ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 656 July 2008

Quantiles for Fractions and Other Mixed Data

José A. F. Machado and J. M. C. Santos Silva

Quantiles for Fractions and Other Mixed Data*

José A. F. Machado
Faculdade de Economia, Universidade NOVA de Lisboa

J. M. C. Santos Silva
Department of Economics, University of Essex and CEMAPRE

July 26, 2008

Abstract

This paper studies the estimation of quantile regression for fractional data, focusing on the case where there are mass-points at zero or/and one. More generally, we propose a simple strategy for the estimation of the conditional quantiles of data from mixed distributions, which combines standard results on the estimation of censored and Box-Cox quantile regressions. The implementation of the proposed method is illustrated using a well-known dataset.

JEL classification code: C13; C25; C51.

Key words: Censored quantile regression; Mass-points; Mixed distributions.

*Machado is consultant for the Research Department of Banco de Portugal. We thank Daniel Dias and Paulo Parente for helpful suggestions and comments. The usual disclaimer applies. The authors also gratefully acknowledge the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER.

Address for correspondence: João Santos Silva, Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Tel: +44 (0)1206872769. E-mail: jmcss@essex.ac.uk.

1. INTRODUCTION

Empirical researchers are often faced with the need to model fractional data. Modelling this sort of data poses particular problems, which have sometimes been dealt with in a unsatisfactory manner. In a landmark paper, Papke and Wooldridge (1996) have shown that the generalized linear models framework provides a simple and effective way of modelling the conditional expectation of fractional data.

In many practical situations, however, the knowledge of the conditional expectation may not be enough. For example, if the researcher needs to construct a confidence interval for the value of the variate of interest, conditional on a given value of the covariates, knowledge of the conditional expectation is of little use because the textbook assumptions of normality and homoskedasticity do not hold. On the other hand, conditional quantiles provide a direct way of constructing this sort of confidence intervals. More generally, conditional quantiles of fractional data are interesting because, due to the bounded nature of the data, the features of the conditional distribution of interest will often depend on the regressors in a complex way.

The specification and estimation of conditional quantile functions for fractional data raises some interesting problems and the estimation strategy depends on the specific nature of the distribution being considered. In the simplest situation, the variate of interest has a continuous distribution in the $[0;1]$ interval. However, it is often the case that there is an "inflation" of zeros and/or ones, and therefore the distribution is mixed.[1] In this paper we look at the estimation of quantile regression for fractional data, focusing particular attention on the case of mixed distributions.

The results we obtain for fractional data with mass-points can easily be extended to other types of mixed data, like the non-negative data with a mass-point at zero, which are often found in health economics (Duan, Manning, Morris and Newhouse, 1983), trade

[1] A third situation is possible. If the fractional variate of interest is defined as the ratio of two integers, it has a discrete distribution. In this case the results of Machado and Santos Silva (2005) can be used to model the numerator of the ratio, conditional on the denominator.

(Santos Silva and Tenreyro, 2006), finance (La Porta, López-de-Silanes, and Zamarripa, 2003), and in many other areas.

The reminder of the paper is organized as follows. Section 2 details our approach to the estimation of quantile regression for fractional data. Section 3 gives details on the proposed method for the estimation of conditional quantiles when the data has a mass-point at zero, and presents the essential asymptotic results. Section 4 illustrates the application of the main results. Finally, section 5 contains some concluding remarks and discusses the extension of the main results to other settings.

2. QUANTILE REGRESSION FOR FRACTIONAL DATA

Due to the equivariance property of the quantiles, estimation of quantile regression functions for continuous fractional data is relatively simple. In particular, let y be the fractional variate of interest and assume that, for any $\theta \in (0,1)$, the researcher specifies the following parametric model

$$Q_y(\theta|x) = \Lambda(x'\beta),$$

where $\Lambda(x'\beta)$ is a function bounded between 0 and 1.[2] Then, β can be estimated by performing the usual linear quantile regression of $\Lambda^{-1}(y)$ on x. For example, when $\Lambda(x'\beta)$ is the logit, β can be estimated by performing a linear quantile regression of the log-odds ratio $\ln\left(\frac{y}{1-y}\right)$ on x. The function $\Lambda^{-1}(y)$ will not be defined for the (rare) observations in which y is zero or one. However, the properties of quantile functions imply that for the cases in which $y = 0$ or $y = 1$, the values of y can be nudged away from the boundaries without affecting the results.

When there are mass-points at zero or one, estimation is complicated by the fact that the quantiles are not necessarily smooth functions of the regressors. For expository purposes, we will consider only the case of a mass-point at zero, but handling a mass-point at one (or both mass-points) is similar.

[2]Naturally, the parameters of $Q_y(\theta|x)$ vary with θ, but that dependence is not made explicit to simplify the notation.

When there is a mass-point at zero, there are conditional quantiles that become identically zero for some values of the covariates. Specifically, the conditional quantiles have the form

$$Q_y(\theta|x) = I(\theta > \Pr(y=0|x)) Q_{y>0}\left(\frac{\theta - \Pr(y=0|x)}{1-\Pr(y=0|x)}\middle| x\right).$$

Therefore, in general, the conditional quantiles are not smooth functions of the regressors. However, because the dependent variable has support on $[0;1]$, the quantiles have to be continuous functions of the regressors. This suggests that, as in Powell (1984, 1986), the quantiles will have the form

$$Q_y(\theta|x) = \max\{0, g(x'\beta)\}, \tag{1}$$

where $g(x'\beta)$ is a function such that $g(z) < 1$, $\forall z$.

The choice of $g(x'\beta)$ is an empirical matter. In the spirit of Papke and Wooldridge (1996), we suggest the following specification

$$g(x'\beta) = (1+\gamma)\Lambda(x'\beta) - \gamma, \tag{2}$$

where $\Lambda(x'\beta)$ is a CDF and $\gamma > 0$ is an unknown shape parameter.[3]

The specification of $g(x'\beta)$ in (2) can be quite flexible. For example, if $\Lambda(x'\beta)$ is the CDF of a symmetric distribution, the quantiles will be s-shaped for $\gamma < 1$ and concave otherwise. Naturally, identification depends on the curvature of $\Lambda(x'\beta)$. If the data is such that $g(x'\beta)$ is essentially linear, identification will be difficult. Therefore, in applications, difficulty in identifying γ suggests that simple censored linear model of the form $Q_y(\theta|x) = \max\{0, x'\beta\}$ will be adequate.

Using (1) and (2), estimation of the parameters of interest is relatively easy. Indeed, using the equivariance properties of the quantiles, it is easy to see that $Q_y(\theta|x) = \max\{0, (1+\gamma)\Lambda(x'\beta) - \gamma\}$ implies

$$Q_{\Lambda^{-1}\left(\frac{y+\gamma}{1+\gamma}\right)}(\theta|x) = \max\left\{\Lambda^{-1}\left(\frac{\gamma}{1+\gamma}\right), x'\beta\right\}.$$

[3]The parameter γ may also be specified as a function of x, but that avenue is not pursued here.

That is, conditional on the value of γ, β can be estimated by the linear censored quantile regression of $\Lambda^{-1}\left(\frac{y+\gamma}{1+\gamma}\right)$ on x, with the dependent variable censored from below at $\Lambda^{-1}\left(\frac{\gamma}{1+\gamma}\right)$. The next section discusses the joint estimation of γ and β.

In the leading case where $\Lambda(\cdot)$ is specified as the logit, $\Lambda^{-1}\left(\frac{y+\gamma}{1+\gamma}\right) = \ln\left(\frac{y+\gamma}{1-y}\right)$ and $\Lambda^{-1}\left(\frac{\gamma}{1+\gamma}\right) = \ln(\gamma)$. Therefore, for a given γ, β can be estimated by the linear censored quantile regression defined by

$$Q_{\ln\left(\frac{y+\gamma}{1-y}\right)}(\theta|x) = \max\{\ln(\gamma), x'\beta\}.$$

The model, defined by (1) and (2), for the case of a mass-point at zero can easily be modified to accommodate other situations. If the mass-point is at one, $Q_y(\theta|x)$ can be specified as

$$Q_y(\theta|x) = \min\{1, (1+\gamma)\Lambda(x'\beta)\}. \tag{3}$$

In the same spirit, in case there are mass-points at both zero and one, the following specification can be adopted

$$Q_y(\theta|x) = \max\{0, \min\{1, (1+\gamma+\xi)\Lambda(x'\beta) - \gamma\}\},$$

where $\xi > 0$ is a second shape parameter.

3. ESTIMATION

In this section we adapt Chamberlain's (1994) two-steps estimation strategy of the Box-Cox quantile model to our present setting, and discuss some details of the implementation of the proposed estimator.

3.1. Theory

The basic intuition for the population is as follows. The assumption that, for a given $\theta \in (0,1)$, there exist β_0 and γ_0 such that

$$Q_y(\theta|x) = \max\{0, (1+\gamma_0)\Lambda(x'\beta_0) - \gamma_0\}$$

implies, under standard conditions on the conditional density of y, that (β_0, γ_0) is the sole solution of

$$\min_{\beta,\gamma} E\left[\rho_\theta\left(y - \max\left\{0, (1+\gamma)\,\Lambda\left(x'\beta\right) - \gamma\right\}\right)\right],$$

with $\rho_\theta(z) = z\left[\theta 1(z \geq 0) + (1-\theta)1(z < 0)\right]$, (Koenker and Bassett, 1978, Powell, 1984, 1986). As noted before, the equivariance property of quantile functions implies that

$$Q_{\Lambda^{-1}\left(\frac{y+\gamma_0}{1+\gamma_0}\right)}(\theta|x) = \max\left\{\Lambda^{-1}\left(\frac{\gamma_0}{1+\gamma_0}\right), x'\beta_0\right\}$$

and, thus, the solution $\beta(\gamma)$ of the program

$$\min_{\beta} E\left[\rho_\theta\left(\Lambda^{-1}\left(\frac{y+\gamma}{1+\gamma}\right) - \max\left\{\Lambda^{-1}\left(\frac{\gamma}{1+\gamma}\right), x'\beta\right\}\right)\right]$$

is such that $\beta_0 = \beta(\gamma_0)$. Notice that, for any given γ, this program defines a standard linear censored quantile regression problem. Finally, γ_0 will be the solution of

$$\min_{\gamma} E\left[\rho_\theta\left(y - \max\left\{0, (1+\gamma)\,\Lambda\left(x'\beta\left(\gamma\right)\right) - \gamma\right\}\right)\right].$$

The proposed estimator will be the sample analogue of the procedure described above. In a first step, for fixed values of γ, $\hat{\beta}(\gamma)$ will be the estimator of β in a linear censored quantile regression of $t_i \equiv \Lambda^{-1}\left(\frac{y_i+\gamma}{1+\gamma}\right)$ on x_i with known censoring points $t_0 = \Lambda^{-1}\left(\frac{\gamma}{1+\gamma}\right)$, that is, it will solve

$$\min_{\beta}\ S_1(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\theta\left(t_i - \max\left\{t_0, x_i'\beta\right\}\right).$$

Then, a one dimensional search over γ to minimize

$$S_2(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \rho_\theta\left(y_i - \max\left\{0, (1+\gamma)\,\Lambda\left(x_i'\hat{\beta}\left(\gamma\right)\right) - \gamma\right\}\right)$$

will yield $\hat{\gamma}$ and thus $\hat{\beta} = \hat{\beta}(\hat{\gamma})$.

To simplify notation let

$$\begin{aligned} w(y, x, \beta, \gamma) &= 1\left(t_0 < x'\beta\right)\left[\theta - 1\left(t_i < x'\beta\right)\right] \\ &= 1\left(0 < g\left(x'\beta, \gamma\right)\right)\left[\theta - 1\left(y < g\left(x'\beta, \gamma\right)\right)\right] \end{aligned}$$

where we now make explicit that $g(\cdot)$ depends on $x'\beta$ and γ. Also put,

$$d(x,\beta,\gamma) = \begin{pmatrix} x \\ \partial g(x'\beta,\gamma)/\partial\beta \\ \partial g(x'\beta,\gamma)/\partial\gamma \end{pmatrix} = \begin{pmatrix} x \\ (1+\gamma)\lambda(x'\beta)x \\ \Lambda(x'\beta)-1 \end{pmatrix} = \begin{pmatrix} x \\ d_2(x,\beta,\gamma) \end{pmatrix}.$$

with $\lambda(x'\beta) = \frac{\partial\Lambda(z)}{\partial z}\Big|_{x'\beta}$. Finally, let

$$A(\gamma) = \begin{pmatrix} I_k & 0_{k\times k} & 0_k \\ 0_k' & \partial\beta(\gamma)/\partial\gamma' & 1 \end{pmatrix}.$$

The function $\beta(\gamma)$ is implicitly defined by $E[w(y,x,\beta(\gamma),\gamma)x] = 0$. If there is no bunching at censoring points (i.e., $g(x'\beta_0,\gamma_0) \neq 0$ with probability one) (see Powell, 1984, 1986, and Fitzenberger, 1997), and if the inverse matrix below exists, we have by the implicit function theorem

$$\begin{aligned} \beta(\gamma)/\partial\gamma &= -\left[E\left\{1(0 < g(x'\beta(\gamma),\gamma))\, f_y(g(x'\beta,\gamma))(1+\gamma)\lambda(x'\beta(\gamma))xx'\right\}\right]^{-1} \times \\ &\quad \left[E\left\{1(0 < g(x'\beta(\gamma),\gamma))\, f_y(g(x'\beta,\gamma))(\Lambda(x'\beta(\gamma))-1)x\right\}\right] \end{aligned}$$

with $f_y(g(x'\beta,\gamma))$ denoting the conditional density of y evaluated at the θth conditional quantile.

Under suitable regularity conditions (Powell, 1991, and Fitzenberger, 1997), the estimator is consistent and has the following linear representation

$$L(\beta_0,\gamma_0)\begin{pmatrix} \sqrt{n}\left(\hat\beta-\beta_0\right) \\ \sqrt{n}(\hat\gamma-\gamma_0) \end{pmatrix} = -A(\gamma_0)\frac{1}{\sqrt{n}}\sum_i w(y_i,x_i,\beta_0,\gamma_0)\, d(x_i,\beta_0,\gamma_0) + o_P(1),$$

where

$$L(\beta_0,\gamma_0) = A(\gamma_0)\, E\left[1(0 < g(x'\beta_0,\gamma_0))\, f_y(g(x'\beta_0,\gamma_0))\, d(x,\beta_0,\gamma_0)\, d_2(x,\beta_0,\gamma_0)'\right].$$

Under the conditions of a Central Limit Theorem, the left hand side has an asymptotic normal distribution with mean zero and covariance matrix

$$M(\beta_0,\gamma_0) = \theta(1-\theta)A(\gamma_0)\, E\left[1(0 < g(x'\beta_0,\gamma_0))\, d(x,\beta_0,\gamma_0)\, d(x,\beta_0,\gamma_0)'\right] A(\gamma_0)'.$$

Therefore, if $L(\beta_0, \gamma_0)$ is non-singular,

$$\begin{pmatrix} \sqrt{n}\left(\hat{\beta} - \beta_0\right) \\ \sqrt{n}\left(\hat{\gamma} - \gamma_0\right) \end{pmatrix} \rightarrow N(0, V),$$

with $V = L_0^{-1} M_0 \left(L_0^{-1}\right)'$.

The asymptotic covariance matrix may be estimated by standard plug-in procedures. As usual, the only critical issue is the estimation of the conditional density of the response variable. A possible solution is to use the kernel methods proposed by Powell (1984) and described, for instance, in Fitzenberger (1997).

Since, in practice, (1) and (2) only provide an approximation to the functional form of the conditional quantiles, misspecification robust estimators of the covariance matrix should be used (see Chamberlain, 1994, Kim and White, 2002, and Angrist, Chernozhukov and Fernandez-Val, 2004).

3.2. Implementation Issues

Estimation of β and γ using the algorithm described above requires repeated estimation of censored linear quantile regressions. Although other methods are available (see Powell, 1986, Fitzenberger, 1997, and Buchinsky and Hahn, 1998), the three-step algorithm for the estimation of censored linear quantile regression proposed by Chernozhukov and Hong (2002), hereinafter CH, is particularly well suited to this particular application.[4] Indeed, this algorithm is generally well-behaved, has a good performance even in moderate samples and is computationally very simple.

The first step of the algorithm proposed by CH is the estimation of a binary model, say a logit, for the probability that a given observation is not censored. The two next steps are linear quantile regressions on selected sub-samples. Let $\hat{p}_i$ be the estimated probability that the ith observation is not censored, obtained in the first-step, and let

[4]Strictly speaking, for the estimator produced by this algorithm to be as efficient as Powell's benchmark estimator of the censored quantile regression model, somewhat more restrictive conditions are needed. Specifically, it is necessary to ensure the validity of the initial step estimating the probability of non-censoring.

$\tilde{\beta}$ denote the estimator of β obtained from the second-step, that is, the linear quantile regression estimated with the first sub-sample.

The selection of the two sub-samples to use in the second and third steps is critical for the performance and properties of the estimator. Indeed, CH prove the consistency of the estimator obtained when the two sub-samples are selected as follows. For the case in which the data are censored from below at C_i, the first sub-sample is such that it contains the observations for which $\hat{p}_i > 1 - \theta + c$, where c is strictly between 0 and θ. As for the second sub-sample, CH defined it so that it contains the observations with $x_i'\tilde{\beta} > C_i + \delta_n$, where δ_n is such that $\delta_n\sqrt{n} \to \infty$ and $\delta_n \searrow 0$.

For the practical implementation of the estimator, CH propose that c can be set as the qth quantile of all $\hat{p}_i$ such that $\hat{p}_i > 1 - \theta$. In their simulations, CH used this rule with $q = 10\%$. As for δ_n, CH suggest that, in practice, it can be chosen like c, but discarding a smaller percentage of observations. Strictly speaking, this method of choosing δ_n does not lead to a consistent estimator because it does not ensure that $\delta_n \searrow 0$, at least if q is not defined as a function of n.

To address this issue, we propose the use of an adaptive cut-off point for the determination of the second sub-sample which takes into account the variance of $x_i'\tilde{\beta}$. The objective of the second cut-off point is to select the observations with $x_i'\beta < C_i$, which is equivalent to $x_i'\tilde{\beta} < C_i + x_i'(\tilde{\beta} - \beta)$. In the cut-off point suggested by CH, δ_n can be viewed as a guess for the maximum value of $x_i'(\tilde{\beta} - \beta)$ over the entire sample. Alternatively, we define the cut-off point as $C_i + \delta_i$, where δ_i is a function of $V(x_i'\tilde{\beta})$, the variance of $x_i'\tilde{\beta}$. In particular, we set $\delta_i = k_n V(x_i'\tilde{\beta})^{0.5}$, where k_n is such that $k_n/\sqrt{n} \to 0$ and $k_n \to \infty$. Following Leamer (1978),[5] we set $k_n = \sqrt{n\left(n^{\frac{1}{n}} - 1\right)}$.

The use of this cut-off point can be interpreted as selecting for the second sub-sample the observations for which the lower bound of a confidence interval for $x_i'\beta$ is above C_i, with the confidence level of the interval going to 1 as n goes to infinity. A similar approach can be used to determine the cut-off point used to define the first sub-sample, but we do not pursue that issue here.

[5] See also Tersvirta and Mellin (1986).

4. AN EMPIRICAL ILLUSTRATION

In this section, the dataset studied by Papke and Wooldridge (1996) is used to illustrate the application of the proposed estimator. This is a dataset with 4734 firm-level observations on employee participation rates in 401(k) pension plans. As explained by Papke and Wooldridge (1996), participation in 401(k) pension plans is voluntary and therefore the participation rate (PRATE) depends on the characteristics of the plan, especially on the rate at which firms match the employees contributions (MRATE). Other regressors available in this dataset include the firm total employment (EMP), age of the plan (AGE), and a dummy indicating whether the 401(k) plan is the sole plan offered by the employer (SOLE). Further details on the data, including descriptive statistics, can be found in Papke and Wooldridge (1996).

In this sample, PRATE is relatively high, and it is equal to 1 for over 40% of the observations.[6] This suggests that the higher quantiles of the distribution will be flat at 1 for most observations. Therefore, we expect the role of the covariates to be particularly important for the lower quantiles.

Given the characteristics of the data, in this particular example we model the quantiles of PRATE as in (3). That is, we specify

$$Q_{\text{PRATE}}(\theta|x) = \min\{1, (1+\gamma)\,\Lambda(x'\beta)\}, \tag{4}$$

where $\Lambda(x'\beta) = \exp(x'\beta)/(1+\exp(x'\beta))$. As in Papke and Wooldridge (1996), $x'\beta$ is specified as

$$x'\beta = \beta_0 + \beta_1 \text{MRATE} + \beta_2 \text{MRATE}^2 + \beta_3 \ln(\text{EMP}) + \beta_4 \ln(\text{EMP})^2 + \beta_5 \text{AGE} + \beta_6 \text{AGE}^2 + \beta_7 \text{SOLE}.$$

Since $\Lambda(x'\beta)$ is the CDF of a distribution symmetric around zero, (4) is equivalent to

$$Q_{1-\text{PRATE}}(1-\theta|x) = \max\{0, (1+\gamma)\,\Lambda(-x'\beta) - \gamma\}.$$

[6] PRATE is computed as the ratio between two integers. Therefore, strictly speaking, it has a discrete distribution. However, the number of support points is so large that it is reasonable to model PRATE as if its distribution were mixed.

Therefore, the framework described in Section 2 for modelling the quantiles of fractional data with a mass-point at zero will be used here to model fractional data with a mass-point at one.

The estimation procedure was implemented as follows. For each value of θ, we performed a grid search over γ to minimize

$$S(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \rho_\theta \left(\text{PRATE}_i - \min\left\{1, (1+\gamma)\,\Lambda\left(x_i'\hat{\beta}(\gamma)\right)\right\}\right).$$

The search was performed for values of γ from 0.001 to 5, in steps of 0.001. For each value of γ, β was estimated by censored linear quantile regression, using the CH three-step algorithm. The algorithm was implemented using $q = 10\%$ for the selection of the first sub-sample and the adaptive cut-off point for the second sub-sample. All computations were performed using TSP 5.0 (Hall and Cummins, 2005).

Table 1 displays the estimated parameters and corresponding standard errors, for $\theta \in \{0.10, 0.25, 0.40\}$. For $\theta > 0.4$, estimation is difficult because $g(x'\beta, \gamma)$ becomes almost linear, making the identification of γ very tenuous. Figures 1 to 3 display the plots of $n \times (S(\gamma) - S(\hat{\gamma}))$ for the most relevant range of values of γ and for the different values of θ. The shape of the objective function for $\theta = 0.4$ clearly reveals the difficulty in identifying γ for the upper quantiles. As mention before, this suggests that a model of the form $Q_y(\theta|x) = \min\{1, x'\beta\}$ may be adequate. For completeness, Table 1 also includes the parameter estimates for $\theta = 0.4$ obtained with the linear specification, as well as the estimates of the parameters of the conditional mean obtained using the Bernoulli pseudo maximum likelihood estimator of Papke and Wooldridge (1996).

In order to assess the adequacy of the proposed specification, RESET-type tests (Ramsey, 1969) were performed. The corresponding test statistics were computed as t-ratios for the significance of $\left(x'\hat{\beta}\right)^2$ in the third step of the CH procedure.[7] This implementation of the RESET test for censored quantile regression, which is slightly different from

[7] In principle, given that $f(x'\beta)$ can be s-shaped, it would be advisable to test for the joint significance of $\left(x'\hat{\beta}\right)^2$ and $\left(x'\hat{\beta}\right)^3$. However, for the range of values included in the sample, $f(x'\beta)$ is concave, at least for values of γ close to the optimum. Therefore, inclusion of the cubic term is likely to reduce the power of the test.

that suggested by Otsu (2007), is adopted here due to its computational simplicity. The bottom row of Table 1 gives the tests statistics for $\theta \in \{0.10, 0.25, 0.40\}$, computed at the estimated value of γ, as well as the analogous test statistic for the conditional mean regression. In all cases, the test statistics provide no evidence of misspecification.

Table 1: Parameter estimates

	Quantile regression				Mean
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.40$	$\theta = 0.40$ Linear	regression
Intercept	3.5401	2.8571	0.4823	1.3651	5.1053
	(0.6848)	(0.5398)	(1.9143)	(0.0613)	(0.4156)
MRATE	1.2897	1.4035	0.4857	0.2477	1.6650
	(0.1548)	(0.2500)	(0.2729)	(0.0484)	(0.1042)
MRATE^2	−0.2688	−0.2727	0.0077	0.0066	−0.3321
	(0.0740)	(0.0492)	(0.0781)	(0.0356)	(0.0256)
$\ln(\text{EMP})$	−0.9740	−0.7347	−0.3105	−0.1658	−1.0306
	(0.1627)	(0.1247)	(0.1735)	(0.0153)	(0.1097)
$\ln(\text{EMP})^2$	0.0525	0.0401	0.0166	0.0089	0.05363
	(0.0096)	(0.0070)	(0.0093)	(0.0009)	(0.0071)
AGE	0.0499	0.0354	0.0146	0.0070	0.0548
	(0.0091)	(0.0083)	(0.0087)	(0.0017)	(0.0077)
AGE^2	−0.0006	−0.0004	−0.0002	−0.0001	−0.0006
	(0.0001)	(0.0002)	(0.0001)	(0.0000)	(0.0002)
SOLE	−0.0319	0.0252	0.0437	0.0206	0.0643
	(0.0490)	(0.0333)	(0.0320)	(0.0086)	(00498)
γ	0.0750	0.1650	1.1940	—	—
	(0.0517)	(0.0632)	(1.8571)	—	—
Objective function	146.3723	234.2997	254.4501	254.4622	—
RESET	−0.3677	−0.5553	−1.1113	−0.36440	1.0091

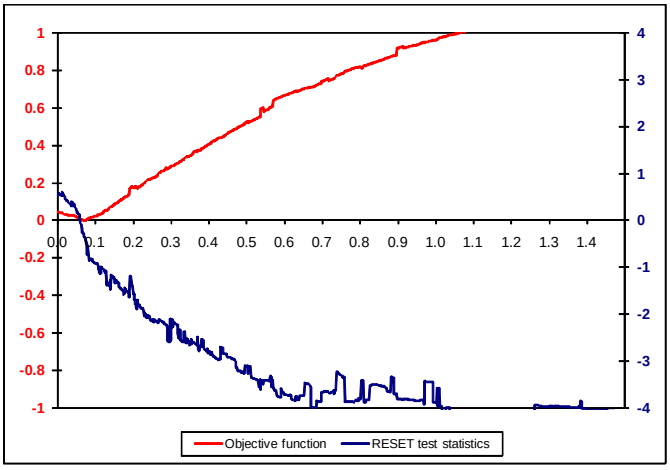

Fig. 1: Objective function and RESET test statistics for different values of γ and $\theta = 0.10$.

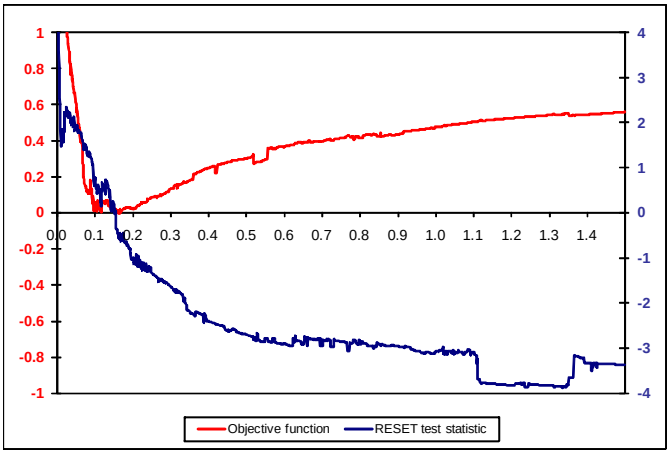

Fig. 2: Objective function and RESET test statistics for different values of γ and $\theta = 0.25$.

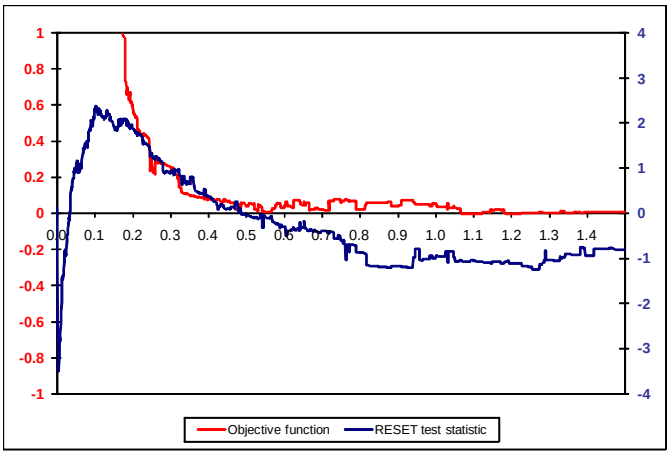

Fig. 3: Objective function and RESET test statistics for different values of γ and $\theta = 0.40$.

Figures 1 to 3 also plot the RESET-type test statistics for a range of values of γ and for $\theta \in \{0.10, 0.25, 0.40\}$. It is interesting to notice that these statistics tend to decrease as γ increases, changing sign, and therefore being close to zero, for values of γ close to the optimum. These plots are particularly useful in cases where $S(\gamma)$ is relatively flat (as for $\theta = 0.4$), helping to pinpoint the region where the objective function is minimized.

For $\theta = 0.10$, the estimated value of γ is close to zero, suggesting that $Q_{\text{PRATE}}(0.10|x)$ hardly reaches one. Indeed, in the sample, the estimated value of $Q_{\text{PRATE}}(0.10|x)$ equals one for a single observation. The estimates of γ increase with θ, reflecting the fact that higher quantiles become flat at one for smaller values of $x'\beta$. The estimated values of $Q_{\text{PRATE}}(0.25|x)$ and $Q_{\text{PRATE}}(0.40|x)$ are equal to one for about 10% and 25% of the observations, respectively.

Using conventional significance levels, Papke and Wooldridge (1996) find that, with the exception of SOLE, all regressors are statistically significant in the mean regression. The results in Table 1 show that, for the lower quantiles, SOLE is also the only regressor not statistically significant. In contradistinction, for $\theta = 0.40$, with the linear specification, all the estimated parameters are statistically significant, including the one associated with SOLE. However, with the non-linear model, all parameters become statistically insignificant as a result of the difficulty in identifying γ. More importantly, for $\theta = 0.40$, the sign of the coefficient of MRATE^2 changes. This suggests that MRATE, which is the more interesting regressor in the model, has very different effects on different regions of the conditional distribution of PRATE.

To better illustrate the effect of MRATE on the conditional distribution of PRATE, Figure 4 displays the estimated conditional quantiles (from top to bottom, for θ equal to 0.40, 0.25 and 0.10, respectively) and conditional expectation (dashed red line) of PRATE, as a function of MRATE, evaluated at the sample means of ln (EMP) and AGE, and for SOLE $= 0$.[8]

[8] For $\theta = 0.4$, the conditional quantiles estimated with the linear and non-linear models are virtually indistinguishable.

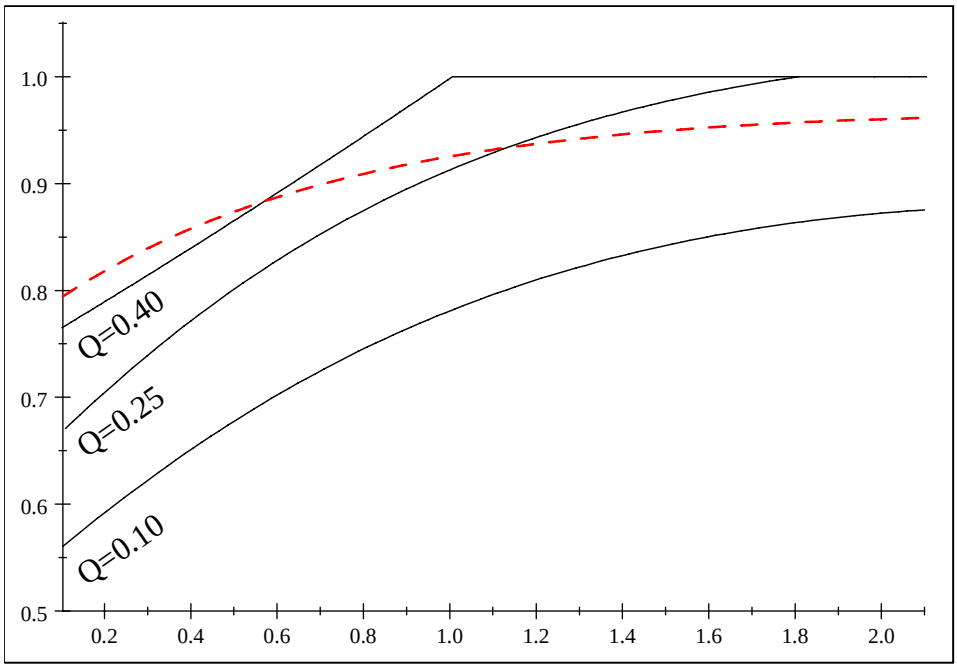

Fig. 4: Quantiles and expectation of PRATE as a function of MRATE

Figure 4 clearly shows that the mean regression of PRATE on MRATE and other control variables is not enough to unveil the complexity of the effects of MRATE on the conditional distribution of PRATE. For example, the plot shows that MRATE has a non-monotonic effect on the dispersion of PRATE. Moreover, for an important part of the sample, the depicted quantile functions are steeper than the conditional mean. Therefore, the mean regression masks the very different effects that changes in MRATE have on different areas of the conditional distribution.

The difference between the marginal effects of the regressors on the mean and on different quantiles is not specific to MRATE. Table 2 presents estimates of the marginal effects of all regressors, for the estimated quantile functions and for the conditional mean, evaluated at the sample means of MRATE, ln (EMP) and AGE, and for SOLE $= 0$. The results in Table 2 confirm that the marginal effects of the regressors vary widely across the different conditional quantiles. These figures also show that, for $\theta = 0.4$, the linear

and non-linear model essentially lead to the same estimates, with the ones from the linear model being much more precise.[9]

Table 2: Estimated marginal effects

	Quantile regression				Mean
	$\theta = 0.10$	$\theta = 0.25$	$\theta = 0.40$	$\theta = 0.40$ Linear	regression
MRATE	0.2108	0.2276	0.2653	0.2576	0.1016
	(0.0173)	(0.0176)	(0.2170)	(0.0124)	(0.0047)
ln (EMP)	−0.0612	−0.0430	−0.0450	−0.0449	−0.0260
	(0.0065)	(0.0035)	(0.0360)	(0.0033)	(0.0018)
AGE	0.0084	0.0059	0.0057	0.0053	0.0033
	(0.0012)	(0.0006)	(0.0045)	(0.0007)	(0.0032)
SOLE	−0.0076	0.0057	0.0224	0.0206	0.0054
	(0.0118)	(0.0073)	(0.0219)	(0.0086)	(0.0042)

5. DISCUSSION

In this paper we propose simple methods to estimate conditional quantiles of fractional data. We show that the particular estimator to be used depends on the specific nature of the variate of interest, and develop a procedure to estimate conditional quantiles for fractional mixed data with mass-points at zero or one. The implementation of the proposed method is illustrated using a well-known dataset.

The estimator developed for the case of fractional data with mass-points can easily be extended to other kinds of data from mixed distributions. For example, consider non-negative data with a mass-point at zero, like that often found in the study of the determinants of medical expenditures or international trade. In this case, it may still be appropriate to specify

$$Q_y(\theta|x) = \max\{0, g(x'\beta, \gamma)\},$$

[9] For θ equal to 0.10 and 0.25, the linear and non-linear models lead to substantially different results and the linear model fails the RESET test.

where now $g(x'\beta, \gamma)$ in not bounded above by 1. For example, we may specify $g(x'\beta, \gamma)$ as

$$g(x'\beta, \gamma) = \exp(x'\beta) - \gamma,$$

where γ is again a non-negative shape parameter. Mutatis mutandis, all the results in Section 3 are valid in this case and therefore inference presents no additional problems.

REFERENCES

Angrist, J.D.; Chernozhukov, V. and Fernandez-Val, I. (2006). "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539-563.

Buchinsky, M. and Hahn, J. (1998). "An Alternative Estimator for the Censored Quantile Regression Model," *Econometrica*, 66, 653-672.

Chamberlain, G. (1994). "Quantile Regression, Censoring and the Structure of Wages," in Sims, C.A. (ed.), *Advances in Econometrics*, Cambridge: Cambridge University Press.

Chernozhukov, V. and Hong, H. (2002). "Three-Step Censored Quantile Regression and Extramarital Affairs," *Journal of American Statistical Association*, 97, 872–882.

Duan N., Manning, W.G. Jr, Morris, C.N. and Newhouse J.P. (1983). "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business and Economic Statistics*, 1, 115-126.

Fitzenberger, B. (1997). "A Guide to Censored Quantile Regressions," in Maddala, G.S. and Rao, C.R. (eds.), *Handbook of Statistics, Volume 15: Robust Inference*, New York: North-Holland.

Hall, B.H. and Cummins, C. (2005). *TSP 5.0 User's Guide*, Palo Alto (CA): TSP International.

La Porta, R., Lopez-de-Silanes, F. and Zamarripa, G. (2003). "Related Lending," *Quarterly Journal of Economics*, 118, 231-268.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference With Nonexperimental Data*, New York: John Wiley & Sons.

Machado, J.A.F. and Santos Silva, J.M.C. (2005). "Quantiles for Counts," *Journal of the American Statistical Association*, 100, 1226-1237.

Kim, T.-H. and White, H. (2002). "Estimation, Inference and Specification Testing and Possibly Misspecified Quantile Regression," Advances in Econometrics (forthcoming).

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

Otsu, T. (2007). "RESET for quantile regression," mimeo, Cowles Foundation, Yale University.

Papke, L.E. and J.M. Wooldridge (1996). "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 619-632.

Powell, J.L. (1984). "Least Absolute Deviation Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303-325.

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

Powell, J.L. (1991). "Estimation of Monotonic Regression Models Under Quantile Restrictions," in Barnett, W.A., Powell, J.L. and Tauchen, G.E. (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press.

Ramsey, J.B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society B*, 31, 350-371.

Santos Silva, J.M.C. and Tenreyro, S. (2006), "The Log of Gravity," *The Review of Economics and Statistics*, 88, 641-658.

Tersvirta, T. and Mellin, I. (1986). "Model Selection Criteria and Model Selection Tests in Regression Models," *Scandinavian Journal of Statistics*, 13, 159-171.