UDC 519.22
DOI: https://doi.org/10.17721/1812-5409.2024/1.8

Rostyslav MAIBORODA, DSc (Phys.&Math.), Prof.
ORCID ID: 0000-0002-3899-558X
e-mail: rostmaiboroda@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Vitaliy MIROSHNYCHENKO, PhD, (Phys.&Math.), Assist. Prof.
ORCID ID: 0009-0003-9799-8902
e-mail: vitaliy.miroshnychenko@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Olena SUGAKOVA, DSc (Phys.&Math.), Assoc. Prof.
ORCID ID: 0000-0002-6529-0788
e-mail: olenasuhakova@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

QUANTILE ESTIMATORS FOR REGRESSION ERRORS IN MIXTURE MODELS WITH VARYING CONCENTRATIONS

We consider the data described by the mixture model with variable concentrations. Models of mixtures of several components naturally arise in the analysis of statistical data of medical, biological and sociological studies. In these models the observed objects are selected from a finite set of subpopulations (components of the mixture) that have different distributions of the observed characteristics. The concentrations of the components in the mixture (mixing probabilities) may vary from observation to observation. In the considered data, the dependence between variables is described by a nonlinear regression model. The coefficients of this model are different for different components. The distributions of the regression errors can also be different. Analysis of these distributions is of significant interest in regression analysis. Quantile-quantile (QQ) plots are a tools of visually comparing error distributions with known distributions. Such charts are useful in the model diagnostics. For homogeneous data, QQ-diagrams can be easily constructed by the regression residuals. In a mixture model, in the presence of several regression models, the residuals cannot be determined unambiguously. Therefore, there is a need to construct estimators for the quantiles of the distribution of residuals. To obtain estimators of quantiles, one can use estimators of the functions of the corresponding distributions. Such estimators in models of mixtures with variable concentrations are constructed using weighted empirical distribution functions (EDF). For independent observations, the asymptotic behavior of the weighted EDFs is well studied. However, in the case of regression error analysis, the researcher is dealing with residuals that are dependent across observations. Therefore, the task of the asymptotic study of weighted EDFs and the corresponding estimators of quantiles in this case is relevant.

The paper considers the construction of weighted EDFs and quantile estimates for regression errors based on observations from the mixture. Consistency of these estimators is demonstrated. To show the consistency results we applied the asymptotic theory of weighted EDFs. Consistency of the estimators for distribution functions and quantiles of errors distributions is de-monstrated. The results can be applied to construction of QQ-plots and model diagnostics.

K e y w o r d s: ***mixture with varying concentrations, nonlinear regression, residuals, quantiles.***

AMS 2020 classification: 62G30, 62J02.

Introduction

The quantile-quantile (QQ) plot is a useful tool for analysis of random variables distribution, see Section 10.2 in (Gilchrist, 2000) . Standard technique of regression models residuals analysis includes QQ-plots drawling for residuals to analyze departure from normality, see Section 6.4.1 in (Groß, 2003). If the observed sample is homogeneous it is not difficult to define empirical quantile as order statistics of the residuals. But it is a much more challenging problem when the observations are drawn from a mixture of different components with different dependencies between observed variables.

In this paper we consider the model of mixture with varying concentrations (MVC) at which the mixing probabilities (concentrations of the mixture components) are different for different observations. You can see (Maiboroda, & Sugakova, 2008) for general theory of MVC and (Pidnebesna, Fajnerová, Horáček, & Hlinka, 2023) on their applications to neurobiological data analysis. Parametric models of regression mixtures with varying concentrations were considered in (Grün, & Leisch, 2006). Nonparametric estimation for MVC regression models based on the modified least squares estimation is discussed in (Maiboroda, & Miroshnychenko, 2020; Maiboroda, Miroshnychenko, & Sugakova, 2022).

QQ-plots for observations from mixture of regressions with varying concentrations were introduced in (Miroshnychenko, 2019). They are based on an estimator for regression errors quantiles which allows to perform estimation for each mixture component separately. In (Miroshnychenko, 2019) performance of the proposed estimator was assessed by simulations.

The object of research is the estimator of regression errors distribution quantiles by data obtained from mixture with varying concentrations.

The aim of the research is to derive conditions of consistency of this estimator.

To derive the estimator for quantiles we, at first, consider the weighted empirical distribution function (WEDF) based on regression residuals. Consistency of WEDF as estimator of corresponding error distribution is demonstrated by the Vapnik – Chervonenkis inequality. The estimator of quantiles is a transform of WEDF, so the WEDF consistency implies consistency of the quantile estimator.

1. Results

Let us consider a statistical data set at which observed subjects O_1, …, O_n can belong to one of M different subpopulations (mixture components). The true number κ_j of the population which O_j belongs to is unknown, but we know the probabilities

$$p^k_{j;n} = \mathsf{P}\{\kappa_j = k\},$$

which are called the mixing probabilities or concentrations of the k-th component in the mixture.

For each subject O_j we observe a set of real valued variables $\xi_j = (Y_j, X_j^1, \dots, X_j^d)$, where Y is the response and X^i, $i = 1, \dots, d$ are regressors of the regression model

$$Y_j = g(\mathbf{X}_j; \boldsymbol{\vartheta}^{(\kappa_j)}) + \varepsilon_j,\ j = 1, \dots, n, \tag{1}$$

where $g : \mathbb{R}^d \times \Theta \to \mathbb{R}$ is a known regression function, $\mathbf{X}_j = (X_j^1, \dots, X_j^d)^T$, $\boldsymbol{\vartheta}^{(k)} \in \Theta \in \mathbb{R}^m$ is a vector of unknown regression parameters for the k-th mixture component, ε_j is an error term.

In what follows we assume that ξ_j are independent random vectors and that $\mathbf{X}_j$ and ε_j are conditionally independent for fixed $\kappa_j = k$, $k = 1, \dots, M$. Let

$$F_\varepsilon^{(k)}(x) = \mathsf{P}\{\varepsilon_j < x \mid \kappa_j = k\}$$

be the cumulative distribution function (CDF) of the error term for the k-th component. We assume that

$$\mathsf{E}[\varepsilon_j | \ \kappa_j = k] = \int x F_\varepsilon^{(k)}(dx) = 0.$$

Functions $F_\varepsilon^{(k)}(x)$ are unknown, our aim is to estimate them and corresponding quantiles by the observations ξ_j, $j = \overline{1, n}$.

We will also denote the CDF of $\mathbf{X}_j$ for the kth component by $F_{\mathbf{X}}^{(k)}(\mathbf{x}) = \mathsf{P}\{\mathbf{X}_j < \mathbf{x} \mid \kappa_j = k\}$. Note that $F_{\mathbf{X}}^{(k)}$ and $F_\varepsilon^{(k)}$, $k = 1, \dots, K$ does not depend on j.

The model represented by (1) is a mixture of regressions with varying concentrations. It is a partial case of the model of mixtures vith varying concentrations (Maiboroda, & Sugakova, 2008).

The usual way to estimate CDFs of components in MVC models is to use weighted empirical CDFs with the minimax weights $a_{j;n}^k$. These weights are defined as

$$a_{j;n}^k = \frac{1}{\det \boldsymbol{\Gamma}_n} \sum_{m=1}^{M} (-1)^{(m+k)} \gamma_{km;n} p_{j;n}^m, \tag{2}$$

where

$$\boldsymbol{\Gamma}_n = \left(\sum_{j=1}^{n} p_{j;n}^i p_{j;n}^l \right)_{i,l=1}^{M},$$

$\gamma_{km;n}$ is the (k, m)th minor of $\boldsymbol{\Gamma}_n$. Of course, we assume here that $\det \boldsymbol{\Gamma}_n \neq 0$.

Minimax properties of CDF estimators with weights $a_{j;n}^k$ are discussed in section 2.1 of (Maiboroda, & Sugakova, 2008).

To construct estimators for $F_\varepsilon^{(k)}$ we need first to estimate unknown parameters $\boldsymbol{\vartheta}^{(k)}$ in (1). In (Maiboroda, & Miroshnychenko, 2020) a modified least squares estimator (mLSE) is considered for this purpose. The mLSE $\hat{\boldsymbol{\vartheta}}_n^{(k)}$ for $\boldsymbol{\vartheta}^{(k)}$ is defined as a stationary point of the weighted LS functional

$$J_n^k(\boldsymbol{\gamma}) = \sum_{i=1}^{n} a_{j;n}^k (Y_j - g(\mathbf{X}_j, \boldsymbol{\gamma}))^2,$$

where $\boldsymbol{\gamma} = (\gamma^1, \dots, \gamma^m) \in \Theta$. I.e. $\hat{\boldsymbol{\vartheta}}_n^{(k)}$ is a solution to the system of equations

$$\frac{\partial}{\partial \gamma^i} J_n^k(\boldsymbol{\gamma}) = 0,\ i = 1, \dots, m.$$

It is shown in (Maiboroda, 2020) that under suitable assumptions mLSE $\hat{\boldsymbol{\vartheta}}_n^{(k)}$ is a $\sqrt{n}$ consistent estimator to $\boldsymbol{\vartheta}^{(k)}$, i.e. the sequence $\sqrt{n}(\hat{\boldsymbol{\vartheta}}_n^{(k)} - \boldsymbol{\vartheta}^{(k)})$, $n = 1, 2, \dots$ is stochastically bounded.

In what follows $\hat{\boldsymbol{\vartheta}}_n^{(k)}$ can be any $\sqrt{n}$-consistent estimator to $\boldsymbol{\vartheta}^{(k)}$, not necessarily the mLSE.

With $\hat{\boldsymbol{\vartheta}}_n^{(k)}$ at hand we can define residuals for the kth component regression model

$$U_j^{(k)} = Y_j - g(\mathbf{X}_j; \hat{\boldsymbol{\vartheta}}_n^{(k)}). \tag{3}$$

Due to consistency of $\hat{\boldsymbol{\vartheta}}_n^{(k)}$, for large n, $U_j^{(k)}$ is near to ε_j if $\kappa_j = k$. But if the jth subject does not belong to the kth component, $U_j^{(k)}$ can be very far from ε_j. To suppress the influence of wrong components we use the mimimax weightings.

So our estimate to $F_\varepsilon^{(k)}(x)$ is

$$\tilde{F}_n^{(k)}(x) = \sum_{j=1}^{n} a_{j;n}^k \mathbb{1}\{U_j^{(k)} < x\}, \tag{4}$$

where $\mathbb{1}\{A\}$ is the indicator function of an event A.

The quantile of level α $(0 < \alpha < 1)$ for the CDF $F_\varepsilon^{(k)}$ is defined as

$$Q_\varepsilon^{(k)}(\alpha) = \inf\{x \in \mathbb{R}\ :\ F_\varepsilon^{(k)}(x) > \alpha\}. \tag{5}$$

To estimate $Q_\varepsilon^{(k)}(\alpha)$ we replace the true CDF by its estimator:

$$\hat{Q}_n^{(k)}(\alpha) = \inf\{x \in \mathbb{R} \ : \ \tilde{F}_\varepsilon^{(k)}(x) > \alpha\}. \tag{6}$$

We will start formulation of main theorems introducing some notations and assumptions. Let

$$\mathbf{\Gamma}_\infty = \lim_{n\to\infty} \frac{1}{n}\mathbf{\Gamma}_n,$$

assuming that this limit exists.

For a function $h : \mathbb{R}^d \to [0, +\infty)$ and $\alpha > 0$ we will say that the **Assumption G**$_{h,\alpha}$ holds if for all $m = 1, \dots, M$

$$\int (h(\mathbf{x}))^\alpha F_{\mathbf{X}}^{(m)}(\mathbf{x}) < \infty \tag{7}$$

and

$$|g(\mathbf{x}; \gamma_1) - g(\mathbf{x}; \gamma_2)| \leqslant h(\mathbf{x})|\gamma_1 - \gamma_2|$$

for all $\mathbf{x} \in \mathbb{R}^d$ and all $\gamma_1, \gamma_2 \in \Theta$.

Theorem 1. *Assume that the following conditions hold.*
(1) det $\mathbf{\Gamma}_\infty > 0$.
(2) Assumption **G**$_{h,\alpha}$ *holds with some function h and $\alpha > 2$.*
(3) There exist probability densities $f_\varepsilon^m(x)$ for $F_\varepsilon^{(m)}$ and $f_\varepsilon^m(x)$ are bounded on $\mathbb{R}$, for $m = 1, \dots, M$.
(4) $\hat{\boldsymbol{\vartheta}}_n^{(m)}$, $m = 1, \dots, M$ are $\sqrt{n}$-consistent.
Then

$$\sup_{x\in\mathbb{R}} |\tilde{F}_n^{(k)}(x) - F_\varepsilon^{(k)}(x)| \xrightarrow{\mathsf{P}} 0 \text{ as } n \to \infty$$

for all $k = 1, \dots, M$.

Theorem 2. *Assume that the conditions of Theorem 1 hold and function $F_\varepsilon^{(k)}(x)$ is strictly incresing in some neighbourhood of $Q_\varepsilon^{(k)}(\alpha)$ then*

$$\hat{Q}_n^{(k)}(\alpha) \xrightarrow{\mathsf{P}} Q_\varepsilon^{(k)}(\alpha) \text{ as } n \to \infty.$$

2. Proofs

To prove the theorems we need two lemmas.

Lemma 1. *Let $h : \mathbb{R}^d \to [0, +\infty)$ be some function. Assume that (7) holds for some $\alpha > 0$ and $m = 1, \dots, M$.*
Then

$$\max_{j=1,\dots,n} h(\mathbf{X}_j) = o_P(n^{1/\alpha}).$$

Proof. This is a simple corollary of lemma 2 in (Maiboroda et al., 2022). □

Now, let ζ_j $j = 1, \dots, n$ be a sample of MVC model, i.e.

$$\mathsf{P}\{\zeta_j < x\} = \sum_{k=1}^M p_{j;n}^k F_\zeta^{(k)}(x).$$

Denote by $\hat{F}_{\zeta;n}^k(x)$ the weighted empirical CDF of the form

$$\hat{F}_{\zeta;n}^{(k)}(x) = \sum_{j=1}^n a_{j;n}^{(k)} \mathbb{1}\{\zeta_j < x\},$$

Lemma 2. *There exists such a random variable $\Lambda < \infty$ a.s. that*

$$\sup_{x\in\mathbb{R}} |\hat{F}_{\zeta;n}^{(k)}(x) - F_\zeta^{(k)}(x)| \leqslant \Lambda\sqrt{n \ln n} \sup_{j=1,\dots n} |a_{j;n}^{(k)}|$$

for all $n = 1, 2, \dots$.

Proof. See corollary 2.2.4 in (Maiboroda et al., 2008). □

Proof of Theorem 1. Let

$$z_j = \varepsilon_j + g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(\kappa_j)}) - g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(k)}),$$
$$\delta_j = g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(k)}) - g(\mathbf{X}_j, \hat{\boldsymbol{\vartheta}}_n^{(k)}).$$

Then by (3)

$$U_j^{(k)} = g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(\kappa_j)}) + \varepsilon_j - g(\mathbf{X}_j, \hat{\boldsymbol{\vartheta}}_n^{(k)}) = z_j + \delta_j.$$

Let Δ_n be some nonrandom numerical sequence, $\Delta_n \to 0$ as $n \to \infty$. Consider a random event

$$A_n = \{\sup_{j=1,\dots,N} |\delta_j| < \Delta_n\}.$$

Assume that A_n holds. Then

$$\mathbb{1}\{z_j < x - \Delta\} \leqslant \mathbb{1}\{U_j^{(k)} < x\} = \mathbb{1}\{z_j < x - \delta_j\} \leqslant \mathbb{1}\{z_j < x + \Delta_n\}.$$

So

$$|\mathbb{1}\{U_j^{(k)} < x\} - \mathbb{1}\{z_j < x\}| \leqslant \mathbb{1}\{z_j < x + \Delta_n\} - \mathbb{1}\{z_j < x - \Delta\} = \mathbb{1}\{x - \Delta < z_j \leqslant x + \Delta\}. \quad (8)$$

Let

$$\hat{F}_{z;n}^{(m)}(x) = \sum_{j=1}^{n} a_{j;n}^{m} \mathbb{1}\{z_j < x\}.$$

Assumption (1) of the Theorem implies that there exists a $C_a < \infty$ such that

$$\max_{j=1,\dots,n, m=1,\dots,M} |a_{j;n}^m| \leqslant C_a/n. \quad (9)$$

So, by (8),

$$\sup_x |\tilde{F}_n^{(k)}(x) - \hat{F}_{z;n}^{(k)}(x)| \leqslant \sup_x \sum_{j=1}^{n} |a_{j;n}^k| \mathbb{1}\{x - \Delta_n < z_j \leqslant x + \Delta_n\} \leqslant \frac{C_a}{n} \sup_x \sum_{j=1}^{n} \mathbb{1}\{x - \Delta_n < z_j < x + \Delta_n\}, \quad (10)$$

if A_n holds.

Let

$$\check{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{z_j < x\}.$$

We will show that

$$\check{F}_n(x) = \sum_{m=1}^{M} \bar{p}_{;n}^m \hat{F}_{z;n}^{(m)}(x), \quad (11)$$

where $\bar{p}_{;n}^m = \frac{1}{n}\sum_{j=1}^n p_{j;n}^m$, $\bar{\mathbf{p}}_{;n} = (\bar{p}^1, \dots, \bar{p}^M)^T$.

Let $\mathbf{a}_{;n} \in \mathbb{R}^{n\times M}$ be the matrix $\mathbf{a}_{;n} = (a_{j;n}^m)_{j=1,\dots,n}^{m=1,\dots M}$ and $\mathbf{p} = (p_{j;n}^m)_{j=1,\dots,n}^{m=1,\dots M}$. Then (2) is equivalent to

$$\mathbf{a}_{;n}\mathbf{\Gamma}_n^{-1} = \mathbf{p}_{;n}.$$

Let for any natural r, $\mathbf{u}_r = (1, 1, \dots, 1)^T \in \mathbb{R}^r$. Then with the equality $\sum_{m=1}^M p_{j;n}^m = 1$ in mind we have

$$\mathbf{a}_{;n}\mathbf{\Gamma}_n\mathbf{u}_r = \mathbf{p}_{;n}\mathbf{u}_r = \mathbf{u}_n$$

and

$$\mathbf{\Gamma}_n\mathbf{u}_r = n\bar{\mathbf{p}}_{;n}.$$

So $\sum_{i=1}^M a_{j;n}^m \bar{p}_{j;n}^m = 1$ and by the definitions of $\check{F}_n(x)$ and $\hat{F}_{z;n}^{(m)}(x)$ we obtain (11).

So, by Lemma 2 and (9) we obtain that, for some random variable $\Lambda < \infty$,

$$|\check{F}_n(x) - \bar{F}_n(x)| \leqslant \Lambda \ell_n, \quad (12)$$

where

$$\bar{F}_n(x) = \sum_{m=1}^{M} \bar{p}^m F_z^{(m)}(x),$$

$$F_z^{(m)}(x) = \mathsf{P}\{z_j < x \mid \kappa_j = m\},$$

$$\ell_n = \sqrt{\frac{\ln n}{n}}$$

Combining (10) and (12) we obtain

$$\sup_x |\tilde{F}_n^{(k)}(x) - \hat{F}_{z;n}^{(k)}(x)| \leqslant C_a(\check{F}_n(x + \Delta_n) - \check{F}_n(x - \Delta_n)) \leqslant$$

$$\leqslant C_a\left(\bar{F}_n(x + \Delta_n) - \bar{F}_n(x - \Delta_n) + 2\Lambda\ell_n\right) \leqslant C_a \sum_{m=1}^{M} (F_z^{(m)}(x + \Delta_n) - F_z^{(m)}(x - \Delta_n)) + 2C_a\Lambda\ell_n.$$

Let us bound

$$J(x, \Delta_n) = F_z^{(m)}(x + \Delta_n) - F_z^{(m)}(x - \Delta_n).$$

For $m = 1, \dots, M$ let

$$\eta_j^{(m)} = g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(m)}) - g(\mathbf{X}_j, \boldsymbol{\vartheta}^{(k)}),$$
$$F_\eta^{(m)}(x) = \mathsf{P}\{\eta_j^{(m)} < x \mid \kappa_j = m\},$$

$f_\eta^{(m)}$ be the PDF of $F_\eta^{(m)}$. Observe that

$$F_z^{(m)}(x) = \mathsf{P}\{z_j < x \mid \kappa_j = m\}$$
$$= \mathsf{P}\{\varepsilon_j + \eta_j^{(\kappa_j)} < x \mid \kappa_j = m\} = \int_{-\infty}^{\infty} F_\varepsilon^{(m)}(x-t) F_\zeta^{(m)}(dt)$$

since $\mathbf{X}_j$ and ϵ_j are conditionally independent for fixed κ_j and

$$f_\eta^{(m)}(x) = \int_{-\infty}^{\infty} f_\varepsilon^{(m)}(x-t) F_\zeta^{(m)}(dt) \leqslant c_f,$$

where

$$c_f = \max_{k=1,\dots,M} \sup_{t\in\mathbb{R}} f_\varepsilon^{(m)}(t).$$

So

$$J(x, \Delta_n) \leqslant 2\Delta_n c_f,$$

So from (10) we obtain

$$\sup_{x\in\mathbb{R}} |\tilde{F}_n^{(k)(x)} - \hat{F}_{z;n}^{(k)}(x)| \leqslant C_a(2Mc_f\Delta_n + 2\Lambda\ell_n) \tag{13}$$

(If A_n holds).

Let us choose Δ_n to obtain

$$\mathsf{P}\{A_n\} \to 1 \text{ as } n \to \infty. \tag{14}$$

Assumption (2) implies $|\delta_j| \leqslant h(\mathbf{X}_j)|\hat{\boldsymbol{\vartheta}}_n^{(k)} - \boldsymbol{\vartheta}^{(k)}|$. By Lemma 1,

$$\max_{j=1,\dots,n} h(\mathbf{X}_j) = o_P(n^{1/\alpha})$$

and by Assumption (4), $|\hat{\boldsymbol{\vartheta}}_n^{(k)} - \boldsymbol{\vartheta}^{(k)}| = O_P(n^{-1/2})$.

So $|\delta_j| = o_P(n^{1/\alpha-1/2})$ and with $\Delta = n^{1/\alpha-1/2}$ we obtain (13). Equations (10) and (14) imply

$$\sup_{x\in\mathbb{R}} |\tilde{F}_n^{(k)}(x) - \hat{F}_{z;n}^{(k)}(x)| \xrightarrow{\mathsf{P}} 0 \text{ as } n \to \infty.$$

Since $\eta_j^{(k)} = 0$, we obtain $F_z^{(k)}(x) = F_\varepsilon^{(k)}(x)$, so Lemma 2 yields

$$\sup_{x\in\mathbb{R}} |\hat{F}_{z;n}^{(k)}(x) - F_\varepsilon^{(k)}| \xrightarrow{\mathsf{P}} 0, \text{ as } n \to \infty.$$

Combining the last two convergences we obtain the statement of the Theorem. □

Proof of Theorem 2. By the definition of $\hat{Q}_\varepsilon^{(k)}(\alpha)$, if $\tilde{F}_n^{(k)}(x) > \alpha$, then $\hat{Q}_\varepsilon^{(k)}(\alpha) \leqslant x$. So

$$\mathsf{P}\{\hat{Q}_\varepsilon^{(k)}(\alpha) \leqslant x\} \geqslant \mathsf{P}\{\tilde{F}_n^{(k)}(x) > \alpha\}. \tag{15}$$

Let us consider any $x > Q_\varepsilon^{(k)}(\alpha)$. By Theorem 1, $\mathsf{P}\{\tilde{F}_n^{(k)}(x) > F_\varepsilon^{(k)}(x) - \delta\} \to 1$ as $n \to \infty$ for all $\delta > 0$. By the assumption of Theorem 2, $F_\varepsilon^{(k)}(x) > \alpha$. Letting $\delta = F_\varepsilon^{(k)}(x) - \alpha$ we obtain $\mathsf{P}\{\tilde{F}_\varepsilon^{(k)}(x) > \alpha\} \to 1$. So

$$\mathsf{P}\{\hat{Q}_n^{(k)}(\alpha) < x\} \to 1 \text{ for } x \geqslant Q_\varepsilon^{(k)}(\alpha). \tag{16}$$

Now consider any $x < Q_\varepsilon^{(k)}(\alpha)$. If $\hat{Q}_n^{(k)}(\alpha) < x$ than

$$S_{x;n} = \sup_{t<x} \tilde{F}_n^{(k)}(t) > \alpha.$$

Put $\delta = \alpha - F_\varepsilon^{(k)}$, $J_n = \sup_t |\tilde{F}_n^{(k)}(t) - F_\varepsilon^{(k)}(t)|$. Then $S_{x;n} \leqslant \sup_{t<x} F_\varepsilon^{(k)}(t) + J_n$ and $\mathsf{P}\{|J_n| > \delta\} \to 0$ by Theorem 1. So

$$\mathsf{P}\{\hat{Q}_n^{(k)}(\alpha) < x\} \leqslant \mathsf{P}\{S_{x;n} > \alpha\} \leqslant \mathsf{P}\{F_\varepsilon^{(k)}(x) + J_n > \alpha\} = \mathsf{P}\{J_n > \delta\} \to 0,$$

as $n \to \infty$. Combining this with (16) we obtain the statement of the Theorem. □

Discussion and conclusions

We discussed estimators for error terms distributions and quantiles in mixture of regressions models and show their consistency. The results can be applied to construct QQ-plots for visual analysis of these distributions.

Authors' contribution: Rostyslav Maiboroda – conceptualization, writing – viewing and editing; Vitaliy Miroshnychenko – methodology, writing original draft; Olena Sugakova – methodology, writing – viewing and editing.

References

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. Chapman & Hall/CRC. doi: https://doi.org/10.1201/9781420035919

Groß, J. (2003). *Linear regression*. Berlin Heidelberg Springer. doi: https://doi.org/10.1007/978-3-642-55864-1

Grün, B., & Leisch, F. (2006). Fitting finite mixtures of linear regression models with varying and fixed effects in R. In A. Rizzi & M. Vichi (Eds.), *Compstat 2006 - Proceedings in Computational Statistics* (pp. 853–860). Heidelberg Physica Verlag.

Maiboroda, R., & Miroshnychenko, V. (2020). Asymptotic normality of modified ls estimator for mixture of nonlinear regressions. *Modern Stochastics: Theory and Applications*, *7*(4), 435–448. doi: https://doi.org/10.15559/20-VMSTA167

Maiboroda, R., Miroshnychenko, V., & Sugakova, O. (2022). Jackknife for nonlinear estimating equations. *Modern Stochastics: Theory and Applications*, *9*(4), 377–399. doi: https://doi.org/10.15559/22-VMSTA208

Maiboroda, R., & Sugakova, O. (2008). *Estimation and classification by observations from a mixture*. Kyiv University [in Ukrainian].

Miroshnychenko, V. O. (2019). Residual analysis in regression mixture model. *Bulletin of Taras Shevchenko National University of Kyiv, Series: Physics and Mathematics*, *3*, 8–16. doi: https://doi.org/10.17721/1812-5409.2019/3.1

Pidnebesna, A., Fajnerová, I., Horáček, J., & Hlinka, J. (2023). Mixture components inference for sparse regression: Introduction and application for estimation of neuronal signal from fMRI BOLD. *Applied Mathematical Modelling*, *116*, 735–748. doi: https://doi.org/10.1016/j.apm.2022.11.034

Отримано редакцією журналу / Received: 22.02.24
Прорецензовано / Revised: 02.04.24
Схвалено до друку / Accepted: 15.05.24

Ростислав МАЙБОРОДА, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-3899-558X
e-mail: rostmaiboroda@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Віталій МІРОШНИЧЕНКО, д-р філософії, асист.
ORCID ID: 0009-0003-9799-8902
e-mail: vitaly.miroshnychenko@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Олена СУГАКОВА, д-р фіз.-мат. наук, доц.
ORCID ID: 0000-0002-6529-0788
e-mail: olenasuhakova@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

ОЦІНКИ КВАНТИЛІВ РОЗПОДІЛУ ПОХИБОК РЕГРЕСІЇ У МОДЕЛЯХ СУМІШЕЙ ЗІ ЗМІННИМИ КОНЦЕНТРАЦІЯМИ

У статті розглянуто дані, що описуються моделлю суміші зі змінними концентраціями. Моделі сумішей кількох компонент приро-дно виникають під час аналізу статистичних даних медико-біологічних та соціологічних досліджень. В цьому випадку об'єкти спосте-реження вибирають зі скінченного набору підпопуляцій (компонентів суміші), які мають різні розподіли спостережуваних характери-стик. Концентрації компонентів у суміші (ймовірності змішування) можуть змінюватись від спостереження до спостереження. У досліджуваних даних залежність між змінними описується моделлю нелінійної регресії. Коефіцієнти цієї моделі є різними для різних компонентів. Розподіли похибок регресії також можуть бути різними. Вивчення цих розподілів відіграє значну роль у регресійному аналізі. Для візуального порівняння розподілів похибок з відомими розподілами використовують діаграми квантиль-квантиль (QQ). Такі діаграми є важливим засобом діагностики регресійної моделі. Для однорідних даних QQ-діаграми зручно будувати, використо-вуючи залишки регресії. У моделі суміші за наявності кількох регресійних моделей залишки неможливо визначити однозначно. Тому виникає потреба у побудові оцінок для квантилів розподілу залишків. Для отримання оцінок квантилів можна скористатись оцінками функцій відповідних розподілів. Такі оцінки у моделях сумішей зі змінними концентраціями будуються з використанням навантажених емпіричних функцій розподілу (нефр). Для незалежних спостережень асимптотична поведінка нефр добре досліджена. Однак у випадку аналізу похибок дослідник має справу із залишками, які є залежними для різних спостережень. З огляду на це, задача асимптотичного дослідження нефр і відповідних оцінок квантилів у цьому випадку є актуальною.

У пропонованій роботі розглянуто побудову нефр та оцінок квантилів для похибок регресії за спостереженнями із суміші і до-ведено їхню консистентність. Для побудови оцінок використовується нефр з мінімаксними навантаженнями, які розраховуються за концентраціями компонент у суміші. Для доведення консистентності оцінок використано асимптотичну теорію таких нефр, зокре-ма нерівність Вапника – Червоненкіса. Визначено умови консистентності оцінок функцій розподілу та квантилів похибок регресії у моделях сумішей зі змінними коцентраціями. Отримані оцінки можуть бути використані для побудови QQ-діаграм для візуального аналізу розподілу похибок.

К л ю ч о в і с л о в а: ***модель суміші зі змінними концентраціями, нелінійна регресія, залишки, квантилі.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis or interpretation of data; in the writing of the manuscript; in the decision to publish the results.