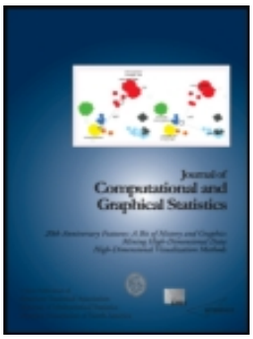

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Practical Confidence Intervals for Regression Quantiles

Masha Kocherginsky, Xuming He & Yunming Mu

To cite this article: Masha Kocherginsky, Xuming He & Yunming Mu (2005) Practical Confidence Intervals for Regression Quantiles, Journal of Computational and Graphical Statistics, 14:1, 41-55, DOI: 10.1198/106186005X27563

To link to this article: https://doi.org/10.1198/106186005X27563

Published online: 01 Jan 2012.

Practical Confidence Intervals for Regression Quantiles

Masha KOCHERGINSKY, Xuming HE, and Yunming MU

Routine applications of quantile regression analysis require reliable and practical algorithms for estimating standard errors, variance-covariance matrices, as well as confidence intervals. Because the asymptotic variance of a quantile estimator depends on error densities, some standard large-sample approximations have been found to be highly sensitive to minor deviations from the iid error assumption. In this article we propose a time-saving resampling method based on a simple but useful modification of the Markov chain marginal bootstrap (MCMB) to construct confidence intervals in quantile regression. This method is compared to several existing methods with favorable performance in speed, accuracy, and reliability. We also make practical recommendations based on the `quantreg` package contributed by Roger Koenker and a new package `rqmcmb2` developed by the first two authors. These recommendations also apply to users of the new SAS procedure PROC QUANTREG, available from Version 9.2 of SAS.

Key Words: Confidence interval; Markov chain marginal bootstrap; Regression quantile; Standard error.

1. INTRODUCTION

The least squares method dominates regression analysis, but quantile regression, as introduced by Koenker and Bassett (1978), has gradually emerged as a powerful complement. In the context of linear models, the least squares method assumes that the conditional mean $E(Y|X = x)$ is a linear function of x. We assume that Y is real-valued and X is R^p-valued. Unless one is willing or able to make strong distributional assumptions on $Y|X$, the least squares analysis does not provide any information beyond the conditional mean. A more comprehensive approach to the statistical analysis of linear models is to analyze the τth quantile of Y given $X = x$, where $\tau \in (0, 1)$. The conditional median corresponds to $\tau = .5$.

Masha Kocherginsky is Research Associate, Department of Health Studies, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637. Xuming He is Professor, and Yunming Mu is Research Assistant, at Department of Statistics, University of Illinois, 725 S. Wright, Champaign, IL 61820 (E-mail: x-he@uiuc.edu).

Journal of Computational and Graphical Statistics, Volume 14, Number 1, Pages 41–55
DOI: 10.1198/106186005X27563

Assume that the τth conditional quantile function is

$$Q_Y(\tau|x) = x'\beta_\tau \tag{1.1}$$

for some parameter vector $\beta_\tau \in R^p$. An estimate $\hat{\beta}_\tau$ of β_τ based on a sample of (x_i, y_i) $(i = 1, \ldots, n)$ is obtained by solving

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta) = \text{minimum}, \tag{1.2}$$

where $\rho_\tau(r) = r(\tau - I(r < 0))$, and $I(r < 0)$ is the indicator for $r < 0$. In the special case of $\tau = .5$, we have $\rho_{.5}(r) = .5|r|$, corresponding to the least absolute deviation regression.

The asymptotic normality for the quantile estimate $\hat{\beta}_\tau$ was established by Koenker and Bassett (1978) under the assumption of iid error models, that is,

$$y_i = x_i'\beta_\tau + e_i, \tag{1.3}$$

where e_i are iid variables with the τth quantile at 0. Asymptotic representations were given by He and Shao (1996) for independently, but not necessarily, identically distributed errors. If e_i are independent with the probability density functions f_i, then the asymptotic variance-covariance matrix of $\hat{\beta}_\tau$ is given by

$$V_\tau = (\tau(1-\tau))(X'FX)^{-1}(X'X)(X'FX)^{-1}, \tag{1.4}$$

where X is the n by p design matrix whose columns are x_i and $F = \text{diag}\{f_1(0), \ldots, f_n(0)\}$. In an iid error model, the expression (1.4) reduces to

$$V_\tau = (\tau(1-\tau)/f^2(0))(X'X)^{-1}, \tag{1.5}$$

where $f(0)$ is the common error density evaluated at 0.

Direct methods of estimating V_τ are available, and large-sample confidence intervals for β_τ can then be constructed based on asymptotic normality. Koenker (1994) discussed some simple ways to estimate $1/f(0)$ and recommended specific choices of the smoothing parameter needed there. It was found through a Monte Carlo experiment that the results based on (1.5) are nonrobust against heteroscedastic errors. A number of alternative methods for constructing confidence intervals were considered by Koenker (1994) for iid error models and later by Koenker and Machado (1999) for heteroscedastic error models. Existing research suggests that the rank-score method used by Koenker (1994) (using inversion of regression rank tests) is a promising approach to the construction of confidence intervals. It circumvents any explicit estimation of the error densities and is resilient under a broad class of models. However, it is practically feasible only for confidence intervals of one-dimensional parameters, and it does not produce an estimate of the variance-covariance matrix. Even for component-wise confidence intervals, the method becomes computationally prohibitive for large datasets.

Another approach to the inference of regression quantiles is to use resampling. Classical bootstrap methods such as pairwise bootstrap are quite reliable, but they are time-consuming

for moderate to large datasets. In this article, we adapt the Markov chain marginal bootstrap (MCMB) approach with a simple but very helpful modification to estimate the variance-covariance matrix of the quantile estimates $\hat{\beta}_\tau$ and to show, through a carefully designed experiment, that the proposed method is a preferred choice for moderate to large datasets.

In this article, the comparisons of different methods are based on the lengths and the coverage probabilities of confidence intervals. We note here that speed is also an important factor in deciding which approach to recommend. For a test dataset of $n = 10{,}000$ observations and $p = 50$ coefficients, we used the `quantreg` package in R (Version 1.81) on a 2.4 GHz Dell computer with 512 MB of RAM. For the median regression with $\tau = .5$, the default option using the Barrodale-Roberts simplex algorithm used 5.4 seconds of CPU time to get the estimate. The faster algorithm, based on an interior point method (Portnoy and Koenker 1997), produced the estimate in 2.4 seconds. Depending on which algorithm is available, several minutes may be needed to perform the usual bootstrap with 100 bootstrap samples, compared to about 3.6 seconds to get a direct estimate of V_τ in the form of (1.4). An attempt to use the rank-score method kept the computer busy for 29 minutes. The MCMB method, using the `rqmcmb2` package in R, produced the variance-covariance matrix with 100 resamples in 31 seconds.

This experiment indicates that, for large datasets, the rank-score method might be less attractive with the typical computer power available today, but the direct method and the MCMB method with a relatively short chain are much more competitive from the timing perspective. The Monte Carlo experiment described in this article shows that the MCMB method is statistically satisfactory, as it is highly competitive with both the rank-score method and the pairwise bootstrap, but often more reliable than the direct method of estimating V_τ.

The rest of the article is organized as follows. We review some existing methods in Section 2 and detail a simple but useful modification to the MCMB method in Section 3. The modified method will be called the MCMB-A method. Time-saving strategies are provided for the implementation of the MCMB-A approach, making it substantially faster than both the usual bootstrap methods and the rank-score method for large datasets. A carefully designed Monte Carlo study is reported in Section 4, and practical recommendations are given in Section 5 for making use of various methods of inference for quantile regression.

An R package `rqmcmb2` based on the the MCMB-A method for quantile regression was recently implemented by the first two authors and can be downloaded and installed directly from the Comprehensive R Archive Network (CRAN) at cran.r-project.org. The same method has been included by SAS Institute, Inc. in PROC QUANTREG, which is available from Version 9.1 (or higher) of SAS/STAT.

2. REVIEW OF EXISTING METHODS

We can classify the methods of inference for quantile regression into three categories: direct estimation of V_τ, rank-score method, and resampling.

2.1 DIRECT ESTIMATION

For the iid error model, a direct estimate of $1/f(0)$, called the sparsity, can be obtained by kernel smoothing or by using a simple difference quotient of the empirical quantile function $[\hat{F}^{-1}(\tau + h_n) - \hat{F}^{-1}(\tau - h_n)]/2h_n$, where $\hat{F}^{-1}(a)$ is the ath quantile of the residuals, and h_n is a bandwidth tending to 0 as $n \to \infty$. In the `quantreg` package, the latter is used in the se = "iid" option of the R function `summary.rq`. The default bandwidth is based on the result of Hall and Sheather (1988), and is optimal for the inference on univariate quantiles. We also refer to Koenker (1994) for more details.

For the non-iid error model, we note that the matrix V_τ in (1.4) can be estimated consistently if one replaces $f_i(0)$ in the matrix $X'FX$ by an asymptotically unbiased estimator. The `quantreg` package includes two options in the `summary.rq` function, se = "nid" and se = "ker". The nid option replaces $f_i(0)$ by $2h_n/(x_i'\hat{\beta}_{\tau+h_n} - x_i'\hat{\beta}_{\tau-h_n})$; see Koenker and Machado (1999) for more detailed discussions on how to handle the cases where $x_i'\hat{\beta}_{\tau+h_n} < x_i'\hat{\beta}_{\tau-h_n}$. The ker option replaces $f_i(0)$ by kernel smoothing. He, Zhu, and Fung (2002) gave another alternative by replacing $f_i(0)$ by $(\psi_\tau(r_i + d_n) - \psi_\tau(r_i - d_n))/2d_n$ for some sequence d_n tending to 0 as $n \to \infty$, where $\psi_\tau(r) = \tau I(r > 0) + (\tau - 1)I(r < 0)$ is the derivative of ρ_τ, but this option is not yet available in any package.

As the sample size increases to infinity and h_n goes to 0 under some mild restrictions, the estimate of V_τ is consistent under the assumption (1.1). The asymptotic validity of direct estimation holds generally, but the finite sample performance of the method varies quite a lot with the choice of h_n.

2.2 RANK-SCORE METHOD

To avoid the need to select a smoothing parameter, Koenker (1994) proposed an attractive method to obtain confidence intervals by inverting a rank score test. Regression rank scores were proposed by Gutenbrunner and Jurečková (1992). They were used by Gutenbrunner, Jurečková, Koenker, and Portnoy (1993) to construct a rank test for the null hypothesis that $A\beta_\tau = 0$ for some m by p matrix A. As a special case, the rank test can be used to test whether the jth component $\beta_{\tau,j}$ of β_τ is equal to b_0 or not. Because the test statistic is monotone in b_0, the set of b_0 values that will not be rejected by the test at level α will be an interval. The rank-score method will use this interval as the level $1 - \alpha$ confidence interval for $\beta_{\tau,j}$. Due to the properties of linear programming involved in calculating regression quantiles and ranks, the end points of the confidence intervals can be obtained through parametric programming, that is, one does not have to perform the test for all values of b_0 in order to construct the confidence intervals. The computational details were given by Koenker (1994), and extensions to the location-scale regression model of the form

$$y_i = x_i'\beta_\tau + (x_i'\gamma)e_i \tag{2.1}$$

with iid errors e_i were given by Koenker and Machado (1999).

Four points are worth making regarding the rank-score method. First, the rank-score

method implemented in `quantreg` has the iid and nid options, corresponding to the assumption of iid error model (1.3) and the location-scale model (2.1), respectively. Although neither option is asymptotically valid under the general assumption of (1.1), the method is quite robust against deviations from the model assumptions. Second, parametric programming involving a total of $O(np \log n)$ simplex pivots is used to find the end points of p confidence intervals, which makes the rank-score method very time consuming for large datasets. Third, the computational complexity of the rank-score method grows exponentially with the dimension if one wishes to obtain confidence sets on a (sub-)vector of the parameter β_τ. Finally, it is not clear how to provide an estimate for the variance-covariance matrix of $\hat{\beta}_\tau$.

2.3 RESAMPLING METHODS

Resampling is another way to avoid direct estimation of the error densities. Bootstrapping pairs and bootstrapping residuals are two common methods, but the latter pertains only to iid error models; see Efron and Tibshirani (1998). Pairwise bootstrap is a rather effective way for estimating the variance-covariance matrix and for constructing confidence intervals. For quantile regression, bootstrapping estimating equations have been found to be equally effective; see Parzen, Wei, and Ying (1993). These bootstrap methods, however, require repeated calculations of the regression quantile estimates. When n and p are large, using 50 bootstrap replications could be very time consuming. The rule of thumb given by Efron and Tibshirani (1998, p. 52) is that 50 bootstrap replications are needed to obtain a decent estimate of the variance-covariance matrix, which agrees with our own experience. More replicates are needed to construct confidence intervals based on the percentiles of the bootstrap estimates; see Andrews and Buchinsky (2002). On the other hand, for large datasets a confidence interval in the form of $\hat{\beta}_\tau \pm z_{\alpha/2}\mathrm{SD}(\hat{\beta}_\tau)$ is generally adequate. If one cannot afford a much larger number of bootstrap replications, this SD-based confidence interval is preferred to other percentile-based methods, even though the latter methods are known to be better with a larger number of bootstrap replications. In this article, the performance of the bootstrap methods is assessed for the SD-based confidence intervals.

We should also recognize that bootstrapping quantile regression is not a panacea. As pointed out by De Angelis, Hall, and Young (1993) and Knight (2003), the covariance matrix of the quantile estimates is not easy to estimate with a high order of accuracy. This issue is not fully considered in the present article.

3. THE MCMB METHOD

Recognizing that computational time is a major hurdle for the use of bootstrap methods, He and Hu (2002) proposed a new resampling method. Instead of solving a p-dimensional system (or its equivalent) for each replication as required by the usual bootstrap method, the Markov chain marginal bootstrap completes each replication by solving p one-dimensional equations. For moderate to large-dimensional problems, the MCMB can be performed with

a fraction of the time needed for a usual bootstrap method. First, we adapt the MCMB method of He and Hu (2002) to the regression quantile setting.

3.1 BASIC ALGORITHM

The MCMB approach requires only the linearity of the τth quantile for one given level of τ, not for other percentiles.

For convenience, we write $x_{i,j}$ as the jth component of x_i, $x_{i,(j-)}$ and $x_{i,(j+)}$ as the first $j-1$ and the last $p-j$ components of x_i, respectively. This convention in notation will be used for any p-dimensional vector. As a consequence, we can write $x_i'\beta = x_{i,j}\beta_j + x_{i,(j-)}'\beta_{(j-)} + x_{i,(j+)}'\beta_{(j+)}$ for any $1 \le j \le p$. Recall that ψ_τ is the derivative of ρ_τ.

Let $r_i = y_i - x_i'\hat{\beta}_\tau$ be the residuals, and $z_i = \psi_\tau(r_i)x_i - \bar{z}$ with $\bar{z} = n^{-1}\sum_{i=1}^n \psi_\tau(r_i)x_i$. The MCMB algorithm starts from the quantile estimate $\beta^{(0)} = \hat{\beta}_\tau$ with step $k = 0$ and iterates through the following steps.

1. $k \leftarrow k+1$.
2. For each integer $j \in [1,p]$ in the ascending order, draw with replacement from $\{z_1,\ldots,z_n\}$ to obtain $\{z_1^{k,j},\ldots,z_n^{k,j}\}$, and then solve $\beta_j^{(k)}$ as the root to

$$\sum_{i=1}^n \psi_\tau(y_i - x_{i,(j-)}'\beta_{(j-)}^{(k)} - x_{i,j}\beta_j^{(k)} - x_{i,(j+)}'\beta_{(j+)}^{(k-1)})x_{i,j} = \sum_{i=1}^n z_i^{k,j}. \quad (3.1)$$

3. Repeat Steps 1 and 2 until we reach a prespecified number of replications K.

Step 2 is the core of the MCMB algorithm. At the kth step, an independent sample $\{z_1^{k,j},\ldots,z_n^{k,j}\}$ needs to be drawn for each j. In Equation (3.1), we are solving for $\beta_j^{(k)}$ by using the most recent values of other parameters. We must note, however, the left side of (3.1) is a monotone step function, so the root to (3.1) is interpreted as the point of sign change. The resulting sequence $\beta^{(1)},\ldots,\beta^{(K)}$ is a Markov chain. Most importantly, it was shown by He and Hu (2002) that, under model (1.3), the sequence $\beta^{(k)}$ ($k = 1,\ldots,K$) is an asymptotically multivariate AR(1) process

$$n^{1/2}(\beta^{(k)} - \hat{\beta}_\tau) = A_n n^{1/2}(\beta^{(k-1)} - \hat{\beta}_\tau) + u_{n,k}, \quad (3.2)$$

and the sample variance of $\beta^{(k)}$ consistently approximates V_τ for large n and K.

The MCMB method shares two things in common with some of the better known MCMC algorithms (e.g., the Gibbs sampler), that is, they turn a high-dimensional problem into several one-dimensional problems, and they return a Markov chain. However, the basic idea and theory behind the MCMB method are rather distinct. The joint distributions of $\beta^{(k)}$ are not generated through the conditionals. Instead, linear approximations to estimating equations are used, and resampling is introduced to each component of the estimating equations, not the raw data point. Furthermore, the asymptotic validity of the MCMB method requires K to grow with n at a controlled rate; we refer to He and Hu (2002) for details.

The basic MCMB algorithm proposed by He and Hu (2002) has some obvious shortcomings. First, high autocorrelations of the MCMB sequence $\beta^{(k)}$ will result in a loss of

accuracy in estimating V_τ. That is to say, a correlated MCMB sequence of length K is not as good as an independent sequence of the same length from the pairwise bootstrap method. Second, if A_n of (3.2) is ill-conditioned, the correlation among the components of $\beta^{(k)}$ can be so high that the asymptotic stationarity of the sequence does not manifest itself for any chain of practical length. In this case, the sample variance will be a poor estimate of V_τ.

Fortunately, these problems can be eliminated by a very simple modification of the basic algorithm.

3.2 THE MCMB-A ALGORITHM

What we call the MCMB-A method is simply an affine transformation of the parameter space to alleviate the problems of autocorrelation in the MCMB sequence. The MCMB-A method for general parametric models was discussed by Kocherginsky (2003), the first author's Ph.D thesis. For the regression quantile model (1.3), the MCMB-A method reduces to a very simple transformation. It standardizes X by $\tilde{X} = (X'X)^{-1/2}X$ before applying the basic MCMB algorithm. The resulting MCMB sequence is denoted as $\tilde{\beta}^{(k)}$. The simple transformation at the end $\beta^{(k)} = (X'X)^{-1/2}\tilde{\beta}^{(k)}$ takes us back to the original parameter space.

It is evident that the above transformation removes the asymptotic correlation among the components of $\tilde{\beta}$. To see why it also removes autocorrelation, we note that the A_n matrix of (3.2) is actually determined by the $X'X$ matrix. More specifically, let $L(X'X)$ be the matrix obtained from $X'X$ by replacing all the upper off-diagonal elements by 0. Then, it follows from He and Hu (2002, p. 793) that $A_n = -(L(X'X))^{-1}(X'X - L(X'X))$. With X replaced by $\tilde{X}$, we have $\tilde{X}'\tilde{X} = I$ and therefore $A_n = 0$.

With the MCMB-A method, the resulting sequence $\tilde{\beta}^{(k)}$ is nearly independent due to the asymptotic normality of the chain. The transformation itself does not add any significant amount of computing time. It is only used at the start and the end of the program, but not in all the iterations. In addition, the same standardization of the x variables may be desirable in computing the estimator $\hat{\beta}_\tau$ for numerical stability, so there is really no additional cost in using it for the MCMB-A algorithm.

Like the rank-score method of Koenker (1994), the MCMB-A method is valid under the iid error model. Note that the A-transformation in the MCMB-A method is an affine transformation of the parameter space. As shown by Kocherginsky (2003), a less straightforward transformation (called MCMB-AB) can be used together with the MCMB algorithm to make the inference valid for non-iid error models. In addition to an A-transformation, the MCMB-AB method uses a simultaneous B-transformation on the space $\psi_\tau(\cdot)x$ of the estimating equations to ensure that the asymptotic variance-covariance of $\beta^{(k)}$ approximates that of $\hat{\beta}_\tau$ for non-iid error models.

Partly because of the robustness of the MCMB-A method against heteroscedasticity, and partly because of the time-savings that can be achieved without estimating the B-transformation, we choose not to use the MCMB-AB method in this article, but hope to include the MCMB-AB method in a future update to our R package `rqmcmb2`. For some

theoretical discussion of the desired robustness of the MCMB method, see He and Hu (2002, sec. 2.4), but also see Section 4.3 for a cautionary note.

3.3 TIME-SAVING STRATEGIES IN THE MCMB ALGORITHM

The basic operation in the MCMB algorithm is solving (3.1) in each iteration. For simplicity, we can write each equation in the following generic form

$$\sum_{i=1}^{n} \psi_\tau(y_i^* - x_{i,j}b)x_{i,j} = c^*, \tag{3.3}$$

where $y_i^* = y_i - x'_{i,(j-)}\beta^{(k)}_{(j-)} - x'_{i,(j+)}\beta^{(k-1)}_{(j+)}$ and c^* is a constant as the right side of (3.1). Any simple search algorithm may be used to solve (3.3), but we find it useful to recognize that the solution is actually a minimizer to

$$\sum_{i=1}^{n+1} \rho_\tau(y_i^* - x_{i,j}b), \tag{3.4}$$

where $x_{n+1,j} = -c^*/\tau$ and y^*_{n+1} is taken to be a very large positive number (e.g., half of the maximum allowed by the computer). Let $z_i = y_i^*/x_{i,j}$. Then (3.4) is equal to $\sum_{i=1}^{n+1}\{|z_i - b||x_{i,j}| - (2\tau - 1)(z_i - b)x_{i,j}\}$, which can be rewritten as

$$\sum_{i=1}^{n+1} |x_{i,j}|\{|z_i - b| - (2\tau^* - 1)(z_i - b) + c(x_{i,j}, z_i)\}$$
$$= \sum_{i=1}^{n+1} |x_{i,j}|\rho_{\tau^*}(z_i - b) + \sum_{i=1}^{n+1} |x_{i,j}|c(x_{i,j}, z_i),$$

where $c(x_{i,j}, z_i)$ does not depend on b, and $\tau^* = .5 + (\tau - .5)\sum_{i=1}^{n+1} x_{i,j}/\sum_{i=1}^{n+1} |x_{i,j}|$. Therefore, the solution to (3.4) is the weighted τ^*th quantile of z_i $(i = 1, \ldots, n + 1)$ with weights $w_i = |x_{i,j}|$. In the special case of $\tau = .5$, we have $\tau^* = .5$.

With this computational trick in mind, we can say that each iteration of the MCMB algorithm replaces a p-dimensional regression quantile problem with p one-dimensional quantile problems. The computational complexity of a p-dimensional quantile problem is in the order of $np^{5/2}$, but through MCMB it is reduced to the order of np. The savings in computer time increases rapidly with p.

There are several ways to calculate a weighted quantile. We use the intuitively appealing method of sorting. To simplify notation, suppose that we need the τ^*th quantile of $\{z_i\}$ with weights $\{w_i\}$, $(i = 1, \ldots, n)$, and the weights add up to 1. If we sort $\{z_i\}$ into an array of ascending order and write the resulting arrays as $\{z_{(i)}, w_{(i)}\}$, the smallest integer I such that $\sum_{i=1}^{I} w_{(i)} \geq \tau^*$ locates the quantile as $z_{(I)}$. The computational complexity of this algorithm can be as good as $O(n)$. For efficient sorting algorithms, we refer to Knuth (1998).

Because the MCMB algorithm needs to compute the weighted quantile Kp times, any savings in time to perform this calculation is helpful. We suggest that after we obtain an MCMB sequence of length 10, we use a timing saving strategy as follows.

Based on the first 10 vectors $\{\beta^{(k)},\ k = 1, \ldots, 10\}$, we can obtain a rough estimate of the SE s_j of each $\hat{\beta}_j$ $(j = 1, \ldots, p)$. In all future iterations, we will target our search of the weighted quantile in a small interval $(\hat{\beta}_j - 5s_j, \hat{\beta}_j + 5s_j)$. We only need to decide if a point z_i is inside, to the left, or to the right of this interval, and then sort the z_i's in this interval to obtain the desired quantile. In our experience, this strategy reduces the computational time quite substantially when n is large. To ensure that the weighted quantile is indeed inside the interval, we calculate $p_1 = \sum_{z_i < \hat{\beta}_j - 5s_j} w_i$ and $p_2 = \sum_{z_i > \hat{\beta}_j + 5s_j} w_i$. If $p_1 < \tau^* \leq 1 - p_2$, then this interval contains the weighted quantile. Otherwise, we enlarge the interval by another $5s_j$ on both sides of $\hat{\beta}_j$ until the interval is large enough to contain the desired quantile. The savings of time comes from the fact that we only need to sort $O(n^{1/2})$ points in the smaller interval.

4. MONTE CARLO COMPARISONS

As described in the preceding sections, several methods to construct the confidence intervals of quantile regression coefficients are available in the `quantreg` package. In this section, we turn to Monte Carlo simulations to assess the performance of the MCMB-A method and compare it with other methods described in Section 2. These methods, ranked from the least to the most expensive in terms of computing time for moderately large problems, are

iid:	based on a direct estimation of (1.5)
nid:	based on a direct estimation of (1.4)
ker:	based on a kernel-based estimate of (1.4)
mcmb:	based on the MCMB-A method
boot:	based on pairwise bootstrap
ciid:	based on the rank-score method under the iid error assumption
cnid:	based on the rank-score method under model (2.1)

The methods iid, nid, and ker involve choosing smoothing parameters. The default values in the `quantreg` package as of April 1, 2003, were used in the computations of this section. The mcmb and boot methods use $K = 200$ replications. In the first five cases, the SD-based intervals are used. A smaller number of $K = 50$ is also tested for mcmb; see the discussion near the end of this section.

To make an informative comparison, we consider models of different characteristics. The criteria we use are the average lengths (L) and coverage probabilities (C) of the confidence intervals for each coefficient. In order to avoid excessive displays of the simulation results, we will present the lengths and coverage probabilities for only a selected set of coefficients from each model. In designing the experiment, we consider the following factors: sample size n, percentile level τ, nominal levels (.9, .95, and .99), distributions of x, model

Table 1. Results for Models 1 and 2 with $n = 400$ and $\tau = .5$. The column L is average length and column C is coverage probability. The nominal coverage is .9.

	Model 1		*Model 2*		*Model 2b*	
Method	*L*	*C*	*L*	*C*	*L*	*C*
iid	.208	.904	.297	.867	.309	.880
nid	.208	.904	.294	.890	.304	.878
ker	.250	.950	.331	.935	.338	.922
ciid	.206	.893	.292	.870	.300	.885
cnid	.207	.900	.294	.870	.301	.888
mcmb	.212	.910	.304	.898	.311	.880
boot	.217	.925	.312	.913	.319	.888

structure and error distributions. Obviously, we can cover only a small number of levels per factor, but we choose them to be as informative as possible. In each case, the number of Monte Carlo samples is 400 for estimating the coverage probabilities with a standard error of less than 2% (when the coverage probabilities exceed .8). Because similar conclusions could be made at various nominal levels, we only report the results for the 90% confidence intervals.

4.1 SIMPLE MODELS

The first two models we consider are very simple:

Model 1. $y = 1 + \beta_1 x_1 + \beta_2 x_2 + e$,

where x_1, x_2, e are independently distributed as standard normal variables, and

Model 2. $y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (1 + x_3)e$,

where x_1 is standard normal, x_3 is uniform on (0,1), $x_2 = x_1 + x_3 + z$ with z being standard normal, and e is standard normal. The variables x_1, x_2, z and e are mutually independent. All the coefficients β_j take the value of 1.

Table 1 gives the performance of seven confidence intervals for the slope parameters when $n = 400$ and $\tau = .5$ for Model 1. Due to the symmetry in x_1 and x_2 in the model, we average the results for β_1 and β_2 in the table. In this simple case, all methods with the only exception of ker provide desirable coverage probabilities with similar lengths. The ker method tends to overestimate the SD in simple models. For Model 2, the results are also given in Table 1 with respect to β_2. The comparisons we obtained for Model 1 hold here. The slight variation in error variance does not seem to hurt the methods under the iid working assumption. To see the robustness of confidence intervals against outliers, we also report the results in Table 1 under Model 2b where 8 positive values of y in each sample are moved to an outlying value of 50. With this data contamination, the length and coverage are essentially unchanged.

4.2 VARIETY IN *x*

The next two models include independent variables with different types of distributions

Table 2. Results for Model 3 with $\tau = .25$. The nominal coverage is .9.

	β_3, $n = 200$		β_5, $n = 200$		β_3, $n = 500$		β_5, $n = 500$	
Method	*L*	*C*	*L*	*C*	*L*	*C*	*L*	*C*
iid	.218	.860	.722	.857	.134	.873	.467	.897
nid	.238	.837	.720	.850	.140	.805	.452	.890
ker	.244	.890	.805	.907	.136	.837	.473	.915
ciid	.287	.885	.765	.880	.156	.903	.469	.907
cnid	.313	.903	.793	.887	.163	.910	.476	.915
mcmb	.305	.945	.784	.903	.159	.930	.472	.913
boot	.291	.962	.857	.943	.156	.935	.504	.932

and take the form

Model 3. $y = 1 + \sum_{j=1}^{7} \beta_j x_j + e$,

Model 4. $y = 1 + \sum_{j=1}^{7} \beta_j x_j + (1 + x_3 + x_5 + x_7)e$,

where x_1 and x_2 are Bernoulli(0.4), x_3 and x_4 are lognormal, (x_5, x_6) is bivariate normal with mean (2,2), variance (1,1), and correlation .8, x_7 is chi-square distributed with one degree of freedom, and e is distributed as t_2. Except for the correlation between x_5 and x_6, all other variables are independently generated. Again, we take all β_j to be 1.

The confidence intervals for β_3 and β_5 will be included in Tables 2 and 3 with $\tau = .25$. Two sample sizes, $n = 200$ and $n = 500$, are considered here. It becomes clear that the iid and ker methods are not reliable, with the coverage probabilities going as low as 51% and 65% for these two methods. The nid method is a little better, but it still under-covers β_3 by a wide margin for Model 3. In fact, its performance does not get better with even larger n in our experiment with Model 3, which suggests that the procedure has room for improvement. The rank-score method aiming for heteroscedastic errors of Model 4 did not do any better than ciid. The pairwise bootstrap method is not clearly better than mcmb. The performance of ciid and mcmb continue to look good even at $\tau = .1$ with $n = 500$; these results are not tabulated here.

Table 3. Results for Model 4 With $\tau = .25$. The nominal coverage is .9.

	β_3, $n = 200$		β_5, $n = 200$		β_3, $n = 500$		β_5, $n = 500$	
Method	*L*	*C*	*L*	*C*	*L*	*C*	*L*	*C*
iid	1.044	.550	3.430	.892	.632	.512	2.198	.890
nid	2.277	.818	3.419	.880	1.442	.852	2.119	.888
ker	1.967	.652	3.632	.870	1.414	.725	2.256	.892
ciid	2.467	.892	3.606	.902	1.468	.895	2.175	.885
cnid	2.612	.900	3.879	.922	1.506	.902	2.252	.898
mcmb	2.529	.890	3.997	.928	1.483	.882	2.317	.902
boot	2.598	.892	4.211	.942	1.555	.885	2.397	.918

Table 4. Results for Model 5 with $\tau = .25$. The nominal coverage is .9.

	β_5, $n = 200$		β_6, $n = 200$		β_5, $n = 500$		β_6, $n = 500$	
Method	*L*	*C*	*L*	*C*	*L*	*C*	*L*	*C*
iid	1.569	.713	1.575	.853	.732	.688	.732	.890
nid	3.006	.943	2.685	.955	1.528	.960	1.324	.980
ker	2.337	.895	1.929	.927	1.261	.927	.958	.978
ciid	2.042	.798	1.867	.910	.980	.772	.865	.893
cnid	2.433	.857	2.088	.925	1.185	.843	.914	.905
mcmb	2.670	.925	2.413	.945	1.224	.880	1.092	.958
boot	2.922	.945	2.491	.968	1.376	.930	1.094	.980

4.3 MORE DEMANDING MODELS

Models 1–4 have a global linear quantile structure. We now consider a model for which the quantile function is linear for only one value of τ.

Model 5. $y = 1 + \sum_{j=1}^{7} \beta_j x_j + x_5^2(e - F^{-1}(\tau))$,

where the x and e variables are the same as in Model 3, and $F^{-1}(\tau)$ is the τth quantile of e. If $\tau = .25$, the .25-quantile of Y is linear in x, but other quantile functions are quadratic in x_5. Table 4 gives the results for this model with $\tau = .25$ and $n = 200$ or 500. Only β_5 and β_6 are included in the table.

This is clearly a more demanding model. The iid method continues to look unfavorable, and ciid has a low (77%) coverage for β_5. The nid intervals are unusually wide. The cnid method holds up better at 84%, and the mcmb is even better. The mcmb intervals look quite conservative for β_6, but slightly less so than the pairwise bootstrap.

Because the MCMB-A method is not designed for severe heteroscedastic errors, we include the following

Model 6. $y = 1 + x + (1.1 + x)e$,

where x is uniformly distributed on $(-1, 1)$ in Model 6, and e is standard normal. The results are given in Table 5. Under this model, the variance of y is about 400 times larger at x near 1 than at x near -1. As a result, the mcmb confidence interval for the slope parameter has a low coverage probability of around 70%, which indicates that the MCMB-AB method mentioned in Section 3.2 is desirable for such problems. Somewhat interestingly, the ciid method, with coverage of around 80%, holds up better for this type of heteroscedasticity than the mcmb method.

4.4 A CASE OF ASYMPTOTIC NONNORMALITY

The asymptotic normality of $\hat{\beta}$ in quantile regression would fail to hold if one or more x variables are too heavy tailed. The following model is meant to test how confidence intervals would do with a heavier tailed independent variable.

Model 7. $y = 1 + x_1 + x_2 + x_3 + e$,

Table 5. Results for Model 6 with $n = 200$ and $\tau = .5$. The nominal coverage is .9.

	β_0		β_1	
Method	*L*	*C*	*L*	*C*
iid	.204	.610	.355	.732
nid	.374	.898	.491	.868
ker	.439	.935	.739	.985
ciid	.319	.822	.394	.792
cinid	.358	.880	.523	.895
mcmb	.249	.712	.342	.708
boot	.376	.870	.513	.890

where x_1, x_3 and e are standard normal, but x_2 is taken to be the absolute value of the t-variate with two degrees of freedom. These variables are taken to be independent of each other. Table 6 provides the results for the coefficient of x_2 with $n = 200$ and $\tau = .25$ or .5. For a small number of times, the rank-score method returned confidence intervals with practically infinite length, so the average lengths reported in this table were computed with those cases excluded. We added a star to the average length when this problem happened.

In this model, x_2 has infinite variance so the standard asymptotic normality theory fails for the quantile estimate. Table 6 suggests that neither the rank-score method nor the MCMB method would be disastrous, but caution is needed when one of the independent variables is heavy tailed or highly skewed. The validity of both methods in situations like this requires further study.

4.5 SMALLER *K* IN MCMB-A

The simulation study reported above used $K = 200$ replications for the mcmb method. To see if the method can still give decent results with a smaller K, we run the simulations for Model 4 with $K = 50$ and $n = 500$, and compare the results with $K = 200$ in Table 7. We note that the smaller K tends to produce slightly smaller SD estimates, but the performance remains acceptable, especially when n is quite large. When tested with an even smaller $K = 20$, we found less reliable results for some models. Therefore, we recommend using K between 50 and 200 in typical applications. For the consideration of time, one may use smaller K (within the recommended range) with larger n.

Table 6. Results for Model 7 with $n = 200$. The nominal coverage is .9. The values marked by * indicate that intervals of infinite length have been excluded in the averaging.

	$\tau = .5$		$\tau = .25$	
Method	*L*	*C*	*L*	*C*
iid	.150	.880	.153	.860
nid	.147	.702	.166	.718
ker	.186	.945	.176	.878
ciid	.184*	.858	.252	.868
cnid	.191*	.872	.285	.875
mcmb	.159	.852	.198	.888
boot	.180	.892	.189	.925

Table 7. MCMB Results for Model 4 with $\tau = .25$ and $n = 500$. The nominal coverage is .9.

Parameter	*L (K = 50)*	*C (K = 50)*	*L (K = 200)*	*C (K = 200)*
β_0	3.368	.872	3.460	.880
β_1	2.692	.900	2.723	.905
β_2	2.688	.865	2.725	.868
β_3	1.462	.868	1.483	.882
β_4	.775	.900	.786	.912
β_5	2.290	.908	2.317	.902
β_6	2.273	.900	2.307	.898
β_7	1.604	.882	1.643	.882

5. RECOMMENDATIONS AND CONCLUSIONS

Combining asymptotic theory, computational complexity (timing), the Monte Carlo comparison as demonstrated in the previous section, and some other empirical experience we have gained, we make the following recommendations.

R1. To have reliable inference, it is desirable to have $n \min\{\tau, 1-\tau\} > 5p$.

R2. For relatively small problems with $n \leq 1{,}000$ and $p \leq 10$, we suggest the ciid method for constructing confidence intervals, and the MCMB-A method (with K between 100 to 200) or the pairwise bootstrap for estimating the variance-covariance matrix. If the ciid method gives a confidence interval of infinite length, a bootstrap method can be used instead.

R3. For moderately large problems with np between 10,000 and 2,000,000, we suggest the MCMB-A method (with K between 50 to 200, depending on your tolerance on speed) for estimating the variance-covariance matrix and constructing the SD-based confidence intervals. (The upper limit on np here should move up as faster computers are available.)

R4. For very large problems where the MCMB-A method is still too time consuming, we suggest using the nid method.

We have seen through the simulation experiment that the cnid method is, in most cases, not much better than the ciid method, but it has a higher computational cost. This is the reason for our recommended use of ciid in **R2**, but we are not against replacing ciid with cnid there.

The iid method can fail quite badly for non-iid error models, so it is not recommended unless one is confident about the iid error assumption. The ker method had the poorest performance in our simulation, but it might be improved on if we construct a better smoothing parameter in the future. The nid method is better than the other direct methods, but its current implementation needs improvement as suggested by its lower coverage at larger n under Model 3 and its wide intervals under Model 5.

We hasten to add that the recommended rank-score method and the MCMB-A method

are not asymptotically valid for general non-iid models. Nevertheless, they are highly robust against many forms of moderate heteroscedasticity. It is still desirable to develop methods that are asymptotically valid for general non-iid models and that are substantially faster than pairwise bootstrap in large problems. Possibilities such as the MCMB-AB method and the m-out-of-n bootstrap will be considered in future updates to the software we are working on.

ACKNOWLEDGMENTS

The research is partially supported by the NSF Grant DMS-0102411. The authors thank Roger Koenker for helpful discussions, Colin Chen and Ying Wei for testing our proposed method and algorithm on SAS, and two anonymous referees for their suggestions to improve presentation.

[Received July 2003. Revised December 2003.]

REFERENCES

Andrews, D. W. K., and Buchinsky, M. (2002), "On the Number of Bootstrap Repetitions for BCa Confidence Intervals," *Econometric Theory*, 18, 962–984.

De Angelis, D., Hall, P., and Young, G. A. (1993), "Analytical and Bootstrap Approximations to Estimator Distributions in L_1 Regression," *Journal of the American Statistical Association*, 88 , 1310–1316.

Efron, B., and Tibshirani, R. J. (1998), *An Introduction to the Bootstrap*, New York: CRC Press.

Gutenbrunner, C., and Jurečková, J. (1992), "Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics," *The Annals of Statistics*, 20, 305–330.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993), "Test of Linear Hypotheses Based on Regression Rank Scores," *Journal of Nonparametric Statistics*, 2, 307–331.

Hall, P., and Sheather, S. (1988), "On the Distribution of a Studentized Quantile," *Journal of the Royal Statistical Society*, Ser. B, 50, 381–391.

He, X., and Hu, F. (2002), "Markov Chain Marginal Bootstrap," *Journal of the American Statistical Association*, 97, 783–795.

He, X., and Shao, Q. M. (1996), "A General Bahadur Representation of M-estimators and its Application to Linear Regression with Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630.

He, X., Zhu, Z. Y., and Fung, W. K. (2002), "Estimation in a Semiparametric Model for Longitudinal Data with Unspecified Dependence Structure," *Biometrika*, 89, 579–590.

Knight, K. (2003), "On the Second Order Behavior of the Bootstrap of L_1 Regression Estimators," preprint.

Kocherginsky, M. (2003), *Extensions of the Markov Chain Marginal Bootstrap*, unpublished Ph.D. thesis, Department of Statistics, University of Illinois at Urbana-Champaign.

Koenker, R. (1994), "Confidence Intervals for Regression Quantiles," in *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, New York: Springer-Verlag, pp. 349–359.

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

Koenker, R., and Machado, J. A. (1999), "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Knuth, D. E. (1998), *Art of Computer Programming, Volume 3: Sorting and Searching* (2nd ed.), New York: Addison-Wesley.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994), "A Resampling Method Based on Pivotal Estimating Functions," *Biometrika*, 81, 341–350.

Portnoy, S., and Koenker, R. (1997), "The Gaussian Hare and the Laplacean Tortoise: Computability of Squared-error vs Absolute Error Estimators" (with discussion), *Statistical Science*, 12, 279–300.