МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет кораблебудування
імені адмірала Макарова

С. Б. ПРИХОДЬКО, Л. М. МАКАРОВА,
Н. В. ПРИХОДЬКО, А. В. ПУХАЛЕВИЧ

МЕТОДИЧНІ ВКАЗІВКИ ТА ЗАВДАННЯ
до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії»

Рекомендовано Методичною радою НУК

ВИДАВНИЦТВО
НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
КОРАБЛЕБУДУВАННЯ
ІМ. АДМІРАЛА МАКАРОВА
2023

УДК 004.412:519.237.5
М 54

Автори: С. Б. Приходько, д-р техн. наук, професор;
Л. М. Макарова, канд. техн. наук; доцент;
Н. В. Приходько, канд. екон. наук, доцент;
А. В. Пухалевич, канд. техн. наук

Рецензент М. В. Ушкац, д-р фіз.-мат. наук, професор

Рекомендовано Методичною радою НУК

М54 **Методичні** вказівки та завдання до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько, Л. М. Макарова, Н. В. Приходько, А. В. Пухалевич. – Миколаїв: НУК, 2023. – 56 с.

Методичні вказівки призначені для студентів четвертого курсу спеціальності 121 «Інженерія програмного забезпечення», які вивчають дисципліну «Емпіричні методи програмної інженерії». Також можуть бути корисними магістрам, аспірантам і усім тим, кому потрібно проводити первинну обробку та регресійний аналіз емпіричних даних, у тому числі багатовимірних.

УДК 004.412:519.237.5

ЗМІСТ

ПЕРЕДМОВА

В Національному університеті кораблебудування імені адмірала Макарова (НУК) дисципліна «Емпіричні методи програмної інженерії» вивчається студентами спеціальності 121 «Інженерія програмного забезпечення» у сьомому семестрі. Вона відноситься до циклу професійно-орієнтованих дисциплін навчального плану підготовки бакалаврів.

Ці методичні вказівки мають допомогти студентам четвертого курсу зазначеної спеціальності у підготовці та виконанні лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії». Вони містять завдання для лабораторних робіт, стислий виклад теоретичних відомостей, етапи виконання робіт, завдання для самостійної роботи, допоміжну літературу та питання для самоконтролю.

При виконанні кожної лабораторної роботи рекомендується:

- Засвоїти теоретичні відомості.
- Розібратися з завданням та етапами виконання роботи.
- Виконати запропоновані завдання та оформити звіт з роботи.
- Перевірити повноту розуміння теми за допомогою питань для самоконтролю, розташованих у кінці кожної роботи.

Лабораторну роботу необхідно виконувати відповідно за наведеними етапами. Звіт про лабораторну роботу оформлюється кожним студентом індивідуально. Звіт повинен містити: назву і мету роботи; завдання (постановку задачі); методику і алгоритм розв’язання задачі; текст програми (у разі якщо використовується оригінальна програма); результати роботи; аналіз результатів та висновки. Текст програми бажано розміщувати у додатку.

ВСТУП

Інженерія програмного забезпечення (software engineering) як дисципліна існує з 1960-х років, коли учасники Конференції НАТО з питань програмного забезпечення (ПЗ) в 1968 році в Гарміші (Німеччина) визнали, що існувала "криза програмного забезпечення" через підвищену складність систем та ПЗ у цих системах. Ця підвищена складність систем стала некерованою розробниками апаратного забезпечення, які також часто були розробниками ПЗ, оскільки вони краще знали апаратне забезпечення ніж процеси, методи та інструменти розробки ПЗ. Багато розробників ПЗ на той час, а також учасники Конференції НАТО усвідомлювали, що для розробки ПЗ потрібні цілеспрямовані процеси, методи та інструменти, і що вони відокремлюються від апаратних систем.

До кризи ПЗ дослідження переважно зосереджувались на теоретичних аспектах програмних систем, зокрема на алгоритмах та структурах даних, що використовуються для написання програмних систем, або на практичних аспектах програмних систем, зокрема на ефективному компілюванні ПЗ для певних систем. Криза ПЗ призвела до визнання того, що програмна інженерія – це більше, ніж обчислювальні теорії та ефективність коду, і що вона вимагає спеціальних досліджень. Таким чином, ця криза стала відправною точкою досліджень програмної інженерії. Це також дало відмінність між дисциплінами комп’ютерні науки (computer science) та програмна інженерія. Дослідження з комп’ютерних наук пов’язано з розумінням та пропонуванням теорій та методів, що стосуються в першу чергу ефективних обчислювальних алгоритмів. Дослідження програмної інженерії стосуються всіх аспектів життєвого циклу ПЗ.

Дослідження програмної інженерії мали величезний вплив на розробку ПЗ протягом останніх десятиліть, сприяючи прогресу в процесах, наприклад, за допомогою гнучких (agile) методів, методів налагодження (debugging), наприклад, в інтегрованих середовищах розробки, а також інструментів, зокрема, з рефакторингу (refactoring). Це також сприяло підвищенню якості програмних систем, об'єднанню досліджень та практики шляхом формалізації, вивчення та популяризації передового досвіду. Дослідження програмної інженерії на початку визнали, що інженерія ПЗ є принципово емпіричною дисципліною – тим

самим ще більше відрізняє комп'ютерні науки від інженерії програмного забезпечення, – оскільки, по-перше, ПЗ безпосередньо не підпорядковується фізичним законам, по-друге, ПЗ написане людьми для людей. Тому на багато аспектів програмної інженерії, за визначенням, впливає людський фактор. Необхідні емпіричні дослідження для виявлення цих людських факторів, що впливають на програмну інженерію, та вивчення впливу цих факторів на програмні системи та розробку ПЗ.

Існує подвійне визначення інженерії ПЗ. Буквально програмна інженерія – це створення та обслуговування (супровід) ПЗ. Але з точки зору досліджень, програмна інженерія – це сукупність знань про створення та обслуговування ПЗ та про явища, що лежать в основі цих двох видів діяльності. Тобто програмна інженерія пов'язана зі створенням та обслуговуванням ПЗ, а дослідження програмної інженерії направлено на інструменти для створення ПЗ, розуміння природи ПЗ та його використання.

Емпірична інженерія програмного забезпечення (empirical software engineering) – це область досліджень, пов'язана з емпіричним спостереженням артефактів програмної інженерії та емпіричним підтвердженням теорій та припущень програмної інженерії. Підобласті програмної інженерії, які звикли до емпіричних досліджень, включають еволюцію ПЗ, обслуговування ПЗ та майнінг сховищ (repositories) ПЗ.

Емпірична програмна інженерія – це область, яка охоплює декілька методів дослідження та зусиль, включаючи, але не обмежуючись цим, опитування для збору даних про якесь явище та контрольовані експерименти для вимірювання кореляції між змінними. Отже, емпіричні дослідження в програмній інженерії – це дослідження, в яких використовується будь-який із емпіричних методів дослідження: опитування (включаючи систематичні огляди літератури), кейси, квазі-експерименти та контрольовані експерименти. І навпаки, ми стверджуємо, що дослідницькі роботи з програмної інженерії, які не включають опитування, тематичні дослідження чи експерименти, не є емпіричними дослідженнями (або не містять емпіричних досліджень). Проте сьогодні рідко коли дослідницькі роботи з ПЗ не включають деяких емпіричних досліджень. Дійсно, емпіричні дослідження є однією з найпопулярніших тем таких конференцій, як Міжнародна конференція ACM/IEEE з програмної інженерії. Тому ми вважаємо корисним на

даний момент часу та задля повноти цього розділу згадати виправдання та поняття, пов'язані з емпіричною програмною інженерією.

Отже ми бачимо, що дослідження програмної інженерії мають чіткий поворот до емпіричних досліджень. Причини цього чіткого повороту подвійні: вони стосуються природи досліджень програмної інженерії та переваг, які забезпечують емпіричні дослідження. З одного боку, дослідження програмної інженерії прагне слідувати науковому методу, щоб запропонувати обґрунтовані результати, і тому вимагає спостережень та експериментів для створення та оцінювання гіпотез та теорій. Емпіричне дослідження пропонує методи, необхідні для здійснення цих спостережень та експериментів, з яких одними із відомих є: опитування, тематичне дослідження (case study) та контрольований експеримент.

Тематичне дослідження (case study) – це емпіричний метод, спрямований на дослідження сучасних явищ у їх контексті.

Контрольований експеримент являє собою дослідження гіпотези, що перевіряється, коли однією або кількома незалежними змінними маніпулюють для вимірювання їх впливу на одну або кілька залежних змінних.

Як було зазначено раніше, на багато аспектів програмної інженерії, за визначенням, впливає людський фактор, який за своєю природою є випадковим. Тому при проведенні емпіричних досліджень в програмній інженерії застосовують різноманітні статистичні методи, у тому числі, методи регресійного аналізу, які розглядаються у наведених лабораторних роботах.

Лабораторна робота № 1
Первинна обробка емпіричних даних з метрик програмного забезпечення

Мета роботи: отримати практичні навички первинної обробки емпіричних даних з метрик програмного забезпечення.

Завдання: виконати первинну обробку емпіричних даних з метрик програмного забезпечення, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Первинну обробку емпіричних даних виконують у наступному порядку:

1) Визначають вибіркове середнє $\bar{x}$ (точкову оцінку математичного сподівання) N значень x_i $(i = 1,2,\ldots,N)$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

2) Визначають вибіркову дисперсію S_x^2 (незміщену точкову оцінку дисперсії) та точкову оцінку середнього квадратичного відхилення $\hat{\sigma}_x$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 ;\ \hat{\sigma}_x = \sqrt{S_x^2} .$$

3) Перевіряють нормальність закону розподілу емпіричних даних.

4) Для заданого значення довірчої ймовірності знаходять довірчі інтервали для вибіркового середнього та середнього квадратичного відхилення.

5) Визначають наявність викидів та, якщо останні знайдені, відповідні результати відкидають і повторюють обчислення.

Нагадаємо деякі відомості з теорії ймовірностей. Під випадковою величиною розуміють величину, яка у результаті досліду з випадковим результатом приймає те або інше значення. Оскільки закономірностей у появі цих значень немає, аналіз таких величин може виконуватися тільки методами теорії ймовірностей і математичної статистики. Для характеристики випадкової величини необхідно

знати сукупність значень цієї величини, а також ймовірності, з якими ці значення можуть появитися.

Випадкова величина називається дискретною, якщо множина її можливих значень кінцева або лічильна. Неперервні (не дискретні) випадкові величини характеризуються тим, що множина їх можливих значень не лічильна.

Законом розподілу випадкової величини називається будь-яке правило, яке дозволяє знаходити ймовірності можливих подій, зв'язаних з випадковою величиною.

Найбільш загальною формою закону розподілу випадкової величини є функція розподілу, яка представляє собою ймовірність того, що випадкова величина X прийме значення менше ніж задане x:

$$F(x) = P\{X < x\}.$$

Якщо функція розподілу $F(x)$ випадкової величини X при будь-якому x неперервна і, крім того, має похідну $F'(x)$ будь-де, крім, можливо, окремих точок, то випадкова величина є неперервною.

Щільністю ймовірності неперервної випадкової величини X називається похідна функції розподілу:

$$f(x) = F'(x).$$

Найбільш розповсюдженим для неперервних випадкових величин є нормальний розподіл (розподіл Гауса) з щільністю ймовірності

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}},$$

де m_x – математичне сподівання випадкової величини X.

Нормально розподілені випадкові величини часто зустрічаються на практиці при проведенні вимірів. Так, випадкові похибки багатократних вимірів як правило розподілені за нормальним законом, навіть коли закони розподілу ймовірностей складових відрізняються від нормального.

Крім того історично склалося так, що багато статистичних критеріїв, методів і оцінок розроблені тільки для випадку нормального початкового розподілу. Тому при первинній обробці експериментальних даних перевіряють нормальність закону розподілу результатів спостережень.

У разі великої вибірки значень випадкової величини (коли $N > 30$) перевірку нормальності закону розподілу результатів спостережень в тому числі виконують за критерієм Пірсона (критерієм χ^2). Відповідно за цим критерієм спочатку обчислюють значення χ^2

$$\chi^2 = \sum_{j=1}^{m} \frac{\left(n_j - Np_j\right)^2}{N p_j},$$

де m – кількість підінтервалів (часток, кліток), на які розбивається інтервал $[x_{\min}, x_{\max}]$; $x_{\min}$ і $x_{\max}$ – відповідно мінімальне і максимальне значення випадкової величини X; n_j – абсолютна частота в j-му підінтервалі (кількість значень випадкової величини, які попадають у j-ий підінтервал); p_j – ймовірність того, що значення випадкової величини X попадають у j-ий підінтервал.

Кількість підінтервалів (кліток), на які розбивається інтервал $[x_{\min}, x_{\max}]$ може бути визначений за наступними формулами: $m = \log_2 N + 1 = 3{,}3 \lg N + 1$ (формула Старджеса) або $m = 5 \lg N$ (формула Брукса і Карузера).

Ймовірність того, що значення випадкової величини X попадають у j-ий підінтервал можуть бути визначені як

$$p_j = \int_{x_{j-1}}^{x_j} f(x)dx,$$

де x_{j-1} і x_j – ліва і права границі j-го підінтервалу.

Після цього в залежності від рівня значимості α та кількості ступенів вільності ν за таблицею верхніх 100α %-вих точок розподілу χ^2 (таблиця А1 у додатку А) знаходять значення $\chi^2_{кр}$. Якщо $\chi^2 \le \chi^2_{кр}$, то з ймовірністю $1-\alpha$ можна прийняти гіпотезу про те, що закон розподілу результатів спостережень є нормальним, якщо $\chi^2 > \chi^2_{кр}$ – цю гіпотезу потрібно відкинути.

Кількість ступенів вільності ν визначається як

$$\nu = m - k - 1,$$

де k – кількість параметрів, від яких залежить закон розподілу. Для нормального розподілу $k = 2$.

У разі малої вибірки значень випадкової величини (коли $N < 30$) перевірку нормальності закону розподілу результатів спостережень необхідно виконувати за іншими критеріями, наприклад, за критерієм Колмогорова-Смирнова.

У разі, якщо із заданою довірчою ймовірністю закон розподілу результатів спостережень можна вважати нормальним, для цього ж значення ймовірності знаходять довірчу похибку результату виміру та довірчий інтервал для середнього квадратичного відхилення.

Завдяки випадковому характеру похибки $\Delta\hat{\theta}$ оцінки $\hat{\theta}$ параметра θ для конкретизації точності наближеної рівності $\hat{\theta} \approx \theta$ необхідно мати ймовірність p_∂ того, що $\left|\Delta\hat{\theta}\right|$ перейде деяку границю $\Delta > 0$:

$$P\left(\left|\Delta\hat{\theta}\right| \le \Delta\right) = p_\partial .$$

Інтервал від $\hat{\theta} - \Delta$ до $\hat{\theta} + \Delta$, в якому з ймовірністю p_∂ знаходиться справжнє значення θ, називається довірчим інтервалом, а його границі – довірчими границями, ймовірність p_∂ – довірчою ймовірністю.

Величина $\alpha = 1 - p_\partial$ в загальному випадку називається рівнем значимості. Під рівнем значимості якої-небудь статистичної гіпотези розуміють найбільшу ймовірність α, з якою ця гіпотеза може дати помилковий результат.

Для нормальної генеральної сукупності $(1-\alpha)$%-вий довірчий інтервал вибіркового середнього визначається як

$$\left[\bar{x} - t_{N-1}S_x/\sqrt{N},\ \bar{x} + t_{N-1}S_x/\sqrt{N}\right],$$

де t_{N-1} – квантіль *t*-розподілу Стьюдента, визначається за таблицею верхніх 100α %-вих точок *t*-розподілу Стьюдента (таблиця Б1 у додатку Б) в залежності від рівня значимості $\alpha/2$ та кількості ступенів свободи ν, $\nu = N - 1$. Величину $t_{N-1}S_x/\sqrt{N}$ розглядають як довірчу похибку результату визначення вибіркового середнього. Вона зменшується зі збільшенням *N*.

Приклад 1.1. Знайти 95%-вий довірчий інтервал вибіркового середнього нормальної сукупності із 10 значень випадкової величини, якщо $\bar{x} = 52$ і $S_x^2 = 96$.

Для рівня значимості $\alpha/2=(1-0{,}95)/2=0{,}025$ і кількості ступенів свободи $\nu=10-1=9$ визначають квантіль *t*-розподілу Стьюдента $t_9(0{,}025)=2{,}26$. Тоді 95%-вий довірчий інтервал вибіркового середнього визначається як $\left[52-2{,}26\cdot 9{,}80/\sqrt{10},\ 52+2{,}26\cdot 9{,}80/\sqrt{10}\right]$, або остаточно $[45;59]$.

Для нормальної генеральної сукупності $(1-\alpha)$%-вий довірчий інтервал точкової оцінки середнього квадратичного відхилення визначають як

$$\left[S_x\sqrt{N/\chi^2_{\alpha/2}},\ S_x\sqrt{N/\chi^2_{1-\alpha/2}}\right].$$

Значення $\chi^2_{\alpha/2}$ і $\chi^2_{1-\alpha/2}$ визначають в залежності від α та ν за таблицею верхніх 100α %-вих точок розподілу χ^2 (таблиця А1 у додатку А). Кількість ступенів свободи ν визначається як $\nu=N-1$.

Приклад 1.2. Знайти 95%-вий довірчий інтервал точкової оцінки середнього квадратичного відхилення нормальної сукупності із 11 значень випадкової величини, якщо $S_x=0{,}130$.

Для кількості ступенів свободи $\nu=11-1=10$ та рівнів значимості $\alpha/2=(1-0{,}95)/2=0{,}025$ і $1-\alpha/2=1-(1-0{,}95)/2=0{,}975$ знаходять значення $\chi^2_{10}(0{,}025)=20{,}48$ і $\chi^2_{10}(0{,}975)=3{,}25$. Тоді 95%-вий довірчий інтервал точкової оцінки середнього квадратичного відхилення визначають як $\left[0{,}13\sqrt{11/20{,}48},\ 0{,}13\sqrt{11/3{,}25}\right]$ або $[0{,}095;0{,}239]$.

Для нормальної генеральної сукупності значення z^* є викидом, якщо z^* виходить за границі інтервалу цензурування вибірки

$$\left[\bar{z}-\Delta_z,\ \bar{z}+\Delta_z\right], \tag{1.1}$$

де $\Delta_z=\begin{cases}4{,}0S_z, & 50<N\le 100;\\ 4{,}5S_z, & 100<N\le 1000;\\ 5{,}0S_z, & 1000<N\le 10000.\end{cases}$

Кращій результат отримують у разі пошуку Δ_z за Граббом (Grubb)

$$\Delta_z=S_z\,\frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}, \tag{1.2}$$

де $t_{\alpha/(2N),N-2}$ – квантіль *t*-розподілу Стьюдента з $N-2$ ступенями свободи та $\alpha/(2N)$ рівнем значимості.

Але всі зазначені методи визначення границь цензурування вибірки розраховані на гаусівський розподіл даних.

У разі, якщо генеральна сукупність не є нормальною, значення Δ_x може бути обчислено як

$$\Delta_x = \left[1{,}55 + 0{,}8\sqrt{\varepsilon - 1}\lg(N/10)\right]S_x, \tag{1.3}$$

де ε – ексцес, $\varepsilon = \mu_4(x)/\sigma_x^4$; $\mu_4(x)$ – центральний статистичний момент четвертого порядку випадкової величини X.

Оцінка центрального статистичного моменту четвертого порядку може бути знайдена за формулою

$$\hat{\mu}_4(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4.$$

Але формула (1.3) та подібні їй не враховують можливу асиметрію закону розподілу негаусівських даних. В [2] запропоновано статистичний метод визначення викидів в емпіричних даних на основі нормалізуючих перетворень, суть якого полягає у наступному. Спочатку значення негаусівської випадкової величини *x* перетворюють у такі, що мають нормальний розподіл, за перетворенням

$$z = \psi(x). \tag{1.4}$$

Далі за значенням вибіркового середнього $\bar{x}$ випадкової величини *x* обчислюємо

$$\bar{z}_{\bar{x}} = \psi(\bar{x}).$$

Потім для нормалізованих даних знаходимо значення Δ_z за (1.2) та визначаємо границі інтервалу (1.1)

$$\left[\bar{z}_{\bar{x}} - \Delta_z, \bar{z}_{\bar{x}} + \Delta_z\right]. \tag{1.5}$$

І нарешті визначаємо границі цензурування вибірки значень випадкової величини *x* шляхом перетворення границь інтервалу (1.5) використовуючи перетворення зворотне до (1.4) $x = \psi^{-1}(z)$

$$\left[\psi^{-1}(\bar{z}_{\bar{x}} - \Delta_z), \psi^{-1}(\bar{z}_{\bar{x}} + \Delta_z)\right]. \tag{1.6}$$

За (1.6) можна визначати границі цензурування вибірки значень негаусівської випадкової величини *x*. Всі значення *x*, які виходять за інтервал (1.6) вважаються аномальними або грубими помилками.

В загальному випадку у якості перетворення (1.4) ми рекомендуємо використовувати перетворення Джонсона [2], яке розглядається у лабораторній роботі № 2. Також в певних випадках можна використовувати інші нормалізуючі перетворення, наприклад, перетворення Бокса-Кокса (Box-Cox), перетворення у вигляді десяткового логарифму.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Визначення вибіркового середнього $\bar{x}$ (точкової оцінки математичного сподівання) n результатів спостережень x_i $(i = 1,2,\ldots,n)$.

2) Визначення вибіркової дисперсії S_x^2 (незміщеної точкової оцінки дисперсії), точкової оцінки середнього квадратичного відхилення $\hat{\sigma}_x$.

3) Побудова гістограми за вибіркою значень випадкової величини *X*. Побудова нормального закону розподілу експериментальних даних (функції щільності ймовірності) на гістограмі.

4) Перевірка нормальності закону розподілу значень випадкової величини *X*.

5) Знаходження довірчих інтервалів для вибіркового середнього та середнього квадратичного відхилення випадкової величини *X* у разі, якщо закон розподілу результатів спостережень вважається нормальним.

6) Визначення наявності викидів (якщо останні знайдені, відповідні результати відкидають і повторюють обчислення).

7) Формулювання висновків за результатами виконання лабораторної роботи.

Завдання для самостійної роботи

Порівняти отримані результати первинної обробки емпіричних даних з відповідними результатами у випадку врахування того, що емпіричні дані не є нормальними.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Навіщо виконують первинну обробку емпіричних даних?
2) Назвіть послідовність виконання первинної обробки емпіричних даних.
3) Що таке випадкова величина?
4) Що таке функція розподілу випадкової величини?
5) Що таке щільність ймовірності неперервної випадкової величини?
6) Від яких параметрів залежить нормальний розподіл (розподіл Гауса)?
7) Як визначити вибіркову дисперсію (незміщену точкову оцінку дисперсії) та точкову оцінку середнього квадратичного відхилення випадкової величини?
8) Для чого перевіряють гіпотезу відносно нормальності закону розподілу ймовірностей?
9) Поясніть критерій χ^2 (критерій Пірсона).
10) Яку вибірку вважають малою (великою)?
11) За яким критерієм можна перевірити гіпотезу відносно нормальності закону розподілу ймовірностей у разі великої (малої) вибірки?
12) Що таке довірчий інтервал, довірча ймовірність, рівень значимості?
13) Як визначають для нормальної генеральної сукупності α%-ві довірчі границі вибіркового середнього?
14) Як визначають для нормальної генеральної сукупності α%-ві довірчі границі точкової оцінки середнього квадратичного відхилення?
15) Як перевіряють наявність викидів у разі коли закон розподілу є нормальним?
16) Що таке інтервал цензурування вибірки?
17) Як можна перевірити наявність викидів у разі коли закон розподілу не є нормальним?
18) Що роблять у разі, якщо у виборці знайдені викиди?

Лабораторна робота № 2
Побудова аналітичної моделі закону розподілу емпіричних даних з метрик програмного забезпечення

Мета роботи: отримати практичні навички побудови аналітичної моделі закону розподілу емпіричних даних з метрик програмного забезпечення.

Завдання: виконати побудову аналітичної моделі закону розподілу емпіричних даних з метрик програмного забезпечення, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Підбор аналітичної моделі закону розподілу є засобом узагальненого представлення емпіричних даних у тих випадках, коли відсутнє теоретичне обґрунтування закону розподілу випадкової величини або функції.

Для опису емпіричних даних аналітичну модель закону розподілу підбирають зазвичай або із сімей розподілів Джонсона, або із класів розподілів Пірсона. Ми рекомендуємо застосовувати розподіли Джонсона. Їх перевага полягає у тому, що вони побудовані на основі нормалізуючих перетворень, які дозволяють перейти від значень негаусівської випадкової величини до нормально розподіленої випадкової величини. А історично склалося так, що велика кількість статистичних критеріїв і методів розроблені в основному тільки для випадку нормального розподілу.

Сім'ї розподілів Джонсона отримані шляхом перетворення нормованої нормально розподіленої випадкової величини z. В загальному випадку перетворення має вигляд [2]

$$z = \gamma + \eta q(x, \varphi, \lambda); \quad \eta > 0; \quad -\infty < \gamma < \infty; \quad \lambda > 0; \quad -\infty < \varphi < \infty, \qquad (2.1)$$

де q – довільна функція; γ, η, λ, φ – параметри розподілу Джонсона. Слід відзначити, що γ і η – це параметри форми, λ – це параметр масштабу, параметр φ характеризує зсув розподілу у напряму осі абсцис.

Джонсон запропонував три різні сім'ї функцій q:

1) $q(x,\varphi,\lambda)=\ln\left(\dfrac{x-\varphi}{\lambda}\right)$, $x>\varphi$ (сім'я S_L);

2) $q(x,\varphi,\lambda)=\ln\left(\dfrac{x-\varphi}{\lambda+\varphi-x}\right)$, $\varphi<x<\varphi+\lambda$ (сім'я S_B);

3) $q(x,\varphi,\lambda)=Arsh\left(\dfrac{x-\varphi}{\lambda}\right)$, $-\infty\le x\le+\infty$ (сім'я S_U),

де $Arsh\left(\dfrac{x-\varphi}{\lambda}\right)=\ln\left[\dfrac{x-\varphi}{\lambda}+\sqrt{\left(\dfrac{x-\varphi}{\lambda}\right)^2+1}\right]$.

Для сім'ї S_L функція щільності ймовірності задається як

$$f_L(x)=\frac{\eta}{\sqrt{2\pi}(x-\varphi)}\exp\left\{-\frac{\eta^2}{2}\left[\frac{\gamma-\eta\ln\lambda}{\eta}+\ln(x-\varphi)\right]^2\right\},$$

де $x>\varphi$; $\eta>0$; $-\infty<\gamma<\infty$; $\lambda>0$; $-\infty<\varphi<\infty$.

Ця функція щільності ймовірності зростає від φ до точки максимуму і потім більш повільно спадає у разі спрямування x у нескінченність $(x\to\infty)$. Функція щільності ймовірності для сім'ї S_L завжди унімодальна, асиметрія завжди додатна, а ексцес більший за 3 (значення ексцесу нормального розподілу).

Для сім'ї S_B функція щільності ймовірності задається формулою

$$f_B(x)=\frac{\eta\lambda}{\sqrt{2\pi}(x-\varphi)(\lambda+\varphi-x)}\exp\left\{-\frac{1}{2}\left[\gamma+\eta\ln\left(\frac{x-\varphi}{\lambda+\varphi-x}\right)\right]^2\right\},$$

де $\varphi<x<\varphi+\lambda$; $\eta>0$; $-\infty<\gamma<\infty$; $\lambda>0$; $-\infty<\varphi<\infty$.

Функції щільності ймовірності сім'ї S_B можуть бути як унімодальними, так і бімодальними. Необхідні та достатні умови для бімодальності полягають у тому, що

$$\eta<1/\sqrt{2}\,,\qquad |\gamma|<\eta^{-1}\sqrt{1-2\eta^2}-2\eta\,arth\sqrt{1-2\eta^2}\,.$$

Для сім'ї S_U функція щільності ймовірності визначається як

$$f_U(x) = \frac{\eta}{\sqrt{2\pi\left\{(x-\varphi)^2 + \lambda^2\right\}}} \exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{x-\varphi}{\lambda} + \sqrt{\left(\frac{x-\varphi}{\lambda}\right)^2 + 1}\right)\right]^2\right\},$$

де $-\infty < x < \infty$; $\eta > 0$; $-\infty < \gamma < \infty$; $\lambda > 0$; $-\infty < \varphi < \infty$.

Графіки функції щільності ймовірності сім'ї S_U мають високий порядок стику з віссю абсцис на кінцях інтервалу зміни аргументу, є унімодальними та їх моди лежать між медіаною та нулем. Тим самим ці розподіли мають додатну або від'ємну асиметрію в залежності від того, від'ємна або додатна величина параметра γ.

Сім'ї розподілів Джонсона відрізняються багатостатністю форм і у площині асиметрії у квадраті A^2 та ексцесу ε займають значні області. На рисунку 2.1 представлена діаграма Джонсона (області комбінацій A^2 і ε для різних розподілів Джонсона). Ця діаграма дозволяє підібрати сім'ю розподілів Джонсона за значеннями оцінок A^2 та ε вибіркового розподілу.

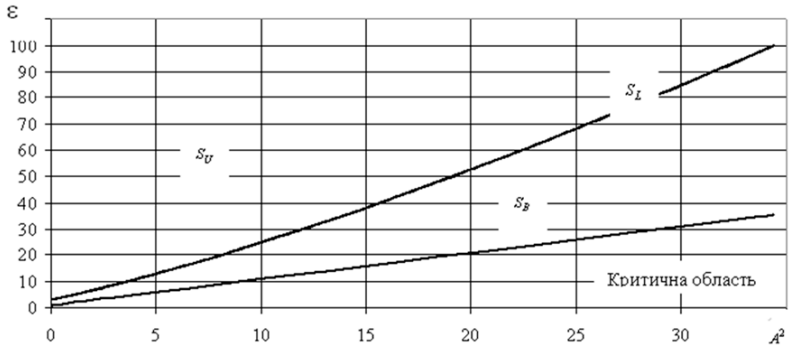


Рисунок 2.1 – Комбінації A^2 і ε для різних розподілів Джонсона [2]

Якщо комбінація A^2 і ε знаходиться біля лінії S_L, то вибірковий розподіл можна апроксимувати розподілом із сім'ї S_L (логарифмічно нормальним розподілом). Розподіл зі значеннями A^2 і ε, які лежать вище лінії S_L, апроксимуються розподілом із сім'ї S_U, а які лежать

нижче лінії S_L (до лінії критичної області) – розподілом із сім'ї S_B. Якщо комбінація A^2 і ε попадає в критичну область, то вибірковий розподіл не можна апроксимувати розподілом із сімей Джонсона.

Лінія верхньої границі критичної області задається залежністю

$$\varepsilon = A^2 + 1.$$

Лінію S_L в межах $A^2 \in [0, 27]$ можна апроксимувати наступною функцією [2]:

$$\varepsilon(A^2) = 7{,}2315 \cdot 10^{-6} A^8 - 6{,}9860 \cdot 10^{-4} A^6 + 4{,}5460 \cdot 10^{-2} A^4 + \\ + 1{,}7979 A^2 + 2{,}9891.$$

Практично завжди при підборі аналітичної моделі закону розподілу емпіричних даних асиметрія A та ексцес ε не бувають відомі. У цьому випадку вибір сім'ї розподілів Джонсона здійснюється за оцінками асиметрії A та ексцесу ε, які знаходяться за гістограмою. Після вибору необхідної сім'ї розподілів Джонсона обчислюють його параметри, і виконують перевірку адекватності обраної моделі експериментальним даним за критерієм згоди, наприклад, критерієм Пірсона (χ^2- критерієм).

Параметри функції щільності ймовірності γ, η, λ та φ для обраної сім'ї розподілів Джонсона в загальному випадку можна знайти шляхом рішення наступної задачі математичного програмування:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \left\{ \sum_{j=1}^{m} \left[y(x_j) - f(x_j, \boldsymbol{\theta}) \right]^2 \right\}, \qquad (2.2)$$

де $\boldsymbol{\theta}$ – вектор невідомих параметрів, $\boldsymbol{\theta} = \{\gamma, \eta, \lambda, \varphi\}$; x_j – значення випадкової величини x в середині j-го підінтервалу; $y(x_j)$ – значення ординати гістограми при значенні x_j; $f(x_j, \theta)$ – вираз функції щільності ймовірності при значенні x_j; m – кількість підінтервалів гістограми. Зазначимо, що при використанні формули (2.2), у якості значення $y(x_j)$ потрібно брати відношення відносної частоти в j-му підінтервалі до довжини j-го підінтервалу Δx.

Параметри функції щільності ймовірності γ, η, λ та φ для обраної сім'ї розподілів Джонсона також можна знайти із рішення системи рівнянь, яка складена з рівностей відповідних аналітичних виразів для

статистичних моментів $\alpha_1(x)$, $\mu_2(x)$, $\mu_3(x)$ і $\mu_4(x)$ та їх оцінок, обчислених за вибірковими даними гістограми [2].

Відповідні оцінки статистичних моментів $\alpha_1(x)$, $\mu_2(x)$, $\mu_3(x)$ і $\mu_4(x)$ знаходяться за гістограмою за наступними формулами:

$$\hat{\alpha}_1(x) = \bar{x} = \sum_{j=1}^{m} y(x_j)\Delta x x_j;$$

$$\hat{\mu}_k(x) = \sum_{j=1}^{m} y(x_j)\Delta x\left[x_j - \bar{x}\right]^k, \quad k = 2,3,4.$$

Для сім'ї розподілів S_L статистичні моменти $\alpha_1(x)$, $\mu_2(x)$, $\mu_3(x)$ і $\mu_4(x)$ можна представити у вигляді наступних виразів:

$$\alpha_1(x) = \varphi + \lambda\omega\rho;$$

$$\mu_2(x) = \lambda^2\omega^2\rho^2\left(\omega^2 - 1\right);$$

$$\mu_3(x) = \lambda^3\omega^3\rho^3\left(\omega^6 - 3\omega^2 + 2\right);$$

$$\mu_4(x) = \lambda^4\omega^4\rho^4\left(\omega^2 - 1\right)\left(\omega^8 + 2\omega^6 + 3\omega^4 - 3\right),$$

де $\omega = \exp\left(0{,}5/\eta^2\right)$; $\rho = \exp(-\gamma/\eta)$.

Тоді параметри функції щільності ймовірності γ, η, λ та φ для сім'ї S_L розподілів Джонсона можна знайти із рішення наступної системи рівнянь:

$$\varphi + \lambda\omega\rho = \hat{\alpha}_1(x);$$

$$\lambda^2\omega^2\rho^2\left(\omega^2 - 1\right) = \hat{\mu}_2(x);$$

$$\lambda^3\omega^3\rho^3\left(\omega^6 - 3\omega^2 + 2\right) = \hat{\mu}_3(x);$$

$$\lambda^4\omega^4\rho^4\left(\omega^2 - 1\right)\left(\omega^8 + 2\omega^6 + 3\omega^4 - 3\right) = \hat{\mu}_4(x).$$

Найпростішу процедуру підгонки параметрів для сім'ї розподілів S_L запропонував Уіксел і вона полягає у наступному. Спочатку знаходять $t = \left(\omega^2 - 1\right)^{0,5}$ як додатній корінь рівняння

$$t^3 + 3t - A = 0.$$

Після чого параметр зсуву φ визначається за формулою

$$\varphi = \alpha_1(x) - \frac{\sqrt{\mu_2(x)}}{t}.$$

Параметр η визначається значенням ω, а саме

$$\eta = \frac{1}{\sqrt{2\ln\omega}}.$$

Вираз $\gamma - \eta\ln\lambda$ розглядається як єдиний параметр і обчислюється після логарифмування добутку $\lambda\rho$.

Для сім'ї розподілів S_U статистичні моменти $\alpha_1(y)$, $\mu_2(y)$, $\mu_3(y)$ і $\mu_4(y)$ можна представити у вигляді наступних виразів:

$$\alpha_1(y) = -\omega sh\Omega;$$

$$\mu_2(y) = \frac{1}{2}\left(\omega^2 - 1\right)\left(\omega^2 ch2\Omega + 1\right);$$

$$\mu_3(y) = -\frac{1}{4}\omega^2\left(\omega^2 - 1\right)^2\left\{\omega^2\left(\omega^2 + 2\right)sh3\Omega + 3sh\Omega\right\}; \qquad (2.3)$$

$$\mu_4(y) = \frac{1}{8}\left(\omega^2 - 1\right)^2\left\{\omega^4\left(\omega^8 + 2\omega^6 + 3\omega^4 - 3\right)ch4\Omega + \right.$$

$$\left. + 4\omega^4\left(\omega^2 + 2\right)ch2\Omega + 3\left(2\omega^2 + 1\right)\right\},$$

де $y = (x - \varphi)/\lambda$; $\omega = \exp\left(0{,}5/\eta^2\right)$; $\Omega = \gamma/\eta$; $sh\Omega = 0{,}5\left(e^{\Omega} - e^{-\Omega}\right)$; $ch\Omega = 0{,}5\left(e^{\Omega} + e^{-\Omega}\right)$.

Врахувавши, що $y = (x - \varphi)/\lambda$, зв'язок між статистичними моментами для випадкової величини y в (2.3) та величини x є таким:

$$\alpha_1(y) = \frac{\alpha_1(x) - \varphi}{\lambda};$$

$$\mu_2(y) = \frac{1}{\lambda^2}\left[\alpha_2(x) - \alpha_1^2(x)\right];$$

$$\mu_3(y) = \frac{1}{\lambda^3}\left[\alpha_3(x) - 3\alpha_2(x)\alpha_1(x) + 2\alpha_1^3(x)\right]; \qquad (2.4)$$

$$\mu_4(y)=\frac{1}{\lambda^4}\left[\alpha_4(x)-4\alpha_3(x)\alpha_1(x)+6\alpha_2(x)\alpha_1^2(x)-3\alpha_1^4(x)\right],$$

де $\alpha_2(x)$, $\alpha_3(x)$ і $\alpha_4(x)$ – початкові статистичні моменти відповідно 2-го, 3-го та 4-го порядку випадкової величини x. Їх оцінки знаходяться за гістограмою як

$$\hat{\alpha}_k(x)=\sum_{j=1}^{m} y(x_j)\Delta x\, x_j^k.$$

Тоді параметри функції щільності ймовірності γ, η, λ та φ для сім'ї S_U розподілів Джонсона можна знайти таким чином. Із рішення системи рівнянь

$$\frac{\lambda^2}{2}\left(\omega^2-1\right)\left(\omega^2\frac{e^{2\Omega}+e^{-2\Omega}}{2}+1\right)=\alpha_2(x)-\alpha_1^2(x);$$

$$-\frac{\lambda^3}{8}\omega^2\left(\omega^2-1\right)^2\left\{\omega^2\left(\omega^2+2\right)\left(e^{3\Omega}-e^{-3\Omega}\right)+3\left(e^{\Omega}-e^{-\Omega}\right)\right\}=$$

$$=\alpha_3(x)-3\alpha_2(x)\alpha_1(x)+2\alpha_1^3(x); \qquad (2.5)$$

$$\frac{\lambda^4}{8}\left(\omega^2-1\right)^2\left\{\omega^4\left(\omega^8+2\omega^6+3\omega^4-3\right)\frac{e^{4\Omega}+e^{-4\Omega}}{2}+\right.$$

$$\left.+4\omega^4\left(\omega^2+2\right)\frac{e^{2\Omega}+e^{-2\Omega}}{2}+3\left(2\omega^2+1\right)\right\}=$$

$$=\alpha_4(x)-4\alpha_3(x)\alpha_1(x)+6\alpha_2(x)\alpha_1^2(x)-3\alpha_1^4(x)$$

визначають параметри λ, ω та Ω. Рішення системи (2.5) може бути знайдено методом Ньютона. Далі визначають параметри η і γ як

$$\eta=\sqrt{0{,}5/\ln\omega}, \qquad \omega>0;$$

$$\gamma=\Omega\eta.$$

І останній параметр φ знаходять із рішення першого рівняння системи (2.4)

$$\varphi=\lambda\omega sh\Omega+\alpha_1(x).$$

Для сім'ї розподілів S_B вирази для статистичних моментів отримати досить складно у прийнятній формі. Тому для підбору параметрів

розподілу в деяких випадках простіше використовувати інші величини, такі як, наприклад, квантілі [2].

Оцінки параметрів перетворення Джонсона краще визначати за методом максимальної правдоподібності

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}),$$

де $l(\boldsymbol{\theta})$ – логарифмічна функція правдоподібності. Для перетворення Джонсона сім'ї S_U логарифмічну функцію правдоподібності можна записати як [2]

$$l(\boldsymbol{\theta}) = N\ln(\eta) - \frac{N\ln(2\pi)}{2} - \frac{1}{2}\sum_{i=1}^{N}\ln\left[(x_i-\varphi)^2+\lambda^2\right] -$$

$$-\frac{1}{2}\sum_{i=1}^{N}\left[\gamma+\eta\,\mathrm{Arsh}\left(\frac{x_i-\varphi}{\lambda}\right)\right]^2.$$

Для перетворення Джонсона сім'ї S_B, логарифмічну функцію правдоподібності можна записати як

$$l(\boldsymbol{\theta}) = N\ln(\eta\lambda) - \frac{N\ln(2\pi)}{2} - \sum_{i=1}^{N}\ln(x_i-\varphi) - \sum_{i=1}^{N}\ln(\varphi+\lambda-x_i) -$$

$$-\frac{1}{2}\sum_{i=1}^{N}\left[\gamma+\eta\ln\frac{x_i-\varphi}{\varphi+\lambda-x_i}\right]^2.$$

Зауважте, при оцінюванні параметрів γ, η, λ та φ потрібно враховувати відповідні обмеження, що накладаються на ці параметри для обраної сім'ї розподілів Джонсона.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Побудова гістограми за вибіркою значень випадкової величини x.

2) Знаходження оцінки центру вибіркового розподілу або оцінки математичного сподівання випадкової величини x за гістограмою.

3) Обчислення оцінок центральних статистичних моментів $\mu_2(x)$, $\mu_3(x)$ і $\mu_4(x)$ за гістограмою або за вибіркою значень випадкової величини.

4) Обчислення оцінок асиметрії A та ексцесу ε за оцінками моментів $\mu_2(x)$, $\mu_3(x)$ і $\mu_4(x)$.

5) Вибір сім'ї розподілів Джонсона за оцінками A^2 і ε.

6) Знаходження параметрів функції щільності ймовірності γ, η, λ та φ для обраної сім'ї розподілів Джонсона.

7) Перевірка адекватності знайденої аналітичної моделі вибірковому розподілу емпіричних даних за критерієм згоди.

8) Побудова обраної аналітичної моделі закону розподілу емпіричних даних (функції щільності ймовірності) на гістограмі.

9) Формулювання висновків за результатами виконання лабораторної роботи.

Завдання для самостійної роботи

Знайти оцінки параметрів аналітичної моделі закону розподілу емпіричних даних з метрик програмного забезпечення за іншим методом. Порівняти отримані аналітичні моделі закону розподілу емпіричних даних з метрик програмного забезпечення для різних оцінок їх параметрів.

Допоміжна література

1. Приходько С.Б. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Приходько, С.Б. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Навіщо здійснюють підбор аналітичної моделі розподілу емпіричних даних?

2) Які розподіли зазвичай використовують для опису емпіричних даних?

3) У чому перевага розподілів Джонсона?

4) Як знайти оцінку центру вибіркового розподілу або оцінку математичного сподівання випадкової величини x за гістограмою?

5) Як знайти оцінки центральних статистичних моментів випадкової величини x за гістограмою?

6) Як обрати сім’ю розподілів Джонсона?
7) Які параметри необхідні для підбору розподілу Джонсона?
8) Як знайти оцінки параметрів функції щільності ймовірності для обраної сім’ї розподілів Джонсона в загальному випадку?
9) Як можна знайти параметри розподілу Джонсона?
10) Як перевірити адекватність знайденої аналітичної моделі вибірковому розподілу емпіричних даних?

Лабораторна робота № 3
Побудова лінійного рівняння регресії для оцінювання метрик програмного забезпечення

Мета роботи: отримати практичні навички побудови лінійного рівняння регресії для оцінювання метрик програмного забезпечення.

Завдання: виконати побудову лінійного рівняння регресії для оцінювання метрик програмного забезпечення за емпіричними даними, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Як відомо, лінійне рівняння регресії визначає залежність оцінки умовного математичного сподівання $\hat{Y}$ залежної випадкової величини Y від k факторів (незалежних змінних) $X_1, X_2, \ldots, X_k$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k, \tag{3.1}$$

де $b_0, b_1, b_2, \ldots, b_k$ – параметри лінійного регресійного рівняння (3.1), оцінки яких як правило знаходять за методом найменших квадратів за емпіричними даними шляхом рішення наступної задачі математичного програмування

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} \left[Y_i - \hat{Y}_i \right]^2 \right\}, \tag{3.2}$$

де $\boldsymbol{\theta}$ – вектор параметрів, що потребують оцінювання, $\boldsymbol{\theta} = \{b_0, b_1, b_2, \ldots, b_k\}$; $\hat{\boldsymbol{\theta}}$ – вектор оцінок параметрів, $\hat{\boldsymbol{\theta}} = \{\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k\}$; N – кількість точок емпіричних даних; Y_i – i-те значення випадкової величини Y; $\hat{Y}_i$ – i-те значення оцінки $\hat{Y}$, $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}$.

Рівняння (3.1) називають множинним (багатофакторним) лінійним рівнянням регресії. Для оцінювання його параметрів замість методу найменших квадратів (3.2) можуть використовуватися інші методи, наприклад, метод максимальної правдоподібності.

У разі однофакторного лінійного рівняння регресії

$$\hat{Y} = b_0 + b_1 X_1 \tag{3.3}$$

знаходження оцінок параметрів рівняння регресії (3.3) за методом найменших квадратів (3.2) можна спростити і обчислювати за наступними формулами:

$$\hat{b}_1 = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(X_{1i} - \overline{X}_1\right)}{\sum_{i=1}^{N}\left(X_{1i} - \overline{X}_1\right)^2}; \tag{3.4}$$

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}_1, \tag{3.5}$$

де $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$; $\overline{X}_1 = \frac{1}{N}\sum_{i=1}^{N} X_{1i}$.

Якість регресійного рівняння та регресійної моделі, як правило, перевіряють за коефіцієнтом детермінації R^2 (the coefficient of determination), середньою величиною відносної помилки *MMRE* (Mean Magnitude of Relative Error) і відсотком прогнозованих результатів *PRED* (Percentage of Prediction), для яких величини відносної помилки *MRE* (Magnitude of Relative Error) менші за 0,25, *PRED*(0,25). Ці показники зазвичай використовуються для оцінювання якості прогнозування за допомогою регресійних моделей і в інженерії програмного забезпечення.

Вибірковий коефіцієнт детермінації R^2 визначається як

$$R^2 = 1 - \frac{SS_E}{SS_T}, \tag{3.6}$$

де SS_E – сума квадратів залишків (the residual sum of squares) або сума квадратів помилок (the error sum of squares), $SS_E = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$; SS_T – загальна сума квадратів (the total sum of squares), $SS_T = \sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$. Нагадаємо, що $SS_T = SS_R + SS_E$. Тут SS_R – сума квадратів регресії (the regression sum of squares), $SS_R = \sum_{i=1}^{N}\left(\hat{Y}_i - \overline{Y}\right)^2$.

Коефіцієнт детермінації R^2 інтерпретується як частка мінливості у спостережуваній залежній змінній Y, що пояснюється моделлю лінійної регресії. Значення R^2 вказує наскільки емпіричні дані підтверджують математичну модель або, чи краща наша модель за

вибіркове середнє. З формули (3.6) ми бачимо, $0 \le R^2 \le 1$. Значення R^2 вважається прийнятним, якщо воно більше за 0,75. Велике значення R^2 свідчить про те, що модель успішно пояснює мінливість *Y*. Коли R^2 малий, це може свідчити про те, що потрібно знайти альтернативну модель, яка може врахувати більшу мінливість *Y*.

Значення величини відносної похибки *MRE* обчислюється як

$$MRE_i = \left| \frac{\left(Y_i - \hat{Y}_i\right)}{Y_i} \right|. \tag{3.7}$$

Середня величина відносної похибки *MMRE* визначається як

$$MMRE = \frac{1}{N} \sum_{i=1}^{N} MRE_i \,. \tag{3.8}$$

Зазвичай $MMRE \le 0{,}25$ вважається прийнятною точністю моделі.

Значення $PRED(0{,}25)$ обчислюється як

$$PRED(0{,}25) = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} 1 & if\ MRE_i \le 0{,}25 \\ 0 & otherwise. \end{cases} \tag{3.9}$$

Зазвичай $PRED(0{,}25) \ge 0{,}75$ вважається прийнятною точністю прогнозування за допомогою регресійних моделей.

Часто корисно перевірити гіпотези щодо нахилу (slope) та перетину (intercept) в моделі лінійної регресії. Припущення про нормальність щодо помилок моделі і, отже, щодо залежної змінної *Y* продовжує застосовуватися у подальшому.

Стандартними похибками *se* (the standard errors) нахилу та перетину при однофакторній лінійній регресії є

$$se\left(\hat{b}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \tag{3.10}$$

та

$$se\left(\hat{b}_0\right) = \hat{\sigma} \sqrt{\frac{1}{N} + \frac{\overline{X}^2}{S_{XX}}} \tag{3.11}$$

відповідно. Тут $\hat{\sigma}^2 = SS_E / (N-2)$, $\overline{X}_1 = \frac{1}{N} \sum_{i=1}^{N} X_{1_i}$, $S_{X_1X_1} = \sum_{i=1}^{N} \left(X_{1_i} - \overline{X}_1\right)^2$.

Перевірити нульову гіпотезу про те, що нахил або параметр b_1 дорівнює константі, скажімо, 0 можна за допомогою тестової статистики

$$T_0 = \frac{\hat{b}_1}{se\left(\hat{b}_1\right)}, \tag{3.12}$$

яка має *t*-розподіл з *N*–2 ступенями свободи. Нульова гіпотеза відхиляється, якщо обчислене значення статистики (3.12), таке, що $|T_0| > t_{\alpha/2,N-2}$. Тут $t_{\alpha/2,N-2}$ – квантиль *t*-розподілу Стьюдента з *N*–2 ступенями свободи та $\alpha/2$ рівнем значущості.

Як і в попередньому випадку, перевірити нульову гіпотезу про те, що перетин або параметр b_0 дорівнює константі, скажімо, 0 можна за допомогою тестової статистики

$$T_0 = \frac{\hat{b}_0}{se\left(\hat{b}_0\right)}, \tag{3.13}$$

Нульова гіпотеза відхиляється, якщо обчислене значення статистики (3.13), таке, що $|T_0| > t_{\alpha/2,N-2}$.

Ці дві гіпотези стосуються значущості лінійної регресії. Неможливість відхилити нульову гіпотезу про те, що нахил або параметр b_1 дорівнює 0, рівнозначно висновку про відсутність лінійної залежності між X_1 та Y.

Підкреслимо, Т-тест на значимість регресії тісно пов'язаний з F-тестом ANOVA. Якщо нульова гіпотеза щодо значення регресії, $H_0 : b_1 = 0$, відповідає дійсності, SS_R/σ^2 є випадковою величиною з розподілом χ^2 з 1 ступенем свободи. Ми також можемо показати, що SS_E/σ^2 є випадковою величиною з розподілом χ^2 із *N*–2 ступенями свободи для однофакторної лінійної регресії, і що SS_E та SS_R незалежні. Нульова гіпотеза відхиляється, якщо значення статистики

$$F_0 = \frac{MS_R}{MS_E}, \tag{3.14}$$

таке, що $F_0 > F_{\alpha,1,N-2}$. Тут $MS_R = \frac{SS_R}{1}$; $MS_E = \frac{SS_E}{N-2}$; $F_{\alpha,1,N-2}$ – квантиль *F*-розподілу Стьюдента з α рівнем значущості та 1 і *N*–2 ступенями

свободи. Якщо нульова гіпотеза відхиляється, то існує лінійна залежність між X_1 та Y.

Нагадаємо, лінійна регресійна модель має наступний вигляд:

$$Y = \hat{Y} + \varepsilon = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \varepsilon, \qquad (3.15)$$

де ε – випадкова величина з розподілом Гаусу, яка характеризує відхилення, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$. Тобто однією із умов застосування лінійної регресії є нормальність розподілу відхилення ε в (3.15). Тому для теоретичного обґрунтування можливості застосування лінійної регресії потребує перевірки нульова гіпотеза про нормальність розподілу випадкової величини ε.

Визначення значень відхилення між емпіричними даними та лінійним рівнянням регресії здійснюється виходячи з (3.15) як

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i} - \ldots - \hat{b}_k X_{ki}. \qquad (3.16)$$

Зауважимо, для теоретичного обґрунтування можливості застосування лінійної регресії окрім нормальності розподілу відхилення ε потребують виконання ще інші умови, наприклад, відсутності гетероскедастичності (heteroscedasticity) або неоднорідності у емпіричних даних, що пов'язано з неоднаковою дисперсією випадкової величини ε в регресійній моделі при різних значеннях факторів. Наявність гетероскедастичності випадкових помилок ε призводить до неефективності оцінок, отриманих за допомогою методу найменших квадратів.

Також слід вказати ще на одну проблему, яка може виникнути при побудові множинного (багатофакторного) рівняння регресії – це наявність мультиколінеарності між факторами (предикторами). Під мультиколінеарністю розуміють наявність лінійної залежності між двома або більше незалежними змінними (факторами) у регресійній моделі. Мультиколінеарність – це негативне явище множинного регресійного аналізу, яке не дозволяє здійснити оцінювання окремого впливу кожного фактору на залежну змінну. Наявність мультиколінеарності свідчить про те, що в множинній регресійній моделі два або більше факторів (незалежних змінні) пов'язані між собою або мають високий ступінь кореляції. Тому при побудові множинного рівняння лінійної регресії потрібно перевірити майбутні фактори на наявність мультиколінеарності.

Наявність мультиколінеарності, як правило, визначають за коефіцієнтами впливу дисперсії (Variance Inflation Factors, VIFs) серед майбутніх предикторів (факторів) в моделі множинної лінійної регресії. Для лінійної моделі множинної регресії з k-предикторами X_i, $i = 1,\ldots, k$, VIFs – це діагональні елементи оберненої коваріаційної матриці kxk k-предикторів. 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 більш точного відображення впливу додавання іншого фактору (регресора) до множинної регресійної моделі можна використати скориговану статистику R^2. Відкоригований коефіцієнт множинної детермінації для моделі множинної регресії з k регресорами становить

$$R^2_{\text{Adjusted}} = 1 - \frac{(N-1)SS_E}{(N-k-1)SS_T} = \frac{(N-1)R^2 - k}{(N-k-1)}. \qquad (3.17)$$

Скоригована статистика R^2 (3.17) істотно зменшує звичайну статистику R^2, беручи до уваги кількість факторів в моделі. Загалом, скоригована статистика R^2 не завжди збільшується, коли до моделі додається змінна. Скоригований R^2 збільшиться лише в тому випадку, якщо додавання нової незалежної змінної призведе до досить великого зменшення залишкової суми квадратів, щоб компенсувати втрату одного залишкового ступеня свободи.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Перевірка факторів множинної лінійної регресії на наявність мультиколінеарності. У разі, якщо для певних факторів є наявність мультиколінеарності, то вони повинні бути відкинуті, і рівняння лінійної регресії будується для зменшеної кількості факторів. Для однофакторної лінійної регресії цей етап не виконується, виконання лабораторної роботи починається з другого етапу.

2) Визначення значень оцінок параметрів лінійного рівняння регресії для оцінювання метрик програмного забезпечення.

3) Визначення значень коефіцієнту детермінації R^2, середньої величини відносної похибки *MMRE* та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25).

4) Побудова лінії регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

5) Визначення значень відхилення між емпіричними даними та лінійним рівнянням регресії (лінією регресії).

6) Перевірка гіпотези про нормальність закону розподілу значень відхилення між емпіричними даними та лінією регресії для довірчої ймовірності 0,95.

7) Висновки про якість оцінювання метрик програмного забезпечення за побудованим лінійним рівнянням регресії та можливість теоретичного обґрунтування застосування побудованого лінійного рівняння регресії.

Завдання для самостійної роботи

Перевірити Т-тест та F-тест на значимість лінійної регресії.

Побудувати аналітичну модель закону розподілу значень відхилення між емпіричними даними та лінією регресії.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни "Обробка експериментальних даних на комп`ютері" / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Що визначає лінійне рівняння регресії?

2) У чому полягає різниця між лінійним рівнянням регресії і лінійною регресійною моделлю?

3) За якими методами можна знаходити оцінки параметрів лінійного регресійного рівняння?

4) Поясніть суть методу найменших квадратів для оцінювання параметрів лінійного рівняння регресії.

5) Сформулюйте задачу математичного програмування для оцінювання параметрів лінійного рівняння регресії.

6) Які показники зазвичай використовуються для оцінювання якості прогнозування за допомогою регресійних моделей в інженерії програмного забезпечення?

7) Як обчислюється значення коефіцієнту детермінації?

8) На що вказує значення коефіцієнту детермінації?

9) В яких межах може змінюватися значення коефіцієнту детермінації?

10) Що означає значення коефіцієнту детермінації 0 (0,2; 0,8; 1)?

11) Як обчислюється середня величина відносної похибки?

12) Яка середня величина відносної похибки вважається прийнятною для моделі?

13) Як обчислюється відсоток прогнозування за моделлю на рівні величини відносної похибки, який дорівнює 0,25?

14) Який відсоток прогнозування на рівні величини відносної похибки, що дорівнює 0,25, вважається прийнятним для моделі?

15) Навіщо потребує перевірки нульова гіпотеза про нормальність розподілу значень відхилення між емпіричними даними та лінійним рівнянням регресії?

16) Поясніть поняття мультиколінеарності між факторами у множинній лінійній регресії.

17) Навіщо при побудові множинного рівняння лінійної регресії потрібно перевірити майбутні фактори на наявність мультиколінеарності?

18) Як визначають наявність мультиколінеарності між факторами у множинній лінійній регресії?

Лабораторна робота № 4

Визначення довірчих інтервалів та інтервалів передбачення лінійної регресії для оцінювання метрик програмного забезпечення

Мета роботи: отримати практичні навички визначення довірчих інтервалів та інтервалів передбачення лінійної регресії для оцінювання метрик програмного забезпечення.

Завдання: визначити довірчі інтервали та інтервали передбачення лінійної регресії для оцінювання метрик програмного забезпечення за емпіричними даними, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Довірчий інтервал (confidence interval) регресії – це такий інтервал, в якому із заданою довірчою імовірністю можна чекати значення оцінки умовного математичного сподівання $\hat{Y}$ залежної випадкової величини Y або більш стисле визначення – це інтервальна оцінка $\hat{Y}$. Фактично довірчий інтервал визначає точність оцінки $\hat{Y}$ за рівнянням регресії.

Інтервал передбачення (prediction interval) регресії – це такий інтервал, в якому із заданою довірчою імовірністю можна чекати на появу значення залежної випадкової величини Y. Фактично інтервал передбачення регресії визначає той інтервал, в якому із заданою довірчою імовірністю знаходяться значення залежної випадкової величини Y, і вихід за який вважається викидом.

Довірчий інтервал лінійної регресії визначається як

$$\hat{Y}_i \pm t_{\alpha/2,\nu} S_Y \left\{ \frac{1}{N} + \left(\mathbf{x}^+\right)^T \left[\left(\mathbf{X}^+\right)^T \mathbf{X}^+ \right]^{-1} \left(\mathbf{x}^+\right) \right\}^{1/2}, \qquad (4.1)$$

а інтервал передбачення лінійної регресії – як

$$\hat{Y}_i \pm t_{\alpha/2,\nu} S_Y \left\{ 1 + \frac{1}{N} + \left(\mathbf{x}^+\right)^T \left[\left(\mathbf{X}^+\right)^T \mathbf{X}^+ \right]^{-1} \left(\mathbf{x}^+\right) \right\}^{1/2}. \qquad (4.2)$$

Тут $\hat{Y}_i$ – результат передбачення за лінійним регресійним рівнянням (3.1); $\mathbf{X}^+$ – матриця центрованих факторів (регресорів), яка містить значення

$X_{1_i} - \overline{X}_1,\ X_{2_i} - \overline{X}_2,\ \ldots,\ X_{k_i} - \overline{X}_k;\ S_Y^2 = \dfrac{1}{\nu}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2,\ \nu = N - k - 1;\ \left(\mathbf{X}^+\right)^T\mathbf{X}^+$ – $k \times k$ матриця

$$\left(\mathbf{X}^+\right)^T\mathbf{X}^+ = \begin{pmatrix} S_{X_1X_1} & S_{X_1X_2} & \ldots & S_{X_1X_k} \\ S_{X_1X_2} & S_{X_2X_2} & \ldots & S_{X_2X_k} \\ \ldots & \ldots & \ldots & \ldots \\ S_{X_1X_k} & S_{X_2X_k} & \ldots & S_{X_kX_k} \end{pmatrix}, \tag{4.3}$$

де $S_{X_qX_r} = \sum_{i=1}^{N}\left[X_{q_i} - \overline{X}_q\right]\left[X_{r_i} - \overline{X}_r\right],\ q,r = 1,2,\ldots,k$.

У разі однофакторної лінійної регресії (3.3) довірчий інтервал (4.1) можна записати як

$$\hat{Y}_i \pm t_{\alpha/2,N-2}S_Y\sqrt{\frac{1}{N} + \frac{\left(X_{1_i} - \overline{X}_1\right)^2}{S_{X_1X_1}}}, \tag{4.4}$$

а інтервал передбачення (4.2) – як

$$\hat{Y}_i \pm t_{\alpha/2,N-2}S_Y\sqrt{1 + \frac{1}{N} + \frac{\left(X_{1_i} - \overline{X}_1\right)^2}{S_{X_1X_1}}}. \tag{4.5}$$

де $t_{\alpha/2,N-2}$ – квантиль *t*-розподілу Стьюдента з *N*–2 ступенями свободи та $\alpha/2$ рівнем значущості; $S_Y^2 = \dfrac{1}{N-2}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$; $\overline{X}_1 = \dfrac{1}{N}\sum_{i=1}^{N}X_{1_i}$; $S_{X_1X_1} = \sum_{i=1}^{N}\left(X_{1_i} - \overline{X}_1\right)^2$.

В (4.4) та (4.5) значення $\hat{Y}_i$ визначається за однофакторним лінійним рівнянням регресії (3.3) в залежності від значення фактору X_{1_i}.

Зауважимо, що формули (4.1), (4.2), (4.4) та (4.5) побудовані виходячи з припущення нормальності закону розподілу залежної величини *Y*. У разі коли закон розподілу не є нормальним, ці формули як правило дають хибні результати. Для уникнення цього потрібно переходити до нелінійних регресійних моделей.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Визначення довірчих інтервалів лінійної регресії для оцінювання метрик програмного забезпечення для довірчої ймовірності 0,95.

2) Визначення інтервалів передбачення лінійної регресії для оцінювання метрик програмного забезпечення для довірчої ймовірності 0,95.

3) Побудова лінії регресії, довірчих інтервалів, інтервалів передбачення лінійної регресії для оцінювання метрик програмного забезпечення та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

Завдання для самостійної роботи

Визначити довірчі інтервали та інтервали передбачення лінійної регресії для оцінювання метрик програмного забезпечення для довірчої ймовірності 0,9 та порівняти з результатами, що отримані у цій лабораторній роботі для довірчої ймовірності 0,95.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Що таке довірчий інтервал регресії?

2) Що визначає довірчий інтервал?

3) Що таке інтервал передбачення регресії?

4) Що означає вихід значення залежної випадкової величини за інтервал передбачення регресії?

5) Як визначити довірчий інтервал лінійної регресії у разі декількох факторів?

6) Як визначити інтервал передбачення лінійної регресії у разі декількох факторів?

7) Як визначити довірчий інтервал однофакторної лінійної регресії?

8) Як визначити інтервал передбачення однофакторної лінійної регресії?

9) До чого призводить невиконання припущення про нормальність закону розподілу залежної величини у разі застосування формул (4.1), (4.2), (4.4) та (4.5)?

Лабораторна робота № 5
Побудова нелінійного рівняння регресії для оцінювання метрик програмного забезпечення

Мета роботи: отримати практичні навички побудови нелінійного рівняння регресії для оцінювання метрик програмного забезпечення.

Завдання: виконати побудову нелінійного рівняння регресії для оцінювання метрик програмного забезпечення за емпіричними даними, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Як відомо, існують чотири основні припущення, які обґрунтовують використання лінійних регресійних моделей, одним з яких є нормальність розподілу похибок, і, отже, нормальність залежної змінної *Y*. Але для реальних даних це припущення виконується лише в поодиноких випадках. Що призводить до необхідності побудови нелінійних регресійних моделей.

На сьогодні для побудови нелінійних регресійних рівнянь та моделей використовуються методи на основі простого перебору, лінеаризуючих та нормалізуючих перетворень. Для уникнення простого перебору при побудові нелінійних регресійних рівнянь використовують методи, що базуються на нормалізуючих взаємо-зворотних перетвореннях як одновимірних, так і багатовимірних. Методи на основі нормалізуючих перетворень не потребують використання процедури простого перебору, а, по-друге, при застосуванні бієктивних (bijective) перетворень не відбувається втрата частини інформації як, наприклад, при лінеаризації.

Будь-який метод для побудови нелінійних регресійних рівнянь на основі нормалізуючих перетворень складається з трьох етапів. На першому етапі негаусівські дані нормалізують, тобто перетворюють у гаусівські дані із застосуванням взаємо-зворотного нормалізуючого перетворення (бажано із застосуванням багатовимірного перетворення). Далі на другому етапі будується лінійне регресійне рівняння для нормалізованих даних. І на третьому етапі за лінійним регресійним рівнянням для нормалізованих даних із застосуванням обраного взаємо-

зворотного нормалізуючого перетворення отримують нелінійне регресійне рівняння.

Перш ніж скористатися відповідним методом для побудови нелінійних регресійних рівнянь на основі нормалізуючих перетворень потрібно визначитися з таким перетворенням. Нехай існує взаємо-зворотне нормалізуюче перетворення негаусівського випадкового вектору $\mathbf{P} = \{Y, X_1, X_2, \ldots, X_k\}^T$ у гаусівський випадковий вектор $\mathbf{T} = \{Z_Y, Z_1, Z_2, \ldots, Z_k\}^T$, яке задається як

$$\mathbf{T} = \boldsymbol{\psi}(\mathbf{P}), \tag{5.1}$$

і зворотне перетворення для (5.1)

$$\mathbf{P} = \boldsymbol{\psi}^{-1}(\mathbf{T}). \tag{5.2}$$

Тут $\boldsymbol{\psi}$ – вектор, $\boldsymbol{\psi} = \{\psi_Y, \psi_1, \psi_2, \ldots, \psi_k\}^T$.

На першому етапі негаусівські дані нормалізують, тобто перетворюють у гаусівські дані із застосуванням перетворення (5.1). Далі на другому етапі будується лінійне регресійне рівняння для нормалізованих даних

$$\hat{Z}_Y = b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k. \tag{5.3}$$

І на третьому етапі за лінійним регресійним рівнянням (5.3) із застосуванням обраного взаємо-зворотного нормалізуючого перетворення (5.1) отримують нелінійне регресійне рівняння

$$\hat{Y} = \psi_Y^{-1}\left(\hat{Z}_Y\right) = \psi_Y^{-1}\left(b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k\right), \tag{5.4}$$

де ψ_Y – перша компонента вектору $\boldsymbol{\psi}$ перетворення (5.1).

Наведемо приклад побудови однофакторного нелінійного рівняння регресії у разі застосування одновимірного перетворення у вигляді десяткового логарифму до кожної змінної

$$Z_Y = \lg Y; \tag{5.5}$$

$$Z_1 = \lg X_1. \tag{5.6}$$

Спочатку емпіричні дані нормалізують за перетвореннями (5.5) і (5.6). Далі будується однофакторне лінійне регресійне рівняння для нормалізованих даних

$$\hat{Z}_Y = b_0 + b_1 Z_1. \qquad (5.7)$$

За рівнянням (5.7) отримують однофакторне нелінійне рівняння регресії

$$\hat{Y} = 10^{\hat{b}_0} X_1^{\hat{b}_1}. \qquad (5.8)$$

Рівняння виду (5.8) з відповідними параметрами застосовується у якості відомих моделей в інженерії програмного забезпечення: COCOMO (COnstructive COst MOdel) та ISBSG (International Software Benchmarking Standards Group).

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Вибір перетворення для нормалізації емпіричних даних з метрик програмного забезпечення.

2) Нормалізація емпіричних даних з метрик програмного забезпечення за обраним перетворенням.

3) Побудова лінійного рівняння регресії для нормалізованих емпіричних даних.

4) Визначення значень відхилення між нормалізованими емпіричними даними та лінійним рівнянням регресії для нормалізованих емпіричних даних.

5) Перевірка гіпотези про нормальність закону розподілу значень відхилення між нормалізованими емпіричними даними та лінійним рівнянням регресії для нормалізованих емпіричних даних для довірчої ймовірності 0,95. У разі, якщо зазначена вище гіпотеза буде відкинута, то потрібно перейти до етапу 1, інакше (у разі прийняття гіпотези) перейти до виконання наступного етапу 6.

6) Побудова нелінійного рівняння регресії за лінійним рівнянням регресії для нормалізованих емпіричних даних та обраним перетворенням.

7) Визначення значень коефіцієнту детермінації R^2, середньої величини відносної похибки *MMRE* та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25).

8) Побудова лінії регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

9) Висновок про якість оцінювання метрик програмного забезпечення за побудованим нелінійним рівнянням регресії у порівнянні з лінійним з л.р.№3.

Завдання для самостійної роботи

Виконати побудову нелінійного рівняння регресії для оцінювання метрик програмного забезпечення із використанням іншого нормалізуючого перетворення та порівняти отримані нелінійні рівняння регресії.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Що призводить до необхідності побудови нелінійних регресійних моделей?

2) Які методи використовуються для побудови нелінійних регресійних рівнянь та моделей?

3) Які методи використовують для уникнення простого перебору при побудові нелінійних регресійних рівнянь?

4) У чому перевага методів на основі нормалізучих перетворень у порівнянні з іншими методами для побудови нелінійних регресійних рівнянь?

5) З яких етапів складається будь-який метод для побудови нелінійних регресійних рівнянь на основі нормалізуючих перетворень?

6) Як побудувати однофакторне нелінійне рівняння регресії у разі застосування одновимірного перетворення у вигляді десяткового логарифму до кожної змінної?

7) Як перевірити гіпотезу про нормальність закону розподілу значень відхилення між нормалізованими емпіричними даними та лінійним рівнянням регресії для нормалізованих емпіричних даних?

8) Навіщо перевіряють гіпотезу про нормальність закону розподілу значень відхилення між нормалізованими емпіричними даними та лінійним рівнянням регресії для нормалізованих емпіричних даних?

9) Які показники зазвичай використовуються для оцінювання якості прогнозування за допомогою регресійних моделей в інженерії програмного забезпечення?

10) Як обчислюється значення коефіцієнту детермінації?

11) На що вказує значення коефіцієнту детермінації?

12) Як визначити величину відносної похибки *MMRE*?

13) Яка середня величина відносної похибки вважається прийнятною для моделі?

14) Як визначити відсоток прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25)?

15) Який відсоток прогнозування на рівні величини відносної похибки, що дорівнює 0,25, вважається прийнятним для моделі?

16) Що собою представляє скоригована статистика R^2?

17) Навіщо визначати у множинній регресійній моделі скориговану статистику R^2?

18) Як визначити відкоригований коефіцієнт множинної детермінації для моделі множинної регресії?

Лабораторна робота № 6

Побудова довірчих інтервалів та інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення

Мета роботи: отримати практичні навички визначення довірчих інтервалів та інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення.

Завдання: визначити довірчі інтервали та інтервали передбачення нелінійної регресії для оцінювання метрик програмного забезпечення за емпіричними даними, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Як було зазначено раніше, формули (4.1), (4.2), (4.4) та (4.5) для визначення довірчих інтервалів та інтервалів передбачення лінійних регресій побудовані виходячи з припущення нормальності закону розподілу залежної величини *Y*. У разі коли закон розподілу не є нормальним, ці формули як правило дають хибні результати. Для уникнення цього потрібно переходити до нелінійних регресійних моделей та застосування відповідних методів для побудови довірчих інтервалів та інтервалів передбачення нелінійних регресії.

Довірчий інтервал множинної нелінійної регресії визначається як

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{\frac{1}{N}+\left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right), \qquad (6.1)$$

а інтервал передбачення множинної нелінійної регресії – як

$$\psi_Y^{-1}\left(\hat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y}\left\{1+\frac{1}{N}+\left(\mathbf{z}_X^+\right)^T\left[\left(\mathbf{Z}_X^+\right)^T\mathbf{Z}_X^+\right]^{-1}\left(\mathbf{z}_X^+\right)\right\}^{1/2}\right). \qquad (6.2)$$

Тут $\hat{Z}_Y$ – результат передбачення за лінійним регресійним рівнянням (5.3) для нормалізованих даних; $\mathbf{Z}_X^+$ – матриця центрованих нормалізованих факторів (регресорів), яка містить значення $Z_{1_i} - \overline{Z}_1$,

$Z_{2_i} - \overline{Z}_2$, …, $Z_{k_i} - \overline{Z}_k$; $S^2_{Z_Y} = \frac{1}{\nu}\sum_{i=1}^{N}\left(Z_{Y_i} - \hat{Z}_{Y_i}\right)^2$, $\nu = N - k - 1$; $\left(\mathbf{Z}^+_X\right)^T \mathbf{Z}^+_X$ – $k \times k$ матриця

$$\left(\mathbf{Z}^+_X\right)^T \mathbf{Z}^+_X = \begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \dots & S_{Z_1Z_k} \\ S_{Z_1Z_2} & S_{Z_2Z_2} & \dots & S_{Z_2Z_k} \\ \dots & \dots & \dots & \dots \\ S_{Z_1Z_k} & S_{Z_2Z_k} & \dots & S_{Z_kZ_k} \end{pmatrix}, \tag{6.3}$$

де $S_{Z_qZ_r} = \sum_{i=1}^{N}\left[Z_{q_i} - \overline{Z}_q\right]\left[Z_{r_i} - \overline{Z}_r\right]$, $q, r = 1,2,\dots,k$.

У разі однофакторної нелінійної регресії довірчий інтервал (6.1) можна записати як

$$\psi_Y^{-1}\left(\hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_Z \sqrt{\frac{1}{N} + \frac{\left(Z_{1_i} - \overline{Z}_1\right)^2}{S_{Z_1Z_1}}}\right), \tag{6.4}$$

а інтервал передбачення (6.2) – як

$$\psi_Y^{-1}\left(\hat{Z}_{Y_i} \pm t_{\alpha/2,N-2} S_Z \sqrt{1 + \frac{1}{N} + \frac{\left(Z_{1_i} - \overline{Z}_1\right)^2}{S_{Z_1Z_1}}}\right). \tag{6.5}$$

де $t_{\alpha/2,N-2}$ – квантиль *t*-розподілу Стьюдента з *N*–2 ступенями свободи та $\alpha/2$ рівнем значущості; $S_Z^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(Z_i - \hat{Z}_i\right)^2$; $\overline{Z}_1 = \frac{1}{N}\sum_{i=1}^{N} Z_{1_i}$; $S_{Z_1Z_1} = \sum_{i=1}^{N}\left(Z_{1_i} - \overline{Z}_1\right)^2$.

В (6.4) та (6.5) значення $\hat{Z}_{Y_i}$ визначається за однофакторним лінійним рівнянням регресії (5.7) для нормалізованих даних в залежності від значення фактору Z_{1_i}.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) Визначення довірчих інтервалів лінійної регресії для нормалізованих емпіричних даних, які отримані у роботі №5, для довірчої ймовірності 0,95.

2) Визначення інтервалів передбачення лінійної регресії для нормалізованих емпіричних даних, які отримані у роботі № 5, для довірчої ймовірності 0,95.

3) Визначення довірчих інтервалів нелінійної регресії для оцінювання метрик програмного забезпечення за допомогою довірчих інтервалів лінійної регресії для нормалізованих емпіричних даних, які отримані у роботі № 5, для довірчої ймовірності 0,95 та перетворення, яке є зворотнім до перетворення, що було обрано у роботі № 5.

4) Визначення інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення за допомогою інтервалів передбачення лінійної регресії для нормалізованих емпіричних даних, які отримані у роботі № 5, для довірчої ймовірності 0,95 та перетворення, яке є зворотнім до перетворення, що було обрано у роботі № 5.

5) Побудова лінії регресії, довірчих інтервалів, інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

6) Висновок про ширини довірчих інтервалів та інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення у порівнянні з відповідними ширинами лінійної регресії з роботи 4.

Завдання для самостійної роботи

Визначити довірчі інтервали та інтервали передбачення нелінійної регресії для оцінювання метрик програмного забезпечення для довірчої ймовірності 0,9 та порівняти з результатами, що отримані у цій лабораторній роботі для довірчої ймовірності 0,95.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Що таке довірчий інтервал регресії?
2) Що визначає довірчий інтервал?
3) Що таке інтервал передбачення регресії?
4) Що означає вихід значення залежної випадкової величини за інтервал передбачення регресії?
5) Як визначити довірчий інтервал нелінійної регресії у разі декількох факторів?
6) Як визначити інтервал передбачення нелінійної регресії у разі декількох факторів?
7) Як визначити довірчий інтервал однофакторної нелінійної регресії?
8) Як визначити інтервал передбачення однофакторної нелінійної регресії?

Лабораторна робота № 7
Побудова нелінійної регресійної моделі у разі наявності викидів у емпіричних даних з метрик програмного забезпечення

Мета роботи: отримати практичні навички побудови нелінійної регресійної моделі у разі наявності викидів у емпіричних даних з метрик програмного забезпечення.

Завдання: виконати побудову нелінійної регресійної моделі у разі наявності викидів у емпіричних даних з метрик програмного забезпечення за даними, які наведені у файлі. (Примітка: файл з емпіричними даними з метрик програмного забезпечення видає викладач).

Зробити висновки щодо отриманих результатів.

Загальні теоретичні відомості

Зазвичай викид визначають як спостереження, яке настільки відхиляється від інших спостережень, що викликає підозру в тому, що воно було породжене іншим механізмом. Точки даних, які не описуються регресійною моделлю, розглядаються як викиди.

Побудову нелінійної регресійної моделі у разі наявності викидів у емпіричних даних здійснюють за наступними етапами.

На першому етапі перевіряють, чи є серед сукупності емпіричних даних такі, які можна вважати викидами. Якщо є, то вони відкидаються, і нова сукупність емпіричних даних перевіряється на наявність викидів. У разі відсутності викидів переходять далі до другого етапу.

На другому етапі будують нелінійну регресійну модель на основі нормалізуючих перетворень.

На третьому етапі визначають інтервали передбачення нелінійної регресії та перевіряють, чи є серед емпіричних даних такі, що виходять за межі інтервалу передбачення нелінійної регресії. Якщо є, то вони відкидаються. Етапи 1–3 повторюють для нових даних до відсутності викидів. За відсутності викидів нелінійна регресійна модель вважається побудованою.

Для визначення викидів у негаусівських даних на першому етапі ми рекомендуємо використовувати відповідний метод на основі квадрату відстані Махаланобіса (Mahalanobis) для нормалізованих даних. Для цього негаусівські дані нормалізують (бажано із застосуванням

багатовимірних перетворень) та для них визначають значення квадрату відстані Махаланобіса, яке для кожної точки багатовимірних даних *i*, $i = 1,2,\ldots,N$, позначається як d_i^2 і обчислюється за наступною формулою:

$$d_i^2 = (\mathbf{Z}_i - \overline{\mathbf{Z}})^T S_N^{-1} (\mathbf{Z}_i - \overline{\mathbf{Z}}), \tag{7.1}$$

де $\mathbf{Z}_i$ – *i*-та точка багатовимірних даних гаусівського вектору **Z**; $\overline{\mathbf{Z}}$ – вектор вибіркових середніх багатовимірних даних гаусівського вектору **Z**; S_N – матриця вибіркових коваріацій

$$S_N = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{Z}_i - \overline{\mathbf{Z}})(\mathbf{Z}_i - \overline{\mathbf{Z}})^T . \tag{7.2}$$

Відомо, якщо дані слідують багатовимірному нормальному розподілу, тоді розподіл квадрата відстані Махаланобіса поводиться як розподіл χ^2. Для великих $N - m$ (принаймні 25) відстань повинна поводитися приблизно як незалежні $\chi^2_{m,\alpha}$ випадкові величини. Тут $\chi^2_{m,\alpha}$ – це квантиль розподілу χ^2, α – рівень значущості. Ми приймаємо α за 0,005, як це зазвичай робиться. Значення даних, для яких значення квадрата відстані Махаланобіса більше, ніж квантиль розподілу χ^2, вважаються викидами, і ці значення відкидаються. Після відкидання зменшена кількість точок багатовимірних даних нормалізується за допомогою нормалізуючого перетворення знову, поки всі значення квадрата відстані Махаланобіса (7.1) не будуть меншими або рівними квантилю розподілу χ^2.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи.

1) З'ясування, чи є серед сукупності емпіричних даних такі, які можна вважати викидами. Якщо є, то вони відкидаються, і нова сукупність емпіричних даних перевіряється на наявність викидів. У разі відсутності викидів переходять далі до етапу 2.

2) Побудова нелінійного рівняння регресії як це робилося у роботі № 5.

3) Визначення інтервалів передбачення нелінійної регресії для оцінювання метрик програмного забезпечення (як це робилося у роботі № 6) та перевірка, чи є серед емпіричних даних такі, що виходять

за межі інтервалу передбачення нелінійної регресії. Якщо є, то вони відкидаються. Повторювати етапи 1–3 для нових даних до відсутності викидів. За відсутності викидів перейти до виконання наступного етапу 4.

4) Обчислення значень коефіцієнту детермінації R^2, середньої величини відносної похибки *MMRE* та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25).

5) Побудова лінії регресії, інтервалів передбачення нелінійної регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

6) Висновок про якість оцінювання метрик програмного забезпечення за побудованим нелінійним рівнянням регресії у порівнянні з лінійним з роботи № 3 та нелінійним з роботи № 5.

Завдання для самостійної роботи

Виконати побудову нелінійної регресійної моделі для оцінювання метрик програмного забезпечення у разі наявності викидів у емпіричних даних із використанням іншого нормалізуючого перетворення та порівняти отримані нелінійні регресійні моделі.

Допоміжна література

1. Навчально-методичні матеріали до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» / С. Б. Приходько. – Миколаїв: НУК, 2020. – 48 с.

2. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп`ютері» / С. Б. Приходько, Л. М. Макарова, К. С. Пугаченко. – Миколаїв: НУК, 2018. – 76 с.

Питання для самоконтролю

1) Що таке викид у даних?

2) Як з’ясування, чи є серед сукупності емпіричних даних такі, які можна вважати викидами?

3) Що роблять у разі знаходження викидів у даних?

4) Що таке відстань Махаланобіса?

5) Як визначають значення квадрату відстані Махаланобіса для кожної точки багатовимірних даних?

6) Що таке квантиль розподілу?

7) Від яких параметрів залежить квантиль розподілу χ^2?

8) Що є результатом множення вектору рядка на квадратну матрицю? Яка умова при цьому повинна виконуватися?

9) Що є результатом множення вектору рядка на вектор стовпчик? Яка умова при цьому повинна виконуватися?

10) Що є результатом транспонування вектору рядка?

11) Як визначити зворотну матрицю?

12) Як визначаються елементи матриці коваріацій?

ДОДАТКИ

Додаток А

Верхні 100α %-ві точки розподілу χ^2

В таблиці А1 наведені верхні 100α %-ві точки розподілу χ^2, тобто такі значення x, що $P\left(\chi^2_\nu > x\right) = \alpha$.

Таблиця А1

ν	α							
	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
1	2	3	4	5	6	7	8	9
1	$0,0^4 39$	$0,0^3 16$	$0,0^3 98$	$0,0^2 39$	3,84	5,02	6,63	7,88
2	0,010	0,0201	0,0506	0,103	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	37,65	40,65	44,31	46,93

Продовження табл. А1

1	2	3	4	5	6	7	8	9
26	11,16	12,20	13,84	15,38	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	90,53	95,02	100,4	104,2
80	51,17	53,54	57,15	60,39	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	124,3	129,6	135,8	140,2

Нижні $100\alpha\,\%$-ві точки розподілу χ^2 дорівнюють верхнім $100(1-\alpha)\%$-вим точкам.

Звернемо увагу на те, що якщо випадкова величина X має розподіл χ^2 з ν ступенями вільності і ν достатньо велико (скажімо, більше 30), то розподіл величини $Z=\sqrt{2X}-\sqrt{2\nu-1}$ є наближено нормальним з нульовим математичним сподіванням і одиничною дисперсією. Це дає змогу застосовувати таблиці нормального розподілу і попередню формулу для знаходження значення x для достатньо великих ν.

Додаток Б

Верхні 100α %-ві точки *t*-розподілу Стьюдента

В таблиці Б1 наведені верхні 100α %-ві точки *t*-розподілу Стьюдента, тобто такі значення x, що $P(t_\nu > x) = \alpha$.

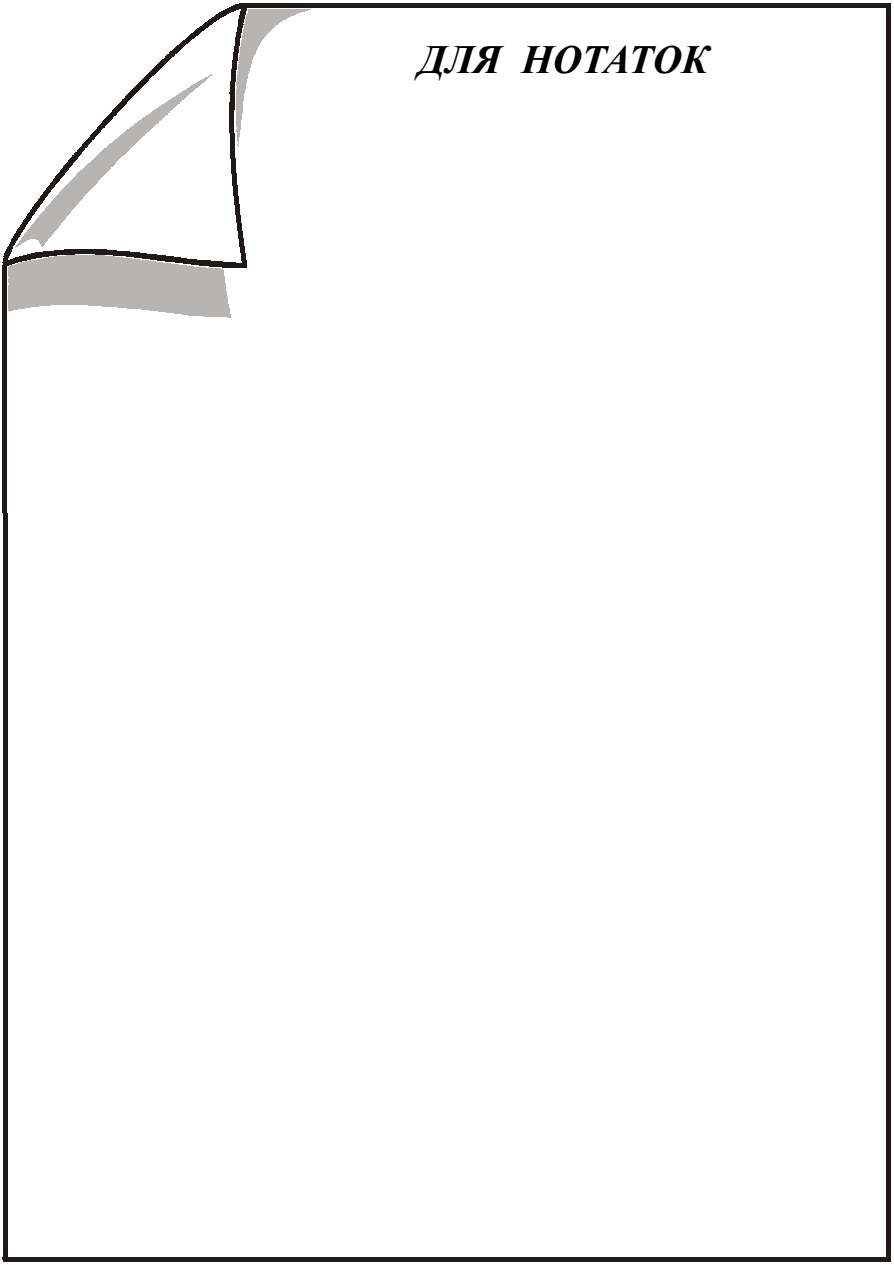
Таблиця Б1

ν	α				
	0,005	0,01	0,025	0,05	0,1
1	2	3	4	5	6
1	63,66	31,82	12,71	6,31	3,08
2	9,92	6,96	4,30	2,92	1,89
3	5,84	4,54	3,18	2,35	1,64
4	4,60	3,75	2,78	2,13	1,53
5	4,03	3,36	2,57	2,01	1,48
6	3,71	3,14	2,45	1,94	1,44
7	3,50	3,00	2,36	1,90	1,42
8	3,36	2,90	2,31	1,86	1,40
9	3,25	2,82	2,26	1,83	1,38
10	3,17	2,76	2,23	1,81	1,37
11	3,11	2,72	2,20	1,80	1,36
12	3,06	2,68	2,18	1,78	1,36
13	3,01	2,65	2,16	1,77	1,35
14	2,98	2,62	2,14	1,76	1,34
15	2,95	2,60	2,13	1,75	1,34
16	2,92	2,58	2,12	1,75	1,34
17	2,90	2,57	2,11	1,74	1,33
18	2,88	2,55	2,10	1,73	1,33
19	2,86	2,54	2,09	1,73	1,33
20	2,84	2,53	2,09	1,72	1,32
21	2,83	2,52	2,08	1,72	1,32
22	2,82	2,51	2,07	1,72	1,32
23	2,81	2,50	2,07	1,71	1,32
24	2,80	2,49	2,06	1,71	1,32
25	2,79	2,48	2,06	1,71	1,32
26	2,78	2,48	2,06	1,71	1,32
27	2,77	2,48	2,05	1,70	1,31
28	2,76	2,47	2,05	1,70	1,31

Продовження табл. Б1

1	2	3	4	5	6
29	2,76	2,47	2,04	1,70	1,31
30	2,75	2,46	2,04	1,70	1,31
40	2,70	2,46	2,02	1,68	1,30
50	2,68	2,42	2,01	1,67	1,30
60	2,66	2,40	2,00	1,67	1,30
80	2,64	2,39	1,99	1,66	1,29
100	2,63	2,37	1,98	1,66	1,29
200	2,60	2,36	1,97	1,65	1,29
500	2,59	2,34	1,96	1,65	1,28
∞	2,576	2,326	1,960	1,645	1,282

Для отримання нижніх 100α %-вих точок *t*-розподілу Стьюдента необхідно змінити знак у верхній 100α %-вій точці.


ДЛЯ НОТАТОК

Навчальне видання

ПРИХОДЬКО Сергій Борисович
МАКАРОВА Лідія Миколаївна
ПРИХОДЬКО Наталія Василівна
ПУХАЛЕВИЧ Андрій Володимирович

МЕТОДИЧНІ ВКАЗІВКИ ТА ЗАВДАННЯ
до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії»

Коректор *О. Є. Вакула*
Комп'ютерне верстання *В. В. Москаленко*

Формат 60×84/16. Ум. друк. арк. 3,6. Тираж 100 прим. Вид. № 03. Зам. № 0902-06.
Видавець і виготівник Національний університет кораблебудування
імені адмірала Макарова
просп. Героїв України, 9, м. Миколаїв, 54025
E-mail : publishing@nuos.edu.ua
Свідоцтво суб'єкта видавничої справи ДК № 6402 від 19.09.2018 р.