# *Шеремет Валерій Сергійович, аспірант, Дніпровський національний університет імені Олеся Гончара, м. Дніпро;*

# *Мацуга Ольга Миколаївна, кандидат технічних наук, доцент, Дніпровський національний університет імені Олеся Гончара, м. Дніпро*

**ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ВІДНОВЛЕННЯ КУСКОВИХ РЕГРЕСІЙ З ОДНИМ ВУЗЛОМ**

**Вступ.** Відновлення регресійних залежностей за експериментальними даними є актуальною задачею, яка часто постає в інформаційних системах і технологіях аналізу даних. Важливим класом таких залежностей є кускові регресії з одним вузлом. Існуюче програмне забезпечення здебільшого орієнтоване на відновлення кусково-лінійних регресій [3–5], в той час як кусково-нелінійні можуть бути більш адекватними та вірогідними. У зв’язку з цим у роботі було поставлено за мету розробити програмне забезпечення для відновлення кусково-нелінійних регресій з одним вузлом.

**Постановка задачі.** Розглядається модель кускової регресії з одним вузлом $v$

$\overbar{y}\left(x\right)=\left\{\begin{array}{c}f\left(x;a\_{1},b\_{1}\right),x\leq v,\\f\left(x;a\_{2},b\_{2}\right),x\geq v,\end{array}\right.$ (1)

яка є неперервною у вузлі $v$, тобто $f\left(v;a\_{1},b\_{1}\right)=f\left(v;a\_{2},b\_{2}\right)$.

В якості функції $f\left(x;a,b\right)$ розглядається лінійна та одна з 11-ти квазілінійних функцій:

* $f\left(x;a,b\right)=ax+b$;
* $f\left(x;a,b\right)=\frac{a}{\frac{1}{x}+b}$;
* $f\left(x;a,b\right)=\frac{1}{aexp\left(-x\right)+b}$;
* $f\left(x;a,b\right)=ax^{3}+b$;
* $f\left(x;a,b\right)=bexp\left(\frac{a}{x}\right)$;
* $f\left(x;a,b\right)=bexp\left(ax\right)$;
* $f\left(x;a,b\right)=\sqrt{ax+b}$;
* $f\left(x;a,b\right)=\frac{a}{x}+b$;
* $f\left(x;a,b\right)=\frac{1}{ax+b}$;
* $f\left(x;a,b\right)=\frac{x}{ax+b}$;
* $f\left(x;a,b\right)=alnx+b$;
* $f\left(x;a,b\right)=bx^{a}$.

Ставиться задача оцінити параметри $a\_{1},b\_{1},a\_{2},b\_{2}$ та вузол $v$ моделі (1) за даними $\left\{x\_{i},y\_{i};i=\overbar{1,N}\right\}$ за заданої функції $f\left(x;a,b\right)$. Враховуючи умову неперервності моделі у вузлі, кількість параметрів, що підлягають оцінюванню в дійсності на один менше (параметр $b\_{2}$ можна виразити через $a\_{1},b\_{1},a\_{2},v$).

Для розв’язання поставленої задачі необхідно розробити програмне забезпечення, яке забезпечить знаходження оцінок параметрів і вузла склеювання кусково-лінійної та 11-ти кусково-нелінійних моделей, а також перевірку адекватності відновлених моделей.

**Основний матеріал.** У процесі роботи було розроблено програмне забезпечення засобами .NET Framework 4.6.1 та мови програмування C# 6.0. Програмне забезпечення надає користувачу наступні можливості:

1. Завантаження даних з файлу.
2. Генерація даних згідно кусково-лінійної та 11-ти вищеперелічених кусково-нелінійних регресій. Залишки генеруються за нормальним розподілом з нульовим математичним сподіванням і заданим користувачем середньоквадратичним відхиленням.
3. Оцінювання параметрів $a\_{1},b\_{1},a\_{2},b\_{2}$ та вузла $v$ моделі (1) за завантаженими або згенерованими даними за заданої функції $f\left(x;a,b\right)$.

Для відновлення кусково-лінійної регресії (тобто коли $f\left(x;a,b\right)$ лінійна) реалізовано два алгоритми [1]. Класичний алгоритм передбачає, що вибірка відсортована за зростанням показника *х* і вузол співпадає з одним з елементів вибірки. Вузол склеювання знаходиться шляхом перебору елементів вибірки, а оцінки параметрів визначаються за фіксованого вузла з умови мінімуму залишкової дисперсії $S\_{Зал}^{2}\left(v\right)$. Модифікований алгоритм відрізняється від класичного припущенням про те, що $S\_{Зал}^{2}\left(v\right)$ можна представити як суму двох незалежних доданків, які мінімізуються окремо.

Для обох алгоритмів пропонуються три варіанти реалізації [2]:

* «наївна» – в цій реалізації для кожного вузла $v=x\_{k}, k=\overbar{3,N-2}$ здійснюється прохід по всій вибірці, щоб сформувати систему лінійних алгебраїчних рівнянь (СЛАР), з якої визначаються оцінки параметрів; така реалізація має квадратичну складність;
* ефективна – з метою більш швидкого формування СЛАР пропонується попередньо розрахувати суми $\sum\_{i=1}^{k}x\_{i}$, $\sum\_{i=1}^{k}x\_{i}^{2}$, $\sum\_{i=1}^{k}y\_{i}$, $\sum\_{i=1}^{k}y\_{i}^{2}$, $\sum\_{i=1}^{k}x\_{i}y\_{i}$ (це можна зробити за один прохід відсортованою вибіркою); тоді на етапі формування СЛАР потрібні для неї суми від 1 до $k$ будуть вже відомі, а суми від $k$ + 1 до $N$ можна обчислити як різниці між сумами від 1 до $N$ та сумами від 1 до $k$, що знижує складність алгоритму до лінійної;
* реалізація, в якій для пошуку вузла склеювання використовується метод оптимізації; у роботі реалізовано метод золотого перетину.

Відновлення кускових регресій у випадку різних квазілінійних функцій$f\left(x;a,b\right)$ базується на їх приведенні до кусково-лінійних [1]. Здійснюється перетворення початкових даних до вигляду $\left\{X\_{i},Y\_{i};i=\overbar{1,N}\right\}$, де $X\_{i}=φ\_{1}\left(x\_{i}\right)$,
$Y\_{i}=φ\_{2}\left(y\_{i}\right)$, $φ\_{1}$ і $φ\_{2}$ залежать від виду квазілінійної функції. На основі перетворених даних оцінюються параметри, включаючи вузол, кусково-лінійної регресії. Після зворотного перетворення одержуються оцінки параметрів кусково-нелінійної моделі.

1. Візуалізація даних та відновленої моделі регресії на графіку, побудова діаграми залишків для оцінювання адекватності відновленої моделі.
2. Обчислення та виведення на екран метрик якості (корінь із залишкової дисперсії та коефіцієнт детермінації R2) і часу роботи алгоритмів відновлення.

Після запуску програми користувач бачить вікно з декількома вкладками. На першій вкладці можна завантажити або згенерувати дані (рис. 1). Для генерації даних потрібно обрати модель регресії та вказати її параметри.



Рисунок 1 – Вкладка для генерації та завантаження даних

На вкладці «Результати відновлення» можна обрати алгоритм для відновлення (рис. 2). Знайдені оцінки параметрів виводяться нижче. Поруч на графіку відображаються дані (сині точки) та відновлена регресія (червона лінія).



Рисунок 2 – Вкладка з результатами відновлення регресії
(параметри, з якими генерувалися дані, вказані на рисунку 1)

Тестування розробленого програмного забезпечення було проведено на згенерованих даних і засвідчило адекватність та швидкість його роботи.

**Висновки.** У роботі було розроблено програмне забезпечення мовою програмування C# на платформі .Net, яке дозволяє проводити відновлення кусково-лінійної та 11-ти кусково-квазілінійних регресії з одним вузлом. Для відновлення кусково-лінійної регресії реалізовано класичний та модифікований алгоритми в «наївній» та ефективній версіях, а також з використанням методу оптимізації. Відновлення усіх кусково-квазілінійних регресій реалізовано на основі їх зведення до кусково-лінійної регресії.

**Література**

1. Приставка О.П., Байбуз О.Г., Ємел’яненко Т.Г. Методи та алгоритми сплайн-регресійного аналізу. Дніпропетровськ: Вид-во ДНУ, 2012. 144 с.
2. Мацуга О.М., Шеремет В.С. Ефективні алгоритми відновлення кусково-лінійної регресії з одним вузлом. *Актуальні проблеми автоматизації та інформаційних технологій*: зб. наук. пр. Дніпро, 2021. Т. 25. С. 119-130.
3. Jekel C., Venter G. pwlf. A Python Library for Fitting 1D Continuous Piecewise Linear Functions. 2019. 15 p. URL: <https://jekel.me/assets/papers/pwlf_Jekel_Venter_v2.pdf> (дата звернення: 10.01.2023).
4. Stadnik V. Segmented Linear Regression. URL: [https://www.codeproject.com/Articles/5282014/Segmented-Linear-Regression](https://www.codeproject.com/Articles/5282014/Segmented-Linear-Regression#segmented) (дата звернення: 10.01.2023).
5. Muggeo V.M.R. segmented: An R Package to Fit Regression Models with Broken-Line Relations. *R News*. Vol. 8/1. May 2008. P. 20-25.