

## FOCUSSED MODEL SELECTION IN QUANTILE REGRESSION

Peter Behl<sup>1</sup>, Gerda Claeskens<sup>2</sup> and Holger Dette<sup>1</sup>

<sup>1</sup>*Ruhr-Universität Bochum* and <sup>2</sup>*KU Leuven*

*Abstract:* We consider the problem of model selection for quantile regression analysis when 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 and data analysis.

*Key words and phrases:* Quantile regression, model selection focused information criterion.

### 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 that include such important areas as medicine, economics, environment modeling, toxicology, and engineering (see Buchinsky (1994); Cade, Terrell, and Schroeder (1999); 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); Chernozhukov, Fernández-Val, and Galichon (2010) among many others). However, nonparametric methods involve the choice of a regularization parameter and, for high dimensional predictors, these methods are not feasible. Parametric models provide an attractive alternative, but 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, Lu, and Zhang (2010)).

The present paper is motivated by recent applications of nonlinear median regression to the EMAX model in pharmacokinetics (see for example Callies et al. (2004)). In studies of this type model identification is not the primary goal of the statistical analysis, but 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 1 is one in which 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, the minimal dose for which a specified minimum effect is achieved.

Model selection methods such as the Akaike information criterion or the Schwarz-Bayesian information criterion operate in an ‘overall’ mode. Indeed, it is not required and even not possible to specify beforehand which purpose the selected model should serve. This is convenient since, for prediction beyond the last observation as well as for estimation of the variability and for estimation of a 10% quantile, one and the same selected model could be used. This, however, implies that in situations where there is a specific purpose in mind, there could be better search methods that lead to more efficient models. An example is in phase II dose finding studies where the sole purpose is to find the minimal effective dose; there is usually no specific interest in other aspects such as predictions or variability estimation.

The focussed information criterion (FIC, Claeskens and Hjort (2003, 2008b)) is designed for targeted model searches. It explicitly takes the purpose of the modeling procedure into account. The underlying idea is to start by specifying the focus and then to select a model for which the focus estimator has the smallest estimated mean squared error (MSE). Other loss functions than squared error can be used, e.g., linex loss (Claeskens and Hjort (2008a)) or  $\ell_p$  loss (Claeskens, Croux, and van Kerckhoven (2006)). The use of the FIC has been extended from the parametric regression models with maximum likelihood estimation for which it was first defined toward semiparametric models (Claeskens and Carroll (2007)), generalized additive partial linear models (Zhang and Liang (2011)), capture-recapture models (Bartolucci and Lupporelli (2008)), time series models (Claeskens, Croux, and van Kerckhoven (2007)), Cox proportional hazard regression models (Hjort and Claeskens (2006)) and volatility forecasting (Brownlees and Gallo (2008)), to name a few.

The purpose of the present paper is to develop a methodology for focussed model selection in quantile regression analysis. The basic terminology is introduced in Section 2, along with 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 focussed information criterion for quantile regression models. The methodology is illustrated by a simulation study and the analysis of a data example in Section 5. Some concluding remarks are given in Section 6; the more technical arguments are deferred 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 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  $\beta := (\beta_1, \dots, \beta_p, \beta_{p+1}, \dots, \beta_q)^t \in \Theta \subset \mathbb{R}^q$  and an explanatory variable  $x \in \mathcal{X}$ . We follow Claeskens and Hjort (2003) and assume that the specification of the parameter  $\beta$  generates several sub-models, where each of the sub-models contains the first part of the vector  $\beta$ ,  $\beta_0 := (\beta_1, \dots, \beta_p)^t$ . Claeskens and Hjort (2003) call this the narrow model and call these parameters “protected” parameters.

**Example 1.** Consider the Hill model

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

that 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)). The “simplest” model to describe the velocity of a chemical reaction or a dose response relationship is a sub-model of (2.1) with  $\beta_3 = 1$  and  $\beta_4 = 0$ , the Michaelis Menten-model

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

Here (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 that arise as special cases of the Hill model. A famous competitor is the EMAX model with  $\beta_3 = 1$ ,

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

If no placebo effect is assumed, this can be addressed by the choice  $\beta_4 = 0$ ,

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

The models (2.1)–(2.4) are frequently used for modeling dose response relationships.

In dose finding studies, a typical problem is to estimate the minimal effective dose (MED) so that a minimum effect, say  $\Delta$ , is achieved. For the purpose of model selection, the aim is to find the model which best estimates the MED. ‘Best’ is here understood in mean squared error sense. In more detail, the focus of the model search procedure is the quantity  $\mu(\beta) = g^{-1}(\Delta, \beta)$ . For the four models in the current example, the focus is

$$\left( \frac{\beta_2^{\beta_3} (\Delta - \beta_4)}{\beta_1 + \beta_4 - \Delta} \right)^{1/\beta_3}, \quad \frac{\beta_2 \Delta}{\beta_1 - \Delta}, \quad \frac{\beta_2 (\Delta - \beta_4)}{\beta_1 + \beta_4 - \Delta}, \quad \left( \frac{\beta_2^{\beta_3} \Delta}{\beta_1 - \Delta} \right)^{1/\beta_3}, \quad (2.5)$$

for (2.1), (2.2), (2.3), and (2.4), respectively. It is typically the case in phase II clinical trials or in toxicological studies that the estimation of the minimum effective dose is the main goal of the experiment.

We derive a focussed model choice criterion for quantile regression analysis that addresses problems of this type in more generality. We propose to choose a subset from  $(\beta_{p+1}, \dots, \beta_q)$  such that the MSE for estimating a certain focus parameter

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

by the chosen quantile regression model is minimal. In order to find this “best” model, we determine the MSE of the estimator  $\hat{\mu}_S$  for each possible sub-model, where  $S$  denotes any subset from  $(\beta_{p+1}, \dots, \beta_q)^t$ . Throughout the text,  $\beta_S$  denotes a parameter vector for the model that includes all parameters from the narrow model plus the parameters contained in a set  $S \subset \{p+1, \dots, q\}$ ,  $\beta_S = (\beta_1, \dots, \beta_p, (\beta_j)_{j \in S})^t$ . Note that  $\beta_S \in \Theta_S$ , where  $\Theta_S \subset \mathbb{R}^{p+|S|}$  denotes the canonical projection of  $\Theta$  corresponding to the parameters from the sub-model  $S$ . We use the notation  $g(x; \beta_S)$  for the model  $g(x; \beta)$ , obtained for the vector  $\beta = (\beta_1, \dots, \beta_p, \gamma_{0, S^c}, (\beta_j)_{j \in S})^t$ , where for a given set  $S$  the vector  $\gamma_{0, S}$  consists of the parameters of a  $q-p$ -dimensional vector  $\gamma_0$  corresponding to the sub-model  $S$  and  $S^c$  denotes the complement of  $S$ . Here, the values of  $\gamma_0$  are always chosen such that  $g(x; \beta_1, \dots, \beta_p, \gamma_0)$  gives the narrow model. For example, in a linear regression model where  $\gamma$  corresponds to the regression coefficients, we choose

$\gamma_0 = (0, \dots, 0)^t$ , whereas in Example 1 where the narrow and full model are given by (2.2) and (2.1), respectively, we have  $(\gamma_{0,1}, \gamma_{0,2}) = (1, 0)$ . Other functions of the parameter  $\beta$  are interpreted in the same way if their argument is  $\beta_S$ . In order to emphasize the case where all parameters are included in the quantile regression model we use the notation  $g(x; \beta_{\text{full}})$  and take

$$\beta_{0,\text{full}} = (\beta_1, \dots, \beta_p, \gamma_0)^t \text{ and } \beta_{0,S} = (\beta_1, \dots, \beta_p, \gamma_{0,S})^t.$$

Throughout, let  $n$  denote the sample size and  $\delta$  be a vector of dimension  $q - p$ . Following Claeskens and Hjort (2003), we assume that the unknown “true” parameter  $\beta_{\text{true}}$  is of the form

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

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

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

where  $\rho_{\tau}(z) := \tau I(z \geq 0)z + (\tau - 1)I(z < 0)z$  denotes the check function (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), required for the derivation of a focussed information criterion for quantile regression. We make some assumptions.

(A0) The parameter space  $\Theta$  is compact.

(A1) (i) The random variables  $Y_1, \dots, Y_n$  are independent random variables with densities  $f_{1n}(\cdot|x_1), \dots, f_{nn}(\cdot|x_n)$  such that for each  $x \in \mathcal{X}$ ,  $f_{in}(\cdot|x)$  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) There exists a constant  $w > 0$  such that for  $i = 1, 2, \dots, n$ ;  $n \in \mathbb{N}$  the densities  $f_{in}(\cdot|x_i)$  are uniformly bounded away from 0 by a constant  $0 < K_0 < \infty$  in a neighbourhood  $W := [g(x_i, \beta_{\text{true}}) - w, g(x_i, \beta_{\text{true}}) + w]$  of  $g(x_i, \beta_{\text{true}})$ .

- (iii) The densities  $\tilde{f}_{in}$  are uniformly bounded from above by a constant  $0 < K_1 < \infty$ .
- (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)  $g(x; \beta_{\text{full}})$  is twice continuously differentiable with respect to the parameter vector  $\beta_{\text{full}}$  for all  $x \in \mathcal{X}$ . For a given sub-model  $S$  and  $\beta'_S \in \Theta_S$ , the corresponding derivatives are

$$m(x_i, \beta'_S) = \left. \frac{\partial g(x_i; \beta_S)}{\partial \beta_S^t} \right|_{\beta_S = \beta'_S}, \quad \mathcal{M}(x_i, \beta'_S) = \left. \left( \frac{\partial^2 g(x_i; \beta_S)}{\partial \beta_S \partial \beta_S^t} \right) \right|_{\beta_S = \beta'_S}.$$

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

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

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_{in}(g(x_i; \beta_{\text{true}})) m(x_i, \beta_{0,\text{full}}) m(x_i, \beta_{0,\text{full}})^t = Q := \begin{pmatrix} Q_{00} & Q_{01} \\ Q_{10} & Q_{11} \end{pmatrix},$$

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

- (iii) There exist constants  $0 < C_1, C_2 < \infty$  and  $u > 0$  such that

$$\max_{i=1, \dots, n} \|m(x_i, \tilde{\beta})\| < C_1, \quad \max_{i=1, \dots, n} \|\mathcal{M}(x_i, \tilde{\beta})\| < C_2$$

for all  $\tilde{\beta}$  in the neighbourhood  $U := \{\beta \in \Theta \mid \|\beta - \beta_{0,\text{full}}\| \leq u\}$  of  $\beta_{0,\text{full}}$ .

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

- (A5) (i) There exists a constant  $0 < k_1 < \infty$  such that, for all  $\beta \in \Theta$  and for  $n > n_0$ ,

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

- (ii) There exists a constant  $0 < k_2 < \infty$  such that, for all  $\beta, \beta' \in \Theta$  and for  $n > n_0$ ,

$$\frac{1}{n} \sum_{i=1}^n [g(x_i; \beta') - g(x_i; \beta)]^2 \leq k_2 \|\beta' - \beta\|^2.$$

Note that the second subscript  $n$  is used here for the distribution functions  $F_{in}$  (and corresponding densities  $f_{in}$ ) in order to emphasize that we are working under (2.7). A similar assumption as (A5) was used by Jureckova and Prochazka (1994) in order to ensure identifiability of the parameter  $\beta_0$ ,

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

for all  $\beta, \beta' \in \Theta$ . For some important nonlinear models, the left inequality 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  $\beta_2$  and  $\beta_3$ . However, for the derivation of the asymptotic results it is actually enough to assume that (3.1) holds only for the “pseudo-true” parameter  $\beta_{0,\text{full}}$ , which corresponds to (A5)(i).

### 3.1. Consistency of the quantile regression estimator

In this section, we prove that under (2.7) the estimated regression quantile  $\hat{\beta}_{n,S}$  in a given submodel  $S$  converges in probability to  $\beta_{0,S}$ .

**Theorem 1.** *Assume that (A0)–(A5) and (2.7) are satisfied. For any submodel  $S$ ,  $\hat{\beta}_{n,S} - \beta_{0,S} = o_P(1)$  as  $n \rightarrow \infty$ .*

**Proof.** Let

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

and note that under the local alternatives (2.7),  $\Delta_i(\beta_{\text{true}})$  tends to zero as  $n \rightarrow \infty$ . Let  $r_{n,\tau}(x_i) := \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - \tilde{F}_{in}(0)$ . Using (A1), (A3)(iii), and (2.7), we obtain, for some  $\alpha$  satisfying  $|\alpha| \leq |\Delta_i(\beta_{\text{true}})|$

$$\begin{aligned} |r_{n,\tau}(x_i)| &= |\tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - \tilde{F}_{in}(0)| = |\tilde{f}_{in}(\alpha)\Delta_i(\beta_{\text{true}})| \\ &\leq K_1 \max_{i=1,\dots,n} \frac{\|m(x_i, \beta_{0,\text{full}})\|}{\sqrt{n}} \|\tilde{\delta}\| + o\left(\frac{1}{\sqrt{n}}\right) = o(1), \end{aligned} \tag{3.3}$$

where  $\tilde{\delta} := (0, \dots, 0, \delta)^t$  denotes a vector of length  $q$  that is zero in the first  $p$  components and takes the value  $\delta$  from (2.7) in the last  $q - p$  components. With the  $u_{i,S}$  in (A1),  $\hat{\beta}_{n,S}$  minimizes the objective function

$$\begin{aligned} Z_n(\beta_S) &:= \frac{1}{n} \sum_{i=1}^n [\rho_\tau(Y_i - g(x_i; \beta_S)) - \rho_\tau(u_{i,S})] \\ &= \frac{1}{n} \sum_{i=1}^n [\rho_\tau(u_{i,S} - \Delta_i(\beta_S)) - \rho_\tau(u_{i,S})]. \end{aligned} \tag{3.4}$$

We have

$$E[Z_n(\beta_S)] = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} [(\tau - \mathbf{1}_{\{s \leq \Delta_i(\beta_S)\}})(s - \Delta_i(\beta_S)) + (\mathbf{1}_{\{s \leq 0\}} - \tau)s] d\tilde{F}_{in}(s)$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \left\{ - \int_{-\infty}^{\Delta_i(\beta_S)} s d\tilde{F}_{in}(s) + \int_{-\infty}^0 s d\tilde{F}_{in}(s) \right. \\
&\quad \left. + \Delta_i(\beta_S) \tilde{F}_{in}(\Delta_i(\beta_S)) - \tau \Delta_i(\beta_S) \right\} \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ \int_{\Delta_i(\beta_S)}^0 s d\tilde{F}_{in}(s) + \Delta_i(\beta_S) (\tilde{F}_{in}(\Delta_i(\beta_S)) - \tilde{F}_{in}(0)) \right. \\
&\quad \left. + \Delta_i(\beta_S) (\tilde{F}_{in}(0) - \tilde{F}_{in}(\Delta_i(\beta_{\text{true}}))) \right\} \\
&= \frac{1}{n} \sum_{i=1}^n \int_{\Delta_i(\beta_S)}^0 (s - \Delta_i(\beta_S)) d\tilde{F}_{in}(s) + o(1), \tag{3.5}
\end{aligned}$$

where the last identity follows from (3.3) and the fact that  $(1/n) \sum_{i=1}^n \Delta_i(\beta_S)$  is bounded due to (A5) and (A0). 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, (A5)(i) guarantees that, for sufficiently large  $n$  and any parameter  $\beta_S \in \Theta_S$  different from  $\beta_{0,S}$ , we have

$$\frac{1}{n} \sum_{i=1}^n \left( \int_{\Delta_i(\beta_S)}^0 (s - \Delta_i(\beta_S)) d\tilde{F}_{in}(s) \right) > 0. \tag{3.6}$$

This implies that, for sufficiently large  $n$ , the sum in (3.5) is only zero for  $\beta_S = \beta_{0,S}$  and is strictly positive otherwise. We show in the Appendix that

$$\sup_{\beta_S \in \Theta_S} |Z_n(\beta_S) - E[Z_n(\beta_S)]| \xrightarrow{P} 0. \tag{3.7}$$

Because  $Z_n$  is minimized at  $\hat{\beta}_{n,S}$ , we have

$$Z_n(\hat{\beta}_{n,S}) \leq Z_n(\beta_{0,S}) = 0. \tag{3.8}$$

Then from (3.6), (3.7), and (3.8) one has  $\|\hat{\beta}_{n,S} - \beta_{0,S}\| = o_P(1)$ .

### 3.2. Weak convergence under local alternatives

Here we derive the asymptotic distribution of  $\hat{\beta}_{n,S}$  for each sub-model  $S$  under (2.7).

**Theorem 2.** *Under (A0)–(A5) and (2.7) we have*

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

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



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

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

and  $\pi_S$  is a  $|S| \times (q - p)$ -projection matrix consisting of ones and zeros that extracts from  $Q_{11}$  the rows corresponding to the sub-model  $S$ .

**Proof.** The estimator  $\hat{\beta}_{n,S}$  minimizes the objective function  $G_n(\beta_S) := \sum_{i=1}^n [\rho_\tau(Y_i - g(x_i; \beta_S)) - \rho_\tau(u_{i,S})]$ . We use a Taylor expansion at the point  $\beta_{0,S}$  to write  $G_n$  in the slightly modified form

$$G_n(\beta_S) = \sum_{i=1}^n \left[ \mathbb{1}_{\{u_{i,S} < 0\}}(1 - \tau)\Delta_i(\beta_S) - \mathbb{1}_{\{u_{i,S} \geq 0\}}\tau\Delta_i(\beta_S) \right. \\ \left. + \mathbb{1}_{\{0 \leq u_{i,S} \leq \Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S}) + \mathbb{1}_{\{\Delta_i(\beta_S) \leq u_{i,S} < 0\}}(u_{i,S} - \Delta_i(\beta_S)) \right] \\ = -\sqrt{n}(\beta_S - \beta_{0,S})^t(\Gamma_{n,S} + R_{n,S}(\beta_S)) + \sum_{i=1}^n b_i(\beta_S), \tag{3.9}$$

where

$$\Gamma_{n,S} := \sum_{i=1}^n \psi_\tau(u_{i,S}) \frac{1}{\sqrt{n}} m(x_i, \beta_{0,S}), \tag{3.10}$$

$$R_{n,S}(\beta_S) := \sum_{i=1}^n \psi_\tau(u_{i,S}) \frac{1}{\sqrt{n}} \left[ m(x_i, \tilde{\beta}_{i,S}) - m(x_i, \beta_{0,S}) \right],$$

$$b_i(\beta_S) := \mathbb{1}_{\{0 \leq u_{i,S} \leq \Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S}) + \mathbb{1}_{\{\Delta_i(\beta_S) \leq u_{i,S} < 0\}}(u_{i,S} - \Delta_i(\beta_S))$$

and  $\tilde{\beta}_{i,S}$  in the definition of  $R_{n,S}$  denotes a suitable value between  $\beta_S$  and  $\beta_{0,S}$ . Here

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

denotes the “derivative” of the check function  $\rho_\tau$ . In the Appendix we derive the following asymptotic properties of  $G_n$ .

- For  $\Gamma_{n,S}$  as in (3.10) we have

$$\Gamma_{n,S} \xrightarrow{D} W_S, \tag{3.11}$$

where

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

- For every  $\beta_S \in U$ ,

$$\begin{aligned} G_n(\beta_S) &= -v^t \Gamma_{n,S} + \frac{1}{2} v^t Q_{n,S} v + O_P(n^{-1/2} \|v\|^2) + O(n^{-1/2} \|v\|^3) \\ &\quad + O(n^{-1} \|v\|^4) + O_P(n^{-1/6} \|v\|^{3/2}), \end{aligned} \quad (3.12)$$

where  $U$  denotes the neighbourhood of  $\beta_{0,\text{full}}$  in A3(iii) and

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

The approximation (3.12) is used to establish a Bahadur-type representation for the statistic  $\hat{T}_n := \sqrt{n}(\hat{\beta}_{n,S} - \beta_{0,S})$ . More precisely, we show in the appendix that

$$\|\hat{T}_n\| = O_P(1). \quad (3.13)$$

Theorem 1 implies that  $P(\hat{\beta}_{n,S} \in U) \rightarrow 1$  for  $n \rightarrow \infty$ , and therefore, by (3.12) and (3.13),

$$G_n(\hat{\beta}_{n,S}) = -\hat{T}_n^t \Gamma_{n,S} + \frac{1}{2} \hat{T}_n^t Q_{n,S} \hat{T}_n + o_P(1). \quad (3.14)$$

Take  $\beta_{n,S}^* := \beta_{0,S} + U_n/\sqrt{n}$  with  $U_n := Q_{n,S}^{-1} \Gamma_{n,S}$ . By (3.11),  $U_n$  is asymptotically normal distributed and, in particular,  $\|U_n\|$  is also stochastically bounded. It follows that  $P(\beta_{n,S}^* \in U) \rightarrow 1$  for  $n \rightarrow \infty$ . Moreover,  $U_n$  satisfies  $U_n^t \Gamma_{n,S} = U_n^t Q_{n,S} U_n$  and consequently (3.12) yields

$$G_n(\beta_{n,S}^*) = -\frac{1}{2} U_n^t Q_{n,S} U_n + o_P(1). \quad (3.15)$$

From (3.14) and (3.15),

$$\begin{aligned} G_n(\hat{\beta}_{n,S}) - G_n(\beta_{n,S}^*) &= -\hat{T}_n^t \Gamma_{n,S} + \frac{1}{2} \hat{T}_n^t Q_{n,S} \hat{T}_n + \frac{1}{2} U_n^t Q_{n,S} U_n + o_P(1) \\ &= \frac{1}{2} (\hat{T}_n - U_n)^t Q_{n,S} (\hat{T}_n - U_n) + o_P(1). \end{aligned} \quad (3.16)$$

By the definition of  $\hat{\beta}_{n,S}$  the left-hand-side of (3.16) is always non-positive, while the dominating term on the right-hand-side is always positive due to the positive definiteness of  $Q_{n,S}$ . Consequently,

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

The asymptotic normality of  $\hat{T}_n$  follows directly from (3.11).

#### 4. The FIC for Quantile Regression

From Theorem 2, an expression for the FIC can be derived by similar arguments to those in Claeskens and Hjort (2003). By applying the Delta method we get

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

with

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

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

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

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

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

It remains to estimate the unknown quantities in this expression so that the FIC can be calculated from the data. The key step is to find an estimator of the matrices  $Q_S$  that is consistent under local alternatives. Using the regression “errors”  $\hat{\epsilon}_i = Y_i - g(x_i; \hat{\beta}_{\text{full}})$ , similarly as in Kim and White (2003), the matrix  $Q_S$  may be estimated by

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

where  $\hat{\beta}_{0,S}$  is calculated by taking estimates  $\hat{\beta}_1, \dots, \hat{\beta}_p$  from the full model, and  $\hat{c}_n$  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 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 2,

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

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  $\hat{\delta}\hat{\delta}^t = D_n D_n^t - \hat{K}$ , truncated to zero if the result is negative definite. An estimator  $\hat{K}$  can be obtained directly by taking the corresponding rows and columns of  $\tau(1-\tau)\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}$ , the estimated covariance matrix of the full model. The derivatives of  $\mu$  can be estimated by plug-in-estimators, using estimates for  $\beta_{0,\text{full}}$  from the full model. Summarizing these calculations, we obtain for every submodel  $S$  an expression for the estimated FIC that can be calculated from the data:

$$\begin{aligned} \widehat{FIC}_S &= \frac{\partial\mu(\beta)^t}{\partial\beta_S} \Big|_{\beta=\hat{\beta}_{0,\text{full}}} \hat{Q}_S^{-1} \begin{pmatrix} \hat{Q}_{01} \\ \pi_S \hat{Q}_{11} \end{pmatrix} \delta\hat{\delta}^t \begin{pmatrix} \hat{Q}_{01} \\ \pi_S \hat{Q}_{11} \end{pmatrix}^t (\hat{Q}_S^{-1})^t \frac{\partial\mu(\beta)}{\partial\beta_S} \Big|_{\beta=\hat{\beta}_{0,\text{full}}} \\ &\quad + \frac{\partial\mu(\beta)^t}{\partial\beta_S} \Big|_{\beta=\hat{\beta}_{0,\text{full}}} \tau(1-\tau)\hat{Q}_S^{-1}\hat{V}_S\hat{Q}_S^{-1} \frac{\partial\mu(\beta)}{\partial\beta_S} \Big|_{\beta=\hat{\beta}_{0,\text{full}}} \\ &\quad - 2 \frac{\partial\mu(\beta)^t}{\partial\beta_S} \Big|_{\beta=\hat{\beta}_{0,\text{full}}} \hat{Q}_S^{-1} \begin{pmatrix} \hat{Q}_{01} \\ \pi_S \hat{Q}_{11} \end{pmatrix} \delta\hat{\delta}^t \frac{\partial\mu(\beta)}{\partial\gamma} \Big|_{\beta=\hat{\beta}_{0,\text{full}}}, \end{aligned} \quad (4.4)$$

where  $\gamma := (\beta_{p+1}, \dots, \beta_q)^t$  denotes the last  $(q-p)$  components of the parameter vector  $\beta$ .

The largest difficulty in specifying a focussed information criterion is the derivation of the mean squared error expressions under local misspecification. Once these expressions are obtained, the MSE values of several models may be compared in order to decide on a best model. Such comparisons give rise to inequalities in terms of the local misspecification neighborhood defined by  $\delta$ , the chosen focus  $\mu$  and  $K$ , related to the lower-right part of the inverse Fisher information matrix. The result of Theorem 5.3 of Claeskens and Hjort (2008b) where the MSE values of two models are compared, is applicable to this setting.

It previously has been obtained that some averaged versions of the FIC behave asymptotically similar to the AIC, see Claeskens and Hjort (2008a). We do not repeat these calculations.

## 5. Finite Sample Properties

Different model selection methods lead to different models. Theoretical properties, such as efficiency, might provide other reasons for practitioners to prefer

one criterion over another. It can be shown that no model selection method can be universally best, a criterion that is efficient cannot at the same time be strongly consistent (Yang (2005)).

With the construction of the FIC, we start from a focus that is to be estimated and we seek the best model for this purpose. Dose finding studies are ideally suited for the use of the FIC since the focus, the MED, plays the prominent role in the modeling process. More generally, any focus  $\mu$  that is expressible in terms of the model parameters  $\beta$  and is differentiable with respect to these parameters can be taken as the starting point for the FIC; one proceeds to (4.4) and the model with the smallest such value gets selected.

### 5.1. Linear quantile regression

We report on a simulation study of model selection by the FIC criterion in a linear quantile regression model. We illustrate the practical application of the FIC for quantile regression in a detailed way, and compare the performance of the FIC for estimation of the focus parameter to such model selection criteria as AIC and BIC. We consider the model

$$g(x; \beta_0, \beta_1, \gamma_1, \gamma_2, \gamma_3, \gamma_4) = \beta_0 + \beta_1 x_1 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4. \quad (5.1)$$

Here,  $\beta_0$  and  $\beta_1$  denote the “protected” parameters that are included in every candidate model, while  $\gamma_1$  to  $\gamma_4$  may be included or not. Consequently, there are 16 candidate models to choose from which all contain  $\beta_0$  and  $\beta_1$ , but differ with respect to the  $\gamma$ -parameters. For example, the narrow model only contains  $\beta_0$  and  $\beta_1$  while the  $\gamma$ s are set to zero. The procedure starts by specifying the focus parameter, for which we chose the prediction of  $Y$  at covariate value  $(x_1, z_1, z_2, z_3, z_4) = (10, 10, 10, 10, 10)$ . The focus is written in terms of the model notation as

$$\mu_1(\beta_0, \beta_1, \gamma_1, \gamma_2, \gamma_3, \gamma_4) = \beta_0 + 10\beta_1 + 10\gamma_1 + 10\gamma_2 + 10\gamma_3 + 10\gamma_4.$$

Each candidate model is fitted to the data and the resulting parameter estimates are used to estimate the MSE of the focus estimator in the considered models. This yields an FIC value defined by (4.4) for every candidate model. The model with the lowest FIC value gets selected and an estimator of the focus parameter is obtained by taking the estimated focus from the chosen model.

For our simulation study, data were generated from (5.1) with parameter values  $\beta_0 = 1$ ,  $\beta_1 = 1$ , and  $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1/\sqrt{n}$ . A set of covariate values of size  $n$  was generated as

$$\begin{aligned} X_1 &\sim N(20, 25), Z_1 \sim N(20, 6.25), Z_2 \sim N(-10, 6.25), \\ Z_3 &\sim N(10, 1), Z_4 \sim N(5, 2.25). \end{aligned}$$

Table 1. Median and median absolute deviation (MAD) of the absolute errors of the estimates of the focus  $\mu_1$  obtained from the FIC, AIC and BIC.

	Median			MAD		
	FIC	AIC	BIC	FIC	AIC	BIC
$n = 50, \quad \epsilon \sim N(0, 4)$	1.90	2.11	2.09	0.94	1.12	1.17
$n = 50, \quad \epsilon \sim C(0, 2)$	2.36	2.64	2.62	1.37	1.53	1.53
$n = 100, \quad \epsilon \sim N(0, 4)$	1.45	1.74	1.79	0.70	0.94	1.02
$n = 100, \quad \epsilon \sim C(0, 2)$	1.65	2.01	2.05	0.87	1.18	1.20

Those values were then considered as fixed and used in all simulation runs. For the distribution of the “error”  $\epsilon = Y - g(x, \beta_{\text{true}})$  we took two scenarios: A normal distribution with mean 0 and variance  $\sigma^2 = 4$ , and a Cauchy distribution with location parameter  $a = 0$  and scale parameter  $b = 2$ . We considered sample sizes  $n = 50$  and  $n = 100$ . The parameters were estimated using median regression. We conducted 2,000 simulation runs where in each run model selection was performed using the FIC. From the chosen model, we obtained a post-selection-estimator  $\hat{\mu}_{1,FIC}$  for the focus parameter  $\mu_1$ .

For comparison we also estimated  $\mu_1$  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

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

where  $\hat{\sigma} = (1/n) \sum_{i=1}^n |y_i - g(x_i; \hat{\beta}_{n,S})|$ ,  $p$  denotes the number of parameters in the model  $S$ , and  $n$  the number of observations; for details see Hurvich and Tsai (1990). We computed the absolute errors of the post-selection-estimators for  $\mu_1$ ,

$$|\hat{\mu}_{1,FIC} - \mu_{\text{true}}|, \quad |\hat{\mu}_{1,AIC} - \mu_{\text{true}}|, \quad |\hat{\mu}_{1,BIC} - \mu_{\text{true}}|. \quad (5.2)$$

We calculated the median absolute error and the median absolute deviation (MAD) from the 2,000 replications separately for FIC, AIC, and BIC. The results are in Table 1.

In a second setting, data were again generated from (5.1) with  $\beta_0 = \beta_1 = 1$  and  $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.3$ . The corresponding results are shown in Table 2. From Tables 1 and 2 it can be seen that in most considered scenarios FIC performs best in terms of median absolute error and MAD, or at least as well as AIC and BIC.

## 5.2. Nonlinear quantile regression: application of the FIC for dose-response-modeling

Here we consider the class of quantile regression models introduced in Example 1. All results are again based on 2,000 simulation runs and we consider three

Table 2. Median and median absolute deviation of the absolute errors of the estimates of the focus  $\mu_1$  obtained from the FIC, AIC and BIC.

	Median			MAD		
	FIC	AIC	BIC	FIC	AIC	BIC
$n = 50, \epsilon \sim N(0, 4)$	2.60	2.72	3.34	1.24	1.47	1.47
$n = 50, \epsilon \sim C(0, 2)$	3.42	3.96	3.83	1.59	1.98	2.12
$n = 100, \epsilon \sim N(0, 4)$	2.39	1.87	2.76	1.19	1.16	1.55
$n = 100, \epsilon \sim C(0, 2)$	2.90	4.05	4.01	1.38	2.04	2.38

Table 3. Median and median absolute deviation of the absolute errors of the estimates of the focus  $\mu_2$  obtained from the FIC, AIC and BIC.

	Median			MAD		
	FIC	AIC	BIC	FIC	AIC	BIC
$\mathcal{N}(0, 0.01)$	3.98	5.07	5.71	2.10	1.68	1.07
$\mathcal{N}(0, \sigma^2(x_i))$	3.69	4.46	5.18	2.04	1.80	1.39
$\mathcal{C}(0, 0.07)$	4.33	5.62	5.69	2.35	1.06	1.00

scenarios for the error distribution: errors are normal with mean 0 and variance  $\sigma^2 = 0.01$ , Cauchy with location parameter  $a = 0$  and scale parameter  $b = 0.07$ , or normal with a heteroscedastic variance structure. In the heteroscedastic case the errors are normal with mean 0 and standard deviation, depending on the explanatory variable  $x$ ,

$$\sigma(x) = \tau_0 + \frac{\tau_1}{1 + e^{-\tau_2 x}}, \quad (5.3)$$

where  $\tau_0 = -0.1$ ,  $\tau_1 = 0.24$ , and  $\tau_2 = 0.15$ . This variance function was proposed by Lim, Sen, and Peddada (2010) for dose-response-modeling. We consider two competing models, the Michaelis-Menten-model (2.2) and the Hill model without intercept (2.4). We generated data from (2.4) with parameter values  $\beta_1 = 0.417$ ,  $\beta_2 = 25$ , and  $\beta_3 = 1.75$ . As experimental design we chose six dose levels equidistantly over the dose range [1mg, 150mg] and assigned 32 observations to each dose level. The parameters were estimated using median regression. From these results we obtained a robust estimate for the focus parameter  $\mu_2$ , the minimal effective dose (MED) defined in (2.5) with  $\Delta = 0.1$ . We investigated the performance of the FIC for choosing between (2.2) and (2.4). As before, we compared FIC to AIC and BIC. In Table 3 we display the median and median absolute deviation of the absolute errors (5.2) of estimators obtained from the different model selection procedures.

The median absolute error of FIC is clearly the smallest in all cases, while the BIC yields the lowest MAD values, but also the largest median of the absolute errors.

For this nonlinear example, we also compared FIC to AIC and BIC by counting how many times in 2,000 simulation runs the FIC obtained a better estimator

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).

$\varepsilon_i$	FIC<AIC	FIC=AIC	AIC<FIC	FIC<BIC	FIC=BIC	BIC<FIC
$\mathcal{N}(0, 0.01)$	637	1083	280	1148	487	365
$\mathcal{N}(0, \sigma^2(x_i))$	618	1192	190	1143	496	361
$\mathcal{C}(0, 0.07)$	1064	511	425	1202	268	530

(in terms of absolute deviation) than AIC (FIC<AIC) and BIC (FIC<BIC), and vice-versa. See Table 4.

Here in the majority of cases the FIC selected a model which is better than the model chosen by AIC and BIC.

### 5.3. Application of the FIC in a clinical dose response study

We consider a data example from a dose response study that 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). The intercept  $\beta_4$  in (2.1) is assumed to be zero, so that the Michaelis Menten model (2.2) and 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, here  $\Delta = 90$ . Figure 1 shows the data, the fitted median regression curves and the location of the  $IC_{90}$  for both models. We observe substantial differences between the estimates of the  $IC_{90}$  obtained from them, so model selection is of importance in this study. We used the FIC to decide whether the Hill slope  $\beta_3$  is included in the model or not. The resulting FIC values are  $1.29 \cdot 10^7$  for (2.2) and  $3.17 \cdot 10^6$  for (2.4). Thus, the  $IC_{90}$  is estimated using the Hill model with no intercept, which gives a value of  $IC_{90} = 183.19$ . 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

Our work 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. We have extended the available theory for estimation under local misspecification from a likelihood setting towards quantile regression models and developed a focussed information criterion (FIC) that takes the specific



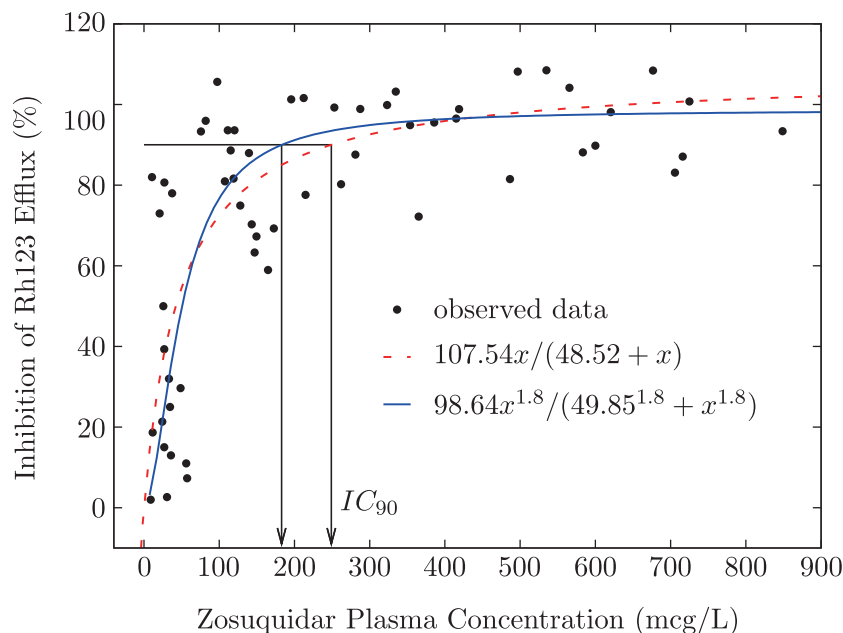


Figure 1. *Zosuquidar data with estimated median regression curves from the Hill (solid line) and Michaelis Menten model (dashed line).*

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 such standard selection methods as AIC and BIC.

The presented FIC is applicable in nonlinear quantile regression models in general. The procedure is to specify the focus and write it in terms of the model parameters. The MSE of the focus estimator is then estimated in each considered model under a local misspecification assumption. This yields a value of the FIC for every model, and the model with the smallest FIC gets selected.

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, we can average the risk function over a wanted domain of values for the covariate  $x$ .

An interesting topic for future research is a study of asymptotic properties of the estimators under a different local misspecification setting than (2.7), 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, Sec. 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  $\int f(y; \theta_0, \gamma_0)|r(y)|dy < \infty$  and  $\int f(y; \theta_0, \gamma_0)r(y)dy = 0$ . It is expected that theoretical properties similar to those in the present paper can be developed for such a situation.

### Acknowledgement

We thank all reviewers of this paper for their constructive remarks and 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/FlexStatRob). The work of Peter Behl has been funded by a doctoral scholarship of the Hanns-Seidel-Foundation.

### Appendix: Proof of technical results

**Proof of (3.7).** The proof of the uniform convergence property can be established using results of Liese and Vajda (1994). However, we have to keep in mind that we work under (2.7). For convenience, take  $\delta_n(\beta_S) := Z_n(\beta_S) - E[Z_n(\beta_S)]$ , where  $Z_n$  is defined by (3.4). We begin with a proof of the following properties.

(B1) The class of functions  $\{\delta_n(\beta_S) | n \in \mathbb{N}, n > n_0\}$  is equicontinuous on  $\Theta_S$ .

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

Observe that, for  $\beta_{S,1}, \beta_{S,2} \in \Theta_S$ ,

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

by the Lipschitz continuity of the check function; (B1) is then implied by (A5) and (A0). For (B2) let

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

which gives

$$\frac{1}{n} \sum_{i=1}^n z_i(\beta_S)^2 = \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{1}_{\{u_{i,S} \leq 0\}}(1 - \tau)^2 \Delta_i^2(\beta_S) + \mathbb{1}_{\{u_{i,S} > 0\}}\tau^2 \Delta_i^2(\beta_S) \right]$$

$$\begin{aligned}
 & +\mathbb{1}_{\{0 < u_{i,S} \leq \Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S})^2 \\
 & +\mathbb{1}_{\{\Delta_i(\beta_S) \leq u_{i,S} \leq 0\}}(u_{i,S} - \Delta_i(\beta_S))^2 \\
 & -2\mathbb{1}_{\{0 < u_{i,S} \leq \Delta_i(\beta_S)\}}\tau\Delta_i(\beta_S)(\Delta_i(\beta_S) - u_{i,S}) \\
 & +2\mathbb{1}_{\{\Delta_i(\beta_S) \leq u_{i,S} \leq 0\}}(1 - \tau)\Delta_i(\beta_S)(u_{i,S} - \Delta_i(\beta_S)) \Big]. \tag{A.2}
 \end{aligned}$$

Taking e.g., the expectation of  $(1/n) \sum_{i=1}^n \mathbb{1}_{\{0 < u_{i,S} \leq \Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S})^2$ , the third term in (A.2) is

$$\begin{aligned}
 & E \left[ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{0 < u_{i,S} \leq \Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S})^2 \right] \\
 & = \frac{1}{n} \sum_{i=1}^n \int_0^{\Delta_i(\beta_S)} (\Delta_i(\beta_S) - s)^2 \tilde{f}_{in}(s) ds \leq \frac{1}{n} \sum_{i=1}^n \Delta_i^2(\beta_S)
 \end{aligned}$$

which is bounded due to (A0) and (A5). Since the expectations of the other terms in (A.2) can be similarly bounded, we obtain that  $(1/n) \sum_{i=1}^n E[z_i(\beta_S)^2]$  is bounded. It follows from Chebychev’s inequality that

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

which establishes (B2). The uniform convergence in (3.7) can be derived from (B1) and (B2) using similar arguments as presented in Liese and Vajda (1994). (B1) yields for any  $\epsilon > 0$  the existence of a  $\delta > 0$  such that, for every  $\beta^* \in \Theta_S$ ,

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

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

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

for some  $i \in 1, \dots, k$ . As a consequence,

$$\lim_{n \rightarrow \infty} P(\sup_{\beta_S \in \Theta_S} |\delta_n(\beta_S)| > \epsilon) \leq \lim_{n \rightarrow \infty} P(\max_{1 \leq i \leq k} |\delta_n(\beta_i)| > \frac{\epsilon}{2}) = 0.$$

where the last equation follows from (B2), which implies (3.7).

**Proof of (3.11).** With the definition of  $\tilde{F}$  and  $\tilde{f}$  in (A1), a straightforward calculation yields

$$E[\psi_\tau(u_{i,S})] = \tau(1 - \tilde{F}_{in}(0)) + (\tau - 1)\tilde{F}_{in}(0) = \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - \tilde{F}_{in}(0).$$

This gives

$$E[\Gamma_{n,S}] = \sum_{i=1}^n \left[ \left( \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - \tilde{F}_{in}(0) \right) \frac{1}{\sqrt{n}} m(x_i, \beta_{0,S}) \right]. \quad (\text{A.3})$$

Note that for some  $\alpha_i$  satisfying  $|\alpha_i| \leq |\Delta_i(\beta_{\text{true}})|$  and  $\tilde{\beta}_i$  between  $\beta_{0,\text{full}}$  and  $\beta_{\text{true}}$ , by using (A1), (A2), and (2.7), we obtain

$$\begin{aligned} \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - \tilde{F}_{in}(0) &= \tilde{f}_{in}(\alpha_i) \left( m(x_i, \tilde{\beta}_i)^t \frac{\tilde{\delta}}{\sqrt{n}} \right) \\ &= \tilde{f}_{in}(0) \left( m(x_i, \beta_{0,\text{full}})^t \frac{\tilde{\delta}}{\sqrt{n}} \right) + o\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Together with (A.3) and (A1)(iv), (A3)(i), and (3.3), this yields

$$E[\Gamma_{n,S}] = \frac{1}{n} \sum_{i=1}^n \tilde{f}_{in}(0) m(x_i, \beta_{0,S}) m(x_i, \beta_{0,\text{full}})^t \tilde{\delta} + o(1) \quad (\text{A.4})$$

and (A3)(ii) implies

$$\lim_{n \rightarrow \infty} E[\Gamma_{n,S}] = \begin{pmatrix} Q_{01} \\ \pi_S Q_{11} \end{pmatrix} \delta. \quad (\text{A.5})$$

For the calculation of the variance of  $\Gamma_{n,S}$ , with  $r_{n,\tau}$  in (3.3) and use of (A.3) and (A4) we get

$$\begin{aligned} \text{Var}[\psi_\tau(u_{i,S})] &= \tilde{F}_{in}(0) - 2\tau \tilde{F}_{in}(0) + \tau^2 - (\tau - \tilde{F}_{in}(0))^2 \\ &= \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - r_{n,\tau}(x_i) - \left[ \tilde{F}_{in}(\Delta_i(\beta_{\text{true}})) - r_{n,\tau}(x_i) \right]^2 \\ &= \tau(1 - \tau) + r_{n,\tau}(x_i)(2\tau - 1) - (r_{n,\tau}(x_i))^2 \\ &= \tau(1 - \tau) + o(1). \end{aligned} \quad (\text{A.6})$$

Using (3.3) we have

$$\text{Var}[\Gamma_{n,S}] = \sum_{i=1}^n \tau(1 - \tau) \left( \frac{1}{n} m(x_i, \beta_{0,S}) m(x_i, \beta_{0,S})^t \right) + o(1).$$

which yields by (A3)(i)

$$\lim_{n \rightarrow \infty} \text{Var}[\Gamma_{n,S}] = \tau(1 - \tau) V_S. \quad (\text{A.7})$$

Note that, due to (A0), (A2) and (A3), the process  $\Gamma_{n,S}$  satisfies a Lindeberg-Condition. From this result and (A.5), (3.11) is then obvious.

**Proof of (3.12).** We first establish the asymptotic properties of the terms in (3.9) for  $\beta_S \in U$ . For the expectation of  $b_i(\beta_S)$ , assuming that  $\Delta_i(\beta_S) > 0$  (the

case where  $\Delta_i(\beta_S) \leq 0$  can be treated analogously) we obtain that for some  $\xi_i$  with  $|\xi_i| \leq |\Delta_i(\beta_S)|$ ,

$$E[b_i(\beta_S)] = \int_0^{\Delta_i(\beta_S)} (-s + \Delta_i(\beta_S)) \tilde{f}_{in}(s) ds = \tilde{f}_{in}(\xi_i) \frac{(\Delta_i(\beta_S))^2}{2}.$$

Note that for  $\beta_S \in U$  by (A3)(iii), we have

$$\Delta_i(\beta_S) = m(x_i, \beta_{0,S})^t \frac{v}{\sqrt{n}} + \frac{1}{2n} v^t \mathcal{M}(x_i, \tilde{\beta}_i) v = O(n^{-1/2} \|v\|) + O(n^{-1} \|v\|^2), \tag{A.8}$$

where  $\tilde{\beta}_i \in U$  denotes a suitable value between  $\beta_S$  and  $\beta_{0,S}$ . Thus, using (A.8) together with assumption (A1)(iv) we obtain

$$\begin{aligned} E\left[\sum_{i=1}^n b_i(\beta_S)\right] &= \sum_{i=1}^n \left(\tilde{f}_{in}(0) \frac{(\Delta_i(\beta_S))^2}{2}\right) + \sum_{i=1}^n \left((\tilde{f}_{in}(\xi_i) - \tilde{f}_{in}(0)) \frac{(\Delta_i(\beta_S))^2}{2}\right) \\ &= \frac{1}{2n} \sum_{i=1}^n \left(\tilde{f}_{in}(0) v^t m(x_i, \beta_{0,S}) m(x_i, \beta_{0,S})^t v\right) + O(n^{-1/2} \|v\|^3) \\ &\quad + O(n^{-1} \|v\|^4) \\ &= \frac{1}{2} v^t Q_{n,S} v + O(n^{-1/2} \|v\|^3) + O(n^{-1} \|v\|^4). \end{aligned} \tag{A.9}$$

Similarly, for the variance of  $b_i(\beta_S)$  (we again consider the case  $\Delta_i(\beta_S) > 0$ , calculations for  $\Delta_i(\beta_S) \leq 0$  yield the same result) it holds that

$$\text{Var}[b_i(\beta_S)] \leq \int_0^{\Delta_i(\beta_S)} (\Delta_i(\beta_S) - s)^2 \tilde{f}_{in}(s) ds \leq K_1 \frac{|\Delta_i(\beta_S)|^3}{3},$$

and consequently, for  $\beta_S \in U$ ,

$$\text{Var}\left[\sum_{i=1}^n b_i(\beta_S)\right] \leq K_1 \sum_{i=1}^n \frac{|\Delta_i(\beta_S)|^3}{3} = O(n^{-1/2} \|v\|^3) \tag{A.10}$$

where the last equality follows from (A3)(iii). An application of Chebychev's inequality using (A.10) yields

$$\sum_{i=1}^n b_i(\beta_S) = E\left[\sum_{i=1}^n b_i(\beta_S)\right] + O_P(n^{-1/6} \|v\|^{3/2}). \tag{A.11}$$

Finally, we determine the asymptotical behavior of the term  $R_{n,S}(\beta_S)$  for  $\beta_S \in U$ . Using (A3)(i), a similar argument as in the proof of (3.11) can be applied in order to show that  $(1/\sqrt{n}) \sum_{i=1}^n \psi_\tau(u_{i,S})$  is asymptotically normal and stochastically

bounded. Then, under (A3)(iii), one obtains  $v^t R_{n,S}(\beta_S) = O_P(n^{-1/2}\|v\|^2)$ , and this completes the proof of (3.12).

**Proof of (3.13).** We need that  $\|\hat{T}_n\| = O_P(1)$ . Note that Theorem 1 implies  $\|\hat{T}_n\|/\sqrt{n} = o_P(1)$ . It follows from (3.12) and  $P(\hat{\beta}_{n,S} \in U) \rightarrow 1$  for  $n \rightarrow \infty$ , that

$$G_n(\hat{\beta}_{n,S}) = A_n + B_n \quad (\text{A.12})$$

with

$$A_n := -\hat{T}_n^t \Gamma_{n,S} + o_P(\|\hat{T}_n\|^2) + O_P(n^{-1/6}\|\hat{T}_n\|^{3/2}) + o_P(1), \quad (\text{A.13})$$

$$B_n := \frac{1}{2}\hat{T}_n^t Q_{n,S} \hat{T}_n = O(\|\hat{T}_n\|^2). \quad (\text{A.14})$$

By (3.11), the term  $\Gamma_{n,S}$  in (A.13) is asymptotically normal and satisfies  $\hat{T}_n^t \Gamma_{n,S} = O_P(\|\hat{T}_n\|)$ . Moreover, under A1(ii) and (A3)(i) we have  $|B_n| > c\|\hat{T}_n\|^2$  for some positive constant  $c$  and  $n$  sufficiently large and  $B_n \geq 0$  due to the positive definiteness of the matrices  $Q_{n,S}$ . Observing that  $G_n(\hat{\beta}_{n,S}) \leq G_n(\beta_{0,S}) = 0$  by the definition of  $\hat{\beta}_{n,S}$ , we obtain

$$c\|\hat{T}_n\|^2 < |B_n| \leq |A_n|. \quad (\text{A.15})$$

Considering the stochastic order of the terms in (A.13), this implies  $\|\hat{T}_n\| = O_P(1)$ .

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Edited by B. Petrov and F. Csáki), 267-281. Akadémiai Kiadó, Budapest.
- Bartolucci, F. and Lupparelli, M. (2008). Focused information criterion for capture-recapture models for closed populations. *Scand. J. Statist.* **35**, 629-649.
- 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**, 981-989.
- Brownlees, C. T. and Gallo, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *J. Finan. Econometrics* **6**, 513-539.
- Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica* **62**, 405-458.
- Cade, B., Terrell, J. and Schroeder, R. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology* **80**, 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**, 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 J.* **7**, E544-E559.
- Claeskens, G. and Carroll, R. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249-265.
- Claeskens, G., Croux, C., and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics* **62**, 972-979.
- 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. L. (2003). The focussed information criterion. *J. Amer. Statist. Assoc.* **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. *J. Roy. Statist. Soc. Ser. B* **70**, 609-627.
- Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *J. Amer. Statist. Assoc.* **101**, 1449-1464.
- Hurvich, C. and Tsai, I. (1990). Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.* **9**, 259-265.
- Jureckova, J. and Prochazka, B. (1994). Regression quantiles and trimmed least squares estimator in nonlinear regression model. *J. Nonparametr. Stat.* **3**, 201-222.
- Kim, T. and White, H. (2003). Estimation, inference and specification testing for possibly misspecified quantile regression. In *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later* (Advances in Econometrics, Volume 17) (Edited by T. Fomby and R. Carter Hill), 107-132. Elsevier, New York.
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.
- Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33-50.
- Liese, F. and Vajda, I. (1994). Consistency of  $M$ -estimates in general regression models. *J. Multivariate Anal.* **50**, 93-114.
- Lim, C., Sen, P. and Peddada, S. (2010). Statistical inference in nonlinear regression under heteroscedasticity. *Sankhyā 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., Schmourer, 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**, 683-694.
- Ronchetti, E. (1985). Robust model selection in regression. *Statist. Probab. Lett.* **3**, 21-23.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.
- Shows, J. H., Lu, W., and Zhang, H. H. (2010). Sparse estimation and inference for censored median regression. *J. Statist. Plann. Inference* **140**, 1903-1917.
- Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statist. Medicine* **25**, 1369-1382.
- Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19**, 801-817.

- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika* **92**, 937-950.
- Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228-237.
- Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.* **39**, 174-200.
- Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108-1126.

Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany.

E-mail: Peter.Behl@ruhr-uni-bochum.de

KU Leuven, ORSTAT and Leuven Statistics Research Center, 3000 Leuven, Belgium.

E-mail: Gerda.Claeskens@kuleuven.be

Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany.

E-mail: Holger.Dette@rub.de

(Received March 2012; accepted March 2013)