

# Quantiles for Fractional Data\*

José A. F. Machado

Faculdade de Economia, Universidade NOVA de Lisboa

J. M. C. Santos Silva

ISEG/Universidade Técnica de Lisboa and CEMAPRE

July 2006

## 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. The main results are illustrated with an application.

*JEL classification code:* C13; C25; C51.

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

---

\*VERY PRELIMINARY; DO NOT QUOTE. Machado is consultant for the Research Department of Banco de Portugal. Santos Silva is grateful for the hospitality, working conditions and financial support provided by Banco de Portugal, which made this work possible. The authors also gratefully acknowledge the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. The usual disclaimer applies.

Address for correspondence: João Santos Silva, ISEG, R. do Quelhas 6, 1200 Lisboa, Portugal. Fax: 351 213922781. E-mail: jmcass@iseg.utl.pt.

## 1. INTRODUCTION

Empirical researchers are often faced with the need to model fractional data. Modeling 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. Conditional quantiles provide a simple way of constructing this sort of confidence intervals and, in general, are interesting because they provide a complete description of the way the conditional distribution of the fractional variate depends on the regressors.

The specification and estimation of conditional quantile functions for fractional data raises some interesting problems. The strategy to adopt for the estimation of quantile regression for fractional data depends on the specific nature of its distribution. 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.<sup>1</sup> In this paper we look at the estimation of quantile regression for fractional data, focusing particular attention on the case of mixed distributions.

The remainder 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 estimation method for the case in which the data has a mass-point at zero. Section 4

---

<sup>1</sup>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.

illustrates the application of the main results. Finally, section 5 contains some concluding remarks.

## 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  $\alpha \in (0, 1)$ , the researcher specifies the following parametric model

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

where  $\Lambda(x'\beta)$  is a function bounded between 0 and 1. Then,  $\beta$  can be estimated by performing the usual linear quantile regression estimation of  $\Lambda^{-1}(y)$  on  $x$ . For example, when  $\Lambda(x'\beta)$  is the logit,  $\beta$  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) observation in which  $y$  is zeros 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 nudge 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.

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

$$Q_y(\alpha|x) = \max\{0, f(x'\beta)\}$$

where  $y$  is the (fractional) variate of interest and  $f(x'\beta)$  is a function such that  $f(z) < 1$ ,  $\forall z$ .

The choice of  $f(x'\beta)$  is an empirical matter. One possibility is to define  $f(x'\beta) = 1 - \exp(x'\beta)$ . However, this function is always concave and in most cases a  $s$ -shaped function will be preferable. In the spirit of Papke and Wooldridge (1996), we suggest the following specification

$$f(x'\beta) = (1 + \gamma) \Lambda(x'\beta) - \gamma \quad (1)$$

where  $\Lambda(x'\beta)$  is a CDF and  $\gamma > 0$  is an unknown parameter.<sup>2</sup>

Depending on the specification of  $\Lambda(x'\beta)$  and on the value of  $\gamma$ , the specification of  $f(x'\beta)$  in (1) leads to  $s$ -shaped or concave quantiles. 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 of  $\gamma$  depends on the curvature of  $\Lambda(x'\beta)$ . If the data is such that  $f(x'\beta)$  is essentially linear, identification of  $\gamma$  will be difficult. Therefore, in applications, difficulty in identifying  $\gamma$  suggests that simple censored linear quantile regression will be adequate.

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

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

That is, conditional on the value of  $\gamma$ ,  $\beta$  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  $\gamma$  and  $\beta$ .

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

---

<sup>2</sup>Notice that if the mass-point is at one, this specification can be used to model  $1 - y$ . In case there are mass-points at both one and zero, the following specification can be adopted

$$f(x'\beta) = (1 + \gamma + \theta) \Lambda(x'\beta) - \gamma.$$

quantile regression defined by

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

### 3. ESTIMATION

In this section we will adapt Chamberlain's (1994) two-steps estimation strategy of the Box-Cox quantile model to our present setting. The basic intuition for the population is as follows. The assumption that, for a given  $\alpha \in (0, 1)$ , there exist  $\beta_0$  and  $\gamma_0$  such that

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

implies, under standard conditions on the conditional density of  $y$ , that  $(\beta_0, \gamma_0)$  is the sole solution of

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

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

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

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

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

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

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

The proposed estimator will be the sample analogue of the procedure described above. In a first step, for fixed values of  $\gamma$ ,  $\hat{\beta}(\gamma)$  will be the estimator of  $\beta$  in a linear CQR 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, will solve

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

Then, a one dimensional search over  $\gamma$  to minimize

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

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

To simplify notation let

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

where  $g(x'\beta, \gamma) = (1 + \gamma) \Lambda(x'\beta) - \gamma$ . Also put,

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

Finally, let

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

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

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

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

Under suitable regularity conditions (Powell, 1990, and Fitzenberger, 1997), the estimator is consistent and has the following linear representation

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

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

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

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

$$\begin{pmatrix} \sqrt{n}(\hat{\beta} - \beta_0) \\ \sqrt{n}(\hat{\gamma} - \gamma_0) \end{pmatrix} \rightarrow N(0, L_0^{-1} M_0 L_0^{-1}).$$

The asymptotic covariance matrix may be estimated by standard plug-in procedures. As usual for quantile regression methods, the only critical issue is the estimation of the conditional density of the response variable. A possible solution is to use the kernel methods described, for instance, in Fitzenberger (1997). Since (1) is just an approximation to the functional form of the quantile regression functions, misspecification robust estimators of the covariance matrix should be used (see Chamberlain, 1994, Kim and White, 2002, and Angrist, Chernozhukov and Fernandez-Val, 2004).

#### 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 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), the plans average 12 years in age (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. This suggests that the higher quantiles of the distribution will be flat as 1 for most observations. Therefore, we expect the role of the covariates to be particularly important for the lower quantiles of the distribution.

Given the characteristics of the data, in this particular example we adopt the following specification for the quantiles of PRATE

$$\begin{aligned} Q_{\text{PRATE}}(\alpha|x) &= \min\{1, f(x'\beta)\}, \\ f(x'\beta) &= (1 + \gamma) \Lambda(x'\beta), \end{aligned}$$

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

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

Since  $\Lambda(x'\beta)$  is specified as the CDF of a symmetric distribution, this specification is equivalent to

$$\begin{aligned} Q_{1-\text{PRATE}}(1 - \alpha|x) &= \max\{0, f(x'\beta)\}, \\ f(x'\beta) &= (1 + \gamma) \Lambda(-x'\beta) - \gamma. \end{aligned}$$

That is, the framework described in Section 2 for modeling 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. Table 1 displays the estimated parameters and corresponding standard errors, for  $\alpha \in \{0.10, 0.25, 0.40\}$ .

The estimation procedure was implemented as follows. For each value of  $\alpha$ , we first performed a grid search over  $\gamma$  to minimize

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

The search was performed for values of  $\gamma$  from 0.001 to 5, in steps of 0.001. For each value of  $\gamma$ ,  $\beta$  was estimated by CQR, using the algorithm described in Fitzenberger (1997).

Then, starting from the optimal value of  $\gamma$  found in the grid search, the Broyden-Fletcher-Goldfarb-Shanno algorithm was used to estimate  $\gamma$  and  $\beta$ . At this moment, the standard errors for  $\hat{\beta}$  are estimated by bootstrap and are conditional on the estimated value of  $\gamma$ . All computations were performed using TSP 5.0 (Hall and Cummins, 2005).

Table 1: Non-linear quantile regression results

	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.40$
Intercept	3.54516 (0.41468)	2.97708 (0.30408)	1.39652 (0.16681)
MRATE	1.29734 (0.12156)	1.44968 (0.07284)	0.23959 (0.22305)
MRATE <sup>2</sup>	-0.26993 (0.03673)	-0.27663 (0.01957)	0.49547 (0.24797)
ln (EMP)	-0.97511 (0.11116)	-0.75836 (0.08045)	-0.41478 (0.04083)
ln (EMP) <sup>2</sup>	0.05253 (0.00731)	0.04115 (0.00520)	0.02225 (0.00255)
AGE	0.05033 (0.00781)	0.04063 (0.00605)	0.02025 (0.00405)
AGE <sup>2</sup>	-0.00056 (0.00018)	-0.00047 (0.00013)	-0.00022 (0.00010)
SOLE	-0.03448 (0.05073)	0.03924 (0.03747)	0.04578 (0.02238)
$\gamma$	0.073200	0.13605	0.64466
$S(\gamma)$	146.36499	234.27571	254.32390

Papke and Wooldridge (1996, p. 631) find that, with the exception of SOLE, all regressors are statistically significant in the mean regression. The results in Table 1 show that SOLE is statistically significant for  $\alpha = 0.40$ , although it is non significant for the lower quantiles. On the other hand, MRATE, which is the more interesting regressor, appears

to loose significance as one moves to the higher quantiles. What is more interesting, the coefficient of  $\text{MRATE}^2$  changes from negative to positive, indicating that the way  $\text{MRATE}$  affects the upper quantiles is substantially different. This suggests that the effect of this regressor on the conditional distribution of  $\text{PRATE}$  is relatively complex.

To better illustrate the effect of  $\text{MRATE}$  on the conditional distribution of  $\text{PRATE}$ , Figure 1 displays the estimated conditional quantiles (from top to bottom, for  $\alpha$  equal to 0.40, 0.25 and 0.10, respectively) and conditional expectation (in red) of  $\text{PRATE}$ , as a function of  $\text{MRATE}$ , evaluated at the sample means of  $\ln(\text{EMP})$  and  $\text{AGE}$ , and for  $\text{SOLE} = 0$ . Clearly, the mean regression of  $\text{PRATE}$  on  $\text{MRATE}$  and other control variables is not enough to unveil the complexity of the effects of  $\text{MRATE}$  on the conditional distribution of  $\text{PRATE}$ . For example, Figure 1 shows that  $\text{MRATE}$  has a non-monotonic effect on the dispersion of  $\text{PRATE}$ , with the displayed quantile functions first getting closer to each other, but then diverging.

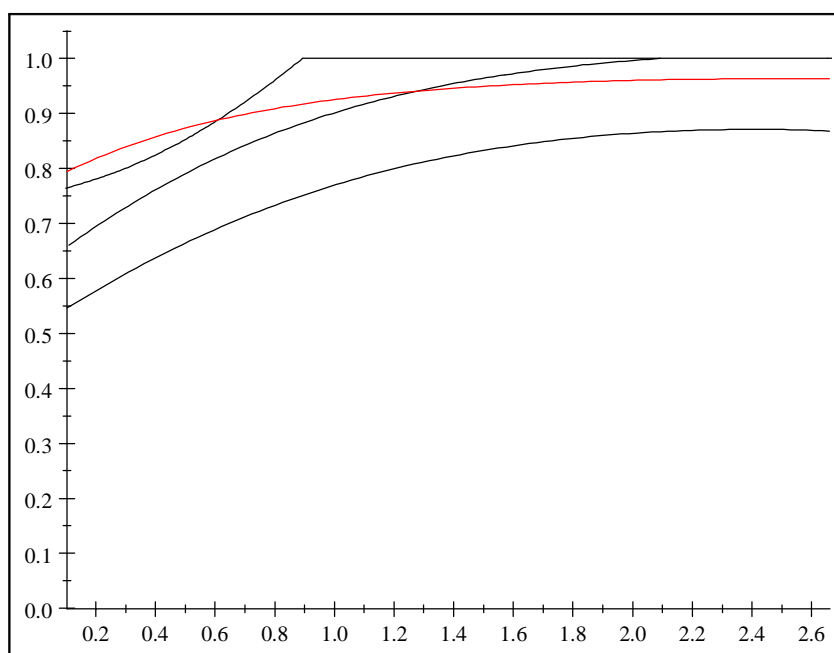


Fig. 1: Quantiles and expectation of  $\text{PRATE}$  as a function of  $\text{MRATE}$

## 5. CONCLUDING REMARKS

In this paper we propose a simple method to estimate conditional quantiles of fractional data, with possible mass-points at zero and/or one. The implementation of the proposed method is illustrated using a well-known dataset.

## REFERENCES

- Angrist, J.D.; Chernozhukov, V. and Fernandez-Val, I. (2004). “Quantile Regression Under Misspecification, with an Application to the US Wage Structure”, National Bureau of Economic Research Working Paper 10428.
- Chamberlain, G. (1994). “Quantile Regression, Censoring and the Structure of Wages”, in C.A. Sims (ed.), *Advances in Econometrics*, Cambridge University Press, Cambridge.
- Fitzenberger, B. (1997). “A Guide to Censored Quantile Regressions”, in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics, Volume 15: Robust Inference*, 405-437.
- Hall, B. H. and Cummins, C. (2005). *TSP 5.0 User's Guide*, TSP International, Palo Alto (CA).
- 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.
- Papke, L. E., 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.