65 (2020) APPLICATIONS OF MATHEMATICS No. 3, 257–269

EMPIRICAL REGRESSION QUANTILE PROCESSES

Jana Jurečková, Praha, Jan Picek, Martin Schindler, Liberec

Received November 8, 2019. Published online May 25, 2020.

Abstract. We address the problem of estimating quantile-based statistical functionals, when the measured or controlled entities depend on exogenous variables which are not under our control. As a suitable tool we propose the empirical process of the average regression quantiles. It partially masks the effect of covariates and has other properties convenient for applications, e.g. for coherent risk measures of various types in the situations with covariates.

Keywords: averaged regression quantile; one-step regression quantile; R-estimator; functionals of the quantile process

MSC 2020: 62J02, 62G30, 90C05, 65K05, 49M29

1. Introduction

The empirical quantile process and its functionals are applied in many domains of everyday life. Our main tool is the process of averaged α-regression quantiles, introduced in [8], which is useful for its ability to mask the influence of covariates. The trajectories of this process approximate the quantile function of the model errors even in the presence of nuisance covariates. Their inversions in turn approximate the parent distribution function. Another related empirical process is that of two-step regression quantiles, which first estimates the slope components (the effects of covariates) by means of R-estimate (rank-estimate) and supplements it with estimating the intercept component as a quantile of residuals. Both processes are asymptotically equivalent, and their finite-sample properties suitably supplement each other. We complete the study with a numerical illustration of both processes, and mention some possible applications.

The authors gratefully acknowledge the support of the Grant GAČR 18-01137S.

DOI: 10.21136/AM.2020.0295-19

2. Notation and basic concepts

We consider the linear regression model

$$Y_{ni} = \beta_0 + \mathbf{x}_{ni}^\top \boldsymbol{\beta} + e_{ni}, \quad i = 1, \dots, n, \tag{2.1}$$

where $Y_{n1}, \dots, Y_{nn}$ are observed responses, $e_{n1}, \dots, e_{nn}$ are independent model errors, possibly non-identically distributed with unknown distribution functions F_i, $i = 1, \dots, n$. The covariates $\mathbf{x}_{ni} = (x_{i1}, \dots, x_{ip})^\top$, $i = 1, \dots, n$, are random or non-random, and $\boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta}^\top)^\top = (\beta_0, \beta_1, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ is an unknown parameter. We also use the notation $\mathbf{x}_{ni}^* = (1, x_{i1}, \dots, x_{ip})^\top$, $i = 1, \dots, n$.

Recall that the regression α-quantile, $0 \leqslant \alpha \leqslant 1$,

$$\hat{\boldsymbol{\beta}}_n^*(\alpha) = (\hat{\beta}_{n0}(\alpha), (\hat{\boldsymbol{\beta}}_n(\alpha))^\top)^\top = (\hat{\beta}_{n0}(\alpha), \hat{\beta}_{n1}(\alpha), \dots, \hat{\beta}_{np}(\alpha))^\top$$

is a $(p+1)$-dimensional vector defined as a minimizer

$$\hat{\boldsymbol{\beta}}_n^*(\alpha) = \arg \min_{\mathbf{b} \in \mathbb{R}^{p+1}} \left\{ \sum_{i=1}^n [\alpha (Y_i - \mathbf{x}_i^{*\top} \mathbf{b})^+ + (1-\alpha)(Y_i - \mathbf{x}_i^{*\top} \mathbf{b})^-] \right\}, \tag{2.2}$$
$$\text{where } z^+ = \max(z, 0) \text{ and } z^- = \max(-z, 0),\ z \in \mathbb{R}^1.$$

The solution $\hat{\boldsymbol{\beta}}_n^*(\alpha) = (\hat{\beta}_0(\alpha), \hat{\boldsymbol{\beta}}(\alpha)^\top)^\top$ minimizes the sum of $(\alpha, 1-\alpha)$ convex combinations of positive and negative parts of residuals $(Y_i - \mathbf{x}_i^{*\top} \mathbf{b})$ over $\mathbf{b} \in \mathbb{R}^{p+1}$. If the response Y_i represents a loss, then the choice of α depends on the balance between underestimating and overestimating the respective losses Y_i, $i = 1, \dots, n$. Increasing $\alpha \nearrow 1$ reflects a greater concern about underestimating losses Y, compared to overestimating.

The *averaged regression α-quantile* is the specific weighted mean of components of $\hat{\boldsymbol{\beta}}_n^*(\alpha)$, $0 \leqslant \alpha \leqslant 1$:

$$\overline{B}_n(\alpha) = \overline{\mathbf{x}}_n^{*\top} \hat{\boldsymbol{\beta}}_n^*(\alpha) = \hat{\beta}_{n0}(\alpha) + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p x_{ij} \hat{\beta}_j(\alpha), \quad \overline{\mathbf{x}}_n^* = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^*.$$

If the model errors e_{ni} are identically distributed with a continuous distribution function, then the residual $\overline{B}_n(\alpha) - \beta_0 - \bar{\mathbf{x}}_n^\top \boldsymbol{\beta}$ is asymptotically equivalent to the $[n\alpha]$-quantile $e_{n:[n\alpha]}$ of the e_{ni}, $i = 1, \dots, n$, as $n \to \infty$. The advantage of the methodology based on the *averaged regression α-quantile* is that it partially suppresses the role of the unobservable covariates, and so it enables an inference on functionals of $\mathbf{Y}$ even under the nuisance regression with unobservable coefficients.

The behavior of $\overline{B}_n(\alpha)$ with $0 < \alpha < 1$ has been illustrated in [1] and [2], and summarized in [12]; they showed that $\overline{B}_n(\alpha)$ is a nondecreasing step function of

$\alpha \in (0,1)$ with a finite number J_n of breakpoints. The upper bound of J_n is generally equal to $\binom{n}{p+1} = \mathcal{O}(n^{p+1})$. This is a huge number for n increasing; however, Portnoy in [16] showed that J_n can be much smaller, namely $J_n = \mathcal{O}_p(n \log n)$ as $n \to \infty$, under some conditions on the design matrix $\mathbf{X}_n$. The most general conditions for this rate are still an open question.

An alternative *two-step regression α-quantile*, introduced in [7], differs from $\hat{\boldsymbol{\beta}}_n^*(\alpha)$ in that the slope components $\boldsymbol{\beta}$ are estimated by a specific R-estimate $\tilde{\boldsymbol{\beta}}_{nR}$. The R-estimate is invariant to the shift in location and hence independent of the intercept. The intercept component is estimated by the α-quantile of residuals of Y_i's from $\tilde{\boldsymbol{\beta}}_{nR}$. The averaged two-step regression quantile $\widetilde{B}_n(\alpha)$ is defined analogously to $\overline{B}_n(\alpha)$. Both sequences are asymptotically equivalent; however, the number of breakpoints of $\widetilde{B}_n(\alpha)$ exactly equals to n (as in the location model), while the number of breakpoints of $\overline{B}_n(\cdot)$ can be much larger. The averaged regression quantile $\overline{B}_n(\alpha)$ is monotone in α, while the two-step averaged regression quantile $\widetilde{B}_n(\alpha)$ is monotone under a suitable R-estimate $\tilde{\boldsymbol{\beta}}_{nR}$. Both $\overline{B}_n(\cdot)$ and $\widetilde{B}_n(\cdot)$ behave like an empirical quantile function. As such, they can be inverted and their inversions approximate the parent distribution function F of the model errors.

The methods based on $\overline{B}_n$ and on its modifications are nonparametric, thus applicable also to heavy-tailed and skewed distributions. An extension to autoregressive models is also possible and will be a subject of further study. The autoregression quantile reflects the behavior based on the past assets, while the averaged regression quantile tries to mask the past history.

Section 3 is devoted to the finite-sample properties as well as to the asymptotic properties of the averaged regression quantile process $\overline{B}_n(\cdot)$, along with its applications. Section 4 deals with the process of the two-step average regression quantile $\widetilde{B}_n(\cdot)$. An intensive numerical illustration of both processes is given in Section 5. Section 6 briefly mentions some functionals of the regression quantile process.

3. Behavior of $\overline{B}_n(\alpha)$ over $\alpha \in (0,1)$

The minimization (2.2) with $\alpha \in [0,1]$ fixed was treated in [14] as a special linear programming problem; various modifications of this algorithm were later developed. Its dual program is a parametric linear program, which can be written as

$$
\begin{array}{lll}
\text{maximize} & \mathbf{Y}_n^\top \hat{\mathbf{a}}(\alpha) \\
\text{under} & \mathbf{X}_n^{*\top} \hat{\mathbf{a}}(\alpha) = (1-\alpha)\mathbf{X}_n^{*\top} \mathbf{1}_n^\top \\
& \hat{\mathbf{a}}(\alpha) \in [0,1]^n,\ 0 \leqslant \alpha \leqslant 1,
\end{array}
\tag{3.1}
$$

where

$$\mathbf{X}_n^* = \begin{bmatrix} \mathbf{x}_{n1}^{*\top} \\ \dots \\ \mathbf{x}_{nn}^{*\top} \end{bmatrix}$$

is of order $n \times (p+1)$. The components of the optimal solution $\hat{\mathbf{a}}(\alpha) = (\hat{a}_{n1}(\alpha), \dots, \hat{a}_{nn}(\alpha))^\top$ of (3.1), called *regression rank scores*, were studied in [3], where it is shown that $\hat{a}_{ni}(\alpha)$ is a continuous, piecewise linear function of $\alpha \in [0,1]$ and $\hat{a}_{ni}(0) = 1$, $\hat{a}_{ni}(1) = 0$, $i = 1, \dots, n$. Moreover, it follows from (3.1) that $\hat{\mathbf{a}}(\alpha)$ is invariant in the sense that it does not change if $\mathbf{Y}$ is replaced with $\mathbf{Y} + \mathbf{X}_n^* \mathbf{b}^*$ for all $\mathbf{b}^* \in \mathbb{R}^{p+1}$.

Let $\{\mathbf{x}_{i_1}^*, \dots, \mathbf{x}_{i_{p+1}}^*\}$ be the optimal base in (3.1) and let $\{Y_{i_1}, \dots, Y_{i_{p+1}}\}$ be the corresponding responses in model (2.1). The following theorem shows that $\overline{B}_n(\alpha)$ equals a weighted mean of $\{Y_{i_1}, \dots, Y_{i_{p+1}}\}$, with the weights based on the regressors.

Theorem 3.1. *Assume that the regression matrix $\mathbf{X}_n^*$ has full rank $p+1$ and that the distribution functions $F_1, \dots, F_n$ of model errors are continuous and increasing in $(-\infty, \infty)$. Then with probability* 1

$$\overline{B}_n(\alpha) = \sum_{k=1}^{p+1} w_{k,\alpha} Y_{i_k}, \quad \sum_{k=1}^{p+1} w_{k,\alpha} = 1 \tag{3.2}$$

and

$$\overline{B}_n(\alpha) \leqslant \overline{B}_n(1) < \max_{i \leqslant n} Y_i, \tag{3.3}$$

where the vector $\mathbf{Y}_n(1) = (Y_{i_1}, \dots, Y_{i_{p+1}})^\top$ corresponds to the optimal base of the linear program (3.1). *The vector $\mathbf{w}_\alpha = (w_{1,\alpha}, \dots, w_{p+1,\alpha})^\top$ of coefficients equals $\mathbf{w}_\alpha = [n^{-1} \mathbf{1}_n^\top \mathbf{X}_n^* (\mathbf{X}_{n1}^*)^{-1}]^\top$, where $\mathbf{X}_{n1}^*$ is the submatrix of $\mathbf{X}_n^*$ with the rows $\mathbf{x}_{i_1}^{*\top}, \dots, \mathbf{x}_{i_{p+1}}^{*\top}$.*

P r o o f. The regression quantile $\hat{\boldsymbol{\beta}}_n^*(\alpha)$ is a step function of $\alpha \in (0,1)$. If α is a continuity point of the regression quantile trajectory, then we have the following identity, proven in [8]:

$$\overline{B}_n(\alpha) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{*\top} \hat{\boldsymbol{\beta}}_n^*(\alpha) = -\frac{1}{n} \sum_{i=1}^n Y_i \hat{a}'_{ni}(\alpha), \tag{3.4}$$

where

$$\hat{a}'_{ni}(\alpha)) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \hat{a}_{ni}(\alpha).$$

Moreover, (3.1) implies

$$\sum_{i=1}^{n} \hat{a}'_{ni}(\alpha) = -n \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}\hat{a}'_{ni}(\alpha) = -\sum_{i=1}^{n} x_{ij} \quad \forall\, j \in \{1, \ldots, p\}.$$

Notice that $\hat{a}'_{ni}(\alpha) \neq 0$ if and only if α is the point of continuity of $\hat{\boldsymbol{\beta}}^*_n(\cdot)$ and $Y_i = \mathbf{x}_i^{*\top}\hat{\boldsymbol{\beta}}^*_n(\alpha)$. To every fixed continuity point α there correspond exactly $p+1$ such components with the property that the corresponding $\mathbf{x}_i^*$ belongs to the optimal base of program (3.1). Hence, there exist coefficients $w_{k,\alpha}$, $k = 1, \ldots, p+1$, such that

$$\overline{B}_n(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} Y_i \hat{a}'_{ni}(\alpha) = \sum_{k=1}^{p+1} w_{k,\alpha} Y_{i_k}.$$

The equalities $Y_i = \mathbf{x}_i^{*\top}\hat{\boldsymbol{\beta}}^*_n(\alpha)$ hold just for $p+1$ components of the optimal base $\mathbf{x}^*_{i_1}, \ldots, \mathbf{x}^*_{i_{p+1}}$. Let $\mathbf{X}^*_{n1}$ be the submatrix of $\mathbf{X}^*_n$ with the rows $\mathbf{x}_{i_1}^{*\top}, \ldots, \mathbf{x}_{i_{p+1}}^{*\top}$ and let $(\hat{\mathbf{a}}'_1(\alpha))^\top = (\hat{a}'_{i_1}(\alpha), \ldots, \hat{a}'_{i_{p+1}}(\alpha))$. Then $\mathbf{X}^*_{n1}$ is regular with probability 1 and

$$\mathbf{w}_\alpha^\top = -\frac{1}{n}(\hat{\mathbf{a}}'(\alpha))^\top = \frac{1}{n}\mathbf{1}_n^\top \mathbf{X}^*_n (\mathbf{X}^*_{n1})^{-1}.$$

This and (3.4) imply (3.2). Inequality (3.3) was proven in [6]. □

Let us now consider $\overline{B}_n(\alpha)$ as a process in $\alpha \in (0,1)$ under the condition that all model errors e_{ni}, $i = 1, \ldots, n$, are independent and equally distributed according to joint continuous increasing distribution function F. We are interested in the *average regression quantile process*

$$\overline{\mathcal{B}}_n(\alpha) = \{n^{1/2}\bar{\mathbf{x}}_n^{*\top}(\hat{\boldsymbol{\beta}}^*_n(\alpha) - \check{\boldsymbol{\beta}}(\alpha));\ 0 < \alpha < 1\},$$

where $\check{\boldsymbol{\beta}}(\alpha) = (F^{-1}(\alpha) + \beta_0, \beta_1, \ldots, \beta_p)^\top$ is the population counterpart of the regression quantile. As proven in [3], the process $\overline{\mathcal{B}}_n$ converges to a Gaussian process in the Skorokhod topology as $n \to \infty$, under mild conditions on F and $\mathbf{X}_n$. More precisely,

$$\overline{\mathcal{B}}_n(\cdot) \xrightarrow{\mathcal{D}} (f(F^{-1}(\cdot)))^{-1} W^*(\cdot) \quad \text{as } n \to \infty,$$

where W^* is the Brownian bridge on (0,1). The convergence holds on every subinterval $[\varepsilon, 1-\varepsilon] \subset (0,1)$.

Under a finite number of observations, the trajectories of $\overline{\mathcal{B}}_n(\cdot)$ are step functions, nondecreasing in $\alpha \in (0,1)$, and they have a finite number of discontinuities for each

fixed n. As shown in [2], if $\overline{B}_n(\alpha_1) = \overline{B}_n(\alpha_2)$ for $0 < \alpha_1 < \alpha_2 < 1$, then $\alpha_2 - \alpha_1 \leqslant (p+1)/n$ with probability 1. It means that the length of the interval on which $\overline{B}_n(\alpha)$ is constant tends to 0 for $n \to \infty$ and fixed p. Let $0 < \alpha_1 < \ldots < \alpha_{J_n} < 1$ be the breakpoints of $\overline{B}_n(\alpha)$, $0 < \alpha < 1$, and let $-\infty < Z_1 < \ldots < Z_{J_n+1} < \infty$ be the corresponding values of $\overline{B}_n(\alpha)$ between the breakpoints. Then we can consider the inversion $\widehat{F}_n(z)$ of $\overline{B}_n(\alpha)$, namely $\widehat{F}_n(z) = \inf\{\alpha\colon\ \overline{B}_n(\alpha) \geqslant z\},\ -\infty < z < \infty$.

It is a bounded nondecreasing step function and, given $Y_1, \ldots, Y_n$ satisfying (2.1), $\widehat{F}_n$ is a distribution function of a random variable, attaining values $Z_1, \ldots, Z_{J_n+1}$ with probabilities equal to the spacings of the vector $(0, \alpha_1, \ldots, \alpha_{J_n}, 1)$. Portnoy [15] studied the tightness of the empirical process $\widehat{F}_n(\cdot)$ and showed its convergence to $F(\cdot)$ under some specific conditions. We recommend $\widehat{F}_n$ as an estimate of F and also recommend to weaken the conditions for the convergence. As we illustrate in the numerical study of Section 4, the approximation is excellent. The process $\overline{\mathcal{B}}_n(\cdot)$ has many applications; besides estimating various functionals of F as the risk measures, it enables goodness-of-fit testing about F even in the presence of a nuisance regression.

4. The averaged two-step regression quantile $\widetilde{\mathcal{B}}_n(\alpha)$

The quantile $\widetilde{\mathcal{B}}_n(\alpha)$ has exactly n breakpoints, which is an advantage compared to $\overline{B}_n(\alpha)$. In spite of that, both processes are asymptotically equivalent as $n \to \infty$.

The two-step regression α-quantile treats the slope components $\boldsymbol{\beta}$ and the intercept β_0 separately. The slope component part is an R-estimate $\tilde{\boldsymbol{\beta}}_{nR}$ of $\boldsymbol{\beta}$, which is invariant to the shift in location, hence independent of β_0. Its determination starts with the selection of a nondecreasing rank-score function $\varphi(u)$, $u \in (0,1)$, square-integrable on (0,1). We can consider two types of rank scores, generated by φ:

(1) Exact scores: $\mathcal{A}_n(i) = \mathbb{E}\{\varphi(U_{n:i})\}$, $i = 1, \ldots, n$, where $U_{n:1} \leqslant \ldots \leqslant U_{n:n}$ is an ordered random sample of size n from the uniform (0,1) distribution.

(2) Approximate scores:

$$\text{either}\quad \text{(i)}\quad A_n(i) = n\int_{(i-1)/n}^{i/n} \varphi(u)\,\mathrm{d}u, \tag{4.1}$$

$$\text{or}\quad \text{(ii)}\quad A_n(i) = \varphi\Big(\frac{i}{n+1}\Big),\quad i = 1, \ldots, n.$$

The test criteria and estimates based on either of these scores are asymptotically equivalent as $n \to \infty$; but the rank tests based on the exact scores are locally most powerful under finite n against pertinent alternatives. The R-estimator $\tilde{\boldsymbol{\beta}}_{nR}$ of the

slopes is defined as a minimizer of the Jaeckel [5] measure of rank dispersion $\mathcal{D}_n(\mathbf{b})$:

$$\tilde{\boldsymbol{\beta}}_{nR} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \mathcal{D}_n(\mathbf{b}), \tag{4.2}$$

$$\text{where} \quad \mathcal{D}_n(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \mathbf{b}) A_n(R_{ni}(Y_i - \mathbf{x}_i^\top \mathbf{b})).$$

Here $R_{ni}(Y_i - \mathbf{x}_i^\top \mathbf{b})$ is the rank of the ith residual, $i = 1, \dots, n$. $\mathcal{D}_n$ is a convex function, piecewise linear in $\mathbf{b} \in \mathbb{R}^p$.

The intercept component $\tilde{\beta}_{n0}(\alpha)$ of the two-step regression α-quantile is defined as the $[n\alpha]$th order statistic of the residuals $Y_i - \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}_{nR}$, $i = 1, \dots, n$. The two-step α-regression quantile is then the vector

$$\tilde{\boldsymbol{\beta}}_n^*(\alpha) = \begin{pmatrix} \tilde{\beta}_{n0}(\alpha) \\ \tilde{\boldsymbol{\beta}}_{nR} \end{pmatrix} \in \mathbb{R}^{p+1}. \tag{4.3}$$

The typical choice of φ is the following:

$$\varphi_\lambda(u) = \lambda - I[u < \lambda], \quad 0 < u < 1, \quad \lambda \in (0,1) \text{ fixed}, \tag{4.4}$$

combined with the approximate scores (ii) in (4.1). These scores were introduced in [4]; Hájek used the following scores (now known as Hájek's rank scores):

$$a_i(\lambda, \mathbf{b}) = \begin{cases} 0 & \dots \quad R_{ni}(Y_i) < n\lambda, \\ R_{ni}(Y_i) - n\lambda & \dots \quad n\lambda \leqslant R_{ni}(Y_i) < n\lambda + 1, \\ 1 & \dots \quad n\lambda + 1 \leqslant R_{ni}(Y_i). \end{cases} \tag{4.5}$$

The solutions of (4.2) are generally not uniquely determined. We can, e.g., take the center of gravity of the set of all solutions; however, the asymptotic representations and distributions apply to any solution.

Define the *averaged two-step regression α-quantile* as $\widetilde{B}_n(\alpha) = \bar{\mathbf{x}}_n^{*\top} \tilde{\boldsymbol{\beta}}_n^*(\alpha)$. By (4.2) and (4.3)

$$\widetilde{B}_n(\alpha) = (Y_i - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \tilde{\boldsymbol{\beta}}_{nR})_{n:[n\alpha]}, \tag{4.6}$$

hence, it is equal to the $[n\alpha]$th order statistic of the residuals $Y_i - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \tilde{\boldsymbol{\beta}}_{nR}$, $i = 1, \dots, n$. Then $\widetilde{B}_n(\alpha)$ is obviously scale equivariant and regression equivariant. The R-estimator of the slopes in (4.4) with a fixed $\lambda \in (0,1)$, independent of α, guarantees that $\widetilde{B}_n(\alpha)$ is monotone in α, thus invertible. Under general conditions, $\widetilde{B}_n(\alpha)$ is asymptotially equivalent to $\overline{B}_n(\alpha)$, hence also asymptotically equivalent to

$e_{n:[n\alpha]} + \beta_0 + \bar{\mathbf{x}}_n^\top \boldsymbol{\beta}$, we shall prove under the following mild conditions on F and on $\mathbf{X}_n = [\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nn}]^\top$:

(A1) *Smoothness of F:* The errors e_{ni}, $i = 1, \ldots, n$, are independent and identically distributed. Their distribution function F has an absolutely continuous density and positive and finite Fisher information.

(A2) *Noether's condition on regressors:*

$$\lim_{n\to\infty} \mathbf{Q}_n = \mathbf{Q}, \text{ where } \mathbf{Q}_n = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n)(\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top$$

and $\mathbf{Q}$ is a positive definite $p \times p$ matrix; moreover,

$$\lim_{n\to\infty} \max_{1\leqslant i\leqslant n} n^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \mathbf{Q}_n^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_n) = 0.$$

(A3) *Rate of regressors:* $\max_{1\leqslant i\leqslant n} \|\mathbf{x}_{ni} - \bar{\mathbf{x}}_n\| = o(n^{1/4})$ as $n \to \infty$.

We shall prove the asymptotic equivalence for R-estimators based on the score function φ_λ, $0 < \lambda < 1$. However, an analogous proof applies to an R-estimator generated by any nondecreasing and square-integrable function φ.

Theorem 4.1. *Let $\widetilde{B}_n(\alpha) = (Y_i - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \tilde{\boldsymbol{\beta}}_{nR})_{n:[n\alpha]}$ be the two-step averaged α-regression quantile (TARQ) in the model* (2.1)*, with the R-estimator $\tilde{\boldsymbol{\beta}}_{nR}$ generated by φ_λ in* (4.4)*, $\lambda \in (0,1)$ fixed. Then, under the conditions* (A1)–(A3)*,*

$$\text{(4.7)} \qquad \text{(i)} \quad n^{1/2}[(\widetilde{B}_n(\alpha) - \beta_0 - \bar{\mathbf{x}}_n \boldsymbol{\beta}) - e_{n:[n\alpha]}] = o_p(1)$$

$$\text{(4.8)} \qquad \text{(ii)} \quad n^{1/2} |\widetilde{B}_n(\alpha) - \overline{B}_n(\alpha)| = o_p(1)$$

as $n \to \infty$, uniformly over $\alpha \in (\varepsilon, 1-\varepsilon) \subset (0,1)$ for all $\varepsilon \in (0, \frac{1}{2})$.

P r o o f. Consider the $[n\alpha]$-quantile of residuals

$$r_{ni} = e_{ni} - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top (\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta}) = Y_i - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \tilde{\boldsymbol{\beta}}_{nR} - \beta_0 - \bar{\mathbf{x}}_n^\top \boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

Under conditions (A1)–(A3), the R-estimator $\tilde{\boldsymbol{\beta}}_{nR}$ admits the following asymptotic representation:

$$\text{(4.9)} \quad n^{1/2}(\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta}) = n^{-1/2} (f(F^{-1}(\lambda)))^{-1} \mathbf{Q}_n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n)(\lambda - I[e_{ni} < F^{-1}(\lambda)]) + o_p(n^{-1/4}),$$

hence $\|n^{1/2}(\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta})\| = \mathcal{O}_p(1)$. The details for (4.9) can be found in [10]. The $[n\alpha]$-quantile $\tilde{a}_n(\alpha)$ of $r_{n1}, \ldots, r_{nn}$ is a solution of the minimization

$$\tilde{a}_n(\alpha) = \arg\min_{a \in \mathbb{R}^1} \sum_{i=1}^{n} \varrho_\alpha(r_{ni} - a),$$

where $\varrho_\alpha(z) = |z|\{\alpha \mathcal{I}[z > 0] + (1-\alpha)\mathcal{I}[z < 0]\}$, $z \in \mathbb{R}^1$. Using Lemma A.2 in [17], we can show that

$$\begin{aligned} &n^{-1/2} \sum_{i=1}^{n} \psi_\alpha(r_{ni} - \tilde{a}_n(\alpha)) \to 0, \quad \text{i.e.,} \\ &n^{-1/2} \sum_{i=1}^{n} (\alpha - \mathcal{I}[e_{ni} - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top (\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta}) < \tilde{a}_n(\alpha)]) \to 0 \end{aligned} \tag{4.10}$$

in probability as $n \to \infty$, where ψ_α is the right-hand derivative of ϱ_α, i.e. $\psi_\alpha(z) = \alpha - \mathcal{I}[z < 0]$, $z \in \mathbb{R}$. Moreover, we have the Bahadur representation of the sample α-quantile $e_{n:[n\alpha]}$ of $e_{n1}, \ldots, e_{nn}$:

$$\begin{aligned} &n^{1/2}[e_{n:[n\alpha]} - F^{-1}(\alpha)] \\ &\quad = n^{-1/2}[f(F^{-1}(\alpha))]^{-1} \sum_{i=1}^{n} \{\alpha - \mathcal{I}[e_{ni} < F^{-1}(\alpha)]\} + o(1) \text{ a.s. as } n \to \infty. \end{aligned} \tag{4.11}$$

Notice that $\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n) = \mathbf{0}$; similarly as in [10] we conclude that under $n \to \infty$

$$\sup_{\|\mathbf{b}\| \leqslant C} \left\{ n^{-1/2} \left| \sum_{i=1}^{n} (\mathcal{I}[e_{ni} - n^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top \mathbf{b} < \tilde{a}_n(\alpha)] - \mathcal{I}[e_{ni} < \tilde{a}_n(\alpha)]) \right| \right\} = o_p(1)$$

for every C, $0 < C < \infty$. Inserting $\mathbf{b} \mapsto n^{1/2}(\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta}) = \mathcal{O}_p(1)$, we obtain for $n \to \infty$

$$n^{-1/2} \left| \sum_{i=1}^{n} (\mathcal{I}[e_{ni} - (\mathbf{x}_i - \bar{\mathbf{x}}_n)^\top (\tilde{\boldsymbol{\beta}}_{nR} - \boldsymbol{\beta}) < \tilde{a}_n(\alpha)] - \mathcal{I}[e_{ni} < \tilde{a}_n(\alpha)]) \right| = o_p(1).$$

Combining the above arguments, we conclude that $n^{1/2}(\tilde{a}_n(\alpha)) - e_{n:[n\alpha]}) = o_p(1)$ as $n \to \infty$, hence,

$$n^{1/2}[\widetilde{B}_n(\alpha) - \beta_0 - \bar{\mathbf{x}}_n^\top \boldsymbol{\beta} - e_{n:[n\alpha]}] = o_p(1) \text{ as } n \to \infty, \tag{4.12}$$

which gives (4.7). This combined with Theorem 2 in [8] implies the propositions of Theorem 3.1. □

5. Computation and numerical illustrations

The simulation study illustrates the behavior of the proposed estimates and describes their computation. For the computation of the averaged regression quantile $\overline{B}_n(\alpha)$, the R package quantreg [13] and its function rq(·) is used; it makes use of a variant of the simplex algorithm.

Concerning the two-step averaged regression quantile $\widetilde{B}_n(\alpha)$, the most difficult is the first step—the computation of the R-estimator $\tilde{\boldsymbol{\beta}}_{nR}$ of the slopes. In the numerical illustration below the score function $\varphi_\lambda(u) = \lambda - I[u < \lambda]$ from (4.4) and the approximate scores (i) from (4.1) are applied. In this case the function rfit(·) from the R package Rfit ([11] version 0.23.0) could be directly used. For the rfit(·) function the score function corresponding to (4.4) and (i) of (4.1) has to be defined, i.e. at point $\lceil \lambda n \rceil/(n+1)$ attaining the value $\lceil \lambda n \rceil - 1 - \lambda(n-1) = A_n(i)$. The function rfit(·) uses the minimization routine optim(..., method="BFGS"), which is a quasi-Newton optimizer. This method works well for the simple linear regression model ($p = 1$) and fairly well if $p \ll n$ but is less precise in other cases. For a numerical illustration we refer to Problem 4.9 in [9]. So, it is better to use the fact that when employing the score function (4.4) with $\lambda = \alpha$ and the approximate scores (i) from (4.1), the slope components of the regression α-quantile and the two-step regression α-quantile coincide, $\hat{\boldsymbol{\beta}}_n(\alpha) = \tilde{\boldsymbol{\beta}}_{nR}$ for every fixed $\alpha \in (0,1)$, see [8]. Therefore, the rq(·) function from the quantreg package is then used to find the exact solution $\tilde{\boldsymbol{\beta}}_{nR}$.

The averaged regression quantile $\overline{B}_n(\alpha)$ and the two-step averaged regression quantile $\widetilde{B}_n(\alpha)$ (and their inversions) can be used as the estimates of the quantile function (and of the distribution function, respectively) of the model errors. The behavior of the proposed estimates is illustrated in the following simulation study.

The regression model (2.1) is simulated with the following parameters: sample size $n = 25$, $\beta_0 = 5$, $\boldsymbol{\beta} = (\beta_1, \beta_2) = (-3, 2)$. The columns of the regression matrix $(x_{11}, \ldots, x_{n1})^\top$ and $(x_{12}, \ldots, x_{n2})^\top$ are generated as two independent samples from the uniform distributions $U(0,4)$ and $U(-4,2)$, respectively, and are standardized so that $\sum_{i=1}^{n} x_{ij} = 0$, $j = 1, 2$. The errors e_{ni} are generated from the standard normal, standard Cauchy or generalized extreme value (GEV) distribution with the shape parameter $k = -0.5$. For each case, 10 000 replications of the model were simulated and $\overline{B}_n(\alpha)$ and $\widetilde{B}_n(\alpha)$ and their inversions were computed. For the two-step version $\widetilde{B}_n(\alpha)$, the score-generating function (4.4) with fixed $\lambda = 0.5$ or 0.9 was used. For a comparison, the empirical quantile function of the errors e_{ni} and its inversion were calculated as well. Empirical quantile estimates based on $\overline{B}_n$ and on $\widetilde{B}_n$ were then calculated and plotted. The statistical software R was used for all the calculations.

The figures showing estimates of the true quantile functions and of the true distribution functions look very similar, up to the inversion. Only the figures for the normal and Cauchy distribution functions are presented, see Figures 1–2. Other figures can be found at `http://robust.tul.cz/figures.pdf`.

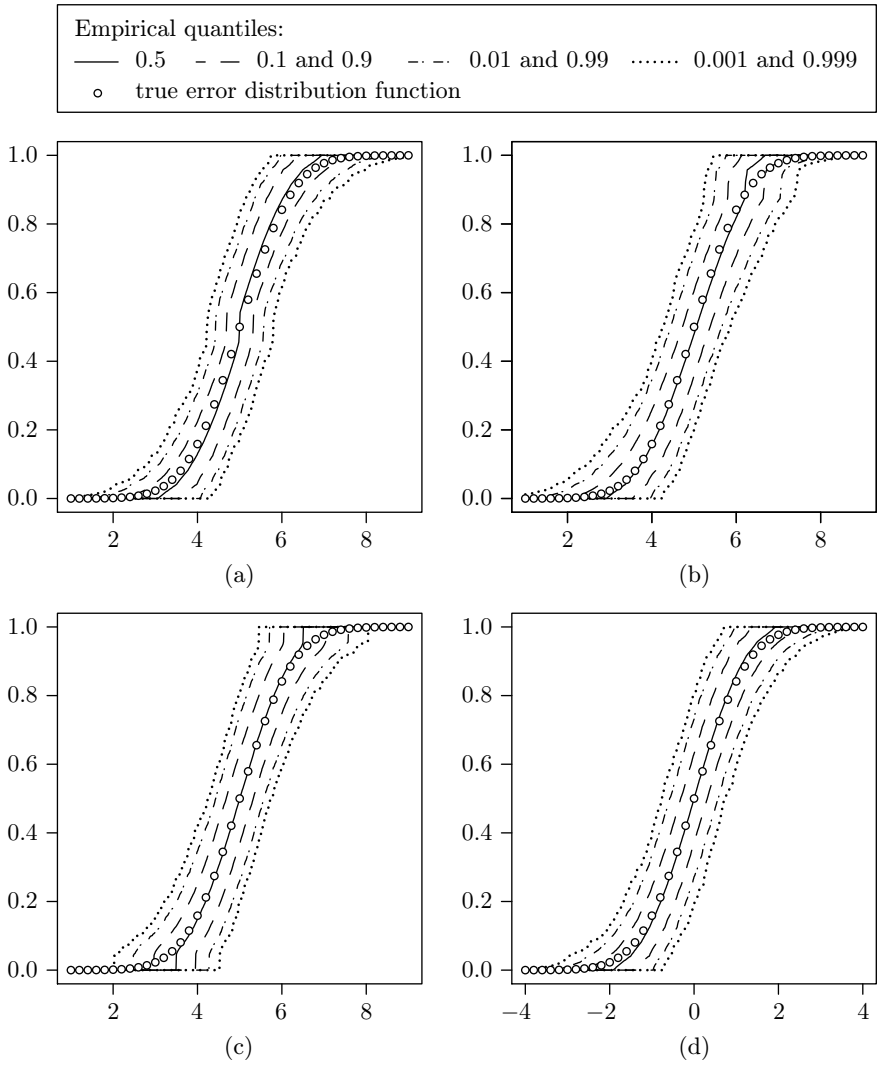

Figure 1. Empirical quantile estimates of the normal distribution function based on (a) $\widetilde{B}$ ($\lambda = 0.5$), (b) $\widetilde{B}$ ($\lambda = 0.9$), (c) $\overline{B}$, and (d) on the empirical quantile function (EQF) of errors.

The approximation of the distribution functions appears to be very good. We notice that in the case of the two-step regression quantile $\widetilde{B}_n(\alpha)$ with λ fixed, the

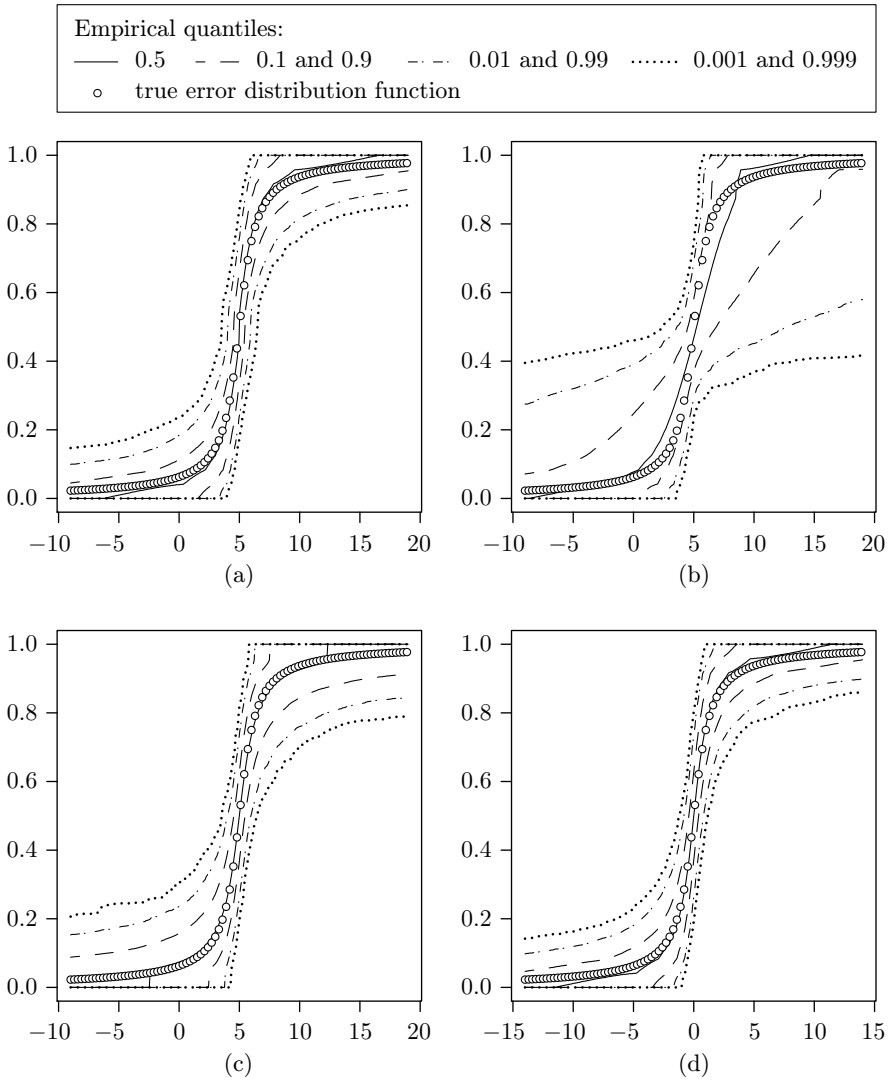

Figure 2. Empirical quantile estimates of the Cauchy distribution function based on (a) $\widetilde{B}$ ($\lambda = 0.5$), (b) $\widetilde{B}$ ($\lambda = 0.9$), (c) $\overline{B}$, and (d) on the empirical quantile function (EQF) of errors. (Vertical axis: values of the distribution f.)

quality of the estimate is sensitive to the choice of λ, especially for skewed or heavy-tailed distributions. The choice around $\lambda = 0.5$ is generally recommended.

A c k n o w l e d g e m e n t . We thank the reviewer for careful reading of our manuscript and for valuable comments.

References

[1] *G. W. Bassett,* Jr.: A property of the observations fit by the extreme regression quantiles. Comput. Stat. Data Anal. *6* (1988), 353–359.

[2] *G. W. Bassett,* Jr.*, R. Koenker*: An empirical quantile function for linear models with iid errors. J. Am. Stat. Assoc. *77* (1982), 405–415.

[3] *C. Gutenbrunner, J. Jurečková*: Regression rank scores and regression quantiles. Ann. Stat. *20* (1992), 305–330.

[4] *J. Hájek*: Extension of the Kolmogorov-Smirnov test to regression alternatives. Proceedings of the International Research Seminar (L. LeCam, ed.). University of California Press, Berkeley, 1965, pp. 45–60.

[5] *L. A. Jaeckel*: Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Stat. *43* (1972), 1449–1458.

[6] *J. Jurečková*: Averaged extreme regression quantile. Extremes *19* (2016), 41–49.

[7] *J. Jurečková, J. Picek*: Two-step regression quantiles. Sankhyā *67* (2005), 227–252.

[8] *J. Jurečková, J. Picek*: Averaged regression quantiles. Contemporary Developments in Statistical Theory (S. Lahiri et al., eds.). Springer Proceedings in Mathematics and Statistics 68, Springer, Cham, 2014, pp. 203–216.

[9] *J. Jurečková, J. Picek, M. Schindler*: Robust Statistical Methods With R. CRC Press, Boca Raton, 2019.

[10] *J. Jurečková, P. K. Sen, J. Picek*: Methodology in Robust and Nonparametric Statistics. CRC Press, Boca Raton, 2013.

[11] *J. D. Kloke, J. W. McKean*: Rfit: Rank-based estimation for linear models. The R Journal *4* (2012), 57–64.

[12] *R. Koenker*: Quantile Regression. Econometric Society Monographs 38, Cambridge University Press, Cambridge, 2005.

[13] *R. Koenker*: quantreg: Quantile Regression. R package version 5.51. Available at `https://CRAN.R-project.org/package=quantreg`.

[14] *R. Koenker, G. Bassett,* Jr.: Regression quantiles. Econometrica *46* (1978), 33–50.

[15] *S. Portnoy*: Tightness of the sequence of empiric c.d.f. processes defined from regression fractiles. Robust and Nonlinear Time Series Analysis (J. Franke et al., eds.). Lecture Notes in Statistics 26, Springer, New York, 1984, pp. 231–245.

[16] *S. Portnoy*: Asymptotic behavior of the number of regression quantile breakpoints. SIAM J. Sci. Stat. Comput. *12* (1991), 867–883.

[17] *D. Ruppert, R. J. Carroll*: Trimmed least squares estimation in the linear model. J. Am. Stat. Assoc. *75* (1980), 828–838.

Authors' addresses: *Jana Jurečková*, Department of Probability and Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic and The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Praha 8, Czech Republic, e-mail: `jurecko@karlin.mff.cuni.cz`, `jureckova@utia.cas.cz`; *Jan Picek*, *Martin Schindler* (corresponding author), Department of Applied Mathematics, Technical University, Studentská 2, 461 17 Liberec, Czech Republic, e-mail: `jan.picek@tul.cz`, `martin.schindler@tul.cz`.