Implementing Box-Cox Quantile Regression*

Bernd Fitzenberger†
Ralf A. Wilke‡
Xuan Zhang§

November 2008

Abstract

The Box-Cox quantile regression model introduced by Powell (1991) is a flexible and numerically attractive extension of linear quantile regression techniques. Chamberlain (1994) and Buchinsky (1995) suggest a two stage estimator for this model but the objective function in stage two of their method may not be defined in an application. We suggest a modification of the estimator which is easy to implement. A simulation study demonstrates 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 two anonymous referees and Blaise Melly for helpful suggestions and comments. The authors gratefully acknowledge financial support by the German Research Foundation through the research project "Microeconometric modelling of unemployment durations under consideration of the macroeconomic situation" which was conducted at the Centre for European Economic Research, Mannheim, Germany.

†Corresponding author: Bernd Fitzenberger, Department of Economics, Albert Ludwigs-University Freiburg, 79085 Freiburg, Germany, Email: bernd.fitzenberger@vwl.uni-freiburg.de
‡University of Nottingham, E–mail: Ralf.Wilke@nottingham.ac.uk
§Goethe-University Frankfurt, E–mail: x.zhang@gmx.de

1 Introduction

This paper studies the practical implementation of the Box-Cox quantile regression (BCQR) model introduced by Powell (1991). We consider a numerical difficulty with the two step estimation approach for Box-Cox quantile regression as suggested by Chamberlain (1994) and Buchinsky (1995), denoted here as CBTS. Applications of the BCQR model in Buchinsky (1995) and Machado and Mata (2000) use the CBTS approach. In these applications, the estimated Box–Cox transformation parameter λ appears to vary too much and often hits the upper or lower bound, -2 or 2, which is set ex ante by the authors.[1] It may be possible that these problems reflect a numerical problem when implementing the estimator.

The CBTS approach involves an attractive two step procedure for the nonlinear BCQR model. In the first step, one runs a standard linear quantile regression for the given Box–Cox transformation parameter λ. Then, the second step involves a nonlinear optimization problem which is one–dimensional and which can be solved effectively by a grid search over λ. However, in the second step of the CBTS approach, the objective function may not be defined because the inverse of the Box-Cox transformation may not be computed for a share of the observations.

This numerical problem is by no means negligible because it can emerge in typical data situations. As a motivation, we illustrate the problem by an empirical example. We take an administrative data set of individual unemployment periods for West-Germany during the period 1981 to 1997. Lüdemann, Wilke and Zhang (2006) use similar data and provide all relevant details about it. The sample contains more than 79000 observations. From these data, we draw 500 independent random samples of size $n = 5000$ and estimate Box-Cox quantile regression with non–censored unemployment duration as dependent variable. We use the same 25 regressors that are used by Lüdemann et al. (2006) and estimate the model at the 0.1, 0.5 and 0.8-quantile. Table 1 reports the mean share of inadmissible observations for the original CBTS estimator. The mean is taken over the 500 samples. It is evident that the CBTS estimator fails at all quantiles since there is always a positive share of observations (between 0.4% and 1.7%) which are inadmissible for the computation of the inverse Box-Cox transformation.

This paper suggests a modified objective function which takes care of almost all inadmissible observations. Our modification is based on an exact theoretical result for the bivariate regression. However, this result may be violated when there are multiple regressors. For such cases, we suggest an additional modification which ensures that the objective function is still well defined. We perform several simulations which show that our modified estimator works well in finite samples for multiple regressions. We also sketch that the basic asymptotic properties of the original

[1] This observation was pointed out to us by an anonymous referee.

Table 1: Mean share of inadmissible observations in samples of unemployment duration data for the 0.1, 0.5 and 0.8-quantile. Standard deviations are in parentheses.

	$\theta = 0.1$		$\theta = 0.5$		$\theta = 0.8$	
CBTS[a]	0.4%	(0.003)	1.5%	(0.005)	1.7%	(0.004)

a: Two step estimation approach for Box-Cox quantile regressions as suggested by Chamberlain (1994) and Buchinsky (1995).

estimator carry over after the modification.

The remainder of this paper is structured as follows: Section 2 describes the Box–Cox quantile regression model and the CBTS estimation approach. Section 3 describes the modified estimation procedure and Section 4 presents simulation results to demonstrate the applicability of the modified estimator.

2 Two Stage Estimation

The BCQR model is a special case of a monotonic, possibly nonlinear regression model under quantile restrictions as introduced by Powell (1991). Let $\text{Quant}_\theta(y|x)$ denote the θ-quantile of the conditional distribution of a positive variable y given x. Using a single index assumption, we assume that the conditional θ–quantile of y given x depends upon a linear index $x'\beta_\theta$ through a nonlinear function $g(.)$, i.e.

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

where $g(.)$ is strictly monotonically increasing in $x'\beta_\theta$ and also depends upon the parameter λ_θ. Furthermore, $y > 0$, $x \in \mathbb{R}^K$ are observed, while the parameters $\beta_\theta \in \mathcal{B} \subset \mathbb{R}^K$ and $\lambda_\theta \in \mathbb{R}$ are unknown, and $\theta \in (0, 1)$.

Box and Cox (1964) introduce a specific functional form for models as in Equation (1). This Box–Cox transformation has gained popularity in applied econometrics as it contains the linear and the log-linear case as special forms. Powell (1991) discusses using the Box–Cox transformation for the model in (1).[2] For the Box–Cox transformation, the inverse of $g(x'\beta_\theta, \lambda_\theta)$ in its first argument is given by:

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

[2]To be precise, Powell (1991) analyzes the Bickel and Doksum (1981) transformation which is a generalization of the original transformation by Box and Cox (1964).

assuming $\lambda \in \mathcal{R}$ and $\mathcal{R} = [\underline{\lambda}, \overline{\lambda}]$ to be a finite closed interval. The Box–Cox transformation is quite attractive as it preserves the ordering of the observations due to the equivariance of quantiles with respect to the monotonically increasing transformation y_λ, i.e. $\text{Quant}_\theta(y_\lambda|x) = (\text{Quant}_\theta(y|x))_\lambda$. Using the Box–Cox transformation as the inverse of $g(x'\beta_\theta, \lambda_\theta)$ in its first argument, Equation (1) becomes

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

It is also important to note that a linear model results for

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

The estimation of β_θ and λ_θ is considered by Powell (1991), Chamberlain (1994), Buchinsky (1995), and Machado and Mata (2000). A Box–Cox quantile regression amounts to minimize the following distance function

$$\min_{\beta\in\mathcal{B},\lambda\in\mathcal{R}} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(y_i - (\lambda x_i'\beta + 1)^{1/\lambda}), \tag{3}$$

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. The resulting parameter estimates are denoted by $\hat{\lambda}_\theta$ and $\hat{\beta}_\theta$. 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 discussed below arises along the optimization process.

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

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

$$\widehat{\beta}_\theta(\lambda) \;=\; \text{argmin}_{\beta\in\mathcal{B}} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(y_{\lambda i} - x_i'\beta) \tag{4}$$

2. estimate λ_θ by solving

$$\min_{\lambda\in\mathcal{R}} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(y_i - (\lambda x_i'\widehat{\beta}_\theta(\lambda) + 1)^{1/\lambda}). \tag{5}$$

Note that the objective function in (4) cannot be used to estimate both β_θ and λ_θ (this would result in the degenerate estimator $\widehat{\beta}_\theta = 0$ and $\widehat{\lambda}_\theta = -\infty$). Chamberlain (1994) sketches the large sample theory of the CBTS estimator (see also Machado and Mata, 2000, Appendix B). Buchinsky (1995) derives large sample properties of this estimator for discrete regressors when applying the minimum distance method.

When implementing the CBTS procedure, we encountered the following numerical problem. It is not guaranteed that for every λ and all observations, the basis of the inverse Box-Cox transformation $\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1$ is strictly positive. However, this is necessary to conduct the second step of the above procedure.[3] Table 1 suggests that this numerical problem can easily occur in an application. The applied researcher therefore faces the question how to proceed in such cases. It may seem natural to omit the observations for which this condition is not satisfied and we suspect that this was done in applications in the past. But doing so raises some problems. First, the set of omitted observations 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 properties of the estimator. Third, should still the full set of observations be used in the first step? The main purpose of this paper is to suggest a structured way on how to implement the necessary omission of data points as a modification of CBTS and to analyze the consequences of doing so.

3 Modified Estimation

The estimation problem given in (5) can only be solved if

$$\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1 > 0 \tag{6}$$

for all $i = 1, \ldots, n$ and for all $\lambda \in \mathcal{R}$. As this condition depends on the first stage estimate and the specific value of λ it can be violated for several reasons. One reason is the finite sampling variation of $\widehat{\beta}_\theta(\lambda)$. Even at the true value of λ_θ the condition may therefore fail. Another important reason is that the objective function is evaluated for many values of λ during the iterative procedure to obtain the estimator. For all $\lambda \in \mathcal{R}$ except the true value, step 1 results in a generally misspecified linear quantile regression of y_λ on x_i. This misspecification can lead to a violation for several values

[3] The issue also arises for any other available computation methods in the literature such as 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 that paper. Koenker and Park (1996) provide an algorithm to solve the nonlinear minimization problem given in (3) directly through an iterative procedure. Both studies involve the inverse Box–Cox transformation $(\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1)^{1/\lambda}$, which is likely to fail for some observations along the search process for the parameter estimates due to the same reason as discussed for the CBTS estimator.

of λ. Since the true model is not known in an application, a general misspecification of the model may lead to a violation for all λ. Therefore, the condition is likely to fail for some observations in typical applications.

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 for a more efficient estimator.

Define the set of admissible observations $\mathcal{N}_{\theta,n}$ as those for which $\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in \mathcal{R}$. Note that $\mathcal{N}_{\theta,n}$ changes with n both due to the additional observations and due to the variation of $\widehat{\beta}_\theta$. A method for finding $\mathcal{N}_{\theta,n}$ in applications is suggested below. Instead of solving (5), we now solve in the second step

$$\min_{\lambda \in \mathcal{R}} \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{i \in \mathcal{N}_{\theta,n}} \cdot \rho_\theta(y_i - \tilde{g}_i[\lambda, \widehat{\beta}_\theta(\lambda)]), \tag{7}$$

where $\mathbb{1}_\omega$ is the indicator function for the event ω and for any $c \in \mathbb{R}$

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

Note it does not matter what value of c is chosen because the indicator function in Equation (7) is always zero in these cases. This notation is introduced to guarantee that the objective function involves a well defined sum from 1 to n. As sketched in Appendix A, the modified estimator has similar asymptotic properties as the original CBTS estimator. Specifically, it remains consistent and it shows an asymptotic distribution which only involves a slight modification.

How to choose $\mathcal{N}_{\theta,n}$ and $\mathcal{R}$ in an application?

As a purely hypothetical rule, one could simply choose $\mathcal{N}_{\theta,n}$ as the set of observations for which $\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1 > 0$ is true for all $\lambda \in \mathcal{R}$. However, this is not a practical rule because in an application one cannot determine in advance whether the condition holds for all $\lambda \in \mathcal{R}$. Another difficulty is that the elements of $\mathcal{N}_{\theta,n}$ also depend on the actual set $\mathcal{R} = [\underline{\lambda}, \overline{\lambda}]$, i.e. on the specific choice of $\underline{\lambda}$ and $\overline{\lambda}$. For these reasons, a practical alternative is needed.

Analogous to the model for means (least squares regression), exact guidance on the choice of $\underline{\lambda}$ and $\overline{\lambda}$ cannot be provided but similarly, both limits should not be too large in absolute values. Given that two focal parameter values of the Box-Cox model are $\lambda = 0$ (logarithmic) and $\lambda = 1$ (linear), it is advisable to include them in the interval. In our applications and simulations, we use different such intervals (in the simulations all these intervals include the true parameter). Our

simulations do not provide evidence that the specific choice of the interval affects the estimation results in a noticeable way despite the fact that the crucial condition is more likely to be violated if $|\lambda|$ tends to be large. A large interval, however, increases the computation time significantly and should therefore be avoided. Based on our experience, we suggest the interval $[-1.5, 2]$ as a practical first rule of thumb for applications.

The choice of $\mathcal{N}_{\theta,n}$ is more difficult and we cannot present a rule that detects all relevant observations in any case. In the following we suggest a rule that is, however, strictly valid in the bivariate regression case $K = 2$ involving an intercept. In this case, 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 $\tilde{g}_i[\lambda, \widehat{\beta}_\theta(\lambda)]$ is well defined (see Proposition 1 and its proof in Appendix B). For the case $K > 2$, Appendix B provides arguments supporting that the rule generally works well for all practical purposes. This is also confirmed by our simulations in Section 4. More precisely, we suggest the following simple heuristic rule for the choice of $\mathcal{N}_{\theta,n}$:

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

Unfortunately, (HR) does not necessarily detect all inadmissible observations in models with $K \geq 3$. Appendix B formalizes why (HR) does not detect all possibly inadmissible observations and it provides an example. Moreover, it argues why (HR) is able to detect most of the inadmissible observations in typical data situations. This also confirmed by the empirical example of the introduction where we observe that only very few inadmissible observations are not detected by HR, see Table 2. Moreover, our simulations in Section 4 show only a very small number of cases for which applying (HR) fails during the search for estimating λ_θ.

Table 2: Mean share of not detected inadmissible observations in samples of unemployment duration data for the 0.1, 0.5 and 0.8-quantile. Standard deviations are in parentheses.

	$\theta = 0.1$		$\theta = 0.5$		$\theta = 0.8$	
Modified estimator using (HR)	0.007%	(0.000)	0.000%	(0.000)	0.000%	(0.000)

In case our rule (HR) is violated, the researcher needs more guidance how to proceed with the remaining inadmissible observations. We suggest as a practical modification to set

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

for some small $\epsilon > 0$ in order to make the objective function well defined.[4] Based on our experience, a violation of (HR) is a rare event for $K > 3$. For this reason, the modification is unlikely to affect the final estimates. This becomes apparent in the next section.

4 Simulations

This section assesses the finite sample performance of the modified estimator (7) through Monte Carlo simulations. For this purpose 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 standard normal distribution with bounds[5] $[-1, 1]$ and it is independent of x. We use both a homoscedastic and a heteroscedastic design.

For the homoskedastic design, the scale function $\sigma(x'\beta)$ is 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 and we choose $\mathcal{R} = [-0.5, 2.5]$. We draw 1000 independent random samples from this model. Estimates for β are obtained using the algorithm implemented in TSP Version 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.[6] Table 3 presents the results for four experiments based on 1000 replications with sample sizes $n = 100$ and $n = 1000$.[7]

[4]This modification is based on a suggestion by Blaise Melly. Note that the additional modification (8) for admissible observations differs from the modified objective function given in (7) involving setting an arbitrary c for the non–admissible observations, which are irrelevant for the optimization.

[5]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).

[6]We also replicate the simulation study by using the Koenker and Park (1996) algorithm for MATLAB provided by Hunter (2002). 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 is much faster than for the grid search and the results only marginally changed. These results are available upon request.

[7]We also considered simulation designs with more than three regressors and different marginal distributions of the covariates. The performance of the modified estimator is very similar in these designs and results are therefore not presented.

Table 3 shows that the proposed modified estimator performs well even in designs where the numerical problem of the original CBTS estimator is by no means negligible. On average, between 16 and 17 percent of all observations are excluded by our modified estimation approach for these simple data generating processes. The results confirm that our modified estimator works well as the averages of the estimates are close to the true parameter values. 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 1000. The distribution of $\hat{\lambda}$ is nicely concentrated around the true parameter $\lambda = 1$ and the variance decreases with the sample size.

To analyze how the performance of the modified estimator changes with different values of λ and with various search intervals, we consider two additional simulation designs. The distributions of x and ϵ are the same in these two scenarios, but the parameters are different. We restrict the model to the homoskedastic case. x_1 is standard normally distributed, and truncated in $[-2, 2]$. x_2 follows an exponential distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$. The error term ϵ is uniformly distributed between -0.5 and 0.5. The parameters are chosen such that the sample variance of y_λ is similar. Table 4 describes the details of the simulation designs.

The simulation results are presented in Tables 5 and 6. Again, the modified estimator performs well in all cases. This is remarkable, because in the case of a wider $\mathcal{R}$, there are on average up to 17% inadmissible observations, if we apply the original CBTS estimator. This demonstrates that the modified estimator shows satisfactory statistical properties and also that the choice $\mathcal{R}$ has only an effect on the variance but not on the average of the estimated coefficients. The smaller standard deviations in the case of a smaller interval $\mathcal{R}$ are likely to be due to the smaller support of $\widehat{\lambda}$ or due to the fact that less observations have been excluded.

We observed only a few violations of our heuristic rule (HR). Table 7 shows the relative frequency that the violation occurs for any observation at each grid point (denoted by 'Total') of λ considered or at one of the grid points in each estimation (denoted by 'Observations') presented above, respectively. Table 7 also reports the frequencies for the simulated sample from the real data in Lüdemann et al. (2006) as described in the introduction. For all simulated data sets, the number of observed violations is extremely small. In such cases we apply the additional modification suggested at the end of Section 3. Apparently, they do not distort the pattern of the simulation results.

Appendix

Appendix A: Asymptotic Properties of the modified estimator

We sketch 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) and Machado and Mata (2000). For a given θ-quantile, 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[8] in the population (Angrist, Chernozhukov, and Fernández–Val, 2006, 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 $\widehat{\beta}_\theta(\lambda)$ 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

$$\int_x \left\{ \int_y x(\mathbb{1}_{(y_\lambda < x'\beta)} - \theta) f(y|x) dy \right\} f(x) dx = Ex(\mathbb{1}_{(y_\lambda < x'\beta)} - \theta) = 0$$

   as a population moment condition. 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 $\text{Quant}_\theta(y_\lambda|x) = x'\beta_\theta(\lambda)$ for general λ (Angrist et al., 2006) the population moment condition suffices for $\widehat{\beta}_\theta(\lambda)$ to be a $\sqrt{n}$–consistent estimator of $\beta_\theta(\lambda)$, as suggested by Chamberlain (1994) and shown explicitly in Fitzenberger (1998).

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

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

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

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

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

Note that there is no a priori reason why the misspecified model should not involve inadmissible observations for the population coefficients $\beta_\theta(\underline{\lambda}), \beta_\theta(\overline{\lambda})$ unless of course one of the two values $\underline{\lambda}, \overline{\lambda}$ corresponds to the true population value of λ. Therefore, I_i may be zero for some observations.
$\sqrt{n}$–consistency of $\widehat{\beta}_\theta(\lambda)$ 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.[9]

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 (7) because the difference

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

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 $\widehat{\beta}_\theta(\lambda)$), 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 $\hat{\eta} = (\widehat{\beta}_\theta(\hat{\lambda}_\theta)', \widehat{\lambda}_\theta)$ is given by

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

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 $m(\eta) = E \begin{pmatrix} [\mathbb{1}_{(y_{\lambda,i} < x_i\beta)} - \theta] \cdot x_i \\ I_i \cdot [\mathbb{1}_{(y_{\lambda,i} < x_i\beta)} - \theta] \cdot \frac{\partial \tilde{g}_i}{\partial \eta} \end{pmatrix}$.

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

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

However, strictly speaking, the implied alternative rule is infeasible in practical applications.

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 – except in very rare exceptions – when a smaller set $\mathcal{R}$ is used (i.e. when $\overline{\lambda}$ 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.

Allowing for misspecification of the Box–Cox quantile regression analogous to Angrist, Chernozhukov, and Fernández–Val (2006), the asymptotic covariance matrix of the joint modified estimator is given by

$$\left[A_0 \frac{\partial m(\eta_0)}{\partial \eta'}\right]^{-1} A_0 \; E\left[(\mathbb{1}_{(y_{\lambda,i}<x_i\beta)}-\theta)^2 \begin{pmatrix} x_i x_i' & I_i \frac{\partial \tilde{g}_i}{\partial \eta} x_i' \\ x_i I_i \frac{\partial \tilde{g}_i}{\partial \eta'} & I_i \frac{\partial \tilde{g}_i}{\partial \eta'} \frac{\partial \tilde{g}_i}{\partial \eta} \end{pmatrix}\right] A_0{}' \left[A_0 \frac{\partial m(\eta_0)}{\partial \eta'}\right]^{-1'} .$$

Appendix B: Theoretical Motivation for the HR

We first present and prove a result for the bivariate regression model. We then discuss why it may not hold for a model with more regressors and we present a counter example. Finally we provide some reasoning why these violations are rare.

Proposition 1: *For the bivariate regression model $K = 2$ (one regressor plus an intercept) and some $\underline{\lambda} \leq 0$ and $\overline{\lambda} \geq 0$, consider the set of observations $\mathcal{N}_{\theta,n}$ for which $\underline{\lambda} x_i' \widehat{\beta}_\theta(\underline{\lambda}) + 1 > 0$ and $\overline{\lambda} x_i' \widehat{\beta}_\theta(\overline{\lambda}) + 1 > 0$. Assume that the rank of the design matrix in the sample $\mathcal{N}_{\theta,n}$ is equal to K. For $i \in \mathcal{N}_{\theta,n}$, it follows that $\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. In case of non–uniqueness of the coefficients, the conditions are supposed to hold for all possibly non–unique $\widehat{\beta}_\theta(\lambda)$.*

Proof of Proposition 1: Without loss of generality, assume that $\overline{\lambda} > 0$. In the following, we will show that $\overline{\lambda} x_i' \widehat{\beta}_\theta(\overline{\lambda}) + 1 > 0$ implies $\lambda x_i' \widehat{\beta}_\theta(\lambda) + 1 > 0$ for all $\lambda \in [0, \overline{\lambda}]$. Therefore, assume $\lambda \geq 0$ in the following. The proof proceeds in a number of steps (steps 1 to 8 below).

The flow of argument in the proof is as follows: Step 1 characterizes the quantile regression estimates as a function of λ. The set of interpolated data points in basic solutions only change at a finite number of values for λ. In case of nonuniqueness, all estimates are convex linear combinations of such estimates. Step 2 states that there are no inadmissible observations both for $\lambda = 0$ and for small positive λ. Steps 3 to 7 show that $\partial x_i' \widehat{\beta}_\theta(\lambda)/\partial\lambda < 1/\lambda^2 \equiv \partial(-1/\lambda)/\partial\lambda$ for $x_i' \widehat{\beta}_\theta(\lambda)$ being close to $-1/\lambda$, which suffices to show the result of Proposition 1 between critical values of λ where the set of interpolated data points changes. This is because the distance between the fitted values and the critical limit $-1/\lambda$ where an observation becomes inadmissible increases when λ falls. Step 8 concludes the proof by showing that the absolute distance between the fitted

values and the critical limit $-1/\lambda$ does not strictly decrease when λ reaches a critical value where the set of interpolated data points changes. Step 9 discusses the typical case of nonuniqueness in a bit more detail.

Details of the proof:

For the proof, define $\mathcal{N}_{\theta,n}$ the set of observations for which $\overline{\lambda}x_i'\widehat{\beta}_\theta(\overline{\lambda}) + 1 > 0$ and assume that the rank of the design matrix in the sample $\mathcal{N}_{\theta,n}$ is equal to K.

1. This point establishes the following characterization of the quantile regression estimates as a function of λ: In a given subsample $\mathcal{N}_{\theta,n}$, there exist a finite sequence of $\lambda_j, j = 0, 1, ..., J$ with $\lambda_0 = 0 < \lambda_1 < ... < \lambda_J = \overline{\lambda}$ with the quantile regression coefficients $\widehat{\beta}_\theta(\lambda)$ and the set of K linearly independent (regarding the matrix formed by the regressor vectors) interpolated data points being the same for all $\lambda \in [\lambda_{j-1}, \lambda_j]$. In case of nonuniqueness of the quantile regression coefficients, all solutions of the minimization problem are convex combinations of coefficient vectors satisfying this interpolation property. Therefore, it suffices to analyze these coefficient estimates. At $\lambda = \lambda_j$ (j=1,...,J-1), the quantile regression coefficients $\widehat{\beta}_\theta(\lambda)$ the coefficient vector associated with the estimated line through the set of interpolated data points in $(\lambda_{j-1}, \lambda_j)$ and in $(\lambda_j, \lambda_{j+1})$, respectively, is the same. Thus, the set of interpolated data points only changes at a finite number of $\lambda_j, j = 1, ..., J-1$.

   This characterization is shown as follows: The interpolation property of linear quantile regression (Koenker and Bassett, 1978, Theorem 3.1) implies that there exist K observations with $x_{i(h)}'\widehat{\beta}_\theta(\lambda) = y_{i(h),\lambda}$[10] for $h = 1, ..., K$. Define $X_H = (x_{i(1)}, x_{i(2)}, ..., x_{i(K)})$ and $Y_{H,\lambda} = (y_{i(1),\lambda}, y_{i(2),\lambda}, ..., y_{i(K),\lambda})'$ $(H = (i(1), i(2), ..., i(K)))$. The indices $i(1), ..., i(K) \in \mathcal{N}_{\theta,n}$ denote K distinct individual observations with linearly independent $x_{i(h)}$. The interpolation property is summarized as

$$(IP) \qquad \widehat{\beta}_\theta = X_H^{-1} Y_{H,\lambda} \quad \text{with} \quad rank[X_H] = K \quad .$$

   The interpolation property is implied by the fact that estimating a linear quantile regression involves solving a standard linear program.

   $\widehat{\beta}_\theta$ is a possibly nonunique solution to the estimation problem, if the following subgradient condition holds (see Koenker and Bassett, 1978, p. 40, and Koenker and d'Orey, 1978, p. 385)

$$\psi(\widehat{\beta}_\theta, \lambda, w) = \sum_{i=1}^{N} [1/2 - 1/2 sgn^*(y_{i,\lambda} - x_i'\widehat{\beta}_\theta; -x_i'w) - \theta] x_i'w \geq 0 \qquad (10)$$

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

for all $w \in \mathbb{R}^K$, where $sgn^*(u;z) = sgn(u)$ if $u \neq 0$ and $sgn^*(u;z) = sgn(z)$ if $u = 0$. $\widehat{\beta}_\theta$ is unique when the inequality in (10) holds strictly for all $w \in \mathbb{R}^K$. Nonuniqueness occurs when $\psi(\widehat{\beta}_\theta, w) = 0$ for some w. Then, there exists a basic solution in direction w with K linearly independent (regarding the matrix formed by the regressor vectors) interpolated data points. In fact, due to the linear programming nature of the problem, all coefficient estimates can be written as convex combinations of such basic solutions (Koenker, 2005, p. 36). Therefore, it suffices to analyze solutions $\widehat{\beta}_\theta$ satisfying (IP).

Note that in Equation (10), λ only affects the first argument of $sgn^*(.;.)$. For data points, which are not interpolated ($y_{i,\lambda} - x_i'\widehat{\beta}_\theta \neq 0$), the sign of the residuals does not change in response to an infinitesimally small change in λ. Thus, a change in the first argument of $sgn^*(.;.)$ in Equation (10) in response to a change in λ does not result in a change of the subgradients for those observations which are not interpolated at the current $\widehat{\beta}_\theta$. In contrast, for interpolated data points, which remain interpolated despite an infinitesimally small change in λ, the gradient condition (10) remains unchanged for all w.

For the subsequent arguments, one needs to distinguish various cases, see Koenker and D'Orey (1987, p. 385) and Koenker (2005, Section 6.3) for a similar argument when estimating the entire process of linear quantile regressions coefficients as a function of θ ('Parametric Programming'). Analogously, with a variation of λ, the set of interpolated observations H continues to define the 'basic' solution for the estimator, i.e. $\widehat{\beta}_\theta = X_H^{-1} Y_{H,\lambda}$, until the inequality in Equation (10) is violated for one w. Nevertheless, with constant H, $\widehat{\beta}_\theta$ changes with λ because $\partial y_{i(h),\lambda}/\partial\lambda = -log(y_{i(h)}) y_{i(h)}^\lambda/\lambda^2$ resulting in a change in $Y_{H,\lambda}$.

A change in the set of interpolated data points H occurs at the critical values λ_j ($j = 1, ..., J-1$). At such a critical value, the residual $y_{l,\lambda} - x_l'\widehat{\beta}_\theta$ for some data points $l \in L$, becomes interpolated at $\lambda = \lambda_j$, i.e. $y_{l,\lambda_j} = x_l'\widehat{\beta}_\theta$. Note that the set L may involve more than one element. The data points $l \in L$ are not interpolated for $\lambda_{j-1} < \lambda < \lambda_j$. The change in the subgradients (10) at λ_j is

$$\psi(\widehat{\beta}_\theta, \lambda_j, w) - \psi(\widehat{\beta}_\theta, \lambda, w) = \sum_{l \in L} 1/2 \left[sgn(x_l'w) + sgn(y_{i,\lambda} - x_l'\widehat{\beta}_\theta) \right] x_l'w \geq 0 \quad ,$$

which is nonnegative because $sgn(x_l'w)x_l'w = |x_l'w| \geq 0$ and $|sgn(y_{i,\lambda} - x_l'\widehat{\beta}_\theta)| \leq 1$. Thus, $\widehat{\beta}_\theta$ remains optimal but also any other combination of K linearly independent data points in $H \cup L$ now defines the coefficient vector $\widehat{\beta}_\theta$ via the interpolation property (IP).

Now, when λ increases above λ_j, the set of interpolating data points typically changes because not all elements in $H \cup L$ can remain interpolated at the same resulting coefficient vector

$\widehat{\beta}_\theta$.[11] To show this, note that x_l can be represented as a linear combination $\sum_{h=1}^{K} g_h x_{i(h)}$ of the regressor vectors in H. The weights g_h sum up to one, i.e. $\sum_{h=1}^{K} g_h = 1$, because the first element of the regressor vector involves 1 which reflects the intercept. Then, we have $y_{l,\lambda_j} = x_l'\widehat{\beta}_\theta = \sum_{h=1}^{K} g_h x_{i(h)}'\widehat{\beta}_\theta = \sum_{h=1}^{K} g_h y_{i(h),\lambda_j}$. Since we assume $\lambda > 0$, this implies $y_l^{\lambda_j} = \sum_{h=1}^{K} g_h y_{i(h)}^{\lambda_j}$. Define the difference $D = y_l^{\lambda_j} - \sum_{h=1}^{K} g_h y_{i(h)}^{\lambda_j}$, then

$$\frac{\partial D}{\partial \lambda} = \log(y_l) y_{l)}^{\lambda} - \sum_{h=1}^{K} g_h \log(y_{i(h)}) y_{i(h)}^{\lambda} = \frac{1}{\lambda}\left[\log(y_l^{\lambda}) y_l^{\lambda} - \sum_{h=1}^{K} g_h \log(y_{i(h)}^{\lambda}) y_{i(h)}^{\lambda}\right]. \tag{11}$$

To complete this step, we now show that typically $\partial D/\partial\lambda \neq 0$ and therefore not all elements in $H \cup L$ can remain interpolated when λ increases above λ_j. Because the function $m(z) = zlog(z)$ is a strictly convex function with $m''(z) < 0$, it follows that

$$\left[\log(y_l^{\lambda}) y_l^{\lambda} - \sum_{h=1}^{K} g_h \log(y_{i(h)}^{\lambda}) y_{i(h)}^{\lambda}\right] < 0$$

unless $y_{i(h)}$ is the same for all h and coincides with y_l. This is because y_l^{λ} is the weighted average of $y_{i(h)}^{\lambda}$ for $h = 1, ..., K$.

In the case, where $y_{i(h)}$ is the same for all h and coincides with y_l, y_l^{λ} would have been interpolated for λ smaller than λ_j. This would be in contradiction to the rationale behind the identification of the lth observation above.

Since observation l becomes interpolated at $\lambda = \lambda_j$, it follows that $\partial D/\partial\lambda < 0$ in an interval around λ_j and therefore the quantile regression for λ slightly above λ_j can not interpolate all points in $H \cup L$. The set of K interpolated data points may change at λ_j. In fact, the data point l will change sides of the regression still interpolating the data points in the former basic set H because $\partial D/\partial\lambda$ is strictly negative. Since the share of observations lying strictly on one side of the fitted regression line is bounded (Koenker and Bassett, 1978, Theorem 3.4), this change of sides of observation l typically results in a new set of interpolated data points associated with the coefficient estimates $\widehat{\beta}_\theta$ for λ increasing above λ_j.

2. For $\lambda = 0$ corresponding to the logarithmic transformation, the invertibility of the Box–Cox transformation is given and there are no inadmissible observations. There exists a sufficiently small $\lambda_0 > 0$ such that $\lambda x_i'\widehat{\beta}_\theta(\lambda) + 1 > 0$ holds for all data points i and $0 < \lambda \leq \lambda_0$. Such a λ_0 exists because the set of interpolated data points remains unchanged (under the provisions

[11] A potential apparent exception would be a case where all interpolated points take the same values of the response variable and, therefore, of the fitted values. In this case, the slope parameter will be zero. The discussion below shows that this case leads to a contradiction.

of point 1) for an infinitesimal increase of λ from 0 and the change in $\widehat{\beta}_\theta(\lambda)$ is therefore continuous.

3. The condition $\lambda x_i'\widehat{\beta}_\theta(\lambda)+1>0$ is equivalent to $x_i'\widehat{\beta}_\theta(\lambda)>-1/\lambda$. To prove the our result, we show in the following that $\partial x_i'\widehat{\beta}_\theta(\lambda)/\partial\lambda<1/\lambda^2\equiv\partial(-1/\lambda)/\partial\lambda$ for $x_i'\widehat{\beta}_\theta(\lambda)$ being close to $-1/\lambda$. This suffices because point 2 implies that $\lambda x_i'\widehat{\beta}_\theta(\lambda)+1>0$ holds for all data points for small positive λ.

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

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

and

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

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

5. 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 $y_{(h),\lambda}+1/\lambda=x_{(h)}'\widehat{\beta}_\theta(\lambda)+1/\lambda$ increases.

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

7. This step requires $K=2$: 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 no further consideration). For the critical data point i in the previous step, it follows that $g_1=y_{(2)}^\lambda/(y_{(2)}^\lambda-y_{(1)}^\lambda)$ and $g_2=1-g_1=y_{(1)}^\lambda/(y_{(1)}^\lambda-y_{(2)}^\lambda)$. Then, after some straightforward manipulations, we obtain

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

The inequality holds because $[\log(y_{(1)})-\log(y_{(2)})]$ and $[\lambda(y_{(2)}^{\lambda} - y_{(1)}^{\lambda})]$ have opposite signs. This is due to both $log(y)$ and y^{λ} being increasing functions of y.

8. After more than an infinitesimally small change of λ, it may occur that the set of interpolated observations changes. For the specific $\lambda = \lambda_j$, where this occurs, the linear quantile regression through the data points, which are interpolated so far, will typically interpolate other data points indexed by $l \in \{1, ..., n\}$ with $x_l'\hat{\beta}_\theta(\lambda) = y_{l,\lambda}$ in addition to $i(h), h = 1, ..., K$. If λ changes infinitesimally further, then the data points l will typically replace some of the interpolated $i(h)$ in the set of interpolated data points with linearly independent regressor vectors (see step 1).
   The fact, that at the critical λ_j the estimated quantile regression interpolates both the data points l and the data points $i(h), h = 1, ..., K$, is due to the continuous nature of the optimization problem and the local monotonicity of the Box–Cox–transformation (see step 1). Thus, at the critical λ_j, data points can not just discontinuously switch sides of the estimated quantile regressions.
   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 the derivative to the left $(\partial\Delta/\partial\lambda)_{d\lambda<0}$ and the derivative to the right $(\partial\Delta/\partial\lambda)_{d\lambda>0}$ are non–positive for critical observations where the quantile regression interpolates $-1/\lambda$.

9. The case of non–uniqueness of the coefficient estimates $\widehat{\beta}_\theta(\lambda)$ results from a rare feature of the design matrix in the case of purely discrete regressors, see Koenker (2005, p. 36). In such a case, non–uniqueness applies to all λ and the all non–unique $\widehat{\beta}_\theta(\lambda)$ are convex combinations of a fixed number of coefficient estimates with K linearly independent interpolated data points (see Koenker and D'Orey, 1987). The latter change with λ as analyzed above.

The proof proceeds in an analogous way for $\underline{\lambda} < 0$ showing that if $\lambda x_i'\widehat{\beta}_\theta(\lambda) + 1 > 0$ 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 $x_1'\widehat{\beta}_\theta(\lambda) = -1/\lambda = -0.5$ for $\lambda = 2$. For $\lambda = 1.99$, the fitted value is $x_1'\widehat{\beta}_\theta(\lambda) = -0.50310 < -0.50251 = -1/\lambda$ and for $\lambda = 2.01$, the fitted value is $x_1'\widehat{\beta}_\theta(\lambda) = -0.49691 > -0.49751 = -1/\lambda$. 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 (12) 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.

Discussion of the HR

Proposition 1 can be explained as follows: Assume that for some $\lambda > 0$ and some critical observation j the linear quantile regression in step 1 of the estimation procedure yields $x_j'\widehat{\beta}_\theta(\lambda) = -1/\lambda$. The fitted value $x_j'\widehat{\beta}_\theta(\lambda)$ is a weighted average of $y_{i(h),\lambda}$ for the two interpolated, linearly independent observations $(i(h), h = 1, 2)$, which are fitted perfectly by the estimated regression line, see Theorem 3.1 in Koenker and Bassett (1978).[12] The regressor vector for the critical observation j can be expressed as $x_j = g_1 x_{(1)} + (1 - g_1) x_{(2)}$ with weight g_1 and $x_{(h)} = x_{i(h)}$. Typically, an infinitesimally small change in λ does not change the set the interpolated observations, but it

[12]The interpolation property is due to the fact that estimating a linear quantile regression involves solving a linear program.

affects differentially the interpolated values $y_{i(h),\lambda}$.[13] We show in the proof that a reduction in λ results in an increase of the distance $\Delta \equiv x_j'\widehat{\beta}_\theta(\lambda) + 1/\lambda$, a result which is formally expressed in Equation (12) below. Thus, if an observation j is admissible for some $\lambda > 0$ it is also admissible for all $\tilde{\lambda} \in [0, \lambda]$. This intuition is rigorously formalized in the proof.

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 (simulation 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.

Next, we argue why HR is still a good choice in models with $K > 2$ even when it does not detect all inadmissible observations.

The proof of Proposition 1 considers critical observations with regressor values x_j resulting in fitted values $x_j'\widehat{\beta}_\theta(\lambda)$ equal to $-1/\lambda$ for some λ. The fitted values are weighted averages of the fitted values of the K interpolated observations with $x_j = \sum_{h=1}^{K} g_h x_{(h)}$ defining the weights g_h for the interpolated observations $h = 1, ..., K$. To investigate the change in the set of observations satisfying the condition as given in (6) in response to a change in λ, the following condition is critical (see proof of Proposition 1)

$$\frac{\partial \Delta}{\partial \lambda} \equiv \sum_{h=1}^{K} g_h \log(y_{(h)}) y_{(h)}^{\lambda} < 0 \tag{12}$$

with $\Delta = \sum_{h=1}^{K} g_h y_{(h)}^{\lambda} = 0$ and $\sum_{h=1}^{K} g_h = 1$, where Δ corresponds to the distance between $x_j'\widehat{\beta}_\theta(\lambda)$ and $-1/\lambda$. If the condition given in (12) is satisfied for $K > 2$ and all $\lambda \in \{\underline{\lambda}, \overline{\lambda}\}$, then the result in Proposition 1 applies in this case as well. The proof of Proposition 1 above is formulated for the case with general K and condition (12) is only needed in step 7 of the proof.

Note that the condition in (12) 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 $y_{(h),\lambda}$ 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

[13] Unless $y_{i(h),\lambda}$ coincide and there would be no critical observation j.

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 the condition in (12). This typical setup does not hold in our counter example 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. Our extensive simulation results in the Section 4 are consistent with our reasoning here.

Tables and Figures

Table 3: Simulation results for 1000 Monte Carlo samples ($\theta = 0.5$). Averages with standard deviations in parentheses.

	Homoskedastic				Heteroskedastic			
	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
% of i not in $\mathcal{N}_{0.5,n}$	17.8%	(0.023)	18.3%	(0.007)	17.6%	(0.023)	18.2%	(0.007)
$\widehat{\beta}_0$	10.068	(1.217)	9.989	(0.352)	10.069	(1.046)	9.987	(0.273)
$\widehat{\beta}_1$	1.010	(0.164)	0.999	(0.047)	1.009	(0.131)	0.999	(0.033)
$\widehat{\beta}_2$	2.017	(0.367)	2.001	(0.105)	2.012	(0.269)	1.999	(0.065)
$\widehat{\lambda}$	0.999.	(0.067)	0.999	(0.019)	1.001	(0.059)	0.999	(0.015)

Table 4: Simulation designs

	β_0	β_1	β_2	λ	$\mathcal{R}$ 1	$\mathcal{R}$ 2	grid
Design A	3	-1.5	1.5	0.5	$[-0.5, 2]$	$[0, 1.5]$	0.005
Design B	-3	1.5	-1.5	-0.5	$[-1.5, 1.5]$	$[-1, 0.5]$	0.005

Table 5: Design A: Finite sample evidence from 1000 Monte Carlo samples. Averages with standard deviations in parentheses.

A	$\mathcal{R} = [-0.5, 2]$				$\mathcal{R} = [0, 1.5]$			
	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
% of i not in $\mathcal{N}_{0.5,n}$	14.1%	(0.027)	14.5%	(0.008)	8.0%	(0.023)	8.3%	(0.007)
$\widehat{\beta}_0$	3.008	(0.262)	3.000	(0.082)	3.001	(0.225)	3.000	(0.067)
$\widehat{\beta}_1$	−1.527	(0.324)	−1.498	(0.096)	−1.516	(0.278)	−1.500	(0.076)
$\widehat{\beta}_2$	1.536	(0.372)	1.497	(0.111)	1.523	(0.319)	1.500	(0.090)
$\widehat{\lambda}$	0.499	(0.086)	0.499	(0.027)	0.498	(0.075)	0.500	(0.022)

Table 6: Design B: Finite sample evidence for 1000 Monte Carlo samples. Averages with standard deviations in parentheses.

B	$\mathcal{R} = [-1.5, 1.5]$				$\mathcal{R} = [-1, 0.5]$			
	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
% of i not in $\mathcal{N}_{0.5,n}$	16.4%	(0.035)	16.7%	(0.011)	1.8%	(0.013)	1.9%	(0.004)
$\widehat{\beta}_0$	−3.080	(0.560)	−3.008	(0.127)	−3.009	(0.344)	−3.002	(0.108)
$\widehat{\beta}_1$	1.655	(0.713)	1.514	(0.150)	1.526	(0.383)	1.506	(0.125)
$\widehat{\beta}_2$	−1.695	(0.835)	1.519	(0.175)	−1.540	(0.447)	−1.509	(0.144)
$\widehat{\lambda}$	−0.508	(0.175)	−0.502	(0.043)	−0.495	(0.110)	−0.500	(0.036)

Table 7: Relative Frequency of violations of HR

500 simulated samples of size $n = 5000$ from real data in Lüdemann et al. (2006), see Introduction, and 301 grid points for $\lambda \in \mathcal{R} = [0, 1.5]$.

	$(\theta = 0.1)$	$(\theta = 0.5)$	$(\theta = 0.8)$
Total[a]	$\frac{336}{5000*500*301} = 4.5 \cdot 10^{-7}$	$\frac{725}{5000*500*301} = 9.6 \cdot 10^{-7}$	$\frac{7240}{5000*500*301} = 9.6 \cdot 10^{-6}$
Observations[b]	$\frac{48}{5000*500} = 1.9 \cdot 10^{-5}$	$\frac{181}{5000*500} = 7.2 \cdot 10^{-5}$	$\frac{742}{5000*500} = 3.0 \cdot 10^{-4}$

1000 Monte Carlo samples of size n as described in Table 3 and 601 grid points for $\lambda \in \mathcal{R} = [-0.5, 2.5]$.

Simulation	Homo.$(n = 100)$	Homo.$(n = 1000)$	Hetero.$(n = 100)$	Hetero.$(n = 1000)$
Total[a]	$\frac{320}{100*1000*601} = 5.3 \cdot 10^{-6}$	$\frac{5}{1000*1000*601} = 8.3 \cdot 10^{-8}$	$\frac{483}{100*1000*601} = 8.0 \cdot 10^{-6}$	$\frac{1}{1000*1000*601} = 1.6 \cdot 10^{-8}$
Observations[b]	$\frac{25}{100*1000} = 2.5 \cdot 10^{-4}$	$\frac{1}{1000*1000} = 1.0 \cdot 10^{-5}$	$\frac{30}{100*1000} = 3.0 \cdot 10^{-4}$	$\frac{1}{1000*1000} = 1.0 \cdot 10^{-5}$

1000 Monte Carlo samples of size n as described in Table 4.

	501 grid points $\lambda \in \mathcal{R}1 = [-0.5, 2.0]$		301 grid points $\lambda \in \mathcal{R}2 = [0, 1.5]$	
Simulation	A1$(n = 100)$	A1$(n = 1000)$	A2$(n = 100)$	A2$(n = 1000)$
Total[a]	$\frac{218}{100*1000*501} = 4.4 \cdot 10^{-6}$	$\frac{18}{1000*1000*501} = 3.6 \cdot 10^{-7}$	$\frac{0}{100*1000*301} = 0$	$\frac{0}{1000*1000*301} = 0$
Observations[b]	$\frac{10}{100*1000} = 1.0 \cdot 10^{-4}$	$\frac{5}{1000*1000} = 5.0 \cdot 10^{-5}$	$\frac{0}{100*1000} = 0$	$\frac{0}{1000*1000} = 0$

1000 Monte Carlo samples of size n as described in Table 4.

	501 grid points $\lambda \in \mathcal{R}1 = [-1.5, 1.5]$		301 grid points $\lambda \in \mathcal{R}2 = [-1, 0.5]$	
Simulation	B1$(n = 100)$	B1$(n = 1000)$	B2$(n = 100)$	B2$(n = 1000)$
Total[a]	$\frac{67}{100*1000*601} = 1.1 \cdot 10^{-6}$	$\frac{3}{1000*1000*601} = 5.0 \cdot 10^{-9}$	$\frac{1}{100*1000*301} = 3.3 \cdot 10^{-9}$	$\frac{0}{1000*1000*301} = 0$
Observations[b]	$\frac{5}{100*1000} = 5.0 \cdot 10^{-5}$	$\frac{3}{1000*1000} = 3.0 \cdot 10^{-6}$	$\frac{1}{100*1000} = 1.0 \cdot 10^{-5}$	$\frac{0}{1000*1000} = 0$

a: Share among all combinations of simulated observations times number of grid points for λ for which violation of HR occurs. Denominator is number of random samples $\times$ sample size $\times$ number of grid points for λ.

b: Share among all simulated observations for which a violation of HR occurs at least for one grid point of λ. Denominator is number of random samples $\times$ sample size.

Figure 1: Distribution of shares of inadmissible observations not in $\mathcal{N}_{0.5,n}$ (left panel) and distribution of $\widehat{\lambda}_{0.5}$ (right panel) for 100 (top panel) and 1000 observations (bottom panel), homoskedastic design

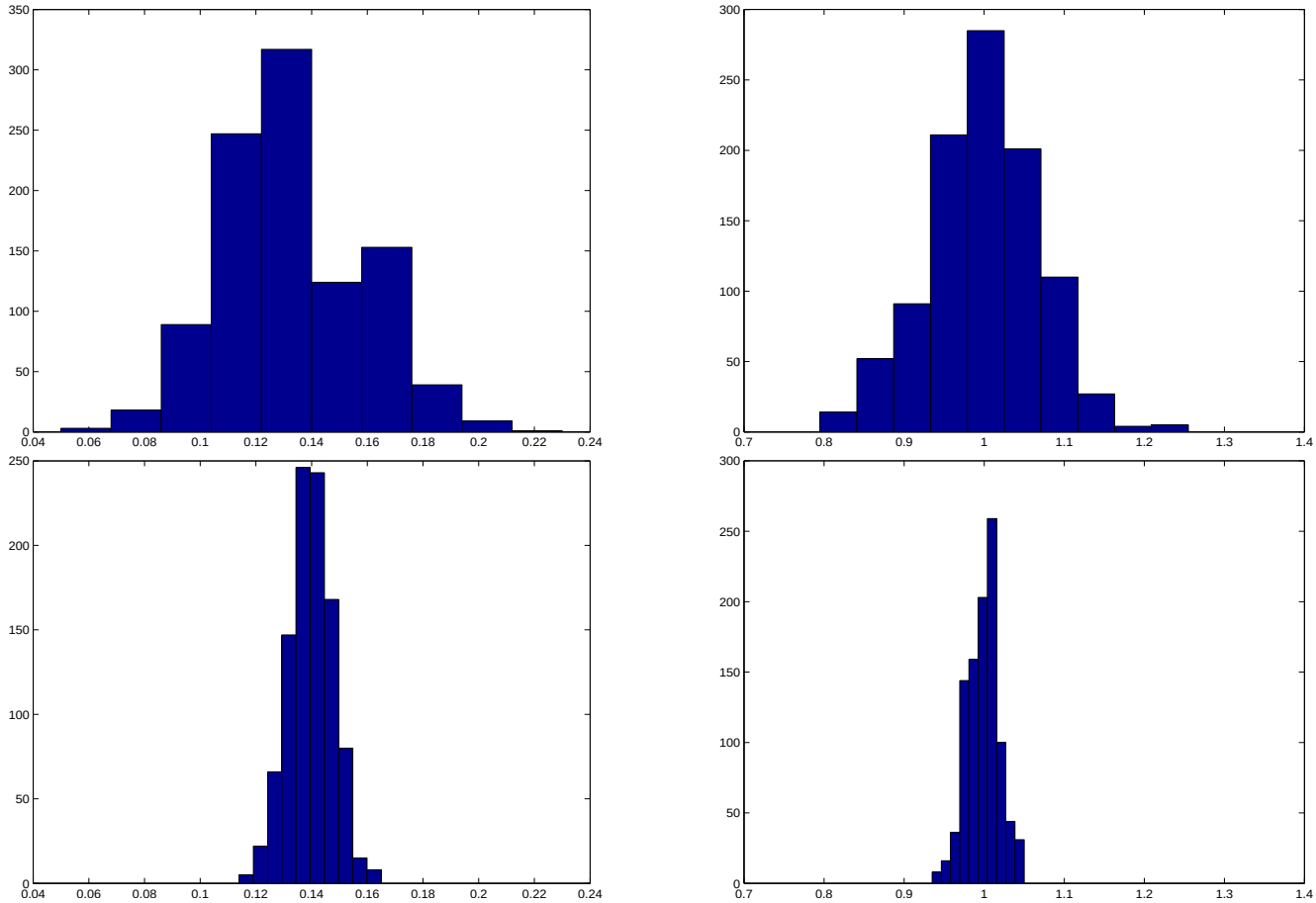
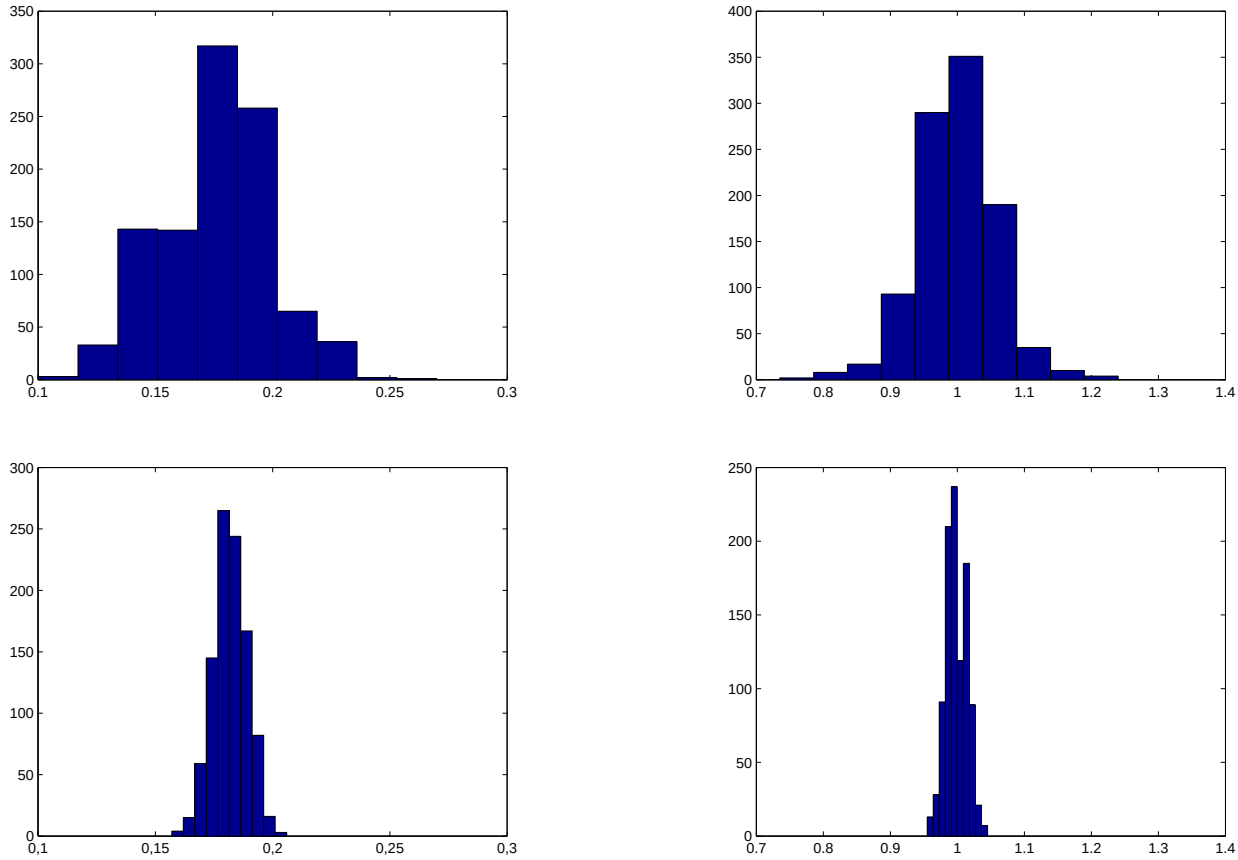


Figure 2: Distribution of shares of inadmissible observations not in $\mathcal{N}_{0.5,n}$ (left panel) and distribution of $\widehat{\lambda}_{0.5}$ (right panel) for 100 (top panel) and 1000 observations (bottom panel), heteroskedastic design



References

[1] Angrist, J., Chernozhukov, V., and Fernández–Val, I. (2006). Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure. *Econometrica* Vol. 74, 539–563.

[2] Bickel, P.J. and Doksum, K.A. (1981). An Analysis of Transformations Revisited. *Journal of the American Statistical Association*, Vol. 76, 296-311.

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

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

[5] 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.

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

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

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

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

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

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

[12] Koenker, R. (2005). Quantile Regression. *Econometric Society Monograph*, Cambridge University Press, New York.

[13] Lüdemann, E., Wilke R.A. and Zhang, X. (2006). Censored Quantile Regressions and the Length of Unemployment Periods in West Germany. *Empirical Economics*, Vol. 31, 1003–1024.

[14] 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.

[15] 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.

[16] 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.

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