МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет кораблебудування
імені адмірала Макарова

С. Б. ПРИХОДЬКО

МЕТОДИЧНІ ВКАЗІВКИ ТА ЗАВДАННЯ
до виконання лабораторних робіт
з дисципліни
«МЕТОДИ АНАЛІЗУ РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ»

Рекомендовано Методичною радою НУК

ВИДАВНИЦТВО
НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
КОРАБЛЕБУДУВАННЯ
ІМ. АДМІРАЛА МАКАРОВА

2026

УДК 004.412:519.237.5
П77

Автор С. Б. Приходько, д-р техн. наук, професор

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

Рекомендовано Методичною радою НУК

Приходько С. Б.

П77 Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Методи аналізу результатів експериментальних досліджень» / С. Б. Приходько. – Миколаїв : НУК, 2026. – 62 с.

Вміщено стислий виклад теоретичних відомостей, методичні вказівки й завдання до виконання лабораторних робіт, етапи виконання робіт та контрольні питання.

Призначено для здобувачів третього (освітньо-наукового) рівня вищої освіти спеціальності 122 «Комп'ютерні науки». Також можуть бути корисні магістрам і аспірантам інших спеціальностей, дослідження яких пов'язані з аналізом як одновимірних, так і багатовимірних даних.

УДК 004.412:519.237.5

ЗМІСТ

ПЕРЕДМОВА

РОЗРОБЛЕНО: кафедрою програмного забезпечення автоматизованих систем (ПЗАС) Національного університету кораблебудування імені адмірала Макарова (НУК) Міністерства освіти і науки України.

РОЗРОБНИК: С. Б. Приходько, д-р техн. наук, професор.

РЕКОМЕНДОВАНО: Методичною радою НУК.

ВСТУП

Підготовка фахівців, які здатні продукувати нові ідеї, розв'язувати комплексні науково-прикладні задачі та/або проблеми в галузі професійної та/або дослідницько-інноваційної діяльності в галузі інформаційних технологій за спеціальністю 122 «Комп'ютерні науки», потребує знань і вмінь, у тому числі з методів аналізу результатів експериментальних досліджень.

У Національному університеті кораблебудування імені адмірала Макарова (НУК) дисципліна «Методи аналізу результатів експериментальних досліджень» вивчається здобувачами вищої освіти спеціальності 122 «Комп'ютерні науки» у третьому семестрі. Вона відноситься до циклу обов'язкових дисциплін під час підготовки докторів філософії. Згідно з робочою програмою на вивчення дисципліни «Методи аналізу результатів експериментальних досліджень» відводиться 3 кредити або 90 годин, з яких: 15 годин – лекції, 15 годин – лабораторні заняття та 60 годин – самостійна робота.

Ураховуючи таку невелику кількість годин, які відведені на аудиторні заняття, як лекційні, так і лабораторні, згідно із зазначеною вище робочою програмою до виконання пропонується чотири лабораторні роботи. Виконання цих лабораторних робіт дозволить здобувачам вищої освіти набути практичних навичок з проведення: початкового аналізу одновимірних та багатовимірних даних; лінійного та нелінійного регресійного аналізу багатовимірних даних, які отримані за результатами експериментальних досліджень та мають у тому числі розподіл, який відхиляється від нормального.

Зазначимо, що під час виконання лабораторних робіт здобувач вищої освіти може запропонувати свої дані, які були одержані за результатами власних експериментальних досліджень, що пов'язані з темою майбутньої дисертації. Але попередньо ці дані потрібно узгодити з викладачем, який проводить заняття з дисципліни «Методи аналізу результатів експериментальних досліджень».

Ці методичні вказівки мають допомогти студентам другого курсу спеціальності 122 «Комп'ютерні науки» у підготовці та виконанні лабораторних робіт з дисципліни «Методи аналізу результатів експериментальних досліджень». Вони містять завдання для лабораторних робіт, стислий виклад теоретичних відомостей, етапи виконання робіт та контрольні питання.

Під час виконання кожної лабораторної роботи рекомендується:

- засвоїти теоретичні відомості;
- розібратися із завданням та етапами виконання роботи;
- виконати запропоновані завдання та оформити звіт;
- перевірити повноту розуміння теми за допомогою контрольних питань, розташованих у кінці кожної роботи.

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

Лабораторна робота № 1

Початковий аналіз одновимірних даних

Мета роботи: набути практичних навичок початкового аналізу одновимірних даних (первинної їх обробки) з використанням комп'ютера.

Завдання: виконати початковий аналіз одновимірних даних, які наведені у файлі. (Файл зі значеннями випадкової величини, залежно від варіанта, видає викладач. Студент може запропонувати свої дані, попередньо узгодивши їх з викладачем).

Загальні теоретичні відомості

При прямих вимірах первинну обробку одновимірних експериментальних даних виконують у наступному порядку [1, 2]:

1. Визначають вибіркове середнє $\overline{x}$ (точкову оцінку математичного сподівання) N результатів спостережень x_i $\left(i = 1, 2, \ldots, N\right)$

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

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

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

3. Визначають оцінку середнього квадратичного відхилення результату виміру

$$\hat{\sigma}_{\overline{x}} = \frac{\hat{\sigma}_x}{\sqrt{N}} \, .$$

4. Перевіряють нормальність закону розподілу результатів спостережень.

5. Для заданого значення довірчої ймовірності знаходять довірчу похибку результату виміру та довірчий інтервал для середнього квадратичного відхилення.

6. Визначають наявність аномальних значень, грубих помилок і промахів та, якщо останні знайдені, відповідні результати відкидають і повторюють обчислення.

У випадку непрямих вимірів шукане значення випадкової величини, яка вимірюється, обчислюється за результатами прямих вимірів величин, зв'язаних з шуканою визначеною функціональною залежністю.

Окремий результат у випадку багаторазового прямого виміру фізичної величини із-за наявності випадкових похибок являє собою випадкову величину.

Нагадаємо деякі відомості з теорії ймовірностей. Під випадковою величиною розуміють величину, яка в результаті досліду з випадковим результатом бере 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 імовірності, з якими ці значення можуть появитися.

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

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 являє собою ймовірність

того, що випадкова величина X візьме значення менше ніж задане x:

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

Якщо функція розподілу $F(x)$ випадкової величини X за будь-якого 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) зі 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 під час проведення вимірів. Так, випадкові похибки багатократних вимірів зазвичай 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 під час первинної обробки експериментальних даних перевіряють нормальність закону розподілу результатів спостережень.

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

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

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

Кількість підінтервалів (клітинок), на які розбивається інтервал $\left[x_{\min}, x_{\max}\right]$ 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 й Карузера).

Імовірність того, що значення випадкової величини X потрапляють 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 залежно від рівня значимості α та кількості ступенів свободи ν за таблицею верхніх 100α %-вих точок розподілу χ^2 (таблиця А1 в дод. А) знаходять значення $\chi^2_{кр}$. Якщо $\chi^2 \le \chi^2_{кр}$, то з імовірністю $1 - \alpha$ можна прийняти гіпотезу про те, що закон розподілу результатів спостережень є нормальним, якщо $\chi^2 > \chi^2_{кр}$ – цю гіпотезу потрібно відкинути.

Кількість ступенів свободи 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 $p_д$ того, що $\left|\Delta\hat{\theta}\right|$ перейде деяку границю $\Delta > 0$:

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

Інтервал від $\hat{\theta} - \Delta$ до $\hat{\theta} + \Delta$, у якому з імовірністю $p_д$ знаходиться справжнє значення θ, називається *довірчим інтервалом*, а його границі – *довірчими границями*, імовірність $p_д$ – *довірчою ймовірністю.*

Величина $\alpha = 1 - p_д$ загалом 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 точкової оцінки математичного сподівання визначається як

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

де t_{N-1} – квантиль t-розподілу Стьюдента, визначається за таблицею верхніх 100α %-вих точок t-розподілу Стьюдента (таблиця Б1 у дод. Б) залежно 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 виміру. Вона зменшується зі збільшенням N.

Приклад 1.1. Знайти 95 %-вий довірчий інтервал точкової оцінки математичного сподівання нормальної сукупності із 10-ти значень випадкової величини, якщо $\overline{x} = 52$ і $S_x^2 = 96$.

За таблицею Б1 у дод. Б для рівня значимості $\alpha/2 = (1-0,95)/2 = 0,025$ і кількості ступенів свободи $\nu = 1-9$ визначають квантиль t-розподілу Стьюдента $t_9(0,025) = 2,26$. Тоді 95 %-вий довірчий інтервал точкової оцінки математичного сподівання визначається як $\left[52 - 2,26 \cdot 9,80/\sqrt{10}, 52 + \right.$ $\left. + 2,26 \cdot 9,80/\sqrt{10}\right]$, 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 залежно від α та ν за таблицею верхніх 100α %-вих точок розподілу χ^2 (таблиця А1 у дод. 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-ти значень випадкової величини, якщо $S_x = 0,130$.

За таблицею А1 у дод. А 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 результат досліду x^* є грубою помилкою, якщо значення x^* виходить за границі інтервалу цензурування вибірки

$$\left[\overline{x}-\Delta_x, \overline{x}+\Delta_x\right]. \tag{1.1}$$

Кращий результат одержують у разі пошуку Δ_x за Граббом (*Grubb*)

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

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

Але більшість існуючих методів визначення границь цензурування вибірки за (1.1), у тому числі із застосуванням (1.2), розраховані на гаусівський розподіл даних.

У разі якщо генеральна сукупність не є нормальною, значення Δ_x може бути обчислено за певною наближеною формулою, наприклад, як

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

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

$$\hat{\mu}_4(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^4.$$

Але формула (1.3) та подібні їй не враховують можливу асиметрію закону розподілу негаусівських даних. У [3] запропоновано статистичний метод визначення аномалій (грубих помилок) в емпіричних даних на основі нормалізуючих перетворень, суть якого полягає в наступному: спочатку значення

негаусівської випадкової величини x перетворюють у такі, що мають нормальний розподіл, за бієктивним перетворенням

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

Далі за значенням вибіркового середнього $\overline{x}$ випадкової величини X обчислюємо

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

Потім для нормалізованих даних знаходимо значення Δ_z за (1.2) лише з тою різницею, що у (1.2) замість S_x потрібно підставити S_z – вибіркове середнє квадратичне відхилення випадкової величини Z.

Далі подібно (1.1) визначаємо границі інтервалу

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

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

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

За (1.6) можна визначати границі цензурування вибірки значень випадкової величини X, розподіл якої відхиляється від гаусівського (нормального). Усі значення x, які виходять за інтервал (1.6) вважаються аномальними або грубими помилками.

Для визначення довірчого інтервалу вибіркового середнього випадкової величини, розподіл якої відхиляється від нормального, застосовують формули (1.4)–(1.6) лише з тою різницею, що у (1.5) та (1.6) замість Δ_z потрібно підставити $\overline{\Delta}_z$, яке визначається за наступною формулою [4]:

$$\overline{\Delta}_z = t_{N-1} S_z / \sqrt{N},$$

де t_{N-1} – квантиль t-розподілу Стьюдента, визначається за таблицею верхніх $100\alpha\,\%$-вих точок t-розподілу Стьюдента

(таблиця Б1 у дод. Б) залежно від рівня значимості $\alpha/2$ та кількості ступенів свободи ν, $\nu = N - 1$.

Загалом як перетворення (1.4), ми рекомендуємо використовувати перетворення Джонсона. Також у певних випадках можна використовувати інші нормалізуючі перетворення, наприклад, перетворення десяткового логарифма або Бокса – Кокса (*Box – Cox*).

Сім'ї розподілів Джонсона одержані шляхом перетворення нормованої нормально розподіленої випадкової величини z. Загалом перетворення має вигляд [1, 2, 5]:

$$z = \gamma + \eta q\left(x, \varphi, \lambda\right);\ \eta > 0;\ -\infty < \gamma < \infty;\ \lambda > 0;\ -\infty < \varphi < \infty, \quad (1.7)$$

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 в напряму осі абсцис.

Джонсон запропонував три різні сім'ї функцій q:

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

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

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

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

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

Ця функція щільності ймовірності зростає від φ до точки максимуму і потім більш повільно спадає в 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

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

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 аргумента, є унімодальними та їх моди лежать між медіаною та нулем. Тим самим ці розподіли мають додатну або від'ємну асиметрію залежно 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 рис. 1.1 подана 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.

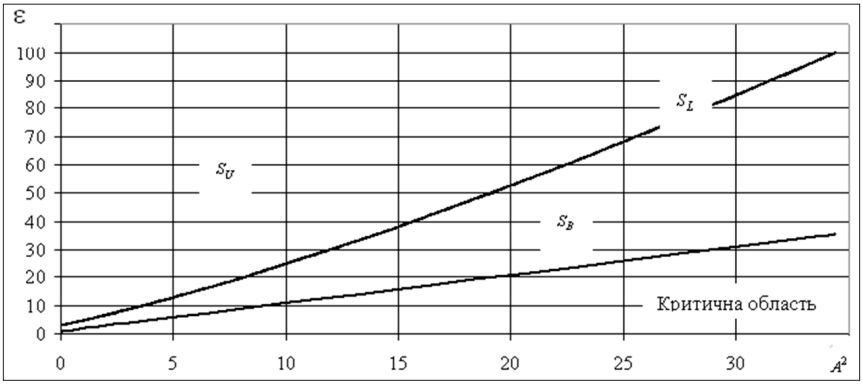

Рис. 1.1. Комбінації A^2 і ε для різних розподілів Джонсона [2]

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, які лежать нижче лінії S_L (до лінії критичної області) – розподілом із сім'ї S_B. Якщо комбінація A^2 і ε попадає у 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]:

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

Оцінки параметрів перетворення Джонсона краще визначати за методом максимальної правдоподібності

$$\hat{\theta} = \arg\max_{\theta} l(\theta), \tag{1.8}$$

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

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

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

Для перетворення Джонсона сім'ї S_B, логарифмічну функцію правдоподібності можна записати як [6]

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

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

Зауважте, під час оцінювання параметрів γ, η, λ, та φ за (1.8) потрібно враховувати відповідні обмеження, що накладаються на ці параметри для обраної сім'ї розподілів Джонсона. Початкові значення параметрів γ, η, λ, та φ під час їх оцінювання за (1.8) ми рекомендуємо або обирати за відповідними обмеженнями, що накладаються на ці параметри для обраної сім'ї розподілів Джонсона (наприклад, сім'ї S_B), або із рішення системи рівнянь для статистичних моментів для розподілу Джонсона сім'ї S_U [1, 2].

В окремих випадках можна використовувати інші нормалізуючі перетворення, наприклад, десяткового логарифма

$$Z = \lg X$$

або перетворення Бокса – Кокса

$$Z = x(\lambda) = \begin{cases} \left(X^{\lambda} - 1\right)/\lambda, & \text{якщо} \quad \lambda \neq 0; \\ \ln(X), & \text{якщо} \quad \lambda = 0. \end{cases}$$

Тут Z – це випадкова змінна з розподілом Гауса (нормалізована змінна); λ – параметр перетворення Бокса – Кокса.

Для оцінювання параметра одновимірного перетворення Бокса – Кокса може бути застосований метод максимальної правдоподібності (ММП) шляхом максимізації логарифма функції правдоподібності

$$l(\lambda) = C - \frac{N}{2}\ln\sum_{i=1}^{N}\frac{\left(x_i(\lambda) - \overline{x}(\lambda)\right)^2}{N} + (\lambda - 1)\sum_{i=1}^{N}\ln(x_i), \quad (1.9)$$

де C – постійна, яку визначають з умови нормування,

$$\overline{x}(\lambda) = \sum_{i=1}^{N} x_i(\lambda) / N.$$

Зауважте, що для довільних даних ми не можемо заздалегідь з'ясувати, чи зможемо ми нормалізувати ці дані за допомогою перетворення Бокса – Кокса або десяткового логарифма, на відміну від перетворення Джонсона.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи:

1. Визначення вибіркового середнього $\overline{x}$ (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) та точкової оцінки середнього квадратичного відхилення $\hat{\sigma}_x$.

3. Побудову гістограми за вибіркою значень випадкової величини X. Побудову нормального закону розподілу

експериментальних даних (функції щільності ймовірності) на гістограмі.

4. Перевірку нормальності закону розподілу значень випадкової величини X.

5. Знаходження довірчого інтервалу для вибіркового середнього випадкової величини X.

6. Визначення наявності аномалій, грубих помилок і промахів (якщо останні знайдені, відповідні результати відкидають і повторюють обчислення).

Завдання для самостійної роботи

Визначте довірчий інтервал для вибіркового середнього випадкової величини X для довірчої ймовірності 0,9 та порівняйте з відповідним результатом, що одержаний у цій лабораторній роботі для довірчої ймовірності 0,95.

Контрольні питання

1. Навіщо виконують первинну обробку експериментальних даних при прямих вимірах?
2. Назвіть послідовність виконання первинної обробки експериментальних даних.
3. Що таке випадкова величина?
4. Що таке функція розподілу випадкової величини?
5. Що таке щільність імовірності 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 за великої (малої) вибірки?

12. Що таке довірчий інтервал, довірча ймовірність, рівень значимості?

13. Як визначають для нормальної генеральної сукупності α %-ві довірчі границі точкової оцінки математичного сподівання?

14. Як визначають для нормальної генеральної сукупності α %-ві довірчі границі точкової оцінки середнього квадратичного відхилення?

15. Як перевіряють наявність аномалій та грубих похибок в одновимірних даних, закон розподілу яких є нормальним?

16. Що таке інтервал цензурування вибірки?

17. Як можна перевірити наявність аномалій та грубих похибок в одновимірних даних, закон розподілу яких не є нормальним?

18. Що роблять у разі, якщо у виборці знайдені аномалії та/або грубі похибки (промахи)?

Лабораторна робота № 2

Початковий аналіз багатовимірних даних

Мета роботи: набути практичних навичок початкового аналізу багатовимірних даних (первинної їх обробки) з використанням комп'ютера.

Завдання: виконати початковий аналіз багатовимірних даних, які наведені у файлі. (Файл зі значеннями випадкової величини, залежно від варіанта, видає викладач. Студент може запропонувати свої дані, попередньо узгодивши їх з викладачем).

Загальні теоретичні відомості

Можна провести певну аналогію між початковим аналізом одновимірних та багатовимірних даних. Початковий аналіз багатовимірних даних виконують у наступному порядку:

1. Визначають вектори оцінок статистичних моментів (часто обмежуються визначенням вектора вибіркових середніх).
2. Визначають вибіркову коваріаційну матрицю.
3. Перевіряють відхилення розподілу багатовимірних даних від нормального.
4. Для заданого значення довірчої ймовірності знаходять довірчий еліпсоїд вибіркових середніх (або довірчий еліпс вибіркових середніх для двовимірних даних).
5. Визначають наявність багатовимірних викидів та, якщо останні знайдені, відповідні результати відкидають і повторюють обчислення.

Вектор вибіркових середніх $\bar{X}$:

$$\bar{X} = \left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k\right)^T,$$

де $\bar{X}_r = \frac{1}{N}\sum_{i=1}^{N} X_{r_i}$, $r = 1, 2, \ldots, k$; N – кількість точок даних.

Однією із важливіших характеристик випадкового вектора X є коваріаційна матриця. У подальшому k-вимірний $(k \times 1)$ вектор X будемо визначати як $X = \left(X_1, X_2, \ldots, X_k\right)^T$. Тут $X_1, X_2, \ldots, X_k$ – це випадкові величини.

Оцінкою коваріаційної матриці Σ є вибіркова коваріаційна матриця S_N:

$$S_N = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T, \qquad (2.1)$$

де $\bar{X}$ – це вектор вибіркових середніх, $\bar{X} = \left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k\right)^T$; $\bar{X}_r = \frac{1}{N}\sum_{i=1}^{N} X_{r_i}$, $r = 1, 2, \ldots, k$; N – кількість точок даних.

Вибіркову коваріаційну матрицю S_N можна також записати як

$$S_N = \begin{pmatrix} S_{X_1X_1} & S_{X_1X_2} & \ldots & S_{X_1X_k} \\ S_{X_1X_2} & S_{X_2X_2} & \ldots & S_{X_2X_k} \\ \ldots & \ldots & \ldots & \ldots \\ S_{X_1X_k} & S_{X_2X_k} & \ldots & S_{X_kX_k} \end{pmatrix}, \qquad (2.2)$$

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

Багатовимірний розподіл Гауса (багатовимірний нормальний розподіл) у теорії ймовірностей та математичній статистиці – це узагальнення одновимірного розподілу Гауса для випадку із багатьма вимірами.

$$f_X\left(x\right) = \frac{1}{\sqrt{\left(2\pi\right)^k \left|\Sigma\right|}} e^{-\frac{1}{2}\left(\mathrm{x}-\mathrm{m_x}\right)^T \Sigma^{-1}\left(\mathrm{x}-\mathrm{m_x}\right)}, \qquad (2.3)$$

де $\left|\Sigma\right| = det\Sigma$ – це детермінант коваріаційної матриці Σ, відомий також як узагальнена дисперсія (*generalized variance*).

Як ми бачимо з (2.3), можна провести певну аналогію між одновимірним та багатовимірним розподілами Гауса.

Історично склалося так, що багато статистичних критеріїв, методів і оцінок розроблені тільки для випадку нормального початкового розподілу. Тому під час первинної обробки багатовимірних даних також перевіряють відхилення від нормальності закону розподілу цих даних. Так у разі відсутності відхилення розподілу багатовимірних даних від нормального як довірчий регіон для компонентів вектора вибіркових середніх застосовують довірчий еліпсоїд вибіркових середніх.

Довірчий еліпсоїд вибіркових середніх визначає певний регіон, у якому із заданою довірчою ймовірністю будуть знаходитися всі можливі значення компонентів вектора вибіркових середніх. Рівняння довірчого еліпсоїда вибіркових середніх базується на статистиці Хотеллинга (*Hotelling's statistic*) T^2 та може бути представлено як

$$\left(\bar{X}-\widehat{m}_x\right)^T S_N^{-1}\left(\bar{X}-\widehat{m}_x\right)=T^2/N\,.$$

Для випадку однієї випадкової величини отримуємо

$$\bar{X}=\bar{x}\pm TS_x/\sqrt{N}\,.$$

Застосувавши замість T, квантиль t-розподілу Стьюдента t_{N-1}, маємо

$$\bar{X}=\bar{x}\pm t_{N-1}S_x/\sqrt{N}\,,$$

де t_{N-1} – квантиль t-розподілу Стьюдента, визначається за таблицею верхніх 100α %-вих точок t-розподілу Стьюдента залежно від рівня значимості α/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.

Рівняння довірчого еліпсоїда вибіркових середніх зазвичай записують у наступній формі:

$$\left(\bar{X}-\widehat{m}_x\right)^T S_N^{-1}\left(\bar{X}-\widehat{m}_x\right)=\frac{k\left(N-1\right)}{N\left(N-k\right)}F_{k,N-k,\alpha}\,, \qquad (2.4)$$

де $F_{k,N-k,\alpha}$ – квантиль F-розподілу з k та $N-k$ ступенями свободи; α – це рівень значущості; S_N – вибіркова коваріаційна матриця.

У разі відсутності відхилення розподілу двовимірних даних від нормального, як довірчий регіон для компонент вектора вибіркових середніх застосовують довірчий еліпс вибіркових середніх. Рівняння довірчого еліпсу вибіркових середніх зазвичай записують у наступній формі:

$$\left(\bar{X}-\widehat{m}_x\right)^T S_N^{-1}\left(\bar{X}-\widehat{m}_x\right)=\frac{2\left(N-1\right)}{N\left(N-2\right)}F_{2,N-2,\alpha}, \qquad (2.5)$$

де $F_{2,N-2,\alpha}$ – квантиль F-розподілу з 2 та $N-2$ ступенями свободи; α – це рівень значущості; S_N – вибіркова коваріаційна матриця.

Як ми бачимо, рівняння довірчого еліпсу вибіркових середніх (2.5) є поодиноким випадком рівняння довірчого еліпсоїда вибіркових середніх (2.4). Довірчі еліпсоїд та еліпс не слід плутати з еліпсоїдом та еліпсом прогнозування (передбачення), які застосовують, у тому числі для визначення наявності викидів у разі відсутності відхилення розподілу відповідно багатовимірних та двовимірних даних, від нормального. Довірчі еліпсоїд та еліпс будуть розглянуті далі.

Для перевірки відхилення розподілу багатовимірних даних від нормального зазвичай використовують міри Мардіа (*Mardia*) багатовимірної асиметрії та ексцесу, оскільки вони найчастіше включені в різноманітні програмні пакети. Ще в 1970 році Мардія визначив багатовимірну асиметрію та багатовимірний ексцес, відповідно, як

$$\beta_{1,k}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\left(X_i-\bar{X}\right)^T S_N^{-1}\left(X_j-\bar{X}\right)\right]^3, \qquad (2.6)$$

$$\beta_{2,k}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(X_i-\bar{X}\right)^T S_N^{-1}\left(X_i-\bar{X}\right)\right]^2, \qquad (2.7)$$

де X – $k\times 1$ вектор випадкових величин, $X=\left(X_1,X_2,\ldots,X_k\right)^T$, а S_N – зміщена вибіркова коваріаційна матриця (2.1). Вибіркову коваріаційну матрицю S_N можна також записати як (2.2).

Обидва показники (2.6) та (2.7) мають індекс k, тому вони специфічні для набору змінних k. Хоча слід зазначити, що інколи для спрощення індекс k не ставлять.

Очікувана асиметрія Мардіа дорівнює 0 для багатовимірного нормального розподілу, а вищі значення вказують на більш серйозне відхилення від нормального закону. Очікуваний ексцес Мардіа становить $k(k+2)$ для багатовимірного нормального розподілу k-змінних. Для двовимірного нормального розподілу очікуваний ексцес Мардіа дорівнює 8.

Для $\beta_{1,k}$ ми використовуємо тестову статистику

$$\frac{N}{6}\beta_{1,k}, \tag{2.8}$$

яка має наближений розподіл χ^2 з $k(k+1)(k+2)/6$ ступенями свободи та рівнем значущості α.

Для $\beta_{2,k}$, як тестової статистики $z_{2,k}$, використовується 1-α квантиль нормального закону розподілу з математичним сподіванням $k(k+2)$ та дисперсією $8k(k+2)/N$.

Щоб перевірити відхилення розподілу від нормального, потрібно порівняти значення статистики (2.8) з квантилем $\chi^2_{k(k+1)(k+2)/6,\alpha}$, а $\beta_{2,k}$ з $z_{2,k}$. Якщо значення статистики (2.8) більше за квантиль $\chi^2_{k(k+1)(k+2)/6,\alpha}$ або, якщо значення $\beta_{2,k}$ більше за $z_{2,k}$, то багатовимірний розподіл даних не є нормальним. Під час перевірки відхилення розподілу від нормального рекомендують брати α, що дорівнює 0,005. Підкреслимо, як і в одновимірному випадку, незначущість тестових статистик для асиметрії та ексцесу не обов'язково означає нормальність.

Під час перевірки відхилення розподілу двовимірних даних ($k=2$) від нормального, як тестової статистики $z_{2,k}$, для $\beta_{2,k}$ використовується 0,995-квантиль нормального закону розподілу з математичним сподіванням 8 та дисперсією $64/N$, а для тесту за $\beta_{1,k}$ – квантиль $\chi^2_{4,0,005}$.

Зважаючи на те, що практично у всіх методах визначення викидів у багатовимірних даних у разі застосування

припущення про їх нормальність, використовується квадрат відстані Махаланобіса. Спочатку розглянемо це поняття.

Відстань Махаланобіса (*the Mahalanobis distance*) – це міра відстані між точкою P і середнім D, з урахуванням кореляції між багатовимірними даними, введена Махаланобісом (*Mahalanobis*) у 1936 році.

Це багатовимірне узагальнення ідеї вимірювання того, на скільки точка P відхиляється від середнього значення D. Ця відстань дорівнює нулю для P за середнього D і зростає, коли P віддаляється від середнього вздовж осі кожної головної компоненти. Якщо кожну з цих осей повторно масштабувати, щоб мати одиницю дисперсії, то відстань Махаланобіса відповідає стандартній евклідовій відстані в перетвореному просторі. Таким чином, відстань Махаланобіса є безрозмірною, масштабно-інваріантною та враховує кореляції набору даних.

Квадрат відстані Махаланобіса (*the squared Mahalanobis distance*), який для кожної точки багатовимірних даних i, $i = 1, 2, \ldots, N$, позначається за допомогою d_i^2, визначається як

$$d_i^2 = \left(X - \bar{X}\right)^T S_N^{-1} \left(X - \bar{X}\right), \tag{2.9}$$

де X – $k \times 1$ вектор випадкових величин, $X = \left(X_1, X_2, \ldots, X_k\right)^T$; $\bar{X}$ – це вектор вибіркових середніх, $\bar{X} = \left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k\right)^T$; $\bar{X}_r = \frac{1}{N} \sum_{i=1}^{N} X_{r_i}$, $r = 1, 2, \ldots, k$; N – кількість точок даних; S_N – вибіркова коваріаційна матриця, яка має вигляд (2.2).

Рівняння еліпсоїда прогнозування зазвичай записують у наступній формі:

$$\left(X - \bar{X}\right)^T S_N^{-1} \left(X - \bar{X}\right) = \chi_{k,\alpha}^2, \tag{2.10}$$

або

$$\left(X - \bar{X}\right)^T S_N^{-1} \left(X - \bar{X}\right) = \frac{k\left(N^2 - 1\right)}{N\left(N - k\right)} F_{k,N-k,\alpha}. \tag{2.11}$$

У (2.10) та (2.11) X – $k \times 1$ вектор випадкових величин, $X = \left(X_1, X_2, \ldots, X_k\right)^T$; $\bar{X}$ – це вектор вибіркових середніх, $\bar{X} = \left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k\right)^T$; $\bar{X}_r = \frac{1}{N}\sum_{i=1}^{N} X_{r_i}$, $r = 1, 2, \ldots, k$; N – кількість точок даних; $\chi^2_{k,\alpha}$ – квантиль χ^2-розподілу з k ступенями свободи; $F_{k,N-k,\alpha}$ – квантиль F-розподілу з k та $N-k$ ступенями свободи; α – це рівень значущості; S_N – вибіркова коваріаційна матриця, яка має вигляд (2.2). Для визначення викидів у багатовимірних даних зазвичай беруть рівень значущості α, що дорівнює 0,005.

Зауважте, що ліва частина рівнянь (2.10) та (2.11) є квадратом відстані Махаланобіса (2.9). Тому для визначення викидів у багатовимірних даних за (2.10) потрібно зробити таку перевірку. Якщо для точки багатовимірних даних i значення d_i^2 буде більше за $\chi^2_{k,\alpha}$, то для рівня значущості α ця точка даних i є багатовимірним викидом та повинна бути видалена із сукупності даних. Якщо ми маємо ситуацію, коли одночасно декілька точок даних мають значення d_i^2 більше за $\chi^2_{k,\alpha}$, то відкидати треба одну точку з найбільшим значенням d_i^2. Після цього перевірку на наявність викидів у багатовимірних даних потрібно продовжити, але вже без відкинутої точки.

Для визначення викидів у багатовимірних даних за (2.11) потрібно зробити таку перевірку. Якщо для точки багатовимірних даних i значення d_i^2 буде більше за $\frac{k\left(N^2-1\right)}{N\left(N-k\right)} F_{k,N-k,\alpha}$, то для рівня значущості α ця точка даних i є багатовимірним викидом та повинна бути видалена із сукупності даних. Якщо ми маємо ситуацію, коли одночасно декілька точок даних мають значення d_i^2 більше за $\frac{k\left(N^2-1\right)}{N\left(N-k\right)} F_{k,N-k,\alpha}$, то відкидати треба одну точку з найбільшим значенням d_i^2. Після цього перевірку

на наявність викидів у багатовимірних даних потрібно продовжити, але вже без відкинутої точки.

Квадрат відстані Махаланобіса та еліпсоїд прогнозування можуть застосовуватися для визначення викидів у багатовимірних даних лише в разі, коли їх розподіл є нормальним (гаусівським).

Але зазвичай розподіл великої кількості наборів багатовимірних даних, наприклад, з метрик програмного забезпечення, не є гаусівським (нормальним). Визначення викидів у багатовимірних даних, розподіл яких не є гаусівським, може бути виконано за допомогою еліпсоїда прогнозування для нормалізованих даних. А визначення викидів у двовимірних даних, розподіл яких не є гаусівським, може бути виконано за допомогою еліпса прогнозування для нормалізованих даних.

У разі якщо компоненти вектора випадкових величин X, $X = \left(X_1, X_2\right)^T$, мають нульову кореляцію, то для їх нормалізації можуть бути застосовані одновимірні взаємозворотні перетворення, наприклад, десяткового логарифма, Бокса – Кокса (*Box – Cox*) або Джонсона (*Johnson*). Так, у нашому випадку перетворення десяткового логарифма для двох змінних можна записати як

$$Z_1 = \lg X_1; \tag{2.12}$$

$$Z_2 = \lg X_2, \tag{2.13}$$

а перетворення Бокса – Кокса буде таким:

$$Z_j = x\left(\lambda_j\right) = \begin{cases} \left(X_j^{\lambda_j} - 1\right) \Big/ \lambda_j, & \text{якщо} \quad \lambda_j \neq 0; \\ \ln\left(X_j\right), & \text{якщо} \quad \lambda_j = 0. \end{cases} \tag{2.14}$$

У (2.14) Z_j – це випадкова змінна з розподілом Гауса (нормалізована змінна), як і у (2.12) та (2.13), λ_j – параметр перетворення Бокса – Кокса, $j = 1,2$.

Для оцінювання параметра одновимірного перетворення Бокса – Кокса для кожної змінної може бути застосований

метод максимальної правдоподібності (ММП) шляхом максимізації логарифма функції правдоподібності (1.9). Для знаходження оцінок параметрів λ_j за ММП у нашому випадку в (1.9) замість λ потрібно підставити відповідний параметр λ_j, $j=1,2$, а також – відповіді значення випадкової величини X_1 або X_2.

У разі якщо компоненти вектора X випадкових величин X_1 та X_2, мають ненульову кореляцію (або мають взаємний вплив одна на одну), то для їх нормалізації краще застосовувати багатовимірні взаємозворотні перетворення, наприклад, Бокса – Кокса або Джонсона. Так, у нашому випадку компоненти двовимірного перетворення Бокса – Кокса також можна визначати за (2.14). Але для оцінювання параметрів двовимірного перетворення Бокса – Кокса за ММП потрібно вже використовувати іншу логарифмічну функцію правдоподібності:

$$l(X,,) = \sum_{j=1}^{k}(»_j - 1)\sum_{i=1}^{N}\ln\left(X_{j_i}\right) - \frac{N}{2}\ln\left[\det\left(S_N\right)\right], \quad (2.15)$$

де $k=2$; $, = \{»_1, »_2\}$; S_N – вибіркова коваріаційна матриця для нормалізованих даних

$$S_N = \frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)\left(Z_i - \bar{Z}\right)^T, \quad (2.16)$$

де Z – це гаусівський випадковий вектор, $Z = \{Z_1, Z_2\}^T$; $\bar{Z}$ – це вектор вибіркових середніх, $\bar{Z} = \{\bar{Z}_1, \bar{Z}_2\}^T$.

Як ми бачимо, у (2.15) взаємний вплив випадкових величин X_1 та X_2 ураховується завдяки коваріаційній матриці (2.16).

Після нормалізації даних треба перевірити відхилення їх двовимірного розподілу від нормального. Для цього можна скористатися тестом Мардіа. Якщо двовимірний розподіл нормалізованих даних не буде наближений до гаусівського, то потрібно використати для нормалізації даних інше перетворення. У цьому випадку ми рекомендуємо застосовувати двовимірне перетворення Джонсона.

Визначення викидів у двовимірних даних, розподіл яких є гаусівським, може бути здійснено за допомогою еліпса прогнозування, рівняння якого зазвичай записують у формі

$$\left(Z-\bar{Z}\right)^T S_N^{-1}\left(Z-\bar{Z}\right)=\frac{2\left(N^2-1\right)}{N\left(N-2\right)}F_{2,N-2,\pm}, \qquad (2.17)$$

де $F_{2,N-2,\alpha}$ – квантиль F-розподілу з 2 та $N-2$ ступенями свободи; α – це рівень значущості; S_N – вибіркова коваріаційна матриця (2.16).

Вихід точки нормалізованих даних за границі еліпса прогнозування буде означати, що ця точка є двовимірним викидом. До того ж як значення α рекомендується брати 0,005.

Зауважте, що ліва частина рівняння (2.17) є квадратом відстані Махаланобіса, який для кожної точки двовимірних даних i, $i=1,2,\ldots,N$, позначається за допомогою d_i^2 і визначається як

$$d_i^2=\left(Z_i-\bar{Z}\right)^T S_N^{-1}\left(Z_i-\bar{Z}\right).$$

Відомо, що тестова статистика для d_i^2 може бути створена наступним чином:

$$N\left(N-2\right)d_i^2\Big/2\left(N^2-1\right),$$

що має приблизний F-розподіл з 2 і $N-2$ ступенями свободи.

Таким чином, якщо для точки нормалізованих двовимірних даних i значення d_i^2 буде більше за $\frac{2\left(N^2-1\right)}{N\left(N-2\right)}F_{2,N-2,\alpha}$, то для рівня значущості α ця точка даних i є двовимірним викидом та повинна бути видалена із сукупності даних. Якщо ми маємо ситуацію, коли одночасно декілька точок даних мають значення d_i^2 більше за $\frac{2\left(N^2-1\right)}{N\left(N-2\right)}F_{2,N-2,\alpha}$, то відкидати треба

одну точку з найбільшим значенням d_i^2. Після цього перевірку на наявність викидів у двовимірних даних потрібно продовжити, але вже без відкинутої точки. До того ж потрібно виконувати нову нормалізацію даних, але вже для зменшеної кількості точок.

Визначення викидів у багатовимірних даних, розподіл яких не є гаусівським, може бути виконано за допомогою еліпсоїда прогнозування для нормалізованих даних. За структурою рівняння еліпсоїда прогнозування буде як і (2.17) з тою різницею, що відповідні вектори Z та $\overline{Z}$ будуть містити замість двох k компонент, а замість $\frac{2\left(N^2-1\right)}{N\left(N-2\right)}F_{2,N-2,\alpha}$ у правій частині (2.17) потрібно підставити $\frac{k\left(N^2-1\right)}{N\left(N-k\right)}F_{k,N-k,\alpha}$. До того ж як значення α рекомендується брати 0,005.

Ураховуючи зазначене, якщо для точки нормалізованих k-вимірних даних i значення d_i^2 буде більше за $\frac{k\left(N^2-1\right)}{N\left(N-k\right)}F_{k,N-k,\alpha}$, то для рівня значущості α ця точка даних i є багатовимірним (k-вимірним) викидом та повинна бути видалена із сукупності даних. Якщо ми маємо ситуацію, коли одночасно декілька точок даних мають значення d_i^2 більше за $\frac{k\left(N^2-1\right)}{N\left(N-k\right)}F_{k,N-k,\alpha}$, то відкидати треба одну точку з найбільшим значенням d_i^2. Після цього перевірку на наявність викидів у багатовимірних даних потрібно продовжити, але вже без відкинутої точки. До того ж потрібно виконувати нову нормалізацію даних, але вже для зменшеної кількості точок. Після нормалізації даних треба перевірити відхилення їх багатовимірного розподілу від нормального. Для цього можна скористатися тестом Мардіа. Якщо k-вимірний розподіл

нормалізованих даних не буде наближений до гаусівського, то потрібно використати для нормалізації даних інше перетворення. У цьому разі ми рекомендуємо застосовувати багатовимірне перетворення Джонсона.

Таким чином відбувається очищення багатовимірних даних, розподіл яких не є гаусівським (нормальним).

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи:

1. Визначення вектора вибіркових середніх.
2. Визначення вибіркової коваріаційної матриці.
3. Перевірку відхилення розподілу багатовимірних даних від нормального.
4. Знаходження довірчого еліпсоїда вибіркових середніх (або довірчого еліпса вибіркових середніх для двовимірних даних) для значення довірчої ймовірності 0,95.
5. Визначення наявності багатовимірних викидів та, якщо останні знайдені, відповідні результати відкидають і повторюють обчислення.

Завдання для самостійної роботи

Знайдіть довірчий еліпсоїд вибіркових середніх (або довірчий еліпс вибіркових середніх для двовимірних даних) для значення довірчої ймовірності 0,9 та порівняйте з відповідним результатом, що одержаний у цій лабораторній роботі для довірчої ймовірності 0,95.

Контрольні питання

1. Для чого виконують початковий аналіз (первинну обробку) багатовимірних даних?
2. Назвіть послідовність виконання початкового аналізу (первинної обробки) багатовимірних даних.
3. Що таке узагальнена дисперсія (*generalized variance*)?

4. Від яких параметрів залежить багатовимірний нормальний розподіл (розподіл Гауса)?

5. Що характеризує коваріаційна матриця?

6. Яку певну аналогію можна провести між одновимірним та багатовимірним розподілами Гауса?

7. Для чого визначають довірчий еліпс вибіркових середніх?

8. Для чого перевіряють відхилення розподілу багатовимірних даних від нормального?

9. Які критерії застосовують для перевірки відхилення розподілу багатовимірних даних від нормального?

10. Як перевіряють наявність багатовимірних викидів у разі, коли закон розподілу даних є нормальним?

11. Як перевіряють наявність багатовимірних викидів у разі, коли закон розподілу даних відхиляється від нормального?

12. Що роблять у разі, якщо в даних знайдені викиди?

Лабораторна робота № 3

Лінійний регресійний аналіз багатовимірних даних

Мета роботи: набути практичних навичок лінійного регресійного аналізу багатовимірних даних.

Завдання:

1. Здійснити перевірку факторів множинної лінійної регресії на наявність мультиколінеарності (у разі однофакторного рівняння регресії цей пункт не виконується).
2. Виконати побудову лінійного рівняння регресії для оцінювання залежної величини за багатовимірними даними, що наведені у файлі.
3. Здійснити перевірку значущості лінійної регресії.
4. Визначити значення коефіцієнта детермінації R^2, середньої величини відносної похибки MMRE та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25).
5. Визначити довірчі інтервали та інтервали прогнозування лінійної регресії для довірчої ймовірності 0,95.
6. Зробити побудову лінії регресії, довірчих інтервалів, інтервалів прогнозування лінійної регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).
7. Перевірити гіпотезу про нормальність закону розподілу значень відхилення між емпіричними даними та лінією регресії для довірчої ймовірності 0,95.
8. Зробити висновки щодо одержаних результатів.

Загальні теоретичні відомості

Регресійний аналіз є важливим статистичним методом для аналізу даних. Регресійний аналіз вивчає вплив кількох незалежних змінних (факторів, предикторів, регресорів) на значення залежної змінної. Регресійний аналіз відповідає на наступні питання:

1. Які фактори мають найбільше значення?

2. Які фактори ми можемо ігнорувати?

3. Як ці фактори взаємодіють один з одним?

4. Наскільки ми впевнені в усіх обраних факторах та у результатах оцінювання (прогнозування) залежної змінної?

Лінійний регресійний аналіз використовується для створення лінійної моделі, яка описує зв'язок між залежною змінною та однією чи кількома незалежними змінними. Залежно від наявності однієї чи кількох незалежних змінних розрізняють простий (*simple*) і множинний (*multiple*) лінійний регресійний аналіз.

Насамперед у лінійному регресійному аналізі визначають лінійну регресію (лінійне регресійне рівняння), а також її (його) статистичну значущість, щоб оцінити вплив кожного фактора на залежну змінну.

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

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

де $b_0, b_1, b_2, \ldots, b_k$ – параметри лінійного регресійного рівняння (3.1), оцінки яких зазвичай знаходять за методом найменших квадратів за емпіричними даними шляхом розв'язання наступної задачі математичного програмування:

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

де θ – вектор параметрів, що потребують оцінювання, $\theta = \{b_0, b_1, b_2, \ldots, b_k\}$; $\hat{\theta}$ – вектор оцінок параметрів, $\hat{\theta} = \{\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k\}$; 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.

При однофакторному лінійному рівнянні 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 й обчислювати за наступними формулами:

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

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

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

Якість регресійного рівняння та регресійної моделі зазвичай 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).

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*), $SS_T = \sum_{i=1}^{N} \left(Y_i - \bar{Y} \right)^2$. Нагадаємо, що $SS_T = SS_R + SS_E$. Тут SS_R – сума квадратів регресії (*the regression sum of squares*), $SS_R = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y} \right)^2$.

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, $0 \le R^2 \le 1$. 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

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

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

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

Стандартними похибками *se* (*the standard errors*) нахилу та перетину за однофакторної лінійної регресії є

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

та

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

відповідно.

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

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 $\left|T_0\right|>t_{\alpha/2,N-2}$. 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, T-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 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:

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

де ε – випадкова величина з розподілом Гауса, яка характеризує відхилення, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$. Тобто однією із умов застосування лінійної регресії є нормальність розподілу відхилення ε у (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

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

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 за різних значень факторів. Наявність гетероскедастичності випадкових помилок ε призводить до неефективності оцінок, одержаних за допомогою методу найменших квадратів.

Також слід указати ще на одну проблему, яка може виникнути під час побудови 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 фактора на залежну змінну. Наявність мультиколінеарності свідчить про те, що у множинній регресійній моделі два або більше факторів (незалежних змінних) пов'язані між собою або мають високий ступінь кореляції. Тому під час побудови множинного рівняння лінійної регресії потрібно перевірити майбутні фактори на наявність мультиколінеарності.

Наявність мультиколінеарності зазвичай визначають за коефіцієнтами впливу дисперсії (*Variance Inflation Factors, VIFs*) серед майбутніх предикторів (факторів) у 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 $k \times k$ k-предикторів. Значення VIFs більше за 10 часто сприймаються як сигнал, що дані мають

проблеми з мультиколінеарністю. У разі якщо значення VIFs знаходяться в межах від 1 до 5, то мультиколінеарності немає.

Для більш точного відображення впливу додавання іншого фактора (регресора) до множинної регресійної моделі можна використати скориговану статистику R^2. Відкоригований коефіцієнт множинної детермінації для моделі множинної регресії з k-регресорами становить

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

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

Довірчий інтервал (*confidence interval*) регресії – це такий інтервал, у якому із заданою довірчою ймовірністю 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 прогнозування (*prediction interval*) регресії – це такий інтервал, у якому із заданою довірчою ймовірністю можна чекати на появу значення залежної випадкової величини Y. Фактично інтервал прогнозування регресії визначає той інтервал, у якому із заданою довірчою ймовірністю знаходяться значення залежної випадкової величини Y, і вихід за який вважається викидом.

Довірчий інтервал лінійної регресії визначається як

$$\hat{Y}_i \pm t_{\pm/2,\nu} S_Y \left\{ \frac{1}{N} + \left(x^+\right)^T S_{XX}^{-1} \left(x^+\right) \right\}^{1/2}, \quad (3.18)$$

а інтервал прогнозування лінійної регресії – як

$$\hat{Y}_i \pm t_{\pm/2,\nu} S_Y \left\{ 1 + \frac{1}{N} + \left(x^+ \right)^T S_{XX}^{-1} \left(x^+ \right) \right\}^{1/2} . \qquad (3.19)$$

Тут $\hat{Y}_i$ – результат прогнозування (передбачення) за лінійним регресійним рівнянням (3.1); x^+ – вектор центрованих факторів (регресорів), який містить значення $X_{1_i} - \bar{X}_1$, $X_{2_i} - \bar{X}_2$, …, $X_{k_i} - \bar{X}_k$; $S_Y^2 = \frac{1}{\nu} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i \right)^2$, $\nu = N - k - 1$; S_{XX} – $k \times k$ матриця

$$S_{XX} = \begin{pmatrix} S_{X_1X_1} & S_{X_1X_2} & \dots & S_{X_1X_k} \\ S_{X_1X_2} & S_{X_2X_2} & \dots & S_{X_2X_k} \\ \dots & \dots & \dots & \dots \\ S_{X_1X_k} & S_{X_2X_k} & \dots & S_{X_kX_k} \end{pmatrix}, \qquad (3.20)$$

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

У разі однофакторної лінійної регресії (3.3) довірчий інтервал (3.18) можна записати як

$$\hat{Y}_i \pm t_{\alpha/2,N-2} S_Y \sqrt{\frac{1}{N} + \frac{\left(X_{1_i} - \bar{X}_1 \right)^2}{S_{X_1X_1}}} , \qquad (3.21)$$

а інтервал передбачення (3.19) – як

$$\hat{Y}_i \pm t_{\alpha/2,N-2} S_Y \sqrt{1 + \frac{1}{N} + \frac{\left(X_{1_i} - \bar{X}_1 \right)^2}{S_{X_1X_1}}} . \qquad (3.22)$$

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

У (3.21) та (3.22) значення $\hat{Y}_i$ визначається за однофакторним лінійним рівнянням регресії (3.3) залежно від значення фактора X_{1_i}.

Зауважимо, що формули (3.18), (3.19), (3.21) та (3.22) побудовані, виходячи з припущення нормальності закону розподілу залежної величини *Y*. У разі коли закон розподілу не є нормальним, ці формули зазвичай дають хибні результати. Для уникнення цього потрібно переходити до нелінійних регресійних моделей.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи:

1. Перевірку 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 роботи починається з другого етапу.
2. Визначення значень оцінок параметрів лінійного рівняння регресії для оцінювання метрик програмного забезпечення.
3. Перевірку значущості лінійної регресії.
4. Визначення значень коефіцієнта детермінації R^2, середньої величини відносної похибки MMRE та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, *PRED*(0,25).
5. Визначення довірчих інтервалів лінійної регресії для довірчої ймовірності 0,95.
6. Визначення інтервалів прогнозування лінійної регресії для довірчої ймовірності 0,95.
7. Побудову лінії регресії, довірчих інтервалів, інтервалів прогнозування лінійної регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).
8. Визначення значень відхилення між емпіричними даними та лінійним рівнянням регресії (лінією регресії).

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

10. Висновки за результатами виконання лабораторної роботи.

Завдання для самостійної роботи

Визначте довірчі інтервали та інтервали прогнозування лінійної регресії для оцінювання метрик програмного забезпечення для довірчої ймовірності 0,9 та порівняти з відповідними результатами, що одержані у цій лабораторній роботі для довірчої ймовірності 0,95.

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?

7. Як обчислюється значення коефіцієнта детермінації?

8. На що вказує значення коефіцієнта детермінації?

9. У яких межах може змінюватися значення коефіцієнта детермінації?

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

11. Що таке скоригований коефіцієнт детермінації?

12. У якому випадку значення скоригованого коефіцієнта детермінації зменшується (збільшується) у порівнянні із значенням коефіцієнта детермінації?

13. Як обчислюється середня величина відносної похибки?

14. Яка середня величина відносної похибки вважається прийнятною для моделі?

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

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

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

18. Поясніть поняття мультиколінеарності між факторами у множинній лінійній регресії.

19. Для чого під час побудови множинного рівняння лінійної регресії потрібно перевірити майбутні фактори на наявність мультиколінеарності?

20. Як визначають наявність мультиколінеарності між факторами у множинній лінійній регресії?

21. Як обчислюється значення відкоригованого коефіцієнта множинної детермінації? На що вказує значення цього коефіцієнта?

22. Що таке довірчий інтервал регресії?

23. Що визначає довірчий інтервал?

24. Що таке інтервал передбачення (прогнозування) регресії?

25. Що означає вихід значення залежної випадкової величини за інтервал передбачення регресії?

26. Як визначити довірчий інтервал лінійної регресії при декількох факторах?

27. Як визначити інтервал прогнозування лінійної регресії при декількох факторах?

28. Як визначити довірчий інтервал однофакторної (багатофакторної) лінійної регресії?

29. Як визначити інтервал прогнозування однофакторної (багатофакторної) лінійної регресії?

Лабораторна робота № 4

Нелінійний регресійний аналіз багатовимірних даних

Мета роботи: набути практичних навичок нелінійного регресійного аналізу багатовимірних даних.

Завдання:

1. Виконати побудову нелінійного рівняння регресії (моделі) для оцінювання залежної величини за багатовимірними даними, які наведені у файлі.
2. Визначити значення коефіцієнта детермінації R^2, середньої величини відносної похибки MMRE та відсотка прогнозування на рівні величини відносної похибки, який дорівнює 0,25, PRED(0,25).
3. Визначити довірчі інтервали та інтервали прогнозування нелінійної регресії для довірчої ймовірності 0,95.
4. Здійснити побудову регресії, довірчих інтервалів, інтервалів прогнозування нелінійної регресії та емпіричних даних на графіку (у разі однофакторного рівняння регресії).
5. Виконати порівняння одержаних результатів з результатами попередньої роботи (лінійного регресійного аналізу).
6. Зробити висновки щодо одержаних результатів.

Загальні теоретичні відомості

Нелінійний регресійний аналіз виконується в разі, коли відсутнє теоретичне обґрунтування проведення лінійного регресійного аналізу. Насамперед у нелінійному регресійному аналізі визначають нелінійну регресію (нелінійне регресійне рівняння), а також її (його) статистичну значущість, щоб оцінити вплив кожного фактора на залежну змінну.

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 приводить до необхідності побудови нелінійних регресійних моделей.

На сьогодні для побудови нелінійних регресійних рівнянь та моделей використовуються методи на основі простого перебору, лінеаризації та нормалізуючих перетворень. Для уникнення простого перебору під час побудови нелінійних регресійних рівнянь та моделей використовують методи, що базуються на нормалізуючих взаємозворотних перетвореннях як одновимірних, так і багатовимірних. Методи на основі нормалізуючих перетворень не потребують використання процедури простого перебору, а, по-друге, у разі застосування бієктивних (*bijective*) перетворень не відбувається втрата частини інформації як, наприклад, під час лінеаризації.

Будь-який метод для побудови нелінійних регресійних рівнянь та моделей на основі нормалізуючих перетворень складається з трьох етапів. На першому етапі негаусівські дані нормалізують, тобто перетворюють у гаусівські дані із застосуванням взаємозворотного нормалізуючого перетворення (бажано із застосуванням багатовимірного перетворення, коли не можна нехтувати кореляцією між змінними). Далі на другому етапі будується лінійне регресійне рівняння (модель) для нормалізованих даних. І на третьому етапі за лінійним регресійним рівнянням (моделлю) для нормалізованих даних із застосуванням обраного взаємозворотного нормалізуючого перетворення отримують нелінійне регресійне рівняння (модель).

Але під час побудови нелінійних регресійних рівнянь та моделей потрібно враховувати можливу наявність викидів у даних. Тому зазначені три етапи бажано доповнити ще додатковими, які дозволяли би перевіряти наявність викидів. Ми рекомендуємо виконувати побудову нелінійних регресійних рівнянь та моделей за наступними етапами:

Етап 1. Нормалізувати багатовимірні дані за допомогою взаємозворотного нормалізуючого перетворення.

Етап 2. Перевірити багатовимірні дані на наявність викидів для довірчої ймовірності 0,995. Якщо багатовимірний викид знайдено, то його потрібно відкинути та повернутися до етапу 1.

Етап 3. Побудувати лінійне регресійне рівняння (модель) для нормалізованих даних.

Етап 4. Перевірити гіпотезу про нормальність закону розподілу значень відхилення між нормалізованими даними та лінією регресії для нормалізованих даних для довірчої ймовірності 0,95. У разі якщо зазначена гіпотеза не приймається, то потрібно відкинути дані, які відповідають максимальному за модулем відхиленню, та повернутися до етапу 1.

Етап 5. За лінійним регресійним рівнянням для нормалізованих даних із застосуванням обраного взаємозворотного нормалізуючого перетворення отримати нелінійне регресійне рівняння (модель).

Етап 6. Визначити інтервали прогнозування нелінійної регресії для довірчої ймовірності 0,95. Якщо серед значень залежної випадкової величини з багатовимірних даних є такі значення, які виходять за межі інтервалу прогнозування нелінійної регресії, то відкинути ці дані та повернутися до етапу 1. Якщо таких значень немає, то на цьому виконання етапів завершують, нелінійне регресійне рівняння (модель) побудовано.

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

$$T=\psi(P), \tag{4.1}$$

і зворотне перетворення для (4.1)

$$P=\psi^{-1}(T). \tag{4.2}$$

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

На першому етапі негаусівські дані нормалізують, тобто перетворюють у гаусівські дані із застосуванням перетворення (4.1). Далі на другому етапі будується лінійне регресійне рівняння

$$\hat{Z}_Y = b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k \qquad (4.3)$$

або лінійна регресійна модель для нормалізованих даних

$$Z_Y = \hat{Z}_Y + \varepsilon = b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k + \varepsilon, \qquad (4.4)$$

де ε – випадкова величина з розподілом Гауса, яка характеризує відхилення, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$.

І на третьому етапі за лінійним регресійним рівнянням (4.3) або за лінійною регресійною моделлю (4.4) із застосуванням обраного взаємозворотного нормалізуючого перетворення (4.1) отримують відповідно нелінійне регресійне рівняння

$$\hat{Y} = \psi_Y^{-1}\left(\hat{Z}_Y\right) = \psi_Y^{-1}\left(b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k\right) \qquad (4.5)$$

або нелінійну регресійну модель

$$Y = \psi_Y^{-1}\left(\hat{Z}_Y + \varepsilon\right) = \psi_Y^{-1}\left(b_0 + b_1 Z_1 + b_2 Z_2 + \ldots + b_k Z_k + \varepsilon\right), \qquad (4.6)$$

де ψ_Y – перша компонента вектора ψ перетворення (4.1).

Підкреслимо різницю між (4.5) та (4.6). За допомогою нелінійного регресійного рівняння (4.5) ми оцінюємо умовне математичне сподівання залежної випадкової величини *Y*. Нелінійна регресійна модель (4.6) дозволяє нам отримувати значення залежної випадкової величини *Y*. Зверніть також увагу на те, що завдяки побудові нелінійної регресійної моделі на основі взаємозворотного нормалізуючого перетворення, розв'язується проблема з визначенням розподілу випадкових відхилень ε, який у (4.6) є гаусівським.

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

$$Z_Y = \lg Y; \qquad (4.7)$$

$$Z_1 = \lg X_1. \qquad (4.8)$$

Спочатку емпіричні дані нормалізують за перетвореннями (4.7) і (4.8). Далі будується однофакторне лінійне регресійне рівняння для нормалізованих даних

$$\widehat{Z}_Y = b_0 + b_1 Z_1 . \qquad (4.9)$$

За рівнянням (4.9) отримують однофакторне нелінійне рівняння регресії

$$\hat{Y} = 10^{\hat{b}_0} X_1^{\hat{b}_1} . \qquad (4.10)$$

Рівняння виду (4.10) з відповідними параметрами застосовується як відомі моделі в інженерії програмного забезпечення: COCOMO (COnstructive COst MOdel) та ISBSG (International Software Benchmarking Standards Group).

Подібно (4.10) будується однофакторна нелінійна регресійна модель у разі застосування одновимірних перетворень (4.7) і (4.8). Ця модель матиме такий вигляд:

$$Y = 10^{\hat{b}_0 + \varepsilon} X_1^{\hat{b}_1} .$$

Довірчий інтервал множинної (багатофакторної) нелінійної регресії визначається як

$$\psi_Y^{-1}\left(\widehat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{ \frac{1}{N} + \left(z_X^+\right)^T S_{ZZ}^{-1} \left(z_X^+\right) \right\}^{1/2} \right), \qquad (4.11)$$

а інтервал прогнозування множинної нелінійної регресії – як

$$\psi_Y^{-1}\left(\widehat{Z}_Y \pm t_{\alpha/2,\nu} S_{Z_Y} \left\{ 1 + \frac{1}{N} + \left(z_X^+\right)^T S_{ZZ}^{-1} \left(z_X^+\right) \right\}^{1/2} \right) . \qquad (4.12)$$

Тут $\widehat{Z}_Y$ – результат прогнозування за лінійним регресійним рівнянням (5.3) для нормалізованих даних; z_X^+ – вектор центрованих нормалізованих факторів (регресорів), який містить значення $Z_{1_i} - \bar{Z}_1$, $Z_{2_i} - \bar{Z}_2$, …, $Z_{k_i} - \bar{Z}_k$; $S_{Z_Y}^2 = \frac{1}{\nu} \sum_{i=1}^{N} \left(Z_{Y_i} - \widehat{Z}_{Y_i} \right)^2$, $\nu = N - k - 1$; S_{ZZ} – $k \times k$ матриця

$$S_{ZZ} = \begin{pmatrix} S_{Z_1Z_1} & S_{Z_1Z_2} & \dots & S_{Z_1Z_k} \\ S_{Z_1Z_2} & S_{Z_2Z_2} & \dots & S_{Z_2Z_k} \\ \dots & \dots & \dots & \dots \\ S_{Z_1Z_k} & S_{Z_2Z_k} & \dots & S_{Z_kZ_k} \end{pmatrix}, \qquad (4.13)$$

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

У разі однофакторної нелінійної регресії довірчий інтервал (4.11) можна записати як

$$\psi_Y^{-1} \left(\hat{Z}_{Y_i} \pm t_{\alpha/2, N-2} S_Z \sqrt{\frac{1}{N} + \frac{\left(Z_{1_i} - \bar{Z}_1 \right)^2}{S_{Z_1Z_1}}} \right), \qquad (4.14)$$

а інтервал прогнозування (4.12) – як

$$\psi_Y^{-1} \left(\hat{Z}_{Y_i} \pm t_{\alpha/2, N-2} S_Z \sqrt{1 + \frac{1}{N} + \frac{\left(Z_{1_i} - \bar{Z}_1 \right)^2}{S_{Z_1Z_1}}} \right) \lim_{x \to \infty}. \qquad (4.15)$$

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

У (4.14) та (4.15) значення $\hat{Z}_{Y_i}$ визначається за однофакторним лінійним рівнянням регресії (4.9) для нормалізованих даних залежно від значення фактора Z_{1_i}.

Етапи виконання роботи

Виконання лабораторної роботи включає в себе наступні етапи:

1. Нормалізацію багатовимірних даних за допомогою взаємозворотного нормалізуючого перетворення.

2. Перевірку багатовимірних даних на наявність викидів для довірчої ймовірності 0,995. Якщо багатовимірний викид знайдено, то його потрібно відкинути та повернутися до етапу 1.

3. Побудову нелінійного регресійного рівняння (моделі) для нормалізованих даних.

4. Перевірку гіпотези про нормальність закону розподілу значень відхилення між нормалізованими даними та лінією регресії для нормалізованих даних для довірчої ймовірності 0,95. У разі якщо зазначена гіпотеза не приймається, то потрібно відкинути дані, які відповідають максимальному за модулем відхиленню, та повернутися до етапу 1.

5. Отримання нелінійного регресійного рівняння (моделі) за лінійним регресійним рівнянням для нормалізованих даних із застосуванням обраного взаємозворотного нормалізуючого перетворення.

6. Визначення інтервалів прогнозування нелінійної регресії для довірчої ймовірності 0,95. Якщо серед значень залежної випадкової величини з багатовимірних даних є такі значення, які виходять за границі інтервалу прогнозування нелінійної регресії, то відкинути ці дані та повернутися до етапу 1. Якщо таких значень немає, то нелінійне регресійне рівняння (модель) побудовано.

7. Визначення довірчих інтервалів нелінійної регресії для довірчої ймовірності 0,95.

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

9. Побудову лінії регресії, довірчих інтервалів, інтервалів прогнозування нелінійної регресії для оцінювання метрик програмного забезпечення та емпіричних даних на графіку (у разі однофакторного рівняння регресії).

10. Висновки за результатами виконання лабораторної роботи.

Завдання для самостійної роботи

Виконати побудову нелінійного регресійного рівняння (моделі) із використанням іншого нормалізуючого перетворення та порівняти одержані нелінійні регресійні рівняння (моделі).

Питання для самоконтролю

1. Що приводить 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 під час побудови нелінійних регресійних рівнянь та моделей?

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

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

6. Як побудувати однофакторне нелінійне рівняння регресії (модель) у разі застосування одновимірного перетворення у вигляді десяткового логарифма 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. Для чого 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?

10. Як обчислюється значення коефіцієнта детермінації?

11. На що вказує значення коефіцієнта 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 являє собою скоригована статистика R^2?

17. Для чого визначати у множинній регресійній моделі скориговану статистику R^2?

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

19. Що таке довірчий інтервал регресії?

20. Що визначає довірчий інтервал?

21. Що таке інтервал прогнозування регресії?

22. Що означає вихід значення залежної випадкової величини за інтервал прогнозування регресії?

23. Як визначити довірчий інтервал нелінійної регресії при декількох факторах?

24. Як визначити інтервал прогнозування нелінійної регресії при декількох факторах?

25. Як визначити довірчий інтервал однофакторної нелінійної регресії?

26. Як визначити інтервал прогнозування однофакторної нелінійної регресії?

СПИСОК ЛІТЕРАТУРИ

1. Приходько С. Б. Методичні вказівки до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на ЕОМ». Миколаїв : НУК, 2005. 52 с. URL: https://rep.nuos.edu.ua/server/api/core/bitstreams/6c8aba57-7ce6-423b-801e-91876d1c7a48/content.

2. Приходько С. Б., Макарова Л. М., Пугаченко К. С. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Обробка експериментальних даних на комп'ютері». Миколаїв : НУК, 2018. 76 с. URL: https://rep.nuos.edu.ua/server/api/core/bitstreams/331f2ec6-5fa7-4425-9585-de6eca5dc3b4/content.

3. Prykhodko S. B. Statistical anomaly detection techniques based on normalizing transformations for non-Gaussian data. *Computational Intelligence* (Results, Problems and Perspectives): Proceedings of the International Conference, May 12–15, 2015, Kyiv-Cherkasy, Ukraine / Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Ye. Snytyuk (Editor). Cherkasy : editor July Chabanenko, 2015. P. 286–287. URL: https://www.researchgate.net/publication/282329599_Statistical_anomaly_detection_techniques_based_on_normalizing_transformations_for_non-Gaussian_data.

4. Prykhodko S., Prykhodko N., Smykodub T. A Statistical Evaluation of the Depth of Inheritance Tree Metric for Open-Source Applications Developed in Java. *Foundations of Computing and Decision Sciences.* 2021. Vol. 46. No. 2. P. 159–172. DOI: https://doi.org/10.2478/fcds-2021-0011.

5. Коваленко І. І., Приходько С. Б., Латанська Л. О. Сучасні методи статистичного аналізу даних : навч. посіб. Миколаїв : НУК, 2011. 189 с.

6. Приходько С. Б., Макарова Л. М., Приходько Н. В., Пухалевич А. В. Методичні вказівки та завдання до виконання лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії». Миколаїв : НУК, 2023. 56 с. URL: https://rep.nuos.edu.ua/server/api/core/bitstreams/a9f56b2e-467d-4427-a0d9-b9ee138d69e6/content.

ДОДАТКИ

ДОДАТОК А

Верхні 100α %-ві точки розподілу χ^2

У таблиці А.1 наведені верхні 100α %-ві точки розподілу χ^2, тобто такі значення x, що $P\left(\chi_v^2 > x\right) = \alpha$.

Таблиця А.1. **Верхні 100α %-ві точки розподілу χ^2**

v	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	$0,0^439$	$0,0^316$	$0,0^398$	$0,0^239$	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

Продовж. табл. А.1

1	2	3	4	5	6	7	8	9
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
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 100α %-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 свободи і ν достатньо велика (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

У таблиці Б.1 наведені верхні 100α %-ві точки *t*-розподілу Стьюдента, тобто такі значення x, що $\overline{x}$.

Таблиця Б.1. **Верхні 100α %-ві точки *t*-розподілу Стьюдента**

v	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

Продовж. табл. Б.1

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

Навчальне видання

ПРИХОДЬКО Сергій Борисович

МЕТОДИЧНІ ВКАЗІВКИ ТА ЗАВДАННЯ
до виконання лабораторних робіт з дисципліни «МЕТОДИ АНАЛІЗУ РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ»

Комп'ютерна правка й коректура *О. Г. Костенко*
Комп'ютерне верстання *Ю. С. Семенченко*

Формат 60×84/16. Ум. друк. арк. 3,60. Вид. № 34. Зам. № 0402-04.
Видавець і виготівник Національний університет кораблебудування
імені адмірала Макарова
просп. Героїв України, 9, м. Миколаїв, 54007
E-mail : publishing@nuos.edu.ua
Свідоцтво суб'єкта видавничої справи ДК № 6402 від 19.09.2018 р.