СИСТЕМНИЙ АНАЛІЗ

УДК 519.21

П.С. КНОПОВ
Інститут кібернетики ім. В.М. Глушкова НАН України, Київ, Україна,
e-mail: *knopov1@yahoo.com*.

А.С. КОРХІН
Придніпровська державна академія будівництв та архітектури, Дніпро, Україна,
e-mail: *a.s.korkhin@gmail.com*.

ВИЗНАЧЕННЯ КУСКОВО-ЛІНІЙНОГО ТРЕНДУ НЕСТАЦІОНАРНОГО ЧАСОВОГО РЯДУ НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ. I. ОПИС ТА ОБҐРУНТУВАННЯ МЕТОДУ

Анотація. Запропоновано розглядати тренд нестаціонарного часового ряду як лінійну регресію з невідомими точками перемикання. Описано та обгрунтовано метод оцінювання точок перемикання, що базується на інтелектуальному аналізі даних з використанням статистичних критеріїв.

Ключові слова: лінійна регресія, алгоритм, часовий ряд, тренд, методи, математичне програмування.

ВСТУП

Більшість методів побудови регресії з перемиканнями, коли точки перемикання невідомі, зводяться до розв’язання різних оптимізаційних задач зі змішаними (цілочисельними та неперервними) шуканими величинами. У цьому напрямку отримано багато загальнотеоретичних результатів, а також розв’язано важливих практичних задач економіки і кліматології [1–4] та [5–8].

Труднощі щодо розв’язання задач оптимізації зі змішаними змінними, які відомі і наведені в зазначених вище роботах, це: необхідність апріорного знання кількості точок перемикань, нелінійність задач оптимізації, наявність у них цілочисельних змінних. Для уникнення цих недоліків потрібно відмовитися від використання схем змішаної оптимізації. Так, у статті [3] запропоновано квазібаєсівську процедуру, яку застосовують в асимптотиці.

У цій роботі розглянуто інший підхід до визначення точок перемикань, що можна застосовувати у реальному часі чи на скінченних часових інтервалах. Цей підхід є продовженням покрокової побудови регресії з перемиканнями, запропонованої в [8]. Він ґрунтується на використанні статистичних критеріїв стосовно парної регресії з перемиканнями, де незалежною змінною є час. Цей вид регресії попри те, що простий на вигляд, нерідко використовують на практиці для визначення тренду нестаціонарних часових рядів. Його окремий випадок, коли лінія реґресії неперервна в точках перемикання, є лінійним сплайном. Підхід знайшов широке застосування для розв’язання задач аналізу та прогнозування в економетрії [9], кліматології для відстеження зміни природних факторів та в інших галузях, а також для апроксимації нелінійних функцій.

1. МОДЕЛЬ РЕГРЕСІЇ

Розглянемо регресію, яка є окремим випадком моделі [5] і має вигляд

$$y_t = \mathbf{x}'_t \boldsymbol{\alpha}^0_{ti} + \varepsilon_t, \; i = 1, 2, \ldots, \; t = 1, 2, \ldots \tag{1}$$

(штрих в індексах означає транспонування), $\mathbf{x}_t = [1 \;\; t]'$ — незалежні змінні (регресори), $\boldsymbol{\alpha}^0_{ti} \in \mathbb{R}^2$ — дійсні невідомі величини параметрів регресії, ε_t — випадкова величина, яка дає змогу враховувати незначний вплив другорядних величин на y_t, при цьому

$$\boldsymbol{\alpha}^0_{ti} = \boldsymbol{\alpha}^0_i = \text{const} \;\text{ для }\; t \in I^0_i = [t^0_{i-1} + 1, t^0_i], \; i = 1, 2, \ldots, \; t^0_0 = 0, \tag{2}$$

де t^0_i — i-та точка перемикання.

На відміну від [5] вважатимемо довжину інтервалу спостереження неперервно зростальною величиною, починаючи з $T = L$, $L \geq 3$. Точки перемикання t^0_i, $i = 1, 2, \ldots$, вважають невідомими, як і їхня кількість, яка може збільшуватись зі зростанням T.

Щодо випадкової величини ε_t в (1) введемо стандартне припущення.

Припущення 1. Випадкові величини ε_t, $t = 1, 2, \ldots$, нормально розподілені, мають однакові математичні сподівання та дисперсії, а саме $E\{\varepsilon_t\} = 0$, $E\{\varepsilon^2_t\} = \sigma^2$, $t = 1, 2, \ldots$; вони попарно некорельовані: $E\{\varepsilon_t \varepsilon_\tau\} = 0$, $t, \tau = 1, 2, \ldots$, $t \neq \tau$.

За цим припущенням маємо

$$E(t) = E\{y_t\} = \mathbf{x}'_t \boldsymbol{\alpha}^0_i, \; i = 1, 2, \ldots, \; t = 1, 2, \ldots \tag{3}$$

З (1), (2) випливає, що математичне сподівання $E(t)$, яке вважають функцією неперервного t, може мати або не мати розриви в точках перемикання.

Введемо ще таке припущення.

Припущення 2. Відстань між двома сусідніми точками перемикань має бути достатньо великою.

Відповідно до [10] інтервал спостереження в 5–7 разів повинен перевищувати кількість параметрів регресії, що оцінюються. Тоді для розглядуваної задачі

$$|t^0_i - t^0_{i-1}| \geq cn, \; i = 1, 2, \ldots, \tag{4}$$

де n — кількість параметрів регресії разом з вільним членом; вважатимемо $c = 3$.

Це співвідношення є орієнтовним. Воно потрібне для визначення помилкових точок перемикання, що зумовлені дією випадкових факторів.

Запропоновано визначати точки перемикань послідовно із зростанням T, використовуючи різні статистичні індикатори. У цій статті розглянемо два індикатори.

2. СТАТИСТИЧНІ ІНДИКАТОРИ

2.1. Регресія зі змінною точкою перемикання (перший індикатор). Нехай відомі значення y_t для $t = 1, \ldots, T + L$, оцінено $i = i(T)$ точок перемикания, до того ж

$$T \geq \hat{t}_i + L, \; L \geq 3, \tag{5}$$

де $T > \hat{t}_i$ — оцінка i-ї точки перемикания.

Потрібно перевірити гіпотезу щодо наявності перемикання в періоді часу T. Для розв’язання цієї задачі побудуємо три регресії:

1) $y_t = \mathbf{x}'_t \boldsymbol{\alpha}_1^0 + \varepsilon_t$, $\boldsymbol{\alpha}_1^0 = \text{const}$, $t \in [\hat{t}_i + 1, T]$, оцінка методом найменших квадратів параметра регресії $\hat{\boldsymbol{\alpha}}_1 = \hat{\boldsymbol{\alpha}}_1(\hat{t}_i + 1, T) = \arg \min\limits_{\boldsymbol{\alpha}_1 \in \mathbb{R}^2} \sum\limits_{t=\hat{t}_i+1}^{T} (y_t - \mathbf{x}'_t \boldsymbol{\alpha}_1)^2$, сума квадратів залишків $S(\hat{t}_i + 1, T) = \sum\limits_{t=\hat{t}_i+1}^{T} (y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}_1(\hat{t}_i + 1, T))^2$;

2) $y_t = \mathbf{x}'_t \boldsymbol{\alpha}_2^0 + \varepsilon_t$, $\boldsymbol{\alpha}_2^0 = \text{const}$, $t \in [\hat{t}_i + 1, T + L]$, оцінка методом найменших квадратів параметра регресії $\hat{\boldsymbol{\alpha}}_2 = \hat{\boldsymbol{\alpha}}_2(\hat{t}_i + 1, T + L) = \arg \min\limits_{\boldsymbol{\alpha}_2 \in \mathbb{R}^2} \sum\limits_{t=\hat{t}_i+1}^{T+L} (y_t - \mathbf{x}'_t \boldsymbol{\alpha}_2)^2$, сума квадратів залишків $S(\hat{t}_i + 1, T + L) = \sum\limits_{t=\hat{t}_i+1}^{T+L} (y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}_2(\hat{t}_i + 1, T + L))^2$;

3) $y_t = \mathbf{x}'_t \boldsymbol{\alpha}_3^0 + \varepsilon_t$, $\boldsymbol{\alpha}_3^0 = \text{const}$, $t \in [T_i + 1, T + L]$, оцінка методом найменших квадратів параметра регресії $\hat{\boldsymbol{\alpha}}_3 = \hat{\boldsymbol{\alpha}}_3(T + 1, T + L) = \arg \min\limits_{\boldsymbol{\alpha}_3 \in \mathbb{R}^2} \sum\limits_{t=T+1}^{T+L} (y_t - \mathbf{x}'_t \boldsymbol{\alpha}_3)^2$, сума квадратів звлишків $S(T + 1, T + L) = \sum\limits_{t=T+1}^{T+L} (y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}_3(T + 1, T + L))^2$.

На інтервалі спостереження $[\hat{t}_i + 1, T + L]$ дані можна записати за другою регресією або двома регресіями — першою і третьою, що утворюють регресію з перемиканням в точці $t = T$.

Не обмежуючи загальності міркувань щодо критерію визначення точки перемикання, далі вважатимемо, що $\mathbf{x}_t \in \mathbb{R}^n$. Позначимо $M_i = T - \hat{t}_i$, $N_i = M_i + L$; $\mathbf{Y}_1 \in \mathbb{R}^{M_i}$, $\mathbf{Y}_2 \in \mathbb{R}^{N_i}$, $\mathbf{Y}_3 \in \mathbb{R}^L$ — вектори, компонентами яких є відповідно y_t, $t \in [\hat{t}_i + 1, T]$; y_t, $t \in [\hat{t}_i + 1, T + L]$; y_t, $t \in [T + 1, T + L]$; $\mathbf{X}_1(M_i \times n)$, $\mathbf{X}_2(N_i \times n)$, $\mathbf{X}_3(L \times n)$ — матриці, рядками яких є відповідно $\mathbf{x}'_t$, $t \in [\hat{t}_i + 1, T]$; $\mathbf{x}'_t$, $t \in [\hat{t}_i + 1, T + L]$; $\mathbf{x}'_t$, $t \in [T + 1, T + L]$; $\boldsymbol{\varepsilon}_1 \in \mathbb{R}^{M_i}$, $\boldsymbol{\varepsilon}_2 \in \mathbb{R}^{N_i}$, $\boldsymbol{\varepsilon}_3 \in \mathbb{R}^L$ — вектори, компонентами яких є відповідно ε_t, $t \in [\hat{t}_i + 1, T]$; ε_t, $t \in [\hat{t}_i + 1, T + L]$; ε_t, $t \in [T + 1, T + L]$.

Тоді побудовані вище три регресії набудуть вигляду

$$\mathbf{Y}_1 = \mathbf{X}_1 \boldsymbol{\alpha}_1^0 + \boldsymbol{\varepsilon}_1, \ \mathbf{Y}_2 = \mathbf{X}_2 \boldsymbol{\alpha}_2^0 + \boldsymbol{\varepsilon}_2, \ \mathbf{Y}_3 = \mathbf{X}_3 \boldsymbol{\alpha}_3^0 + \boldsymbol{\varepsilon}_3.$$

Маємо $\mathbf{Y}_2 = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_3 \end{bmatrix}$, $\mathbf{X}_2 = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_3 \end{bmatrix}$, $\boldsymbol{\varepsilon}_2 = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_3 \end{bmatrix}$.

Звідси випливає формула регресії з перемиканням у точці $t = T$:

$$\mathbf{Y}_2 = \begin{bmatrix} \mathbf{X}_1 & \mathbf{O}_{M_i n} \\ \mathbf{O}'_{M_i n} & \mathbf{X}_3 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1^0 \\ \boldsymbol{\alpha}_3^0 \end{bmatrix} + \boldsymbol{\varepsilon}_2, \tag{6}$$

де $\mathbf{O}_{M_i n}$ — нульова матриця розміру $M_i \times n$.

Ця регресія є альтернативою другої регресії $\mathbf{Y}_2 = \mathbf{X}_2 \boldsymbol{\alpha}_2^0 + \boldsymbol{\varepsilon}_2$, яка збігається з (6), якщо $\boldsymbol{\alpha}_1^0 = \boldsymbol{\alpha}_3^0$, а звідси випливає $\boldsymbol{\alpha}_1^0 = \boldsymbol{\alpha}_3^0 = \boldsymbol{\alpha}_2^0$. Вибір із двох регресій можна зробити, перевіряючи лінійну гіпотезу щодо параметрів регресії (6): $H_0 \colon \boldsymbol{\alpha}_1^0 = \boldsymbol{\alpha}_3^0$.

Альтернативою нульової гіпотези є гіпотеза H_1: $\boldsymbol{\alpha}_1^0 \neq \boldsymbol{\alpha}_3^0$.

Відповідно до припущення 2 величину $\hat{t}_i$ можна розглядати як фіксовану. Якщо припущення 1 також має місце, то відповідно до [11, 12] величина

$$F[\hat{t}_i+1, T+L] = \frac{||\mathbf{Y}_2 - \mathbf{X}_2\hat{\boldsymbol{\alpha}}_2||^2 - ||\mathbf{Y}_1 - \mathbf{X}_1\hat{\boldsymbol{\alpha}}_1||^2 - ||\mathbf{Y}_3 - \mathbf{X}_3\hat{\boldsymbol{\alpha}}_3||^2}{||\mathbf{Y}_1 - \mathbf{X}_1\hat{\boldsymbol{\alpha}}_1||^2 + ||\mathbf{Y}_3 - \mathbf{X}_3\hat{\boldsymbol{\alpha}}_3||^2} \times$$

$$\times \frac{T - \hat{t}_i + L - 2n}{n}, \quad i = 0, 1, \ldots, \tag{7}$$

має F-розподіл з числом ступенів свободи $q_1 = n$ (кількість обмежень на параметри регресії (6), які потрібно ввести для перетворення (6) у регресію без точки перемикання) і $q_2 = T - \hat{t}_i + L - 2n$ (загальна кількість спостережень $T + L - \hat{t}_i$ без кількості оцінюваних параметрів у регресії (6)).

Критерій (7) був запрорононований у [13] для порівняння першої та третьої регресій з другою регресією з метою вибору найбільш коректної. Як випливає з наведених вище міркувань, критерій з [13] є окремим випадком загальної лінійної гіпотези щодо параметрів лінійної регресії.

Нульова гіпотеза H_0 не відхиляється на $100p\%$-му рівні, якщо виконується нерівність

$$F(\hat{t}_i + 1, T + L) \leq F_p(q_1, q_2), \quad i = 0, 1, \ldots, \tag{8}$$

де $F_p(q_1, q_2)$ — верхня $100p\%$-ва точка розподілу Фішера.

У цьому випадку в точці $t = T$ перемикання немає. Тому починається перевірка нульової гіпотези за наявності точки перемикання в точці $t = T + 1$: обчислюється $F(\hat{t}_i + 1, T + 1 + L)$, а потім перевіряється умова (7) для цієї величини F-критерію.

В іншому випадку приймають альтернативну гіпотезу H_1: $\boldsymbol{\alpha}_1^0 \neq \boldsymbol{\alpha}_3^0$. Наступна точка перемикання $\hat{t}_{i+1} = T$. Далі повторюється описана вище процедура визначення оцінки точки перемикання $\hat{t}_{i+2}$.

За наявності декількох періодів часу T, що належать певному інтервалу часу J і задовольняють умову (8), як шукана величина вибирається таке T, якому відповідає максимум $F(\hat{t}_i + 1, T + L)$ на J або перший максимум у разі, коли їх декілька.

Розглянемо можливість рекурентного обчислення F згідно з (7). Запишемо цю формулу по-іншому для $n = 2$:

$$F(\hat{t}_i + 1, T + L) = \frac{S(\hat{t}_i + 1, T + L) - S(\hat{t}_i + 1, T) - S(T + 1, T + L)}{S(\hat{t}_i + 1, T) + S(T + 1, T + L)} \times$$

$$\times \frac{T - \hat{t}_i + L - 4}{2}, \quad i = 0, 1, \ldots \tag{9}$$

Маємо послідовність інтервалів часу:

$$[\hat{t}_i + 1, \hat{t}_i + 3], \; [\hat{t}_i + 1, \hat{t}_i + 4], \; [\hat{t}_i + 1, \hat{t}_i + 5], \ldots, [\hat{t}_i + 1, T + L]. \tag{10}$$

Для кожного інтервалу в цій послідовності методом найменших квадратів знайдемо оцінки багатомірного параметра регресії

$$y_t = \mathbf{x}_t' \boldsymbol{\alpha}^0 + \varepsilon_t, \tag{11}$$

$$\boldsymbol{\alpha}(\hat{t}_i + 1, \hat{t}_i + 3), \; \boldsymbol{\alpha}(\hat{t}_i + 1, \hat{t}_i + 4), \; \boldsymbol{\alpha}(\hat{t}_i + 1, \hat{t}_i + 5), \ldots, \boldsymbol{\alpha}(\hat{t}_i + 1, T + L) \tag{12}$$

та відповідні їм суми квадратів залишків

$$S(\hat{t}_i+1, \hat{t}_i+3),\ S(\hat{t}_i+1, \hat{t}_i+4),\ S(\hat{t}_i+1, \hat{t}_i+5), \dots, S(\hat{t}_i+1, T+L). \qquad (13)$$

Оскільки кожен наступний інтервал в (10) отримано з попереднього додаванням одного спостереження, то послідовність (12) можна побудовати за допомогою рекурентного співвідношення (див., наприклад, [14]).

Послідовність (13) містить величини $S(\hat{t}_i+1, T)$ та $S(\hat{t}_i+1, T+L)$, що фігурують також у (9).

Для обчислення суми $S(T+1, T+L)$ в (9) розглянемо послідовність інтервалів часу з однаковою довжиною, що дорівнює L:

$$[\hat{t}_i+4, \hat{t}_i+4+L-1], \dots, [T+1, T+L],\ [\hat{t}_i+5, \hat{t}_i+5+L-1],$$
$$[\hat{t}_i+6, \hat{t}_i+6+L-1], \dots, [T+1, T+L]. \qquad (14)$$

Кожен з цих інтервалів отримують з попереднього інтервалу з вилученням першого спостереження та додаванням до нього нового спостереження. Оцінки параметра $\boldsymbol{\alpha}^0$ регресії (11) для інтервалів (14) також можна отримувати рекурентно [14]. Відповідні до них суми квадратів залишків утворюють послідовність

$$S[\hat{t}_i+4, \hat{t}_i+4+L-1],\ S[\hat{t}_i+5, \hat{t}_i+5+L-1],$$
$$S[\hat{t}_i+6, \hat{t}_i+6+L-1], \dots, S(T+1, T+L). \qquad (15)$$

Обчислення нового члена в послідовностях (13) і (15) виконується за умов надходження нового спостереження. Для прискорення обчислень можна одночасно обчислювати ці послідовності.

Зауважимо, що для обчислення в реальному часі інформація про появу нової точки перемикання може бути отримана щонайменше через L періодів часу.

2.2. Довірчі інтервали для прогнозування залежної змінної у періоді часу *T* на *L* періодів часу наперед (другий індикатор). Нехай відоме значення y_t для $t = 1, \dots, T+L$, оцінено $i = i(T)$ точок перемикання і виконується перша умова з (5). Прогнозування у точці $t = T$ на L периодів часу наперед:

$$\hat{y}_{T+\tau} = \mathbf{x}'_{T+\tau}\ \hat{\boldsymbol{\alpha}}(\hat{t}_i+1, T),\ \tau = 1, 2, \dots, L, \qquad (16)$$

де $\hat{\boldsymbol{\alpha}}(\hat{t}_i+1, T) = \arg\min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^2} \sum\limits_{t=\hat{t}_i+1}^{T} (y_t - \mathbf{x}'_t\boldsymbol{\alpha})^2$.

Так само, як у п. 2.1, виконання припущення 2 дає змогу наближено вважати $\hat{t}_i$ константою. Тоді за умов виконання припущення 1 довірчий інтервал для прогнозованої величини, який покриє її з ймовірністю $1-p$, визначається за формулою

$$\hat{y}_{T+\tau} - t_p(q)\hat{\sigma}_f(T+\tau) \le y_{T+\tau} \le \hat{y}_{T+\tau} + t_p(q)\hat{\sigma}_f(T+\tau),\ \tau = 1, 2, \dots, L, \qquad (17)$$

де $t_p(q)$ — це $100p$%-ва точка розподілу Стьюдента з числом ступенів свободи $q = T - \hat{t}_i - 2$; $\hat{\sigma}_f(T+\tau)$ — оцінка середньоквадратичної похибки прогнозу.

Оцінка дисперсії прогнозу визначається за формулою $\hat{\sigma}^2_f(T+\tau) = \hat{\sigma}^2(1 + \mathbf{x}'_{T+\tau}\hat{\mathbf{K}}\mathbf{x}_{T+\tau})$, де $\hat{\mathbf{K}}$ — оцінка кореляційної матриці щодо $\boldsymbol{\alpha}^0$ — багатовимірного параметра регресії (11),

$$\hat{\mathbf{K}} = \hat{\sigma}^2 E\{(\hat{\boldsymbol{\alpha}}(\hat{t}_i+1, T) - \boldsymbol{\alpha}^0)(\hat{\boldsymbol{\alpha}}(\hat{t}_i+1, T) - \boldsymbol{\alpha}^0)'\},$$

де $\hat{\sigma}^2$ — оцінка дисперсії випадкової компоненти в регрессії (11).

Твердження 1. Якщо $\hat{\sigma}^2 > 0$, то $\sigma_f^2(T+2) > \hat{\sigma}_f^2(T+1)$.

Доведення. У цьому випадку $\hat{\mathbf{K}} > 0$, $\hat{\mathbf{K}} = [\hat{k}_{ij}]$, $i, j = 1, 2$; $|\hat{k}_{12}| = |\hat{k}_{21}| < 1$, $\hat{k}_{ii} = 1$, $i = 1, 2$. Оскільки $\mathbf{x}_{T+1} = \begin{bmatrix} 1 \\ T+1 \end{bmatrix}$, $\mathbf{x}_{T+2} = \begin{bmatrix} 1 \\ T+2 \end{bmatrix}$, то $\mathbf{x}_{T+2} = \mathbf{x}_{T+1} + \mathbf{\Delta}$, $\Delta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Тоді отримаємо

$$\mathbf{x}'_{T+2}\hat{\mathbf{K}}\mathbf{x}_{T+2} - \mathbf{x}'_{T+1}\hat{\mathbf{K}}\mathbf{x}_{T+1} = 2\mathbf{x}'_{T+1}\hat{\mathbf{K}}\mathbf{\Delta} + \mathbf{\Delta}'\hat{\mathbf{K}}\mathbf{\Delta} = 2(\hat{k}_{12} + \mathrm{T} + 1) + \hat{k}_{22} > 0.$$

Звідси і з формули $\hat{\sigma}_f^2(T+2) - \hat{\sigma}_f^2(T+1) = \hat{\sigma}^2(\mathbf{x}'_{T+2}\hat{\mathbf{K}}\mathbf{x}_{T+2} - \mathbf{x}'_{T+1}\hat{\mathbf{K}}\mathbf{x}_{T+1})$ маємо доведення твердження. ■

З твердження 1 випливає, що довжина інтервалу (17) є більшою для $\tau = 2$, ніж для $\tau = 1$. Аналогічно довжина інтервалу (17) для $\tau = 3$ буде більша за ширину цього інтервалу для $\tau = 2$ і т.д. Таким чином, зі збільшенням τ довжина надійного інтервалу для y_{T+1} збільшується. Отже, для практичного застосування інтервальних оцінок прогнозованої величини залежної змінної у регресії (11) їхня кількість L має бути достатньо малою. У подальшому вважатимемо $L = 2$.

Зазначимо, що ширину довірчого інтервалу (17) можна зменшувати, застосовуючи методологію з [15]. До того ж зменшення буде найбільшим для незначної кількості спостережень.

Уведемо такі позначення:

подія A ={непотрапляння y_{T+1} у довірчий інтервал (17), якщо $\tau = 1$},

подія B ={непотрапляння y_{T+2} у довірчий інтервал (17), якщо $\tau = 2$}.

Згідно з (17) ймовірності цих подій становлять $P\{A\} = P\{B\} = p$.

Оскільки довжина інтервалу (17) згідно з твердженням 1 для $\tau = 2$ більша, ніж для $\tau = 1$, то у разі появи події A ймовірність появи події B буде не більша за $P\{B\}$, тобто $P\{B \mid A\} \le P\{B\}$.

Ймовірність того, що дійсні величини залежної змінної $y_{T+\tau}$, $\tau = 1, 2$, не будуть покриті відповідними інтервалами, становить

$$P\{A \cdot B\} = P\{A\}P\{B \mid A\} \le P\{A\}P\{B\} = p^2. \tag{18}$$

Нехай, наприклад, $p = 0.224$, тоді ймовірності покриття y_{T+1} і y_{T+2} окремо довірчими інтервалами (17) дорівнюють 0.776. Ймовірність того, що ці величин будуть знаходитися поза відповідними довірчими інтервалами (17), згідно (18) не перевищує $0.224^2 = 0.05$.

Якщо y_{T+1} і y_{T+2} будуть знаходитися водночас поза своїми довірчими інтервалами, це означає, що відбулася малоймовірна подія — непотрапляння прогнозованих величин залежної змінної з (11) в довірчі інтервали в двох послідовних періодах часу. Тому на 5%-му рівні не відхиляється нульова гіпотеза:

$$H_0 : \hat{t}_{i+1} = T. \tag{19}$$

Далі обчислюється оцінка наступної точки перемикання за визначеною вище схемою.

Якщо хоча б одна з величин y_{T+1} або y_{T+2} за умов $p = 0.224$ потрапить у відповідний довірчий інтервал (17), то гіпотеза (19) відхиляється на $P = p^2 100\% = 5\%$-му рівні. Приймається альтернативна гіпотеза $H_1 : \hat{t}_{i+1} \ne T$. Починається перевірка гіпотези H_0 щодо наступної точки $t = T + 1$ і т.д.

Уведемо функції:

$$\varphi_{1P}(T)=\begin{cases}1, \text{ якщо } y_{T+1} \text{ задовольняє (17)},\\ 0, \text{ якщо } y_{T+1} \text{ не задовільняє (17)},\end{cases}$$

$$\varphi_{2P}(T)=\begin{cases}1, \text{ якщо } y_{T+2} \text{ задовольняє (17)},\\ 0, \text{ якщо } y_{T+2} \text{ не задовольняє (17)}.\end{cases}$$

Функцію $\varphi_P(T)=\varphi_{1P}(T)+\varphi_{2P}(T)$, $0\le\varphi(T)\le 2$ назвемо індикаторною.

Якщо $\varphi_p(T)=0$, то щодо сказанного вище гіпотеза (19) не відхиляється на $100p\%$-му рівні. В іншому випадку вона відкидається.

2.3. Перевірка гіпотези щодо помилкової точки перемикання. Вважатимемо, що в (2) $\boldsymbol{\alpha}_1^0\in\mathbb{R}^n$, де $n>0$ — довільне ціле число. Подальші міркування аналогічні виведенню критерію 1 з п. 2.1.

Нехай отримані вибіркові оцінки точок перемикання $\hat{t}_{i-2}=\text{const}$, $\hat{t}_{i-1}=\text{const}$, $\hat{t}_i=\text{const}$, при цьому $\hat{t}_i-\hat{t}_{i-1}<cn$ — умова (4) не задовільняє оцінок точок перемикання, $i=2,3,\ldots$ Щоб визначитися щодо точки $\hat{t}_{i-1}$, розглянемо лінійну регресію з однією точкою перемикання $\hat{t}_{i-1}$, яка задана на інтервалі $[\hat{t}_{i-2}+1,\hat{t}_i]$:

$$y_t=\mathbf{x}'_t\boldsymbol{\alpha}_{i-1}^0+\varepsilon_t,\quad t=\hat{t}_{i-2}+1,\ldots,\hat{t}_{i-1};\; y_t=\mathbf{x}'_t\boldsymbol{\alpha}_i^0+\varepsilon_t,\; t=\hat{t}_{i-1}+1,\ldots,\hat{t}_i,$$

де $\boldsymbol{\alpha}_{i-1}^0$, $\boldsymbol{\alpha}_i^0$ — n-вимірні параметри регресії.

Необхідно перевірити гіпотезу про те, що $\hat{t}_{i-1}$ — помилкова точка перемикання; це еквівалентно перевірці нульової гіпотези H_0: $\boldsymbol{\alpha}_{i-1}^0=\boldsymbol{\alpha}_i^0$.

Перевірка полягає у порівнянні сум квадратів залишків з урахуванням чи з неурахуванням цього обмеження.

Сума квадратів залишків без обмеження $\boldsymbol{\alpha}_{i-1}^0=\boldsymbol{\alpha}_i^0$:

$$s(\hat{t}_{i-2}+1,\hat{t}_i,\hat{\boldsymbol{\alpha}})=\sum_{t=\hat{t}_{i-2}+1}^{\hat{t}_{i-1}}(y_t-\mathbf{x}'_t\hat{\boldsymbol{\alpha}}_{i-1})^2+\sum_{t=\hat{t}_{i-1}+1}^{\hat{t}_i}(y_t-\mathbf{x}'_t\hat{\boldsymbol{\alpha}}_i)^2,\; i=2,3,\ldots,$$

де оцінка методом найменших квадратів $\hat{\boldsymbol{\alpha}}=[\hat{\boldsymbol{\alpha}}'_{i-1}\;\;\hat{\boldsymbol{\alpha}}'_i]=\arg\min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^{2n}} s(\hat{t}_{i-2}+1,\hat{t}_i,\boldsymbol{\alpha})$.

Якщо обмеження $\boldsymbol{\alpha}_{i-1}^0=\boldsymbol{\alpha}_i^0$ виконується, то розглянута регресія перетворюється на звичайну лінійну регресію із сумою квадратів залишків, а саме

$$s(\hat{t}_{i-2}+1,\hat{t}_i,\boldsymbol{\alpha}^-)=\sum_{t=\hat{t}_{i-2}+1}^{\hat{t}_i}(y_t-\mathbf{x}'_t\boldsymbol{\alpha}^-)^2,\; i=2,3,\ldots,$$

де $\boldsymbol{\alpha}^-=\arg\min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n}\in\mathbb{R}^{2n}s(\hat{t}_{i-2}+1,\hat{t}_i,\boldsymbol{\alpha})$.

Тоді з виконанням припущення 1 згідно з теорією лінійних гіпотез у регресійному аналізі матимемо, що величина

$$F(\hat{t}_{i-2}+1,\hat{t}_i)=\frac{s^-(\hat{t}_{i-2}+1,\hat{t}_i,\boldsymbol{\alpha}^-)-s(\hat{t}_{i-2}+1,\hat{t}_i,\hat{\boldsymbol{\alpha}})}{s(\hat{t}_{i-2}+1,\hat{t}_i,\boldsymbol{\alpha})}\,\frac{\hat{t}_i-\hat{t}_{i-2}-2n}{n},\; i=2,3,\ldots$$

розподілена за законом Фішера з числом ступенів свободи $q_1 = n$, $q_2 = \hat{t}_i - \hat{t}_{i-2} - 2n$. Нульова гіпотеза H_0 не відхиляється на $100p\%$-му рівні, якщо $F(\hat{t}_{i-2} + 1, \hat{t}_i) \le F_p(q_1, q_2)$, $i = 0, 1, \ldots$ В іншому випадку приймається альтернативна гіпотеза H_1: $\boldsymbol{\alpha}_{i-1}^0 \ne \boldsymbol{\alpha}_i^0$ (точка перемикання $\hat{t}_{i-1}$ існує).

Зазначимо, що наведений результат отримано за умов спрощення — три точки перемикання є фіксованими величинами. Якщо відстані між ними достатньо великі, що потребує припущення 2, то таке припущення можна вважати правомірним.

3. ПОБУДОВА СПЛАЙНУ

Вважатимемо, що довжина інтервалу спостереження обмежена. Нехай за результатами розрахунків відповідно до розд. 2 знайдено оцінки k точок перемикання на часовому інтервалі довжиною Θ:

$$t_0 = 0,\ \hat{t}_i > 0,\ i = 1, \ldots, k,\ \hat{t}_{k+1} = \Theta. \tag{20}$$

Вважаючи ці величини фіксованими, методом найменших квадратів визначимо оцінки параметрів регресії (1):

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{k+1} \sum_{t=\hat{t}_{i-1}+1}^{\hat{t}_i} (y_t - \mathbf{x}_t' \boldsymbol{\alpha}_i)^2 \to \min, \tag{21}$$

де мінімізація виконується щодо $\boldsymbol{\alpha}_i$, $i = 1, \ldots, k+1$.

У результаті отримаємо лінію регресії

$$\hat{y}_t = \mathbf{x}_t' \hat{\boldsymbol{\alpha}}_i, \tag{22}$$

де $\hat{\boldsymbol{\alpha}}_i$ — розв'язок (21), причому

$$\hat{\boldsymbol{\alpha}}_i = \text{const},\ t \in [\hat{t}_{i-1} + 1, \hat{t}_i],\ i = 1, \ldots, k+1,\ t_0^0 = 0.$$

Лінія регресії (21), (22) у точках перемикання може мати розриви.

Нехай маємо апріорну інформацію про те, що тренд розглядуваного часового ряду y_t, $t = 1, 2, \ldots$, вигляду (1) є неперервним у точках перемикання. Інакше кажучи, тренд є лінійним сплайном. З цієї умови випливають обмеження на параметри регресії:

$$\mathbf{x}_{\hat{t}_i}' (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i+1}) = 0,\ i = 1, \ldots, k. \tag{23}$$

Розв'язуючи задачі (21), (23), отримаємо оцінки параметрів сплайну $\boldsymbol{\alpha}_i^-$, $i = 1, \ldots, k+1$, з точками перемикання (20).

Узагальненням (21), (23) є задача оптимізації, в якій шуканими величинами будуть точки перемикання t_i, $i = 1, \ldots, k$, і параметри сплайну $\boldsymbol{\alpha}_i$, $i = 1, \ldots, k+1$:

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{k+1} \sum_{t=\hat{t}_{i-1}+1}^{\hat{t}_i} (y_t - \mathbf{x}_t' \boldsymbol{\alpha}_i)^2 \to \min,\ \mathbf{x}_{\hat{t}_i}' (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i+1}) = 0,\ i = 1, \ldots, k, \tag{24}$$

де $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1'\ \ \boldsymbol{\alpha}_2'\ \ \ldots\ \ \boldsymbol{\alpha}_{k+1}']'$.

Оскільки шукані змінні у цій задачі є цілочисловими і неперервними величинами, її розв’язок навіть для невеликих значень k становить неабиякі труднощі. Розглянемо наближений розв’язок (24).

Нехай за допомогою індикаторів, визначених у розд. 2, отримано точки перемикань (20). Далі будемо поступово їх уточнювати, з’єднуючи неперервною кусково-лінійною лінією. На кожному кроці уточнюється одна точка перемикання, починаючи з першої точки. Уточнення зводиться до розв'язання задачі оптимізації на i-му кроці, $i=1,\ldots,k$.

На першому кроці розв'язується задача оптимізації

$$\left.\begin{aligned} &S(t_1,\ldots,t_k,\boldsymbol{\alpha})=\sum_{i=1}^{k+1}\sum_{t=t_{i-1}+1}^{t_i}(y_t-\mathbf{x}'_t\boldsymbol{\alpha}_i)^2\to\min,\ \mathbf{x}'_{t_1}(\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2)=0;\\ &3\le t_1\le\hat{t}_2-3;\ t_0=0,\ t_j=\hat{t}_j=\text{const},\ j=2,\ldots,k.\end{aligned}\right\}\qquad(25)$$

Розв'язується ця задача у такий спосіб. Для довільного цілого значення $t\in[3,\hat{t}_2-3]$ знаходиться розв'язок задачі (25): $\boldsymbol{\alpha}(t_i)=\boldsymbol{\alpha}(t_i,\ldots,t_k)$, оскільки $t_j=\text{const}, j=2,\ldots,k$. Цей розв'язок є єдиним, оскільки функція $S(t_1,\boldsymbol{\alpha})=S(t_1,\ldots,t_k,\boldsymbol{\alpha})$ є строго опуклою щодо $\boldsymbol{\alpha}$. Шукане уточнення оцінки першої точки перемикання визначається так:

$$t_1^-=\arg\min_{t_0+3\le t_1\le\hat{t}_2-3}\ \min_{\boldsymbol{\alpha}\in\Omega(t_1)}\{S(t_i,\ldots,t_k,\boldsymbol{\alpha})\,|\,\mathbf{x}'_{t_1}(\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_{i+1})=0;$$

$$t_j=\hat{t}_j=\text{const},\ j=\overline{2,k}\},\qquad(26)$$

де $\Omega(t_1)=\{\boldsymbol{\alpha}:\mathbf{x}'_{t_1}(\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2)=0\}$.

Решту точок перемикання отримують з розв'язання аналогічної задачі:

$$\left.\begin{aligned} &S(t_1,\ldots,t_k,\boldsymbol{\alpha})=\sum_{i=1}^{k+1}\sum_{t=t_{i-1}+1}^{t_i}(y_t-\mathbf{x}'_t\boldsymbol{\alpha}_i)^2\to\min,\ \mathbf{x}'_{t_1}(\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_{i+1})=0, j=1,\ldots,i;\\ &t_j=t_j^-=\text{const}, j=1,\ldots,i-1; t_{i-1}^-+3\le t_1\le\hat{t}_{i+1}-3;\ t_j=\hat{t}_j=\text{const},\ j=i+1,\ldots,k;\\ &\qquad i=2,\ldots,k,\end{aligned}\right\}\ (27)$$

де t_j^- — оцінка t_j^0, $1\le j\le i$, знайдена внаслідок розв'язання задач (25), (27).

Нехай $\boldsymbol{\alpha}(t_1,\ldots,t_k)$ — розв'язок задачі (27) для фіксованих точок перемикання. Розв'язок є єдиним, оскільки функція $S(t_1,\ldots,t_k,\ \boldsymbol{\alpha})$ опукла відносно $\boldsymbol{\alpha}$. Тоді для точок перемикання $t_j=\text{const}$, $t_j\ne t_i$, $j=1,\ldots,k$, визначених у (27), маємо уточнювану i-у точку перемикання:

$$t_1^-=\arg\min_{t_{i-1}^-+3\le t_i\le\hat{t}_{i+1}-3}\ \min_{\boldsymbol{\alpha}\in\Omega(t_1,\ldots,t_i)}\{S(t_1,\ldots,t_k,\boldsymbol{\alpha})\,|\,t_j=t_j^-=c,$$

$$j=\overline{1,\ i-1};\ t_j=\hat{t}_j=\text{const},\ j=\overline{i+1,\ k}\},\ i=2,\ldots,k,\qquad(28)$$

де $\Omega(t_1,\ldots,t_i)=\{\boldsymbol{\alpha}:\mathbf{x}'_{t_1}(\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_{i+1})=0, j=1,\ldots,i\}$.

Визначення точності оцінки параметрів сплайну зводиться до визначення точності оцінки параметрів лінійної регресії з лінійними обмеженнями на них [16—18].

Далі для перевірки гіпотез про параметри сплайну використаємо таке твердження.

Теорема 1. Нехай маємо регресію

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\alpha}^0+\boldsymbol{\varepsilon},\qquad(29)$$

де $\mathbf{Y}\in\mathbb{R}^\Theta$, $\boldsymbol{\alpha}^0\in\mathbb{R}^N$, $\boldsymbol{\varepsilon}\in\mathbb{R}^\Theta$, $\mathbf{X}$ — матриця $\Theta\times N$ рангу N, компоненти $\boldsymbol{\varepsilon}$ — випадкові величини ε_t, $t=1,2,\ldots,\Theta$, що задовольняють припущення 1.

На параметри регресії (29) накладено обмеження

$$\mathbf{G}\boldsymbol{\alpha}^0=\mathbf{O}_m,\qquad(30)$$

де $\mathbf{G}$ — матриця $m\times n$ повного рангу, $m<N$, $\mathbf{O}_m\in\mathbb{R}^m$ — нульовий вектор.

Параметр $\boldsymbol{\alpha}^0$ на $100p$%-му рівні задовольнятиме лінійна гіпотеза

$$H_0 : \mathbf{P}\boldsymbol{\alpha}^0 = \mathbf{p},\ \mathbf{P}(\nu \times N),\ \mathbf{p} \in \mathbb{R}^\nu , \tag{31}$$

де $\mathbf{P}$ — матриця повного рангу $\nu\,(m+\nu < N)$, якщо

$$F = \frac{||\mathbf{Y}-\mathbf{X}\boldsymbol{\alpha}^{=}||^2 - ||\mathbf{Y}-\mathbf{X}\boldsymbol{\alpha}^{-}||^2}{||\mathbf{Y}-\mathbf{X}\boldsymbol{\alpha}^{-}||^2}\,\frac{\Theta - N + m}{\nu} \le F_p(q_1, q_2), \tag{32}$$

де $F_p(q_1, q_2)$ — верхня $100p$%-ва точка розподілу Фішера, $q_1 = \nu$, $q_2 = \Theta - N + m$.

В (32) оцінки $\boldsymbol{\alpha}^0$ знаходимо методом найменших квадратів:

$$\boldsymbol{\alpha}^{=} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \{\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} | \mathbf{G}\boldsymbol{\alpha} = \mathbf{O}_m,\ \mathbf{P}\boldsymbol{\alpha}^0 = \mathbf{p}\},$$

$$\boldsymbol{\alpha}^{-} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \{\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} | \mathbf{G}\boldsymbol{\alpha} = \mathbf{O}_m\}.$$

Доведення. Не втрачаючи загальності, оскільки $\mathbf{G}$ і $\mathbf{P}$ — матриці повного рангу, одержуємо

$$\mathbf{G}\boldsymbol{\alpha}^0 = \mathbf{G}_1\boldsymbol{\alpha}_1^0 + \mathbf{G}_2\boldsymbol{\alpha}_2^0 = \mathbf{O}_m,\ \mathbf{P}\boldsymbol{\alpha}^0 = \mathbf{P}_1\boldsymbol{\alpha}_1^0 + \mathbf{P}_2\boldsymbol{\alpha}_2^0 = \mathbf{p}, \tag{33}$$

де $\mathbf{G}_1$ и $\mathbf{P}_1$ — невироджені квадратні матриці порядку m, а $\mathbf{G}_2$ і $\mathbf{P}_2$ — матриці розміру $m \times (n-m)$ та $\nu \times (n-m)$ відповідно, $\boldsymbol{\alpha}_1^0 \in \mathbb{R}^m$, $\boldsymbol{\alpha}_2^0 \in \mathbb{R}^{N-m}$, $\boldsymbol{\alpha}^0 = [(\boldsymbol{\alpha}_1^0)' \ \ (\boldsymbol{\alpha}_2^0)']'$.

Крім того, з (29) маємо

$$\mathbf{Y} - \mathbf{X}\boldsymbol{\alpha}^0 + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\alpha}_1^0 + \mathbf{X}_2\boldsymbol{\alpha}_2^0 + \boldsymbol{\varepsilon}, \tag{34}$$

де $\mathbf{X}_1$ — матриця розміру $\Theta \times m$, $\mathbf{X}_2$ — матриця розміру $\Theta \times (n-m)$.

З першого рівняння (33) отримуємо

$$\boldsymbol{\alpha}_1^0 = -\mathbf{G}_1^{-1}\mathbf{G}_2\boldsymbol{\alpha}_2^0. \tag{35}$$

Підставивши (35) в (34) і в друге рівняння (33), отримаємо рівняння регресії щодо $\boldsymbol{\alpha}_2^0$ та обмеження на цей параметр:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha}_2^0 + \boldsymbol{\varepsilon},\ \mathbf{U}\boldsymbol{\alpha}_2^0 = \mathbf{p}, \tag{36}$$

де $\mathbf{Z} = \mathbf{X}_2 - \mathbf{X}_1\mathbf{G}_1^{-1}\mathbf{G}_2$, $\mathbf{U} = \mathbf{P}_2 - \mathbf{P}_1\mathbf{G}_1^{-1}\mathbf{G}_2$.

Оскільки стовпці матриці $\mathbf{X}$ лінійно незалежні, то $\mathbf{X}_2$, а отже і $\mathbf{Z}$, мають ранг $N-m$. Ранг матриці $\mathbf{U}$ дорівнює ν.

Вирази (36) представляють регресію з параметром $\boldsymbol{\alpha}_2^0$, відносно якого є лінійна гіпотеза — друга рівність з (36). Якщо припущення 1 має місце, то ця гіпотеза не відхиляється на $100p$%-му рівні за умови

$$F = \frac{||\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha}_2^{=}||^2 - ||\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha}_2^{-}||^2}{||\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha}_2^{-}||^2}\,\frac{\Theta - N + m}{\nu} \le F_p(q_1, q_2), \tag{37}$$

де оцінки параметра регресії (36) мають вигляд

$$\boldsymbol{\alpha}_2^{=} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N-m}} \{\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha}_2 | \mathbf{U}\boldsymbol{\alpha}_2 = \mathbf{p}\},\ \boldsymbol{\alpha}_2^{-} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{N-m}} \{\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha}_2\};$$

$$q_1 = \nu,\ q_2 = \Theta - N + m.$$

Наведені оцінки існують, оскільки матриці $\mathbf{Z}$ і $\mathbf{U}$ мають повний ранг. Замінюючи в (35) параметри їхніми оцінками, отримаємо

$$\boldsymbol{\alpha}_1^{=} = -\mathbf{G}_1^{-1}\mathbf{G}_2\,\boldsymbol{\alpha}_2^{=},\ \boldsymbol{\alpha}_1^{-} = -\mathbf{G}_1^{-1}\mathbf{G}_2\,\boldsymbol{\alpha}_2^{-}, \tag{38}$$

при цьому $\boldsymbol{\alpha}^{=} = [(\boldsymbol{\alpha}_1^{=})' \quad (\boldsymbol{\alpha}_2^{=})']'$, $\boldsymbol{\alpha}^{-} = [(\boldsymbol{\alpha}_1^{-})' \quad (\boldsymbol{\alpha}_2^{-})']'$.

Підставивши в (37) вирази для $\mathbf{Z}$ і $\mathbf{U}$ з пояснень до формули (36) з урахуванням рівностей (38), отримаємо (32). ■

Наслідок 1. Нехай деякий сплайн Ω з k точками перемикання згладжує дані $\mathbf{Y}$ на інтервалі часу I довжиною Θ вектором $\mathbf{X}\boldsymbol{\alpha}^0$ за умов (30). На інтервалі I задано сплайн $\overline{\Omega}$ з $\bar{k}$ точками перемикання, при цьому $\bar{k} < k$. Параметри Ω оцінюють методом найменших квадратів з урахуванням обмежень $\mathbf{G}\boldsymbol{\alpha} = \mathbf{O}_m$, $\mathbf{P}\boldsymbol{\alpha} = \mathbf{p}$. Прийняття гіпотези (31) означає можливість заміни Ω сплайном $\overline{\Omega}$ з простішою структурою та меншою кількістю параметрів. При цьому точність отриманих оцінок залишається високою. Якщо гіпотеза (31) відхиляється, дані згладжуються сплайном Ω.

Гіпотеза (31) приймається, якщо виконується нерівність (32), де $N = 2(k+1)$, $k \ge 2$.

ВИСНОВКИ

Розглянуто задачу визначення кусково-лінійного тренда, де ключовим питанням є знаходження точок перемикання. На відміну від традиційного підходу, що полягає у використанні схем математичного програмування, в яких фігурують цілочисельні змінні, тут наведено інший підхід. Він базується на інтелектуальному аналізі, який дає змогу застосовувати статистичні методи і отримувати розв'язки з високою точністю.

Важливою властивістю такого підходу є відсутність вимоги апріорного знання щодо кількості точок перемикання. Підхід дає змогу оцінювати точки перемикання в реальному часі з надходженням нових даних. З невеликим запізненням, що пов'язане з кількістю оцінюваних параметрів тренда, визначаються точки перемикання. У разі побудови сплайну або корегування оцінених точок перемикання таке запізнення збільшується до відстані між сусідніми точками перемикання.

Інша важлива особливість представленого підходу — можливість побудови як тренда з розривами, так і у вигляді лінійного сплайну.

СПИСОК ЛІТЕРАТУРИ

1. Bai J. Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 1997. P. 551–563.
2. Bai J., Perron P. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*. 2003. Vol. 18. P. 1–22.
3. Casini A., Perron P. Continuous record Laplace-based inference about the break date in structural change models. *Journal of Econometrics*. May 2020. P. 37–53.
4. Perron P., Zorta E. Estimation and inference of linear trend slope ratios with an application to global temperature. *Journal of Time Series Analysis*. 2017. Vol. 38, N 5. P. 630–667.
5. Korkhin A.S. Constructing a switching regression with unknown switching points. *Cybernetics and Systems Analysis*. 2018. Vol. 54, N 3. P. 443–455.
6. Knopov P.S., Korkhin A.S. Continuous-time switching regression method with unknown switching points. *Cybernetics and Systems Analysis*. 2020. Vol. 56, N 1. P. 68–80.

7. Korkhin A.S. An approximate method of constructing a switching regression with unknown switch points. *Cybernetics and Systems Analysis*. 2020. Vol. 56, N 3. 426-438.
8. Knopov P.S., Korkhin A.S. Statistical analysis of the dynamics of coronavirus cases using stepwise switching regression. *Cybernetics and Systems Analysi*s. 2020. Vol. 56, N 6. P. 943–952.
9. Розин Б.Б., Котюков В.И., Ягольницер М.А. Экономико-статистические модели с переменной структурой. Новосибирск: Наука, 1984. 242 с.
10. Дрейпер Н., Смит Г. Прикладной регрессионный анализ, 3-е изд. Москва: Издательский дом «Вильямс», 2016. 912 с.
11. Себер Дж. Линейный регрессионный анализ. Москва: Мир, 1980. 456 с.
12. Корхин А., Пржебицин З. Основы теории вероятностей и математической статистики (для экономистов). Днепр: Лира, 2022. 540 с.
13. Chow G.C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28, 1960. P. 591–605.
14. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. Москва: Наука, 1977. 224 с.
15. Knopov P.S., Korkhin A.S, Vovk L.B. On minimum length confidence intervals. Modern optimization methods for decision making under risk and uncertainty. *CRC Press*, 2023. P. 87–101.
16. Корхин А.С., Минакова Е.П. Компьютерная статистика. Ч. 2. Днепропетровск: Национальный горный университет, 2009. 239 с.
17. Korkhin A.S. Parameter estimation accuracy for nonlinear regression with nonlinear constraints. *Cybernetics and Systems Analysis*. 1998. Vol. 34, N 6. P. 663–672.
18. Korkhin A.S. Solution of problems of the nonlinear least-squares method with nonlinear constraints based on the linearization method. *Journal of Automation and Information Sciences*. 1999. Vol. 31, N 6. P. 110–120.

P.S. Knopov, A.S. Korkhin

**DETERMINING A PIECEWISE LINEAR TREND OF A NONSTATIONARY TIME SERIES BASED ON INTELLIGENT DATA ANALYSIS.
I. DESCRIPTION AND JUSTIFICATION OF THE METHOD**

Abstract. The problem of identifying the trend of a non-stationary time series is often encountered in various applications. In the article, this trend is proposed to be represented as a linear regression with unknown switching points. Typically, such a regression is built using mathematical programming methods. Moreover, the desired variables are mixed variables, which significantly complicates the problem's solution. The article proposes a different approach based on data mining using statistical criteria. The algorithms described in the article are used to solve a number of problems, including one practical problem. The calculations showed satisfactory accuracy..

Keywords: linear regression, algorithm, time series, trend, methods, mathematical programming.

Надійшла до редакції 11.08.2023