E C O N O M I C S   B U L L E T I N

An algorithm for censored quantile regressions

Thanasis Stengos
University of Guelph

Dianqin Wang
University of Guelph

Abstract

In this paper, we present an algorithm for Censored Quantile Regression (CQR) estimation problems. Our method permits CQR estimation problems to be solved more efficiently and reliably than was hitherto possible. It guarantees to find a high quality estimator in $O(k \times n^2)$ operations with k regressors and n observations, which is much less than the existing algorithms for CQR problems.

The first author would like to acknowledge financial support from the SSHRC of Canada.
Citation: Stengos, Thanasis and Dianqin Wang, (2007) "An algorithm for censored quantile regressions." *Economics Bulletin*, Vol. 3, No. 1 pp. 1-9
Submitted: October 5, 2006. **Accepted:** January 9, 2007.
URL: http://economicsbulletin.vanderbilt.edu/2007/volume3/EB-06C20071A.pdf

1 Introduction

It is well known that censoring poses serious problems for regression models, see Goldberger (1983) and Honoré (1992), as the resulting estimators in general may not be consistent. However, in the context of quantile regression, censoring can be handled under very weak distributional assumptions. For the Censored Quantile Regression (CQR) model, Powell (1986) obtained estimator consistency and asymptotic normality. However, although the CQR model has many appealing theoretical features, its application has been hampered by the computational complexity of standard methods in obtaining CQR estimators.

The CQR estimation problem is to minimize a non-convex and piecewise linear distance function which may have multiple local optima. The optimal estimators occur at non-smooth points. The feature of the distance function presents a challenge to the standard optimization techniques which require the objective function to be convex and/or differentiable. Several algorithms have been developed by modifying standard optimization algorithms, yet their performance has not been satisfactory. An optimal estimator for a CQR problem is difficult to obtain by modifying other existing algorithms, especially for large-scale problems in practice, see Fitzenberger (1997).

Since the CQR model was introduced by Powell (1986), several algorithms have been presented in the literature to deal with this problem. A detailed survey of the existing algorithms and their performances may be found in Fitzenberger (1997) and Buchinsky (1994). To name a few of the algorithms, they include a modified reduced-gradient algorithm by Womersley (1986), an adaptation of the standard Barrodale-Roberts algorithm, see Barrodale and Roberts, (1973), for Quantile Regressions to the case of Censored Quantile Regressions by Fitzenberger (1997), an interior point approach by Koenker and Park (1996), and an iterative linear programming algorithm by Buchinsky (1994). Emulation algorithms (EA) presented in Pinkse (1993) and Fitzenberger (1997) compute the exact CQR estimator by checking every critical point. Some hybrid algorithms were proposed by Fitzenberger and Winker (1999). However, these algorithms have run into difficulties in solving CQR estimation problems, since they exhibit a high degree of complexity in their implementation. Typically these algorithms achieve convergence to local minima, whereas obtaining a global minimum requires a heavy computational load, something that renders their use in solving real economic problems impractical. Among all these methods, only emulation algorithms could be used to guarantee convergence.

The purpose of this paper is to offer an alternative algorithm that simplifies the computations of the CQR estimation problem. The method presented in this paper uses a systematic procedure to improve the reliability of the estimator, by using approximately $O(k \times n^2)$ operations for a CQR problem with n observations and k regressors. The algorithm is simple to implement. The next section of the paper presents the CQR estimation problem. We then proceed to present the algorithm and then offer some simulation results regarding its relative performance when compared with the EA alternative.

2 The CQR estimation problem

The CQR estimation problem is to minimize a nonconvex and piecewise linear distance function

$$\min \sum_{i=1}^{n} \left(\theta * I(d_i > 0) + (1-\theta) * I(d_i < 0)\right) |d_i| \tag{1}$$

where $d_i = y_i - \max\left(x_i\beta, cy_i\right)$, and $I(x)$ is an indication function, such that

$$I(x) = \begin{cases} 1 & \text{if x is true} \\ 0 & \text{otherwise.} \end{cases}$$

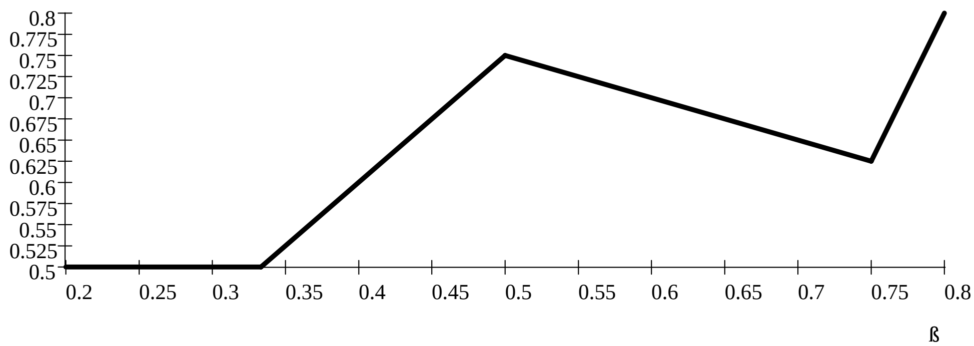
Also $\theta \in (0,1)$ is the quantile and y_i is censored from below at cy_i. The vector β is $k \times 1$ dimensional. Other types of CQR can be easily transformed into the form of Equation (1). We define $\Phi(\beta)$ as

$$\Phi(\beta) = \sum_{i=1}^{n} \left(\theta * I(d_i > 0) + (1-\theta) * I(d_i < 0)\right) |d_i|.$$

The distance function $\Phi(\beta)$ is piecewise linear, non-convex, and it has local minima, and may not have a unique solution as illustrated in the following example. Consider the case where $\theta = 0.5$, and

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} cy_1 \\ cy_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

The distance function is depicted below



The distance function $\Phi(\beta)$

The above figure shows that the distance function $\Phi(\beta)$ has a local minimum $\frac{5}{8}$ at $\beta = \frac{3}{4}$. It is flat when $\beta \leq \frac{1}{3}$, and stays at the minimum of zero. From the above figure, we see that the CQR distance function is not smooth at the critical points $\beta = \frac{y_i}{x_i}$, and $\frac{cy_i}{x_i}$, for $i = 1, \cdots, n$. Otherwise, $\Phi(\beta)$ is linear. The minimum of $\Phi(\beta)$ must occur at one of these points.

3 A new algorithm for CQR estimation

As mentioned above CQR estimation deals with the minimization of a non-convex function. Non-convex functions have been proven very difficult to minimize or maximize. However, from the example above we notice that CQR estimation involves a particular type of non-convexity. Given a direction and a starting point, the objective function for CQR estimation is a piecewise linear function. It consists of at most $2n + 1$ line-segments, or there are at most $2n$ critical points (knots). This unique property guarantees that a local minimum could occur only in $2n$ critical points or knots.

Based on the above observations, our method is designed to find the minimum point by improving the value of the objective function at a given critical point with a smaller objective function value at another critical point. In other words, the method offers a way of moving from a given critical point to the next critical point with a smaller objective function value. This process continues until no critical points with smaller objective function value can be found.

More specifically, our algorithm starts with an initial critical point, and then we choose a direction. Next, we check the $2n$ critical points for a given direction. We choose the critical point with the smallest objective function value to be the new starting point. Then we choose another direction, and we repeat the process, until no new starting point can be found.

4 Implementation of the algorithm

Since the minimum only occurs at critical points, this motivates us to design an algorithm to improve the estimated β through critical points, such that $\Phi(\beta_{l+1})$ is less than $\Phi(\beta_l)$. In other words, given a starting value β_l, we try to find β_{l+1}, such that $\Phi(\beta_{l+1})$ is less than $\Phi(\beta_l)$. We proceed as follows. We first define $\overline{\beta}_{l+1} = \beta_l + d_j * e_j$ for $j \in (1, k)$, where e_j is a $k-$dimensional vector with its j^{th} element being one, the rest zero and d_j is a step size. We next have $x_i\overline{\beta}_{l+1} = x_i\left(\beta_l + d_j * e_j\right)$. The potential critical points are those such that $x_i\overline{\beta}_{l+1} = y_i$ for each $i \in (1, n)$. That is $x_i\left(\beta_l + d_j * e_j\right) = y_i$, or $x_{i,j}d_j = y_i - x_i\beta_l$. If $x_{i,j}$ is not zero, then we have $d_j = \frac{y_i - x_i\beta_l}{x_{i,j}}$. We proceed by computing $\Phi(\overline{\beta}_{l+1})$. We define $\beta_{l+1} = \overline{\beta}^0_{l+1}$ to be that value that minimizes $\left\{\Phi(\overline{\beta}_{l+1})\right\}$ for all $i \in (1, n)$ and $j \in (1, k)$. If $\Phi(\beta_{l+1})$ turns out to be less that $\Phi(\beta_l)$, then a new β_{l+1} is found. We repeat the above process by replacing β_l by β_{l+1}, until a minimum point is reached. The process will end in finite steps since there is only a finite number of critical points. To find an improvement of β_l involves two loops, which need about $O(k \times n^2)$ operations. The above critical point search algorithm is outlined below[1].

1. for a given β_l.
2. for $i = 1, \cdots, n$.
   (a) for $j = 1, \cdots, k$.
      i. if $x_{i,j}$ is not zero, then
         A. $d_j = \frac{y_i - x_i\beta_l}{x_{i,j}}$.
         B. compute $\Phi(\beta_l + d_j \times e_j)$.
         C. if $\Phi(\beta_l + d_j \times e_j) < \Phi(\beta_l)$, then $\beta_{l+1} = \beta_l + d_j \times e_j$.
      ii. end of "if" loop.
   (b) end of loop for j
3. end of loop for i.

[1] The implementation of the above method is written in Gauss code and is available from the authors on request.

4. if β_{l+1} is found during the loop for i, then
   (a) replace β_l by β_{l+1}.
   (b) repeat steps 2 to 3.
5. otherwise the program terminates, a minimum is found.

5 Simulation Study

The timing of finding estimates of CQR problems is reported in Table 1. The method presented in this paper is coded in Gauss, while the computer used to perform the simulations is a Pentium 4 with 1.5 GHZ CPU. We also give the estimated times from using the EA of Pinkse (1993) and Fitzenberger (1997), which is the only one among the standard methods to guarantee convergence. We know that the EA methods checks $\begin{pmatrix} n \\ k \end{pmatrix}$ critical points. The number of critical points checked by our method are reported in Table 2. Based on the critical points checked, we can estimate the time that it would take the EA method to solve the same CQR problems. To compare our method with the EA method, we compute the time ratio of the EA and our method. We randomly generated 2000×20 uniform numbers in the interval $(-2.5, 2.5)$. The CQR problems are solved with $(n = 500, 1000, 1500, 2000)$ and $(k = 2, 5, 10, 15, 20)$. The variable y is censored below a number which is randomly generated between $(-1, 1)$. The dependent variable y is generated as follows:

$$y = \max(x_1 + \cdots + x_k + \varepsilon, cy).$$

Here, the distribution of ε is the Student's t-distribution with 5 degrees of freedom. The results are reported in Table 2. The total time unit is seconds.

From Table 2, we see that the timing ratio increases dramatically as k grows from 2 to 20 for given n. It is a very difficult task to find the optimal points by using the EA method for k over 10 using standard computers. On the other hand our proposed method can solve large scale CQR problems that appear in practice using only $O(k \times n^2)$ operations, which is a fraction of what is required by other existing methods.

Table 1: The timing results for the method presented in this paper.

		Our Method
n	k	Time (Seconds)
500	2	4.98E+01
500	5	7.13E+02
500	10	1.67E+04
500	15	1.43E+04
500	20	3.18E+04
1000	2	5.19E+01
1000	5	1.68E+04
1000	10	2.55E+04
1000	15	1.38E+07
1000	20	3.00E+08
1500	2	2.63E+03
1500	5	2.05E+04
1500	10	3.46E+08
1500	15	7.81E+08
1500	20	1.13E+09
2000	2	3.28E+03
2000	5	1.64E+04
2000	10	6.66E+08
2000	15	2.67E+09
2000	20	1.83E+09

Table 2: The comparison of timings between our method and EA method. NCPC denotes the number of critical points checked.

n	k	Our Method NCPC	EA Method NCPC	Time ratio of EA over ours
500	2	105000	124750	1.19E+00
500	5	125000	2.55245E+11	2.04E+06
500	10	780000	2.45811E+20	3.15E+14
500	15	3427500	1.88779E+28	5.51E+21
500	20	3120000	2.6672E+35	8.55E+28
1000	2	58000	499500	8.61E+00
1000	5	375000	8.25029E+12	2.20E+07
1000	10	2330000	2.6341E+23	1.13E+17
1000	15	5715000	6.88141E+32	1.20E+26
1000	20	19860000	3.39483E+41	1.71E+34
1500	2	78000	1124250	1.44E+01
1500	5	412500	6.28604E+13	1.52E+08
1500	10	5730000	1.54203E+25	2.69E+18
1500	15	9675000	3.12155E+35	3.23E+28
1500	20	12420000	1.20351E+45	9.69E+37
2000	2	112000	1999000	1.78E+01
2000	5	800000	2.65336E+14	3.32E+08
2000	10	6280000	2.75899E+26	4.39E+19
2000	15	10590000	2.37736E+37	2.24E+30
2000	20	12560000	3.91816E+47	3.12E+40

References

[1] Barrodale, I. and F. Roberts, (1973), "An Improved Algorithm for Discrete l_1 Linear Approximation," *SIAM Journal of Numerical Analysis* 10, 839-848.

[2] Buchinsky, M., (1994), "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica* 62, 405-458.

[3] Fitzenberger, B., (1997), "A Guide to Censored Quantile Regressions," in *Handbook of Statistics, Robust Inference* by G.S. Maddala and C.R. Rao eds., North-Holland, Amsterdam.

[4] Fitzenberger, B. and P. Winker, (1999), "Improving the Computation of Censored Quantile Regressions," Universität Mannheim, Institute für VWL und Statistik, Band Nr. 568-99.

[5] Goldberger, A.R., (1983), "Abnormal Selection Bias," in *Studies in Econometrics, Time Series, and Multivariate Statistics* by S. Karlin, ed., Academic Press, New York.

[6] Honoré, B. E., (1992), "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60, 553-565.

[7] Koenker, R. and B.J. Park, (1996), "An Interior Point Algorithm for Nonlinear Quantile Regression," *Journal of Econometrics* 71, 265-283.

[8] Pinkse, C.A.P., (1993), "On the Computation of Semiparametric Estimates in Limited Dependent Variable Models," *Journal of Econometrics* 58, 185-205.

[9] Powell, J., (1986), "Censored Regression Quantiles," *Journal of Econometrics* 32, 143-155.

[10] Womersley, R.S., (1986), "Censored Discrete Linear l_1 Approximation," *SIAM Journal of Scientific and Statistical Computing* 7, 105-122.