УДК 519.21

В.О. Мірошніченко¹, асп.

Узагальнений метод найменших квадратів для суміші нелінійних регресій

¹Київський національний університет імені Тараса Шевченка, 01033, Київ, вул. Володимирська, 64.

Розглянуто оцінку коефіцієнтів нелінійної регресії за спостереженнями з суміші зі скінченною кількістю компонент. Концентрації компонент у суміші є різними для різних спостережень. Для побудови оцінок використовується узагальнений метод найменших квадратів.

Ключові слова: асимптотична поведінка, метод оцінюючих рівнянь, модель суміші, нелінійна регресія, мінімаксні ваги.

We consider data in which each observed subject belongs to one of different subpopulations (components). The true number of component which a subject belongs to is unknown, but the researcher knows the probabilities that a subject belongs to a given component (concentration of the component in the mixture). The concentrations are different for different observations. So the distribution of the observed data is a mixture of components' distributions with varying concentrations. A set of variables is observed for each subject. Dependence between these variables is described by a nonlinear regression model. The coefficients of this model are different for different components. An estimator is proposed for these regression coefficients estimation based on the least squares and generalized estimating equations. Consistency of this estimator is demonstrated under general assumptions. A mixture of logistic regression models with continuous response is considered as an example. It is shown that the general consistency conditions are satisfied for this model under very mild assumptions. Performance of the estimator is assessed by simulations.

Key Words: asymptotic behavior, generalized estimation equations, mixture model, non-linear regression, minimax weights

Communicated by Prof. Kozachenko Yu.V.

1 Вступ

У даній статті розглядається задача оцінки параметрів нелінійної регресії у моделі суміші. Моделі скінченних сумішей були розвинуті у роботах [1], [2]. Існує два підходи до вивчення сумішей: параметричний і непараметричний, при чому результатів по непараметричному підходу більше. У роботі [5] застосований непараметричний підхід для побудови функцій розподілів у моделі скінченної суміші. Модель скінченної суміші використовується для опису соціологічних та медико-біологічних даних [6].

Ми розглядаємо суміш із M компонент зі змінними концентраціями. Від номера компонента залежить розподіл спостережуваних характеристик та параметри залежностей між ними. Для оцінки параметрів можна використати навантажений метод найменших квадратів. У [4] розглядається структурна модель лінійної

V.O. Miroshnychenko¹, PhD stud.

Generalized least squares estimates for mixture of nonlinear regressions

¹Taras Shevchenko National University of Kyiv, 01033, Kyiv, 64 Volodymyrska st.

регресії для моделі суміші, при чому використовуються мінімаксні вагові коефіцієнти [5].

У Теоремі 1 сформульовані загальні умови консистентності цих оцінок, у теоремі 2 вони конкретизовані для випадку логістичної функції регресії. Поведінки оцінок для вибірок фіксованого обсягу досліджено на модельованих вибірках.

2 Модель суміші

Розглянемо модель суміші зі змінними концентраціями. Кожен досліджуваний об'єкт O належить одному із M класів (компонентів суміші). Номер компонента, якому належить об'єкт, позначимо $\kappa(O) \in 1, \dots, M$. Ця характеристика не спостерігається. Вектор спостережуваних змінних об'єкта O позначимо $\xi(O)$. Будемо вважати, що розподіл спостережуваних змінних для

кожного компонента описується моделлю нелінійної структурної регресії.

Отже, $\xi(O) = (Y(O), X^1(O), \dots, X^d(O))^T$, де $Y(O)$ – відгук, $X(O) = (X^1(O), \dots, X^d(O))$ – регресори у моделі.

$$Y(O) = g(X(O), b^{(\kappa(O))}) + \varepsilon(O)$$

де g – деяка відома функція $g : \chi^d \times \Theta \rightarrow R$, $b^{(\kappa)} = (b_1^{(\kappa)}, \dots, b_d^{(\kappa)})^T \in \Theta \subseteq \mathbb{R}^d$, $\kappa = 1, \dots, M$ – невідомі коефіцієнти регресії для κ -ї компоненти суміші, $\varepsilon(O)$ – випадкова похибка. Припускаємо, що $\varepsilon(O)$:

$$E[\varepsilon(O)|\kappa(O) = m] = 0, m = 1, \dots, M,$$

та $\sigma_m^2 = \text{Var}[\varepsilon(O)|\kappa(O) = m] < \infty$. (σ_m^2 невідомі).

$F_{\varepsilon, m}(A) = P(\varepsilon(O) \in A | \kappa(O) = m)$ для довільної вимірної $A \subseteq R$ – розподіл випадкової похибки.

Вектори регресорів $X(O) = (X^1(O), \dots, X^d(O))$ вважаємо випадковими з розподілом, що залежить від $\kappa(O)$. Додатково ми припускаємо, що регресори $X(O)$ та $\varepsilon(O)$ – незалежні при фіксованому $\kappa(O) = m$, $m = 1, \dots, M$. Позначимо невідому функцію розподілу спостережуваних змінних $X(O)$

$$F_{X, m}(A) = P(X(O) \in A | \kappa(O) = m)$$

для довільної вимірної $A \subseteq \mathbb{R}^d$.

Вибірка Ξ_n , що спостерігається, складається зі значень $\xi_j = (Y_j, X_j^T)^T = \xi(O_j)$, $j = 1, \dots, n$, де $O_1, \dots, O_n$ незалежні об'єкти, які можуть належати до різних компонентів з ймовірностями $p_j^m = P(\kappa(O_j) = m)$, $m = 1, \dots, M$; $j = 1, \dots, n$. (ці ймовірності суміші відомі для кожного об'єкту)

3 Оцінки, отримані за допомогою оціночних рівнянь

Нехай далі k – фіксований номер компонента, для якого оцінюються параметри b_k . Для оцінки параметрів різних компонент використаємо мінімаксні ваги, визначені [5] і мають вигляд

$$a_{j:n}^{(k)} = \frac{1}{\det \Gamma_n} \sum_{m=1}^M (-1)^{m+k} \gamma_{km:n} p_{j,m},$$

де i – номер об'єкту у вибірці, а $\gamma_{km:n} = k m$ -й міnor матриці $\Gamma_n = (\langle p_k, p_m \rangle)_{k,m=1}^M$, де $p_k = (p_{i,k})_{i=1}^n$.

Для оцінювання невідомих параметрів регресії можна природно використати навантажений метод найменших квадратів. Для цього складається функціонал МНК із доданків, що помножені на наведені раніше навантаження.

$$G_n^k(\gamma) = \sum_{j=1}^n a_{j:n}^k (Y_j - g(X_j, \gamma))^2 \rightarrow \min_{\gamma \in \Theta}$$

Оцінки параметрів регресії є розв'язком оціночного рівняння, яке отримується шляхом диференціювання узагальненого функціоналу найменших квадратів $G_n^k(b_k)$:

$$\begin{aligned} \nabla G_n^k(\gamma) &= \sum_{j=1}^n \nabla (a_{j:n}^k (Y_j - g(X_j, \gamma))^2) = \\ &= \sum_{j=1}^n (-2) a_{j:n}^k s(X_j, Y_j, \gamma) = 0 \quad (1). \end{aligned}$$

Вектор оцінок буде стаціонарною точкою функціоналу але не завжди буде точкою мінімуму. Треба додати, що в деяких випадках розв'язком рівняння (1) буде множина точок. У такому випадку як оцінку можна обрати будь-який елемент цієї множини, але ми вимагаємо, щоб оцінка була вимірною функцією від даних.

Позначимо $J_n^{(k)}(\gamma) \stackrel{\text{def}}{=} \nabla G_n^k(\gamma)$.

Означення: $\hat{b}_n^{(k)}$ – розв'язок оціночного рівняння (1), тобто така вимірна функція від даних Ξ_n , що $J_n^k(\hat{b}_n) = 0$ м.н., називається оцінкою методу найменших квадратів.

Введемо наступні позначення:

$$\Gamma = \lim_{n \rightarrow \infty} \Gamma_n = \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{j=1}^n p_j^i p_j^k \right)_{i,k=1}^M$$

та

$$s(x, y, \gamma) = (y - g(x, \gamma)) \nabla g(x, \gamma)$$

Далі $X^{(m)}, Y^{(m)}, \varepsilon^{(m)}$ – випадкові величини із розподілами:

$$X^{(m)} \sim F_{X,m}, \varepsilon^{(m)} \sim F_{\varepsilon,m} - \text{незалежні,}$$

$$Y^{(m)} = g(X^{(m)}, b^{(m)}) + \varepsilon^{(m)}.$$

Теорема 1: Нехай виконуються такі умови:

0. $\det \Gamma \neq 0$.

1. Θ – компакт в R^d .

2. $E|\varepsilon^{(m)}|^\delta < \infty$ та $E\|X^{(m)}\|^\delta < \infty$ для деякої $\delta > 0 \forall m$.

3. $E|h(X^{(m)}, Y^{(m)})|^{1+\delta} < \infty$ для δ з умови 3, $\forall m$, де

$$h(x, y) = \sup_{\gamma \in \Theta} \|s(x, y, \gamma)\|.$$

4. $g(x, \gamma)$ та $\nabla g(x, \gamma)$ – рівностепеневно неперервні по x , та неперервні по γ .

5. $E s(X^{(k)}, Y^{(k)}, \gamma) = 0 \Leftrightarrow \gamma = b^{(k)}$.

Тоді $\hat{b}_n^{(k)} \xrightarrow{P} b^{(k)}, n \rightarrow \infty$.

Доведення. Доведення консистентності послідовності оцінок спирається на наступне твердження, яке сформульоване на основі теореми 5.5 та леми 5.3 із [3]:

Твердження 1:

1. Θ – компакт в R^d .

2. $\sup_n E\|\xi_n\|^\delta < \infty$ для деякої $\delta > 0$.

3. $\sup_n E|h(\xi_n)|^{1+\delta} < \infty$ для δ з умови 3. Тут $h(\xi) = \sup_{\gamma \in \Theta} \|s(\xi, \gamma)\|$.

4. $s(\xi, \gamma)$ – рівностепеневно неперервна по ξ .

5. $\lim_{n \rightarrow \infty} E n^{-1} S_n(\gamma) = 0 \Leftrightarrow \gamma = b^{(k)}$, де $S_n(\gamma) = \sum_{j=1}^n \psi(\xi_j, \gamma)$, та $\psi(x, \gamma)$ – оцінююча функція.

Тоді $\hat{b}_n^{(k)} \xrightarrow{P} b^{(k)}, n \rightarrow \infty$.

Перевіримо виконання умов наведеного твердження.

Позначимо $\Delta_n^{(k)}(\gamma) = E_{F_{X,k}} \frac{1}{n} J_n^{(k)}(\gamma)$. Покажемо, що виконується умова 5) твердження 1:

$$\lim_{n \rightarrow \infty} \Delta_n^{(k)}(\gamma) = 0 \Leftrightarrow \gamma = b^{(k)}.$$

Покажемо, що $\Delta_n^{(k)}(\gamma) = -2E s(X^{(k)}, Y^{(k)}, \gamma)$. Дійсно,

$$\begin{aligned} \Delta_n^{(k)}(\gamma) &= E \sum_{s=1}^M \sum_{i=1}^n (-2n^{-1} a_{i:n}^s p_i^s) (g(X_i, b_s) - \\ &\quad g(X_i, \gamma)) \nabla g(X_i, \gamma) = \\ &= \sum_{s \neq k} \sum_{i=1}^n (-2n^{-1} a_{i:n}^s p_i^s) E(g(X_i, b_s) - \\ &\quad g(X_i, \gamma)) \nabla g(X_i, \gamma) + \\ &\quad \sum_{i=1}^n (-2n^{-1} a_{i:n}^k p_i^k) E_{F_{X,k}}(g(X_i, b_k) - \\ &\quad g(X_i, \gamma)) \nabla g(X_i, \gamma) \quad (2) \end{aligned}$$

Позначимо:

$$C_s(\gamma) = -2E_{F_{X,k}}(g(X_i, b_k) - g(X_i, \gamma)) \nabla g(X_i, \gamma).$$

З урахуванням [5] незміщеності коефіцієнтів $a_{j:n}^k$: $\sum_{j=1}^n a_{j:n}^k p_j^k = n I_{k=t}$, вираз (2) спрощується:

$$\begin{aligned} \Delta_n^{(k)}(\gamma) &= \sum_{s \neq k} \sum_{i=1}^n (n^{-1} a_{i:n}^s p_i^s) C_s(\gamma) + \\ &\quad \sum_{i=1}^n (n^{-1} a_{i:n}^k p_i^k) C_k(\gamma) = \\ &\quad \sum_{s \neq k} C_s(\gamma) I_{s=k} + C_k(\gamma) = C_k(\gamma) \end{aligned}$$

З умови 5 матимемо, що $\Delta_n^{(k)}(\gamma) = -2E_{F_{X,k}} s(X^{(k)}, Y^{(k)}, \gamma) \rightarrow 0, n \rightarrow \infty \Leftrightarrow \gamma = b^{(k)}$.

Далі перевіримо умову 4 твердження 1. Покажемо що відображення $(y - g(x, \gamma)) \nabla g(x, \gamma)$ є рівностепеневно неперервним. Дійсно, оскільки g та всі її похідні обмежені і неперервні, то і функції $(\nabla((y - g(x, \gamma)) \nabla g(x, \gamma)))_{(s,d)}$; $s = 1, \dots, D$; $d = 1, \dots, D$ будуть обмежені функціями

$$C|1 + y||x_{(s)}x_{(d)}|; d = 1, \dots, D; s = 1, \dots, D;$$

тут C – деяка константа.

Тому $(y - g(x, \gamma)) \nabla g(x, \gamma)$ – рівностепеневно неперервна.

Бачимо, що усі умови твердження 1 виконуються і має місце консистентність оцінки. $\square$

Як приклад розглянемо випадок, коли функція $g(x, \gamma)$ є логістичною функцією:

$$g(x, \gamma) = \frac{1}{1 + e^{-\langle x, \gamma \rangle}} \quad (3)$$

Теорема 2 Нехай виконуються умови 1), 2), 3) теореми 1, функція g має вигляд (3) і $\forall \gamma \in R^d$: $P(\langle X^{(k)}, \gamma \rangle = 0) < 1$.

Тоді $\hat{b}_n^{(k)} \xrightarrow{P} b^{(k)}, n \rightarrow \infty$.

Доведення. Доведення спирається на використання теореми 1, тому перевіримо виконання умов 4), 5), 6). Умова 4) впливає і умови 3), обмеженості функції g та всіх її похідних.

Далі покажемо, що умова 5) теореми 1 виконується, тобто, що функції $g(x, \gamma), \nabla g(x, \gamma)$ – рівностепеневно неперервні по x . Доведемо, що градієнт та гессіан $g(x, \gamma)$ по γ будуть обмеженими. Позначимо

$$\begin{aligned} \nabla g(x, \gamma) &= (r(x, \gamma)_d)_{d=1}^D \text{ та} \\ H(g(x, \gamma)) &= (z(x, \gamma)_{(d,s)})_{d,s=1}^D \end{aligned}$$

Легко бачити, що $r(x, \gamma)$ та $z(x, \gamma)$ – обмежені:

$$\begin{aligned} |r(x, \gamma)_d| &= \\ &= |g(x, \gamma)(1 - g(x, \gamma))(-x_d)| \leq \\ &\leq |x_d|; d = 1, \dots, D; \\ |z(x, \gamma)_{(d,s)}| &= \\ &= |g(x, \gamma)(1 - g(x, \gamma))(1 - 2g(x, \gamma))(-x_d x_s)| \leq \\ &\leq |x_d x_s|; d, s = 1, \dots, D \end{aligned}$$

Ми показали, що градієнт та гессіан обмежені, і тому сімейства функцій $g(x, \gamma), \nabla g(x, \gamma)$ будуть рівностепенно неперервними.

Перевіримо умову 6). Покажемо, що з рівності $\gamma = b^{(k)}$ випливає:

$$Es(X^{(k)}, Y^{(k)}, \gamma) = 0.$$

Дійсно,

$$\begin{aligned} Es(X^{(k)}, Y^{(k)}, b^{(k)}) &= \\ &= E(Y^{(k)} - -g(X^{(k)}, b^{(k)})) \nabla g(X^{(k)}, b^{(k)}) \quad (4) \end{aligned}$$

За умови, що $Y^{(k)} = g(X^{(k)}, b^{(k)}) + \varepsilon^{(k)}$, (4) набуває наступного вигляду:

$$\begin{aligned} Es(X^{(k)}, Y^{(k)}, b^{(k)}) &= \\ &= E(g(X^{(k)}, b^{(k)}) + \varepsilon^{(k)} - -g(X^{(k)}, b^{(k)})) \times \\ &\quad \times \nabla g(X^{(k)}, b^{(k)}) = \\ E\varepsilon^{(k)} \nabla g(X^{(k)}, b^{(k)}) &= E\varepsilon^{(k)} E \nabla g(X^{(k)}, b^{(k)}) = 0, \\ \text{оскільки } X^{(k)} \text{ та } \varepsilon^{(k)} \text{ незалежні та } E(\varepsilon^{(k)}) &= 0. \end{aligned}$$

Для доведення еквівалентності з умови 6) покажемо що з рівності $Es(X^{(k)}, Y^{(k)}, \gamma) = 0$ матимемо рівність $\gamma = b^{(k)}$. Доведемо це методом від супротивного. Припустимо, що $Es(X^{(k)}, Y^{(k)}, \gamma) = 0$, але $\gamma \neq b^{(k)}$.

Оскільки $g(X^{(k)}, \gamma)$ – монотонно зростаюча функція по $\langle X^{(k)}, \gamma \rangle$, а $\nabla g(X^{(k)}, \gamma) > 0$, то вираз

$$\begin{aligned} &\langle (g(X^{(k)}, b^{(k)}) - g(X^{(k)}, \gamma)) \times \\ &\times \nabla g(X^{(k)}, \gamma) X^{(k)}, b^{(k)} - \gamma \rangle \geq 0, \end{aligned}$$

а отже і його математичне сподівання невід'ємне. Далі покажемо, що воно не може бути нулем за наших припущень. Дійсно, математичне сподівання виразу (5) буде нулем при виконанні умови

$$\begin{aligned} &(g(X^{(k)}, b^{(k)}) - g(X^{(k)}, \gamma)) \times \\ &\times \nabla g(X^{(k)}, \gamma) \langle X^{(k)}, b^{(k)} - \gamma \rangle \equiv 0 \end{aligned}$$

майже напевно.

Оскільки градієнт $\nabla g(X^{(k)}, \gamma)$ обмежений і строго додатний та $(g(X^{(k)}, b^{(k)}) - g(X^{(k)}, \gamma))$ не дорівнює нулю через монотонність g по $\langle X^{(k)}, b \rangle$, тоді $\langle X^{(k)}, b^{(k)} - \gamma \rangle = 0$ майже напевно. Нагадаємо, що згідно з умовою теореми $P(\langle X^{(k)}, b \rangle = 0) < 1, \forall b \in \mathbb{R}^d$. Тому початкове припущення хибне і $\gamma = b^{(k)}$.

Умови 4), 5), 6) теореми 1 перевірені та має місце консистентність оцінки для моделі логістичної регресії. $\square$

4 Результати моделювання

Нами була проведена серія експериментів для перевірки збіжності наведених вище оцінок. Суміш складалася із $M = 2$ компонентів. Розподіли регресорів та похибок для компонента $k \in \{1, 2\}$ мали вигляд: $X^{(k)} \sim N(m_k, \sigma_k^2)$, $\varepsilon^{(k)} \sim N(0, \sigma_{\varepsilon, k}^2)$. Функція $g(x, \gamma)$ у моделюванні має вигляд логістичної функції:

$$g(x, b) = \frac{1}{1 + e^{-b_0 - b_1 x}}$$

Відповідно, параметри розподілів та параметри логістичної регресії кожного компонента наведені у таблиці 1:

	КОМПОНЕНТ	
	1	2
m_k	0.0	1.0
ε_k^2	2.0	2.0
$\sigma_{\varepsilon, k}^2$	0.05	0.05
$b_0^{(k)}$	0.5	0.5
$b_1^{(k)}$	2	-1/3

Таблиця 1.

Ймовірності $p_{j, m}$ були згенеровані наступною стохастичною процедурою

$$p_{j, m} = \frac{u_{j, m}}{\sum_{k=1}^M u_{j, k}},$$

де $u_{j, m}$ – незалежні псевдовипадкові числа, розподілені рівномірно на $[0, 1]$.

Для різних розмірів вибірки n експеримент був повторений $H = 1000$ разів. За отриманими

вибірками з оцінок ми порахували зміщення та дисперсію. Результати наведені на рисунках 1 та 2. Чорним пунктиром та суцільною чорною лінією позначені графіки для першого та дру-

гого компонента відповідно. Крапками позначили величину характеристик параметрів $b_1^{(k)}$, квадратами – значення параметру $b_0^{(k)}$.

Список використаних джерел

1. D. M. Titterington, A. F. Smith, U. E. Makov, Analysis of Finite Mixture Distributions / D. M. Titterington, A. F. Smith, U. E. Makov // Wiley, New York (1985)
2. McLachlan, G.J., Peel, D. Finite mixture models / McLachlan, G.J., Peel, D. // Wiley-Interscience (2000)
3. Shao, Jun, Mathematical Statistics / Shao, Jun // Springer, 2007.
4. Liubashenko D., Maiboroda R. Linear regression by observations from mixture with varying concentrations / Liubashenko D., Maiboroda R // Modern Stochastics: Theory and Applications, 2 ,No 4, 343 – 353, (2015)
5. Maiboroda, R., Statistical analysis of mixtures / Maiboroda, R // Kyiv University Publishers, Kyiv (in Ukrainian) (2003)
6. Maiboroda, R., Sugakova, O., Statistics of mixtures with varying concentrations with application to DNA microarray data analysis / Maiboroda, R., Sugakova, O. // Journal of nonparametric statistics. 24 , No 1 201–205 (2012)

References

1. TITTERINGTON, D. M., SMITH, A. F., MAKOV, U. E. (1985) Analysis of Finite Mixture Distributions. Wiley, New York
2. MCLACHLAN, G.J., PEEL, D. (2000) Finite mixture models. Wiley-Interscience
3. SHAO, JUN (2007) Mathematical Statistics. Springer, 2007.
4. LIUBASHENKO, D., MAIBORODA, R. (2015) Linear regression by observations from mixture with varying concentrations. Modern Stochastics: Theory and Applications, 2 ,No 4, 343 – 353
5. MAIBORODA, R. (2003) Statistical analysis of mixtures. Kyiv University Publishers, Kyiv (in Ukrainian)
6. MAIBORODA, R., SUGAKOVA, O. (2012) Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. Journal of nonparametric statistics. 24 , No 1 201–205 (2012)

Received: 24.05.2018