Журнал обчислювальної та прикладної математики **2012, №4(110)** Journal of Computational & Applied Mathematics

УДК 519.71

АЛГОРИТМ ЗНАХОДЖЕННЯ ДОВІРЧИХ МЕЖ ДЛЯ ТОЧКИ ПЕРЕХОДУ МОДЕЛІ НЕЛІНІЙНОЇ СПЛАЙНОВОЇ РЕГРЕСІЇ

М. Ю. Савкіна

Резюме. В роботі побудовано алгоритм знаходження довірчих меж з заданим рівнем значущості для точки переходу в моделі нелінійної сплайнової регресії, де функція регресії — лінійний сплайн на відрізку $[0,1]$ з одним вузлом. Розташування вузла невідоме та підлягає оцінюванню.

Розглянемо модель регресії

$$y_i = f(\tau_i) + \xi_i, \qquad i = 0, 1, ..., n, \tag{1}$$

де $\xi_0, ..., \xi_n$ — незалежні у сукупності нормально розподілені випадкові величини з $E\xi_i = 0$ та $D\xi_i = \sigma^2$, а функція регресії має вигляд

$$f(\tau) = \begin{cases} a\tau + b, \text{ якщо } 0 \le \tau \le \tau^*, \\ c\tau + d, \text{ якщо } \tau^* \le \tau \le 1, \\ a\tau^* + b = c\tau^* + d. \end{cases} \tag{2}$$

Якщо точка з'єднання τ^* відома, модель (1), (2) є лінійною по параметрах a, b, c, d, які підлягають оцінюванню. Якщо точка з'єднання τ^* невідома, модель стає нелінійною по параметрах, а τ^* перетворюється на невідомий параметр моделі, який також треба оцінювати.

Зауважимо, що невідомі параметри моделі (1), (2) будуть залежними, але функцію $f(\tau)$ можна подати у вигляді

$$f(\tau) = (a\tau + b) + c_1(\tau - \tau^*)_+, \tag{2'}$$

де $c_1 = c - a$, а $(\tau - \tau^*)_+$ — зрізана степенева функція [1]. В цих позначеннях a, b, c_1, τ^* — параметри, які підлягають оцінюванню. Вони будуть незалежними.

Найпоширенішим методом оцінювання параметрів в регресійному аналізі (як лінійному, так і нелінійному) є метод найменших квадратів (МНК), який полягає в мінімізації суми

$$\sum_{i=0}^{n} (y_i - (a\tau_i + b) - c_1(\tau_i - \tau^*)_+)^2$$

відносно вектора невідомих параметрів. Оцінювання параметрів нелінійної моделі в загальному випадку є дуже складною задачею, але для моделі (1), (2′) запропоновано алгоритм [2], завдяки якому за скінченну кількість кроків знаходиться оцінка МНК параметрів a, b, c_1, τ^*. Він полягає в такому.

Розіб'ємо множину $\{\tau_0, ..., \tau_n\}$ на дві підмножини $\{\tau_0, ..., \tau_k\}$ та $\{\tau_{k+1}, ..., \tau_n\}$, $k = 1, ..., n-2$. По точках $\{(\tau_0, y_0), ..., (\tau_k, y_k)\}$ за методом найменших квадратів побудуємо пряму $y = a_k\tau + b_k$. Позначимо

$$\rho_1(k) = \sum_{i=0}^{k}(y_i - a_k\tau_i - b_k)^2,$$

а по точках $\{(\tau_{k+1}, y_{k+1}), ..., (\tau_n, y_n)\}$ за методом найменших квадратів побудуємо пряму $y = c_k\tau + d_k$. Позначимо

$$\rho_2(k) = \sum_{i=k+1}^{n}(y_i - c_k\tau_i - d_k)^2.$$

Далі, точку перетину прямих $y = a_k\tau + b_k$ та $y = c_k\tau + d_k$ позначимо τ_k^*. Якщо $\tau_k^* \in [\tau_k, \tau_{k+1}]$, то покладемо $T(k) = \rho_1(k) + \rho_2(k)$, інакше $T(k) = \infty$.

Таким чином, покладаючи $k = 1, 2, ..., n-2$, отримаємо послідовність $T(1), T(2), ..., T(n-2)$, кожному $T(k)$ відповідає τ_k^*.

Тепер вважаємо, що $\tau^* = \tau_k$, знайдемо $\hat{a}_k, \hat{b}_k, \hat{c}_{1,k}$, які мінімізують

$$S(k) = \sum_{i=0}^{n}(y_i - \hat{a}_k\tau_i - \hat{b}_k - \hat{c}_{1,k}(\tau_i - \tau_k)_+)^2.$$

Покладаючи $k = 1, 2, ..., n-1$, маємо ще одну послідовність $S(1), S(2), ..., S(n-1)$, кожному $S(k)$ відповідає τ_k.

Позначимо

$$R^* = \min\{T(1), ..., T(n-2), S(1), ..., S(n-1)\},$$

якщо $R^* = T(l)$, то $\hat{\tau}^* = \tau_l^*$, якщо $R^* = S(l)$, то $\hat{\tau}^* = \tau_l$.

В роботі [2] доведено, що таким чином отримана оцінка $\hat{\tau}^*$ параметра τ^* є оцінкою МНК для нелінійної моделі (1), (2').

Розглянемо тепер задачу перевірки простої статистичної гіпотези

$$H: \ \tau^* = \bar{\tau},$$

де $\bar{\tau}$ — фіксована точка з проміжку $[0, 1]$. Критерій відношення правдоподібності перевірки гіпотези H приводить до множини прийняття гіпотези [3]

$$E_H = \{y \in R^{n+1} : R(\bar{\tau}) - R^* \le \varphi R^*\},$$

де

$$R(\bar{\tau}) = \sum_{i=0}^{n}(y_i - \bar{a}\tau_i - \bar{b} - \bar{c}_1(\tau_i - \bar{\tau})_+)^2.$$

Значення φ для нелінійної регресії можна вибирати різними методами. Один з них збігається з методом вибору в лінійній регресії. Припустимо, модель лінійній регресії має m невідомих параметрів, а гіпотеза H_0 складається з p лінійних рівнянь. Задамося рівнем значущості α та знайдемо для нього відповідне значення $F_\alpha(p, n+1-m)$ таким чином. Позначимо $f(t, p, n+1-m)$ — щільність розподілу Фішера з p та $n+1-m$ ступенями

свободи та знайдемо таке $F_\alpha = F_\alpha(p, n+1-m)$, що $\int_{F_\alpha}^{\infty} f(t, p, n+1-m)dt = \alpha$. Значення F_α знаходять з таблиць. Покладемо

$$\varphi = \frac{p}{n+1-m} F_\alpha(p, n+1-m). \tag{3}$$

Побудуємо тепер довірчу множину D із заданим рівнем значущості α для τ^*, тобто таку множину, яка покриває справжнє значення параметра τ^* з ймовірністю, не меншою $1-\alpha$. Зауважимо, що задача довірчого оцінювання пов'язана з задачею перевірки статистичних гіпотез. Згідно з [3]

$$D = \{\tau \in [0,1] : R(\tau) - R^* \le \varphi R^*\},$$

де

$$R(\tau) = \sum_{i=0}^{n} (y_i - a\tau_i - b - c_1(\tau_i - \tau)_+)^2.$$

Значення φ вибираємо за формулою (3), в нашому випадку $m = 4$ (модель має 4 невідомі параметри) та $p = 1$ (гіпотеза H складається з одного рівняння), тобто

$$\varphi = \frac{1}{n-3} F_\alpha(1, n-3).$$

Відмітимо, що $R(\tau_k) = S(k),\ k = 1, 2, ..., n-1$.

Довірча множина D може складатися з одного інтервалу $(\tau^{(1)}, \tau^{(2)})$, або з їх сукупності

$$\{\cup_{i=1}^{j} (\tau^{(2i-1)}, \tau^{(2i)}), j \ge 2\}.$$

Алгоритм побудови довірчої множини D.

Побудуємо спочатку інтервал $(\tau^{(1)}, \tau^{(2)})$, якому буде належати точка $\hat{\tau}^*$.

Нехай $\hat{\tau}^* \in [\tau_l, \tau_{l+1}]$. Якщо $R(\tau_s) - R^* \le \varphi R^*$ для всіх $s = 1, ..., l$, покладемо $\tau^{(1)} = \tau_0$, інакше існує точка $\tau_s,\ s = 1, ..., l$, в якій $R(\tau_s) - R^* > \varphi R^*$.

Нехай $\tau_{s'}$ — найближча до τ_l точка, в якій $R(\tau_{s'}) - R^* > \varphi R^*$, або $\tau_{s'} = \tau_l$, якщо $R(\tau_l) - R^* > \varphi R^*$. Якщо $\tau^*_{s'} \notin [\tau_{s'}, \tau_{s'+1}]$, то функція $R(\tau)$ спадає на проміжку $[\tau_{s'}, \tau_{s'+1}]$, якщо $\tau^*_{s'} \in [\tau_{s'}, \tau_{s'+1}]$, то функція $R(\tau)$ спадає на проміжку $[\tau_{s'}, \tau^*_{s'}]$ [2].

Покладемо $\tau^{(1)} = \tau'$, де τ' — точка, в якій $R(\tau') - R^* = \varphi R^*$, $\tau' \in [\tau_{s'}, \tau_{s'+1}]$, якщо $\tau^*_{s'} \notin [\tau_{s'}, \tau_{s'+1}]$, та $\tau' \in [\tau_{s'}, \tau^*_{s'}]$, якщо $\tau^*_{s'} \in [\tau_{s'}, \tau_{s'+1}]$.

Далі, якщо $R(\tau_q) - R^* \le \varphi R^*$ для всіх $q = l+1, ..., n-1$, покладемо $\tau^{(2)} = \tau_n$, інакше існує точка $\tau_q,\ q = l+1, ..., n$, в якій $R(\tau_q) - R^* > \varphi R^*$.

Нехай $\tau_{q'}$ — найближча до τ_{l+1} точка, в якій $R(\tau_{q'}) - R^* > \varphi R^*$, або $\tau_{q'} = \tau_{l+1}$, якщо $R(\tau_{l+1}) - R^* > \varphi R^*$.

Якщо $\tau^*_{q'} \notin [\tau_{q'}, \tau_{q'+1}]$, то функція $R(\tau)$ зростає на проміжку $[\tau_{q'}, \tau_{q'+1}]$, якщо $\tau^*_{s'} \in [\tau_{q'}, \tau_{q'+1}]$, то функція $R(\tau)$ зростає на проміжку $[\tau^*_{q'}, \tau_{q'+1}]$ [2].

Покладемо $\tau^{(2)} = \tau''$, де τ'' — точка, в якій $R(\tau'') - R^* = \varphi R^*$, $\tau'' \in [\tau_{q'}, \tau_{q'+1}]$, якщо $\tau^*_{q'} \notin [\tau_{q'}, \tau_{q'+1}]$ та $\tau'' \in [\tau^*_{q'}, \tau_{q'+1}]$, якщо $\tau^*_{q'} \in [\tau_{q'}, \tau_{q'+1}]$.

Далі, якщо в множині

$$\{T(1), ..., T(n-2), S(1), ..., S(n-1)\} \backslash \{T(l), S(l), S(l+1)\}$$

є такі $T(k)$ (або $S(k)$), що $T(k) - R^* < \varphi R^*$ (або $S(k) - R^* < \varphi R^*$), то навколо кожної точки τ_k^* (або τ_k) будуємо інтервали так само, як навколо точки $\hat{\tau}^*$ побудували інтервал $(\tau^{(1)}, \tau^{(2)})$. У результаті отримаємо множину D.

Приклад. Нехай

	i=0	i=1	i=2	i=3	i=4	i=5
τ_i	0	0.2	0.4	0.6	0.8	1
y_i	0.11	0.19	0.42	0.63	0.83	1.04

Знайдемо за наведеним в роботі [2] алгоритмом оцінку МНК точки переходу τ^*. Маємо

$$T(1) = \infty,\ T(2) = 0.00375,\ T(3) = 0.00942,$$
$$S(1) = 0.00027,\ S(2) = 0.00402,\ S(3) = 0.0057,\ S(4) = 0.00648;$$
$$R^* = S(1),\ \hat{\tau}^* = \tau_1 = 0.2.$$

Знайдемо тепер інтервал $(\tau^{(1)}, \tau^{(2)})$ із заданим рівнем значущості $\alpha = 0.05$ для $\hat{\tau}^*$. Маємо

$$n = 5;\ F_{0.05}(1, 2) = 18.5128;\ \varphi R^* = 0.0025;$$
$$R(\tau_k) - R^* \le \varphi R^*,\ k = 2, 3, 4;\ R(\tau_k^*) - R^* \le \varphi R^*,\ k = 1, 2, 3.$$

Таким чином, $\tau^{(1)} = 0,\ \tau^{(2)} = 0.315$, бо $R(0.315) - R^* = \varphi R^*$.

Висновки

В роботі побудовано алгоритм знаходження довірчих меж із заданим рівнем значущості для точки переходу в моделі нелінійної сплайнової регресії та наведено приклад застосування цього алгоритму.

Література

1. Завьялов Ю. С. Методы сплайн-функций. / Ю. С. Завьялов, Б. И. Квасов, В. Л. Мирошниченко. — Москва: Наука, 1980. — 352 с.
2. Hudson Derek J. Fitting Segmented Curves Whose Join Points Have to Be Estimated. / Derek J. Hudson. // JASA. — 1966. — V. 61. — № 316. — P. 1097–1129.
3. Демиденко Е. З. Линейная и нелинейная регрессии. / Е. З. Демиденко. — Москва: Финансы и статистика, 1981. — 304 с.

Інститут математики НАН України, вул. Терещенківська, 3, Київ, 01601, Україна.

Надійшла 15.01.11