The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Quantile Curves without Crossing

Xuming He

To cite this article: Xuming He (1997) Quantile Curves without Crossing, The American Statistician, 51:2, 186-192, DOI: 10.1080/00031305.1997.10473959

To link to this article: https://doi.org/10.1080/00031305.1997.10473959

Statistical Computing and Graphics

Quantile Curves Without Crossing

Xuming HE

Conventional regression techniques focus on the conditional averages, but often of more interest are lower or upper conditional quantiles. A more informative description of the relationship among variables can be obtained through regression quantiles (RQ). The percentile curves are usually computed one level at a time. Associated with great flexibility is the embarrassing phenomenon of quantile crossing. We propose a restricted version of regression quantiles (RRQ) that avoids the occurrence of crossing while maintaining sufficient modeling flexibility. RRQ remains in the general framework of the regression quantiles both conceptually and computationally. Because it relates all quantile functions through the conditional median, it also means substantial savings in computation costs when multiple quantiles for high-dimensional data are needed. Two examples in connection with the physical and engineering sciences are presented to demonstrate the usefulness of RRQ curves.

KEY WORDS: Linear models; Regression quantiles; Smoothing; Splines.

1. INTRODUCTION

Regression quantiles introduced by Koenker and Bassett (1978) provide a novel approach to the statistical analysis of parametric and nonparametric response models by supplementing the almost exclusive focus of least squares based methods on the estimation of conditional mean functions with a general technique for computing percentile curves suggested by Mosteller and Tukey (1977). Linear models have been receiving most of the attention where regression quantiles form the basis for data ordering, efficient L-estimation, and regression rank score tests. For example, Ruppert and Carroll (1980), Portnoy and Koenker (1989), and Welsh (1991) constructed L-estimators of linear models based on regression quantiles, and Gutenbrunner and Jurečková (1992) constructed regression rank scores as a dual to quantiles. Efron (1990) gave several interesting examples where regression percentiles are useful on their own sake.

For example, the lifetime of a certain product or equipment may depend on the conditions (say temperature) under which it is used. One could use conventional regression techniques to determine how the average duration varies with temperature, but it would only tell part of the story. Typically of more interest is how long the majority of such products can be used, that is, the lower percentiles of the lifetime as a function of temperature.

In public health and medical sciences, growth charts are frequently used as a means of establishing whether some measure of interest on an individual lies within a normal range. In discussing Cole (1988) on fitting centile curves to reference data, Cox and Jones proposed a form of quantile splines to the problem. A more thorough study on quantile smoothing splines is made by Koenker, Ng, and Portnoy (1994). A closely related approach using B-spline approximations on selected knots was studied in He and Shi (1994). In a similar spirit Chaudhuri (1991) employed local polynomials for nonparametric estimates of conditional quantile functions.

Given n pairs of observations $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ the essence in nonparametric estimation of quantile curves of the various forms mentioned in the preceding paragraph is to compute the 100αth percentile function f via minimization of $\sum_{i=1}^{n} \rho_\alpha(y_i - f(x_i))$ subject to a certain smoothness constraint on f, where $\rho_\alpha(x) = \alpha x^+ + (1 - \alpha)x^-$, $x^+ = \max\{x, 0\}$, and $x^- = \max\{0, -x\}$. If f lies in a linear space, say $f(x) = \beta' B(x)$ where $B(x)$ is the vector consisting of all the basis functions of the linear space, the minimization problem can be rewritten as

$$\sum_{i=1}^{n} (\alpha r_i^+ + (1 - \alpha) r_i^-) = \text{minimum}$$

subject to

$$r_i^+ \geq 0, \qquad r_i^- \geq 0, \qquad r_i^+ - r_i^- = y_i - \beta' B(x_i).$$

This is a linear programming problem for the expanded parameters β and r_i^+, r_i^- $(i = 1, 2, \ldots, n)$. The availability of efficient linear programming algorithms makes quantile computations highly competitive.

The computational issues aside, at least four desirable features make this approach natural and effective as a general method of computing percentile curves. First, the quantile regression is distribution-free in the sense that no assumption on the conditional distribution of Y given $X = x$ needs to be made. Second, actual measurements on the covariate are used in the calculation, eliminating measurement

Xuming He is Associate Professor, Department of Statistics, University of Illinois, Champaign, IL 61820. The author thanks an associate editor and two referees for their helpful suggestions and comments on an earlier draft, and is grateful to Pin Ng for his assistance in installing an S-PLUS function. This work was supported in part by National Security Agency Grant MDA904-96-1-0011.

errors associated with rounding off or grouping necessary for some other methods. Third, the quantile regression tends to automatically adjust for varying deviation along the X direction, making it a useful tool to detect heteroscedesticity. Last, it ensures, without any calibration, that at each percentage level the quantile curve so obtained divides the observations into the desired ratio.

One disturbing problem with quantile regression is that quantile curves so obtained can cross one another. When this happens a given point (x_0, y_0) may be classified as below the 20th but above the 30th percentile! The crossing phenomenon is already observed in the simple linear regression setting. It occurs more frequently in multiple regression problems.

In a way, crossing of quantile curves reflects a paucity of data in the region concerned. Because the percentile curves are expected to converge to the true conditional quantile functions as the number of observations increases, the crossing phenomenon will eventually disappear (as the sample size increases). However, it is of practical importance to avoid confusion by any chart readers who use the percentile curves at a finite sample level to do ranking or to construct prediction intervals.

The root of the problem to crossing, in my view, is that quantile curves are computed individually so as to consistently estimate the conditional quantile functions in a broader class of models. But for a dataset of modest size this class of quantile curves is so diverse that variability of the estimates become large enough to upset the monotonicity enjoyed by their population counterparts. Even a large dataset may contain scarce observations in certain regions where quantile crossing is likely to occur.

The crossing problem may be avoided by forcing proper ordering of the percentile curves. In general, this remains a challenging task. Koenker (1984) considered parallel quantile planes for linear models, but they do not cater to the needs of heteroscedestic models. A successful approach used by Cole (1988) in smoothing reference centile curves is to work under the assumption that a suitable power transformation would render the underlying distribution of the measurements normal. Cole and Green (1992) described in detail the method of maximum penalized likelihood to provide smooth estimates of the three curves in the LMS method: the Box–Cox power transformation (L), the mean or median function (M), and the coefficient of variation (S). More specifically, the LMS method assumes that the observed y is positive and

$$z = \frac{(y/M(x))^{L(x)} - 1}{L(x)S(x)}$$

has a standard normal distribution where $L(x)$, $M(x)$, and $S(x)$ are smooth functions of x. As usual, z is taken to be $\log(y/M(x))/S(x)$ when $L(x) = 0$.

Strictly speaking, the transformed variable z in the above form cannot be a normal variable for most values of $L(x)$. This is because $y > 0$ implies a lower bound of $z > -1/(L(x)S(x))$. In practice, it is viewed as an approximation in the range of observations. To find the αth quantile one simply takes

$$C_\alpha(x) = M(x)(1 + L(x)S(x)z_\alpha)^{1/L(x)}, \qquad L(x) \neq 0,$$

or

$$C_\alpha(x) = M(x)\exp[S(x)z_\alpha], \qquad L(x) = 0$$

where z_α is the αth quantile of the standard normal distribution. As indicated earlier, such quantile curves are meaningful over the set $\{x : 1 + L(x)S(x)z_\alpha > 0\}$. To estimate the three curves L–M–S used in the transformation Cole and Green (1992) proposed to use maximum likelihood with a roughness penalty in the form of

$$a_1 \int (L''(x))^2\,dx + a_2 \int (M''(x))^2\,dx + a_3 \int (S''(x))^2\,dx$$

where a_j's are the smoothing parameters.

The LMS method has proven to be useful in applications to anthropometry measurements and other variables of a similar nature. Each curve carries useful information about the distributions of the measurements at different levels of the covariate. On the other hand, it relies on a Box–Cox power transformation to induce normality, so the classical debate about the latter applies. For example, several issues may have to be considered carefully if the response variable does not always take positive values. The choice of the transformation is also sensitive to outliers in the data; see Carroll (1982) for some examples on linear models. Determining the effective dimensions in smoothing not just one but three curves here is no easy job.

In this paper, we propose a simpler alternative that requires some commonality of adjacent quantiles without restricting ourselves to any specific class of error distributions. This is made possible with the regression quantile methodology applied to an appropriate subclass of heteroscedestic models. The space of submodels should be sufficiently large to handle most of the practical problems, yet small enough to avoid crossing of empirical quantile regression curves/surfaces. We assume that the response variable Y is real-valued, and the covariate X is p-dimensional. Let $f_\alpha(x)$ denote the conditional αth quantile of Y given $X = x$. We start with linear models to fix the idea, even though nonparametric models may be more practical in some applications.

2. LINEAR MODELS: THE BASIC IDEA

If $f_\alpha(x)$ is a linear function of x for all $\alpha \in (0, 1)$, there are only two possibilities: the domain D_X of the covariate X over which the model holds is bounded or the quantile planes are all parallel. As usual, we assume that X contains 1 as its first component (allowing intercept in the model). To motivate the restricted regression quantiles we focus on the subclass of linear heteroscedastic models

$$y = x'\beta + (x'\gamma)e \tag{1}$$

where e is any error distribution. We assume that all of the conditional quantiles are linear, and therefore $x'\gamma$ does not

change signs for $x \in D_X$. For model (parameter) identifiability we assume without loss of generality that e has median 0, $|e|$ has median 1, and $x'\gamma > 0$ for all x in the interior of D_X.

The linear heteroscedastic model of type (1) has been studied by Gutenbrunner and Jurečková (1982) and Koenker and Zhao (1994) among others. One could argue that they are of limited use as they are applicable only when all of the conditional quantiles are linear. In practice, one may prefer to estimate the scale function differently. The basic ideas of the restricted regression quantiles (RRQ), however, can be conveyed with the simple model (1) for which we propose to compute RRQ as follows.

1. Compute the median regression of Y on X to obtain residuals r_i and coefficient $\hat{\beta}$.
2. Regress $|r_i|$'s on x_i's to obtain the median regression coefficient $\hat{\gamma}$ and fitted values s_i.
3. Find c_α by minimizing $\sum_i \rho_\alpha(r_i - cs_i)$ over c, and obtain the αth quantile plane as $y = x'(\hat{\beta} + c_\alpha\hat{\gamma})$.

Note that RRQ's can be viewed as empirical regression quantiles for a reduced regression-through-the-origin model. If $x'\hat{\gamma}$ does not change sign for all x values concerned, the resulting RRQ planes maintain proper ordering at each value of the covariate by the following.

Proposition 1. Given any set of n pairs $\{(s_i, r_i), i = 1, 2, \ldots, n\}$ with $s_i \geq 0$ for each i and $\sum_{i=1}^n s_i > 0$, the largest (if not unique) c_α that minimizes $\sum_{i=1}^n \rho_\alpha(r_i - cs_i)$ over $c \in R$ is nondecreasing in $\alpha \in (0, 1)$.

Proof. For each $\alpha \in (0, 1)$ it follows from the basic properties of linear programming that the line $r = c_\alpha s$ fits at least one point in $\{(s_i, r_i)\}$. Note that the directional derivative of $\sum_i \rho_\alpha(r_i - cs_i)$ at $c = c_\alpha$ must be nonnegative. This implies that for each $\alpha \in (0, 1)$ $\alpha \sum_{r_i \geq cs_i} s_i \geq (1 - \alpha) \sum_{r_i < cs_i} s_i$ for any $c \leq c_\alpha$. Now suppose $1 > \beta > \alpha > 0$. If $c_\beta < c_\alpha$, there always exists $c^* \in (c_\beta, c_\alpha]$ such that no points fall into the region $c_\beta s < r < c^* s$. It is then clear that

$$
\begin{aligned}
0 &> \sum_i \rho_\beta(r_i - c_\beta s_i) - \sum_i \rho_\beta(r_i - c^* s_i) \\
&\geq (c^* - c_\beta)\left(\beta \sum_{r_i \geq c^* s_i} s_i - (1 - \beta) \sum_{r_i < c^* s_i} s_i\right) \\
&\geq (c^* - c_\beta)\left(\alpha \sum_{r_i \geq c^* s_i} s_i - (1 - \alpha) \sum_{r_i < c^* s_i} s_i\right) \geq 0
\end{aligned}
$$

which is a contradiction.

Proposition 1 may be more easily understood by noting that c_α can be viewed as a weighted quantile of a univariate sample. The median regression in step 2 does not guarantee the fitted function to be nonnegative. If one really believes in the linear model (1), it is better to impose constraints that all fitted values $s_i \geq 0$. The good news is that such constraints pose no computational problems as they simply add linear constraints to a linear program for minimization. However, I suggest to first examine the fitted function without such constraints. If the fitted s-function changes signs in the interior point of the domain of the covariate, it may be taken as diagnostic evidence of the nonlinearity of the quantile functions. In this case efforts need to be made to find a more appropriate model through transformations (Carroll and Ruppert 1988) or to use a nonparametric fit (see Section 3 below).

Note that RRQ is not a weighted version of the usual RQ algorithm. After steps 1 and 2 are performed, additional quantiles are obtained as if we had a simple regression-through-the-origin model regardless of the dimensionality of X. Because of this, RRQ inherits the finite sample properties of the RQ for the reduced model. Some of the basic properties of RRQ can be summarized as follows. Their derivations and proofs follow the work of Koenker and Bassett (1978) and Koenker and Zhao (1994); we omit the details.

(P1) As a result of linear programming used in the minimization 3, each RRQ plane passes through at least one data point.

(P2) There exists a partition $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{k_n} = 1$ for some $k_n \leq n$ such that the αth quantile planes remain the same for all $\alpha \in (\alpha_{i-1}, \alpha_i), (i = 1, 2, \ldots, k_n)$. There are at most n distinct RRQ planes, as with order statistics for the univariate sample.

(P3) Under model (1) we have, as n tends to infinity, $\hat{\beta} \to \beta, c_{.5}\hat{\gamma} \to \gamma$, and the αth quantile is consistent for the true conditional quantile function $x'\beta + F^{-1}(\alpha)x'\gamma$, where $F^{-1}(\alpha)$ is the αth quantile of the error distribution of e.

(P4) Let $\underline{N}_n(\alpha)$ and $\bar{N}_n(\alpha)$ be the number of data points below and above the αth quantile, respectively. Then $\lim_{n\to\infty}(\underline{N}_n(\alpha)/n) \leq \alpha$ and $\lim_{n\to\infty}(\bar{N}_n(\alpha)/n) \leq 1 - \alpha$. In other words, the αth quantile divides the data at least asymptotically in the ratio of $\alpha/(1 - \alpha)$.

One advantage of the regression quantile approach we use is that for model (1), both RQ and RRQ are consistent estimates of the true conditional quantile functions without having to know or to estimate the error distribution in the model.

We now use the idea of RRQ in an experimental study of automobile braking considered in Hald (1960). There are $n = 50$ observations of $x =$ initial speed and $y =$ stopping distance. As Hald (1960) pointed out, there are theoretical grounds for using the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + (\gamma_1 x_i)e_i. \tag{2}$$

The argument is briefly given as follows. The stopping time is mainly determined by the speed at the moment of the signal to stop and the reaction time of the driver. If the speed of the car is x miles per hour and the reaction time of the driver is β_1 hours, the car will proceed $\beta_1 x$ miles before the brake is used. According to the simple physical law, the distance to stop from the moment the brake is used is proportional to the square of the speed. This justifies a quadratic relation in the model. Repeated experiments with the same car and driver reveal that the variation mainly comes from the varying reaction time of drivers. If the reaction time is normally distributed, the conditional variance of the stopping

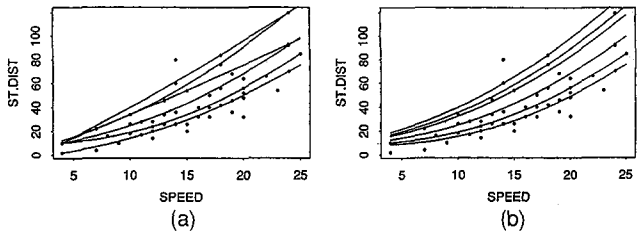

Figure 1. (a) RQ for Automobiles. (b) RRQ for Automobiles.

time would then be in the form of $(\gamma_1 x)^2$ for some constant γ_1. This specific model can be considered as a special case of model (1) by identifying $(1, x, x^2)$ as the covariate vector with $\gamma_0 = \gamma_2 = 0$. The computation of the RRQ curves can be obtained only with some trivial modifications in step 2 above to reflect the built-in constraints that $\gamma_0 = \gamma_2 = 0$.

For safety concerns we are obviously more interested in the upper conditional quantiles of the stopping distance. We give the unrestricted quantile curves for $\alpha = .25, .5, .75, .85, .9, .95$ in Figure 1A, and the RRQ curves at the same percentage levels in Figure 1B. In both cases the quantile curves divide the data into two parts (below and above) by the desired ratios. Under the postulated model (2), both RQ and RRQ curves are consistent estimates of the population quantiles, but RRQ makes more use of the information contained in the model and looks more sensible. The quantile crossing in the RQ plot is mainly a result of larger variability in the individual estimates.

To improve on the asymptotic efficiency one may recalculate $\hat{\beta}$ in step 1 by a weighted median regression using weights estimated in step 2. The median regression needed in steps 1 and 2 may also be replaced by an M estimator (Huber 1981), but care has to be taken to achieve scale equivariance. We omit the details.

3. PERCENTILE CURVES

The same idea applies readily to computation of nonparametric conditional quantile functions. We consider a class of submodels

$$y = f(x) + s(x)e \tag{3}$$

where e represents error, and f and s are two unspecified functions. We restrict ourselves here to a unidimensional covariate $x \in [0, 1]$, but unlike the LMS method described in Section 1, a generalization of RRQ to multivariate covariates is straightforward. Without loss of generality we assume $s(x) \geq 0$ for all x. Suppose that a procedure for estimating the conditional median function is already in order. The αth RRQ curve can then be computed as follows.

1. Obtain the conditional median function $\hat{f}$ of Y given X, and let $r_i = y_i - \hat{f}(x_i)$;
2. Smooth $\{(x_i, |r_i|)\}$ into a nonnegative function $\hat{s}(x)$;
3. Find c_α which minimizes $\sum_i \rho_\alpha(r_i - c_\alpha \hat{s}(x_i))$ over $c \in (-\infty, \infty)$, and obtain the RRQ function as $\hat{f} + c_\alpha \hat{s}$.

The computation of RRQ curves here depends on a particular choice of median smoothing in step 1. The same smoothing technique may be used in step 2, but it is not required. The quantile smoothing spline of Ng, Koenker, and Portnoy (1994) may be used for univariate regressors. The concept of RRQ, however, is not restricted to any specific smoothing technique in use.

The "response" values in step 2 are all nonnegative, so it is generally not difficult to come up with a nonnegative fitted function on the interval of observed data. Any regression fit of the form $\hat{s}(x) = \sum_i w(x)|r_i|$ with $w(x) \geq 0$ and $\sum_i w(x) = 1$ satisfies this requirement. Included are the kernel smoothing and nearest neighbor methods. We shall also see how the quantile smoothing splines of Koenker et al. (1994) preserve positivity.

In generic terms the median smoothing spline $\hat{g}$ is obtained by minimization of

$$\sum_i |y_i - g(x_i)| + \lambda V(g') \tag{4}$$

where $V(\cdot)$ denotes the total variation norm of a function, and λ is a smoothing parameter balancing between fidelity to data and smoothness penalty. Assume without loss of generality that $x_1 < x_2 < \cdots < x_n$. We have the following.

Proposition 2. If $y_i \geq 0$ for each i, then the median smoothing spline of (4) with boundary constraints $g(x_1) \geq 0$ and $g(x_n) \geq 0$ is nonnegative on the whole interval (x_1, x_n).

Proof. By Koenker, Ng, and Portnoy (1994) the solution to (4) is a continuous piecewise linear function on intervals of the form (x_i, x_{i+1}). The smoothing spline is nonnegative on (x_1, x_n) if and only if all fitted values are nonnegative. Now if $g(x_k) < 0$ for some k, redefining the piecewise linear segments by simply changing $g(x_k)$ to 0 would decrease the fidelity function without increasing the penalty in (4). Therefore the solution to (4) must be a nonnegative piecewise linear function.

It was pointed out to me by Professor Esther Portnoy through her experience in applying the RRQ method in her actuarial work that the boundary constraints in Proposition 2 are necessary to ensure positivity.

The result of Proposition 2 remains valid for linear B-spline approximations considered in He and Shi (1994), provided that the knots are selected from the observed covariate values. If higher order splines are used, constraints in the form of $g(x_i) \geq 0$ are needed at all points x_i. Or one may replace $s_i = \hat{s}(x_i)$ by $\max\{s_i, \min_j\{s_j > 0\}\}$ in step 2. If it is believed that there might be an area where the conditional variance is identically zero, one may use $\max\{s_i, 0\}$ instead.

To illustrate the use of RRQ curves we consider the well-known motorcycle data; see Table 1 of Härdle (1989, p. 302) for the tabulation of the observations. The covariate X denotes time (in milliseconds) after a simulated impact with motorcycles. The response Y is the head acceleration of a post-mortem human test object. A total of 133 measurements was taken. The largest covariate value of 65.6 is rather far away from the rest of the data, so we do not include this point. Härdle (1989, p. 57) did the same to lessen the edge effect of a mean smoothing spline fit. The removal of this point also helps the RQ curves to be less likely to cross each other. In this example we shall use B-spline approximations for estimating nonparametric functions.

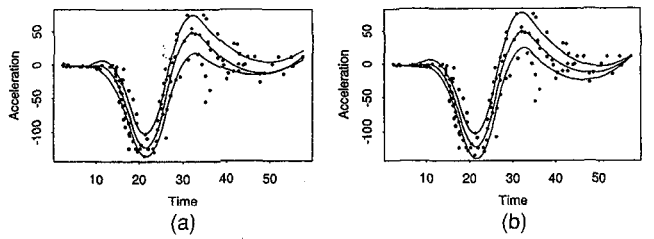

Figure 2. (a) RQ for Motorcycles. (b) RRQ for Motorcycles.

Given a set of knots $\min\{x_i\} = t_0 \leq t_1 \leq \cdots \leq t_k = \max\{x_i\}$, let $N_1(x), N_2(x), \ldots, N_{k+2}(x)$ be the normalized quadratic B-spline basis functions defined on the extended knots; see Schumaker (1980) or He and Shi (1994) for details. We approximate the αth conditional quantile function by $\sum_{j=1}^{k+2} a_j N_j(x)$, and determine the constants a_j's by minimizing $\sum_{i=1}^{n} \rho_\alpha(y_i - \sum_{j=1}^{k+2} a_j N_j(x_i))$.

To be more specific, we choose equally spaced knots $t_j = \min\{x_i\} + j(\max\{x_i\} - \min\{x_i\})/k$ for $j = 0, 1, \ldots, k$, and the number k is first selected by minimizing the following information criterion:

$$\log\left(\sum_i \rho_\alpha(y_i - \hat{g}(x_i))\right) + 3(k+2)/n \qquad (5)$$

where $\hat{g}$ is the fitted function. Once the number k is determined we perform a (backward) stepwise deletion until no single knot removal would lower the value of (5). The information criterion used here aims to balance the fidelity to the data with the smoothness of the function, and it has been proven to be quite successful in our studies. Similar strategies have been used by other authors in spline smoothing; see, for example, Kooperberg and Stone (1992) and Friedman and Silverman (1989).

For clarity we plot only the three quartiles (.25, .50, and .75) for the motorcycle data. Figure 2A shows the RQ curves, and Figure 2B gives the RRQ curves. Behaving very similarly to the RQ curves for the intermediate values of X (time), the RRQ curves clearly indicate that the speed is virtually constant before the motorcycle hits an object. The measurements are more variable in the middle where the motorcycle gets bounced back and forth forcefully before it starts to calm down. Although the three RQ curves plotted here are quite far apart in percentage levels, they already start to suggest the problem of crossing.

Note that the B-spline knots used in this example were chosen by a general information criterion on equally spaced knots without any fine tuning. The method described here applies to other applications as well.

As mentioned in Section 1 the LMS method also yields percentile curves without the crossing problem. A direct comparison of these two methods is rather difficult as they are motivated by and designed for different classes of regression models. The LMS method assumes normality for a transformed response, whereas the RRQ approach is geared for location-scale models without distributional assumptions. Furthermore, the method of Cole and Green (1992) is not completely automated as it requires the users to specify the smoothing parameters or equivalent degrees of freedom. The degree of smoothing also needs to be chosen for the spline method used to compute RRQ, but an explicit selection criterion is given in (5). Both approaches are flexible in the sense that the user has his/her own choice of smoothing techniques. Despite the difficulties we conducted some "unfair" simulation-based comparisons to gain some insights into how the RRQ curves measure up.

The first model we considered falls into the framework of the LMS method, that is,

$$y_i = (\log(x_i) + e_i)^2 - 10, \qquad i = 1, 2, \ldots, 50 \qquad (6)$$

where $x_i = i/50$ and e_i's are independent $N(0,1)$ variables. The L–M–S curves in this case are $L(x) = 1/2, M(x) = (\log x)^2$, and $S(x) = -2/\log x$. The variable $y + 10$ may be taken as response in the model.

In the spirit of LMS we assume that the transformation is "estimated" exactly right, and a smoothing spline is obtained on the transformed data where the smoothing parameter is selected by cross-validation (see the S-PLUS function smooth.spline). For convenience we call this procedure super-LMS. The B-spline approximation described above with automatic knot selection is used to get the RQ and RRQ curves. We measure the quality of the fit by taking the mean-square error as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{f}_\alpha(x_i) - f_\alpha(x_i))^2$$

where $\hat{f}_\alpha$ is the estimated αth quantile function.

A total of 500 Monte Carlo samples was used in estimating the MSE for three methods (super-LMS, RQ, and RRQ) at several different levels of α. The results are summarized in Table 1 for $\alpha = .5, .75, .9$. The numbers given inside parentheses denote the standard errors of the estimates.

Obviously this comparison gives several unfair advantages to the LMS method. The model is not of type (3), the error distribution is normal for which the mean regression based on the least squares principle tends to be more efficient than the median regression, and the variability in estimating the transformation in the LMS method (which could be substantial) is not taken into account. In view of these disadvantages the RRQ does not perform badly at all. The MSE's of RRQ are higher by no more than 70% for the α values we considered between .1 and .9. Note that the variability of the regression quantile would increase indefinitely as α approaches 0 or 1, but the situation is much better for RRQ than for RQ.

Next, we consider a simpler regression model that satisfies the assumptions of both the LMS and RRQ models, that is,

$$y_i = x_i^3 + e_i, \qquad i = 1, 2, \ldots, 50 \qquad (7)$$

Table 1. Estimated MSE Values for Model (6)

α	*Super-LMS*	*RQ*	*RRQ*
.50	1.755 (.053)	2.548 (.065)	2.548 (.065)
.75	3.119 (.094)	4.878 (.193)	4.051 (.115)
.90	4.449 (.161)	10.251 (.364)	7.356 (.240)

Table 2. Estimated 100* MSE's for Model (7)

α	Super-LMS	RQ	RRQ
.50	.688 (.032)	.732 (.027)	.732 (.027)
.75	.650 (.028)	.862 (.032)	.833 (.030)
.90	.670 (.030)	1.641 (.065)	1.259 (.048)

where x_i's are randomly selected from the uniform distribution on (0, 1) and e_i's have the common distribution of $N(0, 1/16)$. The L–M–S curves in this case are $L(x) = 1, M(x) = x^3$, and $S(x) = 1/(4x^3)$. Again, we compare RRQ with super-LMS. We choose the regression function to be cubic in favor of the smoothing splines used in super-LMS. The smoothing spline is indeed a cubic spline, whereas quadratic splines are used when the RQ and RRQ curves are estimated. Based on a Monte Carlo sample size of 500 we obtained the results of Table 2. Somewhat surprisingly, the RRQ method was almost as good at $\alpha = .5$. Even at α as large as .9 the MSE of the RRQ curve is still below two times of that of the super-LMS.

In reality the LMS curves would not do all that well in the comparison because the transformations have to be estimated and the transformed "response" will not be exactly normal. Furthermore, there is room for improvement in the computation of the RRQ curves. The simulation studies indicate that no matter how well the transformations are estimated for LMS, good relative performance can be expected from the RRQ curves in the quality of fit for levels of α not too close to 0 or 1. As we have seen, the RRQ curves are much easier to compute even for the univariate cases. The B-spline approximation used to compute the RRQ curves generalize easily to multivariate quantile functions simply by taking the tensor-product splines; see Stone (1995) and He and Shi (1996) for more details. It is much more difficult to use the LMS method when there are several covariates.

For extreme quantiles when α is close to 0 or 1 the RRQ curves may have increasing variability just like the usual RQ curves. This cannot be avoided without knowing the form of the error distribution in the model, but the variability of the RRQ curves is often substantially lower.

4. CONCLUDING REMARKS

The RRQ curves proposed in the present paper are easy to compute, but do not suffer from the problem of quantile crossing. The regression quantile methodology enables us to do this at least for location-scale models without assuming any specific form of the underlying distribution. The idea can be applied to any smoothing technique for uni- or multivariate regression.

Automatic choice of smoothing parameters is an important part of any smoothing technique. For a large dataset, it is often very demanding in memory requirement and computation time. The problem becomes far more evident when the covariate is multidimensional. When multiple percentile curves are needed the proposed RRQ offers a dramatic reduction in computation costs. This advantage may be sufficient to offset its limitation on flexibility as compared to the usual unrestricted quantiles, not to mention that the latter can often leave us with the perplexing phenomenon of quantile reversal in certain areas. We believe that the proposed RRQ has its best potential in computing multiple quantiles when there are not sufficiently dense data in certain regions to justify greater flexibility of RQ and when the number of regressors in a linear model is large.

The crossing of RQ curves may partly be attributed to several different sources such as the choice of knots in the B-spline approximations or the well-known edge effect in curve fitting. However, by examining the problem in a linar regression setting we come to believe that the problem can only be solved by imposing certain restrictions on the space of possible solutions to conditional quantiles. The proposed method in the present paper is one implementation of such restrictions.

If the observations (x_i, y_i) constitute a random sample from the model (1) or (3), the proposed RRQ functions are clearly consistent. A refined asymptotic theory demands further research. We note that the usefulness of the quantile curves is far beyond the usual iid framework assumed for most asymptotic analyses. On the other hand, the RRQ, as the name implies, carries a restriction. If the error distributions in our motivating models still depend strongly on x, the results may not be satisfactory.

[Received June 1995. Revised July 1996.]

REFERENCES

Carroll, R. J. (1982), "Two Examples of Transformations When There are Possible Outliers," *Applied Statistics*, 31, 149–152.

Carroll, R. J., and Ruppert, D. (1988), *Transformations and Weighting in Regression*, New York: Chapman and Hall.

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *Annals of Statistics*, 19, 760–777.

Cole, T. J. (1988), "Fitting Smoothed Centile Curves to Reference Data," *Journal of the Royal Statistical Society*, Ser. A, 151, 385–418.

Cole, T. J., and Green, P. J. (1992), "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood," *Statistics in Medicine*, 11, 1305–1319.

Efron, B. (1991), "Regression Percentiles Using Asymmetric Squared Error Loss," *Statistica Sinica*, 1, 93–125.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–21.

Gutenbrunner, C., and Jurečková, J. (1992), "Regression Rank Scores and Regression Quantiles," *Annals of Statistics*, 20, 305–330.

Hald, A. (1960), *Statistical Theory with Engineering Applications*, New York: Wiley.

Härdle, W. (1989), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

He, X., and Shi, P. (1994), "Convergence Rate of B-Spline Estimators of Nonparametric Conditional Quantile Functions," *Journal of Nonparametric Statistics*, 3, 299–308.

——— (1996), "Bivariate Tensor-Product B-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Koenker, R. (1984), "A Note on L-Estimators for Linear Models," *Statistics and Probability Letters*, 2, 323–325.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

Koenker, R., and Zhao, Q. (1994), "L-Estimation for Linear Heteroscedastic Models," *Journal of Nonparametric Statistics*, 3, 223–235.

Kooperberg, C., and Stone, C. J. (1992), "Logspline Density Estimation

for Censored Data," *Journal of Computational and Graphical Statistics*, 1, 301–328.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression: A Secondary Course in Statistics*, Reading, MA: Addison-Wesley.

Portnoy, S., and Koenker, R. (1989), "Adaptive L-Estimation for Linear Models," *Annals of Statistics*, 17, 362–381.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

Schumaker, L. L. (1980), *Spline Functions: Basic Theory*, New York: Wiley.

Stone, C. (1994), "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation," *Annals of Statistics*, 22, 118–184.

Welsh, A. H. (1991), "Asymptotically Efficient Adaptive L-Estimators in Linear Models," *Statistica Sinica*, 1, 203–228.