang. Proceedings of the 28th International Conference on Software Engineering (ICSE 06), Shanghai, China, May 20-28, 2006. – P. 321-330.
2. Tan H.B.K. Conceptual data model-based software size estimation for information systems. H. B. K. Tan, Y. Zhao, H. Zhang. Transactions on Software Engineering and Methodology. – 2009. – Vol. 19. – Issue 2. – October 2009. – Article No. 4. DOI: 10.1145/1134285.1134331
3. Prykhodko S.B. Constructing the non-linear regression equation to estimate the software size of open source PHPbased information systems / S. B. Prykhodko, N. V. Prykhodko, T. G. Smykodub, A. V. Spinov. Problemy informatsiinykh tekhnolohii. – 2018. – № 1 (023). – S.118-125. – ISSN 1998-7005
4. Prykhodko S. Application of the Squared Mahalanobis Distance for Detecting Outliers in Multivariate NonGaussian Data. S. Prykhodko, N. Prykhodko, L. Makarova, A. Pukhalevych.Proceedings of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, February 20–24, 2018. – P. 962-965. DOI: 10.1109/TCSET.2018.8336353
5. Boehm B. Software engineering economics. B. Boehm. – New Jersey: Prentice-Hall, 1981. – 42 p.
6. Boehm V. Software Cost Estimation with Cocomo II. New Jersey, Prentice-Hall. 2000. 544 p
7. Prikhodko NV, Prikhodko SB A nonlinear regression equation for estimating the size of open source information systems software in PHP. Prikhodko NV, Prikhodko SB. NEWSLETTER OF THE EASTERN UKRAINIAN NATIONAL UNIVERSITY imeni Volodymyra Dalia. – 2018. - № 6 (247).