УДК 517.5

М. Ю. Савкіна, канд. фіз.-мат. наук, с. н. с.

Інститут математики НАН України, м. Київ

АЛГОРИТМ ПЕРЕВІРКИ НА КОРЕКТНІСТЬ МОДЕЛІ СПЛАЙНОВОЇ РЕГРЕСІЇ

Побудовано алгоритм перевірки на коректність моделі двофазної лінійної регресії з невідомою точкою перемикання у випадку, коли треба зробити вибір між такою моделлю та лінійною. Алгоритм заснований на загальних принципах перевірки статистичних гіпотез у регресійному аналізі.

Ключові слова: *метод найменших квадратів, регресійна модель, точка перемикання.*

Вступ. Класичний регресійний аналіз засновано на тому, що вигляд моделі регресії відомий з точністю до параметрів, тобто, набір незалежних змінних (факторів) задано однозначно, всі істотні змінні присутні та ніяких альтернативних способів вибору факторів немає. Насправді вибір регресорів, тісно пов'язаний з вибором моделі об'єкта, — одна з найскладніших проблем. В середині минулого сторіччя поява ЕОМ значно спростила цю проблему. «...Поступово з'ясувалося, що ЕОМ допускає відмову від жорсткої моделі дослідження та підбір під час обробки даних деякої «найкращої» моделі...» [1]. У даний час розроблено багато статистичних методів відбору змінних, такі як метод всіх можливих регресій, метод виключення, кроковий регресійний метод, ступінчастий регресійний аналіз, гребенева регресія тощо. В одній і тій ж задачі їх застосування не завжди веде до отримання тієї ж самої моделі, хоча в багатьох випадках виходить однаковий результат. Кожен метод має свої недоліки, свої переваги. Жоден метод не буде добре працювати в усіх випадках. У деяких з цих методів застосовується критерій Фішера для перевірки гіпотези рівності нулю невідомого параметра регресії, і на його підставі приймається рішення про видалення відповідного фактора з регресії.

Розглянемо модель регресії

$$y_i = at_i + b + c_1\left(t_i - t^*\right)_+ + \varepsilon_i,\ i = 0,1,...,n, \tag{1}$$

де $\varepsilon_0, \varepsilon_1, ..., \varepsilon_n$ — незалежні у сукупності нормально розподілені випадкові величини з $E\varepsilon_i = 0$ та $D\varepsilon_i = \sigma^2$, а $\left(t_i - t^*\right)_+$ — зрізана степенева функція [2]. Згідно з [3] точка t^* називається точкою перемикання моделі. Якщо вона відома, модель (1) є лінійною по параметрах a, b, c_1, які підлягають оцінюванню. Якщо t^* невідома, модель стає нелінійною по параметрах, а t^* перетворюється на невідомий параметр моделі, який також треба оцінювати.

Далі висуваємо гіпотезу

$$H : c_1 = 0.$$

Якщо вона підтвердиться з великою ймовірністю, фактор $\left(t_i - t^*\right)_+$ видаляємо з регресії, тобто модель (1) перетвориться на таку модель

$$y_i = at_i + b + \varepsilon_i,\ i = 0,1,...,n, \qquad (2)$$

Позначимо $\vec{y} = \left(y_0, y_1, ..., y_n\right)$. Зазвичай перевірка статистичної гіпотези здійснюється за допомогою критерія відношення правдоподібності, який приводить до множини прийняття гіпотези H [4]

$$E_H = \left\{ \vec{y} \in R^{n+1} : \frac{S_2^2 - S_4^2}{S_4^2} < \varphi \right\},$$

де S_2^2 та S_4^2 — залишкові суми квадратів моделі (2) та нелінійної моделі (1) відповідно. Значення φ для нелінійної регресії можна вибирати різними методами, один з них дає

$$\varphi = \frac{1}{n-2} F_\alpha\left(1, n-3\right),$$

де α — рівень значущості, $F_\alpha = F_\alpha\left(1, n-3\right)$ — значення, при якому $\int\limits_{F_\alpha}^{\infty} f\left(t,1,n-3\right)dt = \alpha$, $f\left(t,1,n-3\right)$ — щільність розподілу Фішера з 1 та $n-3$ ступенями свободи. Значення F_α знаходять з таблиць.

У роботі [5] у випадку, коли $t_i = \dfrac{i}{n}$, $i = 0,1,...,n,$ побудовано алгоритм, завдяки якому можна відхиляти гіпотезу H не знаходячи S_4^2.

Позначимо $z_k, k = 1,2,...,n-1,$ — останній діагональний елемент матриці $\left(X_k' X_k\right)^{-1}$, де

$$X_k' = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \frac{1}{n} & \dots & \frac{k}{n} & \frac{k+1}{n} & \dots & 1 \\ 0 & 0 & \dots & \frac{1}{n} & \frac{2}{n} & \dots & \frac{n-k+1}{n} \end{pmatrix}.$$

Твердження. Має місце рівність

$$S_2^2 - S_{3,k}^2 = \frac{\left(\widehat{c}_1^{(k)}\right)^2}{z_k}, \tag{3}$$

де $S_{3,k}^2$ та $\widehat{c}_1^{(k)}$ — залишкова сума квадратів та оцінка МНК параметра c_1 лінійної регресійної моделі (1), коли $t^* = t_k$.

Доведення випливає з загального результату [4].

У випадку, коли $t_i = \frac{i}{n}$, $i = 0,1,...,n$, маємо

$$\widehat{c}_1^{(k)} = \frac{6n}{(2k+1)n - 2\left(k^2-1\right)}\left(\frac{1}{k(k+1)}\sum_{i=0}^{k} y_i\left[kn - i(2k+n+2)\right] + \right.$$
$$\left. + \frac{1}{(n-k)(n-k+1)}\sum_{i=k+1}^{n} y_i\left[(k-2n-2)n - i(2k-3n-2)\right]\right). \tag{4}$$

$$z_k = \frac{6n^3(n+1)(n+2)}{k(k+1)(n-k)(n-k+1)\left((2k+1)n - 2\left(k^2-1\right)\right)}. \tag{5}$$

Далі, у випадку $t_i = \frac{i}{n}$, $i = 0,1,...,n$, оцінку МНК $\widehat{a}, \widehat{b}$ параметрів a, b регресійної моделі (1) можна знайти за формулами [5]

$$\widehat{a} = \frac{12n}{(n+1)(n+2)}\sum_{i=0}^{n} y_i\left(\frac{i}{n} - \frac{1}{2}\right),\ \widehat{b} = \overline{y} - \frac{1}{2}\widehat{a},\ \overline{y} = \frac{1}{n+1}\sum_{i=0}^{n} y_i, \tag{6}$$

а залишкову суму квадратів —

$$S_2^2 = \sum_{i=0}^{n}\left(y_i - \overline{y}\right)^2 - \widehat{a}^2\sum_{i=0}^{n}\left(\frac{i}{n} - \frac{1}{2}\right)^2. \tag{7}$$

Розглянемо застосування цього алгоритму на прикладах.

Приклад 1. Чи можна відхилити гіпотезу H з рівнем значущості $\alpha = 0.05$ для нелінійної регресійної моделі (1), якщо $\vec{y} = (0, 0.2, 0.25, 0.3, 0.35, 0.55, 0.7, 0.95, 1.1, 1.35, 1.5)$?

В даному випадку $n = 10$, $\varphi = \frac{1}{8}\cdot 5.59 = 0.7375.$

1. Знаходимо $\widehat{a}, \widehat{b}, S_2^2$ за формулами (6), (7):

$$\widehat{a}4 = 1.482,\ \widehat{b} = -0.082,\ S_2^2 = 0.099.$$

2. Знаходимо $M_1 = \max\left\{ \mid y_0 - \widehat{b} \mid , \mid y_{10} - \widehat{a} - \widehat{b} \mid \right\} : M_1 = 0.1$;

   Оскільки $\dfrac{M_1^2}{S_2^2 - M_1^2} < \varphi,$ переходимо до.

3. Знаходимо

   $M_2 = \max\left\{ \mid y_i - 0.1\widehat{a}\, i - \widehat{b} \mid , i = 1, 2, ..., n-1, \right\} : M_2 = 0.161,\ k = 4.$

   Знаходимо $\widehat{c}_1^{(4)}, z_4, S_{3,4}^2$ за формулами (4), (5), (3): $S_{3,4}^2 = 0.01352$;

   Оскільки $\dfrac{S_2^2 - S_{3,4}^2}{S_{3,4}^2} > \varphi,$ гіпотезу H відхиляємо.

Приклад 2. Чи можна відхилити гіпотезу H з рівнем значущості $\alpha = 0.05$ для нелінійної регресійної моделі (1), якщо $\vec{y} = (0, 0.2, 0.25, 0.3, 0.35, 0.55, 0.6, 0.75, 0.85, 0.9, 1.05)$?

В даному випадку також $n = 10,\ \varphi = 0.7375.$

1. Знаходимо $\widehat{a}, \widehat{b}, S_2^2$ за формулами (6), (7):

$$\widehat{a} = 1,\ \widehat{b} = 0.027,\ S_2^2 = 0.017.$$

2. Знаходимо $M_1 = \max\left\{ \mid y_0 - \widehat{b} \mid , \mid y_{10} - \widehat{a} - \widehat{b} \mid \right\} : M_1 = 0.027$; $\dfrac{M_1^2}{S_2^2 - M_1^2} < \varphi,$

   переходимо до

3. Знаходимо

   $M_2 = \max\left\{ \mid y_i - 0.1\widehat{a}\, i - \widehat{b} \mid , i = 1, 2, ..., n-1, \right\} : M_2 = 0.077,\ k = 4.$

   Знаходимо $\widehat{c}_1^{(4)}, z_4, S_{3,4}^2$ за формулами (3), (4), (5): $S_{3,4}^2 = 0.01427$;

   Оскільки $\dfrac{S_2^2 - S_{3,4}^2}{S_{3,4}^2} < \varphi,$ переходимо до

4. На цьому кроці для $k = 1, 2, ..., 8$ будуємо пари прямих по точкам $\left\{ (t_i, y_i), i = 0, 1, ..., k \right\}$ та $\left\{ (t_i, y_i), i = k+1, ..., 10 \right\}$ за методом найменших квадратів та знаходимо їх точку перетину; якщо вона не належить проміжку (t_k, t_{k+1}), цю пару прямих відкидаємо. В нашому прикладі жодна з цих пар прямих не має перетину на відповідному проміжку, тому переходимо до

5. Знаходимо $S_{3,1}^2 = 0.01573,\ S_{3,2}^2 = 0.01668,\ S_{3,3}^2 = 0.01519,$

$$S_{3,5}^2 = 0.01565,\ S_{3,6}^2 = 0.01588,\ S_{3,7}^2 = 0.01653,$$

$$S_{3,8}^2 = 0.01668, \quad S_{3,9}^2 = 0.01606.$$

Таким чином, $S_4^2 = S_{3,4}^2 = 0.01427$; гіпотезу H приймаємо.

Висновки. На прикладах можна побачити, що відхилення гіпотези H майже завжди буде відбуватися на кроці 3, тобто знаходити оцінку МНК невідомих параметрів та залишкову суму квадратів нелінійної моделі (1) немає потреби. Для прийняття гіпотези H треба знаходить S_4^2. Бажано довести, що отримана на кроці 3 $S_{3,k}^2$, що відповідає M_2, буде збігатися з S_4^2. Тоді кроків 4, 5 не треба робити ні в якому разі.

Список використаних джерел:

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Финансы и статистика, 1986. 366 с.
2. Завьялов Ю. С., Квасов Б. И., Мирошниченко В.Л. Методы сплайн-функций. М.: Наука, 1980. 352 с.
3. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980. 456 с.
4. Демиденко Е. З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1986. 304 с.
5. Савкіна М. Ю. Алгоритм перевірки на коректність моделі двофазної нелінійної регресії. Вісник Київського університету. 2015. № 3. С. 115–120.

The algorithm of checking for correctness of two-phase regression model with unknown switch point is constructed in the case when it is necessary to do a choice between such model and linear. The algorithm is based on the general principles of statistical hypothesis testing in regression analysis.

Key words: *least square method, regression model, switch point.*

Одержано 24.02.2017

УДК 517.9

Г. В. Сандраков, д-р фіз.-мат. наук, с. н. с.

Київський національний університет імені Тараса Шевченка, м. Київ

ОПТИМІЗАЦІЯ ПАРАМЕТРІВ МАСИВІВ МІКРОГОЛОК

Оптимізація параметрів пружної взаємодії масивів мікроголок з поверхнею розглянута як задача наближення розв'язків проблем мінімізації для інтегральних функціоналів.

Ключові слова: *оптимізація параметрів, масиви мікроголок, проблеми мінімізації, інтегральні функціонали.*

Вступ. Масиви мікроголок для ін'єкцій ліків все частіше використовуються в сучасній медицині при лікуванні різних захворювань. Такі масиви формуються досить великою кількістю мікроголок, за-