Wilke, Ralf A.; Fitzenberger, Bernd; Zhang, Xuan

Working Paper

A Note on Implementing Box-Cox Quantile Regression

ZEW Discussion Papers, No. 04-61 [rev.]

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Wilke, Ralf A.; Fitzenberger, Bernd; Zhang, Xuan (2005) : A Note on Implementing Box-Cox Quantile Regression, ZEW Discussion Papers, No. 04-61 [rev.], Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

This Version is available at:
https://hdl.handle.net/10419/24695

Discussion Paper No. 04-61

A Note on Implementing Box-Cox Quantile Regression

Bernd Fitzenberger, Ralf A. Wilke and Xuan Zhang

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 04-61

A Note on Implementing Box-Cox Quantile Regression

Bernd Fitzenberger, Ralf A. Wilke and Xuan Zhang

Non–technical Summary

Quantile regression is gradually evolving into a comprehensive approach to the statistical analysis of linear and nonlinear response models for conditional quantile functions. Just as classical linear regression methods based on minimizing sums of squared residuals allow one to estimate a general class of models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the conditional median function and the full range of other conditional quantile functions.

The Box-Cox function is a nonlinear monotonic transformation including the log-linear and the linear function as special cases. The Box-Cox quantile regression model therefore provides an attractive extension of linear quantile regression techniques. Chamberlain (1994) and Buchinsky (1995) introduce a computationally convenient two stage method. However, a major numerical problem exists when implementing this method which has not been addressed so far in the literature. We suggest a simple solution modifying the estimator slightly. This modification is easy to implement. We derive the asymptotic distribution of the modified estimator and show that it has still standard statistical properties. Simulation studies confirm that the modified estimator works well in finite samples.

A Note on Implementing Box-Cox Quantile Regression*

Bernd Fitzenberger†

Ralf A. Wilke‡

Xuan Zhang§

December 2005

Abstract

The Box-Cox quantile regression model using the two stage method suggested by Chamberlain (1994) and Buchinsky (1995) provides a flexible and numerically attractive extension of linear quantile regression techniques. However, the objective function in stage two of the method may not exists. We suggest a simple modification of the estimator which is easy to implement. The modified estimator is still $\sqrt{n}$–consistent and we derive its asymptotic distribution. A simulation study confirms that the modified estimator works well in situations, where the original estimator is not well defined.

Keywords: Box-Cox quantile regression, iterative estimator

JEL: C13, C14

*We thank Blaise Melly for suggestions and comments. Financial support of the German Research Foundation (DFG) through the project "Microeconometric modelling of unemployment durations under consideration of the macroeconomic situation" is gratefully acknowledged.

†Corresponding author: Goethe-University Frankfurt, ZEW Mannheim, IZA Bonn and IFS London. E–mail: fitzenberger@wiwi.uni-frankfurt.de

‡ZEW Mannheim, P.O. Box 10 34 43, 68034 Mannheim, Germany, E–mail: wilke@zew.de

§ZEW Mannheim, Mannheim University, E–mail: x.zhang@gmx.de

1 Introduction

This note considers a numerical difficulty with the two step estimation approach for Box-Cox quantile regressions as suggested by Chamberlain (1994) and Buchinsky (1995).[1] In the second step, the objective function may not be defined and this problem arises in typical data situations. We suggest a simple modification of the objective function in order to ensure that it is well defined. The approach is motivated by a theoretical result, which we prove for the bivariate regression case. Simulations show that the modification works well in finite samples both in bivariate and multiple regression settings. We show that the standard asymptotic properties of the original estimator carry over after the modification and we derive the limit distribution of the modified estimator.

2 Model

Let us denote $\text{Quant}_\theta(y|x)$ as the θ's conditional quantile of y given x and g is a strictly monotonically increasing transformation function. We consider

$$\text{Quant}_\theta(y|x) = g(x'\beta_\theta), \tag{1}$$

where $y > 0$, the observable regressors $x \in \mathbb{R}^K$, the unknown parameters $\beta_\theta \in \mathcal{B} \subset \mathbb{R}^K$, and the quantile $\theta \in (0,1)$. We restrict our analysis to the transformation of the dependent variable introduced by Box and Cox (1964) :

$$y_\lambda = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ log(y) & \text{if } \lambda = 0, \end{cases}$$

as the inverse mapping to $g(.)$ where $\lambda \in \mathcal{R}$ where we assume $\mathcal{R} = [\underline{\lambda}, \overline{\lambda}]$ to be a finite closed interval. This transformation is quite attractive since it preserves the ordering of the observations because of the invariance of quantiles with respect to the monotonically increasing transformation g, i.e. $\text{Quant}_\theta(g(y)|x) = g(\text{Quant}_\theta(y|x))$. Thus, we obtain a linear model for

$$\text{Quant}_\theta(y_\lambda|x) = x'\beta_\theta$$

and equation (1) becomes

$$\text{Quant}_\theta(y|x) = (\lambda x'\beta_\theta + 1)^{1/\lambda} \quad . \tag{2}$$

However, equation (2) is in general no longer a valid representation for a conditional quantile of a nonnegative random variable, if the term $\lambda x'\beta_\theta + 1$ is negative. For $\lambda = 0$, there is no problem to

[1] The Box–Cox quantile regression model was introduced by Powell (1991).

map an unrestricted linear predictor $x'\beta_\theta$ to nonnegative quantiles $Q(y|x)$. But, for $\lambda < 0$ or for $\lambda > 0$, there are implicit restrictions on the possible values that $x'\beta_\theta$ may take in order to keep $Q(y|x)$ positive, as required.

The possibility to estimate λ allows for flexibility in estimating the model in (1). Powell (1991), Chamberlain (1994), Buchinsky (1995), and Machado and Mata (2000) provide further details on the model.

3 Estimation Problem

A Box–Cox quantile regression amounts to minimize the following objective

$$\min_{\beta\in\mathcal{B},\lambda\in\mathcal{R}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho_\theta(y_i - (\lambda x_i'\beta + 1)^{1/\lambda}),$$

for observations $i = 1, \ldots, n$ where the check function is given by $\rho_\theta(t) = \theta|t|\mathbb{1}_{t\geq 0} + (1-\theta)|t|\mathbb{1}_{t<0}$ and $\mathbb{1}$ denotes the indicator function. Powell (1991) shows that this nonlinear estimator is consistent and asymptotically normal, see also Machado and Mata (2000) for a concise discussion of the asymptotic distribution. In principle, the estimator could be obtained directly using an algorithm for nonlinear quantile regressions, e.g. Koenker and Park (1996). However, this is likely to be computationally demanding and the same numerical problem as dicussed below arises along the optimization process.

Chamberlain (1994) and Buchinsky (1995) suggest the following numerically attractive simplification in form of a two step procedure which exploits the equivariance property of quantiles:

1. estimate $\beta_\theta(\lambda)$ conditional on λ by

$$\hat{\beta}_\theta(\lambda) \;=\; \operatorname{argmin}_{\beta\in\mathcal{B}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho_\theta(y_{\lambda i} - x_i'\beta) \tag{3}$$

2. estimate λ by solving

$$\min_{\lambda\in\mathcal{R}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho_\theta(y_i - (\lambda x_i'\hat{\beta}_\theta(\lambda) + 1)^{1/\lambda}). \tag{4}$$

Note that the objective in (3) cannot be used to estimate both β_θ and λ (this would result in the degenerate estimator $\hat{\beta}_\theta = 0$ and $\hat{\lambda} = -\infty$). Chamberlain (1994) sketches the large sample theory of the two step estimator. Buchinsky (1995) derives large sample properties of this estimator for discrete regressors when applying the minimum distance method.

When implementing the two step procedure, we encountered the following general numerical problem which is due to the implicit restrictions on the feasible values of $x'\beta_\theta$. For every λ, it is not guaranteed that for all observations $i = 1, ..., n$ the inverse Box-Cox transformation $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1$ is strictly positive. However, this is necessary to conduct the second step of the above procedure.[2] It is natural to omit the observations for which this condition is not satisfied. But this raises a number of problems. First, the set of observations omitted changes when going through an iterative procedure to find the optimal λ. Second, it is not a priori clear how such an omission of observations affects the asymptotic distribution of the resulting estimator. Third, should still the full set of observations be used in the first step? The purpose of this note is to suggest a structured way on how to implement the necessary omission of data points and to clarify the consequences of doing so.

4 Modified Estimation

Stage two can only be solved if

$$\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0 \tag{5}$$

for all $i = 1, \ldots, n$. This clearly depends on the first stage estimates and the specific value of λ. A violation of this condition may occur due to the finite sample bias of the estimates, by misspecification of the model, or equivalently, when the second step is evaluated during an iterative procedure to obtain the estimator.[3] Therefore, in finite samples the inequality (5) may not hold for all observations.

Our modification of the estimator consists of using only those observations in the second step for which the second stage of the estimation is always well defined for all $\lambda \in \mathcal{R}$. The first step is still implemented based on all observations which allows asymptotically for a more efficient estimator.

Define the set of admissible observations $\mathcal{N}_{\theta,n}$ as those $i = 1, ..., n$ for which $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in \mathcal{R}$. Note that $\mathcal{N}_{\theta,n}$ may change with the number of observations due to variation of $\hat{\beta}_\theta$ and due to additional observation. A method for finding $\mathcal{N}_{\theta n}$ in applications is suggested below.

[2]The issue also arises for any other available computation method in the literature when evaluating $(\lambda x_i'\hat{\beta}_\theta(\lambda) + 1)^{1/\lambda}$, i.e. the algorithm by Koenker and Park (1996) for nonlinear quantile regression or the minimum–distance approach of Buchinsky (1995), see equation (10), page 117 of the latter paper.

[3]For some λ during the iteration process, step 1 results in the generally misspecified linear quantile regression of y_λ on x_i (see appendix).

Instead of problem (4), we now solve in the second step

$$\min_{\lambda \in \mathcal{R}} \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{i \in \mathcal{N}_{\theta,n}} \cdot \rho_\theta(y_i - \tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)]), \tag{6}$$

where for any $c \in \mathbb{R}$

$$\tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)] = \begin{cases} c & \text{if } \lambda > 0 \text{ and if } x_i' \hat{\beta}_\theta(\lambda) \leq -1/\lambda \\ c & \text{if } \lambda < 0 \text{ and if } x_i' \hat{\beta}_\theta(\lambda) \geq -1/\lambda \\ (\lambda x_i' \hat{\beta}_\theta(\lambda) + 1)^{1/\lambda} & \text{otherwise.} \end{cases}$$

Note it does not matter what value of c is chosen because the indicator function in equation (6) is always zero in these cases. This notation is introduced in order to have an objective function with a well defined sum from 1 to n. It is shown in the appendix that the modified estimator is consistent and asymptotically normally distributed. The asymptotic variance matrix for $(\hat{\beta}_\theta', \hat{\lambda})$ just uses the observations in $\mathcal{N}_{\theta,n}$.

How to choose the set of admissible observations $\mathcal{N}_{\theta,n}$?

As a purely theoretical rule, one could simply choose $\mathcal{N}_{\theta,n}$ as the set of observations i for which $\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in \mathcal{R} = [\underline{\lambda}, \overline{\lambda}]$. However, this is not a rule which can be applied in actual estimation because one can not determine whether the condition holds for all $\lambda \in \mathcal{R}$. For this reason, a practical alternative is needed.

We suggest a simple heuristic rule for the choice of $\mathcal{N}_{\theta,n}$ during the iteration process in $\lambda \in \mathcal{R}$. We show that this rule is strictly valid in the bivariate regression case $K = 2$ involving an intercept. For the case $K > 2$, we argue why the rule generally works for practical purposes and we confirm this by extensive simulation evidence. In the case $K = 2$, it turns out that it is only necessary to check for the smallest and the largest values $\underline{\lambda}$ and $\overline{\lambda}$ in $\mathcal{R}$, respectively, whether $\tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)]$ is well defined.

> *(HR) Our heuristic selection rule defines $\mathcal{N}_{\theta,n}$ as the set of observations i for which the condition $\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 > 0$ holds for both $\lambda = \underline{\lambda}$ and $\lambda = \overline{\lambda}$.*

This rule is based on the following result (the proof can be found in the appendix).[4]

[4]Note that proposition 1 does not hold for censored Box-Cox quantile regressions because the result hinges critically on the interpolation of actual data points for linear quantile regressions. This is not necessarily the case for censored quantile regressions, see Fitzenberger (1997). Limited simulation evidence (simualtion results are available upon request) suggests that our selection rule works for censored Box-Cox quantile regressions only up to an upper and lower bound of λ. These bounds seem to depend on the simulation design. Further research is necessary on this issue.

Proposition 1: *For the bivariate regression model $K = 2$ (one regressor plus an intercept) assume that $F_{\epsilon_\theta}(u|x)$ is a continuous distribution function almost surely and that the design matrix has full rank 2. If, for some observation i, $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0$ for $\lambda \in \{\underline{\lambda}, \overline{\lambda}\}$, then $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ with probability one.*

Proposition 1 can be motivated as follows: If for some $\lambda > 0$ and some data point i the linear quantile regression in step 1 of the estimation procedure yields $x_i'\hat{\beta}_\theta(\lambda) = -1/\lambda$. Then, the fitted value is a weighted average of two interpolated observations with perfect fit, see Theorem 3.1 in Koenker and Bassett (1978). This is due to the linear quantile regression involving a linear program. Since the predicted values for the latter two interpolated observations lie strictly above $-1/\lambda$ the weight on the observation with the higher value of y must be negative. A reduction in λ reduces the distance between the fitted value and $-1/\lambda$ more strongly for the latter observation compared to the observation with positive weight. Therefore, the linear combination of the fitted values must increase.

Unfortunately, Proposition 1 does not hold for the case with $K \geq 3$. In the appendix, we provide a counter example. However, in our subsequent simulations, we found no case where applying the selection rule based on proposition 1 did not work perfectly during the search for estimating λ. In the following, we will argue why this is the case in typical estimation problems.

For the proof of Proposition 1, one has to consider critical observations with regressor values x_i^c resulting in fitted values $x_i^{c\prime}\hat{\beta}_\theta(\lambda)$ close to $-1/\lambda$ for some λ. The fitted values are weighted averages of the fitted values of the K interpolated observations (Theorem 3.1 in Koenker and Bassett, 1978). To investigate the change in the set of regressor values satisfying condition (5) in response to a change in λ, the following condition is critical (see proof of Proposition 1)

$$\frac{\partial \Delta}{\partial \lambda} = \sum_{h=1}^{K} g_h log(y_{(h)}) y_{(h)}^{\lambda} < 0 \tag{7}$$

for interpolated observations $h = 1, ..., K$ with $\Delta = \sum_{h=1}^{K} g_h y_{(h)}^{\lambda} = 0$ and $\sum_{h=1}^{K} g_h = 1$. The weights are given by the regressor vector for the critical observation being a linear combination of interpolated design points, $x_i^c = \sum_{h=1}^{K} g_h x_{(h)}$ (see appendix for details). If condition (7) is satisfied for $K > 2$, then the result in Proposition 1 applies (the proof in the appendix is formulated for the case with general K and condition (7) is only needed in step 5 of the proof).

Note that condition (7) holds strictly if the minimum of the dependent variable for all observations with negative weights is not smaller than the maximum of the dependent variable for all observations with positive weights, i.e. $min\{y_{(h)}, g_h < 0\} \geq max\{y_{(h)}, g_h > 0\}$. This is a useful benchmark, since $-1/\lambda$, which is the fitted value at the critical data points, is strictly below $y_{(h),\lambda}$

Table 1: Finite sample evidence from 1.000 Monte Carlo experiments ($\theta = 0.5$). Means with standard deviations in parentheses.

	Homoskedastic				Heteroskedastic			
	$n = 100$		$n = 1.000$		$n = 100$		$n = 1.000$	
% of i not in $\mathcal{N}_{0.5,n}$	17.7%	(0.02)	18.3%	(0.01)	17.6%	(0.02)	18.2%	(0.01)
$\hat{\beta}_1$	10.067	(1.21)	9.990	(0.35)	10.0197	(1.03)	10.011	(0.27)
$\hat{\beta}_2$	1.010	(0.16)	0.999	(0.05)	1.003	(0.13)	1.001	(0.03)
$\hat{\beta}_3$	2.016	(0.36)	2.001	(0.10)	2.002	(0.26)	2.000	(0.07)
$\hat{\lambda}$	0.999	(0.07)	0.999	(0.02)	0.998	(0.06)	1.000	(0.02)

for all h. For this reason, some of the weights have to be negative because, at the critical point, the regression predicts a smaller value than at all the interpolating point. Typically the weights are positive for the interpolating points, which are closer to the critical point in the covariates space, and the closer interpolating points are typically associated with smaller predicted values, thus being closer to the predicted value at the critical point. Therefore, it is typically the case that g_h is positive, if $y_{(h)}$ is small, and g_h is negative, if $y_{(h)}$ is large. This generally holds in practical data designs implying condition (7).[5] The extensive simulation results in the next section are consistent with our reasoning here.

In case our rule (HR) is violated, i.e. we find for some observation $i \in \mathcal{N}_{\theta,n}$ and some $\lambda \neq \underline{\lambda}, \overline{\lambda}$ that $\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 < 0$, we suggest as a practical modification of (HR) to set

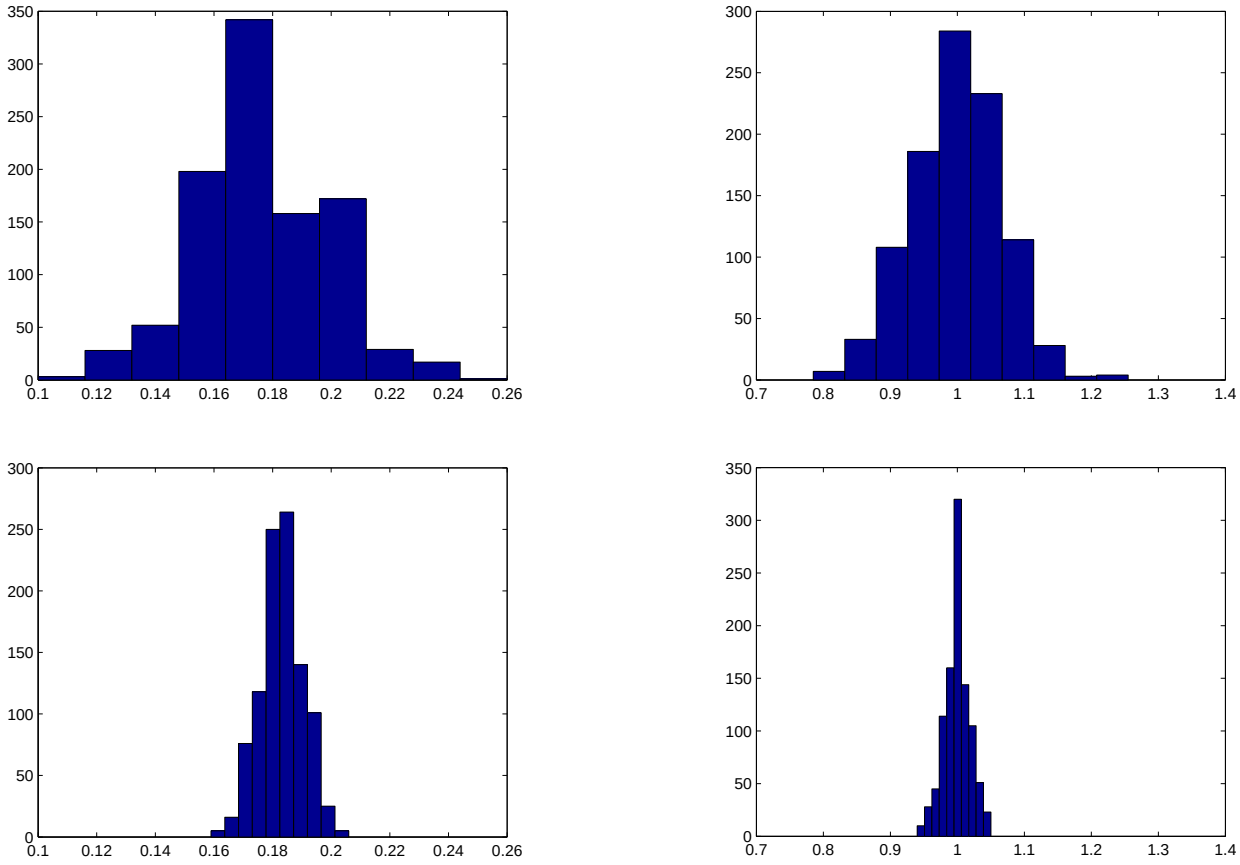
$$\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 = \epsilon \tag{8}$$

for some small $\epsilon > 0$ in order to make the objective function well defined.[6] Based on our simulation results, a violation of (HR) is likely to be a very rare event. The impact of this additional modification is likely to be negligible.

[5]This typical setup does not hold in our counter example in the appendix since none of the interpolating data points is close to the critical point in the covariates space (all interpolating points lie in different quadrants). In this situation, the observation with the largest value of the dependent variable also has the largest positive weight resulting in a strong "leverage effect" on the critical data point.

[6]This modification is based on a suggestion by Blaise Melly. Note that the additional modification (8) for admissible observations differs from from the modification in (6) involving setting an arbitrary c for the non–admissible observations which are irrelevant for the optimization.

Figure 1: Distribution of shares of inadmissible observations not in $\mathcal{N}_{0.5,n}$ (left panel) and distribution of $\hat{\lambda}$ (right panel) for 100 (top panel) and 1.000 observations (bottom panel), homoskedastic design



5 Simulations

This section assesses the finite sample performance of the modified estimator (6) through Monte Carlo studies. We use the following model:

$$y_\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma(x'\beta)\epsilon,$$

where x_1 is uniformly distributed between -10 and 10, $x_2 \in \{0, 1\}$ with $Prob(x_2 = 0) = Prob(x_2 = 1) = 0.5$ and $\beta = (10, 1, 2)'$. The error term ϵ follows a truncated normal distribution with bounds $[-1, 1]$[7] and it is independent of x. For the homoskedastic design, the scale function $\sigma(x'\beta)$ is

[7]Note that $y_\lambda > -\lambda^{-1}$ if $\lambda > 0$ and $y_\lambda < -\lambda^{-1}$ if $\lambda < 0$ are required for the inverse of the Box-Cox transformation to be well defined for the true λ. Thus, we use a truncated error term distribution. For further details see Poirier (1978).

set to 1, and for the heteroskedastic design the scale function is set to $exp(x'\beta/10)/4$. Note that both for the homoskedastic and the heteroskedastic design the residuals have very similar sample variances. The "true" value of λ is set to 1. We base our modified estimator on the admissible interval $\mathcal{R} = [-0.5, 2.5]$ for λ. We draw 1.000 independent random samples from this model. Estimates for β are obtained using the algorithm implemented in TSP 4.5. We apply a grid search in λ on the interval $[-0.5, 2.5]$ with step size 0.005 because the objective function may be locally non-convex.[8] Table 1 presents the results for four experiments based on 1.000 replications with sample sizes $n = 100$ and $n = 1.000$.[9]

Table 1 indicates that the proposed modified estimator performs well at both sample sizes in the homoskedastic design, and moderately well in the heteroskedastic design. The results show that the numerical problem addressed in this note may be in an application by no means negligible. On average, between 16 and 17 percent of all observations are affected for this simple data generating process. The results also show that our modification of the estimator works well in practice. The averages of the estimates are close to the true parameter values and the estimator appears to be unbiased even in small samples.

Figures 1 and 2 depict the empirical distributions of the share of observations not falling in $\mathcal{N}_{0.5,n}$ and of the estimates of λ. It turns out that in some samples more than 20 percent of the observations are affected by the numerical problem addressed here when the sample size is 100. As to be expected, the share of critical observations is much more concentrated around 17 percent when the sample size is 1.000. The distribution of $\hat{\lambda}$ is nicely concentrated around the true parameter $\lambda = 1$ and as to be expected the variance decreases with the sample size.

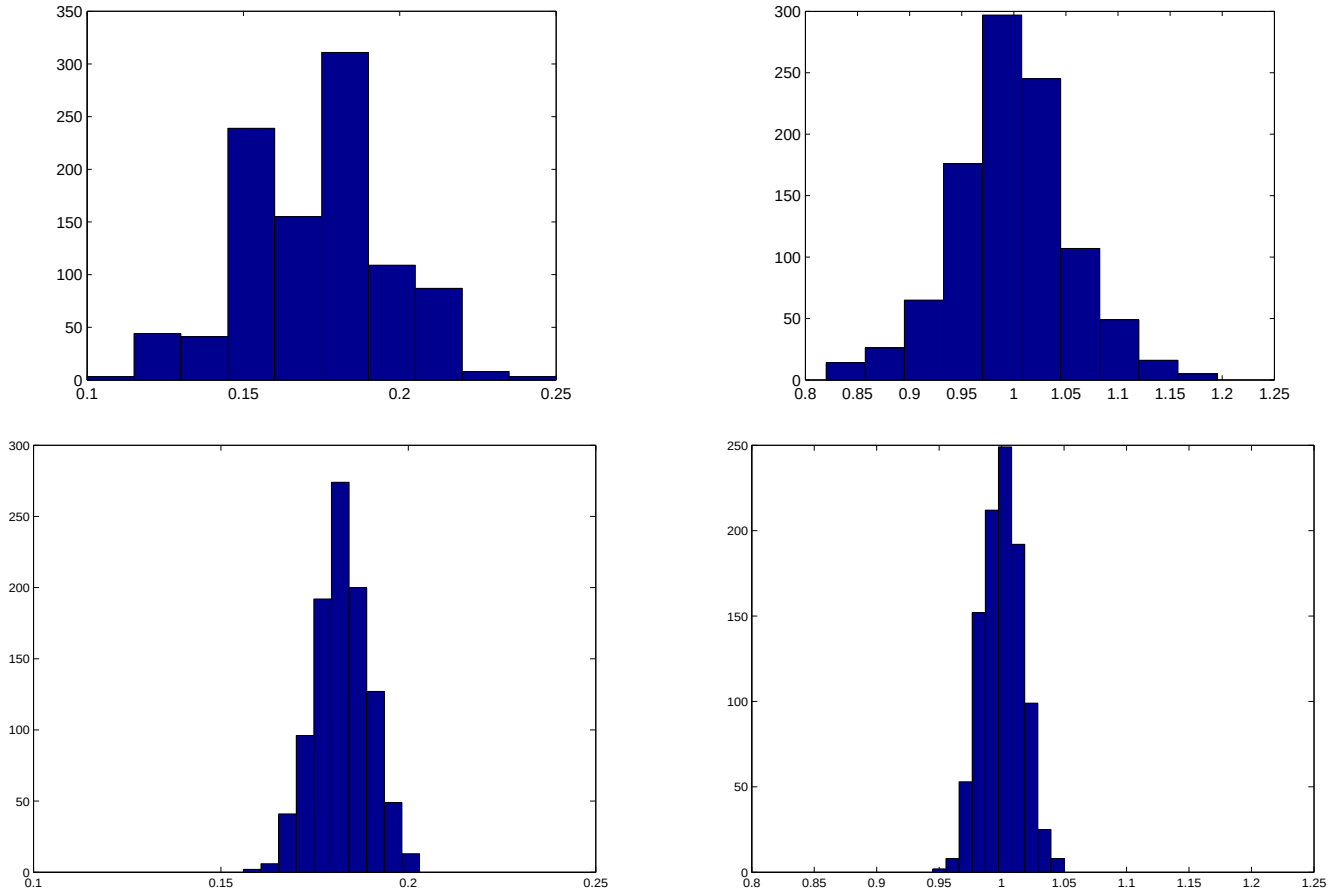
During our simulation study (using both the grid search and the numerical optimization method), we do not observe any violation of our heuristic rule (HR), although violations can in principle exist for our data generating process.[10] Therefore, in our simulation study, we did not have to apply the additional modification suggested at the end of section 4 in any case. If a violation occurs in applying our modified estimation approach, we recommend to apply the additional modification.

[8] We also replicate the simulation study by using the Koenker and Park (1996) algorithm for MATLAB provided by Hunter (2002) which results in a local optimum. The second stage is solved by using the *fminsearch* function of MATLAB which uses the Nelder-Mead simplex method for non-differentiable objective functions. We use a randomly chosen initial start point. The computation time is much faster than for the grid search and the results only marginally change. These results are available upon request.

[9] We also considered simulation designs with more than three regressors and different marginal distributions of the covariates. In all cases we did not observe any violation of our heuristic rule.

[10] We are grateful to Blaise Melly for pointing this out.

Figure 2: Distribution of shares of inadmissible observations not in $\mathcal{N}_{0.5,n}$ (left panel) and distribution of $\hat{\lambda}$ (right panel) for 100 (top panel) and 1.000 observations (bottom panel), heteroskedastic design



Appendix

Proof of Proposition 1: Without loss of generality, assume that $\overline{\lambda} > 0$. In the following, we will show that $\overline{\lambda} x_i' \hat{\beta}_\theta(\overline{\lambda}) + 1 > 0$ implies $\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in (0, \overline{\lambda}]$.

Therefore, assume $\lambda > 0$ in the following. The proof proceeds in a number of steps.

1. The condition $\lambda x_i' \hat{\beta}_\theta(\lambda) + 1 > 0$ is equivalent to $x_i' \hat{\beta}_\theta(\lambda) > -\frac{1}{\lambda}$ and our result is implied by $\frac{\partial x_i' \hat{\beta}_\theta(\lambda)}{\partial \lambda} < \frac{1}{\lambda^2}$ for $x_i' \hat{\beta}_\theta(\lambda)$ being close to $-\frac{1}{\lambda}$, which is to be shown.

2. We omit for this step the index i. Note that

$$f(y, \lambda) \equiv \frac{\partial y_\lambda}{\partial \lambda} = \frac{1}{\lambda^2} + \frac{y^\lambda (\lambda ln(y) - 1)}{\lambda^2}$$

and

$$f(y,\lambda)\left(\begin{array}{c}>\\=\end{array}\right)0 \text{ for } y\left(\begin{array}{c}\neq\\=\end{array}\right)1 \text{ and } f(y,\lambda)\left(\begin{array}{c}<\\=\\>\end{array}\right)\frac{1}{\lambda^2} \text{ for } y\left(\begin{array}{c}<\\=\\>\end{array}\right)exp\left(\frac{1}{\lambda}\right).$$

Starting at some λ, for y being small, i.e. $y < exp(1/\lambda)$, reducing λ will result in an increase and for y being large, i.e. $y > exp(1/\lambda)$, in a decline of $y_\lambda + 1/\lambda$.

3. The interpolation property of linear quantile regression (Koenker and Bassett, 1978, Theorem 3.1) implies that $x'_{(h)}\hat{\beta}_\theta(\lambda) = y_{(h),\lambda}$[11] for $h = 1, ..., K$ individual observations with linearly independent $x_{(h)}$ and $i(h) \in \{1, ..., n\}$ representing individual distinct observations $(x_{(h)} = x_{i(h)}, y_{(h)} = y_{i(h)})$. This interpolation property is the consequence of the fact that estimating a linear quantile regression involves solving a standard linear program. A reduction in λ for $\lambda > 0$ results in a stronger decline of the interpolated $y_{(h),\lambda}$ the higher its value. In particular, for a small $y_{(h),\lambda}$ it follows that $y_{(h),\lambda} + 1/\lambda = x'_{(h)}\hat{\beta}_\theta(\lambda) + 1/\lambda$ increases. Note, that for an infinitesimally small reduction in λ, the set of interpolated data points $i(h), h = 1, ..., K$ does not change (only the interpolated values $y_{(h),\lambda}$ do change), see Koenker and D'Orey (1987, p. 385) for a similar argument.

4. Suppose for some $\lambda \leq \overline{\lambda}$ and some observation i with $x_i = \sum_{h=1}^{K} g_h x_{(h)}$ (the weights g_h are given by the fact that every x_i can be represented as a linear combination of K linearly independent vectors $x_{(h)}$) it is the case that $x'_i\hat{\beta}_\theta(\lambda) = -1/\lambda$. Due to the presence of an intercept, it is clear that $\sum_{h=1}^{K} g_h = 1$. By the interpolation property, it follows that $\sum_{h=1}^{K} g_h y_{(h),\lambda} = -1/\lambda$. The latter statement is equivalent to $\Delta \equiv \sum_{h=1}^{K} g_h y^\lambda_{(h)} = 0$, where the left–hand–side denotes the difference between the fitted value for observation i and the critical value $-1/\lambda$. We will show that $\partial\Delta/\partial\lambda < 0$.

5. Assume without loss of generality $y_1 \neq y_2$ (for the case $y_1 = y_2$ there are no critical data point with fitted values not lying strictly above $-1/\lambda$ thus requiring not further consideration). For the critical data point i in the previous step, it follows that $g_1 = y^\lambda_{(2)}/(y^\lambda_{(2)} - y^\lambda_{(1)})$ and $g_2 = 1 - g_1 = y^\lambda_{(1)}/(y^\lambda_{(1)} - y^\lambda_{(2)})$. Then, after some straightforward manipulations, we obtain

$$\frac{\partial\Delta}{\partial\lambda} = \sum_{h=1}^{2} g_h log(y_{(h)}) y^\lambda_{(h)} = \frac{y^\lambda_{(2)} y^\lambda_{(1)}[log(y_{(1)}) - log(y_{(2)})]}{\lambda(y^\lambda_{(2)} - y^\lambda_{(1)})} < 0\,.$$

The inequality holds because $[log(y_{(1)}) - log(y_{(2)})]$ and $[\lambda(y^\lambda_{(2)} - y^\lambda_{(1)})]$ have opposite signs.

[11] With $y_{(h),\lambda} = (y^\lambda_{(h)} - 1)/\lambda$ for $\lambda \neq 0$ and $y_{(h),\lambda} = log(y_{(h)})$ for $\lambda = 0$.

6. After more than an infinitesimal change of λ it may occur that the set of interpolating observations changes. For the specific λ, where this occurs, the linear quantile regression will interpolate another data point $l = 1, ..., n$ with $x_l'\hat{\beta}_\theta(\lambda) = y_{l,\lambda}$ in addition to $i(h), h = 1, ..., K$, again see Koenker and D'Orey (1987, p. 385) for a similar argument. If λ moves infinitesimally further, then the data point l will replace one of the interpolated $i(h)$ in the set of interpolated data points. For the new set of interpolated data points, the regressor vectors will again be linearly independent. Since the quantile regression interpolates all $y_{(h),\lambda}$ as well as $y_{l,\lambda}$ and all except one of the $i(h)$ data points remain interpolated when λ moves beyond the critical value, the same argument applies as in the previous step. Thus, also for such critical values of λ, where the set of interpolated data points changes, it is clear that both one directional derivatives $(\partial\Delta/\partial\lambda)_{d\lambda<0}$ and $(\partial\Delta/\partial\lambda)_{d\lambda>0}$ are non-positive for critical observations where the quantile regression interpolates $-1/\lambda$.

The proof proceeds in an analogous way for $\underline{\lambda} < 0$ showing that if $\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0$ holds for $\lambda = \underline{\lambda}$, then it holds for all $\lambda \in [\underline{\lambda}, 0)$.

□

Counter example for the result in Proposition 1 for $K = 3$

Consider the following data set with $n = 10$ observations and 2 regressors x_{1i} and x_{2i}:

i	$x_{i,1}$	$x_{i,2}$	y_i
1	-2	-2	0.3
2	1	3	0.2
3	1	3	0.2
4	1	3	0.2
5	2	-3	2.0
6	2	-3	2.0
7	2	-3	2.0
8	3	-1	1.9600354921
9	3	-1	1.9600354921
10	3	-1	1.9600354921

Note that three times three observations are the same respectively and that for $\lambda = 2$ the Box–Cox quantile regression at the median ($\theta = 0.5$) interpolates observations 2(=3,4), 5(=6,7), and

8(=9,10). Observation 1 is a critical observation for our purpose with $x_1'\hat{\beta}_\theta(\lambda) = -1/\lambda = -0.5$ for $\lambda = 2$. For $\lambda = 1.99$, the fitted value is $x_1'\hat{\beta}_\theta(\lambda) = -0.50310 < -0.50251 = -1/\lambda$ and for $\lambda = 2.01$, the fitted value is $x_1'\hat{\beta}_\theta(\lambda) = -0.49691 > -0.49751 = -1/\lambda$. For $\lambda = 2$, one obtains $(g_1, g_2, g_3) = (1.125, 2.75, -2.875)$ as weights for observation 1 with g_1, g_2, g_3 referring to observations 2, 5, and 8, respectively. Furthermore, $\partial\Delta/\partial\lambda = \sum_{h=1}^{K} g_h log(y_{(h)}) y_{(h)}^{\lambda} = 0.11932 > 0$ for $\lambda = 2$. The critical condition (7) is violated in this case, because of the large positive weight g_2 for the observation with the highest value of the dependent variable $y_5 = 2.0$ resulting in a strong "leverage effect" on the critical observation 1.

Asymptotic Properties of modified estimator

We establish the asymptotic properties of our modified estimator based on the following four steps, following the analysis of the asymptotic distribution of Box–Cox quantile regression in Chamberlain (1994, appendix A.2) and building on the analysis in Powell (1991). For a given quantile θ, λ_0 and $\beta_{0,\theta}$ are the true parameter values.

1. For a possibly misspecified linear quantile regression define the best linear quantile predictor[12] in the population (Angrist et al., 2004, section 2.1) under asymmetric loss by

$$\beta_\theta(\lambda) = \text{argmin}_\beta \, E\rho_\theta(y_\lambda - x'\beta) \quad .$$

   For a given λ and under standard regularity conditions, the linear quantile regression estimator $\hat{\beta}_\theta(\lambda)$ is $\sqrt{n}$–consistent and it converges to the coefficients of the best linear quantile predictor. Under standard regularity conditions as in Powell (1991) or Chamberlain (1994), in particular y is continuously distributed conditional on x guaranteeing differentiability of the population objective function, and analogous to the least squares case, it can be shown then that $\beta_\theta(\lambda)$ satisfies the following first order condition

$$\int_x \left\{ \int_y x(I(y_\lambda < x'\beta) - \theta) f(y|x) dy \right\} f(x) dx = Ex(I(y_\lambda < x'\beta) - \theta) = 0$$

   as a population moment condition, where $I(.)$ is the indicator function. It is clear that for the true λ_0, we obtain $\beta_\theta(\lambda_0) = \beta_{0,\theta}$. Even though, the linear quantile predictor as an approximation does not satisfy $Quant(y_\lambda|x) = x'\beta_\theta(\lambda)$ for general λ (Angrist et al., 2004) the population moment condition suffices for $\hat{\beta}_\theta(\lambda)$ to be a $\sqrt{n}$–consistent estimator of $\beta_\theta(\lambda)$, as suggested by Chamberlain (1994) and shown explicitly in Fitzenberger (1998).

[12]This definition is analogous to the linear projection for least squares, see Wooldridge (2002), chapters 2 and 3.

2. The dummy variable indicating the admissible observations for the modified estimator is given by

$$\mathbb{1}_{i\in\mathcal{N}_{\theta,n}} = I(\{\overline{\lambda}x_i'\hat{\beta}_\theta(\overline{\lambda}) + 1 > 0\} \text{ and } \{\underline{\lambda}x_i'\hat{\beta}_\theta(\underline{\lambda}) + 1 > 0\})$$

which is based on the estimated linear quantile predictors for both $\underline{\lambda}$ and $\overline{\lambda}$. For the population quantile predictors, define

$$I_i = I(\{\overline{\lambda}x_i'\beta_\theta(\overline{\lambda}) + 1 > 0\} \text{ and } \{\underline{\lambda}x_i'\beta_\theta(\underline{\lambda}) + 1 > 0\}) \ .$$

$\sqrt{n}$–consistency of $\hat{\beta}_\theta(\lambda)$ implies that $E(\mathbb{1}_{i\in\mathcal{N}_{\theta,n}} - I_i) = O_p(n^{-1/2})$ and $Var(\mathbb{1}_{i\in\mathcal{N}_{\theta,n}} - I_i) = O_p(n^{-1})$ for uniformly bounded moments (higher than second) of x_i.[13]

3. For the asymptotic analysis, we can replace $\mathbb{1}_{i\in\mathcal{N}_{\theta,n}}$ by I_i in the objective function for the second step of the modified estimator in equation (6) because the difference

$$\frac{1}{n}\sum_{i=1}^{n} I_i \cdot \rho_\theta(y_i - \tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)]) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{i\in\mathcal{N}_{\theta,n}} \cdot \rho_\theta(y_i - \tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)]). \tag{9}$$

uniformly converges to zero in probability. Note that $\mathbb{1}_{i\in\mathcal{N}_{\theta,n}}$ and I_i do not depend upon λ (and therefore $\hat{\beta}_\theta(\lambda)$), because $\underline{\lambda}$ and $\overline{\lambda}$ are fixed a priori. Thus, the asymptotic properties of the modified estimator can simply be derived as resulting from minimizing the first term in equation (9), i.e. the estimation error in $\mathbb{1}_{i\in\mathcal{N}_{\theta,n}}$ does not matter asymptotically.

4. Since conditional on x_i, I_i is not random, the asymptotic analysis in Powell (1991) and Chamberlain (1994) applies analogously to the modified estimator provided that $E(1/n)\sum_i I_i x_i x_i'$ is uniformly positive definite in order to guarantee identification. For finite $\overline{\lambda}$ and $\underline{\lambda}$ this condition is satisfied for non-degenerate distributions of x_i. Under this assumption and standard regularity conditions as in Powell (1991), consistency and $\sqrt{n}$ asymptotic normality of the modified estimator follows immediately based on the analysis in Powell (1991) and Chamberlain (1994). Denoting $\eta' = (\beta', \lambda)$ and following Chamberlain's (1994, p. 204) notation (see also the appendix in Machado and Mata, 2000) as closely as possible, the asymptotic covariance matrix of the joint modified estimator $\hat{\eta} = (\hat{\beta}(\hat{\lambda})', \hat{\lambda})$ is given by

$$\left[A_0 \frac{\partial m(\eta_0)}{\partial \eta'}\right]^{-1} A_0 \ \theta(1-\theta) \ E \begin{pmatrix} x_i x_i' & I_i \frac{\partial \tilde{g}_i}{\partial \eta} x_i' \\ x_i I_i \frac{\partial \tilde{g}_i}{\partial \eta'} & I_i \frac{\partial \tilde{g}_i}{\partial \eta'} \frac{\partial \tilde{g}_i}{\partial \eta} \end{pmatrix} A_0{}' \left[A_0 \frac{\partial m(\eta_0)}{\partial \eta'}\right]^{-1'} \quad ,$$

[13] Alternatively, in cases, when our heuristic rule does not work, one can define

$$\mathbb{1}_{i\in\mathcal{N}_{\theta,n}} = I(\lambda x_i'\hat{\beta}_\theta(\lambda) + 1 > 0) \ \text{ and } \ I_i = I(\lambda x_i'\beta_\theta(\lambda) + 1 > 0) \ \text{ for all } \lambda \in [\underline{\lambda}, \overline{\lambda}].$$

However, this rule can not be easily applied in practical applications.

where $A_0 = \begin{pmatrix} E_K & 0 & 0 \\ 0 & \frac{\partial \beta_\theta(\lambda_0)}{\partial \lambda} & 1 \end{pmatrix}$, E_K is the $K \times K$ identity matrix,

and $m(\eta) = E \begin{pmatrix} [I(y_{\lambda,i} < x_i\beta) - \theta] \cdot x_i \\ I_i \cdot [I(y_{\lambda,i} < x_i\beta) - \theta] \cdot \frac{\partial \tilde{g}_i}{\partial \eta} \end{pmatrix}$.

The asymptotic results derived here differ from Chamberlain (1994) only by the fact that the dummy I_i enters the asymptotic first order condition for the second step of the estimator when optimizing over λ. Since I_i is nondecreasing for all observations when a smaller set $\mathcal{R}$ is used (i.e. $\overline{\lambda}$ decreases or $\underline{\lambda}$ increases) still containing λ_0, the asymptotic variance decreases (in the usual matrix sense), i.e. the modified estimator becomes asymptotically more efficient.

References

[1] Angrist, J., Chernozhukov, V., and Fernández–Val, I. (2004). Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure. *Unpublished Manuscript,* MIT.

[2] Box, G. and Cox, D. (1964). An Analysis of Transformation. *Journal of the Royal Statistical Society* B 26, 211–252.

[3] Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics* Vol.65, 109–154.

[4] Chamberlain, G. (1994) Quantile Regression, Censoring, and the Structure of Wages. In: Sims, C. (ed.), Advances in Econometrics: Sixth World Congress, Volume 1, Econometric Society Monograph.

[5] Fitzenberger, B. (1997) A Guide to Censored Quantile Regressions. In: G.S. Maddala and C.R. Rao, eds., *Handbook of Statistics*, 15, 405–437, North–Holland.

[6] Fitzenberger, B. (1998) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. *Journal of Econometrics*, 82, 235–287.

[7] Hunter, D. (2002) MATLAB CODE for (Non-)Linear Quantile Regressions. *http://www.stat.psu.edu/~dhunter/qrmatlab/* .

[8] Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica* Vol. 46, 33–50.

[9] Koenker, R. and D'Orey, V. (1987). Algorithm AS 229. Computing Regression Quantiles. *Statistical Algorithms, Royal Statistical Society* 383–393.

[10] Koenker, R. and Park, B. (1996). An Interior Fixed Point Algorithm for Quantile Regressions. *Journal of Econometrics* Vol. 71, 265–283.

[11] Machado, J. and Mata, J. (2000). Box-Cox Quantile Regressions and the Distribution of Firm Sizes. *Journal of Applied Econometrics*, Vol. 15, No.1, 253–264.

[12] Poirier, J. D. (1978). The Use of the Box-Cox Transformation in Limited Dependent Variable Models. *Journal of the American Statistical Association*, Vol.73, 284-287.

[13] Powell, J. (1991). Estimation of monotonic regression models under quantile restrictions. In: W.Barnett, J.Powell, and G.Tauchen, eds., Nonparametric and semiparametric methods in Econometrics, (Cambridge University Press, New York, NY) 357–384.

[14] Wooldridge, J.M. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, Massachusetts.