

Predictions in Quantile Regressions

Marilena Furno

Department of Agriculture, Università degli Studi di Napoli "Federico II", Napoli, Italy
Email: marfurno@unina.it

Received 6 June 2014; revised 10 July 2014; accepted 18 July 2014

Copyright © 2014 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Two different tools to evaluate quantile regression forecasts are proposed: MAD, to summarize forecast errors, and a fluctuation test to evaluate in-sample predictions. The scores of the PISA test to evaluate students' proficiency are considered. Growth analysis relates school attainment to economic growth. The analysis is complemented by investigating the estimated regression and predictions not only at the centre but also in the tails. For out-of-sample forecasts, the estimates in one wave are employed to forecast the following waves. The reliability of in-sample forecasts is controlled by excluding the part of the sample selected by a specific rule: boys to predict girls, public schools to forecast private ones, vocational schools to predict non-vocational, etc. The gradient computed in the subset is compared to its analogue computed in the full sample in order to verify the validity of the estimated equation and thus of the in-sample predictions.

Keywords

Predictions, Quantile Regressions, Gradient

1. Introduction

Quantile regression is a robust procedure particularly helpful in dealing with asymmetric and with non-normal distributions. It is very useful to evaluate data sets where it is relevant to analyse and estimate a regression model not only at the mean/median of the conditional distribution but also in the tails, measuring the impact of the regressors on the dependent variable at different quantiles.

Besides the estimates of the regression coefficients at different quantiles, the predictions of the model at a given quantile and their summary measures are the focus of the analysis. As the goodness of fit index, an indicator of quantile regression predictive power is a local measure, relative to the specific quantile under analysis. The regression estimates at the selected quantile provide forecasts for that specific quantile, and it may very well be the case that the predictive ability of the model changes at a different quantile: a model may predict the median better than the upper or the lower quantile. The literature on quantile regression predictions is not very ex-

tensive. Mayr *et al.* [1], for instance, consider quantile regressions to compute predictive intervals. When a new observation for the explanatory variables is available, x_{n+1} , the fitted value computed at the $\alpha/2$ and $1 - \alpha/2$ regression quantiles provide the prediction interval for the explanatory variable, y_{n+1} . In the weather forecast, the “quantile verification score” (QVS) is implemented, as for instance in Friederichs and Hense, [2]. It is a quantile-specific approach where the forecaster optimises the performance by predicting the “true” quantiles. In what follows a different approach is considered. Different tools to evaluate quantile regression forecasts are proposed and their characteristics are explored and compared. The first is a summary statistic measuring forecast errors. The second is a fluctuation test, based on the quantile regression gradient, which is here considered to evaluate forecasts.

The math scores at the international OECD-PISA test of 15-year-old Italian students are analysed here. The OECD-PISA test has been administered in year 2000, 2003, 2006 and 2009 to verify students’ proficiency. The selected model is estimated by OLS and quantile regression. The purpose is to link student performance to school efficiency. There is a long lasting debate on the role of school variables on student proficiency. Generally speaking, the rate of return to additional years of schooling is large. In addition, growth analysis emphasizes school attainment and shows its high relation to differences in economic growth across countries. Analyses of the relationship between schooling and economic growth by Barro [3] and Mankiw *et al.* [4] find a significant positive link between quantitative measures of schooling and economic growth. Hanushek [5] relates one standard deviation difference of student performance to a one percent difference in the annual growth rate of per capita gross domestic product, although higher spending in school resources does not necessarily involve higher scores, as discussed in Hanushek and Woessmann [6].

Therefore, prediction of student performance is one of the possible indicators of economic growth and schooling is one of the elements to improve in order to increase the rate of growth of a country. These analyses generally rely on average values, while policy may be more effective when focusing on the tails, away from the mean. This would be the case of schemes seeking to improve the low scoring in order to rise the general performance. But the analysis of the model on average is not very helpful when the target is to improve the lower tail.

The predictive power of the model is considered by analysing both in-sample and out-of-sample forecasts. To evaluate out-of-sample forecasts the estimates in one wave are employed to forecast the following waves, under the assumption of unchanged model from one wave to the other, that is assuming a fixed coefficient model. The comparison of realizations and predictions based on the estimates of the previous wave provides a measure of the validity of the out-of-sample forecasts at different quantiles. The validity of in-sample forecasts is analysed by comparing the gradient computed in a subset with the gradient computed in the full sample. The following procedure is here proposed:

- 1) Exclusion of q data from the sample, possibly chosen by a specific selection rule,
- 2) Estimation of the model in the selected subset,
- 3) Computation of the residuals with respect to this set of estimates for the entire sample,
- 4) Comparison of the gradient using these residuals, with the gradient computed in the entire sample.

The model yields good in-sample predictions when such a difference goes to zero. In this case, the model in the subset provides a good approximation for the excluded data, and predictions are good. Vice versa, if the difference between the two sets of results is sizable and cannot be neglected, the subset poorly approximates the excluded observations and the in-sample forecasts are not reliable.

Section 2 summarizes the quantile regression estimator whereas the data and the model are discussed in Section 3. Section 4 presents the data set and Section 5 discusses the results. The last section draws the conclusions. In the Appendix are the Stata codes used to implement the comparison of partial and total sum of the gradient proposed here.

2. Quantile Regression

Consider the linear regression model $y_i = x_i^T \beta + e_i$, where y_i is the dependent variable, x_i^T is the p -dimensional row vector of a single observation for the p explanatory variables, e_i is the error term having continuous and strictly positive density $f(\cdot)$ in a sample of size n . The quantile regression model provides robust estimates of the coefficients and does not require distributional assumptions. Its objective function is given by¹:

$$V(\theta) = \sum_i w_i |e_i(\theta)| = \sum_{y_i > x_i^T \beta} \theta |y_i - x_i^T \beta| + \sum_{y_i < x_i^T \beta} (1-\theta) |y_i - x_i^T \beta| \quad (1)$$

where the asymmetric weighting system w_i is equal to θ for positive errors and to $1 - \theta$ for negative e_i . When $\theta = 0.5$ the weights are symmetric and the objective function simplifies into $V(0.5) = \sum_i \theta |y_i - x_i^T \beta|$, which provides an estimate of the conditional median regression, or least absolute deviation (LAD). Often the median regression estimates do not significantly differ from the conditional mean regression (OLS), and this is the case with symmetric conditional distributions. The advantage of quantile regression relies in the possibility of investigating the dependent variable at many points of the conditional distribution, not only at the centre but also in the tails.

The predictions of the quantile regression model are then analysed. Forecast in quantile regression, just as the goodness of fit index, can be considered a local measure, relative to the specific quantile under investigation. In the following sections predictions for one wave based on the estimates computed in the previous wave provide out-of-sample forecasts, while predictions within each wave provide in-sample forecasts for a given quantile regression.

A first approach is to compute a summary measure of forecast error and, among the various existing statistics on predictions, the mean absolute deviation of forecasts from realizations, MAD, is here considered. Quantile regressions are based on the l_1 norm, therefore MAD seems the appropriate index to summarise the predictive power in quantile regressions, particularly so at the median. The statistic is given by:

$$\text{MAD}(\theta) = \sum |y_i - \hat{y}_i(\theta)|/n = \sum |\hat{\varepsilon}_i(\theta)|/n \quad (2)$$

where $\hat{y}_i(\theta)$ are the fitted values computed at the selected quantile and $\hat{\varepsilon}_i(\theta)$ are the prediction errors at the quantile θ .

In addition, to verify the reliability of the in-sample forecasts it is possible to consider the comparison between partial and total sum of the gradient. This comparison, considered by Qu [12] and by Oka and Qu [13] to define a changing coefficient test and by Xiao [14] for a test of cointegration, focuses on the quantile regression gradient

$$g(n, \theta) = n^{-1/2} \sum_{i=1}^n x_i \psi_\theta(y_i - x_i \beta(\theta)) = n^{-1/2} \sum_{i=1}^n x_i \psi_\theta(e_i) \quad (3)$$

where $\psi_\theta(e_i) = \theta - 1(e_i \leq 0)$ is the sign function. In a subset of size n_1 , the gradient (n_1, θ) , is:

$$g(n_1, \theta) = n_1^{-1/2} \sum_{i=1}^{n_1} x_i \psi_\theta(e_i) \quad (4)$$

and a function of the gradient, considered for both n and n_1 sample sizes, is defined as

$$H(n, \theta) = (X^T X)^{-1/2} \sum_{i=1}^n x_i \psi_\theta(e_i) \quad (5)$$

$$H(n_1, \theta) = (X^T X)^{-1/2} \sum_{i=1}^{n_1} x_i \psi_\theta(e_i) \quad (6)$$

where X is the (n, p) matrix of explanatory variables. The comparison of $H(n_1, \theta)$ and $H(n, \theta)$ is provided by the following test function defined in Qu [12]:

$$Q_\theta = \sup_{n_1} \left\| (\theta(1-\theta))^{-1/2} [H(n_1, \theta) - \lambda H(n, \theta)] \right\| \quad (7)$$

where $\sup \|\cdot\|$ is the sup norm selecting the maximum difference between $H(n_1, \theta)$ and $H(n, \theta)$ within the entire vector of coefficients and $\lambda = n_1/n$ is the proportion of data in the subset. The gradient estimated in the n_1 subset, yielding $H(n_1, \theta)$, is compared with the gradient estimated in the full sample $H(n, \theta)$. In case, $H(n_1, \theta)$ is a good proxy for $H(n, \theta)$, the in-sample predictions are reliable. Vice versa, when the model is inadequate $H(n_1, \theta)$ and $H(n, \theta)$ diverge and the in-sample forecasts are unreliable. In addition, the sample of size n can be split into $n = n_1 + q$ by choosing a specific selection rule related to the variables of the equation: for instance in n_1 the public funded schools can be selected and compared to the private ones, or the schools located in the southern regions compared to all the other schools. A relevant discrepancy would signal poor forecasts due to a statistically significant gap between the two subsets. The model is separately estimated in the subset and in the full sample. The

¹For further details see Koenker and Bassett [7], Koenker [8], Koenker and Machado [9], Davino *et al.* [10], Hao and Naiman [11].

residuals from the regression estimated in the n_1 subset are compared to the residuals estimated in the entire sample through the $H(n_1, \theta)$ and $H(n, \theta)$ functions. If they do not significantly differ the in-sample forecasts of the model based on that specific selection rule are reliable and there is no gap between the subset and the entire sample.

Furthermore, the comparison between partial and total sum of the gradients is analyzed separately for each regression coefficient and the convergence/divergence between $H(n_1, \theta)$ and $H(n, \theta)$ becomes a diagnostic tool to explore the validity of each explanatory variable for predictive purposes. In detail, the equation is estimated twice, in the n_1 subset and in the full sample. From the two estimated equations two sets of n residuals are computed². The former, computed in the subset, provides an estimate of $H(n_1, \theta)$, while the latter allows to evaluate $H(n, \theta)$. The behavior of this divergence spells out whether the selected subset is fit or unfit to predict the q omitted observations. The comparison of the gradients becomes a diagnostic tool to gauge the validity of each variable and thus the predictive capability of the selected model.

3. The Model

The OECD-PISA test of Italian students has been administered in year 2000, 2003, 2006 and 2009. The different waves of the test are taken into account and the score of the test is related to explanatory variables describing school characteristics like school size, funding, supplied services, student teacher ratio. The purpose is to link student performance to school efficiency. The model is estimated not only at the mean but also in the tails of the conditional distribution, considering low and high proficiency students as well. There is a wide-ranging debate on the role of school variables on student proficiency. Growth analysis emphasizes school attainment and shows its high relation to differences in economic growth across countries. The mathematical skill, in particular, is considered a good proxy of labour force quality. Hanushek [5] relates one standard deviation difference of student performance to a one percent difference in annual growth rate of per capita gross domestic product. However, higher spending in school resources does not necessarily involve higher test scores, as discussed in Hanushek and Woessmann [6].

The model here considered explains the scores on the math tests as a function of region—a dummy variable taking unit value for the southern regions to capture the existing lag in the southern economy; funding—a dummy variable taking unit value for private schools, which in Italy are in a small number and mostly remedial; school track—which distinguishes academic from technical and from vocational fields by means of dummy variables; gender—a dummy taking unit value for boys who, generally speaking, improve upon girls in math while worsen in reading proficiency; school size—providing the number of students enrolled; number of computers in the school; student-teacher ratio; proportion of certified teachers in the school; shortage of teachers in the school; the four categorical variables on poor library facilities, shortage of computers, shortage of textbooks, shortage of laboratory equipment. These four variables take the categories of: a lot, some, a little, not at all. They are transformed in numerical variables by assigning decreasing values from 4 to 1, assigned respectively to the category “a lot” and “not at all”. The sample comprises $n = 52,067$ observations. Table 1 presents few summary statistics of the dependent variable: sample size, mean, standard deviation, 10th, 50th and 90th quantile in each wave of the

Table 1. Math scores, summary statistics.

	n	Mean	Std. Dev.	Median	10 th	90 th
2000	2012	469.56	89.76	474.34	341.86	581.15
2003	10453	499.41	87.75	502.29	383.97	640.78
2006	16014	481.90	90.47	483.21	364.53	597.36
2009	23588	495.02	84.43	497.05	384.18	604.15
all	52067	490.88	87.57	492.76	376.45	602.81

²The coefficients estimated in the n_1 subs *et al.* low to compute the residuals for both the n_1 included and the q excluded observations, so that the comparison between $H(n_1, \theta)$ and $H(n, \theta)$ can be implemented as $n_1 \rightarrow n$.

data set. The mean and the median reach a peak on 2003 then decline to raise again in 2009. The median is always larger than the mean, signalling skewness. The 10th quantile over time has increased more than the 90th, as can be seen by comparing the score in 2000 and 2009 at the selected quantiles: $\text{math}(0.10)_{2009} - \text{math}(0.10)_{2000} = \Delta(0.10)_{2009-00} = 42.32$, $\text{math}(0.50)_{2009} - \text{math}(0.50)_{2000} = \Delta(0.50)_{2009-00} = 22.71$, $\text{math}(0.90)_{2009} - \text{math}(0.90)_{2000} = \Delta(0.90)_{2009-00} = 23$. This shows a greater improvement of the lower scoring students over time. However, the greatest improvement occurs from wave 2000 to 2003 at the top quantile, with a gain in the scores of highly proficient students equal to $\Delta(0.90)_{2003-00} = 59.63$.

The following analysis compares OLS and quantile regression results, and the comparison is particularly useful with asymmetric distributions since it allows to consider the impact of school resources in the tails that is at several levels of student proficiency. The issue of the role and impact of school resources on student proficiency is highly debated, and this impact is hardly homogeneous across quantiles. The selected quantiles-proficiency levels are the 10th, 50th and the 90th quantile³.

4. OLS and Quantile Regression Estimates

Table 2 presents the OLS and the median regression estimates. The regional variable is negative, reflecting the lag of southern regions; the school track yield positive estimates, showing that academic and technical schools improve upon the vocational schools; the gender variable has a positive sign; the private schools have a negative impact, confirming their remedial nature; school size is negative but small; shortage of computers and student teacher ratio have a positive sign; the remaining variables are mostly non-statistically different from zero. The impact of school resources on students' performance is highly debated in the literature. These results are in line with those works ruling out any effect of school inputs, like class size or student-teacher ratio, on students performance—see for instance Hanushek [5] or Hanushek and Woessmann [6]. The comparison of left and right hand sides' results of **Table 2** shows that the median provides results mostly inside the $\pm 2\sigma$ confidence interval of the OLS coefficients, thus the median is not significantly different from the mean regression (OLS). The underlined coefficients, which are those statistically different from their OLS analogues, are the estimates of private, regional and school size parameters.

Table 3 reports the results of the 10th and the 90th quantiles. This table presents many more underlined values that are quantile regression estimates laying outside the $\pm 2\sigma$ OLS confidence intervals. The implication is that both the low and the top scoring students are poorly represented by the conditional mean, the OLS regression. There are however differences from one tail to the other, like for instance the negative impact of private schools is much stronger at the 10th than at the 90th quantile. The underlined and significant coefficients cluster in the upper part of this table, signalling regional variable, gender, school track and funding as the main variables affecting math scores at both ends of the conditional distribution.

This result is confirmed by the interquantile differences, which compute changes across quantiles and verify if a discrepancy between estimated coefficient at two different quantile regressions is statistically significant. The differences between the 90th and the 50th quantile together with the difference between the 10th quantile and the median are reported in **Table 4**. It turns out that these differences are mostly explained by the regional variable—particularly in the 90th - 50th interquantile difference, and by funding, gender, school track in both 90th - 50th and 50th - 10th interquantile comparisons. The regional variable improves in going from the median to the top quantile signalling a reduced regional gap at the higher quantiles. The improvement in math scores granted by the non-vocational schools, both academic and technical, is greater at the first quantile with respect to the median and greater at the median with respect to the 90th quantile; this causes a negative sign for these coefficients in the interquantile differences; furthermore the increments are larger at the median versus the 90th quantile than at the 10th versus the median. Gender improves in going from the median to the top 90th quantile with improvements occurring in every wave but in 2009. The negative impact of private funded schools improves at the median with respect to the first quantile in 2006 and 2009, while is less relevant in explaining the 90th - 50th interquantile difference, since it is significant only in year 2009. The remaining variables either do not explain any of the changes across quantiles—as occurs for the variables number of computers in the school, school size, shortage of teachers and shortage of textbooks—or are significant in explaining the interquantile differences only in one wave—as in the case of proportion of certified teachers, poor library facilities, shortage of laboratory equipment, shortage of computers, student-teacher ratio.

³The estimates computed at additional quantiles are available on request.

Table 2. OLS and median regression estimates in each wave and in the entire sample.

	OLS					MEDIAN				
	2000	2003	2006	2009	all	2000	2003	2006	2009	all
south	-51.35 (11.19)	-64.45 (40.90)	-60.22 (46.31)	-40.10 (40.30)	-52.87 (75.36)	-53.14 (9.55)	<u>-68.28</u> (26.80)	-60.35 (26.69)	<u>-48.38</u> (28.24)	<u>-55.57</u> (32.36)
academic	73.91 (11.83)	91.56 (42.50)	103.37 (52.32)	88.36 (56.28)	90.93 (90.22)	70.06 (7.73)	92.11 (23.36)	103.80 (29.42)	<u>97.001</u> (34.97)	92.30 (36.60)
technical	30.09 (5.85)	56.07 (26.75)	56.13 (32.66)	60.88 (45.72)	53.07 (57.96)	35.61 (4.50)	56.51 (15.88)	59.96 (19.05)	58.75 (26.77)	<u>55.23</u> (24.12)
boy	15.57 (4.06)	33.16 (22.21)	20.64 (17.19)	19.80 (20.09)	23.07 (33.77)	15.03 (2.51)	33.02 (12.66)	21.41 (9.19)	<u>24.78</u> (14.67)	24.22 (14.06)
stud/teac	4.96 (4.52)	1.80 (13.90)	1.57 (4.65)	3.12 (12.63)	2.21 (23.57)	4.34 (2.80)	<u>2.31</u> (8.53)	1.03 (1.78)	2.86 (6.25)	2.34 (7.97)
private	-33.96 (5.16)	-47.05 (18.66)	-51.91 (21.74)	-30.82 (15.77)	-41.11 (33.78)	<u>-19.13</u> (1.50)	<u>-60.45</u> (8.94)	<u>-43.75</u> (7.96)	<u>-41.03</u> (8.23)	-42.79 (9.15)
short lab equip	-2.70 (1.30)	-1.93 (1.87)	-3.28 (4.52)	3.20 (5.47)	-0.93 (2.28)	-2.05 (0.70)	-2.64 (1.64)	-2.87 (2.17)	2.78 (2.74)	-0.76 (0.76)
short textbk	-13.45 (4.81)	-6.51 (5.79)	-5.43 (4.83)	-6.37 (7.47)	-7.12 (13.25)	-14.15 (3.35)	-6.22 (3.33)	-7.64 (4.13)	<u>-4.22</u> (3.11)	-7.09 (5.18)
short compu	2.73 (1.20)	5.71 (5.12)	9.91 (11.81)	-1.76 (2.66)	3.54 (7.83)	1.61 (0.45)	6.20 (3.47)	<u>13.73</u> (8.55)	<u>-4.49</u> (3.94)	4.15 (3.55)
poor library	4.75 (1.88)	2.13 (2.03)	0.49 (0.57)	-2.20 (3.48)	1.73 (3.90)	3.29 (1.01)	2.62 (1.37)	0.43 (0.27)	<u>-0.05</u> (0.05)	1.62 (1.44)
short teacher	0.19 (0.10)	-0.21 (0.27)	-1.38 (2.28)	4.22 (7.00)	0.67 (1.87)	2.39 (0.90)	-0.29 (0.19)	-1.61 (1.22)	4.52 (4.90)	<u>1.82</u> (1.84)
prop cer teac	-0.32 (0.05)	-3.72 (0.85)	7.75 (2.19)	2.83 (1.01)	9.52 (5.78)	6.33 (0.71)	-16.11 (1.63)	7.84 (1.40)	5.99 (1.21)	6.19 (1.27)
school size	0.002 (0.16)	0.003 (1.24)	-0.006 (2.84)	-0.05 (11.86)	-0.007 (5.12)	0.005 (0.57)	0.003 (0.69)	-0.01 (2.67)	<u>-0.06</u> (9.71)	<u>-0.01</u> (3.70)
computer	0.15 (2.74)	0.13 (6.96)	0.18 (13.82)	0.29 (13.65)	0.13 (19.40)	<u>0.06</u> (0.86)	0.16 (5.92)	0.16 (6.62)	<u>0.34</u> (11.28)	<u>0.13</u> (9.55)
constant	399.20 (34.45)	418.89 (83.05)	395.45 (88.71)	414.13 (117.16)	402.75 (193.31)	<u>410.56</u> (21.74)	424.544 (40.84)	397.42 (51.82)	409.07 (66.25)	405.32 (65.58)
n	2012	10453	16014	23588	52067	2012	10453	16014	23588	52067
R ²	0.23	0.35	0.34	0.28	0.30	0.11	0.10	0.16	0.18	0.18
scale $\omega(\theta)$						88.13	77.55	79.44	75.38	97.38

Note: Absolute value of t-statistics in parentheses. Underlined are the quantile regression estimates outside the $\pm 2\sigma$ OLS confidence intervals.

Table 3. First and last quantile regression estimates in each wave and in the entire sample.

	first quantile, 10 th					last quantile, 90 th				
	2000	2003	2006	2009	all	2000	2003	2006	2009	all
south	-58.98 (5.63)	-63.03 (18.41)	<u>-63.01</u> (18.54)	-42.35 (17.77)	-53.61 (21.26)	<u>-36.77</u> (5.99)	-65.55 (22.41)	<u>-51.78</u> (16.72)	<u>-30.09</u> (10.14)	<u>-46.46</u> (29.57)
academic	<u>100.92</u> (6.36)	92.50 (20.26)	<u>95.32</u> (18.94)	87.95 (23.73)	91.74 (26.08)	69.50 (5.69)	87.42 (18.87)	<u>109.70</u> (23.38)	91.11 (18.73)	<u>88.57</u> (35.16)
technical	28.53 (1.75)	<u>48.85</u> (11.52)	<u>46.87</u> (11.98)	<u>57.61</u> (18.20)	<u>48.01</u> (14.22)	26.676 (2.97)	53.58 (12.19)	59.71 (12.49)	<u>68.81</u> (20.92)	52.49 (24.06)
boy	<u>-0.467</u> (0.04)	32.17 (10.60)	<u>16.64</u> (5.52)	18.538 (8.20)	<u>18.80</u> (7.45)	<u>34.02</u> (4.98)	35.32 (11.36)	<u>23.98</u> (6.64)	18.36 (6.16)	<u>25.39</u> (16.01)
stud/teach	5.48 (2.29)	<u>1.14</u> (3.05)	2.06 (1.85)	<u>4.08</u> (7.19)	2.01 (4.76)	3.72 (2.07)	<u>2.46</u> (9.00)	1.81 (2.36)	3.28 (7.23)	2.01 (9.99)
private	<u>-54.14</u> (2.01)	-44.48 (4.82)	<u>-78.48</u> (9.39)	<u>-64.91</u> (10.66)	<u>-54.57</u> (7.35)	-21.90 (1.49)	-48.20 (7.41)	<u>-40.81</u> (5.62)	<u>12.13</u> (1.26)	<u>-27.88</u> (7.51)
sh lab eq	1.41 (0.26)	<u>-6.42</u> (3.13)	-3.28 (1.98)	2.63 (1.74)	-1.56 (1.10)	-6.46 (2.01)	-1.06 (0.63)	-4.40 (2.21)	<u>7.59</u> (4.98)	-1.18 (1.26)
sh text	-14.10 (1.91)	<u>-1.81</u> (0.74)	-7.31 (2.88)	-6.55 (3.59)	<u>-5.85</u> (2.81)	-13.01 (3.22)	-6.21 (3.06)	-7.36 (2.74)	<u>-13.03</u> (6.67)	<u>-9.05</u> (7.79)
sh comp	4.28 (0.70)	-4.702 (2.24)	9.29 (4.36)	-0.48 (0.31)	<u>2.24</u> (1.20)	3.60 (1.09)	<u>13.09</u> (6.04)	<u>5.90</u> (2.43)	<u>0.87</u> (0.47)	<u>5.15</u> (5.16)
poor lib	9.05 (1.95)	<u>7.20</u> (2.90)	1.28 (0.62)	<u>-3.71</u> (2.23)	<u>-0.40</u> (0.24)	2.73 (0.96)	<u>-3.34</u> (1.62)	<u>4.21</u> (1.81)	1.53 (1.08)	1.26 (1.41)
short teacher	<u>1.40</u> (0.29)	-0.15 (0.08)	-2.78 (1.53)	4.44 (3.27)	0.40 (0.29)	<u>-1.39</u> (0.47)	<u>-0.48</u> (0.26)	0.78 (0.47)	<u>7.26</u> (5.39)	0.75 (0.89)
prop cert	-4.38 (0.26)	-5.68 (0.48)	<u>-2.27</u> (0.29)	<u>9.96</u> (1.56)	12.00 (1.69)	13.08 (1.34)	<u>9.70</u> (0.99)	8.29 (0.99)	<u>-10.35</u> (1.93)	<u>16.47</u> (4.51)
school size	<u>0.021</u> (1.13)	0.008 (1.54)	-0.003 (0.69)	<u>-0.06</u> (5.46)	-0.008 (2.14)	<u>-0.008</u> (0.81)	-0.001 (0.24)	-0.008 (1.61)	-0.05 (5.22)	-0.008 (3.41)
computer	0.24 (1.47)	<u>0.08</u> (2.57)	0.17 (6.07)	0.30 (5.37)	0.14 (7.16)	0.15 (1.75)	<u>0.19</u> (5.44)	0.18 (4.77)	0.26 (6.43)	<u>0.11</u> (8.84)
constant	<u>252.8</u> (7.46)	<u>349.5</u> (31.46)	<u>321.8</u> (28.89)	<u>321.1</u> (38.06)	<u>319.5</u> (36.44)	<u>502.9</u> (23.22)	<u>488.8</u> (39.47)	<u>479.7</u> (41.37)	<u>496.2</u> (71.46)	<u>493.1</u> (98.05)
n	2012	10453	16014	23588	52067	2012	10453	16014	23588	52067
pseudo R ²	0.20	0.16	0.16	0.16	0.13	0.18	0.20	0.13	0.15	
scale $\omega(\theta)$	98.02	107.77	103.71	135.48	100.37	90.98	119.14	111.58	90.62	

Note: Absolute value of t-statistics in parentheses. Underlined are the quantile regression estimates outside the $\pm 2s$ OLS confidence intervals.

Table 4. Interquartile difference between 50th - 10th and 90th - 50th quantiles.

	50 th - 10 th					90 th - 50 th				
	2000	2003	2006	2009	all	2000	2003	2006	2009	all
south	15.95 (1.75)	-1.03 (0.44)	1.15 (0.40)	-2.01 (1.30)	-0.21 (0.16)	16.31 (2.87)	6.88 (2.14)	3.50 (1.42)	10.12 (6.21)	5.91 (3.91)
academic	4.91 (0.31)	-5.33 (1.28)	2.49 (0.91)	0.84 (0.23)	-3.85 (2.27)	-13.98 (1.27)	-5.31 (1.12)	-3.23 (0.84)	-10.82 (4.92)	-11.85 (6.07)
technical	6.56 (0.60)	-5.369 (1.36)	4.36 (1.51)	-5.69 (2.51)	-3.96 (2.18)	-15.17 (2.09)	-8.54 (2.85)	-4.873 (1.56)	-4.83 (1.84)	-6.41 (3.92)
boy	-3.38 (0.41)	-1.00 (0.36)	3.00 (1.71)	4.90 (2.97)	4.01 (3.89)	12.85 (2.38)	6.59 (3.22)	5.73 (3.03)	0.76 (0.33)	3.25 (3.40)
stud/teach	-1.49 (0.57)	0.11 (0.50)	0.17 (0.33)	-0.94 (2.39)	0.22 (1.09)	1.22 (0.80)	0.42 (0.96)	1.57 (2.99)	-0.07 (0.23)	1.13 (4.03)
private	-0.92 (0.05)	7.83 (1.03)	30.11 (3.14)	16.96 (3.32)	14.53 (5.01)	-14.85 (0.96)	11.65 (1.48)	9.92 (1.14)	16.60 (3.63)	11.72 (4.21)
sh lab equ	1.56 (0.35)	2.26 (1.47)	0.77 (0.95)	-1.52 (1.70)	-0.59 (1.03)	-6.457 (1.98)	1.151 (0.57)	0.62 (0.54)	1.36 (0.97)	0.90 (1.40)
sh text	4.97 (0.82)	4.12 (1.79)	1.40 (0.78)	2.12 (1.59)	2.31 (2.46)	-7.61 (1.46)	-0.37 (0.14)	3.67 (2.81)	-0.78 (0.47)	1.27 (1.12)
sh comp	-4.19 (0.97)	2.19 (1.01)	0.79 (0.56)	-0.13 (0.16)	0.008 (0.01)	0.41 (0.10)	0.66 (0.38)	-4.33 (3.46)	3.65 (3.44)	0.81 (0.97)
poor libr	1.30 (0.29)	-7.75 (3.74)	0.58 (0.38)	0.03 (0.04)	-0.44 (0.72)	3.96 (1.18)	1.22 (0.43)	-0.88 (0.75)	-1.28 (1.20)	-2.00 (2.53)
sh teacher	1.47 (0.36)	-2.20 (1.38)	2.04 (2.40)	1.49 (1.22)	1.49 (2.60)	1.48 (0.40)	0.43 (0.31)	-1.69 (1.47)	0.11 (0.13)	-0.18 (0.45)
prop cert	0.97 (0.08)	5.73 (0.79)	3.07 (0.74)	-8.46 (1.73)	-5.49 (2.10)	-5.35 (0.45)	-11.55 (0.82)	-17.51 (2.65)	-8.95 (1.70)	-7.79 (2.39)
school size	0.000 (0.01)	-0.004 (1.09)	-0.007 (2.17)	-0.006 (0.94)	-0.004 (2.70)	-0.009 (0.67)	-0.004 (0.68)	-0.008 (2.95)	0.01 (1.85)	-0.001 (0.45)
computer	-0.05 (0.47)	0.05 (2.00)	0.020 (0.84)	0.030 (1.22)	0.008 (0.74)	0.01 (0.21)	-0.002 (0.05)	0.04 (2.00)	-0.04 (1.58)	-0.01 (1.57)
constant	107.31 (3.72)	87.67 (9.42)	78.05 (12.58)	104.47 (19.55)	95.60 (27.53)	105.81 (7.39)	97.756 (7.11)	93.70 (14.64)	91.42 (11.63)	91.88 (23.18)
n	2012	10453	16014	23588	52067	2012	10453	16014	23588	52067

Note: Absolute value of t-statistics in parentheses.

5. Results

Table 5 reports the summary statistics of quantile regressions predictive power of the out-of-sample forecasts. The first three rows of the table consider year 2000 estimated coefficients at the 10th, 50th and 90th quantile regressions to predict years 2003, 2006 and 2009, that is using first wave to predict the following ones. Rows four to six consider year 2003 to predict years 2006 and 2009. In rows seven to nine, the forecast of year 2009 is based on the quantile regression estimated coefficients of year 2006. The comparison of observed and predicted

Table 5. Assessment of OLS and quantile regression out-of-sample forecasts.

year	quantile	MAD(q)			MAD _{OLS}		
		2003	2006	2009	2003	2006	2009
2000	10 th	133.58	120.96	115.84			
	50 th	62.52	59.36	57.82	63.5	59.8	58.7
	90 th	88.65	98.22	96.73			
2003	10 th		90.42	91.65			
	50 th		63.69	64.54		58.8	57.6
	90 th		112.57	116.70			
2006	10 th			98.58			
	50 th			57.60			58.0
	90 th			96.61			

Note: The first three rows use estimates of year 2000 at the 10th, 50th and 90th quantile regression to compute predictions for wave 2003, 2006 and 2009. The second group of three rows utilizes estimates of year 2003, while the third group considers estimates of year 2006 to forecast next waves. The first group of three columns considers quantile regression estimates while the second group of three columns computes predictions based on the OLS estimates.

values in $MAD(\theta)$, reported in the first three columns of this table, shows that the forecasts are more precise at the median than in the tails, the estimated coefficients for year 2000 yield worse forecasts at the lower quantile while the estimates of 2003 worsen at the upper quantile. The second group of three columns of the table report the MAD values computed by OLS, MAD_{OLS} . These results are quite comparable with the median regression values, yielding only minor differences.

In order to analyse in-sample forecasts, consider the model estimated in the entire sample, pooling together year 2000, 2003, 2006, 2009. Last column of **Table 2** reports the results for the median regression, where the estimated coefficients not statistically different from zero are in italics. These variables—shortage of lab equipment, poor library facilities, shortage of teachers and proportion of certified teachers—are dropped from the model, in order to provide a more parsimonious specification. The results for this parsimonious model as estimated at the median regression are reported in the last column of **Table 6**. The elimination of these four variables allows to increase the number of available observations to $n = 56,856$.

The previous analysis shows that regional variable, gender, school track and funding are the main regressors affecting interquantile changes in math scores, and these variables provide the criteria to select the sample. In time series it is easy to exclude the final observations, and the sole decision is on the number of excluded observations q . The data here considered are not time series and, instead of dropping the last observations of the sample, it is more informative to define a criterion to select the q subset. In principle, any explanatory variable could provide a systematic rule to exclude data points, but the analysis is here limited to the four above mentioned variables which proved to cause the greater interquantile differences.

Consider first the regional variable. In this case the score of students of the northern regions, in a sample of size $n_1 = 36,369$, are chosen to forecast southern students performance. This means to select 64% of the sample, in order to forecast the remaining 36% of the data. Data are sorted so that northern students are at the beginning of the sample and the same model is estimated twice: in the full sample and in the subset of northern students. The estimates in the entire sample are reported in the first column of **Table 6**, and the regional variable in this column is dropped from the model⁴. From the estimates in the subset, the n residuals are computed and the difference between total and partial sum of the gradients can be analysed as n_1 converges to n . The difference between $H(n_1, \theta)$ and $H(n, \theta)$ as n_1 approaches n , for each regression coefficient, is reported in **Figure 1** which depicts the pattern for each regression coefficient. In the figure the n_1 subset is to the left of the vertical line, which in the graph marks the passage from the n_1 subset to the q excluded observations. The graph shows that partial and total

⁴The model estimated in the subset will automatically drop the regional variable, since in the subset there are only southern students. In case the regional variable is not excluded from the full sample model, the number of explanatory variables will differ in the two equations, and this would make them unsuitable for the comparison of $H(n_1, \theta)$ and $H(n, \theta)$. The inclusion of the variable clouds the results, since a divergence between the H functions can be due to the different number of coefficients and not, or at least not only, to the predictive capability of the model.

Table 6. Median regression estimates in the entire sample to evaluate predictions.

	sorted by				without sorting
	region	gender	track	funding	
south	-----	-56.22 (40.47)	-57.42 (32.33)	-53.41 (37.26)	-56.46 (41.33)
academic	91.98 (34.83)	86.80 (42.17)	-----	85.69 (39.83)	90.96 (45.07)
technical	52.41 (22.61)	55.17 (30.58)	-----	51.49 (27.28)	53.25 (29.95)
boy	22.53 (12.68)	-----	16.37 (9.24)	22.68 (15.95)	23.45 (17.21)
student/ teacher	1.94 (6.08)	2.49 (9.98)	6.94 (24.23)	1.96 (7.72)	2.71 (11.07)
private	-23.73 (4.90)	-39.67 (10.47)	-24.55 (5.13)	-----	-39.03 (10.53)
short textbook	-12.52 (9.70)	-4.72 (4.73)	-15.48 (12.48)	-4.15 (4.08)	-5.09 (5.19)
school size	-0.018 (6.66)	-0.009 (4.33)	0.001 (0.49)	0.003 (1.58)	-0.01 (4.94)
computers	0.24 (18.03)	0.154 (14.39)	0.114 (8.67)	0.139 (12.50)	0.14 (13.36)
constant	388.55 (91.75)	418.65 (128.31)	432.91 (106.89)	399.97 (111.29)	405.46 (123.73)
n	56856	56856	56856	56856	56856
R ²	0.11	0.16	0.11	0.16	0.17

Note: Absolute value of t-statistics in parentheses. In italics are the estimates with low t-values. The variable selected to define the n_1 subset is dropped from the equation. In the subset it is automatically dropped. If it is included in the full sample equation and excluded in the subset equation, the difference between total and partial sum of gradients could be due to the different number of regression coefficients and not, or not only, to the adequacy of the model.

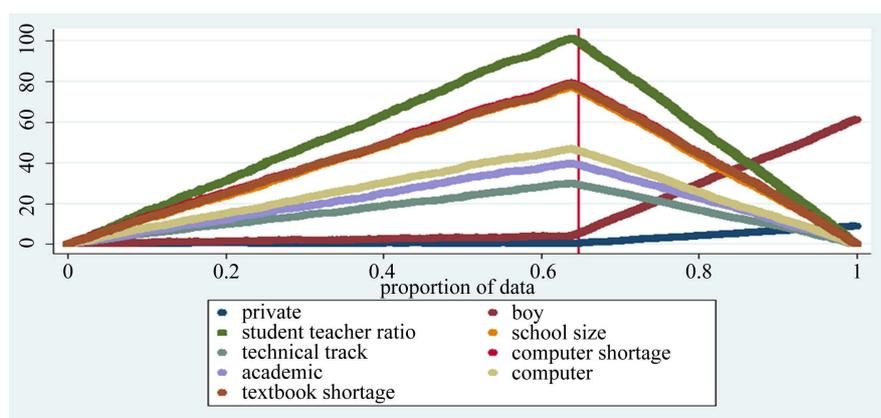


Figure 1. The graph compares partial and total sum of the gradient of each estimated coefficient in case northern student performance is considered to forecast southern student performance for year 2000, 2003 2006, and 2009 pooled together. This leaves out to predict about 64% of the sample. The differences to the right of the vertical bar, which signals the predicted observations, converge to zero for all coefficients but those of funding and gender, which are not good predictors of math score of southern students, data OECD-PISA.

sum of the gradient for the regression coefficients, after reaching a peak at the end of the n_1 subset, converge toward zero. The exceptions are gender and funding, which diverge after the vertical bar signalling the beginning of the q subset. This implies that the performance of northern boys does not predict well southern girl scores and that northern student are not good predictors for southern student enrolled in private schools.

Next the gender variable is selected to define the n_1 subset. Boys' scores in math test are used to forecast girls' performance. Gender is a relevant explanatory variable and this criterion splits in half the sample, excluding $q = 28540$ observations, which is about 50% of the sample. The second column of **Table 6** presents the estimates in the entire sample, where with respect to the parsimonious model the gender variable has been excluded⁵. Boys math scores are sorted at the beginning of the sample. The predictions of girl scores based on boys performance yield an increasing difference between $H(n_1, \theta)$ and $H(n, \theta)$ up to n_1 , and then a decreasing difference after the vertical bar, as can be seen in **Figure 2**. This occurs for all gradients but the regional and the funding coefficients, thus implying that setting aside gender differences in performance, region and funding make a big difference and cause poor forecasts.

The third column of **Table 6** presents the estimated equation excluding academic and technical track variables, where data have been sorted to have scores from the vocational track at the beginning of the sample⁶. The same regression is estimated in a subset of size $n_1 = 12,387$ which considers the vocational schools only. The set of n residuals from this equation are computed in order to evaluate predictions for the remaining 78% of the sample, which represents students enrolled in non-vocational tracks. **Figure 3** presents the comparison of total and partial sum of the gradients as $n_1 \rightarrow n$. In this graph, the discrepancy between the two sets of gradients goes to zero as n_1 increases, with the sole exception of the regional and funding coefficients. The scores of students enrolled in vocational schools are not good predictors of southern students' performance and of private schools learners' scores.

Finally, the same approach is implemented considering the public school sector to forecast the performance of students enrolled in private schools. The number of excluded observations in this experiment is quite small, approximately 3.6% of the sample. The estimates for the entire sample are presented in the last column of the table—headed “funding”⁷. It turns out that students of public schools well predict math scores of both public and

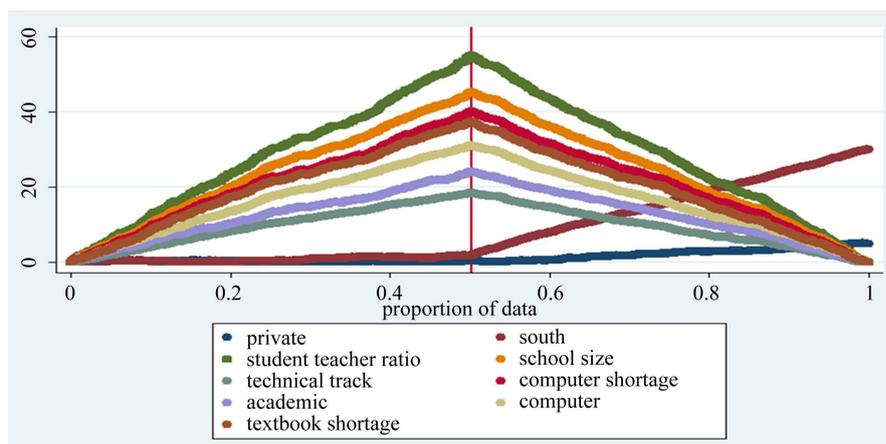


Figure 2. The graph depicts the comparison between partial and total sum of the gradient of each estimated coefficient in the regression considering only boys to forecast girls performance in year 2000, 2003, 2006, 2009 pooled together. This leaves out to predict 49.08% of the sample. The patterns to the right of the vertical bar, which signals the predicted observations, decline toward zero for all the regression coefficients with the exception of the regional and private variables, which do not help in predicting math score of girls, data OECD-PISA.

⁵The model estimated in the subset will automatically drop the gender variable, since in the subset there are only boys. Therefore, the gender variable is dropped from the full sample equation as well, otherwise the number of explanatory variables in the two models would differ and the comparison of $H(n_1, \theta)$ and $H(n, \theta)$ would be inappropriate.

⁶Once again the exclusion of the academic and the technical variable from the parsimonious model estimated in the entire sample is needed in order to have the same number of variables in the two different estimates of the model, so that $H(n_1, \theta)$ and $H(n, \theta)$ are comparable.

⁷The variable private has been excluded from the parsimonious equation estimated in the entire sample, in order to accurately compare $H(n_1, \theta)$ and $H(n, \theta)$.

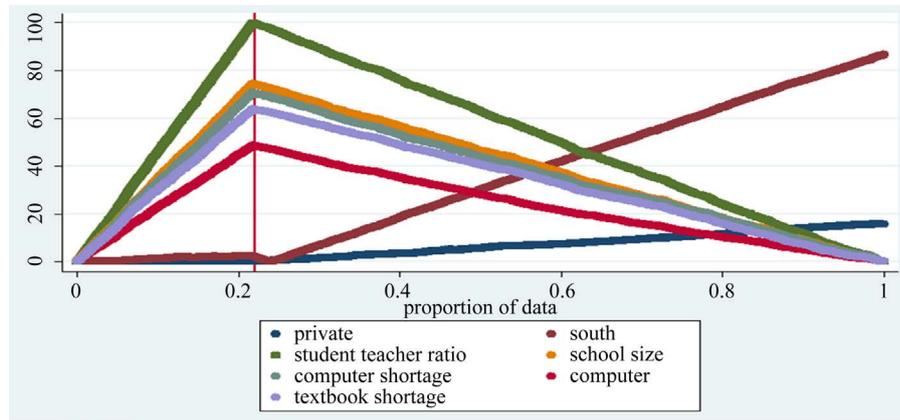


Figure 3. The graph depicts the comparison between partial and total sum of the gradient of each estimated coefficient in the regression considering vocational schools to forecast performance in academic and technical track schools for the years 2000, 2003, 2006, 2009 pooled together. This leaves out to predict 78% of the sample. The patterns to the right of the vertical bar, which signals the q predicted observations, decline toward zero for all the regression coefficients with the exception of the regional and private variables, data OECD-PISA.

private schools with the exception of the regional and gender variables. That is, math scores of public schools students are poor predictors of southern students and of girls performance. The graph in **Figure 4** shows the divergence between total and partial sum of the gradient, and the regional and the gender variable diverge after the vertical line marking the q excluded observations, while the discrepancy between $H(n_1, \theta)$ and $H(n, \theta)$ in all the other coefficients goes to zero after the vertical line.

Summarizing, **Table 6** considers region, gender, track and funding as sample selection rules to evaluate in-sample forecasts, since in the interquantile difference analysis these variables cause the greater changes in the estimated coefficients at the selected quantiles. The first three variables, region, gender and track, exclude about 50%, 64% and 78% of the observations, while funding excludes a very small number of data, only 3%. This is to say that the results do not depend upon the number of excluded observations. On the overall, the graphs show that the coefficients causing the greatest divergence between total and partial sum of the gradient are region, gender and funding, and their impact is mutually influenced: when selecting private schools, the regional and the gender variables are unreliable predictors; when selecting the northern region, gender and funding are the most diverging terms; when gender and school track provide the sample selection rule, both region and funds are poor predictors.

6. Conclusions

The math scores of the OECD-PISA test have been investigated at different quantiles, to analyse the impact of school resources on students' proficiency. School attainment is related to differences in economic growth, while mathematical skill is considered a good proxy of labour force quality. These links are generally discussed in terms of average values while quantile regressions also allow to investigate the tails of the conditional distribution. The estimated coefficients do differ across quantiles, which implies that the tails of the math score distribution behave quite differently from the central model as computed by OLS. Thus, the analysis of the model on average is not very helpful when the aim is to improve one or both tails.

The forecast ability of the model has been considered. The literature on quantile regression predictions is not very extensive. The above analysis considers the median absolute deviation, $MAD(\theta)$, computed as the difference between predicted and observed data to evaluate quantile regression forecasts. The $MAD(\theta)$ results show that the predictive power of quantile regressions improves at the median for out-of-sample forecasts, and that forecasts computed at the median are comparable with OLS predictions.

To evaluate in-sample forecasts some data points are excluded from the sample. The previous sections consider the exclusion based on a specific selection rule, related to the variables of the model. Four different criteria to select the sample have been considered: southern students to predict students living in the rest of the country;

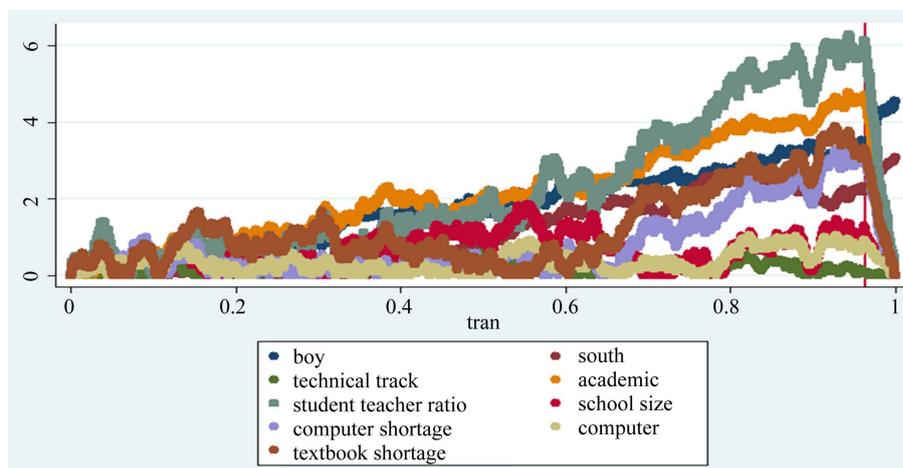


Figure 4. The left hand side of this figure depicts the comparison between partial and total sum of the gradients of each estimated coefficient in the regression considering public schools to forecast private schools for the years 2000, 2003, 2006, 2009 pooled together. This leaves out to predict 3.6% of the sample. The patterns to the right of the vertical bar point out the predicted observations, and show a discrepancy for the regional and the gender variables, data OECD-PISA.

boys to predict girls performance; vocational track to forecast performance of students enrolled in the other curricula; students enrolled in public school to forecast performance of private schools students. Each of the four criteria selects a different subset. The data are appropriately sorted locating the selected subset at the beginning of the sample. The model is estimated in the full sample and in the subset. The vector of gradients estimated in the subset is compared with its analogue in the full sample. The model yields good in-sample predictions when such a difference, computed individually for each regression coefficient, goes to zero. Indeed, if the gradients are similar, the predictions generated by the selected subset are reliable. Vice versa, when the two gradients are far apart, the subset poorly predicts the excluded data, signalling a significant gap between the two subsets. The diagnostic tool here implemented to check in-sample forecasts shows that, while for most variables the gradient in a subset does not diverge much from its analogue in the full sample, thus granting good predictions, the predictive capability of the regional, gender and funding variables is quite unreliable.

References

- [1] Mayr, A., Horthorn, T. and Fenske, N. (2012) Prediction Intervals for Future BMI Values of Individual Children—A Non-Parametric Approach by Quantile Boosting. *BMC Medical Research Methodology*, **12**, 1-13. <http://dx.doi.org/10.1186/1471-2288-12-6>
- [2] Friederichs, P. and Hense, A. (2007) Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, **135**, 2365-2378. <http://dx.doi.org/10.1175/MWR3403.1>
- [3] Barro, R. (1997) *Determinants of Economic Growth: A Cross-Country Empirical Study*. MIT Press, Cambridge.
- [4] Mankiw, N., Romer, D. and Weil, D. (1992) A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics*, **107**, 407-437. <http://dx.doi.org/10.2307/2118477>
- [5] Hanushek, E. (2006) School Resources. In: Hanushek, E.A. and Welch, F., Eds., *Handbook of Economics of Education*, Vol. II, North Holland, Amsterdam.
- [6] Hanushek, E. and Woessmann, L. (2011) The Economics of International Differences in Educational Achievement. In: Hanushek, E.A., Machin, S.J. and Woessmann, L., Eds., *Handbook of Economics of Education*, Vol. III, North Holland, Amsterdam.
- [7] Koenker, R. and Bassett, G. (1978) Regression Quantiles. *Econometrica*, **46**, 33-50. <http://dx.doi.org/10.2307/1913643>
- [8] Koenker, R. (2005) *Quantile Regression*. Cambridge University Press, Cambridge. <http://dx.doi.org/10.1017/CBO9780511754098>
- [9] Koenker, R. and Machado, J. (1999) Goodness of Fit and Related Inference for Quantile Regression. *Journal of the American Statistical Association*, **94**, 1296-1310. <http://dx.doi.org/10.1080/01621459.1999.10473882>

- [10] Davino, C., Furno, M. and Vistocco, D. (2014) *Quantile Regression: Theory and Applications*. Wiley, Hoboken.
- [11] Hao, L. and Naiman, D. (2007) *Quantile Regression*. Sage, London.
- [12] Qu, Z. (2008) Testing for Structural Change in Regression Quantiles. *Journal of Econometrics*, **146**, 170-184.
<http://dx.doi.org/10.1016/j.jeconom.2008.08.006>
- [13] Oka, T. and Qu, Z. (2011) Estimating Structural Changes in Regression Quantiles. *Journal of Econometrics*, **162**, 248-267. <http://dx.doi.org/10.1016/j.jeconom.2011.01.005>
- [14] Xiao, Z. (2009) Quantile Cointegrating Regression. *Journal of Econometrics*, **150**, 248-260.
<http://dx.doi.org/10.1016/j.jeconom.2008.12.005>

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or [Online Submission Portal](#).

