

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

Download this ZEW Discussion Paper from our ftp server:

**<ftp://ftp.zew.de/pub/zew-docs/dp/dp0461.pdf>**

Die Discussion Papers dienen einer möglichst schnellen Verbreitung von neueren Forschungsarbeiten des ZEW. Die Beiträge liegen in alleiniger Verantwortung der Autoren und stellen nicht notwendigerweise die Meinung des ZEW dar.

---

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.

## 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<sup>†</sup>

Ralf A. Wilke<sup>‡</sup>

Xuan Zhang<sup>§</sup>

September 2004

## Abstract

The Box-Cox quantile regression model using the two stage method introduced by Chamberlain (1994) and Buchinsky (1995) provides an attractive extension of linear quantile regression techniques. 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. The modified estimator is still  $\sqrt{n}$ -consistent and its asymptotic distribution can easily be derived. A simulation study confirms that the modified estimator works well.

**Keywords:** Box-Cox quantile regression, iterative estimator

**JEL:** C13, C14

---

\*Financial support of the German Research Foundation (DFG) through the project “Microeconomic modelling of unemployment durations under consideration of the macroeconomic situation” is gratefully acknowledged.

<sup>†</sup>Corresponding author: Goethe-University Frankfurt, ZEW Mannheim, IZA Bonn and IFS London, Department of Economics; Address: Goethe-University Frankfurt, PO Box 11 19 32 (PF 247), 60054 Frankfurt am Main, E-mail: fitzenberger@wiwi.uni-frankfurt.de

<sup>‡</sup>ZEW Mannheim, E-mail: wilke@zew.de

<sup>§</sup>ZEW Mannheim, Mannheim University, E-mail: xuanz98@rumms.uni-mannheim.de

# 1 Introduction

This note considers a general difficulty with the two step estimation approach to Box-Cox quantile regressions as introduced by Chamberlain (1994) and Buchinsky (1995).<sup>1</sup> In the second step, the objective function is not defined in some situations 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. Simulations show that the modification works well in finite samples. Furthermore, consistency and asymptotic normality of the estimator are not affected by the modification.

## 2 Model

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

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

where  $y > 0$ , the observable regressors  $x \in \mathbb{R}^K$ , the unknown parameters  $\beta \in \mathcal{B} \subset \mathbb{R}^K$ , the quantile  $\theta \in (0, 1)$  and the error  $\epsilon_\theta$  s.t.  $F_{\epsilon_\theta}(0|x) = \theta$ . We consider the strictly positively monotonic 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  $g(\cdot)$  where  $\lambda \in \mathcal{R}$  and without loss of generality  $\mathcal{R} = [\underline{\lambda}, \bar{\lambda}]$  is a finite closed interval. This transformation is quite attractive since it preserves the ordering of the observations because of the equivariance property of quantiles. Thus, we obtain a linear model for

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

and equation (1) becomes

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

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

---

<sup>1</sup>The Box-Cox quantile regression model was introduced by Powell (1991).

### 3 Estimation Problem

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

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

for observations  $i = 1, \dots, n$  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. However, these studies do not explicitly address the issue of actually computing the estimates. In principle, this could be estimated directly using the Koenker and Park (1996) algorithm for nonlinear quantile regressions.

Chamberlain (1994) and Buchinsky (1995) suggest the following attractive simplification in form of a two step procedure based on the equivariance property of quantiles with respect to a strictly positively monotonic transformation

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

$$\hat{\beta}_{\theta}(\lambda) = \operatorname{argmin}_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \rho_{\theta}(y_{\lambda i} - x'_i \beta) \quad (2)$$

2. estimate  $\lambda$  by solving

$$\min_{\lambda \in \mathcal{R}} \frac{1}{n} \sum_{i=1}^n \rho_{\theta}(y_i - (\lambda x'_i \hat{\beta}_{\theta}(\lambda) + 1)^{1/\lambda}) \quad (3)$$

Note that the objective in (2) cannot be used to estimate both  $\beta_{\theta}$  and  $\lambda$  (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 not mentioned in any of the cited references above. For every  $\lambda$ , it is not guaranteed that for all observations  $i = 1, \dots, n$  the inverse Box-Cox transformation  $\lambda x'_i \hat{\beta}_{\theta}(\lambda) + 1$  is strictly positive. However, this is necessary to implement the second step of Buchinsky’s procedure.<sup>2</sup> It is natural to omit the observations for which this condition is not satisfied. But

---

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

this raises a number of problems. First, the set of observations omitted changes when going through an iterative procedure to find the optimal  $\lambda$ . 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

$$\lambda x'_i \hat{\beta}_\theta(\lambda) + 1 > 0 \quad (4)$$

for all  $i = 1, \dots, n$ . This clearly depends on the first stage estimates and the specific value of  $\lambda$ . 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.<sup>3</sup> Therefore, in finite samples the inequality (4) 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  $i = 1, \dots, n$  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. Instead of problem (3), we now solve in the second step

$$\min_{\lambda \in \mathcal{R}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{i \in \mathcal{N}_{\theta,n}} \cdot \rho_\theta(y_i - \tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)]), \quad (5)$$

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

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

---

<sup>3</sup>For some  $\lambda$  during the iteration process, step 1 results in the linear quantile projection of  $y_\lambda$  on  $x_i$  as defined in the appendix. Even for the true projection, i.e. the true associated parameter values, it is not guaranteed that condition (4) is satisfied for all observations. The condition is only satisfied for sure for the true  $\lambda$  and under the assumption that the model is not misspecified.

Note it does not matter what value of  $c$  is chosen because the indicator function in equation (5) 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 $\mathcal{N}_{\theta,n}$ ?

To choose the set of admissible observations  $\mathcal{N}_{\theta,n}$  for the iteration process in  $\lambda \in \mathcal{R}$ , we use a very simple heuristic rule which works for sure 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  $\bar{\lambda}$  in  $\mathcal{R}$ , respectively, whether  $\tilde{g}_i[\lambda, \hat{\beta}_{\theta}(\lambda)]$  is well defined. The 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  $\lambda = \bar{\lambda}$ . This rule is based on the following result.<sup>4</sup>

**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  $\lambda \in \{\underline{\lambda}, \bar{\lambda}\}$ , then  $\lambda x'_i \hat{\beta}_{\theta}(\lambda) + 1 > 0$  for all  $\lambda \in [\underline{\lambda}, \bar{\lambda}]$  with probability one.*

Proposition 1 can be motivated as follows (the proof can be found in the appendix). 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 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  $\lambda$  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 simulation exercise, we found no case

---

<sup>4</sup>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). Simulation experience (simulation results are available upon request) indicates that our selection rule works for censored Box-Cox quantile regressions only up to an upper and lower bound of  $\lambda$ . These bounds seem to depend on the simulation design. Further research is necessary on this issue.



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

	$n = 100$		$n = 1.000$	
% of $i$ not in $\mathcal{N}_{0.5,n}$	6.3%	(0.02)	6.6%	(0.01)
$\hat{\beta}_1$	50.020	(0.52)	50.003	(0.15)
$\hat{\beta}_2$	1.001	(0.03)	1.000	(0.01)
$\hat{\beta}_3$	1.000	(0.25)	0.997	(0.08)
$\hat{\lambda}$	1.500	(0.01)	1.500	(0.00)

where applying the selection rule based on proposition 1 did not work perfectly during the search for estimating  $\lambda$ . In the following, we will argue why this is the case under most circumstances.

For the proof of Proposition 1 (see appendix) it is critical whether the following condition holds

$$\frac{\partial \Delta}{\partial \lambda} = \sum_{h=1}^K g_h \log(y_{(h)}) y_{(h)}^\lambda < 0 \quad (6)$$

for interpolated observations  $h = 1, \dots, K$  with  $\Delta = \sum_{h=1}^K g_h y_{(h)}^\lambda = 0$  and  $\sum_{h=1}^K g_h = 1$ .<sup>5</sup> Note that this condition 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  $h$ . It is natural that  $g_h$  is positive if  $y_{(h)}$  is small and  $g_h$  is negative if  $y_{(h)}$  is large. This is likely to hold “on average” and therefore condition (6) is likely to hold for practical designs.

As a theoretical alternative to avoid the above problems and to show the asymptotic validity of the estimator, one could choose  $\mathcal{N}_{\theta,n}$  as the set of observations  $i$  for which  $\lambda x'_i \hat{\beta}_\theta(\lambda) + 1 > 0$  for all  $\lambda \in [\underline{\lambda}, \bar{\lambda}]$ . However, this is not a rule which can be strictly applied in actual estimation.

---

<sup>5</sup>This weights represent a regressor vector, for which the fitted value lies at the critical boundary, by the coefficients in the linear combination of interpolated design points, see appendix for more details.

## 5 Simulations

This section presents finite sample results for the modified estimator. We use the following model:

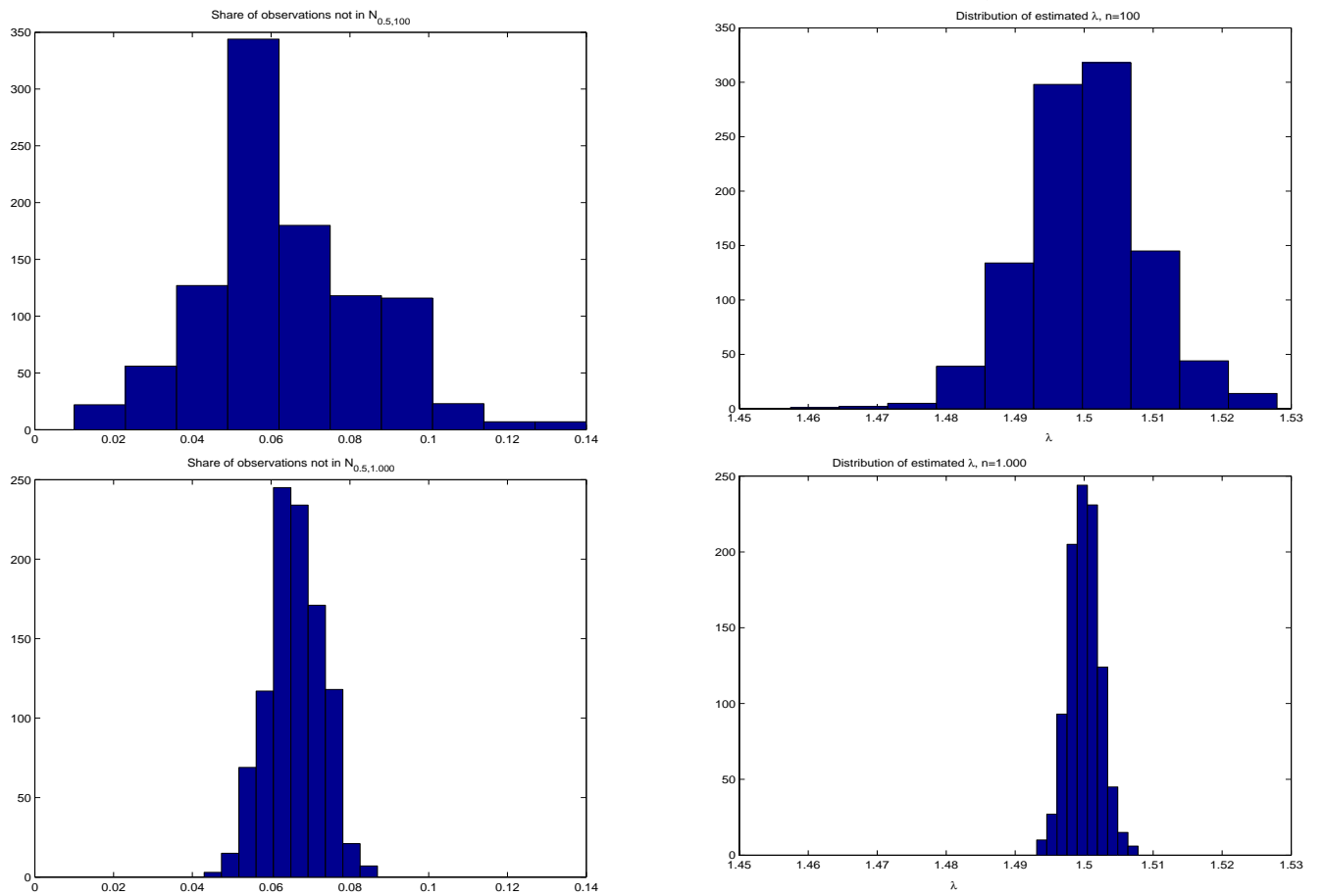
$$y_\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where  $x_1 \sim N(200, 110)$ ,  $x_2 \in \{0, 1\}$ ,  $Prob(x_2 = 0) = Prob(x_2 = 1) = 0.5$ ,  $\beta = (50, 1, 1)'$ ,  $\epsilon \sim N(0, 1)$  and  $\lambda = 1.5$  with  $\mathcal{R} = [-0.5, 2.5]$ . We draw 1.000 independent random samples from this model. Even though in principle, observations based on this model might violate the critical condition (4), in our simulations all observations drawn satisfy this condition for the chosen parameter values. Estimates for  $\beta$  are obtained 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 6.5 which uses the Nelder-Mead simplex method for non differentiable objective functions. Table 1 presents the results for two experiments based on 1.000 replications with sample sizes  $n = 100$  and  $n = 1.000$ .

The results show that the numerical problem addressed in this note is by no means negligible. On average, between 6 and 7 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 very close to the true parameter values and the estimator appears to be unbiased even in small samples.

Figure 1 depicts the empirical distributions of the share of observations not falling in  $\mathcal{N}_{0.5, n}$  and of the estimates of  $\lambda$ . It turns out that in some samples more than 10 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 6 percent when the sample size is 1.000. The distribution of  $\hat{\lambda}$  is nicely concentrated around the true parameter  $\lambda = 1.5$  and as to be expected the variance decreases with the sample size.

Figure 1: Distribution of shares of inadmissible observations not in  $\mathcal{N}_{0.5,n}$  and distribution of  $\hat{\lambda}$



## Appendix

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

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

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

Starting at some  $\lambda$ , for  $y$  being small, i.e.  $y < \exp(1/\lambda)$ , reducing  $\lambda$  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  $x'_{(h)}\hat{\beta}_\theta(\lambda) = y_{(h),\lambda}$ <sup>6</sup> for  $h = 1, \dots, K$  individual observations with linearly independent  $x_{(h)}$  and  $i(h) \in \{1, \dots, n\}$  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  $\lambda$  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  $\lambda$ , the set of interpolated data points  $i(h), h = 1, \dots, K$  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  $\lambda \leq \bar{\lambda}$  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  $\Delta \equiv \sum_{h=1}^K g_h y_{(h)}^\lambda = 0$ , where

---

<sup>6</sup>With  $y_{(h),\lambda} = (y_{(h)}^\lambda - 1)/\lambda$  for  $\lambda \neq 0$  and  $y_{(h),\lambda} = \log(y_{(h)})$  for  $\lambda = 0$ .

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  $g_1 = y_{(2)}^\lambda / (y_{(2)}^\lambda - y_{(1)}^\lambda)$  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.

6. After more than an infinitesimal change of  $\lambda$  it may occur that the set of interpolating observations changes. For the specific lambda, when this occurs, the linear quantile regression will interpolate another data point  $l = 1, \dots, n$  with  $x'_l \hat{\beta}_\theta(\lambda) = y_{l,\lambda}$  in addition to  $i(h), h = 1, \dots, K$ , again see Koenker and D'Orey (1987, p. 385) for a similar argument. If  $\lambda$  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  $\lambda$  moves beyond the critical value, the same argument applies as in the previous step. Thus, also for such critical values of  $\lambda$ , 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 (6) is violated in this case, because of the large positive weight  $g_2$  for the observation with the highest value of the dependent variable  $y_5 = 2.0$ .

### 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  $\theta$ ,  $\lambda_0$  and  $\beta_{0,\theta}$  are the true parameter values.

1. For a given  $\lambda$  and under standard regularity conditions, the linear quantile regression estimator  $\hat{\beta}_\theta(\lambda)$  is  $\sqrt{n}$ -consistent converging to the coefficients of the linear quantile projection (this terminology is analogous to the linear projection for least squares, see Wooldridge, 2002, chapters 2 and 3) in the population

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

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

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

as a population moment condition, where  $I(\cdot)$  is the indicator function. It is clear that for the true  $\lambda_0$ , we obtain  $\beta_\theta(\lambda_0) = \beta_{0,\theta}$ . Even though, the linear quantile projection as an approximation does not satisfy  $Quant(y_\lambda|x) = x'\beta_\theta(\lambda)$  for general  $\lambda$  (analogous to the least squares case, see Wooldridge, 2002) 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).

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

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

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

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

$\sqrt{n}$ -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  $x_i$ . Alternatively, in cases, when our heuristic rule does not work, one can define

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

However, this “rule” can not be easily checked in practical applications.

3. The objective function for the second step of the modified estimator in equation (5) can be written as

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

By the Cauchy–Schwarz inequality, the absolute value of the second term can be bounded from above by

$$\left| \frac{1}{n} \sum_{i=1}^n (\mathbb{1}_{i \in \mathcal{N}_{\theta,n}} - I_i)^2 \right|^{1/2} \cdot \left| \frac{1}{n} \sum_{i=1}^n \rho_\theta(y_i - \tilde{g}_i[\lambda, \hat{\beta}_\theta(\lambda)])^2 \right|^{1/2} .$$

The first term in this product converges to zero in probability uniformly at a rate faster than  $\sqrt{n}$  and the second term is uniformly bounded from above in probability. Thus, the asymptotic properties of the modified estimator can simply be derived as resulting from minimizing the first term in equation (7), i.e. the estimation error in  $\mathbb{I}_{i \in \mathcal{N}_{\theta, n}}$  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  $\bar{\lambda}$  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 consideration 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

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

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

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  $\lambda$ . Since  $I_i$  is nondecreasing for all observations when a finer set  $\mathcal{R}$  is used (i.e.  $\bar{\lambda}$  decreases or  $\underline{\lambda}$  increases) still containing  $\lambda_0$ , the asymptotic variance decreases (in the usual matrix sense), i.e. the modified estimator becomes asymptotically more efficient.



## References

- [1] Box G. and Cox D. (1994). An Analysis of Transformation. *Journal of the Royal Statistical Society B* 26, 211–252.
- [2] Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963-1987. *Journal of Econometrics* Vol.65, 109–154.
- [3] 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.
- [4] 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.
- [5] Fitzenberger, B. (1998) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. *Journal of Econometrics*, 82, 235–287.
- [6] Hunter, D. (2002) MATLAB CODE for (Non-)Linear Quantile Regressions. <http://www.stat.psu.edu/~dhunter/qrmatlab/>.
- [7] Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica* Vol. 46, 33–50.
- [8] Koenker, R. and D’Orey (1987). Algorithm AS 229. Computing Regression Quantiles. *Statistical Algorithms, Royal Statistical Society* 383–393.
- [9] Koenker, R. and Park, B. (1996). An Interior Fixed Point Algorithm for Quantile Regressions. *Journal of Econometrics* Vol. 71, 265–283.
- [10] 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.
- [11] 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.
- [12] Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, Massachusetts.