УДК 519.21

<https://doi.org/10.17721/1812-5409.2021/1.3>

В.О. Мірошніченко¹, асп.

V.O. Miroshnychenko¹, PhD stud.

Швидкі розрахунки оцінок коваріаційних матриць методом складаного ножа

Fast calculations of Jackknife covariance matrix estimator

¹Київський національний університет імені Тараса Шевченка, 01033, Київ, вул. Володимирська, 64, e-mail: vitaliy.miroshnychenko@gmail.com.

¹Taras Shevchenko National University of Kyiv, 01033, Kyiv, 64 Volodymyrska st., e-mail: vitaliy.miroshnychenko@gmail.com.

У роботі розглядається побудова довірчих еліпсоїдів для параметрів нелінійної регресії за спостереженнями з суміші кількох компонентів. Концентрації компонентів є різними для різних спостережень. Для побудови еліпсоїдів використовуються оцінки узагальненого методу найменших квадратів. Коваріаційні матриці цих оцінок оцінюються за допомогою різних модифікацій методу складаного ножа (jackknife). Якість отриманих еліпсоїдів у випадку логістичної функції регресії досліджується за допомогою імітаційного моделювання.

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

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. Normality of estimator for nonlinear regression parameters is demonstrated under general assumptions. A mixture of logistic regression models with continuous response is considered as an example. In the paper we construct confidence ellipsoids for the regression parameters based on the modified least squares estimators. The covariances of these estimators are estimated by the multiple modifications of jackknife technique. Performance of the obtained confidence ellipsoids is assessed by simulations.

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

Статтю представив д.ф.-м.н., доцент Ямненко Р.Є.

1 Вступ

Моделі скінченних сумішей широко застосовуються у математичній статистиці для опису неоднорідних даних [1], [2]. У випадку, коли концентрації компонентів суміші є різними для різних спостережень, непараметричний підхід до оцінювання розподілів компонентів розглянуто у [3], [4]. Для суміші лінійних регресій модифікація оцінок методу найменших квадратів була запропонована у [5]. Для побудови довірчих інтервалів та еліпсоїдів для коефіцієнтів лінійної регресійної моделі можна використовувати техніку складаного ножа (СК, jackknife) [6].

Дана стаття продовжує започатковану у [9], [10] розробку техніки статистичного аналізу да-

них, що описуються моделлю суміші нелінійних регресій. У цих роботах введено узагальнені оцінки методу найменших квадратів для параметричних регресійних моделей компонентів суміші, отримані умови їхньої консистентності та асимптотичної нормальності. Тепер ми дослідимо можливість побудови довірчих еліпсоїдів на основі цих оцінок. Для цього нам будуть потрібні оцінки коваріаційних матриць параметрів регресії для різних компонентів суміші. Для цього ми використаємо метод СК, який вперше був запропонований у роботах Quenouille (1956)[7] та Tukey (1958) [8]. Загальна теорія цього методу для вибірок з незалежних, однаково розподілених спостережень викладена у п.

5.5.2 [11].

У даній роботі розглянуто три модифікації оцінки СК, кожна із яких базується на використанні одного кроку процедури послідовного наближення Ньютона-Рафсона. Якість довірчих еліпсоїдів, побудованих на основі цих оцінок порівнюється у імітаційних експериментах на модельованих даних.

2 Модель суміші

Розглянемо модель суміші зі змінними концентраціями. Кожен досліджуваний об'єкт O належить одному із M класів (компонентів суміші). У кожного об'єкта є спостережувані та неспостережувані змінні. Номер компонента, якому належить об'єкт, не спостерігається. Його позначатимемо символом $\kappa(O) \in 1, \dots, M$. Вектор спостережуваних змінних об'єкта O позначимо $\xi(O)$. Будемо вважати, що розподіл спостережуваних змінних для кожного компонента описується моделлю нелінійної структурної регресії. Таким чином

$$\xi(O) = \left(Y(O), X^1(O), \dots, X^d(O) \right)^T,$$

де $Y(O)$ – відгук, $X(O) = \left(X^1(O), \dots, X^d(O) \right)$ – регресори у моделі

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

Тут g – деяка відома функція $g: \mathbb{R}^D \times \Theta \rightarrow \mathbb{R}$, $b^{(\kappa)} = (b_1^{(\kappa)}, \dots, b_D^{(\kappa)})^T \in \Theta \subseteq \mathbb{R}^D$, $\kappa = 1, \dots, M$ – невідомі коефіцієнти регресії, які відповідають κ -му компоненту суміші. Саме для цих коефіцієнтів далі буде побудовано довірчі еліпсоїди. $\varepsilon(O)$ – центрована випадкова похибка

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

зі скінченною дисперсією

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

(σ_m^2 невідомі).

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

Вектори регресорів $X(O) = \left(X^1(O), \dots, X^D(O) \right)$ вважаємо випадковими з розподілом, що залежить від $\kappa(O)$. Позначимо невідому функцію розподілу спостережуваних змінних $X(O)$

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

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

Додатково ми припускаємо, що регресори $X(O)$ та похибки $\varepsilon(O)$ – незалежні при фіксованому $\kappa(O) = m$, $m = 1, \dots, M$.

Вибірка Ξ_n , що спостерігається, складається зі значень $\xi_i = (Y_i, X_i^T)^T = \xi(O_i)$, $i = 1, \dots, n$, де $O_1, \dots, O_n$ – незалежні об'єкти, які можуть належати до різних компонентів з ймовірностями

$$p_j^k = \mathbf{P}\{\kappa(O_j) = k\},$$

$$P = \left(p^k \right)_{k=1}^M = \left(p_i^k \right)_{i=1, n}^{k=1, M}; \quad \sum_{k=1}^M p_i^k = 1, 1 \leq i \leq n$$

(ці змішуючі ймовірності відомі для кожного об'єкта)

3 Оцінки параметрів, що отримані за допомогою оцінюючих рівнянь

Нехай далі k – фіксований номер компонента, для якого будується оцінка вектору параметрів $b^{(k)}$. Для побудови оцінок параметрів регресії різних компонент використовуємо мінімаксні навантаження, визначені у [3]. Позначимо $\Gamma_n = \left(\langle p^k, p^m \rangle \right)_{k, m=1}^M$. Тоді матриця мінімакських навантажень має вигляд

$$A = \left(a_{i:n}^{(k)} \right)_{k=1, M}^{i=1, n} = P \Gamma_n^{-1}. \quad (1)$$

Для оцінювання невідомих параметрів регресії природно використати навантажений метод найменших квадратів. Для цього складається функціонал МНК

$$G_n^k(\gamma) = \sum_{i=1}^n a_{i:n}^{(k)} \left(Y_i - g(X_i, \gamma) \right)^2.$$

Оцінки параметрів регресії є розв'язком оціночного рівняння, яке отримується шляхом диференціювання $G_n^k(b^{(k)})$:

$$\begin{aligned} \nabla G_n^k(\gamma) &= \sum_{i=1}^n \nabla \left(a_{i:n}^{(k)} (Y_i - g(X_i, \gamma))^2 \right) = \\ &= \sum_{i=1}^n (-2) a_{i:n}^{(k)} s(\xi_i, \gamma) = 0, \end{aligned} \quad (2)$$

де $s(z, \gamma)$ – елементарна оцінювальна функція і

$$s(z, \gamma) = \left(y - g(x, \gamma) \right) \nabla g(x, \gamma), \quad z = (x, y).$$

Вектор оцінок буде стаціонарною точкою функціоналу $\nabla G_n^k(\gamma)$, але не завжди буде точкою мінімуму. У 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 $S_n(\gamma) = \sum_{i=1}^n a_{i:n}^{(k)} s(\xi_i, \gamma)$. Оцінки параметрів розглядатимуться для кожного компонента окремо, тому над оцінювальним виразом S_n не будемо писати індекс компонента k .

Означення: $\hat{b}_n^{(k)}$ – розв'язок оціночного рівняння (2), тобто така вимірна функція від даних Ξ_n , що $S_n(\hat{b}_n^{(k)}) = 0$ м.н., називається оцінкою методу найменших квадратів (МНК).

Введемо наступне позначення

$$\Gamma_\infty = \lim_{n \rightarrow \infty} n^{-1} \Gamma_n$$

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

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

та

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

Також позначимо наступні суми

$$\langle a^{(k)^2} p^t p^f \rangle_n = \sum_{i=1}^n a_{i:n}^{(k)^2} p_i^f p_i^t,$$

$$\langle a^{(k)^2} p^t \rangle_n = \sum_{f=1}^M \langle a^{(k)^2} p^t p^f \rangle_n.$$

4 Складності оцінювання матриці коваріації оцінок параметрів регресії

У статті [10] наведено доведення асимптотичної нормальності оцінок регресії для моделі регресійної суміші. Далі наведено переформульовану теорему про асимптотичну нормальність оцінок МНК.

Позначимо

$$h(z) = \sup_{\gamma \in \Theta} \|\partial s(z, \gamma) / \partial \gamma\|, i = \overline{1, n}.$$

Теорема 4.1. Якщо має місце збіжність за ймовірністю послідовності оцінок $\hat{b}_n^{(k)}$ до $b^{(k)}$ і виконуються такі умови:

1 Θ – компактна множина, що включає $b^{(k)}$

2 $\mathbb{E}_t \|\xi^{(t)}\|^\delta < \infty$, для $1 \leq t \leq M$ та деякого $\delta > 0$.

3 Нехай існують функції $\partial s(z, \gamma) / \partial \gamma$ і для них виконуються наступні такі дві умови:

3.a $\mathbb{E}_t |h(\xi^{(t)})|^{1+\delta} < \infty$ для $\delta > 0$ з умови 2 та $1 \leq t \leq M$.

3.b Для довільних $c > 0$ та послідовності $z_i \in \mathbb{R}^{D+1}$, такої, що $\|z_i\| < c$, і довільної відкритої множини $U \in \Theta$, функції $\partial s(z_i, \gamma) / \partial \gamma$, $\gamma \in U$ є рівностепенно неперервними на U .

4 $\exists \lim_{n \rightarrow \infty} n^{-1} \Gamma_n = \Gamma_\infty$

5 $\mathbb{E}_k \|s(\xi^{(k)}, \theta^{(k)})\|^2 < \infty$, $k = 1, \dots, M$.

6 Існують граничні матриці $\lim_{n \rightarrow \infty} n \text{Var } S_n(b^{(k)})$ та $\lim_{n \rightarrow \infty} L_n(b^{(k)}) = L(b^{(k)})$ і вони невід'язані, де $L_n(b^{(k)}) = -\mathbb{E} \nabla S_n(b^{(k)})$.

Тоді

$$V_n^{-\frac{1}{2}} (\hat{b}_n^{(k)} - b^{(k)}) \rightarrow^w N_D(0, I_D),$$

$$V_n = L_n^{-1}(b^{(k)}) \text{Var } S_n(b^{(k)}) L_n^{-1}(b^{(k)})^T,$$

$$L_n(b^{(k)}) = \mathbb{E}_k \nabla g(X^{(k)}, b^{(k)}) \nabla g^T(X^{(k)}, b^{(k)}),$$

$$\begin{aligned} \text{Var } S_n(b^{(k)}) &= \sum_{t=1}^M (\mathbb{E}_t s(\xi^{(t)}, b^{(k)}) s(\xi^{(t)}, b^{(k)})^T) \langle a^{(k)^2} p^t \rangle_n - \\ &- \sum_{t=1}^M \sum_{f=1}^M (\mathbb{E}_t s(\xi^{(t)}, b^{(k)}) \mathbb{E}_f s(\xi^{(f)}, b^{(k)})^T) \langle a^{(k)^2} p^t p^f \rangle_n. \end{aligned}$$

Теорема 4.1 дає вигляд матриці коваріацій V_n , проте її використання ускладнюється декількома проблемами. Перша полягає в тому, що для деяких функцій $g(x, b)$ (наприклад логістичної, $g(x, b) = \sigma(\langle x, b \rangle)$), важко порахувати математичні сподівання $\mathbb{E}_t s(\xi^{(t)}, b^{(k)}) s(\xi^{(t)}, b^{(k)})^T$ і $\mathbb{E}_t s(\xi^{(t)}, b^{(k)})$, $\forall k, t = 1, \dots, M$. Інша проблема полягає в тому, що справжні значення параметрів для розрахунку матриці V_n невідомі. Експерименти зі статті [10] показали, що ці математичні сподівання можна рахувати наближено, і що замість справжніх параметрів регресії можна брати оцінки $\hat{b}_n^{(k)}$. Наведені дві проблеми можливо обійти, використовуючи техніку СК для оцінки матриці V_n .

Слідуючи [11] п. 5.5.2, опишемо стандартну процедуру СК. Далі $\hat{b}_n^{(k)}$ – оцінки методом оцінюючих рівнянь для k -го компонента, а $\hat{b}_{-1}^{(k)}, \dots, \hat{b}_{-n}^{(k)}$ – оцінки отримані методом оцінюючих рівнянь із вилученими 1-м, ..., n -м спостереженнями відповідно. Символом $S_{i-}(\gamma)$ позначимо оцінювальну функцію із вилученим i -м спостереженням. Оцінки розглядатимуться для кожного компонента окремо, тому над оцінювальним виразом S_{i-} також не будемо писати індекс компонента k .

Означення. СК-оцінкою матриці розсіювання оцінок $\hat{b}_n^{(k)}$ називається матриця

$$\hat{V}_J = \sum_{i=1}^n (\hat{b}_{i-}^{(k)} - \hat{b}_n^{(k)}) (\hat{b}_{i-}^{(k)} - \hat{b}_n^{(k)})^T. \quad (3)$$

У статті [10], точність СК-оцінок визначених (3) була досліджена методом імітаційного моделювання. Потрібно зазначити, що дана процедура оцінки потребує багато обчислень, що є значним недоліком. У наступному розділі описано декілька можливих модифікацій оцінки для розв'язання цієї проблеми. Перша ідея – це використати оцінку, в якій при наближеному обчисленні оцінок $\hat{b}_{-1}^{(k)}, \dots, \hat{b}_{-n}^{(k)}$ використовується лише один крок процедури Ньютона–Рафсона (НР). Інші два способи є модифікацією першого, якщо у однокроковій процедурі Ньютона–Рафсона замість гессіана відображення записати його оцінку.

5 Пришвидшення побудови оцінок матриці розсіювання методом СК

Процедура СК потребує розрахунку оцінок $\hat{b}_{i-}^{(k)}, i = 1, \dots, n$, які в свою чергу будуються за навантаженнями $A_{(i-)}$. Ці навантаження можна рахувати кожен раз заново, проте у роботі [6] наведений алгоритм швидкого перерахунку $A_{(i-)}$ на основі вже відомої матриці навантажень A :

$$\Gamma_{(i-)}^{-1} = \Gamma^{-1} + \frac{1}{1 - h_i} \Gamma_n^{-1} p_i p_i^T \Gamma_n^{-1},$$

тут $p_i = (p_i^1, \dots, p_i^M)$ та $h_i = p_i^T \Gamma_n^{-1} p_i$

Наведена процедура зменшує кількість операцій розрахунку навантажень $A_{(i-)}$ із $o(n^2)$ до $o(n)$.

Для пришвидшення побудови оцінок $\hat{b}_{i-}^{(k)}, i = 1, \dots, n$ далі використаємо однокрокову

процедуру НР. Вона дає нам приблизні значення $\hat{b}_{i-}^{(k)}$ значно швидше, ніж точне розв'язування рівняння (1). Один крок методу НР у сторону зменшення узагальненого функціоналу найменших квадратів виглядає так:

$$\hat{b}_{i-}^{(k)} = \hat{b}_n^{(k)} - \left[\nabla S_{i-}(\hat{b}_n^{(k)}) \right]^{-1} S_{i-}(\hat{b}_n^{(k)}), \quad (4)$$

тут

$$S_{(i-)}(b) = \sum_{j \neq i} a_{j:n(-i)}^{(k)} (Y_j - g(X_j, b)) \nabla g(X_j, b),$$

та

$$\begin{aligned} \nabla S_{(i-)}(b) = & - \sum_{j \neq i} a_{j:n(-i)}^{(k)} \nabla g(X_j, b) \nabla g^T(X_j, b) + \\ & \sum_{j \neq i} a_{j:n(-i)}^{(k)} (Y_j - g(X_j, b)) Hg(X_j, b). \end{aligned}$$

$Hg(X_j, b)$ – гессіан відображення g по параметру b .

$$\text{Позначимо } SS_{JK} = \left[\sum_{i=1}^n S_{i-}(\hat{b}_n^{(k)}) S_{i-}(\hat{b}_n^{(k)})^T \right]$$

Якщо підставити (4) у (3), то буде отримана перша модифікація оцінки $\hat{V}_J$

$$\hat{V}_{n,1} = \sum_{i=1}^n \left[\nabla S_{i-}(\hat{b}_n^{(k)}) \right]^{-1} S_{i-}(\hat{b}_n^{(k)}) S_{i-}(\hat{b}_n^{(k)})^T \left[\nabla S_{i-}(\hat{b}_n^{(k)}) \right]^{-1} \quad (5)$$

Далі ми розглянемо дві модифікації, що отримані з (5) заміною виразу $\nabla S_{i-}(\hat{b}_n^{(k)})$ на $-\nabla S_n(\hat{b}_n^{(k)})$, або $L(\hat{b}_n^{(k)})$. Застосування цих двох модифікацій дають змогу не рахувати матриці $\nabla S_{i-}(\hat{b}_n^{(k)})$.

Заміна $\nabla S_{i-}(\hat{b}_n^{(k)})$ на $\nabla S_n(\hat{b}_n^{(k)})$ дає другу модифікацію оцінки V_n

$$\hat{V}_{n,2} = \left[\nabla S_n(\hat{b}_n^{(k)}) \right]^{-1} SS_{JK} \left[\nabla S_n(\hat{b}_n^{(k)}) \right]^{-1}.$$

Третя модифікація будується заміною $\nabla S_{i-}(\hat{b}_n^{(k)})$ на $-L(\hat{b}_n^{(k)})$. Тоді модифікована оцінка виглядатиме так:

$$\hat{V}_{n,3} = \left[L(\hat{b}_n^{(k)}) \right]^{-1} SS_{JK} \left[L(\hat{b}_n^{(k)}) \right]^{-1}.$$

6 Моделювання

Нехай $\hat{V}_n$ – консистентна оцінка V_n , тобто $V_n^{-1} \hat{V}_n$ збігається за ймовірністю до одиничної матриці при $n \rightarrow \infty$. Позначимо $\chi_D^2(\alpha)$ – квантиль рівня $1 - \alpha$ χ^2 -розподілу з D ступенями вільності,

$$B_{k;n} = \{b \in \mathbb{R}^D : (b - \hat{b}_n^{(k)})^T \hat{V}_n^{-1} (b - \hat{b}_n^{(k)}) \leq \chi_D^2(\alpha)\}.$$

Тоді за теоремою 4.1

$$\mathbf{P}\{b^{(k)} \in B_{k;n}\} \rightarrow 1 - \alpha,$$

тобто $B_{k;n}$ є асимптотичним довірчим еліпсоїдом для вектора справжніх значень коефіцієнтів регресії k -того компонента суміші.

Для перевірки точності таких довірчих еліпсоїдів, які використовують різні модифікації СК-оцінок коваріаційної матриці V_n , було проведено ряд імітаційних експериментів.

У кожному експерименті для заданого обсягу вибірки генерувались $N = 1000$ вибірок із суміші зі змінними концентраціями. По кожній згенерованій вибірці будувалася оцінка $\hat{b}^{(k)}$ вектора параметрів $b^{(k)}$ і їх матриця коваріації V_n оцінювалась трьома оцінками $\hat{V}_{n,t}$, де $t \in \{1, 2, 3\}$. Далі для кожної вибірки перевірялося попадання $b^{(k)}$ до довірчого еліпса із центром у $\hat{b}_n^{(k)}$ та матрицею коваріації, що дорівнювала $\hat{V}_{n,t}$. Номінальним рівнем довірчих еліпсоїдів обрано значення $\alpha = 0.05$. Результатами експериментів є частоти попадання справжніх значень у відповідні еліпсоїди.

Номери компонент генерованих об'єктів $\kappa_i = \kappa(O_i)$, $i = 1, \dots, n$ обиралися відповідно до розподілів $(p_i^1, \dots, p_i^M)_{i=1}^n$, де ймовірності p_i^k згенеровані наступною процедурою:

$$p_i^k = \frac{u_i^k}{\sum_{t=1}^M u_i^t}; u_i^t \sim U[0, 1]$$

Спостережувані характеристики об'єктів для m -го компонента суміші $\xi_i^{(m)} = (Y_i^{(m)}, X_i^{(m)})$ генерувалися наступним чином:

$$Y_i^m = g(X_i^m, b^{(m)}) + \varepsilon_i^{(m)}; X_i^{(m)} \simeq N(\mu^{(m)}, \Sigma^{(m)}),$$

де

$$g(X, \gamma) = \frac{1}{1 + e^{-\gamma_0 - \sum_{d=1}^D X_d \gamma_d}} = \sigma(X, \gamma).$$

Гradient і гессіан функції $g(X, *)$, які потрібні для розрахунку модифікацій відповідно:

$$\nabla g(X, \gamma) = \sigma(X, \gamma) (1 - \sigma(X, \gamma)) X,$$

$$Hg(X, \gamma) = \sigma(X, \gamma) (1 - \sigma(X, \gamma)) (1 - 2\sigma(X, \gamma)) X X^T$$

Розрахунки були проведені окремо для різних розподілів похибки $\varepsilon_i^{(m)}$: нормального, рівномірного, та розподілу Стюдента. Параметри розподілів та регресії наведені у таблиці 1.

	КОМПОНЕНТ	
	1	2
$\mu^{(m)}$	0.0	1.0
$\Sigma^{(m)}$	2.0	2.0
$b_0^{(m)}$	0.5	0.5
$b_1^{(m)}$	2	-1/3

Табл. 1: Таблиця параметрів моделювання

Експеримент 1 Значення параметрів для цього експерименту зазначено у таблиці (1). Похибки компонент ε^k мали один з таких розподілів: $\varepsilon_{1,i}^{(m)} \simeq N(0, 0.05^2)$; $\varepsilon_{2,i}^{(m)} \simeq U(-0.25, 0.25)$; $\varepsilon_{3,i}^{(m)} \simeq T_4/10$. Цей експеримент описується моделлю “повної роздільності”, коли візуально спостереження можуть бути розділені на дві групи відповідно до різних компонент

суміші. Приклад того, як виглядають модельовані дані, зображено на рисунку 1.

Як видно із таблиці 2, частоти включення справжніх значень параметрів до довірчих еліпсоїдів, що були отримані у результаті моделювання, прямує до теоретичного рівня $1 - \alpha = 0.95$ зі збільшенням обсягу вибірки. Цей результат є очікуваним.

Рис. 1: Приклад діаграми розкиду пар X проти Y на даних з експерименту 1.

	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
n	1	2	1	2	1	2
100	0.863	0.899	0.843	0.891	0.85	0.903
500	0.931	0.941	0.928	0.931	0.934	0.939
1 000	0.945	0.945	0.942	0.928	0.946	0.930
5 000	0.948	0.954	0.947	0.940	0.931	0.937
7 500	0.936	0.954	0.949	0.941	0.943	0.936
10 000	0.951	0.952	0.943	0.937	0.953	0.944

Табл. 2: Частоти для $\hat{V}_{n,1}$ у експерименті 1

	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
n	1	2	1	2	1	2
100	0.852	0.897	0.845	0.879	Nan	Nan
500	0.927	0.944	0.92	0.923	0.802	0.959
1 000	0.948	0.942	0.94	0.93	0.850	0.931
5 000	0.939	0.93	0.939	0.936	0.931	0.939
7 500	0.944	0.943	0.93	0.944	0.943	0.933
10 000	0.941	0.937	0.938	0.953	0.944	0.942

Табл. 3: Частоти для $\hat{V}_{n,2}$ у експерименті 1

Експеримент 2 Для того, щоб подивитися вплив розкиду похибок регресії на якість різних методів, ми збільшили параметр дисперсії залишків у другому експерименті. Експерименти проведено окремо для різних розподілів похибки. Таблиці 5, 6, 7 показують, що частоти включення у наведеному експерименті менші, ніж аналогічні частоти у першому експерименті. І збільшення обсягу вибірки не дуже допо-

магає виправити ситуацію. Це означає, що випадок із великими значеннями параметру дисперсії є складним для наведеної моделі.

$$\varepsilon_{1,i}^{(m)} \simeq N(0, 0.25^2); \varepsilon_{2,i}^{(m)} \simeq U(-1.25, 1.25); \varepsilon_{3,i}^{(m)} \simeq T_4/2.$$

Всі інші параметри було залишено як у першому експерименті.

n	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
	1	2	1	2	1	2
100	0.845	0.905	0.82	0.907	0.797	0.919
500	0.93	0.922	0.915	0.939	0.919	0.943
1 000	0.939	0.942	0.931	0.933	0.925	0.944
5 000	0.95	0.938	0.943	0.946	0.944	0.945
7 500	0.951	0.955	0.941	0.943	0.945	0.938
10 000	0.943	0.938	0.955	0.931	0.950	0.930

Табл. 4: Частоти для $\hat{V}_{n,3}$ у експерименті 1

Рис. 2: Приклад діаграми розкиду пар X проти Y на даних з експерименту 2.

Експеримент 3

Параметри розподілу похибок регресії ті ж самі, що і у експерименті 2. Усі інші параметри підібрані таким чином, щоб продемонструвати роботу моделі у випадку, коли залишки регресії об'єктів з інших компонент не врівноважуються навантаженнями $a_{i:n}^{(k)}$. Прикладом цього є випадок, коли всередині k -го компонента, на деякій області B зосереджено близько третини спосте-

режень, і їх відгуки концентруються біля значення 1. В той же час відгуки інших компонент зосереджені біля значення 0 і їхня кількість така, що ними не можна нехтувати. З огляду на таку ситуацію наближення $\hat{b}_n^{(k)}$ до $b^{(k)}$, означатиме збільшення значення функціоналу МНК і відповідно збільшення дисперсії відгуків.

У таблиці 8 наведені параметри розподілів і регресії цього експерименту.

З таблиць (9), (10), (11) видно, що частоти включення значно менші за теоретичне значення $1 - \alpha = 0.95$. Цей експеримент показує, що навіть при гарному виборі функції регресії

і відносно малих значеннях похибок, в деяких випадках побудовані оцінки параметрів регресії і оцінки їх матриць коваріацій значно відрізняються від своїх справжніх значень.

n	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
	1	2	1	2	1	2
100	0.779	0.907	0.502	0.919	0.542	0.898
500	0.914	0.921	0.795	0.930	0.803	0.939
1 000	0.928	0.931	0.873	0.932	0.855	0.912
5 000	0.928	0.930	0.905	0.928	0.923	0.927
7 500	0.938	0.938	0.924	0.919	0.925	0.928
10 000	0.95	0.915	0.906	0.931	0.915	0.917

Табл. 5: Частоти для $\hat{V}_{n,1}$ у експерименті 2

n	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
	1	2	1	2	1	2
100	0.704	0.837	0.508	0.873	0.515	0.894
500	0.883	0.930	0.819	0.923	0.812	0.918
1 000	0.955	0.952	0.867	0.922	0.852	0.915
5 000	0.927	0.917	0.911	0.911	0.909	0.921
7 500	0.931	0.939	0.911	0.920	0.908	0.930
10 000	0.917	0.922	0.932	0.929	0.910	0.911

Табл. 6: Частоти для $\hat{V}_{n,2}$ у експерименті 2

n	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
	1	2	1	2	1	2
100	0.782	0.909	0.687	0.913	0.528	0.842
500	0.905	0.919	0.819	0.933	0.853	0.936
1 000	0.927	0.921	0.864	0.930	0.835	0.942
5 000	0.928	0.942	0.916	0.923	0.926	0.919
7 500	0.946	0.938	0.914	0.923	0.907	0.912
10 000	0.938	0.930	0.914	0.918	0.906	0.933

Табл. 7: Частоти для $\hat{V}_{n,3}$ у експерименті 2

7 Висновки

Нами побудовано декілька модифікацій оцінки матриці коваріації параметрів регресії на основі техніки складаного ножа для моделі суміші зі змінними концентраціями. Якість побудованих оцінок порівнюється за допомогою експери-

ментів на модельованих даних. Результати експериментів показують, що надійності отриманих довірчих еліпсоїдів мало залежать від вибору варіанта оцінки коваріаційної матриці. Тому на практиці доцільно використовувати найбільш швидко з розглянутих оцінок — $\hat{V}_{n,3}$.

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. G.J. Mclachlan, D. Peel, Finite mixture models / G.J. Mclachlan, D. Peel // Wiley-Interscience (2000)
3. R.E. Maiboroda, Statistical analysis of mixtures // R.E. Maiboroda / Kyiv University Publishers, Kyiv (in Ukrainian) (2003)
4. R.E. Maiboroda, O.V. Sugakova, Statistics of mixtures with varying concentrations with application to DNA microarray data analysis // R.E. Maiboroda, O.V. Sugakova / Journal of nonparametric statistics. 24 , No 1 201–205 (2012)
5. R.E. Maiboroda, D. Liubashenko, Linear regression by observations from mixture wi-

	КОМПОНЕНТ	
	1	2
$\mu^{(m)}$	-2.5	2.5
$\Sigma^{(m)}$	2.0	2.0
$b_0^{(m)}$	0.5	-0.5
$b_1^{(m)}$	2	-2

Табл. 8: Таблиця параметрів моделювання

	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
n	1	2	1	2	1	2
100	0.311	0.258	0.197	0.187	0.170	0.170
500	0.609	0.490	0.466	0.397	0.487	0.412
1 000	0.717	0.618	0.597	0.496	0.604	0.514
5 000	0.907	0.840	0.835	0.724	0.831	0.727
7 500	0.929	0.881	0.864	0.766	0.862	0.776
10 000	0.928	0.885	0.872	0.782	0.886	0.799

Табл. 9: Частоти для $\hat{V}_{n,1}$ у експерименті 3

	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
n	1	2	1	2	1	2
100	0.27	0.223	0.172	0.178	0.189	0.159
500	0.599	0.5	0.292	0.266	0.454	0.402
1 000	0.755	0.646	0.609	0.512	0.602	0.492
5 000	0.758	0.643	0.829	0.745	0.849	0.724
7 500	0.927	0.871	0.888	0.798	0.851	0.721
10 000	0.932	0.916	0.875	0.831	0.867	0.778

Табл. 10: Частоти для $\hat{V}_{n,2}$ у експерименті 3

- th varying concentrations // R.E. Maiboroda, D. Liubashenko / Kyiv National Taras Shevchenko University, Kyiv, Ukraine
- R.E. Maiboroda, O.V. Sugakova, Jackknife covariance matrix estimation for observations from mixture // R.E. Maiboroda, O.V. Sugakova / Modern Stochastics: Theory and Applications, 2019
 - M. H. Quenouille, Notes on bias in estimation / M. H. Quenouille // Biometrika, 43, 353-60 (1956).
 - J. W. Tukey, Bias and confidence in not quite large samples / J. W. Tukey // The Annals of Mathematical Statistics. 29 (2): 614 (1958)
 - V.O. Miroshnychenko, Generalized least squares estimates for mixture of nonlinear regressions // V.O. Miroshnychenko / Bulletin of Taras Shevchenko National University of Kyiv; Series: Physics Mathematics, 2019, 5
 - R.E. Maiboroda V.O. Miroshnichenko "Confidence ellipsoids for regression coefficients by observations from a mixture"// R.E. Maiboroda V.O. Miroshnichenko / Modern Stochastics: Theory and Applications, Vol.5, Iss.2 pp. 225 - 245, - 2018
 - V.O. Miroshnichenko, R.E. Maiboroda, Asymptotic normality of modified LS estimator for mixture of nonlinear regressions

	$\varepsilon_{1,i}^{(m)}$		$\varepsilon_{2,i}^{(m)}$		$\varepsilon_{3,i}^{(m)}$	
n	1	2	1	2	1	2
100	0.150	0.173	0.228	0.224	0.444	.445
500	0.689	0.649	0.467	0.502	0.508	0.555
1 000	0.706	0.643	0.657	0.627	0.708	0.681
5 000	0.921	0.888	0.848	0.780	0.838	0.782
7 500	0.933	0.916	0.879	0.835	0.864	0.835
10 000	0.941	0.924	0.863	0.840	0.879	0.860

Табл. 11: Частоти для $\hat{V}_{n,3}$ у експерименті 3

// V.O. Miroshnichenko, R.E. Maiboroda /
Modern Stochastics: Theory and Applicati-
ons, Vol.7, Iss.7 pp. 435 - 448, 2020

12. J. Shao, Mathematical Statistics // J. Shao /
Springer, 2007

References

1. D. M. TITTETINGTON, A. F. SMITH, U. E. MAKOV (1985) Analysis of Finite Mixture Distributions. Wiley, New York
2. G.J. MCLACHLAN, D.PEEL (2000) Finite mixture models. Wiley-Interscience
3. R.E MAIBORODA (2003) Statistical analysis of mixtures. Kyiv University Publishers, Kyiv (in Ukrainian)
4. R.E. MAIBORODA, O.V. SUGAKOVA (2012) Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. Journal of nonparametric statistics. 24 , No 1 201–205 (2012)
5. R.E MAIBORODA, D. LIUBASHENKO (2015) Linear regression by observations from mixture with varying concentrations, Kyiv National Taras Shevchenko University, Kyiv, Ukraine
6. R.E. MAIBORODA, O.V. SUGAKOVA (2019) Jackknife covariance matrix estimation for observations from mixture, Modern Stochastics: Theory and Applications
7. M. H. QUENOUILLE (1956) Notes on bias in estimation. Biometrika, 43, 353-60 .
8. J. W. TUKEY (1958) Bias and confidence in not quite large samples. The Annals of Mathematical Statistics. 29 (2): 614
9. V.O. MIROSHNYCHENKO (2019) Generalized least squares estimates for mixture of nonlinear regressions, Bulletin of Taras Shevchenko National University of Kyiv; Series: Physics Mathematics, 2019, 5
10. V.O. MIROSHNYCHENKO, R.E. MAIBORODA (2020) Asymptotic normality of modified LS estimator for mixture of nonlinear regressions Modern Stochastics: Theory and Applications, Vol.7, Iss.7 pp. 435 - 448
11. J. Shao (2007) Mathematical Statistics, Springer

Received: 18.06.2020