Metrika (2002) 55: 17–26

Metrika

Quantile models and estimators for data analysis

Gilbert W. Bassett Jr.[1], **Mo-Yin S. Tam**[1], **Keith Knight**[2]

[1] University of Illinois at Chicago, 601 South Morgan Chicago, IL 60607, USA
[2] University of Toronto, Department of Statistics, 100 St. George St., Toronto, Ont. M5S3G3, Canada

Abstract. Quantile regression is used to estimate the cross sectional relationship between high school characteristics and student achievement as measured by ACT scores. The importance of school characteristics on student achievement has been traditionally framed in terms of the effect on the expected value. With quantile regression the impact of school characteristics is allowed to be different at the mean and quantiles of the conditional distribution. Like robust estimation, the quantile approach detects relationships missed by traditional data analysis. Robust estimates detect the influence of the bulk of the data, whereas quantile estimates detect the influence of co-variates on alternate parts of the conditional distribution. Since our design consists of multiple responses (individual student ACT scores) at fixed explanatory variables (school characteristics) the quantile model can be estimated by the usual regression quantiles, but additionally by a regression on the empirical quantile at each school. This is similar to least squares where the estimate based on the entire data is identical to weighted least squares on the school averages. Unlike least squares however, the regression through the quantiles produces a different estimate than the regression quantiles.

Key words: Quantile Models, Regression Quantiles, Robustness Student Achievement

1 Introduction

Quantile regression represents an extension of traditional estimation methods that allows for distinct quantile effects; see Koenker and Hallock (2001). The quantile model posits the θ^{th} quantile of y conditional on x to be, $Q(\theta|x) = \alpha(\theta) + x\beta(\theta)$, $0 < \theta < 1$. If $\beta(\theta)$ is a constant β, the model reduces to the standard conditional expectation model, $E(y|x) = \alpha + x\beta$, with constant vari-

ance errors. When $\beta(\theta)$ depends on θ, the model allows the distribution of y to depend on x in different ways at different parts of the distribution.

The traditional linear model can be viewed as a summary of all the quantile effects; that is, $\int Q(\theta|x)\,d\theta = E(y|x)$. Under this interpretation, traditional analysis loses information due to its aggregation of possibly disparate quantile effects. Many different quantile paths, for example, can lead to $\beta_k = 0$. On the one hand, $\beta_k = 0$ can mean x_k does not matter – does not affect the distribution of y. But it can also mean there are important but compensating quantile effects relating y and x. In the latter case the single β statistic obscures information about quantile effects. This is especially important when scientific interest concerns differences in the way regressors affect different parts of the distribution. The details provided by the quantiles discriminate between what would be otherwise identical situations.

While quantile regression and robust estimation are concerned with different aspects of data analysis, they have the shared objective of uncovering relationships missed by traditional data analysis. The robustness criterion translates into estimates that are unaffected by a small fraction of the data, and sampling distributions that stay good when hypothesized models are only approximately valid. Robustness is important because (i) approximate validity is most realistic, and (ii) classical methods, optimal at the hypothesized model, tend to be far from optimal under slight departures from the given model. Robust estimation is designed to deal with mistakes due to discrepant data. Quantile regression is concerned with mistakes due to summarizing potentially disparate quantile effects into a single, potentially misleading, representation of the way y and x are related.

In this paper quantile regression is used to estimate the relationship between student achievement and high school characteristics. (For a financial application of quantile analysis see Bassett and Chen (2000); see, too, references in Koenker and Hallock (2001)). The importance of school characteristics on student achievement has been traditionally framed in terms of the effect on the expected value. With quantile regression the impact of school characteristics is allowed to differ at the mean and various quantiles of the conditional distribution. The quantile estimation allows detection of school effects that would be missed by standard analysis. Recently, quantile regression has been used by Eide and Showalter (1998) to consider how student scores changed over time as a function of school quality, and Levin (2001) considered peer effects and class size.

Past research on school inputs and outputs as measured by ACT scores has often shown negligible effects. (The ACT is a general achievement test taken by US high school students. Either the ACT or the related SAT exam is required by almost all US colleges and universities). Class size, expenditures per student, and other policy variables have been found unimportant after controlling for social and economic differences. Hanushek (1998) and more recently, Hoxby (2000), for example, find little evidence that reduction in class size leads to improvement in student achievement. Hanuschek, Rivkin, and Taylor (1996) and Hanushek and Somers (1999) show that school expenditures do not make a significant difference on student performance. Hanushek, Kain, Rivkin (1999) also find little evidence that higher teacher salaries result in higher quality teachers and conclude that higher expenditures on teachers do not matter. Since these analyses were based on estimates of mean effects, they do not rule out that the school inputs matter in selected parts of the achievement distribution.

We consider ACT scores for all Illinois students in 1996 along with associated school and demographic characteristics. Regression quantiles on the entire data set provides one set of estimates. Since the design consists of multiple responses (individual student ACT scores) at fixed explanatory variables (school characteristics) the quantile model can be also estimated by a regression on the empirical quantile at each school. This is similar to least squares where the estimate based on the entire data is identical to the weighted estimate based on just the school average ACT. This identity for least squares does not however extend to quantile estimates.

The next section presents the quantile model and describes quantile estimation methods. Data and variables are described in Section 3. Comparison of estimates in Section 4 focuses on school characteristics where the quantile and mean results differ.

2 Quantile regression

2.1 Model

Let y_{ij} denote the j^{th} student's ACT score at school i where $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, and let x_i denote a vector of characteristics at the i^{th} school. The proportion of observations at each school is $\lambda_i = n_i/N$, where $N = \sum_{i=1}^{n} n_i$. The y_{ij} are assumed to be independent with distribution, $F(z|x_i)$.

With the standard linear expectation model, student scores are assumed to have an expected value that is a linear function of school characteristics,

$$E(y_{ij}|x_i) = \alpha + x_i\beta. \tag{1}$$

Conditional on x_i, the β coefficient describes the way school characteristics affect average ACT scores.

For the linear quantile model it is the θ^{th} quantile of y_{ij} that depends on school characteristics. That is, let the quantile function or "inverse" of $F(z|x_i)$ be denoted by $Q_i(\theta) = Q(\theta|x_i)$, $0 < \theta < 1$. The linear quantile model specification entails, $Q(\theta|x_i) = \alpha(\theta) + x_i\beta(\theta)$.

When $\beta(\theta)$ does not depend on θ, the quantile model reduces to the standard expectation model with constant variance errors. When $\beta(\theta)$ depends on θ, the model specifies a form of variation such that the quantiles of y_{ij} vary with x_i. This includes heteroscedastic models in which the variance depends on independent variables. But the quantile model is distinct from heteroscedastic specifications. It is more general in not restricting distributions to differ only in their second moment. Indeed, the quantile model allows for the possibility of homoscedasticity even though $\beta(\theta)$ depends on θ (so error distributions are not identical). With the quantile model ACT scores can be influenced by school characteristics in different ways at different parts of the distribution.

2.2 Estimation

Given (y_{ij}, x_i), the quantile model can be estimated by regression quantiles, which are defined by the minimization problem,

$$\beta^*(\theta) = \arg\min_b \sum_{i=1}^{N} \sum_{j=1}^{n_i} w_i[\theta|y_{ij} - x_i b|^+ + (1-\theta)|y_{ij} - x_i b|^-] \tag{2}$$

where the weights w_i are introduced to account for different school variability and the different number of observations at each school.

Since y_{ij}, $j = 1, \ldots n_i$ depends only on x_i, the school design is of interest because it admits other quantile estimators. By analogy with least squares, estimates can be computed using all the data, or alternatively, they can be computed by collapsing scores at each school into a single statistic (the mean, median, or .9 quantile). With least squares the estimate based on all the data is the same as the weighted least squares estimate through the school means. The linearity of the expectation and least squares means the β for the school average, y_i, is the same as the β for individual students: $E(y_{ij}|x_i) = E(\bar{y}_i|x_i)$. In contrast, the quantile $\hat{Q}_i(\theta)$ has a quantile function whose θ^{th} quantile is not the same as $Q_i(\theta)$. Hence, for the quantiles, different estimates are produced by the two methods.[1]

To compare the different quantile estimates it is useful to express the data in terms of the empirical distribution at each school,

$$\hat{F}_i(v) = n_i^{-1} \sum_{j=1}^{n_i} I[y_{ij} < v] \tag{3}$$

and let $\hat{Q}_i(\theta)$ be the associated empirical quantile function. Notice that $\hat{Q}_i$ is an ordinary empirical quantile and hence is asymptotically normal with mean, $\alpha(\theta) + x_i\beta(\theta)$ and variance, $n_i^{-1}\sigma_i^2(\theta)$, where $\sigma_i^2(\theta) = \theta(1-\theta)/(f(Q_i(\theta))^2)$. Hence the data can be written in familiar linear model form as,

$$\hat{Q}_i(\theta) = \alpha(\theta) + x_i\beta(\theta) + \varepsilon_i \quad i = 1, \ldots, n \tag{4}$$

where the error terms are independent and asymptotically normal with mean zero and variance $\lambda_i^{-1}\sigma_i^2(\theta)$. The model is thus seen to be amenable to weighted least squares estimation. Instead of implementing regression quantile estimation on all the (y_{ij}, x_i) data, we can do weighted least squares on the smaller data set, $[\hat{Q}_i(\theta), x_i]$, $i = 1, \ldots, n$.

How does this "regression-through-the-empirical quantiles" estimate compare to regular regression quantiles? What does the usual regression quantile problem look like when expressed in terms of the empirical school distributions? The answer turns out to be given by,

$$\beta^*(\theta) = \arg\min_b \sum_{i=1}^{n} w_i \rho_i(x_i b : \theta). \tag{5}$$

where

$$\rho_i(v) = \int_{-\infty}^{v} \hat{F}_i(t)\, dt - v\theta. \tag{6}$$

[1] These quantile estimates are, in turn, different from those in Bassett and Tam (2000), which were based on available information on the mean and standard deviation of ACT scores at each school.

It can be verified that these ρ_i result in the regression quantiles. This can be most easily verified by taking the (sub)derivative of (2) and noting that the summands reduce to $\frac{\partial \rho_i(v;\theta)}{\partial v} = \hat{F}_i(v) - \theta$.

In the next section we present quantile estimates based on regression quantiles using all the individual data. These are compared to the quantile estimate as in (2) above. The second-order properties of these alternative quantile estimates turn out to be different from one another and provide insights into the workings of quantile estimates; for details see Knight (2001).

3 Data and model

Fifteen variables are included in the model. They are grouped into three categories: (i) resource characteristics, (ii) socioeconomic variables, and (iii) school characteristic variables. Resource characteristics are policy variables. They include percentage of teachers with MA or higher degree, number of students per teacher, school expenditures per student, and school enrollment. Impacts of these variables on student performance have received considerable interest in the literature since they indicate where public policy initiatives might improve student performance. For example, class size that may have little affect on the conditional mean, can be investigated with regression quantiles for its impact on the conditional lower tail of the achievement distribution. Socioeconomic and school characteristics variables are included to control for other factors influencing student performance.

Our analysis uses ACT data of Illinois public high school students in 1996. They were matched to school characteristics data taken from the Illinois Goal Assessment Program (IGAP). Most of the IGAP data contains school-level information for each public school in Illinois. Several variables however are available only at the district level. In most cases there is only one or two schools in a district so that the district values are likely a good proxy for individual schools in the district. An important exception is Chicago where all 62 schools are in the same district. This means that the values of district-level variables are constant across Chicago schools. The data on school characteristics is supplemented with 1990 census data on socioeconomic variables matched to school zip codes. The census data includes information on the percentage of single parent families with children, and the educational achievement of families in the local zip code. Variables and descriptive statistics are listed in Table 1. A dummy variable for Chicago is included in the regressions because Chicago is a single district; there is no variation in its district measured variables. Another reason is that it is likely that Chicago is subject to selection effects. Better students in Chicago are more likely to attend non public schools than better students outside of Chicago.

4 Results

4.1 Least squares compared to regression quantiles

Previous analysis of the impact of policy variables on student performance has examined mean effects. Many past studies have shown negligible effects after controlling for socioeconomic factors and school characteristics. Our least

Table 1. Descriptive statistics

Variables	Data Source	Unit	Mean[2]	St. dev.
DEPENDENT VARIABLE				
ACT (#)	ACT	STUDENT	21.2	4.25
INDEPENDENT VARIABLES				
Resource Characteristics				
Teacher MA Degree %	IGAP	DISTRICT	52.64	19.10
Pupil Teacher Ratio (#)	IGAP	DISTRICT	18.25	2.62
Per Pupil Expenditures ($)	IGAP	DISTRICT	6963	2523
Enrollment (#)	IGAP	SCHOOL	1581	863
Economic Social Characteristics				
Low Income %	IGAP	SCHOOL	19.29	22.74
Single Parent Family	CENSUS	ZIPCODE	0.1847	0.1248
Bachelor and Above	CENSUS	ZIPCODE	0.2353	0.1530
School Characteristics				
White %	IGAP	SCHOOL	74.11	29.28
Black %	IGAP	SCHOOL	13.55	24.20
Asian %	IGAP	SCHOOL	4.18	5.92
Attendance Rate (%)	IGAP	SCHOOL	91.92	4.82
Mobility Rate (%)	IGAP	SCHOOL	14.44	10.03
Dropout Rate (%)	IGAP	SCHOOL	5.05	5.01
ACT of Class %	IGAP	SCHOOL	66.83	14.20
Chicago	IGAP	SCHOOL	0.13	0.34

Table 2. LSE of policy effects (resource characteristics)

Variables	LS Estimate	*t* statistic
Teacher MA Degree %	−0.002	−0.512
Pupil Teacher Ratio	0.02	1.564
Per Pupil Expenditures	−0.000017	−0.58
Enrollment	0.00039	6.417

Table 3. Sign and significance of least squares and regression quantile estimates

Variables	LS	QR				
		0.1	0.3	0.5	0.7	0.9
(Policy Variables)						
Teacher MA Degree %	−	+ *	+ *	− *	− *	− *
Pupil Teacher Ratio (#)	+	− *	+ *	+ *	+ *	+ *
Per Pupil Expenditures ($)	−	− *	− *	− *	−	−
Enrollment (#)	+ *	+ *	+ *	+ *	+ *	+ *

* significant at 5%

squares analysis shows similar results. Table 2 lists the least squares estimates for policy variables. Except for the enrollment variable, none of the least square estimates of the coefficients of the policy variables is statistically significant.

Figure 1 and Table 3 show that quantile regression tells a different story.

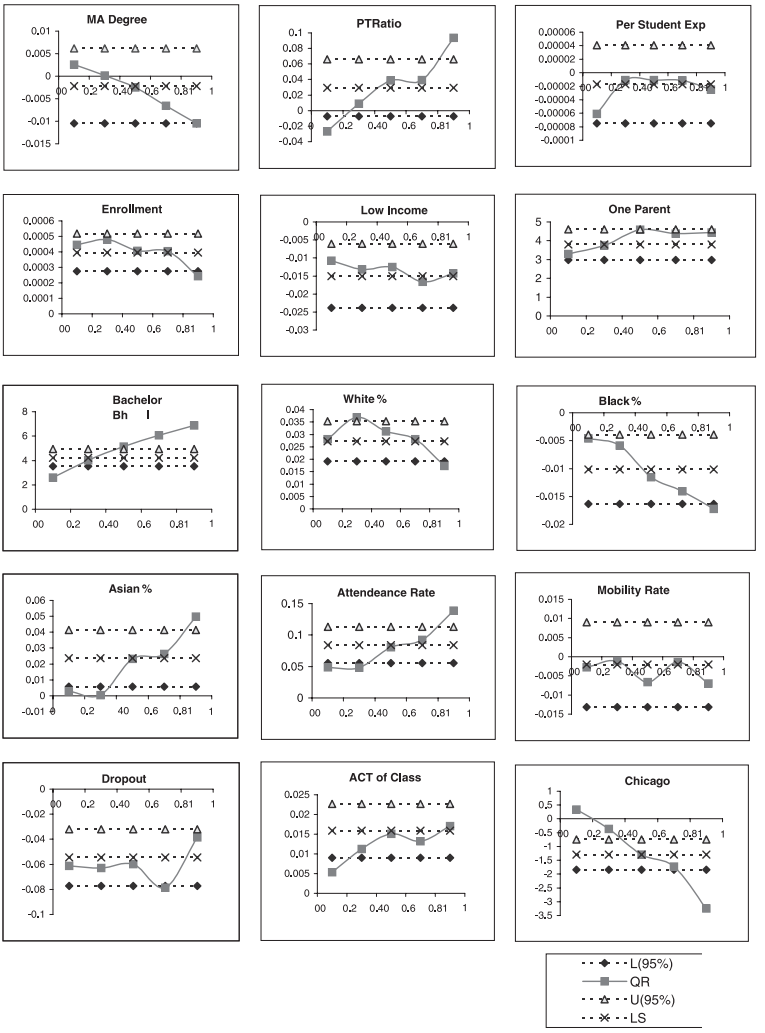

Fig. 1. Comparison of Regression Quantile and Least Squares Estimates.

Each plot of Figure 1 depicts estimates of the coefficient for one variable in the model. The solid line with filled squares shows the five point estimates for the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles. The long-dashed line is the usual least-squares regression estimate of the mean effect. It is bracketed by two short-dashed lines that represent the upper and lower limits of the 90% confidence interval for the ordinary regression estimate.

For three of the policy variables at least one of the quantile estimates lies outside the conditional mean interval. Also, in contrast to the least squares estimates, all the quantile regression effects are significant except for the lower quantiles of the per pupil expenditures variable. Impacts of teacher qualification and pupil-teacher ratio on student performance are significant at all quantile levels though qualitatively different in the two tails of the distribu-

tion. Increasing the percentage of teachers with advanced degrees would help low scorers' performance but harm high achievers. So would a reduction in the pupil-teacher ratio. The positive impact of a higher pupil-teacher ratio on high achievers may incorporate the positive scale effect of larger schools, where a greater variety of higher-level courses are available. Similar reasoning is consistent with the positive impact of total enrollment on student performance at all quantile levels; for similar quantile results on the enrollment effect see; Eide and Showalter (1998).

Finally we note some of the striking differences between quantile regression and least squares for other variables. The quantile estimates are very different than least squares for the Bachelor and the Chicago variables. Parents with Bachelor or higher degree have a much greater positive impact on high achieving students. The Chicago variable displays huge differences across quantiles. At the 0.1 quantile it is close to zero, but becomes very negative at higher quantiles. These differences probably are due to selection effects as top students opt for private schools. Whatever the reason, these different impacts go undetected with least squares analysis.

4.2 Comparison of alternative regression quantile estimates

In the "full" model, quantile estimates are obtained using all 62,839 observations for ACT data. These are the usual regression quantiles. The "Q585" estimates are based on the regression through-the-quantiles at each of the 585 schools. Figure 2 presents a visual comparison of the two alternative approaches. Each plot depicts the quantile regression results of the two approaches for each of the fifteen variables in the model. The solid line with filled squares represents the Full regression estimates. The dotted line with triangles (diamonds) represents the upper (lower) 95% intervals for the "Q585" estimates. The solid line lies within the interval for all variables. This indicates that the "regression through the quantiles" are similar to the usual regression quantiles. Again, this is not too surprising given that the estimates are first-order equivalent.

5 Summary

The relationship between school characteristics and student achievement has traditionally meant the effect on the expected value. This paper has considered a generalization in which characteristics are allowed to differentially affect alternative quantiles of the achievement distribution. Estimates for public high schools in Illinois show instances where the quantile effects differ from the expected value. Thus, covariate influence is sensitive to where on the achievement distribution one is looking. Whereas robustness is concerned with mistakes due to a small fraction of the data, the quantiles deal with the problem of summarizing disparate quantile effects into a single summary measure of influence. One dimensional "influence" is replaced by a multidimensional quantile description of how covariates work, thereby allowing discovery of relationships that would be otherwise missed.

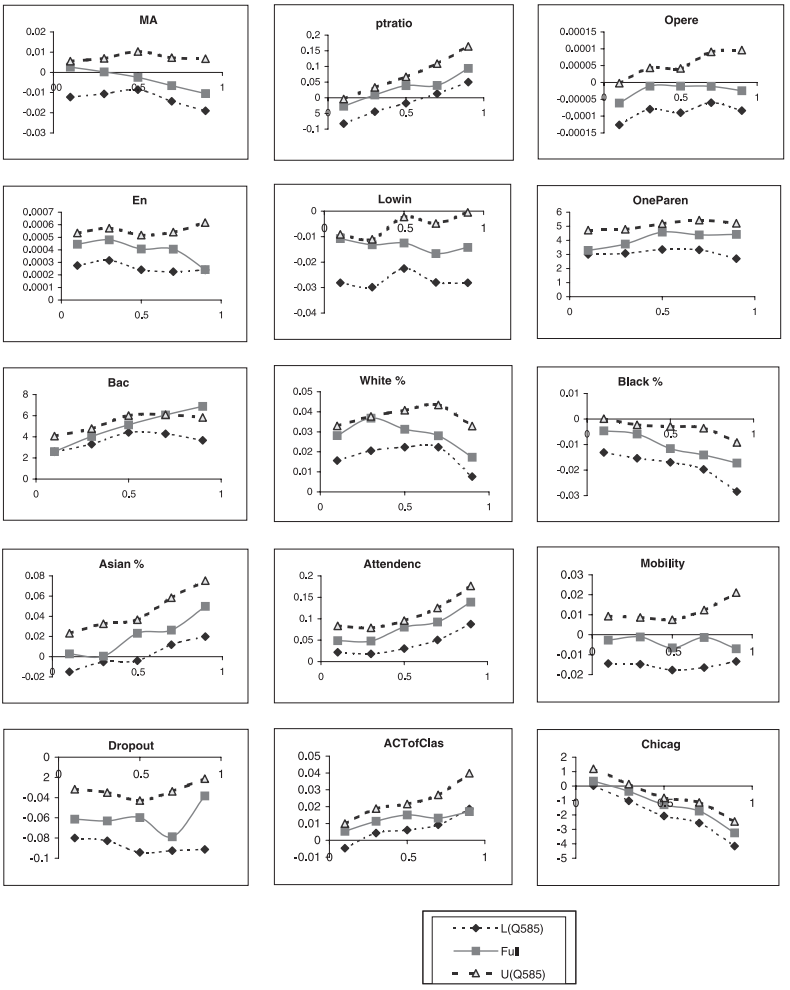

Fig. 2. Regression Quantile compared to upper and lower Q585 Estimates.

References

[1] Bassett GW, Chen H-l (2000) Portfolio style: Return-based attribution using quantile regression. Empirical Economics 26:293–305
[2] Bassett GW, Tam M-Y (2000) School quality and school quantiles. Working paper
[3] Eide E, Showalter M (1998) The effect of school quality on student performance: A quantile regression approach. Economic Letters Volume 58:345–350
[4] Hanushek E (1998) The evidence of class size. Earning and Learning: How Schools Matter, In: Mayer S, et al., editors, Brookings Institution, Washington D.C., 131–168
[5] Hanushek E, Somers J (1999) Schooling, inequality, and the impact of government. Working paper
[6] Hanushek E, Kain J, Rivkin S (1999) Do higher salaries buy better teachers? National Bureau of Economics Research
[7] Hanushek E, Rivkin S, Taylor L (1996) Aggregation and the estimated effects of school resources. Review of Economics and Statistics 78(4):611–627

[8] Hoxby C (2000) The effects of class size on student achievement: New evidence from population variation. Quarterly Journal of Economics, November, 1239–1285
[9] Illinois Goal Assessment Program (1996)
[10] Knight K (2001) Comparing conditional quantile estimators: first and second order considerations. Working paper
[11] Koenker R, Hallock K (2001) Quantile regression: An introduction. Journal of Economic Perspectives forthcoming
[12] Levin J (2001) For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. Empirical Economics 26(1):221–246
[13] U.S. 1990 Census (1990) http://venus.census.gov/cdrom/lookup