AARHUS UNIVERSITY
School of Economics and Management

CREATES
Center for Research in Econometric Analysis of Time Series

Minimax Regression Quantiles

Working paper, August 2010

Stefan Holst Bache∗

Abstract

A new and alternative quantile regression estimator is developed and it is shown that the estimator is $\sqrt{n}$-consistent and asymptotically normal. The estimator is based on a minimax 'deviance function' and has asymptotically equivalent properties to the usual quantile regression estimator. It is, however, a different and therefore new estimator. It allows for both linear- and nonlinear model specifications. A simple algorithm for computing the estimates is proposed. It seems to work quite well in practice but whether it has theoretical justification is still an open question.

Keywords: Quantile regression, non-linear quantile regression, estimating functions, minimax estimation, empirical process theory
JEL Classifications: C1, C4, C5, C6

1. Introduction

The *regression quantiles* methodology, as introduced some thirty years ago by Koenker and Bassett (1978), has become a well established and popular empirical tool amongst researchers and practitioners. It compliments its more mature cousin, that of *least squares*, in providing a (sometimes) more complete picture of distributional relationships between variables of interest. Here, 'more mature' is meant to express that some modelling challenges are better understood and have elegant solutions in the least squares methodology while they still keep researchers busy with searching for quantile regression (QR) analogues. Examples present themselves when faced with selection issues, dependent data, unobserved heterogeneity, etc. Much work has been done in dealing with these (and

∗Aarhus University, School of Economics and Management, CREATES. Bartholins Allé 10, 8000 Aarhus C, Denmark. E-mail: sbache@creates.au.dk.
The Author is grateful to Jørgen Hoffmann-Jørgensen, Michael Jansson, Christian M. Dahl, and Anders Bredahl Kock for helpful comments and suggestions. Financial support by the Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation, is gratefully acknowledged

other) common problems, yet there still seems to be work to do, which is also indicated by the rapidly increasing QR literature.

The contribution of this paper is not a solution to such problems per se; rather, the paper offers a new and alternative approach to QR, which may then serve as a new 'building block' for such investigations. This alternative approach to regression quantiles is based on an artificially constructed 'deviance' criterion function, for which the minimax is a consistent solution to appropriate estimating equations for QR. The minimax approach allows for linear and non-linear regression functions (with certain regularity restrictions). One potential drawback of such minimax estimators is that the powerful linear programming algorithms, in which one usually puts ones faith to provide numerical estimates of the linear QR model, no longer apply. To remedy this, the paper provides a simple intuition-based algorithm. In practice it seems to work quite well for some problems, but the numerical aspect is a most welcomed topic for future research.

2. Some Notation

Let $(\Omega, \mathscr{F}, P)$ denote the underlying probability space. We shall investigate relations between $Y(\omega) \in \mathbb{R}$, a real continuous random variable, and $X(\omega) \in \mathscr{X}$, a k-dimensional random vector. Define $S = \mathbb{R} \times \mathscr{X}$, and let $(S, \mathscr{A})$ be the sample space with distribution law π. By y and x we will refer to the first and the last k elements, respectively, of a point $s \in S$, i.e. points in $\mathbb{R}$ and $\mathscr{X}$. The marginal distribution laws will be written as π_y and π_x respectively. The observed sample, $\{s_i\} = \{y_i, x_i\}$, with $i = 1, \ldots, n$, is assumed independent and identically distributed according to π. Denote by Θ_0 the parameter set, an analytic subset of the compact metric space (Θ, d). We set out to find a $(\mathscr{A}, \mathscr{B}_0)$-measurable criterion function $D : S \times \Theta_0 \to \mathbb{R}$, where $\mathscr{B}_0 = \mathscr{B}_0(\Theta_0)$ is the Borel σ-algebra on Θ_0, with the help of which we can estimate the parameters, $\theta_0 \in \Theta_0$, in a model for the quantiles of $Y(\omega)$, conditional on $X(\omega) = x$. Let $\mu : \mathscr{X} \times \Theta_0$ be such a (known and fixed) model for a given quantile index $\tau \in (0, 1)$. The dependence on τ is usually omitted from notation. For convenience, μ_i^θ, μ_x^θ and μ_ω^θ is sometimes used to abbreviate $\mu(x_i, \theta)$, $\mu(x, \theta)$ and $\mu(X(\omega), \theta)$. Finally, denote by $\mathbb{E}$ the expectation operator, which unless otherwise specified is with respect to π.

3. QR Estimating Functions and the Minimax Criterion

First, consider the task of computing a sample τ-quantile. For the moment we let $F(y) = P(Y(\omega) \leqslant y)$ denote the continuously differentiable distribution function of $Y(\omega)$. Given a sample, the minimiser, $\hat{q}$, of the asymmetric loss-function $\sum_{i=1}^n \rho_\tau(y_i - q)$, with $\rho_\tau(u) =$

$u(\tau - 1\{u \leqslant 0\})$, is well-known as a consistent estimator of the τth sample-quantile of $Y(\omega)$, i.e. $Q_Y(\tau) \equiv F^{-1}(\tau) = \inf\{y : F(y) \geqslant \tau\}$. The argument roughly goes that

$$(1) \qquad \frac{\mathrm{d}}{\mathrm{d}q}\left[\tau\int_q^\infty (Y-q)\mathrm{d}F - (1-\tau)\int_{-\infty}^q (q-Y)\mathrm{d}F\right] = F(q) - \tau,$$

so the population minimum is obtained at $F(q) = \tau$. Koenker and Basset's quite clever and fruitful idea, then, was to extend this to a regression setting, letting $q \equiv \mu(x,\theta)$, with x being covariates of y and $\theta \in \Theta_0$ being the unknown parameter vector of interest. The (conditional quantile) function μ is assumed to be known, and in their original paper it is linear in the parameters. It is perhaps not intuitive (to some), without the above argument, that the resulting regression curve implied by $\hat{\theta} = \hat{\theta}(\tau)$, the minimiser of the regression version of the asymmetric loss-function, 'splits' the regression residuals $\{\epsilon\}_{i=1}^n$ such that a fraction, τ, of these are non-positive. This property, however, can in some sense be taken as a 'defining property' of the quantile regression curve, and it is therefore a natural starting point. We will now explore this in more detail for quantile regression models where the (approximation to the) conditional quantile function for $Y(\omega)$ is known and of the form $Q_\tau(y|x) = \mu(x,\theta_0)$. Here, $\theta_0 = \theta_0(\tau) \in \Theta_0$ denotes the 'true' parameter (or, as μ is most likely viewed as an approximation of the true conditional quantile function, 'population solution' may be a more appropriate term).

Let

$$(2) \qquad H(s,\theta) = \tau - 1\{y \leqslant \mu_x^\theta\}$$

and define

$$(3) \qquad H(\theta) = \int_S H(s,\theta)\pi(\mathrm{d}s) = \int_S \left[\tau - 1\{y \leqslant \mu_x^\theta)\}\right]\pi(\mathrm{d}s) = \mathbb{E}\left[\tau - 1\{Y(\omega) \leqslant \mu_\omega^\theta\}\right]$$

We shall make the identifying assumption that

$$(\mathrm{i}1) \qquad H(\theta_0) = 0 \text{ and } H(\theta) \neq 0 \ \forall \theta \neq \theta_0$$

Hence, as argued, we have the natural starting point in this (unbiased) estimating function. The empirical counterpart, where dependence on the observed sample is suppressed from notation, is

$$(4) \qquad H_n(\theta) = \frac{1}{n}\sum_{i=1}^n h_i(\theta) = \frac{1}{n}\sum_{i=1}^n \left[\tau - 1\{y_i \leqslant \mu_i^\theta\}\right]$$

The term 'estimating function', usually used for $nH_n(\theta)$ and not $H(\theta)$, is here to be understood in the sense of Godambe (1960) and the vast amount of literature following

this work. An obvious question is then whether this estimating function is optimal. It is typical to consider a class of (empirical) estimating functions, where each summand is weighted, i.e. $\mathcal{G} := \{G_n : G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} a_i(\theta)h_i(\theta)\}$. The following proposition states the theoretically optimal (but unfortunately impractical) estimating function within this class, i.e. it defines the optimal $a_i(\theta)$.

Proposition: Jointly optimal estimating functions for regression quantiles. Let $\mathcal{G}$ and $h_i(\theta)$ be defined as above. Then the optimal estimating function $G_n^* \in \mathcal{G}$, is given by

$$G_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} a_i^*(\theta)h_i(\theta), \quad \text{where} \tag{5}$$
$$a_i^*(\theta) = -\frac{f_i(\mu_i^\theta)}{\tau(1-\tau)}\dot{\mu}_i^\theta.$$

Here, f_i denotes the density of $Y(\omega)$ given that $X(\omega) = x_i$; $\dot{\mu}_i^\theta$ is the vector of first-order derivatives of μ_i^θ (w.r.t. θ); and $\tau(1-\tau) = \mathrm{var}(h_i(\theta_0))$. The optimality statement is to be understood in the 'Godambe sense', which for real-valued estimating functions states that

$$\frac{\mathbb{E}\{(G_n^*)^2\}}{\{\mathbb{E}\partial G_n i^*/\partial\theta\}^2} \leqslant \frac{\mathbb{E}\{G_n'^2\}}{\{\mathbb{E}\partial G_n'/\partial\theta\}^2} \quad \forall G_n' \in \mathcal{G}, \tag{6}$$

and for vector-valued estimating functions that $\Sigma_{G_n'} - \Sigma_{G_n^*}$ is non-negative definite, where Σ_{G_n} is the variance-covariance matrix of $\{\mathbb{E}\partial G_n/\partial\theta\}^{-1}G_n$. With non-differentiable functions, as the G_ns in the present case, one exchanges differential and expectation operators, cf Godambe and Thompson (1984).

Proof. A classic result in estimating function theory (we shall not prove this) is that

$$a_i^*(\theta) = \mathbb{E}_{\pi_y}\{h_i^2(\theta_0)\}^{-1}\mathbb{E}_{\pi_y}\{\partial h_i(\theta)/\partial\theta\}, \tag{7}$$

where we treat x_i as fixed. For references, see e.g. Godambe (1985), Godambe (1987), and Ferreira (1982). To make the expression in (7) defined, we shall exchange differential and expectation operators, due to non-differentiability of $h_i(\theta)$. Write

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}}[\tau - 1\{y \leqslant \mu_i^\theta\}]\pi_y(\mathrm{d}y) = \frac{\partial}{\partial\theta}[\tau - F_i(\mu_i^\theta)] = -f_i(\mu_i^\theta)\dot{\mu}_i^\theta. \tag{8}$$

Clearly, $\mathbb{E}_{\pi_y}\{h_i^2(\theta_0)\} = \tau(1-\tau)$, and the result follows. □

This result is similar to the one by Jung (1996) for quasi likelihood estimation of the median, and that by Godambe (2001) for median estimating functions. However, the presence of the density evaluated at the quantile in the optimal estimating function is problematic, since this is unknown, and distributional assumptions are mostly unwanted.

One could consider a semi-paramtetric approach in which the density is estimated non-parametrically, but one should be careful that the estimating function remains unbiased if the density is made a random quantity. This is not dealt with in the present paper. Instead, we drop all but the derivative term from the estimating function, at the cost of some efficiency, and focus on $G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \dot{\mu}_i^\theta[\tau - 1\{y_i \leqslant \mu_i^\theta\}]$

A note on the (vector-) estimating function G_n is that it equals the vector of first-order derivatives of the classical QR objective function, where they exist. The minimiser of the latter, therefore (heuristically speaking), is *one* consistent solution to the functions G_n (but may not necessarily imply exact roots in finite samples). We shall next turn to the main result of this paper, namely *another* consistent solution: the minimax of an artificially constructed deviance function.

The minimax approach presented here is inspired by the results for a variety of estimating functions provided by Li (1996, 1997). The estimating function $G_n(\theta)$ does not, however fall into the category of functions considered there, but it will be shown here that a similar deviance function can be constructed, and that consistency is preserved. The proof, however, requires a small extension. We shall present the full consistency proof for completeness.

Let $G(s,\theta) = \dot{\mu}_x^\theta[\tau - 1\{y \leqslant \mu_x^\theta]$. From the discussion of estimating functions, we have that (only) the parameter of interest, θ_0, will satisfy the vector-valued estimating equation

$$G(\theta_0) = 0, \text{ where} \tag{9}$$
$$G(\theta) = \int_S G(s,\theta)\pi(\mathrm{d}s) = \int_S \dot{\mu}(x,\theta)[\tau - 1\{y \leqslant \mu(x,\theta)\}]\pi(\mathrm{d}s),$$

where, again, $\dot{\mu}$ is the vector of first-order derivatives with respect to θ. The empirical counterpart, which we derived above, is

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \dot{\mu}_i^\theta(\tau - 1\{y_i \leqslant \mu_i^\theta\}) \tag{10}$$

Solving $G_n(\theta) = 0$ is not particularly practical and it may not have an exact solution. The claim, however, is that we can make use of the following function (from $S \times \Theta \times \Theta$ into $\mathbb{R}$):

$$D(s,\theta,\vartheta) = [\mu(x,\vartheta) - \mu(x,\theta)][\tau - 1\{y \leqslant \bar{\mu}(\theta,\vartheta)\}], \quad \text{where} \tag{11}$$
$$\bar{\mu}(\theta,\vartheta) = [\mu(x,\theta) + \mu(x,\vartheta)]/2.$$

Let

$$D(\theta,\vartheta) = \int_S [\mu(x,\vartheta) - \mu(x,\theta)][\tau - 1\{y \leqslant \bar{\mu}(\theta,\vartheta)\}]\pi(\mathrm{d}s). \tag{12}$$

Further, define that

$$(13)\quad D(\theta,\vartheta)=\begin{cases}\infty, & \text{if } \theta\in\Theta\backslash\Theta_0\\ -\infty, & \text{if } \theta\in\Theta_0,\vartheta\in\Theta\backslash\Theta_0.\end{cases}$$

Make the following observations:

(i) $D(\theta_0,\vartheta)<0\ \forall\vartheta\neq\theta_0$.

(ii) $D(\theta,\theta)=0$.

(iii) $D(\theta,\vartheta)=-D(\vartheta,\theta)$.

We now impose an additional, but quite weak identifying assumption[1]:

$$(\text{i2})\quad \forall\theta\neq\theta_0\ \exists\ \vartheta\ :\begin{cases}\mu(x,\theta)>\mu(x,\vartheta)>2\mu(x,\theta_0)-\mu(x,\theta) & \text{if } \mu(x,\theta)>\mu(x,\theta_0)\\ \mu(x,\theta)<\mu(x,\vartheta)<2\mu(x,\theta_0)-\mu(x,\theta) & \text{if } \mu(x,\theta)<\mu(x,\theta_0).\end{cases}$$

This allows us to add the following to the list above:

(iv) $\forall\theta\neq\theta_0\ \exists\vartheta:\ D(\theta,\vartheta)>0$.

Observation (i) follows from the definition of $\mu(x,\theta_0)$ and τ; (ii)-(iii) are evident; and to realise (iv), fix x and choose ϑ according to (i2). This will make $D(\theta,\vartheta)$, conditional on x, positive. By the law of iterated expectations, (iv) then follows. From these observations, it follows that

$$(14)\quad \sup_{\vartheta\in\Theta_0} D(\theta_0,\vartheta)=\inf_{\theta\in\Theta_0}\sup_{\vartheta\in\Theta_0} D(\theta,\vartheta),$$

or in other words:

$$(15)\quad \theta_0=\operatorname*{arg\,inf}_{\theta\in\Theta_0}\sup_{\vartheta\in\Theta_0} D(\theta,\vartheta).$$

Also, we have that

$$(16)\quad \inf_{\theta\in\Theta_0}\sup_{\vartheta\in\Theta_0} D(\theta,\vartheta)=0.$$

[1]For example, including a continuous intercept parameter in the model will ensure that the assumption is valid.

A final, but important note, is the connection between $D(s,\theta,\vartheta)$ and the estimating function $G(s,\theta)$. Now, where the derivative exists, we have that

$$G(s,\theta) = \left.\frac{\partial D(s,\theta,\vartheta)}{\partial \vartheta}\right|_{\vartheta=\theta} \tag{17}$$

The empirical criterion is given by

$$D_n(\theta,\vartheta) = \frac{1}{n}\sum_{i=1}^{n}[\mu_i^{\vartheta} - \mu_i^{\theta}][\tau - 1\{y_i \leqslant \bar{\mu}_i\}] \tag{18}$$

and based on the aforementioned properties, it is natural to estimate θ_0 by

$$\widehat{\theta}_n := \underset{\theta\in\Theta_0}{\arg\inf} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) \tag{19}$$

As Li also points out, the properties of such deviance functions are closely related to properties of the likelihood-ratio. With these constructed deviance functions, however, one gets similar properties without assuming the existence of a likelihood function. Some applications, including the present, do not have a likelihood function as a natural starting point. More generally, estimating functions need not be the score or derivative of any function. Further, we need not require differentiability of D_n, and it is therefore consistent even when the estimating function (when viewed as derivative of D_n) does not exist. Consistency of the minimax estimator is given in theorem 1 below. The asymptotic distribution is given in Theorem 2.

Theorem 1. Consistency of the minimax estimator: Under the assumptions (i1), (i2) and those in Section 2, the minimax estimator

$$\widehat{\theta}_n = \underset{\theta\in\Theta_0}{\arg\inf} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) \tag{20}$$

is consistent, i.e. $\widehat{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$.

Proof: Let $D(\theta) := D(\theta,\theta_0) = \mathbb{E}D_n(\theta,\theta_0)$ and $D_n^{\circ}(\theta) = D_n(\theta,\theta_0) - D(\theta)$.

First, we shall consider the uniform convergence properties of $D_n^{\circ}(\theta)$, which are essential to the proof. To this end, we will utilise some nice results from the theory of empirical processes. Let $\mathscr{C}$ be a collection of subsets of the sample space S. This collection is said to *pick out* a certain subset, Z say, of a finite set $S^{(n)} = \{s_1,\ldots,s_n\} \subset S$ if it can be written as $Z = S^{(n)} \cap C$, for some $C \in \mathscr{C}$. If $\mathscr{C}$ picks out all of the possible 2^n subsets, then it is said to shatter $S^{(n)}$. Let $V(\mathscr{C})$ be the smallest n such that no set of size n is shattered by C. If $V(\mathscr{C})$ is finite, then $\mathscr{C}$ is called a Vapnik-Červonenkis class (or VC-class). A class of real-valued measurable functions on S is said to be a VC-subgraph class if the collection of subgraphs of these functions forms a VC-class of sets in $S \times \mathbb{R}$. We shall find that

the relevant class of functions for D_n is a VC-subgraph class. First, the class of indicator functions $\mathscr{I} := \{1\{y \leqslant (\mu(\theta,x)+\mu(\vartheta,x))/2\} : s \in S, (\theta,\vartheta) \in \Theta_0 \times \Theta_0\}$ is a classic example of a VC-subgraph class. Let $\widetilde{\mu}(\theta,\vartheta,x) = \mu(\vartheta,x) - \mu(\theta,x)$. Now, τ and $\widetilde{\mu}$ are fixed functions from $\Theta_0 \times \Theta_0 \times S$ into $\mathbb{R}$, and thus by Lemma 2.6.18 (iv)-(v) of Van der Vaart and Wellner (1996), the classes $-\mathscr{I} := \{-\iota : \iota \in \mathscr{I}\}$ and $\tau - \mathscr{I} := \{\tau - \iota : \iota \in \mathscr{I}\}$ are VC-subgraph classes. Finally, (v) of the same lemma gives that $\mathscr{D} := \{D_n : s \in S, (\theta,\vartheta) \in \Theta_0 \times \Theta_0\}$ is a VC-subgraph class. This property implies that it is uniformly Glivenko-Cantelli in π (sometimes called a GC-π class), i.e. it is a sufficient condition for the following to hold:

$$\sup_{D_n \in \mathscr{D}} |D_n - D| \to 0 \text{ a.s.}, \quad \text{and thus} \tag{21}$$

$$\sup_{\theta \in \Theta_0} |D_n^\circ(\theta)| \to 0 \text{ a.s.} \tag{22}$$

Now, we will derive the result that for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left\{ \inf_{\theta\in\Theta_0} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) < \varepsilon \right\} = 1. \tag{23}$$

Then we show that this is contradicted if the assertion of the theorem is false. Let $T \subseteq \Theta_0$ and note that

$$\inf_{\theta\in T} D_n(\theta,\theta_0) \geqslant \inf_{\theta\in T} D_n^\circ(\theta) + \inf_{\theta\in T} D(\theta). \tag{24}$$

Since $-\sup_{\theta\in T}|D_n^\circ(\theta)| \leqslant \inf_{\theta\in T} D_n^\circ(\theta) \leqslant \sup_{\theta\in T}|D_n^\circ(\theta)|$, then by (22) the first term on the right-hand side of (24) converges in probability to 0. So, for any $\varepsilon > 0$ we now have that

$$\lim_{n\to\infty} P\left\{ \inf_{\theta\in T} D_n(\theta,\theta_0) \geqslant \inf_{\theta\in T} D_n^\circ(\theta) + \inf_{\theta\in T} D(\theta), \quad \left|\inf_{\theta\in T} D_n^\circ(\theta)\right| < \varepsilon \right\} = 1. \tag{25}$$

If $T = \Theta_0$, then $\inf_{\theta\in T} D(\theta) = 0$ and (25) implies

$$\lim_{n\to\infty} P\left\{ \sup_{\vartheta\in\Theta_0} D_n(\theta_0,\vartheta) < \varepsilon \right\} = \lim_{n\to\infty} P\left\{ \inf_{\theta\in\Theta_0} D_n(\theta,\theta_0) > -\varepsilon \right\} = 1. \tag{26}$$

The first equality is due to the anti-symmetry of D_n. This confirms the validity of (23).

Let O be a small open ball centered at θ_0. Now, if the theorem was false we would have $\limsup_{n\to\infty} P\{\widehat{\theta}_n \notin O\} > 0$. We shall use this to contradict (23). Let $T = \Theta_0 \backslash O$ in (25), which then implies

$$\lim_{n\to\infty} P\left\{ \inf_{\theta\notin O} D_n(\theta,\theta_0) > \delta \right\} = 1, \tag{27}$$

since we can take $\varepsilon = \delta$ and $\inf_{\theta \in T} D(\theta) = 2\delta > 0$ by (i2) and the properties of D and μ. This, in turn, implies

$$(28) \quad \lim_{n\to\infty} P\left\{ \inf_{\theta \notin O} \sup_{\vartheta \in \Theta_0} D_n(\theta, \vartheta) > \delta \right\} = 1.$$

Now, write

$$(29) \quad \limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta_0} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) \geqslant \delta \right\} \geqslant$$
$$\limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta_0} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) = \inf_{\theta\notin O} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta), \quad \inf_{\theta\notin O} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) > \delta \right\}.$$

By (28) the right-hand side reduces to

$$(30) \quad \limsup_{n\to\infty} P\left\{ \inf_{\theta\in\Theta_0} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) = \inf_{\theta\notin O} \sup_{\vartheta\in\Theta_0} D_n(\theta,\vartheta) \right\} \geqslant \limsup_{n\to\infty} P\left\{ \widehat{\theta}_n \notin O \right\} > 0.$$

This is in contradiction to (23) and proves the theorem. □

Theorem 2. Asymptotic normality of the minimax estimator: Let f_i and F_i denote the density and distribution functions of $Y(\omega)$ given $X = x_i$. Write μ_i for $\mu(x_i, \theta_0)$ and μ_x for $\mu(x, \theta_0)$. Assume that

(i) $0 < f_{Y|X=x} < M \ \forall x \in \mathscr{X}$, for some $M \in \mathbb{R}_+$.

(ii) $V \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \dot{\mu}_i \dot{\mu}_i^\top = \int_{\mathscr{X}} \dot{\mu}_x \dot{\mu}_x^\top \pi(\mathrm{d}x)$ has full rank.

(iii) $\int_{\mathscr{X}} \left\| \dot{\mu}(X(\omega), \theta) \right\|^2 \pi(\mathrm{d}x) < \infty \quad \forall \theta \in \Theta_0$.

(iv) $\theta_0 \in \text{interior}(\Theta_0)$.

Then, $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \tau(1-\tau)\Gamma^{-1} V \Gamma^{-1})$, where

$$\Gamma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f_i(F_i^{-1}(\tau)) \dot{\mu}_i \dot{\mu}_i^\top.$$

Proof: We show that the conditions for Theorem 3.3 of Pakes and Pollard (1989) are satisfied. We will refer to these as (p-i) – (p-v) to avoid confusion with the assumptions of this theorem. Let $D_L(\theta)$ denote the vector of left partial derivatives of D_n with respect to ϑ, evaluated at $\vartheta = \theta$. Since $\widehat{\theta}_n$ is the optimiser of D_n, the summands of $D_L(\widehat{\theta}_n)$ for which $y_i \neq \mu(x_i, \widehat{\theta}_n)$ cancel out. Thus, the only 'important' contributions to $G_n(\widehat{\theta}_n)$ occur at the

'kinks' $y_i = \mu_i(x_i, \widehat{\theta}_n)$. We can therefore bound the absolute value of the jth coordinate of $G_n(\widehat{\theta}_n)$ by

$$|G_{n,j}(\widehat{\theta}_n)| \leqslant \max_{1\leqslant i\leqslant n} \left\{\left\|\dot{\mu}(x_i, \widehat{\theta}_n)\right\|\right\} \frac{1}{n}\sum_{i=1}^{n} 1\{y_i = \mu(x_i, \widehat{\theta}_n)\}(\tau - 1) \tag{31}$$

By assumption (iii), $\max_{1\leqslant i\leqslant n}\{\|\dot{\mu}(x_i, \widehat{\theta}_n)\|\} \in o_p(n^{1/2})$, and since $Y(\omega)$ is continuously distributed, the normalised sum is $O_p(n^{-1})$.

Combined, we therefore have that $|G_{n,j}(\widehat{\theta}_n)| \in o_p(n^{-1/2})$, and thus also $\|G_n(\widehat{\theta}_n)\| \in o_p(n^{-1/2})$. This argument is similar to one made by Honoré (1992) and shows that (p-i) is satisfied.

Next, write $G(\theta)$ as

$$G(\theta) = \int_S \dot{\mu}_x^\theta [\tau - 1\{y \leqslant \mu_x^\theta\}]\pi(\mathrm{d}s) = \int_{\mathscr{X}} \dot{\mu}_x^\theta [\tau - F_{Y|X=x}(\mu_x^\theta)]\pi_x(\mathrm{d}x) \tag{32}$$

Assumption (iii) allows us to differentiate under the integral sign (Billingsley, 2008), so:

$$\Gamma \equiv \frac{\partial G}{\partial \theta}\Big|_{\theta=\theta_0} = -\int_{\mathscr{X}} f_{Y|X=x}(\mu_x)\dot{\mu}_x\dot{\mu}_x^\top \pi_x(\mathrm{d}x) = -\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f_i(\mu_i)\dot{\mu}_i\dot{\mu}_i^\top, \tag{33}$$

which by (i) and (ii) exists and is finite. This concludes on (p-ii).

Regarding the condition (p-iii), we will use that the function class $\mathscr{I} := \left\{1\{y_i \leqslant \mu_i^\theta\} : s_i \in S, \theta \in \Theta_0\right\}$ is universally P-Donsker, and so is $-\mathscr{I}$ and $\{\tau - \mu_i^\theta \iota : \iota \in \mathscr{I}\}$. The property is retained under addition, cf Van der Vaart and Wellner (1996, Theorem 2.10.6 and example 2.10.7). Therefore, for all positive sequences $\{\delta_n\}$ with $\delta_n \in o(1)$,

$$\sup_{||\theta-\theta_0||\leqslant\delta_n} \frac{||G_n(\theta) - G(\theta) - G_n(\theta_0)||}{n^{-1/2} + ||G_n(\theta)|| + ||G(\theta)||} \in o_p(1), \tag{34}$$

cf Chen et al. (2003, Section 4).

Finally, for condition (p-iv), the iid assumption allows an application of the standard central limit theorem:

$$G_n(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, J), \quad \text{where} \tag{35}$$
$$J = \tau(1-\tau)\int_{\mathscr{X}} \dot{\mu}_x\dot{\mu}_x^\top \pi(\mathrm{d}x)$$
$$= \tau(1-\tau)V.$$

Condition (p-v) is simply assumption (iv). Thus, all conditions for Theorem 3.3 of Pakes and Pollard (1989) are satisfied, and the result of this theorem follows.

□

4. Example: The linear model

As an illustration, let $\mu(x,\theta) = x^\top\theta$. Then,

$$D_n(\theta,\vartheta) = \frac{1}{n}\sum_{i=1}^{n} x_i^\top[\vartheta-\theta][\tau - 1\{y_i \leqslant (x_i^\top[\vartheta+\theta])/2\}]. \tag{36}$$

The limiting distribution is then

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &\xrightarrow{\mathrm{d}} \mathcal{N}(0, \tau(1-\tau)\Gamma_l^{-1}H_l\Gamma_l^{-1}) \qquad (37)\\ \Gamma_l &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f_i(F_i^{-1}(\tau))x_i x_i^\top\\ H_l &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top.\end{aligned}$$

It is seen that the asymptotic distribution of the minimax estimator is equivalent to that of the usual QR estimator (Koenker, 2005, Section 3.2.3). However, the estimator is not necessarily coinciding with the QR estimator numerically and it may give different estimates.

5. Behavior of the deviance function

The preceding section presented the deviance criterion function and showed the consistency and asymptotic normality of the minimax estimator $\widehat{\theta}_n$. We now inspect the function a little more closely closer to get an idea of the behaviour it displays. At first, it appears intractable, since it is piecewise linear with discontinuities. It does, however, reveal some properties that might be exploited for numerical aspects, something we will return to later. First, what does it look like? For ease of graphical exposition, consider the one-dimensional case of estimating a sample quantile. Then we have $\Theta_0 \subset \mathbb{R}$ and no covarites. Figure 1 shows $D_n(\theta,\vartheta)$ as a function of ϑ for three different fixed values of θ, given the sample points $\{-5,-4,\ldots,5\}$, with $\tau = ½$. As it is seen, the further away θ is from the sample median, the easier it becomes to choose ϑ in order to get a large (positive) function value. As θ approaches the sample median, the function "flattens", and the supremum function value decreases. When θ equals the sample median, here $\theta = 0$, the best one can do, in terms of choosing ϑ is to set $\vartheta = \theta$ and get a function value of 0. Also, note that when $\theta < \theta_0$, the maximum function value is achieved by setting $\vartheta > \theta$, and vice versa. This gives an indication of the direction in which one should direct θ to find, in this case, the sample median.

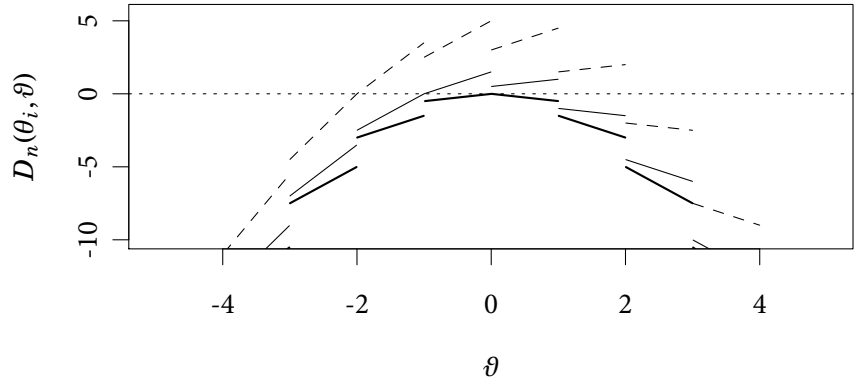

Figure 1: The deviance function $D_n(\theta, \vartheta)$ as a function of ϑ for three different θs. The dashed line represents $\theta_1 = -2$; the solid is for $\theta_2 = -1$; and when $\theta_3 = 0$, the sample median, we get the solid bold graph.

An alternative geometric interpretation is as follows. Again, let us stick with the simplest possible case of finding a sample quantile. Let

$$\Delta(\vartheta) = \#\left\{i \ : y_i \leqslant (\theta + \vartheta)/2\right\} \tag{38}$$

be the number of observations with value less than $\bar{\mu}_i = (\theta + \vartheta)/2$, viewed as a function of ϑ. Then, if the sample is ordered with $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n$, and we fix $\theta = \theta^*$, we can write

$$\begin{aligned} nD_n(\vartheta) = nD_n(\theta^*, \vartheta) &= [\vartheta - \theta^*] \sum_{i=1}^{\Delta(\vartheta)} (\tau - 1) + [\vartheta - \theta^*] \sum_{i=\Delta(\vartheta)+1}^{n} \tau \\ &= [\vartheta - \theta^*][n\tau - \Delta(\vartheta)]. \end{aligned} \tag{39}$$

Given a θ, then ϑ is chosen to maximise the area of a rectangle where one side is increasing in ϑ and the other is decreasing. The job is then to chose a θ for which no ϑ can make this area positive. In higher dimensions, it is a little more complicated. As the $[\mu_i^\vartheta - \mu_i^\theta]$ term no longer can be taken outside the sums, we have many rectangles for which the sides are determined by a common parameter vector. θ must then be chosen such that no ϑ can make the sum of the areas positive (where some can be negative). This does not as such get us closer to how one would go about computing this value of θ. The next section gives a suggestion on how to do this in practice.

6. Computing the minimax regression quantiles.

It is not the aim of this paper to develop a definitive algorithm for computing the minimax estimates. However, there is a simple intuition-based algorithm, by which is meant that convergence is not theoretically guarantied as of yet, but it seems to work quite well in practice. The procedure is discussed briefly below, and further work on numerical aspects of the estimator is encouraged by the author.

In the one-dimensional case, we saw that for a given value of the θ, then by evaluating whether $D_n(\theta, \vartheta)$ takes on positive values for ϑ to the left or to the right of θ we will know in which direction to move θ. A simple algorithm, then, is the following:

(i) first trap the minimiser, $\widehat{\theta}_n$, of $\sup_{\vartheta \in \Theta_0} D_n(\theta, \vartheta)$ beween two points. One point where positive values occur to the left, and one where they occur to the right.

(ii) then, bisect the interval iteratively while updating the boundaries.

(iii) At some convergence-level, stop.

Actually, as seems evident from the discussion, the optimising value of θ is located at a sample-point, i.e $\widehat{\theta}_n = y_i$ for some i. Hence, instead of (iii), simply trap a single sample-point in such an interval, and this point will be the solution.

In high dimension, i.e. $k \geqslant 2$, it may or may not be justified to use a component-wise version of this scheme, i.e. to fix $k-1$ parameters, where $\theta_j = \vartheta_j$ for these, and use the above procedure for the remaining parameter. Then, move on to the next and repeat for all k parameters. Of course, there is a dependence between parameter values, and one should not expect to find the solution by only cycling through the parameters once. However, although I have as yet no theoretical justification, believing that a number of cycles will lead you to the solution is almost irresistible. And it turns out, that in practice this is the case in many situations where the initial guess for θ is reasonable.[2]

The first step of trapping the parameter in some interval is not too hard in many situations: if $\mu(x, \theta)$ is monotone in the parameters, then let $\varepsilon > 0$ be a small number and set $\vartheta_j = \theta_j + \varepsilon$ for some j and $\vartheta_j = \theta_j$ for the remaining $k-1$ parameters. Then if $D_n(\theta, \vartheta) > 0$, one has found the left boundary, otherwise one has found the right. Now choose a value, for the parameter of interest, sufficiently far in the relevant direction to make $D_n(\theta, \vartheta)$ change sign. Now, the other boundary has been found.

Of course, the nature of the specific μ may induce specific considerations, but the method just described often sufficient. The convergence criterion for the bisection steps could initially be set rather loose to quickly get within a reasonable neighbourhood of the solution, and then tightened after the first couple of cycles.

[2] At least for those situations experienced by the author.

7. Directions and motivation for future research.

The minimax estimator of this paper offers a new way to perceive regression quantiles: the estimating functions described provide an intuitive description of the quantity of interest as the roots of these functions. They do not, however, give a practical way to obtain this quantity as they may not have exact roots in finite samples. The deviance function is constructed to give *consistent* roots to these functions and it has restated the problem as that of a certain type of extremum estimators. This is interesting in itself, as it provides a quantile regression framework which incorporates a wide class of regression functions. The framework, as a building block, may prove useful in alleviating some problems related to quantile regression. To round off the discussion, here are two observations which could find use in further developing the ideas of this paper.

First, panel models with dependent data pose trouble for quantile regression estimators, see e.g. Bache et al. (2010) for a review. In the estimating function literature, this is often dealt with using 'generalized estimating equations'. Li (1997) presents a minimax framework for such estimating equations. It is quite possible, in the spirit of these findings and this paper, that given an $n \times T$ panel, one could construct a function, say

$$D_n(\theta,\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{2}W_i^{-1}(\alpha(\theta))[\mu_i^{\vartheta}-\mu_i^{\theta}][\tau - 1\{y_i \leqslant \mu_i^{\theta}\}] + \frac{1}{2}W_i^{-1}(\alpha(\vartheta))[\mu_i^{\vartheta}-\mu_i^{\theta}][\tau - 1\{y_i \leqslant \mu_i^{\vartheta}\}]\Big\}, \tag{40}$$

where W_i is a specified (up to an estimable parameter α that may depend on θ) $T \times T$ working variance-covariance matrix, and y_i, μ_i, and the indicators are T-vectors.

A second potential is for non-linear quantile functions. Here, with classical non-linear QR, one encounters numerical difficulties, since linear programming is no longer an option. Further, some quantile regression models have criterion functions which are flat, and therefore hard to optimise. Examples are Powell's censored regression quantiles (1986) and Manski's maximum score-type estimators (1975, 1985). Perhaps, using the minimax approach, one could construct non-flat criterion functions which could lead to easier optimisation. Also, the numerical aspect of the minimax estimator is still uncharted territory and could very well turn out some positive and useful directions.

References

Bache, S. H., Dahl, C. M., Kristensen, J. T., 2010. Headlights on tobacco road to low birthweight: Evidence from a battery of quantile regression estimators. Working Paper.

Billingsley, P., 2008. Probability and measure. Wiley-India.

Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71 (5), 1591–1608.

Ferreira, P., 1982. Multiparametric estimating equations. Annals of the Institute of Statistical Mathematics 34 (1), 423–431.

Godambe, V., 1960. An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31 (4), 1208–1211.

Godambe, V., 1985. The foundations of finite sample estimation in stochastic processes. Biometrika 72 (2), 419–428.

Godambe, V., 1987. The foundations of finite sample estimation in stochastic processes - ii. Bernouilli 2, 49–59.

Godambe, V., 2001. Estimation of median: Quasi-likelihood and optimum estimating functions. Hypothesis, 07.

Godambe, V., Thompson, M., 1984. Robust estimation through estimating equations. Biometrika, 115–125.

Honoré, B., 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. Econometrica: Journal of the Econometric Society 60 (3), 533–565.

Jung, S., 1996. Quasi-Likelihood for Median Regression Models. Journal of the American Statistical Association 91 (433).

Koenker, R., 2005. Quantile regression. Cambridge Univ Pr.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica: journal of the Econometric Society 46 (1), 33–50.

Li, B., 1996. A minimax approach to consistency and efficiency for estimating equations. The Annals of Statistics, 1283–1297.

Li, B., 1997. On the consistency of generalized estimating equations. Lecture Notes-Monograph Series 32, 115–136.

Manski, C., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3 (3), 205–228.

Manski, C., 1985. Semiparametric analysis of discrete response:: Asymptotic properties of the maximum score estimator. Journal of Econometrics 27 (3), 313–333.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica: Journal of the Econometric Society 57 (5), 1027–1057.

Powell, J., 1986. Censored regression quantiles. Journal of Econometrics 32 (1), 143–155.

Van der Vaart, A., Wellner, J., 1996. Weak convergence and empirical processes. Springer Verlag.