KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Focused model selection in quantile regression

Peter Behl, Gerda Claeskens, Holger Dette

KBI_1301

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

January 25, 2013

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 and data analysis.

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

Running title: Focused model selection in quantile regression

Contact information of corresponding author: G. Claeskens, tel. +32-16-326993, Fax +32-16-326624.
Email addresses of the authors: peter.behl@ruhr-uni-bochum.de, gerda.claeskens@kuleuven.be, holger.dette@rub.de.

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 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 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.

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. On the one hand, 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. On the other hand, this of course implies that in situations when one

has a specific purpose in mind, there could be better search methods, leading toward models that are more efficient for that specific purpose. One such clear example is in phase II dose finding studies where the sole purpose of the modeling procedure is to find the minimal effective dose. In such studies, there is usually no specific interest in other aspects of the model such as predictions or variability estimation.
The focused information criterion (FIC, Claeskens and Hjort, 2003, 2008b) is designed for such 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 from different models that model for which the focus estimator has the smallest estimated mean squared error (MSE). Other loss functions than squared error could be used, e.g. linex loss (Claeskens and Hjort, 2008a) or ℓ_p loss (Claeskens et al., 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 Lupparelli, 2008), time series models (Claeskens et al., 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 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 concluding remarks are given in Section 6 and 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 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 a 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.
In dose finding studies, a typical problem is to estimate the minimal effective dose (MED), that is the smallest dose level such that a minimum effect, say Δ 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 different models, this focus takes different parametric forms. For the four given models in the current example, the focus 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. 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. The focused information criterion is constructed to select the model which estimates the MED in the best way, by taking explicitly this target into account from the start.

The aim of this paper is to derive a focused model choice criterion for quantile regression analysis, which addresses problems of this type in more generality. 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 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 β are interpreted in the same way if their argument is β_S. In order to emphasize the case where all parameters are included in the quantile regression model we use the notation $g(x; \beta_{full})$ and we define the vectors

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

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 $\hat{\beta}_{n,S}$ 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 [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 focused information criterion for quantile regression. For this purpose we assume that the following assumptions are satisfied.

(A0) The parameter space Θ is 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 $x \in \mathcal{X}$, $f_{in}(\cdot|x)$ 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 $i = 1, 2, \ldots, n$; $n \in \mathbb{N}$ the densities $f_{in}(\cdot|x_i)$ are uniformly bounded away from 0 by a constant $0 < K_0 < \infty$ in a neighbourhood $W := [g(x_i, \beta_{true}) - w, g(x_i, \beta_{true}) + w]$ of $g(x_i, \beta_{true})$.

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

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

(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 $(q-p) \times (q-p)$-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

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

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

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

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

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

(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 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 and Prochazka (1994) in order to ensure identifiability of the parameter β_0, that is

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

for all $\beta, \beta' \in \Theta$. However, 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 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)(i).

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 $\hat{\beta}_{n,S}$ in a given submodel S converges in probability to $\beta_{0,S}$. The precise statement is the following result.

Theorem 3.1 *Assume that (A0) – (A5) and (2.7) are satisfied. For any submodel S, the statistic $\hat{\beta}_{n,S}$ is a consistent estimator for $\beta_{0,S}$, i.e. $\hat{\beta}_{n,S} - \beta_{0,S} = o_P(1) \quad as \quad n \to \infty$.*

Proof. Define

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

and note that under the local alternatives (2.7) $\Delta_i(\beta_{true})$ tends to zero for $n \to \infty$. Using assumptions (A1), (A3)(iii) and (2.7), we obtain for some α satisfying $|\alpha| \leq |\Delta_i(\beta_{true})|$ and $\tilde{\beta}_i$ between β_{true} and $\beta_{0,full}$

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

where $\tilde{\delta} := (0,...,0,\delta)^t$ denotes a vector of length q which is zero in the first p components and takes the value δ from (2.7) in the last $q-p$ components. Now recall the definition of $u_{i,S}$ in (A1) and note that the estimated regression quantile $\hat{\beta}_{n,S}$ 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 first calculate the expectation of $Z_n(\beta_S)$ as

$$\begin{aligned}
E[Z_n(\beta_S)] &= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}} \left[(\tau - \mathbb{1}_{\{s\le\Delta_i(\beta_S)\}})(s-\Delta_i(\beta_S)) + (\mathbb{1}_{\{s\le 0\}}-\tau)s\right] d\tilde{F}_{in}(s) \\
&= \frac{1}{n}\sum_{i=1}^n \Big\{ -\int_{-\infty}^{\Delta_i(\beta_S)} s\, d\tilde{F}_{in}(s) + \int_{-\infty}^{0} s\, d\tilde{F}_{in}(s) + \Delta_i(\beta_S)\tilde{F}_{in}(\Delta_i(\beta_S)) - \tau\Delta_i(\beta_S)\Big\} \\
&= \frac{1}{n}\sum_{i=1}^n \Big\{ \int_{\Delta_i(\beta_S)}^{0} s\, d\tilde{F}_{in}(s) + \Delta_i(\beta_S)(\tilde{F}_{in}(\Delta_i(\beta_S)) - \tilde{F}_{in}(0)) \\
&\qquad + \Delta_i(\beta_S)(\tilde{F}_{in}(0) - \tilde{F}_{in}(\Delta_i(\beta_{true})))\Big\} \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\Delta_i(\beta_S)}^{0} (s-\Delta_i(\beta_S))\, d\tilde{F}_{in}(s) + o(1), \qquad (3.5)
\end{aligned}$$

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

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

This implies that for sufficiently large n the sum in (3.5) will only be zero for $\beta_S = \beta_{0,S}$ and is 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

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

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

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

Then from (3.6), (3.7) and (3.8) follows the statement of the Theorem, i.e. $\|\hat{\beta}_{n,S} - \beta_{0,S}\| = o_P(1)$. □

3.2 Weak convergence under local alternatives

In this section we derive the asymptotic distribution of the quantile regression estimator $\hat{\beta}_{n,S}$ 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 (A0) - (A5) and (2.7) we have*

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

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

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

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

$$\begin{aligned} \Gamma_{n,S} &:= \sum_{i=1}^{n} \psi_\tau(u_{i,S}) \frac{1}{\sqrt{n}} m(x_i, \beta_{0,S}), \qquad (3.10) \\ R_{n,S}(\beta_S) &:= \sum_{i=1}^{n} \psi_\tau(u_{i,S}) \frac{1}{\sqrt{n}} \Big[m(x_i, \tilde{\beta}_{i,S}) - m(x_i, \beta_{0,S})\Big], \\ b_i(\beta_S) &:= \mathbb{1}_{\{0<u_{i,S}\leq\Delta_i(\beta_S)\}}(\Delta_i(\beta_S) - u_{i,S}) + \mathbb{1}_{\{\Delta_i(\beta_S)\leq u_{i,S}\leq 0\}}(u_{i,S} - \Delta_i(\beta_S)) \end{aligned}$$

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

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

denotes the "derivative" of the check function ρ_τ. In the Appendix we will derive the following asymptotic properties of G_n:

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

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

  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 every $\beta_S \in U$ it holds that

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

where U denotes the neighbourhood of $\beta_{0,full}$ defined in assumption 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) will be 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 will show in the appendix that $\hat{T}_n$ is stochastically bounded, that is

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

Note that Theorem 3.1 implies that $P(\hat{\beta}_{n,S} \in U) \to 1$ for $n \to \infty$. Therefore, by (3.12) and (3.13) $G_n(\hat{\beta}_{n,S})$ has the following stochastic expansion:

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

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

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

From (3.14) and (3.15) it therefore follows that

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

Note that by the definition of $\hat{\beta}_{n,S}$ the left-hand-side of the above equation is always non-positive, while the first term in the second 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.

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

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

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

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

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" $\hat{\epsilon}_i = Y_i - g(x_i; \hat{\beta}_1, \ldots, \hat{\beta}_p)$, ($\hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimated in the full model) Kim and White (2003) suggested to estimate the matrix Q_S by

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

where $\hat{\beta}_{0,S}$ is calculated by taking estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ 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 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

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

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 $\delta\hat{\delta}^t = D_n D_n^t - \hat{K}$, 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 $\beta_{0,full}$ from the full model. Summarizing these calculations, we obtain for every submodel S the following expression for the estimated FIC which can be calculated from the data:

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

where $\gamma := (\beta_{p+1}, ..., \beta_q)^t$ denotes the last $(q-p)$ components of the parameter vector β.
The largest difficulty in specifying a focused 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 δ, the chosen focus μ 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 shall not repeat these calculations, but rather refer to that paper.

5 Finite sample properties

Different model selection methods each have their own underlying strategy. For example, the AIC aims at minimising the estimated expected Kullback-Leibler distance between the true density of the data and the density corresponding to the specified models. The BIC originates from an approximation to the Bayesian posterior probability of the model given the data, which should be large for the selected model. With these different underlying constructions, it should be no surprise that different models get selected if different types of information criteria are applied. Theoretical properties, such as efficiency, might be other

reasons for practitioners to prefer one criterion above 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 explicitly from a focus that is to be estimated and we try to find the best model for precisely this purpose. It is in such situations that the FIC can be recommended, when models are constructed for a particular purpose. Dose finding studies are ideally suited for the use of the FIC since the focus, the MED, plays the prominent role in all of the modeling process. We wish to stress, however, that the given theory and derivations extend much beyond the dose finding studies. Any focus μ that is expressible in terms of the model parameters β and that is differentiable with respect to these parameters could be taken as the starting point for the FIC. This quantity is to be used in the FIC expression (4.4) and the model with the smallest such value gets selected.

5.1 Linear quantile regression

In this section we present a simulation study for model selection by the FIC criterion in a linear quantile regression model. Moreover, we also 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 more conventional model selection criteria such as AIC and BIC. We consider the following model:

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

Here, β_0 and β_1 denote the "protected" parameters which are included in every candidate model while γ_1 to γ_4 may be included or not. Consequently, there will be 16 candidate models to choose from which all contain β_0 and β_1, but differ with respect to the parameters $\gamma_1 - \gamma_4$. For example, the narrow model only contains β_0 and β_1 while $\gamma_1 - \gamma_4$ are set to zero. The procedure starts by specifying the focus parameter, for which we chose the prediction of Y at covariate value $(x_1, z_1, z_2, z_3, z_4) = (10, 10, 10, 10, 10)$. Next, the focus is written in terms of the model notation as

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

To continue, 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. Finally, 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 are generated from model (5.1) with parameter values $\beta_0 = 1$, $\beta_1 = 1$ and $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1/\sqrt{n}$. First, a set of covariate values of size n is generated by drawing from normal distributions, that is

$$X_1 \sim N(20, 25), Z_1 \sim N(20, 6.25), Z_2 \sim N(-10, 6.25), Z_3 \sim N(10, 1), Z_4 \sim N(5, 2.25).$$

Those values are now considered as fixed and are used in all simulation runs. For the distribution of the "error" $\epsilon = Y - g(x, \beta_{true})$ we assume two different scenarios: A normal distribution with mean 0 and variance $\sigma^2 = 4$ and a Cauchy distribution with location parameter $a = 0$ and scale parameter $b = 2$. Furthermore, we consider two different sample sizes, $n = 50$ and $n = 100$. The parameters are estimated using median regression (i.e. $\tau = 0.5$). We conduct 2000 simulation runs where in each run model selection is performed using the FIC. From the chosen model, we obtain a post-selection-estimator $\hat{\mu}_{1,FIC}$ for the focus parameter μ_1.
In order to compare the FIC to more conventional information measures such as AIC and BIC, in each replication step we also estimate μ_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

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

where $\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} |y_i - g(x_i; \hat{\beta}_{n,S})|$, p denotes the number of parameters in the model S and n the number of observations [for details see Hurvich and Tsai (1990)]. For a comparison of the different model selection procedures we compute the absolute errors of the post-selection-estimators for μ_1

$$|\hat{\mu}_{1,FIC} - \mu_{true}|, \quad |\hat{\mu}_{1,AIC} - \mu_{true}|, \quad |\hat{\mu}_{1,BIC} - \mu_{true}|, \tag{5.2}$$

where $\hat{\mu}_{1,FIC}, \hat{\mu}_{1,AIC}$ and $\hat{\mu}_{1,BIC}$ denote the estimators of the focus μ_1, where the model has been chosen by FIC, AIC and BIC, respectively.
We calculate the median absolute error and the median absolute deviation (MAD) from the 2000 replications separately for FIC, AIC and BIC. The results are displayed in Table 1.

	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

Table 1: *Median and median absolute deviation (MAD) of the absolute errors of the estimates of the focus μ_1 obtained from the FIC, AIC and BIC.*

In a second setting, data are again generated from model (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 Table 1 and Table 2 it can be seen that in nearly all considered scenarios FIC either performs clearly better in terms of median absolute error and MAD or at least equally well as AIC and BIC. For both sample sizes FIC is a clear winner over both AIC and BIC if the errors are Cauchy distributed. For normal errors, FIC shows substantial advantages over AIC and BIC with parameter values $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1/\sqrt{n}$ whereas for the case $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.3$ FIC yields similar results as AIC, but is still considerably better than BIC.

	Median			MAD		
	FIC	AIC	BIC	FIC	AIC	BIC
$n = 50, \quad \epsilon \sim N(0, 4)$	2.60	2.72	3.34	1.24	1.47	1.47
$n = 50, \quad \epsilon \sim C(0, 2)$	3.42	3.96	3.83	1.59	1.98	2.12
$n = 100, \quad \epsilon \sim N(0, 4)$	2.10	1.87	2.75	1.14	1.15	1.53
$n = 100, \quad \epsilon \sim C(0, 2)$	2.90	4.05	4.01	1.38	2.04	2.38

Table 2: *Median and median absolute deviation of the absolute errors of the estimates of the focus μ_1 obtained from the FIC, AIC and BIC.*

5.2 Nonlinear quantile regression: Application of the FIC for dose-response-modeling

As a second example, we consider the class of quantile regression models introduced in Example 2.1. All results are again based on 2000 simulation runs and we consider three scenarios for the error distribution: Errors are assumed to be normal with mean 0 and variance $\sigma^2 = 0.01$, Cauchy distributed with location parameter $a = 0$ and scale parameter $b = 0.07$ or normal with a heteroscedastic variance structure. In the heteroscedastic case we assume that the errors are normal with mean 0 and standard deviation (depending on the explanatory variable x)

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

where $\tau_0 = -0.1$, $\tau_1 = 0.24$ and $\tau_2 = 0.15$. This variance function was proposed by Lim et al. (2010) for dose-response-modeling. 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_2, 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). As in the previous chapter, we compare 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.

	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 3: *Median and median absolute deviation of the absolute errors of the estimates of the focus μ_2 obtained from the FIC, AIC and BIC.*

We observe that the median absolute error of FIC is the smallest in all cases, while the BIC yields the largest median of the absolute errors. However, for the MAD the situation is not so clear. While the FIC also yields the smallest MAD for homoscedastic normal distributed errors, the BIC is superior in the case of the Cauchy distribution.
For this nonlinear example, we also compare FIC to AIC and BIC by counting how many times in 2000 simulation runs the FIC obtains a better estimator (in terms of absolute deviation) than AIC (FIC<AIC) and BIC (FIC<BIC) and vice-versa. The first row of Table 4 shows the results for homoscedastic normal distributed errors, the second row displays the results for heteroscedastic errors with variance function (5.3) 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 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).*

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.

5.3 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)]. The intercept β_4 in model (2.1) is assumed to be 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 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.

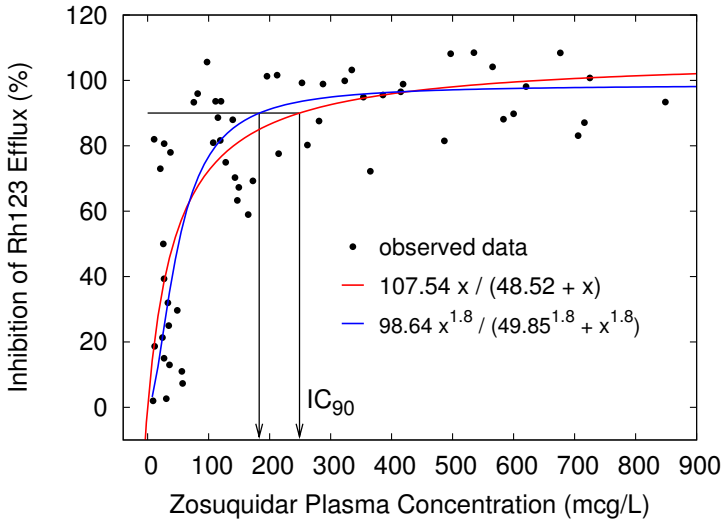

Figure 1: *Zosuquidar data with estimated median regression curves from the Hill and Michaelis Menten model.*

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 answer to the question which criterion to use depends on the research problem, often also on the preference of the researcher for one of the criteria and is not straightforward to answer in general. When a model is to be sought that gives a best performance for the estimation of a particular quantity, the focus, the FIC is a good choice since it is designed for this purpose, and the results of Section 5 indicate some improvement with respect to the precision of the focus estimate. The $\log(n)$ penalty that is used in the BIC may for large sample sizes have the effect of selecting rather small models with only few parameters. The AIC on the other hand has a possibility of overfitting, which might in some situations be advantageous in order not to fail to identify possibly important variables. A good advice would be to always consider the choice of the model search criterion in concert with the further use of the model and to take the ideas underlying the construction of the criteria into consideration.

The presented FIC is applicable in nonlinear quantile regression models in general, hence not only for minimal effective dose determination. The procedure is always the same. First, specify the focus and write it in terms of the model parameters. Estimate the MSE of the

focus estimator 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 (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. 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 $\int f(y; \theta_0, \gamma_0)|r(y)|dy < \infty$ 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.

7 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), 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). For notational convenience, define $\delta_n(\beta_S) := Z_n(\beta_S) - E[Z_n(\beta_S)]$ where Z_n is defined by (3.4). We begin with a proof of the following properties, which will be used later to establish uniform convergence of the objective function:

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

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

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

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

which follows from the Lipschitz continuity of the check function. The equicontinuity (B1) is then implied by assumption (A5) and (A0). For a proof of (B2) we introduce the notation

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

which gives

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

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

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

which is bounded by assumptions (A0) and (A5). Since the expectations of all other terms in the sum (7.2) can be similarly bounded, we obtain that $\frac{1}{n}\sum_{i=1}^n E[z_i(\beta_S)^2]$ is bounded. Therefore it follows from Chebychev's inequality that

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

which establishes (B2). The uniform convergence in (3.7) can now 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 $\beta^* \in \Theta_S$,

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

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

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

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

Proof of (3.11). Recall the definition of $\tilde{F}$ and $\tilde{f}$ in assumption (A1). A straightforward calculation yields

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

This gives for the expectation of $\Gamma_{n,S}$,

$$E[\Gamma_{n,S}] = \sum_{i=1}^{n}\left[\left(\tilde{F}_{in}(\Delta_i(\beta_{true})) - \tilde{F}_{in}(0)\right)\frac{1}{\sqrt{n}}m(x_i,\beta_{0,S})\right]. \tag{7.3}$$

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

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

Together with (7.3) and assumptions (A1)(iv), (A3)(i) and (3.3) this yields

$$E[\Gamma_{n,S}] = \frac{1}{n}\sum_{i=1}^{n}\tilde{f}_{in}(0)m(x_i,\beta_{0,S})m(x_i,\beta_{0,full})^t\tilde{\delta} + o(1) \tag{7.4}$$

and assumption (A3)(ii) implies

$$\lim_{n\to\infty} E[\Gamma_{n,S}] = v^t\begin{pmatrix} Q_{01} \\ \pi_S Q_{11}\end{pmatrix}\delta. \tag{7.5}$$

For the calculation of the variance of $\Gamma_{n,S}$ we recall the definition of $r_{n,\tau}$ in (3.3) and use (7.3) and assumption (A4) to get

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

Therefore, using (3.3) we obtain

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

which yields (by Assumption (A3)(i))

$$\lim_{n\to\infty}\mathrm{Var}[\Gamma_{n,S}] = \tau(1-\tau)v^tV_sv. \tag{7.7}$$

Note that, due to assumptions (A0), (A2) and (A3), the process $\Gamma_{n,S}$ satisfies a Lindeberg-Condition. From this result and (7.5), statement (3.11) is then obvious.

Proof of (3.12). In order to show (3.12), we are going to establish the asymptotic properties of the terms in (3.9) for $\beta_S \in U$. First, 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 with the same result) we obtain for some ξ_i with $|\xi_i| \leq |\Delta_i(\beta_S)|$

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

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

$$\Delta_i(\beta_S) = m(x_i, \beta_{0,S})^t \frac{v}{\sqrt{n}} + \frac{1}{2n} v^t \mathcal{M}(x_i, \tilde{\beta}_i) v = O(n^{-1/2}\|v\|) + O(n^{-1}\|v\|^2) \tag{7.8}$$

where $\tilde{\beta}_i \in U$ denotes a suitable value between β_S and $\beta_{0,S}$. Thus, using (7.8) together with assumption (A1)(iv) we obtain

$$\begin{aligned} E[\sum_{i=1}^n b_i(\beta_S)] &= \sum_{i=1}^n \Big(\tilde{f}_{in}(0)(\Delta_i(\beta_S)^2)/2\Big) + \sum_{i=1}^n \Big((\tilde{f}_{in}(\xi_i) - \tilde{f}_{in}(0))(\Delta_i(\beta_S)^2)/2\Big) \\ &= \frac{1}{2n}\sum_{i=1}^n \Big(\tilde{f}_{in}(0) v^t m(x_i, \beta_{0,S}) m(x_i, \beta_{0,S})^t v\Big) + O(n^{-1/2}\|v\|^3) + O(n^{-1}\|v\|^4) \\ &= \frac{1}{2} v^t Q_{n,S} v + O(n^{-1/2}\|v\|^3) + O(n^{-1}\|v\|^4). \end{aligned} \tag{7.9}$$

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

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

and consequently, for $\beta_S \in U$

$$\mathrm{Var}[\sum_{i=1}^n b_i(\beta_S)] \leq K_1 \sum_{i=1}^n \frac{|\Delta_i(\beta_S)|^3}{3} = O(n^{-1/2}\|v\|^3) \tag{7.10}$$

where the last equality follows from (A3)(iii). An application of Chebychev's inequality using (7.10) yields

$$\sum_{i=1}^n b_i(\beta_S) = E[\sum_{i=1}^n b_i(\beta_S)] + O_P(n^{-1/6}\|v\|^{3/2}). \tag{7.11}$$

Finally, we will determine the asymptotical behavior of the term $R_{n,S}(\beta_S)$ for $\beta_S \in U$. Using assumption (A3)(i) a similar argument as in the proof of (3.11) can be applied in order to show that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_\tau(u_{i,S})$ is asymptotically normal and stochastically bounded. Then, under assumption (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). It remains to prove that $\hat{T}_n$ is stochastically bounded, that is $\|\hat{T}_n\| = O_P(1)$. Note that Theorem 3.1 implies $\frac{\|\hat{T}_n\|}{\sqrt{n}} = o_P(1)$. Therefore it follows from (3.12) and the fact that $P(\hat{\beta}_{n,S} \in U) \to 1$ for $n \to \infty$ that $G_n(\hat{\beta}_{n,S})$ admits a representation

$$G_n(\hat{\beta}_{n,S}) = A_n + B_n \tag{7.12}$$

with

$$A_n := -\hat{T}_n^t \Gamma_{n,S} + o_P(\|\hat{T}_n\|^2) + O_P(n^{-1/6}\|\hat{T}_n\|^{3/2}) + o_P(1), \tag{7.13}$$

$$B_n := \frac{1}{2}\hat{T}_n^t Q_{n,S} \hat{T}_n = O(\|\hat{T}_n\|^2). \tag{7.14}$$

Note that by (3.11) the term $\Gamma_{n,S}$ which appears in (7.13) is asymptotically normal and satisfies $\hat{T}_n^t \Gamma_{n,S} = O_P(\|\hat{T}_n\|)$. Moreover, under assumptions 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 is positive 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 definiton of $\hat{\beta}_{n,S}$, we obtain the inequality

$$c\|\hat{T}_n\|^2 < |B_n| \leq |A_n|. \tag{7.15}$$

Considering the stochastic order of the terms in (7.13), this implies $\|\hat{T}_n\| = O_P(1)$.

Acknowledgements. 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/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.

Bartolucci, F. and Lupparelli, M. (2008). Focused information criterion for capture-recapture models for closed populations. *Scand. J. Statist.*, 35(4):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(5):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(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):249–265.

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. (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. and Prochazka, B. (1994). Regression quantiles and trimmed least squares estimator in nonlinear regression model. *Journal of Nonparametric Statistics*, 3(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.

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. *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.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be