Focused model selection in quantile regression

Peter Behl
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany

Gerda Claeskens
KU Leuven
ORSTAT and Leuven
Statistics Research Center
3000 Leuven, Belgium

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany

March 21, 2012

Abstract

We consider the problem of model selection for quantile regression analysis where a particular purpose of the modeling procedure has to be taken into account. Typical examples include estimation of the area under the curve in pharmacokinetics or estimation of the minimum effective dose in phase II clinical trials. A focused information criterion for quantile regression is developed, analyzed and investigated by means of a simulation study.

Keywords and Phrases: quantile regression, model selection focused information criterion
AMS Subject Classification: 62J02, 62F12

1 Introduction

Quantile regression was introduced by Koenker and Bassett (1978) as an alternative to least squares estimation and yields a far-reaching extension of regression analysis by estimating families of conditional quantile curves. Since its introduction, quantile regression has found great attraction in statistics because of its ease of interpretation, its robustness and its numerous applications which include such important areas as medicine, economics, environment modeling, toxicology or engineering [see Buchinsky (1994); Cade et al. (1999) or Wei et al. (2006) among many others]. For a detailed description of quantile regression analysis we refer to the monograph of Koenker (2005), which also provides a variety of additional examples. In a concrete application the parametric specification of a quantile regression model might be difficult and several authors have proposed nonparametric methods to investigate conditional quantiles [see Yu and Jones (1998), Dette and Volgushev (2008) and Chernozhukov

et al. (2010) among many others]. However, nonparametric methods involve the choice of a regularization parameter and for high dimensional predictors these methods are not feasible because of the curse of dimensionality. Parametric models provide an attractive alternative because they do not suffer from these drawbacks. On the other hand, in the application of these models the problem of model selection and validation is a very important issue, because a misspecification of the regression model may lead to an invalid statistical analysis. Machado (1993) considered a modification of the Schwarz (1978) criterion for general M-estimates, Ronchetti (1985) studied such a variant for the Akaike information criterion [see Akaike (1973)]. Koenker (2005) proposed to use the Akaike criterion for quantile regression, which usually overestimates the dimension but has advantages with respect to prediction. More recently, several authors have worked on penalized quantile regression in the context of variable selection in sparse quantile regression models [see Zou and Yuan (2008); Wu and Liu (2009); Shows et al. (2010)].

The work of the present paper is motivated by some recent application of nonlinear median regression with the EMAX model in pharmacokinetics [see Callies et al. (2004) or Chien et al. (2005) among others]. In studies of this type quantities such as area under the curve (AUC) or minimum effective dose (MED) are of main interest and model selection should take this into account. Example 2.1, see Section 2, is one such situation where a dose response relationship is modeled by nonlinear quantile regression and a clear target is involved. Different dose response models are considered with the specific purpose of using the selected model to estimate the minimal effective dose, i.e. the target, the minimal dose for which a specified minimum effect is achieved.

The existing variable selection methods have in common that they do not take the *purpose* of the modeling procedure into account. The focused information criterion (FIC, Claeskens and Hjort, 2003, 2008b), is especially designed to find the best model for the estimation of such a target. The criterion estimates the mean squared error (MSE) of the focus or target estimator and selects that model for which this quantity is the smallest. The FIC has been developed first for parametric likelihood models with maximum likelihood estimation, and has later been extended towards semiparametric models (Claeskens and Carroll, 2007), generalized additive partial linear models (Zhang and Liang, 2011), time series models (Claeskens et al., 2007), Cox proportional hazard regression models (Hjort and Claeskens, 2006), volatility forecasting (Brownlees and Gallo, 2008), to name a few.

Therefore, the purpose of the present paper is to develop a methodology for focused model selection in quantile regression analysis. The basic terminology is introduced in Section 2, where we also present a motivating example from a phase II dose finding study. Section 3 provides some asymptotic properties of the quantile regression estimate under local alternatives. A rigorous statement of these properties is – to the best knowledge of the authors – not available in the literature. In Section 4 we use these results to define a focused information criterion for quantile regression models. The methodology is illustrated by a small simulation study and by the analysis of a data example in Section 5. Finally, some of the more technical arguments are referred to an appendix in Section 7.

2 Preliminaries

Let $F(y|x)$ denote the conditional distribution function of a random variable Y for a given predictor x. For a given $\tau \in (0,1)$ we consider the common nonlinear quantile regression model

$$Q_\tau(x) = F^{-1}(\tau|x) = g(x;\beta),$$

where the regression function $g(x;\beta)$ depends on a q-dimensional vector of parameters $\beta := (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_q)^t \in \Theta \subset \mathbb{R}^q$ and an explanatory variable $x \in \mathcal{X}$. In order to address the problem of model selection we follow Claeskens and Hjort (2003) and assume that the specification of the parameter β generates several sub-models, where each of the sub-models contains the first part of the vector β, that is $\beta_0 := (\beta_1, \ldots, \beta_p)^t$ (Claeskens and Hjort (2003) call this the narrow model and call these parameters "protected" parameters). The following example illustrates this assumption for the class of competing models.

Example. 2.1 Consider the Hill model

$$g(x;\beta) = \beta_4 + \frac{\beta_1 x^{\beta_3}}{\beta_2^{\beta_3} + x^{\beta_3}}, \tag{2.1}$$

which is widely used in pharmacokinetics and dose response studies [for some applications see Chien et al. (2005); Park et al. (2005); Blake et al. (2008) among many others]. The "simplest" model to describe the velocity of a chemical reaction or a dose response relationship is a sub-model of (2.1) and is obtained by the choice $\beta_3 = 1$ and $\beta_4 = 0$, namely the Michaelis Menten-model

$$g(x;\beta_1,\beta_2,1,0) = \frac{\beta_1 x}{\beta_2 + x}. \tag{2.2}$$

The model (2.2) corresponds to the narrow model (note that we have $p = 2$, $q = 4$ in the general terminology). Moreover, there are several other interesting models which arise as special cases of the Hill model. A famous competitor is the EMAX model which is obtained for $\beta_3 = 1$, that is

$$g(x;\beta_1,\beta_2,1,\beta_4) = \beta_4 + \frac{\beta_1 x}{\beta_2 + x}. \tag{2.3}$$

Similarly, if no placebo effect is assumed, this can by addressed by the choice $\beta_4 = 0$, i.e.

$$g(x;\beta_1,\beta_2,\beta_3,0) = \frac{\beta_1 x^{\beta_3}}{\beta_2^{\beta_3} + x^{\beta_3}}. \tag{2.4}$$

The models (2.1) - (2.4) are frequently used for modeling dose response relationships and a typical problem in this context is to estimate the minimal effective dose (MED), that is the smallest dose level, such that a minimum effect, say Δ is achieved. In the present context this means that we are interested in the quantity $\mu(\beta) = g^{-1}(\Delta,\beta)$, which is given by

$$\Big(\frac{\beta_2^{\beta_3}(\Delta-\beta_4)}{\beta_1+\beta_4-\Delta}\Big)^{1/\beta_3}\ ,\ \frac{\beta_2\Delta}{\beta_1-\Delta}\ ,\ \frac{\beta_2(\Delta-\beta_4)}{\beta_1+\beta_4-\Delta}\ ,\ \Big(\frac{\beta_2^{\beta_3}\Delta}{\beta_1-\Delta}\Big)^{1/\beta_3}\ , \tag{2.5}$$

for the models (2.1), (2.2), (2.3) and (2.4), respectively. If this is the main goal of the experiment (which is typically the case in phase II clinical trials or in toxicological studies), model selection should take this target into account.

The aim of this paper is to derive a focused model choice criterion for quantile regression analysis, which addresses problems of this type. For this purpose we propose to choose a subset from $(\beta_{p+1}, \ldots, \beta_q)$ such that the MSE for estimating a certain focus parameter

$$\mu := \mu(\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_q) \tag{2.6}$$

by the chosen quantile regression model is minimal. In order to find this "best" model, we will determine the MSE of the estimator $\hat{\mu}_S$ for each possible sub-model, where S denotes any subset from $(\beta_{p+1}, \ldots, \beta_q)^t$. Throughout the text, β_S will denote a parameter vector for the model which includes all parameters from the narrow model plus the parameters contained in a set $S \subset \{p+1, \ldots, q\}$, that is $\beta_S = (\beta_1, \ldots, \beta_p, (\beta_j)_{j \in S})^t$. Note that $\beta_S \in \Theta_S$, where $\Theta_S \subset \mathbb{R}^{p+|S|}$ denotes the canonical projection of Θ corresponding to the parameters from the sub-model S. We will use the notation $g(x; \beta_S)$ for the model $g(x; \beta)$, which is obtained for the vector $\beta = (\beta_1, \ldots, \beta_p, \gamma_{0,S^c}, (\beta_j)_{j \in S})^t$, where for a given set S the vector $\gamma_{0,S}$ consists of the parameters of a $q-p$-dimensional vector γ_0 corresponding to the sub-model S and S^c denotes the complement of S. Here, the values of γ_0 are always chosen such that $g(x; \beta_1, \ldots, \beta_p, \gamma_0)$ gives the narrow model. For example, in a linear regression model where γ corresponds to the regression coefficients, we choose $\gamma_0 = (0, \ldots, 0)^t$, whereas in Example 2.1 where the narrow model is given by (2.2) and the full model is given by (2.1) we have $(\gamma_{0,1}, \gamma_{0,2}) = (1, 0)$. Other functions of the parameter β are interpreted in the same way if their argument is β_S. In order to emphasize that all parameters are included in the quantile regression model we use the notation $g(x; \beta_{full})$ and we also introduce the vectors

$$\begin{aligned} \beta_{0,full} &= (\beta_1, \ldots, \beta_p, \gamma_0)^t, \\ \beta_{0,S} &= (\beta_1, \ldots, \beta_p, \gamma_{0,S})^t. \end{aligned}$$

Throughout this paper let n denote the sample size and δ be a vector of dimension $q-p$. Following Claeskens and Hjort (2003) we assume that the unknown "true" parameter, say β_{true}, is of the form

$$\beta_{true} = (\beta_1, \ldots, \beta_p, \gamma_0 + \frac{\delta}{\sqrt{n}})^t. \tag{2.7}$$

If a particular quantile regression model has been specified (by the choice of an appropriate set S), the quantile regression estimate on the basis of n observations $Y_1, \ldots, Y_n$ at experimental conditions $x_1, \ldots, x_n$ is defined as the minimizer of the function

$$\sum_{i=1}^{n} \rho_\tau(Y_i - g(x_i; \beta_S)) \tag{2.8}$$

where $\rho_\tau(z) := \tau I(z \geq 0) z + (\tau - 1) I(z < 0) z$ denotes the check function [see Koenker (2005)].

3 Asymptotic properties

In this section we study the asymptotic properties of quantile regression estimates under local alternatives of the form (2.7), which are required for the derivation of a focussed information criterion for quantile regression. For this purpose we assume that the following assumptions are satisfied.

(A0) The parameter space Θ and the design space $\mathcal{X}$ are compact.

(A1) (i) $Y_1, \ldots, Y_n$ are independent random variables with densities $f_{1n}(\cdot|x_1), \ldots, f_{nn}(\cdot|x_n)$ such that for each $x \in \mathcal{X}$ the function $f_{in}(\cdot|x)$ is continuous. F_{in} denotes the corresponding distribution function, while $\tilde{f}_{in}(u) = f_{in}(u + g(x_i; \beta_{0,S})|x_i)$ is the density of the regression error $u_{i,S} := Y_i - g(x_i; \beta_{0,S})$ with corresponding distribution function $\tilde{F}_{in}$.

(ii) $f_{in}(g(x_i; \beta_{true})) \neq 0$ for $i = 1, 2, \ldots, n$; $n \in \mathbb{N}$.

(iii) The densities f_{in} are uniformly bounded by a constant $0 < K < \infty$.

(iv) The densities $\tilde{f}_{in}(u)$ are differentiable with respect to u and $|\tilde{f}'_{in}(u)| \leq K_2$ in a neighborhood of zero, where the constant K_2 does not depend on n.

(A2) $g(x; \beta_{full})$ is twice continuously differentiable with respect to the parameter vector β_{full} for all $x \in \mathcal{X}$. For a given sub-model S we denote the corresponding derivatives by

$$m(x_i, \beta_{0,S}) = \frac{\partial g(x_i; \beta_S)}{\partial \beta_S^t}\Big|_{\beta_S = \beta_{0,S}}, \quad \mathcal{M}(x_i, \beta_S^*) = \Big(\frac{\partial^2 g(x_i; \beta_S)}{\partial \beta_S \partial \beta_S^t}\Big)\Big|_{\beta_S = \beta_S^*}$$

where β_S^* is a suitable value between β_S and $\beta_{0,S} := (\beta_1, \ldots, \beta_p, \gamma_{0,S})^t$, which will be specified in the concrete applications.

(A3) (i) There exists a positive definite matrix V such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} m(x_i, \beta_{0,full}) m(x_i, \beta_{0,full})^t = V.$$

(ii) There exists a positive definite matrix Q such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_{in}(g(x_i; \beta_{true})) m(x_i, \beta_{0,full}) m(x_i, \beta_{0,full})^t = Q := \left(\begin{array}{c|c} Q_{00} & Q_{10} \\ \hline Q_{01} & Q_{11} \end{array}\right),$$

where Q_{00} is a $p \times p$-matrix which corresponds to the narrow model and Q_{11} denotes a $q \times q$-matrix corresponding to the additional parameters of the full model.

(A4) $F_{in}(g(x_i; \beta_{true})) = \tau$ for all $i = 1, \ldots, n$.

(A5) There exist constants $0 < k_1, k_2 < \infty$ such that for all $\beta_1, \beta_{0,full} \in \Theta$ and for $n > n_0$

$$k_1\|\beta_2 - \beta_{0,full}\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}[g(x_i;\beta_2) - g(x_i;\beta_{0,full})]^2 \leq k_2\|\beta_2 - \beta_{0,full}\|^2.$$

Note that the second subscript n is used here for the distribution functions F_{in} (and corresponding densities f_{in}) in order to point out that we are working under the assumption (2.7) of local alternatives. Moreover, it should be pointed out here that a similar assumption as (A5) was also used by Jureckova (1994) in order to ensure identifiability of the parameter β_0, that is

$$k_1\|\beta_2 - \beta_1\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}[g(x_i;\beta_2) - g(x_i;\beta_1)]^2 \leq k_2\|\beta_2 - \beta_1\|^2, \tag{3.1}$$

for all $\beta_1, \beta_2 \in \Theta$. However, for some important nonlinear models, this condition may not be fulfilled. A typical example is model (2.1), where we have $g(x;0,\beta_2,\beta_3,\beta_4) = \beta_4$ independent of the values of β_2 and β_3. However, for the derivation of the asymptotic results in this chapter it is actually enough to assume that (3.1) holds only for the "pseudo-true" parameter $\beta_{0,full}$, which corresponds to assumption (A5).

3.1 Consistency of the quantile regression estimator

In this section, we will prove that under the local alternatives of the form (2.7) the estimated regression quantile $\hat{\beta}_S$ in a given submodel S converges in probability to $\beta_{0,S}$. The precise statement is the following result.

Theorem. 3.1 *Assume that (A1) – (A5) and (2.7) are satisfied. For any submodel S, the statistic $\hat{\beta}_S$ is a consistent estimator for $\beta_{0,S}$, i.e.*

$$\hat{\beta}_S - \beta_{0,S} = o_P(1) \quad as \quad n \to \infty.$$

Proof. Define

$$\Delta_i(\beta_S) \;=\; g(x_i;\beta_S) - g(x_i;\beta_{0,S}), \tag{3.2}$$

then a Taylor expansion (using assumptions (A0), (A2) and (2.7)) gives

$$\Delta_i(\beta_{true}) \;=\; m(x_i,\beta_{0,full})^t\frac{\tilde{\delta}}{\sqrt{n}} + \frac{\tilde{\delta}}{\sqrt{n}}^t\frac{1}{2}\mathcal{M}(x_i,\tilde{\beta}_i)\frac{\tilde{\delta}}{\sqrt{n}} = O(n^{-1/2}), \tag{3.3}$$

where we used the notation $\tilde{\delta} = (0,\ldots,0,\delta)^t$ and $\tilde{\beta}_i$ satisfies $\|\tilde{\beta}_i - \beta_{0,full}\| \leq \|\beta_{true} - \beta_{0,full}\|$. This yields (using assumptions (A0) - (A2)) for some α satisfying $|\alpha| \leq |\Delta_i(\beta_{true})|$

$$\begin{aligned} r_{n,\tau}(x_i) \;:=\; & \tilde{F}_{in}(\Delta_i(\beta_{true})) - \tilde{F}_{in}(0) = \tilde{f}_{in}(0)\Big(m(x_i,\beta_{0,full})^t + \frac{\tilde{\delta}}{\sqrt{n}}^t\frac{1}{2}\mathcal{M}(x_i,\tilde{\beta}_i)\Big)\frac{\tilde{\delta}}{\sqrt{n}} \\ & + \frac{1}{2}\tilde{f}'_{in}(\alpha)\Big(m(x_i,\beta_{0,full})^t\frac{\tilde{\delta}}{\sqrt{n}} + \frac{\tilde{\delta}}{\sqrt{n}}^t\frac{1}{2}\mathcal{M}(x_i,\tilde{\beta}_i)\frac{\tilde{\delta}}{\sqrt{n}}\Big)^2 = O(\frac{1}{\sqrt{n}}). \end{aligned} \tag{3.4}$$

Now recall the definition of $u_{i,S}$ in (A1) and note that the estimated regression quantile $\hat{\beta}_S$ minimizes the objective function

$$Z_n(\beta_S) := \frac{1}{n}\sum_{i=1}^{n}\left[\rho_\tau(Y_i - g(x_i;\beta_S)) - \rho_\tau(u_{i,S})\right]. \tag{3.5}$$

We first calculate the expectation of $E[Z_n(\beta_S)]$ as

$$\begin{aligned}
E[Z_n(\beta_S)] &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}}\left[(\tau - \mathbb{1}_{\{s\le\Delta_i(\beta_S)\}})(s-\Delta_i(\beta_S)) + (\mathbb{1}_{\{s\le 0\}} - \tau)s\right] d\tilde{F}_{in}(s) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big\{ -\int_{-\infty}^{\Delta_i(\beta_S)} s\, d\tilde{F}_{in}(s) + \int_{-\infty}^{0} s\, d\tilde{F}_{in}(s) + \Delta_i(\beta_S)\tilde{F}_{in}(\Delta_i(\beta_S)) - \tau\Delta_i(\beta_S)\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big\{ \int_{\Delta_i(\beta_S)}^{0} s\, d\tilde{F}_{in}(s) + \Delta_i(\beta_S)(\tilde{F}_{in}(\Delta_i(\beta_S)) - \tilde{F}_{in}(0)) \\
&\qquad + \Delta_i(\beta_S)(\tilde{F}_{in}(0) - \tilde{F}_{in}(\Delta_i(\beta_{true})))\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\Delta_i(\beta_S)}^{0}(s-\Delta_i(\beta_S))\, d\tilde{F}_{in}(s) + O(\frac{1}{\sqrt{n}}),
\end{aligned} \tag{3.6}$$

where the last identity follows from (3.4). Note that the integral in the last line is always positive, except in the case $\Delta_i(\beta_S) = 0$ which corresponds to the choice $\beta_S = \beta_{0,S}$. Furthermore, the identifiability assumption (A5) guarantees that for sufficiently large n and any parameter $\beta_S \in \Theta_S$ different from $\beta_{0,S}$ we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\int_{\Delta_i(\beta_S)}^{0}(s-\Delta_i(\beta_S))\, dF_{ni}(s)\right) > 0. \tag{3.7}$$

This implies that for sufficiently large n the sum in (3.6) will only be zero for $\beta_S = \beta_{0,S}$ and will be strictly positive otherwise. The key step for completing the proof is a uniform convergence property of the criterion function. More precisely, we will show in the Appendix that

$$\sup_{\beta_S\in\Theta_S}|Z_n(\beta_S) - E[Z_n(\beta_S)]| \xrightarrow{P} 0. \tag{3.8}$$

Because Z_n is minimized at $\hat{\beta}_S$, we have

$$Z_n(\hat{\beta}_S) \le Z_n(\beta_{0,S}) = 0. \tag{3.9}$$

Then from (3.7), (3.8) and (3.9) follows the statement of the Theorem, i.e. $\|\hat{\beta}_S - \beta_{0,S}\| = o_P(1)$. □

3.2 Weak convergence under local alternatives

In this section we derive the asymptotic distribution of the quantile regression estimator $\hat{\beta}_S$ for each sub-model S under local alternatives of the form (2.7), which is the key step for defining the FIC in every sub-model.

Theorem. 3.2 *Under assumptions (A1) - (A5) and (2.7) we have*

$$\sqrt{n}(\hat{\beta}_S - \beta_{0,S}) \xrightarrow{D} N_S \sim \mathcal{N}\Big(Q_S^{-1}\begin{pmatrix} Q_{01} \\ \pi_S Q_{11}\end{pmatrix}\delta, \tau(1-\tau)Q_S^{-1}V_S Q_S^{-1}\Big),$$

where $\mathcal{N}(\mu,\Sigma)$ denotes a normal distribution with mean μ and covariance matrix Σ,

$$Q_S = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f_{in}(g(x_i;\theta_{0,S})m(x_i,\beta_{0,S})m(x_i,\beta_{0,S})^t,$$

$$V_S = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m(x_i,\beta_{0,S})m(x_i,\beta_{0,S})^t,$$

and π_S is a $|S|\times p$-projection matrix consisting of ones and zeros which simply extracts from Q_{11} the rows corresponding to the sub-model S.

Proof. By a Taylor expansion at the point $\beta_{0,S}$, the quantity Δ_i defined in (3.2) can be written in terms of $v := \sqrt{n}(\beta_S - \beta_{0,S})$:

$$\Delta_i(v) \;=\; \frac{1}{\sqrt{n}}m(x_i,\beta_{0,S})^t v + \frac{1}{2n}v^t\mathcal{M}(x_i,\beta_S^*)v, \tag{3.10}$$

where $\beta_S, \beta_S^* \in \Theta_S$ satisfy $\|\beta_S^* - \beta_{0,S}\| \le \|\beta_S - \beta_{0,S}\|$. Inserting $\Delta_i(v)$ into the definition (3.5), we obtain the slightly modified objective function

$$\begin{aligned} G_n(v) &:= \sum_{i=1}^{n}\Big[\mathbb{1}_{\{u_{i,S}\le 0\}}(1-\tau)\Delta_i(v) - \mathbb{1}_{\{u_{i,S}>0\}}\tau\Delta_i(v) \\ &\quad + \mathbb{1}_{\{0<u_{i,S}\le\Delta_i(v)\}}(\Delta_i(v)-u_{i,S}) + \mathbb{1}_{\{\Delta_i(v)\le u_{i,S}\le 0\}}(u_{i,S}-\Delta_i(v))\Big] \\ &= -v^t\Gamma_{n,S} + \sum_{i=1}^{n} b_i(v), \end{aligned} \tag{3.11}$$

where the random variables $\Gamma_{n,S}$ and $b_i(v)$ are defined by

$$\begin{aligned} \Gamma_{n,S} &:= \sum_{i=1}^{n}\psi_\tau(u_{i,S})\left[\frac{1}{\sqrt{n}}m(x_i,\beta_{0,S}) + \frac{1}{2n}\mathcal{M}(x_i,\beta_{i,S}^*)v\right], \\ b_i(v) &:= \mathbb{1}_{\{0<u_{i,S}\le\Delta_i(v)\}}(\Delta_i(v)-u_{i,S}) + \mathbb{1}_{\{\Delta_i(v)\le u_{i,S}\le 0\}}(u_{i,S}-\Delta_i(v)), \end{aligned} \tag{3.12}$$

respectively, and

$$\psi_\tau(u_{i,S}) \;:=\; \tau\mathbb{1}_{\{u_{i,S}\ge 0\}} + (\tau-1)\mathbb{1}_{\{u_{i,S}<0\}},$$

denotes the "derivative" of the check function ρ_τ. Note that $G_n(v)$ is now minimized at $\hat{T}_n := \sqrt{n}(\hat{\beta}_S - \beta_{0,S})$. In the Appendix we will derive the following asymptotic properties for the terms in this expansion.

- For any $v \in \mathbb{R}^{p+|S|}$ we have

$$v^t \Gamma_{n,S} \xrightarrow{D} v^t W_S, \tag{3.13}$$

where

$$W_S \sim \mathcal{N}\Big(\begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta, \tau(1-\tau) V_S \Big).$$

- For any $v \in \mathbb{R}^{p+|S|}$ we have

$$\sum_{i=1}^{n} b_i(v) = \frac{1}{2} v^t Q_{n,S} v + O(n^{-1/2}\|v\|^3) + O_P(n^{-1/6}\|v\|^{3/2}) + O(n^{-1}\|v\|^4), \tag{3.14}$$

where

$$Q_{n,S} = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{in}(0) m(x_i, \beta_{0,S}) m(x_i, \beta_{0,S})^t.$$

Note that (3.14) provides a quadratic approximation of the function

$$G_n(v) = -v^t \Gamma_{n,S} + \frac{1}{2} v^t Q_{n,S} v + O(n^{-1/2}\|v\|^3) + O_P(n^{-1/6}\|v\|^{3/2}) + O(n^{-1}\|v\|^4),$$

which will be used to establish a Bahadur-type representation for the statistic $\hat{T}_n = \sqrt{n}(\hat{\beta}_S - \beta_{0,S})$. More precisely, we will show in the Appendix that $\hat{T}_n$ is stochastically bounded, that is

$$\|\hat{T}_n\| = O_P(1). \tag{3.15}$$

Therefore $G_n(\hat{T}_n)$ has the following stochastic expansion

$$G_n(\hat{T}_n) = -\hat{T}_n^t \Gamma_{n,S} + \frac{1}{2} \hat{T}_n^t Q_{n,S} \hat{T}_n + o_P(1). \tag{3.16}$$

By (3.13) the term $U_n := Q_{n,S}^{-1} \Gamma_{n,S}$ is asymptotically normal distributed. In particular $\|U_n\|$ is also stochastically bounded and satisfies $U_n^t \Gamma_{n,S} = U_n^t Q_{n,S} U_n$, which yields

$$G_n(U_n) \;=\; -\frac{1}{2} U_n^t Q_{n,S} U_n + o_P(1). \tag{3.17}$$

From (3.16) and (3.17) it therefore follows that

$$\begin{aligned} G_n(\hat{T}_n) - G_n(U_n) &= -\hat{T}_n^t \Gamma_{n,S} + \frac{1}{2} \hat{T}_n^t Q_{n,S} \hat{T}_n + \frac{1}{2} U_n^t Q_{n,S} U_n + o_P(1) \\ &= \frac{1}{2} (\hat{T}_n - U_n)^t Q_{n,S} (\hat{T}_n - U_n) + o_P(1). \end{aligned} \tag{3.18}$$

Note that by the definition of $\hat{T}_n$ the left-hand-side of the above equation is always non-positive, while the first term in the last row on the right-hand-side is always positive due to the positive definiteness of $Q_{n,S}$. Consequently, we obtain $\|\hat{T}_n - U_n\| = o_P(1)$, i.e.

$$\hat{T}_n = \sqrt{n}(\hat{\beta}_S - \beta_{0,S}) = U_n + o_P(1) = Q_{n,S}^{-1} \Gamma_{n,S} + o_P(1).$$

Therefore the asymptotic normality of $\hat{T}_n$ directly follows from (3.13). □

4 The FIC for quantile regression

From Theorem 3.2, an expression for the FIC can be derived by similar arguments as in Claeskens and Hjort (2003). By applying the Delta method we get the asymptotic distribution of the estimator $\hat{\mu}_S$ in the submodel S

$$\sqrt{n}(\hat{\mu}_S - \mu_{true}) = \sqrt{n}(\mu(\hat{\beta}_S) - \mu(\beta_{0,S})) + \sqrt{n}(\mu(\beta_{0,S}) - \mu(\beta_{true})) \xrightarrow{\mathcal{D}} N_S - \frac{\partial \mu}{\partial \beta_{full}}^t \tilde{\delta} \quad (4.1)$$

with

$$N_S \sim \mathcal{N}\Big(\frac{\partial \mu}{\partial \beta_S}^t Q_S^{-1} \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta, \frac{\partial \mu}{\partial \beta_S}^t \tau(1-\tau) Q_S^{-1} V_S Q_S^{-1} \frac{\partial \mu}{\partial \beta_S}\Big) \quad (4.2)$$

and $\tilde{\delta} = (0, ..., 0, \delta)^t$. Here as well as in the following steps, all partial derivatives $\frac{\partial \mu}{\partial \beta_{full}}$ and $\frac{\partial \mu}{\partial \beta_S}$ are evaluated at $\beta = \beta_{0,full}$ and $\beta = \beta_{0,S}$, respectively. This yields for the MSE of (4.1)

$$\begin{aligned} \text{MSE}_S &= \frac{\partial \mu}{\partial \beta_S}^t Q_S^{-1} \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta\delta^t \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix}^t (Q_S^{-1})^t \frac{\partial \mu}{\partial \beta_S} - 2\frac{\partial \mu}{\partial \beta_S}^t Q_S^{-1} \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta \frac{\partial \mu}{\partial \beta_{full}}^t \tilde{\delta} \\ &\quad + \Big(\frac{\partial \mu}{\partial \beta_{full}}^t \tilde{\delta}\Big)^2 + \frac{\partial \mu}{\partial \beta_S}^t \tau(1-\tau) Q_S^{-1} V_S Q_S^{-1} \frac{\partial \mu}{\partial \beta_S}. \end{aligned}$$

Because the third term in this expression does not depend on the particular sub-model we finally define the FIC for the quantile regression estimator as

$$\begin{aligned} FIC_S &= \frac{\partial \mu}{\partial \beta_S}^t \Big[Q_S^{-1} \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta\delta^t \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix}^t (Q_S^{-1})^t + \tau(1-\tau) Q_S^{-1} V_S Q_S^{-1} \Big] \frac{\partial \mu}{\partial \beta_S} \\ &\quad -2\frac{\partial \mu}{\partial \beta_S}^t Q_S^{-1} \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta \frac{\partial \mu}{\partial \beta_{full}}^t \tilde{\delta}. \end{aligned} \quad (4.3)$$

It remains to estimate the unknown quantities in this expression such that the FIC can be calculated from the data. The key step here is to find an estimator of the matrices Q_S which is consistent under local alternatives. Using the regression "errors"

$$\hat{\epsilon}_i = Y_i - g(x_i; \hat{\beta}_1, \ldots, \hat{\beta}_q),$$

($\hat{\beta}_1, \ldots, \hat{\beta}_q$ are estimated in the full model) Kim and White (2003) suggested to estimate the matrix Q_S by

$$\hat{Q}_S = \frac{1}{2\hat{c}_n n} \sum_{i=1}^n \mathbb{1}_{\{-\hat{c}_n \le \hat{\epsilon}_i \le \hat{c}_n\}} m(x_i, \hat{\beta}_{0,S}) m(x_i, \hat{\beta}_{0,S})^t.$$

Here $\hat{c}_n$ denotes the bandwidth of the estimator which is in some way (e.g. by cross-validation) determined from the data. The other terms in (4.3) can be estimated similarly as in Claeskens and Hjort (2003), e.g.

$$\hat{V}_S = \frac{1}{n} \sum_{i=1}^n m(x_i, \hat{\beta}_{0,S}) m(x_i, \hat{\beta}_{0,S})^t.$$

Finally, we have to estimate the term $\delta\delta^t$. By Theorem 3.2 we have shown that

$$D_n := \sqrt{n}((\hat{\beta}_{p+1} - \gamma_{0,1}), \ldots, (\hat{\beta}_q - \gamma_{0,p-q}))^t \xrightarrow{D} D \sim \mathcal{N}(\delta, K),$$

where K denotes the $(q-p) \times (q-p)$-matrix obtained by taking the last $q-p$ rows and columns from the matrix $\tau(1-\tau)Q^{-1}VQ^{-1}$. Therefore, DD^t has mean $\delta\delta^t + K$, and, following Claeskens and Hjort (2003), we propose to use the estimator

$$\hat{\delta\delta^t} = DD^t - \hat{K},$$

which should be truncated to zero when the result is negative definite. An estimator $\hat{K}$ can be obtained directly by taking the corresponding rows and columns of $\tau(1-\tau)\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}$ of the estimated covariance matrix of the full model. Finally, the derivatives of μ can be estimated by plug-in-estimators, using estimates for $\hat{\beta}_{0,S}$ from the full model.

5 Finite sample properties

5.1 Simulation results

In this section, we present a small simulation study investigating the finite sample properties of the FIC for the class of quantile regression models introduced in Example 2.1. All results are based on 2000 simulation runs. For the distribution of the "error" $\epsilon = Y - g(x, \eta)$ we assume three scenarios: a normal distribution with mean 0 and variance $\sigma^2 = 0.01$, a Cauchy distribution with location parameter $a = 0$ and scale parameter $b = 0.07$, and in a third setting we address the problem of heteroscedasticity. Here we assume that the errors are normal distributed with mean 0 and standard deviation (depending on the explanatory variable x)

$$\sigma(x) = \tau_0 + \frac{\tau_1}{1 + e^{-\tau_2 x}}, \tag{5.1}$$

where $\tau_0 = -0.1$, $\tau_1 = 0.24$ and $\tau_2 = 0.15$. This variance function is proposed by Lim et al. (2010) for dose-response-modeling. In our first example we consider the case of two competing models, the Michaelis-Menten-model defined in (2.2) and the Hill model without intercept given by (2.4). We generate data from the model (2.4) with parameter values $\beta_1 = 0.417$, $\beta_2 = 25$ and $\beta_3 = 1.75$. As experimental design we choose six different dose levels equidistantly over the dose range $[0\text{mg}, 150\text{mg}]$ and assign 32 observations to each dose level. The parameters are estimated using median regression (i.e. $\tau = 0.5$). From these results we obtain a robust estimate for the focus parameter μ, the minimal effective dose (MED) defined in (2.5) with $\Delta = 0.1$. We investigate the performance of the FIC for choosing between the Michaelis-Menten-model (2.2) and the Hill model (2.4). In each step, we calculate the FIC for both models with model (2.2) as the narrow model. An estimator $\hat{\mu}$ is then obtained from the model with the lowest FIC value. In order to compare the FIC to more conventional information measures such as AIC and BIC, in each replication step

we also estimate μ using the model selected by AIC and BIC. In the median regression case, the AIC and BIC for the candidate model S are obtained as

$$AIC_S = n\log(\hat{\sigma}) + p, \qquad BIC_S = n\log(\hat{\sigma}) + \frac{1}{2}p\log(n),$$

where $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}|y_i - g(x_i;\hat{\beta}_S)|$, p denotes the number of parameters in the model S and n the number of observations [for details see Hurvich and Tsai (1990)]. For a comparison of the different model selection procedures we compute the absolute errors of the post-selection-estimators for μ

$$|\hat{\mu}_{FIC,i} - \mu_{true}|, \quad |\hat{\mu}_{AIC,i} - \mu_{true}|, \quad |\hat{\mu}_{BIC,i} - \mu_{true}|, \tag{5.2}$$

where $\hat{\mu}_{FIC}, \hat{\mu}_{AIC}$ and $\hat{\mu}_{BIC}$ denote the estimators of the focus μ, where the model has been chosen by FIC, AIC and BIC, respectively. Finally, it is counted how many times in 2000 simulation runs the FIC obtains a better estimator then AIC (FIC<AIC) and BIC (FIC<BIC) and vice-versa. The first row of Table 1 shows the results for homoscedastic normal distributed errors, the second row displays the results for heteroscedastic errors with variance function (5.1) and the third row shows the results for Cauchy-distributed errors.

ε_i	FIC<AIC	FIC=AIC	AIC<FIC	FIC<BIC	FIC=BIC	AIC<BIC
$\mathcal{N}(0, 0.01)$	657	1085	258	1149	475	376
$\mathcal{N}(0, \sigma^2(x_i))$	579	1222	199	1176	469	355
$\mathcal{C}(0, 0.07)$	1062	571	367	1200	304	496

Table 1: *Comparison of the absolute error of the estimate of the MED, where the model is chosen by FIC and AIC (left part) and FIC and BIC (right part). Data are generated from model* (2.4) *and the models* (2.2) *and* (2.4) *are compared for estimating the MED. The symbol FIC<AIC means that the absolute deviation of the estimate of the focus is smaller for the FIC criterion as for AIC and the symbols FIC=AIC, AIC<FIC, FIC<BIC etc. are interpreted in the same way.*

	Median			MAD		
	FIC	AIC	BIC	FIC	AIC	BIC
$\mathcal{N}(0, 0.01)$	1.76	1.96	3.12	1.04	1.22	1.53
$\mathcal{N}(0, \sigma^2(x_i))$	3.62	4.29	5.24	1.96	1.92	1.38
$\mathcal{C}(0, 0.07)$	4.25	5.70	5.73	2.43	0.98	0.96

Table 2: *Median and median absolute deviation of the absolute errors of the estimates of the focus μ obtained from the FIC-, AIC- and BIC-criterion.*

In this example it is clearly seen that in the majority of cases the FIC selects a model which is better than the model chosen by AIC and BIC. Roughly speaking, FIC finds a better model than BIC in more than half of the simulation runs for all considered error distributions. In

Table 2 we display the median and median absolute deviation of the absolute errors (5.2) of estimators obtained from the different model selection procedures. We observe that the absolute error of FIC is the smallest, while the BIC yields the largest median of the absolute errors. The differences are substantial. For the MAD the situation is not so clear. While the FIC yields the smallest MAD for homoscedastic normal distributed errors, the BIC is superior in the case of the Cauchy distribution.
In the second example we consider a larger class of models including models (2.1) and (2.3) so that model selection is now performed for the four models (2.1) - (2.4). Data are again generated from model (2.4) and the errors are assumed to have the same distributions as in the first example. The experimental design is also identical to the design from the first example. The focus parameter μ is the MED defined by equations (2.5) for $\Delta = 0.1$. Similarly as in the previous example, the Michaelis-Menten-model (2.2) is the narrow model and we count how many times the FIC selects a better model than AIC (FIC<AIC) and BIC (FIC<BIC) and vice-versa. It is seen from Table 3 that also in this example the FIC chooses a better model for the estimation of the MED than AIC or BIC in a large number of cases. It yields twice more often a smaller error than the AIC. Compared to the BIC the superiority of the FIC is even more substantial, in particular under heteroscedasticity.

ε_i	FIC<AIC	FIC=AIC	AIC<FIC	FIC<BIC	FIC=BIC	BIC<FIC
$\mathcal{N}(0, 0.01)$	813	838	349	962	602	436
$\mathcal{N}(0, \sigma^2(x_i))$	395	1418	187	1063	677	260
$\mathcal{C}(0, 0.07)$	835	828	337	988	613	399

Table 3: *Comparison of the absolute error of the estimate of the MED, where the model is chosen by FIC and AIC (left part) and FIC and BIC (right part). Data are generated from model* (2.4) *and models* (2.1) - (2.4) *are compared for estimating the MED. The symbol FIC<AIC means that the absolute deviation of the estimate of the focus is smaller for the FIC criterion as for AIC and the symbols FIC=AIC, AIC<FIC, FIC<BIC etc. are interpreted in the same way.*

Finally, we have also generated data for the case where the true model is the Michaelis-Menten model defined in (2.2) with $\beta_1 = 0.417$ and $\beta_2 = 25$. The corresponding results are displayed in Table 4, where models (2.1) - (2.4) are considered as competitors. While in this case the FIC and AIC have similar properties for normal distributed errors (for iid errors, the FIC seems to be somewhat better), the FIC performs worse than AIC and BIC for Cauchy distributed errors. In particular, BIC is better than FIC for all error distributions. However, taking into account that the FIC performs substantially better than both AIC and BIC in the case where the Hill model is the "true" model we recommend the FIC if the minimum effective dose has to be estimated with one of the models (2.1) - (2.4).

ε_i	FIC<AIC	FIC=AIC	AIC<FIC	FIC<BIC	FIC=BIC	BIC<FIC
$\mathcal{N}(0, 0.01)$	655	797	548	389	796	815
$\mathcal{N}(0, \sigma^2(x_i))$	876	217	907	881	82	1037
$\mathcal{C}(0, 0.07)$	364	905	731	317	908	775

Table 4: *Comparison of the absolute error of the estimate of the MED, where the model is chosen by FIC and AIC (left part) and FIC and BIC (right part). Data are generated from the Michaelis Menten model* (2.2) *and models* (2.2) - (2.4) *are compared for estimating the MED. The symbol FIC<AIC means that the absolute deviation of the estimate of the focus is smaller for the FIC criterion as for AIC and the symbols FIC=AIC, AIC<FIC, FIC<BIC etc. are interpreted in the same way.*

5.2 Application of the FIC in a clinical dose response study

For an empirical illustration we consider a data example from a dose response study, which has recently been investigated by Callies et al. (2004). Zosuquidar is an inhibitor of P-glycoprotein which is administered in combination with chemotherapeutic agents in order to increase tumor cell exposure to chemotherapy. In this study median regression is used to estimate the relationship between the plasma concentration of Zosuquidar and the percentage of P-glycoprotein inhibition [for details see Callies et al. (2004)]. As a consequence the intercept β_4 in model (2.1) is zero, so that either the Michaelis Menten model (2.2) or the Hill model with no intercept (2.4) are candidates to describe the dose response relationship. The focus parameter in question is the IC_{90}, the dose where 90% of maximum P-glycoprotein inhibition are realized, that is $\Delta = 90$. Figure 1 shows the data, the fitted median regression curves and the location of the IC_{90} for both models. We observe substantial differences

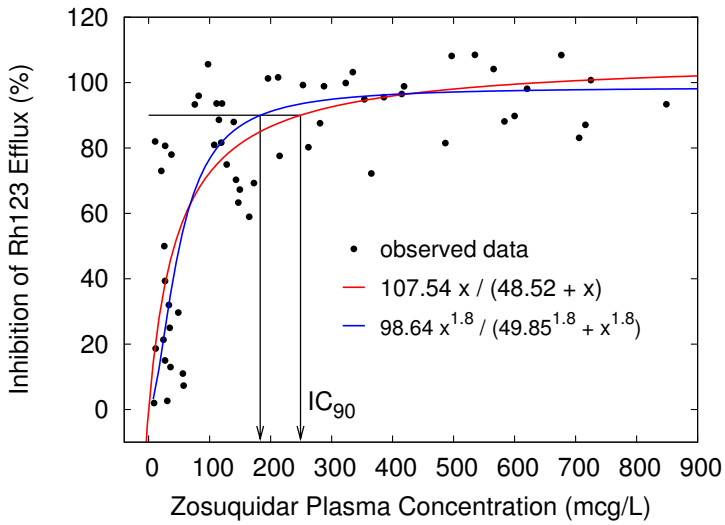

Figure 1: *Zosuquidar data with estimated median regression curves from the Hill and Michaelis Menten model.*

between the estimates of the IC_{90} obtained from the two models and therefore model selection for estimating the IC_{90} is of importance in this study. We use the FIC to decide whether the Hill slope β_3 is included in the model or not. The resulting FIC values are $1.21 \cdot 10^7$ for the Michaelis Menten model (2.2) and $4.38 \cdot 10^6$ for model (2.4). Thus, the IC_{90} is estimated using the Hill model with no intercept, which gives a value of $\hat{IC}_{90} = 183.19$. Finally we note that the AIC also selects the Hill model with no intercept in this example, while BIC favors the Michaelis Menten model with only two parameters.

6 Discussion

The work in this paper was motivated by the problem of selecting a model to determine the minimal effective dose in a dose response study on the basis of median regression analysis. For this purpose we have extended the available theory for estimation under local misspecification from a likelihood setting towards quantile regression models and developed a focused information criterion (FIC), which takes the specific target of the statistical analysis into account for the process of model selection. Simulation studies demonstrate that this way of selection indeed often results in estimators of the effective dose with smaller error than those obtained by standard selection methods such as AIC and BIC. The presented focused information criterion in this paper is applicable in more generality in nonlinear quantile regression models, hence not only for minimal effective dose determination.

In general the focus might depend also on the particular covariate information x, hence $\mu = \mu(\beta; x)$. In such cases, the derived FIC expression is specific to the given value of x, and 'subject-specific' model searches could be performed. When this level of detail is not wanted, a model search may be performed keeping the idea of the focus as in the present paper, though averaging the risk function over a wanted domain of values for the covariate x (e.g. when x represents the age, one could consider a range of values (20, 60), or one could perform the selection for all women in the dataset, or for all treated patients in a clinical trial, etc.). Claeskens and Hjort (2008a) work this out for the class of generalized linear models. Instead of considering the MSE at one particular covariate value, one could consider the loss function for model S in the following way

$$L_n(S) = n \int \{\hat{\mu}_S(\beta; x) - \mu_{\text{true}}(\beta; x)\}^2 \, dW_n(x),$$

where a weight function W_n determines a distribution of relevant x values, which might for example be an empirical distribution over the observed sample. A similar idea could be applied in this setting of nonlinear quantile estimation.

Another interesting topic for future research could be a study of asymptotic properties of the estimators under a different local misspecification setting than (2.7) by no longer assuming misspecification at the coefficient level, but rather at the level of the density functions. This line of thought is explained for likelihood regression models in Claeskens and Hjort (2003, Section 8) where it is assumed that

$$f_{\text{true}}(y) = f(y; \theta_0, \gamma_0)\{1 + r(y)/\sqrt{n}\} + o(1/\sqrt{n}),$$

for some function $r(\cdot)$ that satisfies

$$\int f(y;\theta_0,\gamma_0)|r(y)|dy < \infty \quad \text{and} \quad \int f(y;\theta_0,\gamma_0)r(y)dy = 0.$$

It is expected that theoretical properties similar to those in the present paper can be developed for such a situation.

7 Appendix: Proof of technical results

7.1 Proof of (3.8)

The proof of the uniform convergence property can be established using results of Liese and Vajda (1994), who presented general conditions for consistency of M-estimators and the uniform convergence of the corresponding objective functions. However, we still have to keep in mind that we work under local alternatives of the form (2.7). We begin with a proof of the following properties, which will be used later to establish uniform convergence of the objective function:

(B1) The class of functions $\{g(x;\beta_S)|\beta_S \in \Theta_S\}$ is equicontinuous on Θ_S for all $x \in \mathcal{X}$.

(B2) $|Z_n(\beta_S) - E[Z_n(\beta_S)]| \xrightarrow{P} 0$ for any $\beta_S \in \Theta_S$.

First, the equicontinuity (B1) is implied by assumptions (A0) and (A2) since, by the boundedness of $m(x,\beta)$ we have for $x \in \mathcal{X}$

$$|g(x;\beta_{S,1}) - g(x;\beta_{S,2})| \;=\; \left|m(x,\beta^*)^t(\beta_{S,1} - \beta_{S,2})\right| \le C\|\beta_{S,1} - \beta_{S,2}\|,$$

for some constant $C > 0$ (here β^* denotes a suitable value between $\beta_{S,1}$ and $\beta_{S,2}$).
For a proof of (B2) we introduce the notation

$$\begin{aligned} z_i(\beta_S) &= \rho_\tau(Y_i - g(x_i;\beta_S)) - \rho_\tau(u_{i,S}) \\ &= (\tau - \mathbb{1}_{\{u_{i,S}<\Delta_i(\beta_S)\}})(u_{i,S} - \Delta_i(\beta_S)) + (\mathbb{1}_{\{u_{i,S}<0\}} - \tau)u_{i,S} \\ &= \mathbb{1}_{\{u_{i,S}\le 0\}}(1-\tau)\Delta_i(\beta_S) - \mathbb{1}_{\{u_{i,S}>0\}}\tau\Delta_i(\beta_S) \\ &\quad + \mathbb{1}_{\{0<u_{i,S}\le\Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S}) + \mathbb{1}_{\{\Delta_i(\beta_S)\le u_{i,S}\le 0\}}(u_{i,S} - \Delta_i(\beta_S)). \end{aligned} \tag{7.1}$$

which gives

$$\begin{aligned} z_i(\beta_S)^2 &= \mathbb{1}_{\{u_{i,S}\le 0\}}(1-\tau)^2\Delta_i^2(\beta_S) + \mathbb{1}_{\{u_{i,S}>0\}}\tau^2\Delta_i^2(\beta_S) \\ &\quad + \mathbb{1}_{\{0<u_{i,S}\le\Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S})^2 + \mathbb{1}_{\{\Delta_i(\beta_S)\le u_{i,S}\le 0\}}(u_{i,S} - \Delta_i(\beta_S))^2 \\ &\quad - 2\mathbb{1}_{\{0<u_{i,S}\le\Delta_i(\beta_S)\}}\tau\Delta_i(\beta_S)(\Delta_i(\beta_S) - u_{i,S}) \\ &\quad + 2\mathbb{1}_{\{\Delta_i(\beta_S)\le u_{i,S}\le 0\}}(1-\tau)\Delta_i(\beta_S)(u_{i,S} - \Delta_i(\beta_S)). \end{aligned} \tag{7.2}$$

Taking e.g. the expectation of the third term in the above sum yields

$$E\left[\mathbb{1}_{\{0<u_{i,S}\leq\Delta_i(\beta_S)\}}(\Delta_i(\beta_S)-u_{i,S})^2\right]=\int_0^{\Delta_i(\beta_S)}(\Delta_i(\beta_S)-s)^2\tilde{f}_{in}(s)ds\leq K\left|\Delta_i(\beta_S)^3\right|.$$

which is bounded because $\Delta_i(\beta_S)$ is bounded by assumptions (A0) and (A2). Since the expectations of all other terms in the sum (7.2) can be similarly bounded, we obtain that $E[z_i(\beta_S)^2]$ is bounded (uniformly with respect to $i=1,\ldots,n$). Therefore it follows from Chebychev's inequality that

$$P(|Z_n(\beta_S)-E[Z_n(\beta_S)]|>\epsilon) \;\leq\; \frac{\max_{i=1,\ldots,n}E[z_i(\beta_S)^2]}{n\epsilon^2} \;=\; o(1),$$

which establishes (B2).

The uniform convergence in (3.8) can now be derived from (B1) and (B2) using similar arguments as presented in Liese and Vajda (1994). To be precise define $\delta_n(\beta_S):=Z_n(\beta_S)-E[Z_n(\beta_S)]$ and observe that for $\beta_{S,1},\beta_{S,2}\in\Theta_S$

$$|\delta_n(\beta_{S,1})-\delta_n(\beta_{S,2})| \;\leq\; \frac{2c}{n}\sum_{i=1}^n|g(x_i;\beta_1)-g(x_i;\beta_2)|,$$

which follows from the Lipschitz continuity of the check function. Therefore (B1) yields for any $\epsilon>0$ the existence of a $\delta>0$ such that for every $\beta^*\in\Theta_S$,

$$\sup_{\{\beta_S:|\beta_S-\beta^*|<\delta\}}|\delta_n(\beta_S)|\leq|\delta_n(\beta^*)|+\epsilon/2,\quad n\in\mathbb{N}.$$

By the compactness of Θ_S there exist finitely many points $\beta_1,\ldots,\beta_K\in\Theta_S$ such that

$$\sup_{\beta_S\in\Theta_S}|\delta_n(\beta_S)|\leq|\delta_n(\beta_i)|+\epsilon/2,\quad n\in\mathbb{N},$$

for some $i\in 1,\ldots,k$. As a consequence, we have

$$\begin{aligned}\lim_{n\to\infty}P(\sup_{\beta_S\in\Theta_S}|\delta_n(\beta_S)|>\epsilon) \;&\leq\; \lim_{n\to\infty}P(\max_{1\leq i\leq k}|\delta_n(\beta_i)|+\epsilon/2>\epsilon)\\ &=\; \lim_{n\to\infty}P(\max_{1\leq i\leq k}|\delta_n(\beta_i)|>\epsilon/2)=0.\end{aligned}$$

where the last equation follows from (B2), which implies (3.8).

7.2 Proof of (3.13)

To simplify notation, put $c_i(v):=\psi_\tau(u_{i,S})[\frac{1}{\sqrt{n}}v^tm(x_i,\beta_{0,S})+\frac{1}{2n}v^t\mathcal{M}(x_i,\beta^*_{i,S})v]$, so that we have $v^t\Gamma_{n,S}=\sum_{i=1}^n c_i(v)$. Recall the definition of $\tilde{F}$ and $\tilde{f}$ in assumption (A1) then a straightforward calculation yields

$$\begin{aligned}E[\psi_\tau(u_{i,S})] \;&=\; \tau(1-\tilde{F}_{in}(0))+(\tau-1)\tilde{F}_{in}(0)\\ &=\; F_i(g(x_i;\beta_{true}))-\tilde{F}_{in}(0)=\tilde{F}_{in}(\Delta_i(\beta_{true}))-\tilde{F}_{in}(0).\end{aligned}\tag{7.3}$$

This gives for the expectation of $c_i(v)$,

$$E[c_i(v)] = [\tilde{F}_{in}(0) - \tilde{F}_{in}(\Delta_i(\beta_{true}))][\frac{1}{\sqrt{n}} v^t m(x_i, \beta_{0,S}) + \frac{1}{2n} v^t \mathcal{M}(x_i, \beta^*_{i,S}) v]. \tag{7.4}$$

By plugging the result from (3.4) into (7.4) and applying Assumptions (A3) and (A1)(iii), in the local alternative framework we obtain

$$E[c_i(v)] = \tilde{f}_{in}(0) v^t m(x_i, \beta_{0,S}) m(x_i, \beta_{0,full})^t \frac{\tilde{\delta}}{n} + O(n^{-\frac{3}{2}}),$$

and, assumption (A2) implies

$$\lim_{n\to\infty} \sum_{i=1}^{n} E[c_i(v)] = v^t \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta. \tag{7.5}$$

For the calculation of the variance of $c_i(v)$ we recall the definition of $r_{n,\tau}$ in (3.4) and use (7.3) and assumption (A4) and get

$$\begin{aligned} \mathrm{Var}[\psi_\tau(u_{i,S})] &= \tilde{F}_{in}(0) - 2\tau \tilde{F}_{in}(0) + \tau^2 - (\tau - \tilde{F}_{in}(0))^2 \\ &= \tilde{F}_{in}(\Delta_i(\beta_{true})) - r_{n,\tau}(x_i) - \left[\tilde{F}_{in}(\Delta_i(\beta_{true})) - r_{n,\tau}(x_i)\right]^2 \\ &= \tau(1-\tau) + r_{n,\tau}(x_i)(2\tau - 1) - (r_{n,\tau}(x_i))^2 = \tau(1-\tau) + O(\frac{1}{\sqrt{n}}). \end{aligned} \tag{7.6}$$

Therefore we obtain

$$\mathrm{Var}[c_i(v)] = \tau(1-\tau)(\frac{1}{\sqrt{n}} m(x_i, \beta_{0,S})^t v)^2 + o(n^{-1}).$$

uniformly with respect to $i = 1, \ldots, n$, which yields (by Assumption (A4))

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{Var}[c_i(v)] = \tau(1-\tau) v^t V_s v. \tag{7.7}$$

Note that, due to assumptions (A0), (A2) and (A3)(i), the process $v^t \Gamma_{n,S}$ satisfies a Lindeberg-Condition. From this result and (7.4), statement (3.13) is then obvious.

7.3 Proof of (3.14)

First, we calculate the expectation and variance of $\sum_{i=1}^{n} b_i(v)$. To this end, assume that $\Delta_i(v) > 0$. The case where $\Delta_i(v) \leq 0$ can be treated analogously with the same result. For the expectation of $b_i(v)$, we obtain for some ξ with $|\xi| \leq |\Delta_i(v)|$

$$\begin{aligned} E[b_i(v)] &= \int_0^{\Delta_i(v)} (-s + \Delta_i(v)) \tilde{f}_{in}(s)\, ds = \tilde{f}_{in}(\xi)(\Delta_i(v)^2)/2 \\ &= \tilde{f}_{in}(0)(\Delta_i(v)^2)/2 + O(n^{-3/2}\|v\|^3), \end{aligned} \tag{7.8}$$

uniformly with respect to $i = 1, \ldots, n$, where the last equality follows from assumption (A1)(iv). Using the representation of $\Delta_i(v)$ in (3.10) it now follows

$$E[\sum_{i=1}^{n} b_i(v)] = \frac{1}{2} v^t Q_{n,S} v + O(n^{-1/2}\|v\|^3) + O(n^{-1}\|v\|^4).$$

Similarly, we have for the variance of $\sum_{i=1}^{n} b_i(v)$ (we consider again the case $\Delta_i(v) > 0$ and remark that the calculations for $\Delta_i(v) \leq 0$ yield the same result)

$$\mathrm{Var}[\sum_{i=1}^{n} b_i(v)] \;\leq\; \sum_{i=1}^{n} \int_0^{\Delta_i(v)} (\Delta_i(v) - s)^2 \tilde{f}_{in}(s)\, ds \leq K_1 \sum_{i=1}^{n} \frac{|\Delta_i(v)|^3}{3} \;=\; O(n^{-1/2}\|v\|^3).$$

for some positive constant K_1. Finally, an application of Chebychev's inequality yields for any $c > 0$ and $v \in \Theta_S$

$$P\Big(|\sum_{i=1}^{n} b_i(v) - E[\sum_{i=1}^{n} b_i(v)]| \leq c n^{-1/6}\|v\|^{3/2}\Big) \geq 1 - \frac{\mathrm{Var}[\sum_{i=1}^{n} b_i(v)]}{c^2 n^{-1/3}\|v\|^3} \geq 1 - \frac{O(n^{-1/2}\|v\|^3)}{c^2 n^{-1/3}\|v\|^3} = 1 - O(n^{-1/6}).$$

which completes the proof of (3.14).

7.4 Proof of (3.15)

Assume that $\hat{T}_n$ is not bounded, but $\|\hat{T}_n\|$ (more precisely a subsequence) tends to infinity at some rate, then the criterion function G_n evaluated at $\hat{T}_n$ can be estimated as follows

$$\begin{aligned} G_n(\hat{T}_n) &= A_n + B_n + C_n + O(n^{-1/2}\|\hat{T}_n\|^3) + O_P(n^{-1/6}\|\hat{T}_n\|^{3/2}) + O(n^{-1}\|\hat{T}_n\|^4) \\ &= A_n + B_n + C_n + o_P(\|\hat{T}_n\|^2), \end{aligned} \tag{7.9}$$

where $C_n = \frac{1}{2}\hat{T}_n^t Q_{n,S} \hat{T}_n$,

$$A_n \;=\; -\hat{T}_n^t \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\tau(u_{i,S}) m(x_i, \beta_{0,S}),\; B_n = \hat{T}_n^t \Big(\frac{1}{n} \sum_{i=1}^{n} \psi_\tau(u_{i,S}) \frac{1}{2} \mathcal{M}(x_i, \beta^*_{i,S})\Big) \hat{T}_n$$

and the estimate in (7.9) follows from Theorem 3.1, which implies $\frac{\|\hat{T}_n\|}{\sqrt{n}} = o_P(1)$. We obtain by similar arguments as in the proof of (3.13) that $A_n = O_P(\|\hat{T}_n\|) = o_P(\|\hat{T}_n\|^2)$ and

$$C_n = O_P(\|\hat{T}_n\|^2).$$

In order to determine the order of the term B_n we define the matrices

$$W_i := \psi_\tau(u_{i,S}) \frac{1}{2} \mathcal{M}(x_i, \beta^*_{i,S})$$

and denote by $w_{i,jk}$ the entries of W_i. Chebychev's inequality and assumption (A2) yield

$$\begin{aligned} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} w_{i,jk} - E[w_{i,jk}]\Big| > \epsilon\Big) &\leq \frac{\max_{i=1,\ldots,n}\mathrm{Var}[w_{i,jk}]}{n\epsilon^2} \leq \frac{C_{jk}\mathrm{Var}[\psi_\tau(u_{i,S})]}{n\epsilon^2} \\ &= \frac{C_{jk}\tau(1-\tau) + O(n^{-1/2})}{n\epsilon^2}. \end{aligned} \tag{7.10}$$

for a positive constant C_{jk} (note that the components of the matrix $\mathcal{M}(x_i, \beta^*_{i,S})$ are bounded). On the other hand we have from (7.3) and (3.4) $E[w_{i,jk}] = O(n^{-1/2})$ which implies $w_{i,jk} = o_P(1)$ and as a consequence we obtain $B_n = o_P(\|\hat{T}_n\|^2)$. Combining all these arguments yield that

$$G_n(\hat{T}_n) = \frac{1}{2}\hat{T}_n^t Q_{n,S}\hat{T}_n(1 + o_p(1)).$$

Under assumptions (A0), (A1)(ii) and (A3), the matrices $Q_{n,S}$ are positive definite for sufficiently large n, so that the dominating term will be positive. On the other hand we have $G_n(\hat{T}_n) \leq G_n(0) = 0$, which provides a contradiction. Consequently the assertion that $\hat{T}_n$ is not stochastically bounded is wrong, which establishes (3.15).

Acknowledgements The authors thank Martina Stein, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1) of the German Research Foundation (DFG) and by the research fund of the KU Leuven (project GOA/07/04). The work of Peter Behl has been funded by a doctoral scholarship of the Hanns-Seidel-Foundation.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Blake, K., Madabushi, R., Derendorf, H., and Lima, J. (2008). Population pharmacodynamic model of bronchodilator response to inhaled albuterol in children and adults with asthma. *Chest*, 134(5):981–989.

Brownlees, C. T. and Gallo, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *Journal of Financial Econometrics*, 6(4):513–539.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458.

Cade, B., Terrell, J., and Schroeder, R. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology*, 80(1):311–323.

Callies, S., de Alwis, D. P., Mehta, A., Burgess, M., and Aarons, L. (2004). Population pharmacokinetic model for daunorubicin and daunorubicinol coadministered with zosuquidar.3hcl (ly335979). *Cancer Chemother Pharmacol*, 54:39–48.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chien, J. Y., Friedrich, S., Heathman, M. A., de Alwis, D. P., and Sinha, V. (2005). Pharmacokinetics/pharmacodynamics and the stages of drug development: Role of modeling and simulation. *The AAPS Journal*, 7(3):E544–E559.

Claeskens, G. and Carroll, R. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 94(2):249–ï¿½265.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *Aust. N. Z. J. Stat.*, 49:359–379.

Claeskens, G. and Hjort, N. (2003). The focussed information criterion. *Journal of the American Statistical Association*, 98:900–916.

Claeskens, G. and Hjort, N. L. (2008a). Minimising average risk in regression models. *Econometric Theory*, 24:493–527.

Claeskens, G. and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Dette, H. and Volgushev, S. (2008). Non-crossing nonparametric estimates of quantile curves. *Journal of the Royal Statistical Society, Ser. B*, 70(3):609–627.

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101(476):1449–1464.

Hurvich, C. and Tsai, I. (1990). Model selection for least absolute deviations regression in small samples. *Statistics and Probability Letters*, 9:259–265.

Jureckova, J. (1994). Regression quantiles and trimmed least squares estimator in nonlinear regression model. *Journal of Nonparametric Statistics*, (3):201–222.

Kim, T. and White, H. (2003). Estimation, inference and specification testing for possibly misspecified quantile regression. *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, pages 107–132.

Koenker, R. (2005). *Quantile Regression*. Cambrige University Press, New York.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Liese, F. and Vajda, I. (1994). Consistency of M-estimates in general regression models. *Journal of Multivariate Analysis*, 50:93–114.

Lim, C., Sen, P., and Peddada, S. (2010). Statistical inference in nonlinear regression under heteroscedasticity. *Sankhya B*, 72:202–218.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory*, 9:478–493.

Park, S. I., Felipe, C. R., Machado, P. G., Garcia, R., Skerjanec, A., Schmouder, R., Tedesco-Silva Jr, H., and Medina-Pestana, J. O. (2005). Pharmacokinetic/pharmacodynamic relationships of FTY720 in kidney transplant recipients. *Braz J Med Biol Res*, 38(5):683–694.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics & Probability Letters*, 3:21–23.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shows, J. H., Lu, W., and Zhang, H. H. (2010). Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference*, 140(7):1903–1917.

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25(8):1369–1382.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19:801–817.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics*, 39(1):174–200.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36:1108–1126.