КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет комп'ютерних наук та кібернетики
Кафедра системного аналізу та теорії прийняття рішень

Кваліфікаційна робота

на здобуття ступеня магістра

за спеціальністю 124 Системний аналіз

на тему:

ЛІНІЙНА РЕГРЕСІЯ З ПАРАМЕТРАМИ У ФОРМІ НЕЧІТКИХ ЧИСЛЕЛ ТИПУ-2

Виконав студент 2-го курсу магістратури
Пікула Богдан Андрійович
(підпис)

Науковий керівник:
професор, доктор фіз.-мат. наук
Мащенко Сергій Олегович
(підпис)

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент ______
(підпис)

Роботу розглянуто й допущено до захисту на засіданні кафедри системного аналізу та теорії прийняття рішень
«04» травня 2023 р.,
протокол № 11
Завідувач кафедри
О. Г. Наконечний ______
(підпис)

Київ – 2023

Анотація

Тема: Лінійна регресія з параметрами у формі нечітких чисел типу-2

Автор: Пікула Богдан Андрійович

Науковий керівник: професор, доктор фізико-математичних наук Мащенко Сергій Олегович

Робота складається з: 35 сторінок, 7 рисунків, 1 таблиця, використаних джерел -10.

Ключові слова: лінійна регресія, нечітка множина типу-2, нечітке число типу-2, функція належності, нечітка регресія.

Предмет дослідження: лінійна регресія з параметрами у формі нечітких чисел типу-2

Мета роботи: представити метод побудови лінійної регресії з параметрами у формі нечітких чисел типу-2; продемонструвати застосування розробленого методу на практиці.

Короткий огляд праці:

У першому розділі наведено базові поняття та визначення з теорії нечітких множин, чисел, та лінійної регресії.

У другому розділі представлено підхід до розв’язання задачі побудови лінійної регресії з параметрами у формі звичайний нечітких чисел.

У третьому розділі представлено підхід до розв’язання задачі побудови лінійної регресії з параметрами у формі нечітких чисел типу-2; одержані результати продемонстровано на прикладах.

ЗМІСТ

Вступ

В останні роки набули поширення регресійні моделі з параметрами в формі нечітких чисел, що дозволяють враховувати невизначеність та неоднорідність в даних. Завдяки цьому вони знаходять застосування в різних галузях, таких як економіка, фінанси, екологія, медицина та техніка.

Розвиток методів побудови лінійної регресії з параметрами в формі нечітких чисел типу-2 є актуальною проблемою, оскільки він сприяє підвищенню точності аналізу даних та якості прийняття рішень у різних галузях на основі точного та робастного прогнозування.

Метою даної роботи є розробка методів розв'язання задачі побудови лінійної регресії з параметрами в формі нечітких чисел типу-2.

Завданнями роботи є аналіз сучасних підходів до розв'язання цієї задачі, розробка нових алгоритмів, дослідження впливу різних факторів на точність та стабільність моделей, та застосування розглянутих методів до реальних задач.

Об'єктом дослідження є задача лінійної регресії з параметрами в формі нечітких чисел типу-2.

Методами дослідження є інтервальний метод та генетичні алгоритми, а також розробка нових методів оптимізації на основі цих підходів.

Методи лінійної регресії з параметрами в формі нечітких чисел типу-2 можуть бути успішно використані в різних галузях, таких як економіка, фінанси, екологія, медицина, техніка та інших, де відбувається аналіз даних з невизначеністю та неоднорідністю.

Структура роботи включає вступ, три розділи, висновок та список використаної літератури.

Розділ 1. Огляд з теорії нечітких множин та лінійної регресії

1.1. Лінійна регресія

1.1.1. Загальні означення. Лінійна регресія – це метод моделювання залежності між скалярною змінною y та векторною змінною X. Лінійна регресія називають простою, якщо вектор X є також скалярною змінною. При використанні лінійної регресії взаємозв'язок між даними моделюється за допомогою лінійних функцій, а невідомі параметри моделі оцінюються за вхідними даними. Для розрахунку моделі лінійної регресії зазвичай використовується метод найменших квадратів, хоча можуть бути використані інші методи. Загальна лінійна регресійна модель має вигляд:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \tag{1.1}$$

де y – залежна змінна, $(x_1, x_2, \ldots, x_k)$ - незалежні змінні, u – випадкова похибка, розподіл якої в загальному випадку залежить від незалежних змінних, але математичне сподівання якої дорівнює нулеві. Згідно з цією моделлю математичне сподівання залежної змінної є лінійною функцією незалежних змінних:

$$\mathbb{E}(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u. \tag{1.2}$$

Вектор параметрів $(\beta_0, \beta_1, \ldots, \beta_k)$ є невідомим. Задача лінійної регресії полягає в оцінці цих параметрів на основі деяких експериментальних значень y та $(x_1, x_2, \ldots, x_k)$. Тобто для деяких n експериментів є відомі значення $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ незалежних змінних та їм відповідають значення залежної змінної. Для кожного експериментального випадку залежність між змінними визначається такими формулами:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + u_i \tag{1.3}$$

або у матричних позначеннях $y = X\beta + u$, де:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},\ u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}. \quad (1.4)$$

На основі цих даних задача полягає в оцінці значень параметрів $(\beta_0, \beta_1, \dots, \beta_k)$, а також розподіл величини u. Зважаючи на характеристики досліджуваних змінних можуть додаватися різні додаткові специфікації моделі та застосовуватись різні методи оцінки параметрів. Серед найпоширеніших специфікацій лінійних моделей є класична модель лінійної регресії та узагальнена модель лінійної регресії.

1.1.2. Моделі регресій. Класична модель. Згідно з класичною моделлю додатково вводяться такі вимоги щодо специфікації моделі та відомих експериментальних даних:

1) $\forall i \neq j\ \mathbb{E}(u_i u_j | x_i) = 0$ – відсутність кореляції залишків. Залежність або пов'язаність, є будь-яким статистичним відношенням, чи казуальним (причино-наслідковий зв'язок, «наслідок» випливає з «події»), чи ні, між двома випадковими величинами. Кореляція це будь-яке статистичне відношення, що задає залежність величин.

2) $\forall i\ \mathbb{E}(u_i^2 | x_i) = \sigma^2$ (гомоскедастичність). Незалежність дисперсії випадкових складових від номеру спостереження. При виконанні цієї умови, оцінки звичайного методу найменших квадратів мають найменшу дисперсію серед усіх незміщених оцінок коефіцієнтів лінійної моделі.

3) Ранг матриці X дорівнює $K + 1$.

4) Усі елементи матриці X є невипадковими.

Узагальнена модель. Умови гомоскедастичності та відсутності кореляції між випадковими залишками у моделі часто не виконуються на практиці. Тож, якщо взяти загальнішу умову:

$$\mathbb{V}(u|X) = \sigma^2 W, \quad (1.5)$$

де W – відома додатно визначена матриця, то одержана модель називається узагальненою моделлю лінійної регресії.

Оскільки для кожної додатно визначеної матриці W існує така матриця N, що $W^{-1} = NN$, то модель:

$$N_y = NX\beta + N_u \quad (1.6)$$

вже буде класичною моделлю лінійної регресії.

1.1.3. Методи оцінювання. Залежно від об'єктів, що досліджуються за допомогою лінійної регресії та конкретних цілей дослідження можуть використовуватися різні методи оцінки невідомих параметрів. Найпопулярнішим є звичайний метод найменших квадратів. Він приймає за оцінку параметра значення, що мінімізують суму квадратів залишків по всіх спостереженнях:

$$\hat{\beta} = arg\min_{\beta}\left\{\sum_{i=1}^{n}\left|y_i - \beta_0 - \sum_{j=1}^{k} X_{ij}\beta_j\right|^2\right\} = \arg\min_{\beta}\|y = X\beta\|^2. (1.7)$$

Метод найменших квадратів можна застосовувати у будь-яких задачах, якщо лише ранг матриці X рівний кількості її стовпців. Також цей метод дає простий аналітичний вираз для оцінки параметрів

$$\hat{\beta} = (X'X)^{-1}X'y. \quad (1.8)$$

У випадку класичної моделі лінійної регресії оцінка методу найменших квадратів є незміщеною, змістовною оцінкою. Оцінка $\hat{\theta}$ є незміщеною оцінкою параметра θ, якщо

$$E(\hat{\theta}) = \theta. \quad (1.9)$$

Оцінка $\hat{\theta}$ є змістовною, якщо

$$\hat{\theta} \to \theta, \forall\theta \in \Theta \text{ при } n \to \infty. \quad (1.10)$$

1.2. Нечіткі числа

На практиці виникає потреба у використанні нечітких виразів, наприклад, «приблизно п'ять», «молода людина» та «гаряча температура води». Такі твердження не мають чіткого визначення та для кожної людини мають власний зміст, якщо такий взагалі існує. Тому, для вирішення проблеми надання чіткості таким виразам та створенню інструментів для роботи з ними існує поняття нечіткого числа.

Концепт нечітких чисел та нечіткої арифметики ввів Лотфі Заде.

Розглянемо означення нечіткого числа. Почнемо з вигляду звичайного дійсного числа. Нехай $a \in \mathbb{R}$ дійсне число. Його можна визначити множиною $\{a\}$, що складається лише з єдиного елемента a. Якщо обрати довільну підмножину $S \subset \mathbb{R}$, то легко бачити, що існує функція-індикатор, яка приймає лише два значення – нуль та одиницю $\chi_S : \mathbb{R} \to \{0,1\}$, тобто вказує, чи дійсне число міститься в підмножині S: $x \in S \iff \chi_S(x) = 1, \quad x \notin S \iff \chi_S(x) = 0.$

У випадку, якщо множина S є одноелементною і містить єдиний елемент $a \in \mathbb{R}$, тобто $S = \{a\}$, функція-індикатор $\chi_S(x)$ набуватиме нульового значення для всіх x, окрім $x = a$: $x = a \iff \chi_S(x) = 1$, $x \neq a \iff \chi_S(x) = 0$.

Маючи означення чіткого числа, можливим є введення нечіткості, замінивши функцію-індикатор на функцію належності, яка приймає значення на інтервалі від 0 до 1: $\mu : \mathbb{R} \to [0,1]$. Для визначення нечіткого числа, цього буде недостатньо, тому що варто покласти кілька обмежень на функцію належності.

Нечітким числом називають [1] відображення $\mu_a : \mathbb{R} \to [0,1]$, таке що: $\mu_a(x) = 1 \iff x = a$ та $\lim_{|x| \to \infty} \mu_a(x) = 0$.

Наведемо кілька прикладів вигляду функцій належності нечітких чисел:

Рис. 1.1. Трикутне нечітке число

Рис. 1.2. Трапецієвидне нечітке число

Рис. 1.3. Дзвіно-видне нечітке число

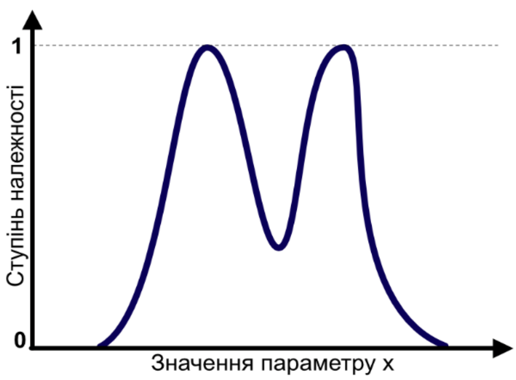

Рис. 1.4. Не опукле нечітке число

Ввівши додаткові умови для функції належності, можемо визначити різні типи нечітких чисел [1], які позначаються як D_i - нечіткі числа для $i = 2,3,\ldots,8$ відповідно. При $i = 1$ матимемо звичайне нечітке число. Функції належності притаманні такі властивості:

а) μ_a – неперервна;

б) μ_a – монотонно зростаюча на $S_a^- = \{x \in (-\infty, a) : \mu_a(x) > 0\}$ та монотонно спадна на $S_a^+ = \{x \in (a, \infty) : \mu_a(x) > 0\}$;

в) μ_a – не спадна на $(-\infty, a)$ та не зростаюча на (a, ∞) ;

г) μ_a – монотонно спадна на $(-\infty, a)$ та монотонно спадна на (a, ∞) ;

ґ) $(\forall x)(\mu_a(a + x) = \mu_a(a - x))$;

д) $(\exists \lambda)(x \notin [a - \lambda, a + \lambda] \Rightarrow \mu_a(x) = 0)$.

Означення D_i - нечітких чисел тепер можна легко подати через комбінацію шести цих властивостей.

Звичайне нечітке число будемо називати D_i - нечітким числом, якщо μ_a задовольняє таким умовам:

- а) для $i = 2$;
- в) для $i = 3$;
- в) та ґ) для $i = 4$;
- а) та г) для $i = 5$;
- а), г) та ґ) для $i = 6$;
- а), г) та д) для $i = 7$;
- б) для $i = 8$.

Називатимемо S_a носієм функції $\mu_a : S_a = \{x \in \mathbb{R} : \mu_a(x) > 0\}$ [1]. Нечіткі числа типу-D_8 з раціональним носієм $S_a \subset \mathbb{Q}$ називатимемо нечіткими раціональними числами. Аналогічно, числа типу-D_8 з цілим носієм $S_a \in \mathbb{Z}$ називатимемо нечіткими цілими числами [1]. Подібним чином можна визначити інші нечіткі числа, наприклад, ірраціональні.

1.3. Нечіткі множини

Поняття нечіткої множини – це спроба математичної формалізації нечіткої інформації з метою побудови математичних моделей за допомогою них. У цьому є потреба, адже в житті часто трапляються ситуації, коли належність елемента до певної множини є не однозначною, й тому потрібно вказувати наскільки цей елемент належить цій множині. Простим способом опису нечіткої множини є характеризація ступеню належності елемента до цієї множини якимось числом, наприклад, взявши числа з інтервалу [0, 1].

Нехай X – довільна множина. Нечіткою множиною A в X називатимемо сукупність пар $\left(x, \mu_A(x)\right)$, де $x \in X$, $\mu_A : X \to [0,1]$ – функція належності елементів до множини A [2]. Значення функції належності $\mu_A(x)$ називається ступенем належності елемента до нечіткої множини A.

Дане означення стосується звичайної нечіткої множини. Її ще можна назвати нечіткою множиною типу 1. У природі існують нечіткі множини різних типів. Нечіткою множиною типу n називають таку нечітку множину, що має функцію належності, яка приймає значення нечіткої множини типу $n - 1$ [2].

Нечітку множину називатимемо порожньою, якщо функція належності тотожно дорівнює нулю: $\mu_{\varnothing}(x) \equiv 0$.

Універсум X можна визначити через функцію належності таким чином: $\mu_X(x) \equiv 1$.

Носієм [2] нечіткої множини A з функцією належності $\mu_A(x)$ називатимемо множину $\operatorname{supp} A = \left\{x \mid x \in X, \mu_A(x) > 0\right\}$.

Нормальною вважатимемо нечітку множину, якщо виконується така нерівність: $\sup\limits_{x \in X} \mu_A(x) = 1$.

Розглянемо відношення між двома нечіткими множинами. Нехай A та B – нечіткі множини на X з функціями належності $\mu_A(x)$ та $\mu_B(x)$ відповідно. Множина A включає множину B, якщо $\forall x \in X$ виконується умова

$\mu_B(x) \leq \mu_A(x)$. Записуємо таке відношення як $B \subseteq A$.

Множини A та B є еквівалентними, якщо $\mu_A(x) = \mu_B(x), \quad \forall x \in X$.

У випадку включення множиною A множини B, справедливе таке відношення:$\operatorname{supp} B \subseteq \operatorname{supp} A$.

Також варто розглянути операції з нечіткими множинами. Аналогічно звичайним множинам, з нечіткими множинами можна здійснювати такі операції, як об'єднання, перетин, доповнення, різницю та ін.

Об'єднанням множин A та B в X називатимемо [2] нечітку множину $A \cup B$, яка має таку функцію належності:

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}, \quad x \in X.$$

Рис. 1.5. Операція об'єднання

Перетином множин A та B в X називатимемо [2] нечітку множину $A \cap B$, яка має таку функцію належності:

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}, \quad x \in X.$$

Рис. 1.6. Операція перетину

Доповненням до множини A в X називатимемо [2] нечітку множину

A^c, яка має таку функцію належності:

$$\mu_{A^c}(x) = 1 - \mu_A(x), \quad x \in X.$$

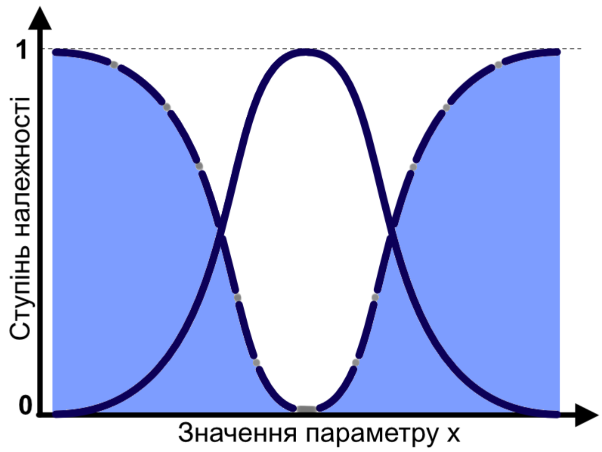

Рис. 1.7. Операція доповнення

Різницею множин A та B в X називатимемо [2] нечітку множину $A \setminus B$, яка має таку функцію належності:

$$\mu_{A \setminus B}(x) = \begin{cases} \mu_A(x) - \mu_B(x), & \mu_A(x) \geq \mu_B(x), \\ 0, & \mu_A(x) < \mu_B(x). \end{cases}$$

Декартовим добутком $A_1 \times \ldots \times A_n$ множин A_i в X_i, $i = 1, \ldots, n$, називатимемо [2] нечіткою множиною A, в декартовому добутку $X = X_1 \times \ldots \times X_n$, яка має таку функцію належності:

$$\mu_A(x) = \min\{\mu_{A_1}(x_1), \ldots, \mu_{A_n}(x_n)\},$$
$$x = (x_1, \ldots, x_n) \in X.$$

Опуклою комбінацією множин $A_1, \ldots, A_n$ в X називатимемо [2] нечітку множину A, яка має таку функцію належності:

$$\mu_A(x) = \sum_{i=1}^{n} \lambda_i \mu_i(x),$$

де $\lambda_i \geq 0, i = 1,2, \ldots, n, \quad \sum_{i=1}^{n} \lambda_i = 1.$

Розділ 2. Нечітка лінійна регресія

Нагадаємо, що статистична лінійна регресія має такий вигляд:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k + x_{ik} + \epsilon, \quad i = 1,2,\ldots,m. \tag{2.1}$$

У цьому виразі, залежна змінна y_i, незалежні змінні x_{ij}, та коефіцієнти β_j є чіткими числами, та ϵ_i є чітким значенням випадкової похибки, котрій належать такі властивості – математичне сподівання $\mathbb{E}(\epsilon_i) = 0$, дисперсія $\sigma^2(\epsilon_i) = \sigma^2$ та коваріація $\sigma(\epsilon_i, \epsilon_j) = 0, \forall i, j, i \neq j$.

Попри те, що статистична регресія має безліч застосувань, можуть виникнути проблеми в таких ситуаціях:

- наявний малий датасет – невелика кількість спостережень;
- складнощі при перевірки припущення розподілу;
- нечіткість у відношенні вхідних та вихідних змінних;
- неоднозначність подій або ступінь їхнього виникнення;
- неточність та спотворення при лінеаризації

Застосування чітких статистичних регресій є проблематичним у випадку, коли присутній малий датасет, складність перевірки того, що похибка має Гаусівський розподіл, нечіткість відповідності залежних та незалежних змінних, неоднозначність подій або припущення лінійності недоречне.

2.1. Складові нечіткої регресії

Загалом можна окреслити два способи побудови нечіткої регресійної моделі:

1) з нечітким відношення змінних;
2) з нечіткими змінними.

Обидві моделі будуть детально описані в літературі, але заради концептуалізації, варто сконцентруватись на моделях, яким властиві чіткі дані, але які мають нечіткі залежності між змінними. Задача концептуалізації нечіткої регресії не є важкою. Для наведення такого прикладу та в подальшому будемо використовувати тестовий набір даних Ішібучі [2]:

Таблиця 1.

Простий набір даних

i	1	2	3	4	5	6	7	8
x_i	2	4	6	8	10	12	14	16
y_i	14	16	14	18	18	22	18	22

Якщо взяти ці дані за основу, ми маємо можливість провести пряму через дві або більше точок таким чином, що всі решта точок опиняться або вище прямої, або нижче неї. Іншими словами, ми можемо побудувати верхню та нижню межу для цього набору даних. Зобразимо дану дію на Рис. 2.1. Підрахувавши, отримаємо дві прямих:

$$\hat{y} = 13 + 0.75x, \quad \hat{y} = 11 + 0.5x. \tag{2.2}$$

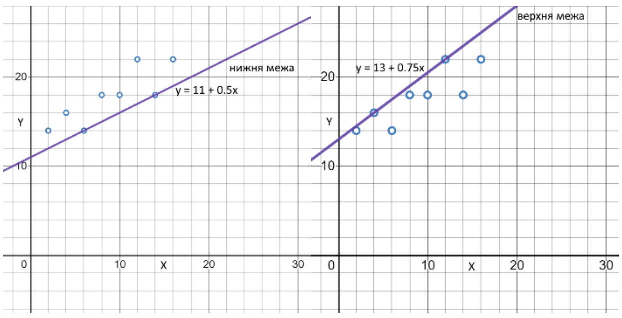

Рис. 2.1. Побудовані верхня та нижня межі (2)

Зробимо припущення, що нечіткі числа, що будуть використовуватись у даному прикладі, матимуть вигляд звичайного трикутного нечіткого числа, тобто супремум ступеня належності завжди припадає на середину інтервалу меж [3].

Для будь-якої пари значень (x_i, y_i), вищезазначена концептуалізація може бути підсумована за допомогою нечіткого регресійного інтервалу

$[Y_i^L, Y_i^U]$ як показано на Рис. 2.2.

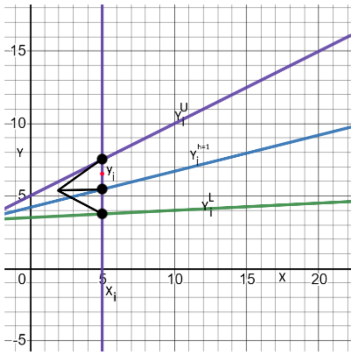

Рис. 2. 2. Нечіткий регресійний інтервал

$Y_i^{h=1}$ є серединою функції належності та за припущенням звичайного трикутного нечіткого числа

$$Y_i^{h=1} = \overline{Y}_i = \frac{(Y_i^U + Y_i^L)}{2}. \quad (2.3)$$

Нам дані параметри $(Y^U, Y^L, Y^{h=1})$, які характеризують поведінку моделі нечіткої регресії. Для кожної i-ої пари значень (x_i, y_i) існують параметри моделі $(Y_i^U, Y_i^L, Y_i^{h=1})$. З точки зору регресії, можна розглядати $Y_i^U - y_i$ і $y_i - Y_i^L$ як складові загальної суми квадратів, та $y_i - Y_i^{h=1}$ як складову похибки суми квадратів, та $Y_i^U - Y_i^{h=1}$ і $Y_i^{h=1} - Y_i^L$ як складові суми квадратів внаслідок регресії [3].

У «можливостійній» регресії, побудованій за допомогою ЗТНЧ, тільки пари значень, залучені у розрахунку верхніх та нижніх меж, вирішують структуру моделі, що зображена на Рис. 2.1. Решта значень не мають жодного впливу на її структуру. Цю проблему варто спробувати розв'язати за допомогою використання асиметричних нечітких чисел.

2.2. Нечіткі коефіцієнти

Розглянемо загальний вигляд моделі для нечіткої регресії [4]:

$$\tilde{Y} = \widetilde{A_0} + \widetilde{A_1}x_1 + \cdots + \widetilde{A_n}x_n, \quad (2.4)$$

де $\tilde{Y}$ є вихідним нечітким значенням, $\widetilde{A_i}, i = 1,2, \dots, n$ – нечіткими коефіцієнтами, а $x = (x_1, \dots, x_n)$ – чітким n-вимірним вхідним вектором.

Об'єднавши (14) та означення трикутних нечітких чисел, розглядаючи стандартні трикутні нечіткі числа, варто заначити вигляд j-ого коефіцієнту функції належності:

$$\mu_{(A_j)}(a) = max\,\{1 - \frac{|a-a_j|}{c_j}, 0\}, \quad (2.5)$$

де a_j є серединою меж та c_j – розмаху (відстані між межею та серединою). Дану функцію належності можемо зобразити на Рис. 2.3.

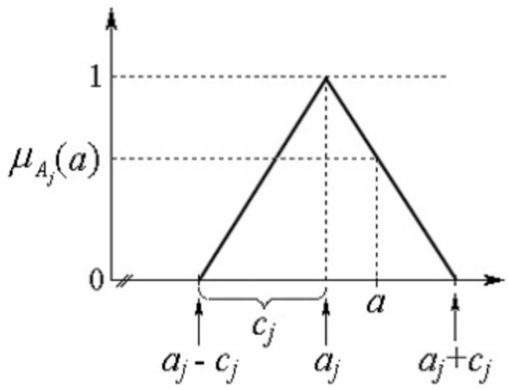

Рис. 2.3. Симетричні нечіткі параметри

Введемо такі позначення:

$$\widetilde{A_j} = \{a_j, c_j\}_L = \{\widetilde{A_j}: a_j - c_j \le \widetilde{A_j} \le a_j + c_j\}_L, \quad j = 0,1,\dots,n \quad (2.6)$$

та обмежимо розглянутий випадок тільки до тих випадків, коли лише коефіцієнти є нечіткими. Таким чином, можемо записати

$$\widetilde{Y_i} = \widetilde{A_0} + \sum_{j=1}^{n} \widetilde{A_j}\, x_{ij} = (a_0, c_0)_L + \sum_{j=1}^{n} \left(a_j, c_j\right)_L x_{ij}. \quad (2.7)$$

Дане формулювання є вкрай важливою, адже вона явно показує середину та розмах нечітких параметрів. У подальшому нашим завданням буде дослідження нечітких незалежних змінних.

2.3. Метод нечітких найменших квадратів

Одним із найочевидніших способів привести нечітку регресію ближче до статистичної – це моделювання нечіткої регресії таким самим чином. У випадку єдиної змінної прогнозованого результату, візьмемо за основу стандартну лінійну регресійну модель:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1,2, \dots, m. \tag{2.8}$$

Цю модель можна переписати в умовах нечіткості таким чином:

$$\widetilde{Y_\iota} = \beta_0 + \beta_1 \widetilde{X_\iota}, \quad i = 1,2, \dots, m. \tag{2.9}$$

Концепційний зв'язок між нечітким i-им відгуком та прогнозованих змінних у (2.9) зображено на Рис. 2.6.

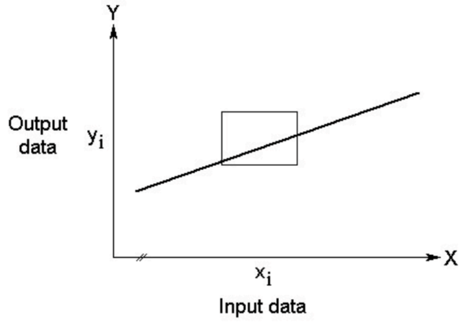

Рис. 2. Приклад нечіткої регресії.

Якщо переставити елементи в рівнянні (2.9), то одержимо

$$\widetilde{\epsilon_\iota} = \widetilde{Y_\iota} - \beta_0 - \beta_1 \widetilde{X_\iota}, \quad i = 1,2, \dots, m. \tag{2.10}$$

З точки зору найменших квадратів, задача перетворюється на таку:

$$min \sum_{i=1}^{n} \left(\widetilde{Y_\iota} - b_0 - b_1 \widetilde{X_\iota}\right)^2. \tag{2.11}$$

На практиці є кілька способів застосувати метод нечітких найменших квадратів, але два найпопулярніших – це вимірюючи відстань та використовуючи заходи сумісності.

2.4. МННК через вимір дистанції

Діамонд (1988) був першим, хто імплементував МННК через використання виміру. Його методологія є найбільш популярною. Він визначив L^2 – метрику $d(.,.)^2$ між двома ТНЧ за допомогою такої формули:

$$d(\langle m_1, l_1, r_1\rangle, \langle m_2, l_2, r_2\rangle)^2 = (m_1 - m_2)^2 + \left((m_1 - l_1) - (m_2 - l_2)\right)^2$$
$$+ \left((m_1 + r_1) - (m_2 + r_2)\right)^2. \tag{2.12}$$

Для заданих ТНЧ, це дозволяє виміряти відстань між двома нечіткими числами для їхніх середніх, лівих та правих поширень.

Цей випадок є найбільш близьким до моделі Санчеза та Гомеза, яка має такий вигляд:

$$\widetilde{Y_\iota} = \widetilde{\beta_0} + \widetilde{\beta_1} x_i + \widetilde{\epsilon_\iota}, \quad i = 1,2,\dots,m \tag{2.13}$$

і потребує такої оптимізації:

$$\min_{A,B} \sum d\left(\tilde{A} + \tilde{B} x_i, \widetilde{Y_\iota}\right)^2. \tag{2.14}$$

Розв’язок слідує з (22), та якщо, $\tilde{B}$ є додатнім, воно має форму:

$$d\left(\tilde{A} + x_i \tilde{B}, \widetilde{Y_\iota}\right)^2 = (a + b x_i - y_i)^2 + \left(a + b x_i - c_A^L - c_B^L x_i - y_i + c_{Y_i}^L\right)^2$$
$$+\left(a + b x_i + c_A^R + c_B^R x_i - y_i + c_{Y_i}^R\right)^2 . \tag{2.15}$$

Аналогічний вираз можна записати у випадку, що $\tilde{B}$ є від’ємним. Якщо розв’язок існує, параметри $\tilde{A}$ та $\tilde{B}$ задовольняють систему шести рівнянь у такій же кількості невідомих, ці рівняння виникають з похідних, що асоціюються з (25), яке дорівнює нулю. Звісно, що ця модель має такі ж загальні характеристики як раніше показано, але зараз ми можемо використовувати залишкову суму d-квадратів, щоб виміряти ефективність моделі.

У випадку найближчому до статистичної регресії, коефіцієнти, коефіцієнти є чіткими та задача перетворюється на оптимізацію найменших квадратів

$$\min_{a,b} \sum d\left(a + b \widetilde{X_\iota}, \widetilde{Y_\iota}\right)^2. \tag{2.16}$$

Зрештою, виникає цікава задача при імплементації підходу Діаманту при моделі такого вигляду

$$\widetilde{Y_\iota} = \widetilde{\beta_0} + \widetilde{\beta_1} \widetilde{X_\iota} + \widetilde{\epsilon_\iota}, \quad i = 1,2,\dots,m, \tag{2.17}$$

для якої не існує загального розв’язку, через те, що з лівого боку, $\widetilde{Y_\iota}$ є ТНЧ, а права частина включає нечіткий добуток $\widetilde{\beta_1}\widetilde{X_\iota}$.

Одним з підходів вирішення цієї задачі є заміна t-норми $min(a, b)$ з t-нормою $T_w(a, b) = a,$ якщо $b = 1;$ $b,$ якщо $a = 1;$ $0,$ інакше. Через те, що

$T_w(a, b)$ має вигляд, що зберігає порядок операцій після множення, він вирішує проблему. Також непоганою ідеєю є апроксимація ТНЧ.

2.5. Нечітка регресія найменших квадратів з використанням заходів сумісності

Альтернативний підхід найменших квадратів базується на заходах сумісності Целмінсу:

$$\gamma\left(\tilde{A}, \tilde{B}\right) = \max_{x} min\{ \mu_A(X), \mu_B(X)\}. \qquad (2.18)$$

З прикладами візуальної репрезентації зображеними на Рис. 2.7.

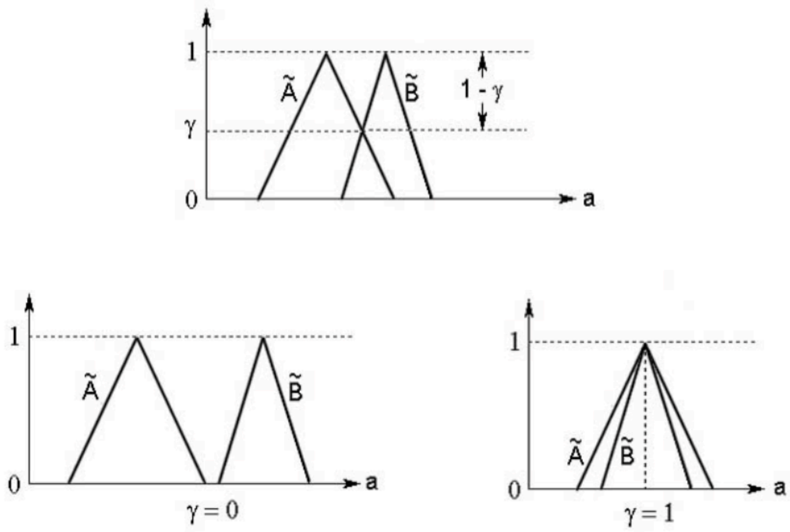

Рис. 2. 7. Заходи сумісності Целмінсу.

Як зазначено, γ приймає значення від 0, коли функції належності є взаємновиключними, та до 1, коли функції нележності співпадають.

Модель сумісності Целмінсу, яка включає максимізацію сумісності даних та приталеної моделі випливає з цього вимірку. Цільова функція має такий вигляд:

$$\sum_{i=1}^{m}(1-\gamma_i)^2. \qquad (2.19)$$

Таким чином, наприклад, коли існує єдина чітка пояснювальна змінна

$$\hat{Y} = \widetilde{A_0} + \widetilde{A_1}x = m_0 + m_1x \pm \sqrt{(c_0 + 2c_{01}x + c_1^2x^2)}, \qquad (2.20)$$

де m_0 та m_1 визначаються через зважену регресію найменших квадратів, та

c_0, c_1 та c_{01} визначаються через ітерацію та необхідний вимір сумісності.

Прикладом використання моделі сумісності Целмінсу застосованого для наших зразкових даних зображено на Рис. 2.8.

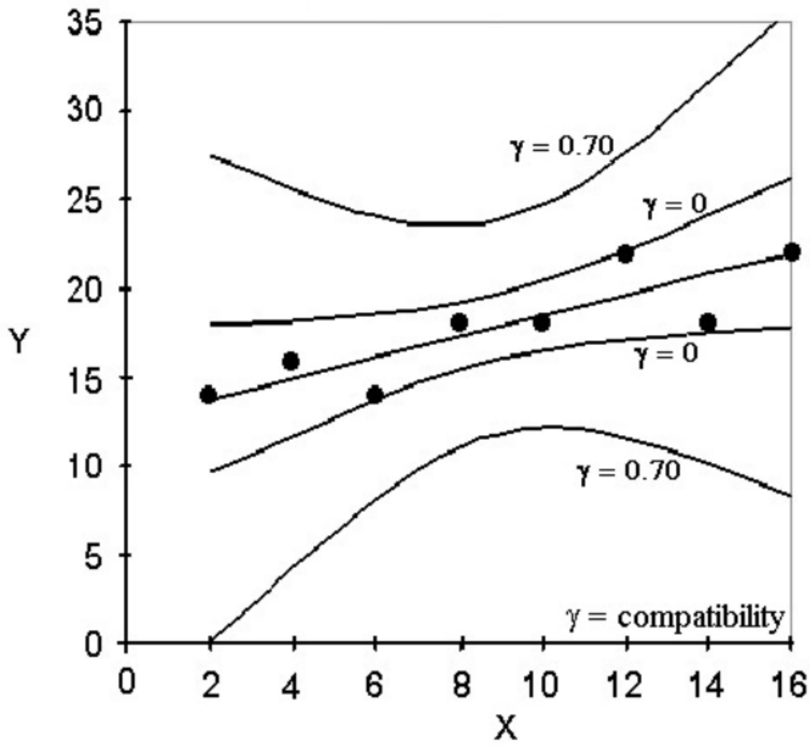

Рис. 2.8. Нечіткі найменші квадрати за допомогою критерію максимальної сумісності

Суттєвими характеристиками моделі в цьому випадку є параболічні криві для верхньої та нижньої границі. Чим вищий рівень сумісності, тим ширші границі.

Розділ 3. Задача про лінійну регресію з параметрами в формі нечітких чисел типу-2

Лінійна регресія є одним з найпопулярніших методів прогнозування і аналізу залежності між змінними. Зазвичай, вона використовується для визначення залежності між однією залежною змінною та однією або кількома незалежними змінними. Однак, в реальних задачах дані часто містять нечіткість, що може ускладнити аналіз і прогнозування. Один зі способів подолання цього недоліку - використання нечітких чисел типу-2 для моделювання нечітких параметрів лінійної регресії. У цьому розділі розглядається задача побудови лінійної регресії з параметрами в формі нечітких чисел типу-2.

Розглянемо задачу побудови лінійної регресії з однією залежною змінною y та n незалежними змінними x1, x2, ..., xn. Модель чіткої лінійної регресії має вигляд:

$$y = a0 + a1 * x1 + a2 * x2 + \dots + an * xn + \varepsilon, \qquad (3.1)$$

де a0, a1, ..., an - параметри моделі, ε - випадкова помилка з нульовим математичним очікуванням і деякою дисперсією. Оскільки реальні дані часто містять нечіткість, ми замінимо параметри моделі (a0, a1, ..., an) нечіткими числами типу-2 A0, A1, ..., An. Таким чином, наша модель матиме наступний вигляд:

$$Y = A0 + A1 * X1 + A2 * X2 + \dots + An * Xn + E. \qquad (3.2)$$

Нечітке число типу-2 можна представити як множину нечітких чисел типу-1, що є його "рівнями" (умовно кажучи, "різними ступенями нечіткості"). Відповідно, кожне число типу-2 має дві границі: верхню і нижню, які також є нечіткими числами типу-1. Для розв'язання задачі побудови лінійної регресії з параметрами в формі нечітких чисел типу-2 можна використовувати різні методи, зокрема інтервальний підхід, алгоритми оптимізації та методи нечіткої логіки. Розглянемо деякі з них:

3.1 Інтервальний підхід

Оскільки кожне нечітке число типу-2 має верхню і нижню границі, можна спробувати розв'язати задачу, апроксимуючи ці границі інтервалами. В цьому випадку, задача лінійної регресії розбивається на дві підзадачі: одна для верхньої границі та одна для нижньої границі. Розв'язки цих підзадач можуть бути знайдені за допомогою традиційних методів лінійної регресії, а потім об'єднані для отримання розв'язку для нечіткої лінійної регресії. Таким чином, ми апроксимуємо нечітке число типу-2 інтервалами для верхньої та нижньої границі:

$$[A_i^L, A_i^U]. \tag{3.3}$$

Тоді задачу лінійної регресії можна розбити на дві підзадачі:

$$Y^L = A_0^L + A_1^L X_1 + \cdots + A_n^L X_n, \tag{3.4}$$
$$Y^U = A_0^U + A_1^U X_1 + \cdots + A_n^U X_n.$$

Ми можемо розв'язати ці підзадачі за допомогою традиційних методів лінійної регресії, таких як метод найменших квадратів. Нехай ми маємо набір даних зі змінною відгуку Y та набором змінних-пояснювачів $X_1, X_2, \ldots, X_n$. Ми хочемо побудувати модель лінійної регресії для прогнозування значень Y.

Верхня границя інтервалу може бути апроксимована лінійною функцією:

$$Y^U = A_0^U + A_1^U X_1 + A_2^U X_2 + \ldots + A_n^U X_n, \tag{3.5}$$

де $A_0^U, A_1^U, A_2^U, \ldots, A_n^U$ є параметрами моделі для верхньої границі.

Аналогічно, нижня границя інтервалу може бути апроксимована лінійною функцією:

$$Y^L = A_0^L + A_1^L X_1 + A_2^L X_2 + \ldots + A_n^L X_n, \tag{3.6}$$

де $A_0^L, A_1^L, A_2^L, \ldots, A_n^L$ є параметрами моделі для нижньої границі.

Нашою метою є знаходження оптимальних значень параметрів

$A_0^U, A_1^U, A_2^U, ..., A_n^U$та $A_0^L, A_1^L, A_2^L, ..., A_n^L$, щоб якнайкраще апроксимувати верхню і нижню границі інтервалу.

Метод найменших квадратів може бути використаний для розв'язання кожної підзадачі окремо. Для цього ми мінімізуємо суму квадратів різниць між спостережуваними значеннями Y та прогнозованими значеннями Y^U або Y^L. Для верхньої границі:

Мінімізуємо суму квадратів різниць:

$$\min \Sigma (Y - Y^U)^2. \tag{3.7}$$

Підставляємо вираз для Y^U:

$$\min \Sigma \Big(Y - (A_0^U + A_1^U X_1 + A_2^U X_2 + \dots + A_n^U X_n)\Big)^2. \tag{3.8}$$

Цю задачу можна розв'язати, наприклад, за допомогою методу найменших квадратів або інших методів оптимізації.

Аналогічно, для нижньої границі:

Мінімізуємо суму квадратів різниць:

$$\min \Sigma (Y - Y^L)^2. \tag{3.9}$$

Підставляємо вираз для Y^L:

$$\min \Sigma \big(Y - (A_0^L + A_1^L X_1 + A_2^L X_2 + \dots + A_n^L X_n)\big)^2. \tag{3.10}$$

Цю задачу також можна розв'язати методом найменших квадратів або іншими методами оптимізації.

Отримавши оптимальні значення параметрів $A_0^U, A_1^U, A_2^U, ..., A_n^U$та $A_0^L, A_1^L, A_2^L, ..., A_n^L$ для верхньої та нижньої границь, ми можемо побудувати модель для нечіткої лінійної регресії, яка апроксимує інтервал:

$$Y = [A_0^L, A_0^U] + [A_1^L, A_1^U]X_1 + [A_2^L, A_2^U]X_2 + \dots + [A_n^L, A_n^U]X_{n,} \tag{3.11}$$

де $\left[A_i^L, A_i^U\right]$ позначає нечітке число типу-2 з верхньою границею A_i^U та нижньою границею A_i^L.

Таким чином, інтервальний підхід до регресійної задачі з нечіткими числами типу-2 полягає у розбитті задачі на дві підзадачі для верхньої та нижньої границі інтервалу, розв'язанні кожної підзадачі за допомогою традиційних методів лінійної регресії, а потім об'єднанні результатів для отримання моделі нечіткої лінійної регресії.

Отримана модель нечіткої лінійної регресії дозволяє здійснювати прогнозування з інтервальною не визначеністю. Замість точного значення прогнозу, ми отримуємо інтервал, який вказує на можливість варіацій в прогнозі.

Слід звернути увагу, що інтервальний підхід до регресійної задачі з нечіткими числами типу-2 є одним з підходів, які можуть бути використані для моделювання нечітких даних. Інші методи, такі як нечітка регресія або гібридні моделі, також можуть бути застосовані залежно від контексту задачі і доступних даних.

Загальний підхід із використанням інтервального підходу до регресійної задачі з нечіткими числами типу-2 полягає у розбитті задачі на підзадачі для верхньої та нижньої границь інтервалу, розв'язанні кожної підзадачі окремо традиційними методами лінійної регресії, а потім об'єднанні результатів для отримання нечіткої моделі регресії.

3.2 Алгоритм оптимізації

Розглянемо підхід, що полягає в застосуванні алгоритмів оптимізації, таких як градієнтний спуск, генетичні алгоритми або алгоритми рою часток, для пошуку оптимальних параметрів моделі. Ці алгоритми можуть бути адаптовані для роботи з нечіткими числами типу-2, що дозволяє знайти оптимальні параметри, мінімізуючи деяку визначену функцію втрат для нечіткої лінійної регресії. Можна використовувати алгоритми оптимізації для мінімізації функції втрат, наприклад середньоквадратичну помилку:

$$\min_{A_0,\dots,A_n} 1/N \sum_{i=1}^{N} \left(Y_i - (A_0 + A_1 X_{1i} + \cdots + A_n X_{ni})\right)^2, \qquad (3.12)$$

де Y_i - істинне значення залежної змінної, N - кількість спостережень.

3.3. Метод на основі нечіткої логіки

Методи нечіткої логіки також можуть бути використані для розв'язання задачі лінійної регресії з параметрами в формі нечітких чисел типу-2. Наприклад, можна використовувати правила нечіткої логіки для встановлення залежностей між змінними, а потім розв'язати задачу, використовуючи методи нечіткої інференції та дефазифікації.

3.4 Генетичний алгоритм.

Генетичний алгоритм (ГА) може бути адаптований для розв'язання задачі лінійної регресії з параметрами в формі нечітких чисел типу-2. Основні кроки ГА можуть бути сформульовані наступним чином:

1. Ініціалізація популяції: створити випадкову популяцію розміром P з нечітких чисел типу-2, де кожний індивід представляє параметри лінійної регресійної моделі.
2. Оцінка пристосованості: обчислити пристосованість кожного індивіда в популяції, використовуючи функцію втрат, таку як середньоквадратична помилка:

$$fitness(A_0, \dots, A_n) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - (A_0 + A_1X_{1i} + \dots + A_nX_{ni})\right)^2. \quad (3.13)$$

3. Селекція: вибрати батьків для наступного покоління на основі їх пристосованості, використовуючи методи селекції, такі як рулеткове колесо або турнірна селекція.
4. Скрещівання: створити нові індивіди шляхом комбінації генів батьків за допомогою операцій скрещівання, таких як одноточкове, двоточкове або уніформне скрещівання.
5. Мутація: з невеликою ймовірністю змінити гени деяких індивідів у популяції, щоб забезпечити різноманітність та уникнути

преждевременной збіжності.

6. Заміна: замінити стару популяцію новою, що складається з нащадків, отриманих на кроках 3-5.
7. Повторіть кроки 2-6 до досягнення критеріїв зупинки, таких як максимальна кількість поколінь або задана точність.

3.5. Приклади застосування розроблених підходів

Приклад 1. Припустимо, у нас є набір даних з відомими значеннями X та Y, де X - це вік старих автомобілів, а Y - їх вартість. Ці значення можуть бути представлені у вигляді нечітких чисел типу-2 через неоднорідність вартості автомобілів та невизначеність у визначенні віку. Наша мета - побудувати модель лінійної регресії з нечіткими параметрами типу-2, яка враховує ці невизначеності та забезпечує точні прогнози вартості автомобілів.

Вхідні дані (X, Y): (5, 15), (7, 12), (12, 8), (15, 6), (20, 5).

Для спрощення прикладу ми використаємо інтервальний підхід для представлення нечітких чисел типу-2 та зосередимося на генетичному алгоритмі для вирішення задачі.

1. Ініціалізація популяції: Створимо випадкову популяцію з 50 індивідів, де кожний індивід представляє параметри A_0 та A_1 лінійної регресії:
$$Y = A_0 + A_1 * X.$$
2. Оцінка пристосованості: Обчислимо пристосованість кожного індивіда в популяції, використовуючи середньоквадратичну помилку:
$$fitness(A_0, A_1) = 1/N * sum\big(Y_i - (A_0 + A_1 * X_i)\big)^2.$$
3. Повторюємо кроки селекції, скрещівання, мутації та заміни, поки не досягнемо критеріїв зупинки, наприклад, після 1000 поколінь або

досягнення заданої точності.

4. Знайдемо кращий індивід у популяції, який представляє оптимальні параметри лінійної регресії для заданого набору даних. Нехай кращий індивід має параметри A_0 та A_1, тоді наша оптимальна лінійна регресійна модель з нечіткими параметрами типу-2 буде мати такий вигляд:

$$Y = A_0 + A_1 * X.$$

5. Використовуючи отриману модель, ми можемо прогнозувати вартість автомобілів для нових значень віку X. Наприклад, якщо ми хочемо прогнозувати вартість автомобіля віком 10 років, ми підставимо значення $X = 10$ у нашу модель:

$$Y = A_0 + A_1 * 10.$$

Варто звернути увагу, що через стохастичну природу генетичних алгоритмів, результати можуть відрізнятися при кожному виконанні алгоритму. Також, для отримання більш точних результатів, можливо знадобиться налаштувати параметри генетичного алгоритму, такі як розмір популяції, кількість поколінь, ймовірності схрещування та мутації тощо.

Приклад 2. Нехай у нас є набір даних з відомими значеннями X та Y, де X - це вік старих автомобілів, а Y - їх вартість. Ці значення можуть бути представлені у вигляді інтервалів через неоднорідність вартості автомобілів та невизначеність у визначенні віку. Наша мета - побудувати модель лінійної регресії з інтервальними параметрами, яка враховує ці невизначеності та забезпечує точні прогнози вартості автомобілів.

Вхідні дані (X, Y):

X: ([4, 6], [6, 8], [11, 13], [14, 16], [19, 21])

Y: ([13, 17], [10, 14], [6, 10], [4, 8], [3, 7]).

Щоб вирішити цю задачу за допомогою інтервального методу, ми можемо використати наступний підхід:

1. Замість стандартної лінійної регресійної моделі $Y = A_0 + A_1 * X$, ми

будемо використовувати інтервальну модель: $Y = [A_0^-, A_0^+] + [A_1^-, A_1^+] * X$.

2. Використовуючи метод найменших квадратів, ми можемо знайти інтервальні оцінки параметрів A_0 та A_1. Для цього ми обчислюємо інтервальні середні значення X та Y, а також інтервальні значення коваріації та дисперсії для X.
3. Визначимо інтервальні оцінки параметрів A_0 та A_1:

$$A_1 = \left(Cov(X, Y)/Var(X)\right), \tag{3.14}$$

$$A_0 = Y_{mean} - A_1 * X_{mean}. \tag{3.15}$$

4. Використовуючи отримані інтервальні оцінки параметрів A_0 та A_1 ми можемо побудувати інтервальну лінійну регресійну модель:

$$Y = [A_0^-, A_0^+] + [A_1^-, A_1^+] * X. \tag{3.16}$$

5. За допомогою цієї інтервальної лінійної регресійної моделі ми можемо прогнозувати вартість автомобілів для нових значень віку X. Наприклад, якщо ми хочемо прогнозувати вартість автомобіля віком 10 років, ми підставимо значення $X = 10$ у нашу модель:

$$Y = [A_0^-, A_0^+] + [A_1^-, A_1^+] * 10. \tag{3.17}$$

Отримане значення Y буде інтервалом, який відображає прогнозовану вартість автомобіля з урахуванням невизначеності в даних.

У цьому прикладі ми розглянули задачу лінійної регресії з інтервальними параметрами, яка дозволяє враховувати невизначеність в даних. Інтервальний метод є простим та ефективним підходом для роботи з такими задачами, але він має свої обмеження, оскільки не завжди забезпечує достатню точність прогнозів у випадках з високою ступенем невизначеності.

Висновки

У цій роботі досліджено задачу побудови лінійної регресії з параметрами в формі нечітких чисел типу-2. Сформульовано постановку задачі та запропоновано методи розв'язання, зокрема такі, що засновані на інтервальному підході, на алгоритмах оптимізації та на методах нечіткої логіки. Використання нечітких чисел типу-2 у лінійній регресії дозволяє враховувати нечіткість даних та отримувати більш адекватні моделі для прогнозування та аналізу залежностей між змінними.

На основі проведених досліджень можна зробити такі висновки.

1. Інтервальний метод є простим та ефективним підходом для роботи з задачами лінійної регресії з нечіткими параметрами. Він дозволяє швидко побудувати регресійну модель, яка відображає невизначеність в даних. Однак, цей метод має свої обмеження, оскільки не завжди забезпечує достатню точність прогнозів у випадках з високою ступенем невизначеності.

2. Генетичний алгоритм є гнучким та адаптивним підходом до розв'язання задачі лінійної регресії з нечіткими параметрами. Він може бути налаштований для різних типів задач та даних, але потребує додаткового часу на обчислення та налаштування параметрів.

На основі проведених досліджень можна запропонувати такі напрямки майбутніх досліджень.

1. Застосування розглянутих методів до інших типів регресійних моделей, таких як поліноміальна, логістична та множинна регресія, для розширення можливостей аналізу та прогнозування даних з невизначеністю та неоднорідністю.

2. Дослідження впливу різних факторів на точність та стабільність регресійних моделей з нечіткими параметрами, таких як розмір даних, рівень невизначеності, види розподілів тощо. Це може допомогти виявити оптимальні умови застосування різних методів та покращити якість аналізу

даних.

3. Розробка методів оцінки та порівняння регресійних моделей з нечіткими параметрами, що базуються на різних критеріях, таких як точність прогнозів, стабільність, робастність, складність моделей та час обчислень. Це дозволить вибрати найкращі методи для конкретних задач та даних.

Список використаних джерел

[1] A. F. Shapiro, "Fuzzy Regression Models," Penn State University, Pennsylvania, 2005.

[2] H. Ishibuchi, "Fuzzy regression analysis," *Fuzzy Theory and Systems,* pp. 137-148, 1992.

[3] H. F. Wang and R. C. Tsaur, "Insight of a fuzzy regression model," *Fuzzy Sets and Systems,* vol. 112, no. 3, pp. 355-369, 2000.

[4] H. Tanaka, S. Uejima and K. Asai, "Linear regression analysis with fuzzy model," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 12, no. 6, pp. 903-907, 1982.

[5] Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—II. Information Sciences, 8(4), 301-357.

[6] Mendel, J. M. (2001). Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall.

[7] Melin, P., & Castillo, O. (2008). Type-2 Fuzzy Logic: Theory and Applications. Springer.

[8] Wu, H., & Mendel, J. M. (2002). Uncertainty bounds and their use in the design of interval type-2 fuzzy logic systems. IEEE Transactions on Fuzzy Systems, 10(5), 622-639.

[9] Karnik, N. N., & Mendel, J. M. (1998). An introduction to type-2 fuzzy logic systems. In Proceedings of the IEEE International Conference on Fuzzy Systems, 735-740.

[10] Nguyen, H. T., Prasad, N. R., Walker, C. L., & Walker, E. A. (2006). A First Course in Fuzzy and Neural Control. CRC Press.