УДК 519.21 https://doi.org/10.17721/1812-5409.2019/3.1

В.О. Мірошниченко[1], *асп.*

Аналіз залишків у моделі регресійної суміші

[1]Київський національний університет імені Тараса Шевченка, 01033, Київ, вул. Володимирська, 64.
vitaliy.miroshnychenko@gmail.com

V.O. Miroshnychenko[1], *PhD stud.*

Residual analysis in regression mixture model

[1]Taras Shevchenko National University of Kyiv, 01033, Kyiv, 64 Volodymyrska st.
vitaliy.miroshnychenko@gmail.com

Розглянуто оцінку коефіцієнтів нелінійної регресії за спостереженнями з суміші зі скінченою кількістю компонент. Концентрації компонент у суміші є різними для різних спостережень. Для побудови оцінок регресії використовується узагальнений метод найменших квадратів Побудовані оцінки використані для оцінки функцій розподілу і дисперсій залишків різних компонент. Оцінки перевірені симуляційним моделюванням і використані для аналізу соціологічних даних. Побудовані діаграми типу квантиль проти квантиля для візуального порівняння розподілу залишків.

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

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 and applied for sociological data analysis. Q-Q diagrams are built for visual comparison of residuals' distributions.

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

Communicated by Prof. Sugakova O.V.

1 Вступ

Моделі регресійних сумішей широко використовуються для опису прикладних статистичних даних. Як правило, на основі цих моделей будують оцінки невідомих параметрів [3], [6] і довірчі множини для них [9],[13]. Важливим елементом статистичного дослідження є також діагностика моделі, тобто перевірка її відповідності реальним даним. Для регресійних моделей однорідних спостережень така діагностика зазвичай спирається на аналіз залишків, що визначаються як відхилення спостережуваних значень відгуку від результатів прогнозування на основі моделі [4]. У випадку регресійної суміші дослідник працює не з одним прогнозом, а з набором кількох прогнозованих значень, що відповідають різним компонентам суміші. Тому модифікація стандартних технік аналізу залишків на випадок регресійних сумішей являє собою нетривіальну задачу.

У даній роботі для нелінійних моделей регресійних сумішей розглядаються дві задачі аналізу залишків: оцінювання дисперсій похибок і візуальна перевірка припущень про розподіл похибок за допомогою модифікованої діаграми квантиль-квантиль (QQ-діаграма). Для розв'язання цих задач використовуються загальні методи статистики сумішей [1], [2] [7].

Опис моделі нелінійної регресійної суміші наведено у розділі 1. У розділі 2 описані мінімаксні коефіцієнти, які використовуються у розділі 3 для отримання оцінок коефіцієнтів регресії. Використання параметрів регресії для аналізу залишків описано у розділі 4. У розділі 5 доведена консистентність оцінок дисперсій похибок регресії для різних компонентів суміші. Поведінки оцінок для вибірок фіксованого обсягу досліджено на модельованих даних і застосовано для аналізу соціологічних даних у розділі 6 і розділі 7.

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 $\kappa(O) \in 1, ..., M$. 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.

Таким чином

$$\xi(O) = (Y(O), X^1(O), ..., X^d(O))^T,$$

де $Y(O)$ – відгук, $X(O) = (X^1(O), ..., X^d(O))$ – регресори у моделі.

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

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

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

та

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

(σ_m^2 невідомі).

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

Вектори регресорів

$$X(O) = (X^1(O), ..., X^d(O))$$

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

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

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

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

$$p_i^m = \mathrm{P}(\kappa(O_i) = m), m = \overline{1, M}; i = \overline{1, n}.$$

(ці ймовірності змішування відомі для кожного об'єкту)

3 Мінімаксні емпіричні навантаження

Нехай далі k – фіксований номер компонента, для якого оцінюються параметри $b^{(k)}$. Для оцінки функцій розподілу і $F_{X,k}$ можна використати навантажену емпіричну функцію розподілу, як це зроблено у [8]:

$$\hat{F}_{X,n}^{(k)}(x) = \sum_{i=1}^n a_{i:n}^{(k)} I[X_i < x]$$

Мінімаксні ваги $a_{i:n}^{(k)}$, які визначені [10] і які мають вигляд

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

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

4 Оцінки параметрів, що отримані за допомогою оціночних рівнянь

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

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

Оцінки параметрів регресії є розв'язком оціночного рівняння, яке отримується шляхом диференціювання наведеного функціоналу

$G_n^k(b^{(k)})$:

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

$$= \sum_{j=1}^{n}(-2)a_{j:n}^{(k)} s(\xi_i,\gamma) = 0, \qquad (1)$$

де

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

Вектор оцінок буде стаціонарною точкою функціоналу 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 $J_n^k(\gamma) =^{def} \nabla G_n^k(\gamma)$.

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

Введемо наступне позначення:

$$\Gamma_\infty = \lim_{n\to\infty} n^{-1}\Gamma_n = \lim_{n\to\infty} n^{-1}(\sum_{j=1}^{n} p_j^i p_j^k)_{i,k=1}^{M}$$

Далі $X_\cdot^{(m)}, Y_\cdot^{(m)}, \varepsilon_\cdot^{(m)}$ – випадкові величини із розподілами:

$$X_\cdot^{(m)} \sim F_{X,m},\ \varepsilon_\cdot^{(m)} \sim F_{\varepsilon,m} \text{ – незалежні,}$$

та

$$Y_\cdot^{(m)} = g(X_\cdot^{(m)}, b^{(m)}) + \varepsilon_\cdot^{(m)}.$$

5 Застосування до діаграми квантиль проти квантиля і оцінки дисперсії залишків

У розділі 3 були побудовані оцінки $\hat{b}_n^{(k)}, k = 1,...,M$. Їх можна використати для оцінки функції розподілу похибок різних компонент. Для цього розглянемо залишки, що відповідають k-тому компоненту: $u_i = Y_i - g(X_i, \hat{b}_n^{(k)})$. Оскільки у суміші присутні об'єкти, що належать різним компонентам, для оцінювання функції розподілу похибок, що відповідає k-тій компоненті, використаємо навантажену ф.р. залишків з мінімаксними коефіцієнтами: $a_{i:n}^{(k)}$:

$$\hat{F}_{\varepsilon,k}(u) = \sum_{i=1}^{n} a_{i:n}^{(k)} I[u_i < u] \qquad (2)$$

Оцінка (2) може бути використана для побудови діаграми типу квантиль проти квантиля і для оцінки моментів залишків, в тому числі і дисперсії.

Для візуальної перевірки гіпотез про розподіл похибок за однорідними даними використовують діаграми квантиль проти квантиля (QQ діаграма).

Щоб побудувати аналогічну діаграму для залишків у моделі суміші, розглянемо оцінки квантилів на основі навантажених ф.р., запропоновані у [6].

$$Q^{\hat{F}_{\varepsilon,k}}(\alpha) = \frac{1}{2}(Q_+^{(k)}(\alpha) + Q_-^{(k)}(\alpha)),$$

тут

$$Q_+^{(k)}(\alpha) = \sup(x \in \mathbb{R} : \hat{F}_{\varepsilon,k}(x) \leqslant \alpha),$$
$$Q_-^{(k)}(\alpha) = \inf(x \in \mathbb{R} : \hat{F}_{\varepsilon,k}(x) \geqslant \alpha).$$

Якщо потрібно перевірити, що ф.р. залишків належить сім'ї розподілів $F_{\varepsilon,k}$, задаємо набір рівнів $\alpha_1,...,\alpha_N \in [0,1]$. На діаграмі відображаються точки з координатами $Q^{\hat{F}_{\varepsilon,k}}(\alpha_i)$ по горизонталі, і $Q^{F_{\varepsilon,k}}(\alpha_i)$ по вертикалі, $i \in 1,...,N$. Тут $Q^{F_{\varepsilon,k}}(\alpha)$ - квантиль розподілу $F_{\varepsilon,k}$ із рівнем α. Додатково на діаграмі проводиться пряма зі зміщенням 0 і коефіцієнтом нахилу 1.

На основі (2) можна також побудувати оцінку для дисперсії похибки регресії k-того компонента суміші.

$$\hat{\sigma}_k^2(\gamma) = \sum_{i=1}^{n} a_{i:n}^{(k)}[Y_i - g(X_i,\gamma)]^2.$$

Тоді статистика $\hat{\sigma}_k^2(\hat{b}_n^{(k)})$ - оцінка дисперсії залишків k-го компонента.

6 Консистентність оцінки дисперсії залишків

Для доведення консистентності $\hat{\sigma}_k^2(\hat{b}_n^{(k)})$ буде використане наступне твердження з [9]:

Твердження 1

Припустимо що $\det\Gamma_\infty > 0$. Тоді

1 . $\sup_{j=1,\bar{n},k=1,\bar{M}} |a_{j:n}^{(k)}| = O(n^{-1})$

2 . $\sup_{i,j=1,\bar{n},k=1,\bar{M},i\neq j} |a_{j:n}^{(k)} - a_{j:i-}^{(k)}| = O(n^{-2})$

Твердження 1 з [12] дає приклади умов консистентності оцінок $\hat{b}_n^{(k)}$. Наступна теорема доводить консистенту збіжність оцінок $\hat{\sigma}_k^2(\hat{b}_n^{(k)})$ до параметрів σ_k^2.

Теорема

Нехай $\hat{b}_n^{(k)} \to^P b^{(k)}$, і $\forall\gamma \in \Theta\ \exists\mathbb{E}[\varepsilon_\cdot^{(k)}]^4 < \infty$ та $\exists\mathbb{E}g(X_\cdot^{(k)},\gamma)^4 < \infty$. Тоді $\forall\alpha > 0$ має місце місце відношення:

$$\hat{\sigma}_k^2(\hat{b}_n^{(k)}) - \sigma_k^2 -$$

$$\mathbb{E}_k[g(X_\cdot^{(k)},\gamma) - g(X_\cdot^{(k)},b^{(k)})]^2|_{\gamma=\hat{b}_n^{(k)}} = o_P(n^{-1/2+\alpha}) \tag{3}$$

Якщо додатково припустити рівноступеневу неперервність $g(x,\gamma)$ у деякому околі точки $\gamma = b_k$ і неперервність по x, то:

$$\hat{\sigma}_k^2(\hat{b}_n^{(k)}) - \sigma_k^2 = o_P(1) \tag{4}$$

Доведення. Для доведення буде використано наступне твердження з [11].

Твердження 2

Нехай ξ_n - послідовність випадкових величин, для якої існує перший момент $\mathbb{E}\xi_n$ і додатна дисперсія $\mathbb{D}\,\xi_n$. Тоді:

$$\xi_n - \mathbb{E}\xi_n = O_P([\mathbb{D}\,\xi_n]^{1/2}).$$

Щоб довести рівність (3) знайдемо математичне сподівання $\mathbb{E}\hat{\sigma}_k^2(\gamma)$, оцінимо дисперсію $\mathbb{D}\,\hat{\sigma}_k^2(\gamma)$, і застосуємо Твердження 2.

Запишемо перший момент оцінки дисперсії похибок в іншому вигляді:

$$\mathbb{E}\hat{\sigma}_k^2(\gamma) = \sum_{i=1}^n a_{i:n}^{(k)}\mathbb{E}[Y_i - g(X_i,\gamma)]^2 =$$

$$\sum_{i=1}^n a_{i:n}^{(k)}\sum_{t=1}^M p_i^t\mathbb{E}_t[Y_\cdot^{(t)} - g(X_\cdot^{(k)},\gamma)]^2 =$$

$$\sum_{t=1}^M \mathbb{E}_t[Y_\cdot^{(t)} - g(X_\cdot^{(k)},\gamma)]^2\sum_{i=1}^n a_{i:n}^{(k)}p_i^t =$$

$$\mathbb{E}_k[Y_\cdot^{(k)} - g(X_\cdot^{(k)},\gamma)]^2 \tag{5}$$

Остання рівність випливає з із властивості незміщеності коефіцієнтів $a_{i:n}^{(k)}$:

$\sum_{i=1}^n a_{i:n}^{(k)}p_i^t = I[k=t]$,

що детальніше описана у [10]. З виразу (5) можна виокремити параметр дисперсії залишків σ_k^2, якщо згадати про центрованість похибки $\varepsilon_\cdot^{(k)}$ і, що $Y_\cdot^{(k)} = g(X_\cdot^{(k)},b^{(k)}) + \varepsilon_\cdot^{(k)}$. Тоді:

$$\mathbb{E}\hat{\sigma}_k^2(\gamma) = \mathbb{E}_k[Y_\cdot^{(k)} - g(X_\cdot^{(k)},\gamma)]^2 =$$

$$\sigma_k^2 + \mathbb{E}_k[g(X_\cdot^{(k)},b^{(k)}) - g(X_\cdot^{(k)},\gamma)]^2$$

Далі оцінимо дисперсію статистики $\hat{\sigma}_k^2(\gamma)$. Якщо винести суму і навантаження зі знаку дисперсії, то

$\mathbb{D}\,\hat{\sigma}_k^2(\gamma) = \sum_i^n |a_{i:n}^{(k)}|^2\,\mathbb{D}[Y_i - g(X_i,\gamma)]^2$.

За припущенням теореми усі дисперсії існують і скінченні. Зауважимо, що використання Твердження 1 для мінімаксних коефіцієнтів дає відношення $\sum_i^n |a_{i:n}^{(k)}|^2 = O(n^{-1})$. Це означатиме, що $\mathbb{D}\,\hat{\sigma}_k^2(\gamma) = O(n^{-1}) = o(n^{2\alpha-1}), \forall\alpha > 0$.

У результаті застосування Твердження 2, маємо $\forall\gamma\in\Theta, \forall\alpha > 0$:

$$\hat{\sigma_k}^2(\gamma) - \sigma_k^2 - \mathbb{E}_k[g(X_\cdot^{(k)},\gamma) - g(X_\cdot^{(k)},b^{(k)})]^2 = o_P(n^{\alpha-1/2})),$$

і це доводить відношення (3) при підставленні $\gamma = \hat{b}_n^{(k)}$.

Доведемо відношення (4). Для цього покажемо, що $R(\gamma) = \mathbb{E}_k[g(X_\cdot^{(k)},\gamma) - g(X_\cdot^{(k)},b^{(k)})]^2$ неперервна у $b^{(k)}$. З рівноступеневої неперервності $g(x,\gamma)$ у точці $\gamma = b_k$:

Для довільного $\lambda > 0$, існує $\delta > 0$, для всіх $x \in \mathbb{R}^d$ і для всіх γ із околу $b^{(k)}$ таких, що $|\gamma - b^{(k)}| < \delta$, випливає:

$$|g(x,\gamma) - g(x,b^{(k)})| < \lambda \tag{6}$$

Отже, з (6) отримуємо

$$|R(\gamma)| = |\mathbb{E}_k[g(X_\cdot^{(k)},\gamma) - g(X_\cdot^{(k)},b^{(k)})]^2| =$$

$$|\int_{R^d}[g(x,\gamma) - g(x,b^{(k)})]^2 dF_k(x)| < \varepsilon^2\int_{R^d} dF_k(x) = \varepsilon^2$$

Тобто $R(\gamma)$ є неперервною у точці $b^{(k)}$. За припущенням консистентності оцінок $\hat{b}_n^{(k)} \to^P b^{(k)}$, $R(\hat{b}_n^{(k)}) \to^P R(b^{(k)}) = 0$. Тому з (3) випливає (4). □

7 Моделювання

Було проведено серію із $H = 1000$ експериментів для різних розмірів вибірки. Цими експериментами перевірялася консистентність оцінки $\hat{\sigma}_k^2(\hat{b}_n^{(k)})$ у рівності (4) та її уточнення у рівності (3). Для цього було пораховано середнє абсолютне зміщення справжніх параметрів від їх оцінок у (4) і окремо з уточненням (3). Додатково рахувалася і дисперсія отриманих оцінок для рівності (4).

Номера компонент генерованих об'єктів $\kappa_i = \kappa(O_i), i = 1, ..., n$ обиралися відповідно до векторів розподілу $(p_i^1, ..., p_i^M)_{i=1}^n$. Тут p_i^k - ймовірності, що згенеровані наступною процедурою:

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

Для моделювання регресії використовувалась логістична модель.

$$g(X, \gamma) = \frac{1}{1+e^{-\gamma_0 - \gamma_1 X_1}}$$

. Характеристики, які спостерігаються для m-го компонента суміші $\xi_i^{(m)} = (Y_i^{(m)}, X_i^{(m)})$ генерувалися наступним чином:

$$Y_i^m = g(X_i^m, b^{(m)}) + \varepsilon_i^{(m)}$$

$$X_i^{(m)} \simeq N(\mu^{(m)}, \Sigma^{(m)}).$$

$$\varepsilon_i^{(m)} \simeq N(0, \sigma_k^2)$$

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
σ_k^2	0.05	0.05

Таблиця 1

Результати моделювання для рівності (3) при $\alpha = 0.25$ і рівності (3) наведено у таблиці 2.

Дисперсії оцінок $\hat{\sigma}_k^2(\hat{b}_n^{(k)})$ наведені у таблиці 3.

	рівність (3), $\alpha = 0.25$		рівність (4)	
n\Компонент	1	2	1	2
100	0.039	0.050	0.01615	0.0099
500	0.0268	0.0191	0.0058	0.0044
1000	0.0145	0.0152	0.0027	0.0028
5000	0.0090	0.0101	0.00099	0.00122
7500	0.0077	0.0097	0.00092	0.00091
10000	0.0073	0.0089	0.00068	0.00076

Таблиця 2. Зміщення

n\Компонент	1	2
100	0.0048	0.00021
500	1.88e-3	4.31e-5
1000	3.005e-5	2.17e-5
5000	3.61e-6	5.03e-6
7500	2.89e-6	3.02e-6
10000	1.68e-6	2.15e-6

Таблиця 3. Дисперсії оцінок

8 Застосування до соціологічних даних

Для демонстрації діаграми квантиль проти квантиля і оцінки дисперсії залишків було обрано соціологічні дані. У нашому прикладі це - результати зовнішнього незалежного оцінювання (ЗНО) за 2016 рік (див [13]), яке складають абітурієнти у виші. Спостережувані характеристики: результати іспитів із математики (регресор) та української мови і літератури (відгук). Абітурієнтів, які подолали мінімальний поріг усього 94 тисячі особи. Ми розглянули випадок, коли функція залежності - логістична:

$$g(X,\gamma) = \frac{1}{1+e^{-\gamma_0-\gamma_1 X_1}}$$

Оскільки результати іспитів - це числа в інтервалі від 100 до 200, то вони були переведені в інтервал від 0 до 1 відніманням і діленням на 100.

Додатково для кожного абітурієнта спостерігається область, у якій він складав іспит. Це знання необхідне для побудови концентрацій компонент. Кожна з компонент у суміші - це українські політичні течії 2014 року, що об'єднані у 3 групи:

1 Коаліція (проукраїнські): БПП, Народний Фронт, Батьківщина, Радикальна партія, Самопоміч.

2 Опозиція (контрукраїнські): Опозиційний блок, маленькі партії, проти всіх.

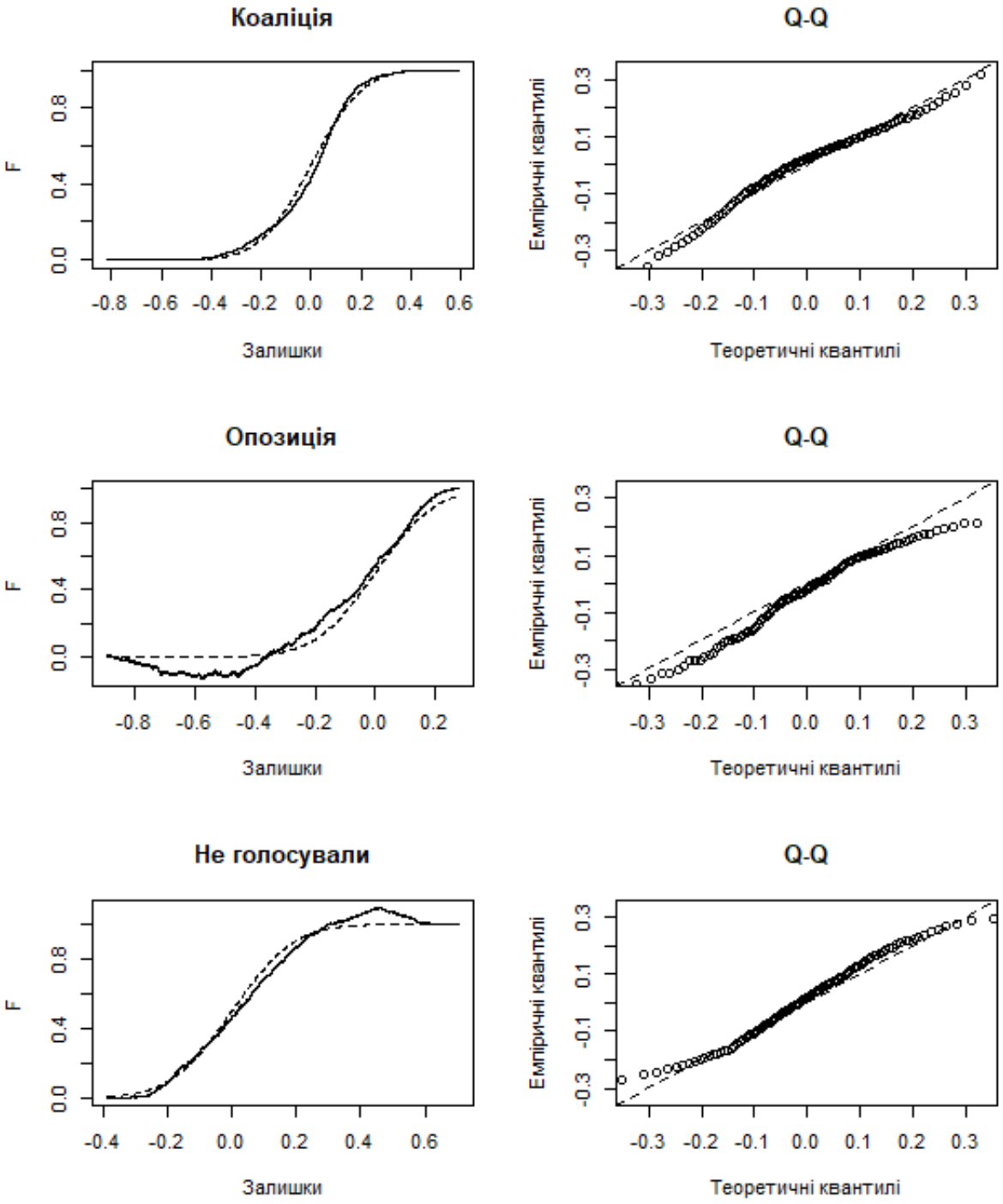
3 Не зацікавлені у політиці, - громадяни, що не проголосували.

Частку отриманих голосів під час парламентських виборів 2014 року по різних областях вважатимемо концентраціями компонент. Далі кожному абітурієнту поставили у відповідність концентрації течій по назві області в якій вони складали іспити.

Наведеної інформації достатньо, щоб побудувати мінімаксні навантаження, і оцінити параметри регресі. Оцінки параметрів регресії і оцінка дисперсії залишків наведені у Таблиці 3. За цими параметрами та мінімаксними навантаженнями були оцінені функції розподілу залишків і побудовані діаграми квантиль проти квантиля у припущенні що апріорний розподіл залишків має гаусів розподіл.

Графіки першої колонки - порівняння оцінки функції розподілу компонент (суцільна лінія) і нормального розподілу (пунктиром) із відповідною дисперсією σ_k^2. У другій колонці зображені діаграми квантиль проти квантиля для оцінок розподілів і нормального розподілу.

Як видно з діаграми, квантилі гаусової функції розподілу з оціненою дисперсією схожі на квантилі оцінених розподілів. Це каже нам про те, що при припущеній логістичній залежності, залишки можеть бути розподілені нормально. Для впевненості слід розробити статистичний тест перевірки гіпотез про рівність розподілів.



9 Висновки

У результаті були побудовані оцінки розподілу залишків у регресійній суміші. Це наближення було використано для побудови діаграм типу квантиль проти квантиля і оцінки дисперсії залишків. Для оцінки дисперсії залишків доведено теорему про консистентність і результати перевірені імітаційним моделюванням. Його результати показують що якість оцінок для вибірок понад 5 000 спостережень не дуже сильно відрізняються від результатів на 5 000. Знайдені оцінки застосовані для аналізу даних ЗНО та виборів 2014 року. Побудовані діаграми квантиль проти квантиля показують, розподіли залишків для логістичної функції досить близькі до нормальних розподілів. Точний результат можуть дати статистичні тести.

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. Grun Bettina and Leisch Friedrich: Fitting finite mixtures of linear regression models with varying & fixed effects in R. / Grun Bettina, Leisch Friedrich // In Alfredo Rizzi and Maurizio Vichi, editors, Compstat 2006 - Proceedings in Computational Statistics, pages 853-860. Physica Verlag, Heidelberg, Germany, 2006.
4. G.A.F Seber, A.J. Lee, Linear Regression Analysys / G.A.F Seber, A.J. Lee //. Wiley (2003)
5. R.E. Maiboroda, Statistical analysis of mixtures // R.E. Maiboroda / Kyiv University Publishers, Kyiv (in Ukrainian) (2003)
6. R.E. Maiboroda, D. Liubashenko, Linear regression by observations from mixture with varying concentrations // R.E. Maiboroda, D. Liubashenko / Kyiv National Taras Shevchenko University, Kyiv, Ukraine
7. R.E. Maiboroda, O.V. Sugakova, Esitmation and classification by observations from a mixture // R.E. Maiboroda, O.V. Sugakova / Kyiv University Publishers, Kyiv, 2008. (In Ukrainian)
8. Р.Є. Майборода, О.В. Сугакова, "Тести для гіпотез про квантилі розподілів компонентів суміші"// Р.Є. Майборода, О.В. Сугакова / Теор. ймов. та мат. статист., Vol.101, Iss. pp. 157 - 168, - 2019
9. 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
10. 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)
11. Y.M. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis Theory and Practice // Y.M. Bishop, S.E. Fienberg, P.W. Holland / Springer, 2007
12. 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
13. 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

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. B. GRUNAND F.LEISCH (2006) Fitting finite mixtures of linear regression models with varying & fixed effects in R. In Alfredo Rizzi and Maurizio Vichi, editors, Compstat 2006 - Proceedings in Computational Statistics, pages 853-860. Physica Verlag, Heidelberg, Germany, 2006
4. G.A.F. SEBER, A.J.LEE (2003) Linear Regression Analysys. Wiley
5. R.E MAIBORODA (2003) Statistical analysis of mixtures. Kyiv University Publishers, Kyiv (in Ukrainian)
6. R.E MAIBORODA, D. LIUBASHENKO (2015) Linear regression by observations from mixture with varying concentrations, Kyiv National Taras Shevchenko University, Kyiv, Ukraine
7. R.E. MAIBORODA, O.V. SUGAKOVA (2008) Esitmation and classification by

observations from a mixture, Kyiv University Publishers, Kyiv, 2008. (In Ukrainian)

8. R.E. MAIBORODA, O.V. SUGAKOVA (2019) "Тести для гіпотез про квантилі розподілів компонентів суміші". Теор. ймов. та мат. статист., Vol.101, Iss. pp. 157 - 168, - 2019

9. R.E. MAIBORODA, O.V. SUGAKOVA (2019) Jackknife covariance matrix estimation for observations from mixture, Modern Stochastics: Theory and Applications, 2019

10. 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)

11. Y.M. BISHOP, S.E. FIENBERG, P.W. HOLLAND (2007) Discrete Multivariate Analysis Theory and Practice, Springer,

12. 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

13. R.E. MAIBORODA, V.O. MIROSHNYCHENKO (2018) "Confidence ellipsoids for regression coefficients by observations from a mixture". Modern Stochastics: Theory and Applications, Vol.5, Iss.2 pp. 225 - 245, - 2018

Received: 18.06.2019