Журнал обчислювальної та прикладної математики **2017, №2(125)** Journal of Numerical & Applied Mathematics

УДК 517.5
MSC 62J02

MODIFICATION OF THE ALGORITHM OF CHECKING FOR CORRECTNESS OF TWO-PHASE NONLINEAR REGRESSION MODEL

Marta Savkina
Institute of Mathematics of NASU, Kyiv, Ukraine, E-mail: marta@imath.kiev.ua

МОДИФІКАЦІЯ АЛГОРИТМУ ПЕРЕВІРКИ НА КОРЕКТНІСТЬ МОДЕЛІ ДВОФАЗНОЇ НЕЛІНІЙНОЇ РЕГРЕСІЇ

М. Ю. Савкіна
Інститут математики НАНУ , Київ, Україна, E-mail: marta@imath.kiev.ua

Abstract. At the paper the problem of exclusion of variables in regression model is examined that arises in the selection of model. More precisely the task of selecting regression model is examined when the choice should be done between linear model and two-phase linear regression model with unknown point switch. Two-phase linear regression model is represented by truncated power function, i.e. truncated power function is one of the factors of the model. Unknown parameter that corresponds to this factor is denoted by c_1. The unknown parameters of both models are estimated using the method of least squares.

In the case of the uniform arrangement of observation points and the location of point switch in the observation point theorem 1 is proved, wherein the estimation of parameter c_1 is represented as a known linear combination of the residues of linear model. Also theorem 2 is proved, wherein the residual sum of squares model two-phase linear regression is represented as the difference between the residual sum of squares of linear model and a linear combination of its residues. Thanks to Theorems 1 and 2 the residual sum of squares of two-phase linear regression model and the estimation parameter c_1 in some cases can be found the formula without solving a system of linear algebraic equations.

A modification the algorithm of checking for correctness of two-phase nonlinear regression model is Proposed, using Theorem 2. The algorithm is based on the general principles of statistical hypothesis testing in regression analysis. Testing of hypothesis for equality of the unknown parameter to zero is carried out by using the likelihood ratio criterion, which determines to accept or reject the hypothesis by the ratio of the sum of squared deviations caused this hypothesis to the residual sum of squares of the two-phase regression model with the unknown switch point. Thanks to Theorems 2 the procedure for

acceptance of hypothesis for equality of the parameter c_1 to zero is greatly simplified.

Keywords: least square method, regression model, point switch.

Резюме. Розглядається задача вибору регресійної моделі у випадку, коли вибір треба зробити між лінійною молеллю та моделлю двофазної лінійної регресії з невідомою точкою перемикання. У випадку рівномірного розташування точок спостереження та розташуванні точки перемикання у точці спостереження доведено теорему, у якій залишкову суму квадратів моделі двофазної лінійної регресії представлено як різницю між залишковою сумою квадратів лінійної моделі та лінійною комбінацією її залишків. Запропоновано модифікацію алгоритму перевірки на коректність моделі двофазної лінійної регресії, яка використовує цю теорему.

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

Вступ

Задача вибору моделі є однією з найактуальніших в регресійному аналізі. У даний час побудовано велику кількість критеріїв якості регресійних моделей, одні дозволяють оцінювати якість моделі самої по собі, інші дозволяють порівнювати моделі одну з одною [1]. Якщо одна модель виходить з іншої видаленням певного фактора (або кількох факторів), то при виборі більш відповідної моделі з цих двох ставиться питання про значимість даного фактора, тобто, перевіряється гіпотеза про рівність нулю невідомого параметра, що відповідає цьому фактору. Для цього зазвичай використовують критерій Фішера.

1. Модель двофазної регресії. Задача перевірки моделі на коректність

Розглянемо модель лінійної регресії

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

де $\epsilon_0, ..., \epsilon_n$ — незалежні у сукупності нормально розподілені випадкові величини з $E\epsilon_i = 0$ та $D\epsilon_i = \sigma^2$, а $(t_i - t^*)_+$ — зрізана степенева функція [2]. Згідно з [3] точка t^* називається точкою перемикання моделі. Якщо вона відома, модель (1) є лінійною за параметрами a, b, c_1, які підлягають оцінюванню. Якщо t^* невідома, модель стає нелінійною за параметрами, а t^* перетворюється на невідомий параметр моделі, який також треба оцінювати.

Крім того, висуваємо гіпотезу

$$H: \ c_1 = 0.$$

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

$$y_i = at_i + b + \epsilon_i, \qquad i = 0, 1, ..., n. \tag{2}$$

Позначимо $y = (y_0, y_1, ..., y_n)$. Критерій відношення правдоподібності перевірки гіпотези H приводить до множини прийняття гіпотези[4]

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

де S_2^2 та S_4^2 — залишкові суми квадратів моделей (2) та (1) відповідно.

Значення φ визначається так. Припустимо, що модель лінійної регресії має m невідомих параметрів, а гіпотеза H_0 складається з p лінійних рівнянь. Задамося рівнем значущості α та знайдемо для нього відповідне значення $F_\alpha(p, n+1-m)$ так. 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. Насамкінець, покладаємо

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

У роботі [5] побудовано алгоритм перевірки на коректність моделі (1) з невідомою точкою перемикання у випадку, коли вибір треба зробити між моделями (1) та (2). Завдяки цьому алгоритму гіпотезу H буде відхилено в переважної більшості випадків без знаходження S_4^2, а для того, щоб гіпотезу H прийняти, доведеться знаходити S_4^2.

2. Теорема про зв'язок оцінки $\hat{c}_1$ із залишками лінійної моделі у випадку відомої точки перемикання

Позначимо

$$e_i = y_i - \hat{a}t_i - \hat{b}, \tag{3}$$

де $\hat{a}$ та $\hat{b}$ — оцінки МНК параметрів a та b моделі (2) відповідно.

Теорема 1. *Нехай $\hat{c}_1^{(k)}$ — оцінка МНК параметра c_1 моделі (1) у випадку, коли $t^* = t_k$.*

Тоді $\hat{c}_1^{(k)}$ можна представити у вигляді

$$\hat{c}_1^{(k)} = D_k\big(ke_0 + (k-1)e_1 + ... + 2e_{k-2} + e_{k-1}\big), \tag{4}$$

$$k = 1, 2, ..., \Big[\frac{n-1}{2}\Big];$$

$$\hat{c}_1^{(k)} = D_k\big((n-k)e_n + (n-k-1)e_{n-1} + ... + 2e_{k+2} + e_{k+1}\big), \tag{5}$$

$$k = n-1, n-2, ..., n - \Big[\frac{n-1}{2}\Big],$$

де

$$D_k = \frac{6n^2(n+1)(n+2)}{\big((2k+1)n - 2(k^2-1)\big)k(k+1)(n-k)(n-k+1)}. \tag{6}$$

Якщо n — парне, то

$$\hat{c}_1^{(0.5n)} = \frac{1}{2} D_{0.5n}\Big((0.5n-1)(e_0+e_n)+(0.5n-2)(e_1+e_{n-1})+\ \ldots\ + \\ +(e_{0.5n-1}+e_{0.5n+1})-e_{0.5n}\Big). \tag{7}$$

Доведення. У роботі [5] отримано

$$\hat{a} = \frac{12n}{(n+1)(n+2)} \sum_{i=0}^{n} y_i \Big(\frac{i}{n} - \frac{1}{2}\Big); \tag{8}$$

$$\hat{b} = \frac{1}{n+1} \sum_{i=0}^{n} y_i - \frac{1}{2}\hat{a}; \tag{9}$$

$$\hat{c}_1^{(k)} = \frac{6n}{(2k+1)n+2(k^2-1)} \ \Bigg(\frac{1}{k(k+1)} \sum_{i=0}^{k} y_i\big[kn - i(2k+n+2)\big] + \\ + \frac{1}{(n-k)(n-k+1)} \sum_{i=k+1}^{n} y_i\big[(k-2n-2)n - i(2k-3n-2)\big]\Bigg), \tag{10}$$

$$k = 1, 2, ..., n-1.$$

Підставимо в (4) формули (3) і (9). Отримуємо

$$\hat{c}_1^{(k)} = D_k\Bigg(\sum_{i=0}^{k-1}(k-i)y_i - \frac{k(k+1)}{2}\cdot\frac{1}{n+1}\sum_{i=0}^{n} y_i + \hat{a}\sum_{i=0}^{k-1}\big(\frac{1}{2}-\frac{j}{n}\big)(k-j)\Bigg). \tag{11}$$

Далі, підставимо (8) в (11). Після низки перетворень отримуємо

$$\hat{c}_1^{(k)} = D_k\Bigg(\sum_{i=0}^{k-1}(k-i)y_i - \frac{k(k+1)}{2(n+1)}\sum_{i=0}^{n} y_i + \\ + \frac{2k(k+1)(1.5n-k+1)}{(n+1)(n+2)}\sum_{i=0}^{n} y_i\big(\frac{i}{n}-\frac{1}{2}\big)\Bigg) =$$

$$= D_k\Bigg(\sum_{i=0}^{k-1} y_i\Bigg(\frac{k(n-k)(n-k+1)}{(n+1)(n+2)} - i\frac{(n-k)(n-k+1)(n+2k+2)}{n(n+1)(n+2)}\Bigg) + \\ + \sum_{i=k}^{n} y_i\Bigg(-\frac{k(k+1)(2n-k+2)}{(n+1)(n+2)} + i\frac{k(k+1)(3n-2k+2)}{n(n+1)(n+2)}\Bigg)\Bigg). \tag{12}$$

Насамкінець, підставимо (6) в (12), отримуємо (10) для $k = 1, 2, ..., \left[\frac{n-1}{2}\right]$.

Формули (5) та (7) можна довести аналогічно. Зауважимо, що $D_k = D_{n-k},\ k = 1, 2, ..., \left[\frac{n-1}{2}\right]$.

Теорему 1 доведено. □

Наслідок 1. *Наслідок Мають місце нерівності*

$$D_1 > D_2 > ... > D_{[\frac{n-1}{2}]}. \tag{13}$$

Доведення. Поліноми від k другого та четвертого порядку $(2k+1)n-2(k^2-1)$ та $k(k+1)(n-k)(n-k+1)$ відповідно зростають на інтервалі $\left(0, \frac{n}{2}\right)$ і в точці $k = \frac{n}{2}$ досягають максимального значення, їх добуток — також. □

3. Теорема про зв'язок залишкової суми квадратів двофазної моделі із залишками лінійної моделі у випадку відомої точки перемикання

Теорема 2. *Нехай $T(k)$ — залишкова сума квадратів моделі (4) у випадку, коли $t^* = t_k$. Тоді $T(k)$ можна представити у вигляді*

$$T(k) = S_2^2 - \frac{D_k}{n}\big(ke_0 + (k-1)e_1 + ... + 2e_{k-2} + e_{k-1}\big)^2, \tag{14}$$

$$k = 1, 2, ..., \Big[\frac{n-1}{2}\Big];$$

$$T(k) = S_2^2 - \frac{D_k}{n}\big((n-k)e_n + (n-k-1)e_{n-1} + ... + 2e_{k+2} + e_{k+1}\big)^2, \tag{15}$$

$$k = n-1, n-2, ..., n - \Big[\frac{n-1}{2}\Big];$$

якщо n - парне, то

$$T(0.5n) = S_2^2 - \frac{D_{0.5n}}{2n}\Big((0.5n-1)(e_0+e_n) + (0.5n-2)(e_1+e_{n-1}) + \ ... \ +$$

$$+(e_{0.5n-1} + e_{0.5n+1}) - e_{0.5n}\Big)^2. \tag{16}$$

Доведення. Позначимо через $z_k,\ k = 1, 2, ..., n-1,$ останній діагональний елемент матриці $(T_k'T_k)^{-1}$,

де

$$T_k = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{n} & 1 & 0 \\ . & . & . \\ \frac{k}{n} & 1 & 0 \\ \frac{k+1}{n} & 1 & \frac{1}{n} \\ . & . & . \\ 1 & 1 & \frac{n-k}{n} \end{pmatrix}.$$

Згідно з [4] різницю $S_2^2 - T(k)$ можна подати у вигляді

$$S_2^2 - T(k) = \frac{(\hat{c}_1^{(k)})^2}{z_k}. \tag{17}$$

Далі, після низки перетворень маємо

$$z_k = \frac{6n^2(n+1)(n+2)}{\big((2k+1)n - 2(k^2-1)\big)k(k+1)(n-k)(n-k+1)} = nD_k. \qquad (18)$$

Підставимо (4) і (18) в (17), отримуємо (14).
Формули (15) та (16) можна довести аналогічно.
Теорему 2 доведено. □

Позначимо

$$w_k = |ke_0 + (k-1)e_1 + ... + 2e_{k-2} + e_{k-1}|,\ k = 1, 2, ..., \Big[\frac{n-1}{2}\Big],$$

$w_k = |(n-k)e_n + (n-k-1)e_{n-1} + ... + 2e_{k+2} + e_{k+1}|, k = n-1, n-2, ..., n - \Big[\frac{n-1}{2}\Big]$;
якщо n — парне, то

$$w_{0.5n} = \frac{1}{2}\Big|(0.5n-1)(e_0+e_n) + (0.5n-2)(e_1+e_{n-1}) + ... + (e_{0.5n-1}+e_{0.5n+1}) - e_{0.5n}\Big|.$$

Очевидно $T(k),\ k = 1, 2, ..., \big[\frac{n-1}{2}\big]$, приймає найменше значення тоді, коли величина $g_k = w_k\sqrt{\frac{D_k}{n}}$ приймає найбільше значення.

4. Алгоритм прийняття або відхилення гіпотези H.

У послідовності

$$w_1,\ w_2,\ ...\ , w_{[\frac{n-1}{2}]}$$

виділимо строго зростаючу підпослідовність. Вона має вигляд

$$w_1 < w_2 < ... < w_{k_1} < w_{k_2} < w_{k_2+1} < ... < w_{k_3} < w_{k_4} < w_{k_4+1} < ...\ ,$$

при цьому

$$\begin{aligned} &w_k \le w_{k_1},\ k = k_1+1, ..., k_2-1, \\ &w_k \le w_{k_3},\ k = k_3+1, ..., k_4-1, \\ &\qquad ... \ . \end{aligned} \qquad (19)$$

Знаходимо g_1; якщо $g_1 > \sqrt{\frac{\varphi}{\varphi+1}S_2^2}$, то гіпотезу H відхиляємо, інакше для $k = 2, 3, ..., k_1, k_2, k_2+1, ..., k_3, k_4, k_4+1, ...$ знаходимо g_k та відразу перевіряємо нерівність

$$g_k > \sqrt{\frac{\varphi}{\varphi+1}S_2^2}. \qquad (20)$$

Якщо нерівність вірна, гіпотезу H відхиляємо, інакше для наступного k знаходимо g_k і так далі. Очевидно, найбільше g_k входить у підпослідовність

$$g_1, g_2, ..., g_{k_1}, g_{k_2}, g_{k_2+1}, ..., g_{k_3}, g_{k_4}, g_{k_4+1}, ...\ , \qquad (21)$$

тому що з урахуванням (13) та (19)

$$\begin{aligned} &g_k = w_k\sqrt{\frac{D_k}{n}} < w_{k_1}\sqrt{\frac{D_{k_1}}{n}} = g_{k_1},\ k = k_1+1, ..., k_2-1; \\ &g_k = w_k\sqrt{\frac{D_k}{n}} < w_{k_3}\sqrt{\frac{D_{k_3}}{n}} = g_{k_3},\ k = k_3+1, ..., k_4-1; \\ &\qquad ... \ . \end{aligned}$$

Якщо у послідовності (21) не знайшлося g_k, яке задовільняє нерівність (20), то у послідовності

$$w_{n-1},\ w_{n-2},\ ...,\ w_{n-[\frac{n-1}{2}]}$$

виділимо строго зростаючу підпослідовність та аналогічно для $k{=}n{-}1$, $n-2, ...$ шукаємо g_k, яке буде задовільняти нерівність (20). Якщо і у послідовності $g_{n-1}, g_{n-2}, ...$ не знайшлося такого g_k, гіпотезу H приймаємо (якщо n — парне, то треба буде перевірити ще нерівність (20) для $k = 0.5n$).

Зауваження 1. Зауваження Якщо n велике ($n \geq 20$), випадок, коли $\hat{t}^*$ не збігається з точкою спостереження, можна не розглядати, S_4^2 майже не зміниться (у багатьох випадках взагалі не зміниться).

5. Приклади

Розглянемо приклади на застосування цього алгоритму.

Приклад 1.

k	t	y	e	w	z	g
0	0	0.25	0.204			
1	0.05	0.33	0.146	0.204	486.3157	0.225
2	0.1	0.47	0.147	0.554	114.9683	0.297
3	0.15	0.54	0.079	1.051	48.7034	0.367
4	0.2	0.61	0.1	1.627	27.1765	0.424
5	0.25	0.72	-0.018	2.213	17.907	0.468
6	0.3	0.83	-0.047	2.781	13.2331	0.506
7	0.35	0,95	-0.065	3.302	10.6658	0.539
8	0.4	1.07	-0.083	3.758	9.226	0.571
9	0.45	1.18	-0.111	4.131	7.7778	0.576
10	0.5	1.25	-0.18	4.403	8,2555	0.633
11	0.55	1.36	-0.209	4.495	7.7778	0.653
12	0.6	1.44	-0.264	4.366	9.226	0.663
13	0.65	1.68	-0.166	3.97	10.6658	0.648
14	0.7	1.89	-0.094	3.408	13.2331	0.62
15	0.75	2.09	-0.033	2.752	17.907	0.582
16	0.8	2.25	-0.011	2.063	27.1765	0.538
17	0.85	2.48	0.08	1.363	48.7034	0.476
18	0.9	2.69	0.132	0.743	114.9683	0.398
19	0.95	2.87	0.193	0.275	486.3157	0.303
20	1	3.09	0.275			

Маємо таку лінійну регресійну модель

$$y = 2.770t + 0.046; \quad S_2^2 = 0.445.$$

Нехай $\alpha = 0.001$. Тоді $\sqrt{\frac{\varphi}{\varphi+1}S_2^2} = 0.284$. Маємо $g_1 < 0.284,\ g_2 > 0.284$. Гіпотезу H відхиляємо.

Якшо треба визначити S_4^2, знаходимо всю підпослідовність (21). Маємо

$$\hat{t}^* = t_{12}; \quad \hat{c}_1^{(12)} = 2.011; \quad S_4^2 = T(12) = 0.00676.$$

Приклад 2.

k	t	y	e	w	z	g
0	0	0	-0.476			
1	0.05	1	0.524	0.476	486.3157	0.525
2	0.1	0	-0.476	0.428	114.9683	
3	0.15	1	0.524	0.856	48.7034	0.299
4	0.2	0	-0.476	0.76	27.1765	
5	0.25	1	0.524	1.14	17.907	0.241
6	0.3	0	-0.476	0.996	13.2331	
7	0.35	1	0.524	1.328	10.6658	0.217
8	0.4	0	-0.476	1.136	9.226	
9	0.45	1	0.524	1.42	7.7778	0.198
10	0.5	0	-0.476	1.182	8,2555	
11	0.55	1	0.524	1.42	7.7778	0.198
12	0.6	0	-0.476	1.136	9.226	
13	0.65	1	0.524	1.328	10.6658	0.217
14	0.7	0	-0.476	0.996	13.2331	
15	0.75	1	0.524	1.14	17.907	0.241
16	0.8	0	-0.476	0.76	27.1765	
17	0.85	1	0.524	0.856	48.7034	0.299
18	0.9	0	-0.476	0.428	114.9683	
19	0.95	1	0.524	0.476	486.3157	0.525
20	1	0	-0.476			

Маємо таку лінійну регресійну модель

$$y = 0 * t + 0.476; \quad S_2^2 = 5.238.$$

Нехай $\alpha = 0.2$. Тоді $\sqrt{\frac{\varphi}{\varphi+1} S_2^2} = 0.601$. Бачимо, що всі $g_k < 0.601$. Гіпотезу H приймаємо.

Крім того,

$$\hat{t}^* = t_1; \quad \hat{c}_1^{(1)} = -11.579; \quad S_4^2 = T(1) = 4.96241.$$

Висновки

У алгоритмі значно спрощено процедуру прийняття гіпотези про рівність нулю параметру c_1. Крім того, завдяки теоремам 1 та 2 залишкову суму квадратів моделі двофазної лінійної регресії та оцінку параметра c_1 в окремих випадках можна знайти за формулою, без розв'язування системи лінійних алгебраїчних рівнянь.

Література

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. — Москва: Финансы и статистика, 1986. — 366 с.

2. Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л. Методы сплайн-функций. — Москва: Наука, 1980. — 352 с.
3. Себер Дж. Линейный регрессионный анализ. — Москва: Мир, 1980. — 456 с.
4. Демиденко Е. З. Линейная и нелинейная регрессии. — Москва: Финансы и статистика, 1981. — 304 с.
5. Савкіна М. Ю. Алгоритм перевірки на коректність моделі двофазної нелінійної регресії // Вісник Київського національного університету імені Тараса Шевченка. — 2015. — №3. — С. 115–120.

Надійшла 31.05.17