Statistics and Its Interface Volume 11 (2018) 541–556

Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on University outcome

Cristina Davino* and Domenico Vistocco

In many real data applications, statistical units belong to different groups and statistical models should be tailored to incorporate and exploit this heterogeneity among units. This paper proposes an innovative approach to identify group effects through a quantile regression model. The method assigns a conditional quantile to each group and provides a separate analysis of the dependence structure inside the groups. The relevance of the proposal is provided through an empirical analysis investigating the impact of students' features on University outcome. The analysis is performed on a sample of graduated students; the degree mark is the response variable, a set of variables describing the students' profile are used as regressors, and the attended School determines the group effects. A working example and a small simulation study are introduced to highlight the main features of the proposed approach.

Keywords and phrases: Quantile regression, Group effects, Statistical models.

1. INTRODUCTION

Many real datasets have a hierarchical or clustered structure, with statistical units grouped at different levels (e.g. students and schools, regions and countries). In such a framework, a statistical model must be tailored to incorporate and exploit the data structure. The analysis of the relationship between a response variable and a set of regressors cannot be carried out by neglecting the membership of the units to the different levels of the hierarchical structure. It is a matter of fact that if two units belong to the same group, the dependence structures of their regression models may be alike.

The present paper refers to the simplest hierarchical structure, which consists of two levels: units (level 1) belong to one of m groups (level 2). The proposed approach aims to estimate group effects in a regression model exploiting quantile regression [11], a method that is able to model the entire conditional distribution of a response variable.

Different approaches have been proposed in the literature to analyze group effects in a dependence model. All of them share the aim of inspecting how the hierarchical data structure affects the impact of the regressors on the dependent variable, although they differ in terms of complexity and ability to detect group effects in depth.

The simplest approach consists of estimating different models for each group, but it obviously does not permit the identification of the impact of the groups on the dependent variable. It also requires ad hoc tools for comparisons of the models estimated on different samples. Fitting a different model for each group may also be inappropriate when groups contain different numbers of units and some groups have very few units. The first issue makes it difficult to use the classical statistical tests to compare models. The latter leads to unreliable estimates [5]. In the quantile regression (QR) framework, the problems deriving from the estimation of separate models are amplified because the comparison must be carried out among models related to the same quantile along the groups and/or among quantiles inside each group.

A second solution exploits the introduction of dummy variables among the regressors to denote group membership [7]. An indicator variable is considered for each level except the reference level. In such a case, the effect of each group is immediately available, but the specific impact of the regressors on the different groups is not available. This effect could be caught up with the inclusion of appropriate interaction terms, but the complexity of the deriving model makes this approach unfeasible. Moreover, the inclusion of interaction terms leads to the estimation of a unique model with residuals that may not be independent [8]. Units belonging to the same group often are often more similar to one another than to units belonging to the other groups. In essence, this approach does not effectively take into account the clustered data structure.

A more widely used solution is provided by multilevel modeling [3, 16, 17], also known as mixed models, hierarchical linear models, nested models and random–effects models, according to the particular field of application. This class of models restricts the analysis of group differences to the mean of the dependent variable, requires classical distributional hypotheses and detects group effects through a single coefficient (the estimation of the between–cluster variability). Recently, many contributions to the literature have extended

[*] Corresponding author.

quantile regression to clustered data. These studies include those published by Koenker [9], Lamarche [14] and Geraci and Bottai [4], although some of these studies apply to longitudinal data. The use of quantile regression overcomes the limitation of classical multilevel modeling, which focuses on the estimate of the conditional mean of the dependent variable. Notwithstanding, such proposals detect the group effects through a single coefficient, in line with multilevel models, and do not provide any details on the dependence structure inside the groups. The use of interaction terms in a QR model could overcome this limit at the cost of increased complexity in the interpretation, as the corresponding coefficients are replicated for each estimated quantile.

Within this framework, the aim of this paper is to introduce a method that can provide a separate analysis of the dependence structure for each group, limiting the number of coefficients that must be estimated, and therefore simplifying its use and interpretation for practitioners. In particular, a quantile regression model is estimated on the whole sample and a conditional quantile is assigned to each group. This approach allows us to detect group effects and easily compare coefficients related to the different groups. The approach is illustrated through an empirical analysis evaluating the effectiveness of the University educational process: data are presented in Section 2, basic notation and a preliminary analysis are described in Section 3, the detection of group effects is carried out using the classical regression approach in Section 4 and the proposed QR approach is presented in Section 5. The potentialities of the proposal are further investigated in Section 6 through a working example on a synthetic dataset and a simulation study. Some concluding remarks and directions for future avenues of research are described in Section 7.

2. THE EFFECTIVENESS OF THE UNIVERSITY EDUCATIONAL PROCESS: DATA DESCRIPTION

The analysis focuses on modeling the final outcome of graduated students in terms of socio–demographic and University experience attributes. The final outcome is measured through the degree mark. As the dependence model can be affected by the particular School students are enrolled in, knowledge of their Schools allows us to consider the School membership as a stratification variable.

The evaluation of the factors influencing the degree mark is based on a random sample of 362 students who graduated from the University of Macerata [2], which is located in the Italian region of Marche. The survey was completed in 2007 and includes students who graduated between 2002 and 2005. The explicative variables included in the model pertain to the student profile. In particular, the following regressors have been considered: gender, place of residence during University education (Macerata and its province, Marche region, outside Marche), course attendance (no attendance, regular), working condition (full-time student, working student), number of years to obtain a degree and diploma mark. The degree mark is measured on a discrete scale ranging between 77 and 109. Students with degree mark equal to 110 and 110 with the 'cum laude' have been excluded from the analysis because a preliminary study revealed that their performance is not affected by the considered features.

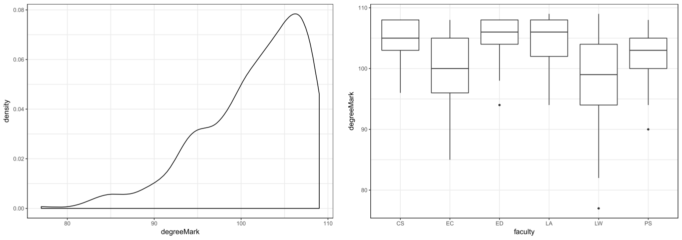

Figure 1. Degree mark density for the whole sample (left) and boxplots according to School membership (right). The degree mark shows a high left-skewed distribution with different intensities among the Schools.

Because the School where students were enrolled can be relevant in determining the final degree mark, it is a natural candidate for the stratification variable in our procedure. The levels of the School variable are: Economics (EC), Law (LW), Liberal Arts (LA), Communication Sciences (CS), Education (ED) and Political Science (PS).

The density plot of the response variable (Figure 1, left) shows the presence of strong left skewness, which also distinguishes the distributions according to School (Figure 1, right). From the boxplots related to the Schools, the differences among the levels of the stratification variable are evident, thus supporting the use of the School to discern group effects. It is worthwhile to notice how the simple analysis of the response variable suggests the presence of two distinct groups: LA, CS and ED, with a strong left skewness, and EC and LW, with less pronounced skewness. The remaining School, PS, shows intermediate behavior.

The main descriptive statistics of the degree mark (Table 1) confirm the asymmetric distribution of the degree mark.

3. QUANTILE REGRESSION: METHODOLOGY AND MAIN RESULTS

Quantile regression, introduced by Koenker and Basset [11], can be considered the extension of ordinary least squares (OLS) to the estimation of a set of conditional quantile functions. QR allows the estimation of the conditional quantiles of a response variable as a function of a set of covariates without requiring assumptions on the error distribution. Although different functional forms can be used, this paper deals only with linear regression models.

Let us consider a vector $\mathbf{y}_{[n]}$ storing the dependent variable and a matrix $\mathbf{X}_{[n\times p]}$ of regressors, where n denotes

Table 1. Descriptive statistics of the degree mark for the whole sample (first column) and according to School (second through seventh). The peculiarities of LW and EC are confirmed by the summary statistics

	total	LA	CS	LW	ED	PS	EC
min	77	94	96	77	94	90	85
I quartile	98	102	103	94	104	100	96
median	103	106	105	99	106	103	100
mean	101.5	105	105	98	105	102	100
III quartile	106	108	108	104	108	105	105
max	109	109	108	109	108	108	108
Frequency	362	71	33	133	37	32	56
% Frequency	100%	20%	9%	37%	10%	9%	15%

the number of units and p the number of regressors. Let the data be partitioned by row, where the partition is determined by a categorical variable (hereafter stratification variable) assuming m groups; the number of units in group g $(g = 1, \dots, m)$ is denoted by n_g; and the total sample size can be expressed as $n = \sum_{g=1}^{m} n_g$.

The QR model for a given conditional quantile θ, with $0 < \theta < 1$, can be formulated as follows:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\beta}(\theta) \tag{1}$$

where $Q_\theta(.|.)$ is the conditional quantile function for the θth quantile. The estimates in QR linear models have the same interpretation as those of any other linear model. Each $\hat{\beta}_j(\theta)$ coefficient represents the rate of change of the θth conditional quantile of the dependent variable per unit change in the value of the jth regressor $(j = 1, \dots, p)$, holding the others constant.

In Table 2, for purely descriptive purposes, the QR coefficients related to the three quartiles and the two extreme quantiles, $\theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$, are shown along with their OLS counterpart (significant coefficients at $\alpha = 0.10$ in bold). The standard errors, used to evaluate the statistical significance of the coefficients, have been estimated using the standard xy–pair bootstrap [15]. The QR coefficients are also depicted in Figure 2. Each panel represents a single regression coefficient, i.e. the intercept and the slopes for the different features of the student profile. The horizontal axis displays the different quantiles, while the effect of each feature holding the others constant is represented on the vertical axis. QR confidence bands (in grey) are obtained through the bootstrap method using $\alpha = 0.1$. The solid lines parallel to the horizontal axis correspond to OLS coefficients, and the dashed lines representing the corresponding confidence intervals using the same significance level.

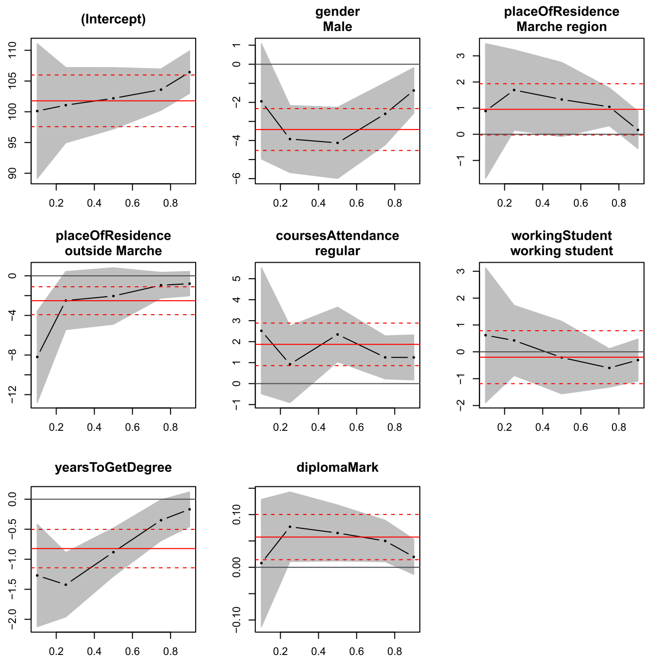

Figure 2. OLS and QR coefficients and related confidence intervals. The horizontal axis displays the different quantiles, while the coefficients are represented on the vertical axis. The shaded region in each subplot shows the confidence band ($\alpha = 0.1$). The lines parallel to the horizontal axis correspond to OLS coefficients and the related confidence intervals are in dashed lines using the same level for α.

The analysis of Table 2 and Figure 2 shows that the effect of the student features on the degree mark is different both

Table 2. OLS (first column) and QR coefficients (from the second to the last column) for five distinct conditional quantiles: $\theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Significant coefficients at α = 0.10 are shown in bold. The impact of the regressors varies across the different parts of the degree mark distribution

	OLS	θ=0.10	θ=0.25	θ=0.50	θ=0.75	θ=0.90
(Intercept)	**101.78**	**100.12**	**101.08**	**102.19**	**103.60**	**106.45**
Gender = Male	**-3.42**	-1.94	**-3.92**	**-4.12**	**-2.60**	**-1.38**
Place of residence = Marche region	0.95	0.89	**1.69**	1.33	**1.05**	0.17
Place of Residence = outside Marche	**-2.51**	**-8.19**	-2.50	-2.04	-0.95	-0.79
Courses attendance = regular	**1.87**	2.52	0.92	**2.34**	**1.25**	**1.25**
Working student = yes	-0.20	0.62	0.42	-0.21	-0.60	-0.31
Numbers of years to get a degree	**-0.82**	**-1.27**	**-1.42**	**-0.88**	-0.35	-0.17
Diploma mark	**0.06**	0.01	0.08	0.07	**0.05**	0.02

in sign and in quantity among the considered features. Moreover, the strong skewness of the response variable engenders differences among the conditional quantile estimates for each feature. Gender and residence during University education have great influence on the lowest quantiles of the distribution: males and residents outside the Marche region show negative coefficients. A foreign experience positively influences the degree mark. This effect decreases in the higher part of the distribution, indicating that very good students are less influenced by their University experiences abroad. Working students are less likely to get high degree marks (the OLS coefficient is equal to –0.50), but the QR results show how this effect becomes relevant at the highest part of the distribution and is negligible elsewhere. All the coefficients of the variable numbers of years to get a degree are negative, particularly for the lowest quantiles. It is worth noticing that only the coefficients related to the 0.25 and 0.5 quantiles are significant. The diploma mark always has a positive effect, but its value is very low for successful students. In the higher part of the response variable distribution, the only positive effect is provided by regular course attendance, while residence outside Marche negatively influences the final degree mark. Regressors do not play any effect on the 90th percentile of the conditional distribution of the degree mark, which is a sign that the highest performances are related to other student features.

Model (1) does not evaluate the difference in the dependence structure with respect to group membership. Two units sharing the same level of the stratification variable could indeed share a more similar dependence structure than two units belonging to different groups. In the following, a strategy aiming to evaluate group effects through the assignment of a particular quantile to each group is introduced.

4. MAIN REGRESSION APPROACHES TO DETECT GROUP EFFECTS

In the framework of the classical regression, the evaluation of the role of the School in predicting the final degree mark can be carried out by exploiting one of the several approaches provided by the scientific literature to detect group effects in a dependence model, as briefly introduced in Section 1.

1) *Different models for each School*

A first attempt consists of estimating different models for each School. Figure 3 shows a bar chart for the intercept and for each regressor, where the height of each bar is equal to the coefficient obtained from a model estimated on the whole sample (first bar of each chart) or separately on each School subsample (from the second to the last bar). Black bars represent coefficients significant at $\alpha = 0.10$. Using the typical dummy coding for the categorical variables, and excluding a level for each variable, it follows that the intercept measures the effect on the degree mark of the reference student with the following features: female, living in Macerata and its province, no course attendance, no foreign experience and full-time student (no working student). Figure 3 easily shows if and how the impact of each regressor changes among the Schools and with respect to the whole sample. However, this interpretation must be done with great caution because the coefficients related to each School are separately estimated

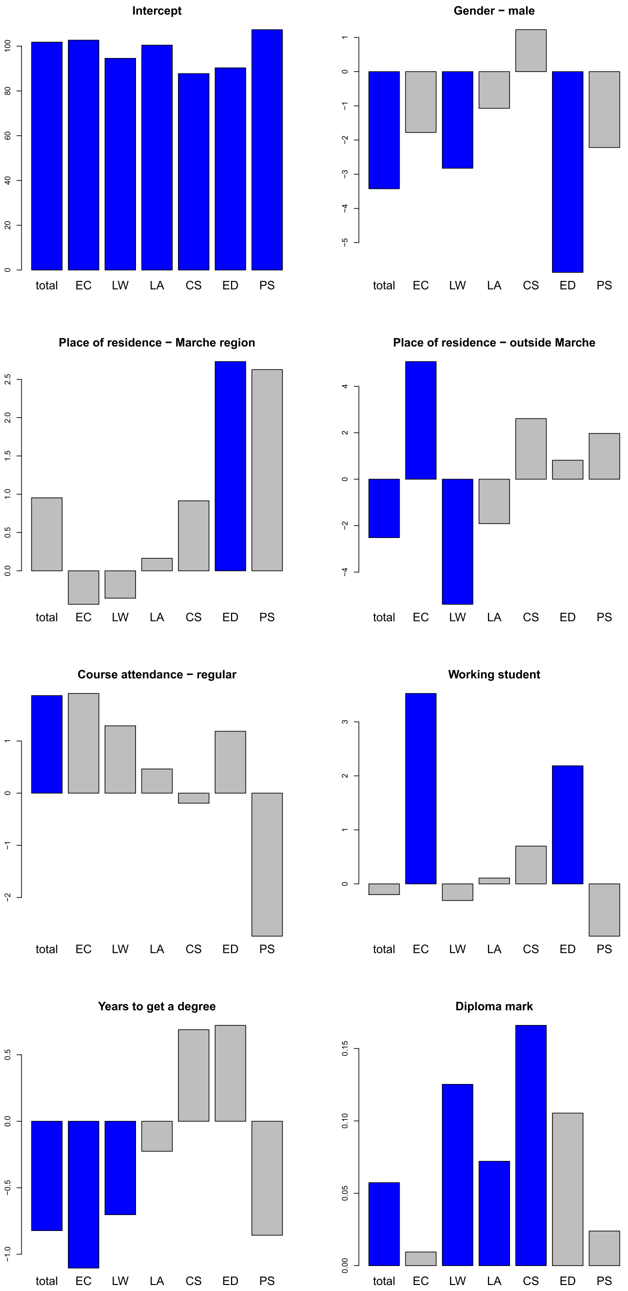

Figure 3. OLS coefficients obtained from separate regressions on the whole sample (first bar of each plot) and for each School sub-sample (from the second to the last bar of each plot). The direct comparison of the coefficients is risky as they derive from the estimation of different models for each School. Black bars depict significant coefficient.

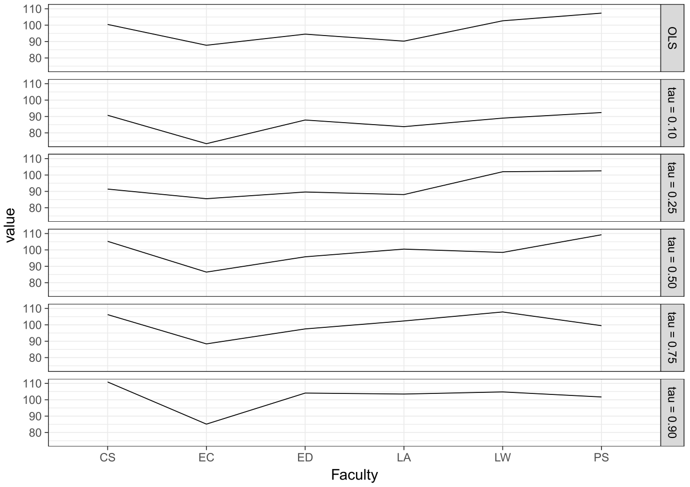

Figure 4. OLS (top panel) and QR coefficients for $\theta = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ *(from the second to the last panel) including School as a dummy variable in the model. The effect played by each School varies in the lowest part of the degree mark distribution.*

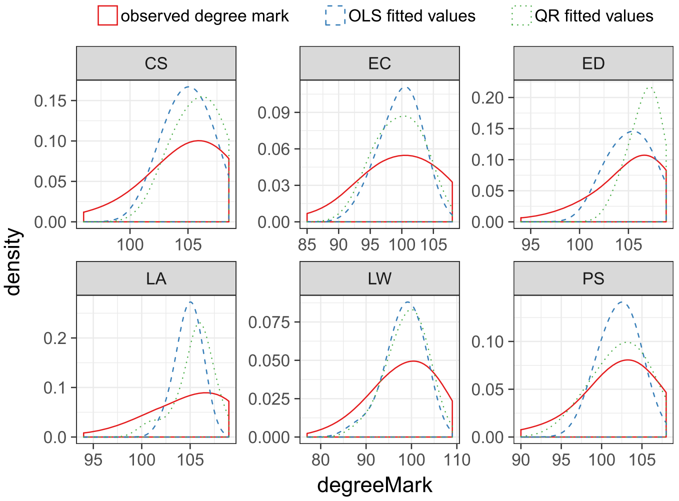

Figure 5. Distribution of the observed and fitted (OLS and QR) degree mark for each School.

and the groups are not of equal size, as shown in the last two rows of Table 1. Finally, this approach becomes almost unfeasible when the QR model is used because the effect of the regressors has to be explored at several quantiles of interest of the conditional distribution of the degree mark.

2) *Model with dummy variables representing each School level*

With the introduction of a dummy variable for each School, OLS coefficients indicate the impact played by each School on the conditional average and QR results indicate the same effect on the considered conditional quantiles of the degree mark. To ease interpretation, the model is estimated without intercept.

Figure 4 shows for each School (horizontal axis) the effect it plays on the degree mark, setting the other regressors to their reference values. The different panels refer to the OLS results (top panel) and the QR results for $\theta = \{0.1, 0.25, 0.50, 0.75, 0.90\}$ (from the second to the bottom panel). Inspection of the obtained estimates (see Table 3) highlights greater differences among the Schools at the lowest quantiles, suggesting that the performance of the best students, i.e. with the highest degree marks, is less affected by the School of origin. With respect to the use of single models for each School, this approach allows estimation of the model only once on the whole sample. Nonetheless, the School effect is captured only by the coefficients associated to the dummy variables. Therefore, the model is not able to capture the different impact of each regressor in the groups.

3) *Model with the group variable and all the interactions*

To capture the group effect, the group variable and all the interactions among the groups and the regressors are considered in the model (hereafter, group interaction model). Because the analysis involves six Schools and six regressors, with a total of 13 categories, the group interaction model includes 48 regressors. A classical OLS regression as well as a QR, using a dense grid of quantiles (from 0.1 to 0.9 with step equal to 0.1), has been carried out. Albeit such an approach allows to take into account the group effect with respect to all regressors, its main remark is related to the huge number of coefficients to be interpreted. This number increases for QR, because the coefficients are estimated in correspondence to each quantile of interest.

Focusing only on prediction, both OLS and QR provide quite similar results, as shown in Figure 5, where the observed degree mark and the fitted values deriving from OLS and QR regression are plotted for each School. QR out–performs OLS with respect to the fitting capability: almost all the corresponding BIC values (with the exception of $\theta = 0.1$ and $\theta = 0.2$) are lower than the BIC obtained for OLS regression (see Figure 6, where the horizontal line depicts the BIC for OLS).

Moreover, the added value of QR in analysing the impact of the regressors on the whole conditional distribution of the response is particularly useful for such data, where the dependent variable is highly skewed. For example, consider two regressors including the interaction terms: CS:Male and CS:CourseAttendanceRegular, i.e. the coefficients linking the effect of the School in Communication Sciences with the gender and the course attendance, respectively. Figures 7 and 8 depict the corresponding QR coefficients: the horizontal axis displays the different quantiles, while the effect of each feature holding the others constant is represented on the vertical axis. The dotted line parallel to the horizontal axis corresponds to OLS coefficient. It is evident how the analysis can be limited or even misrepresented when it is confined to the interpretation of the

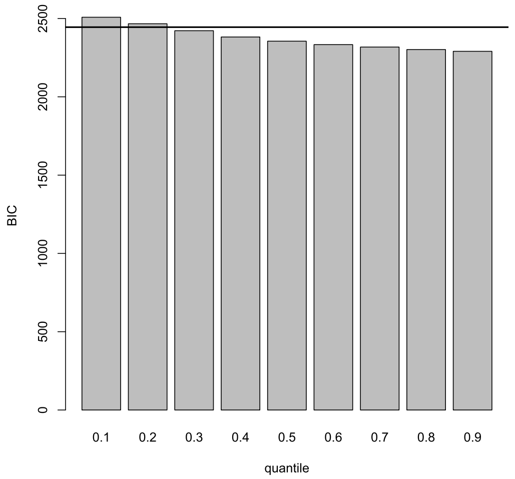

Figure 6. BIC values related to the group interaction model estimated using OLS (horizontal line) and QR (bars) at nine different quantiles.

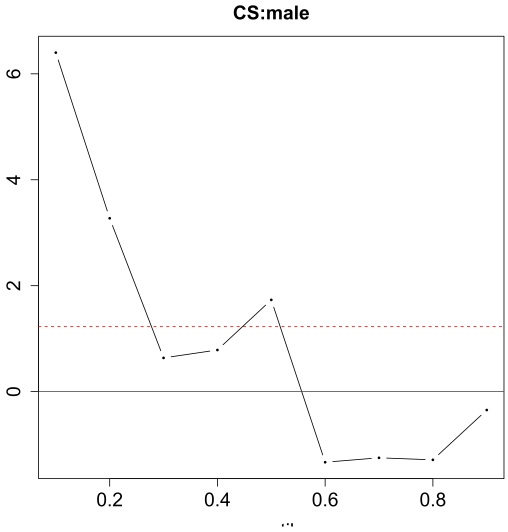

Figure 7. OLS and QR coefficients for the CS:male regressor.

effect played by the two regressors on the conditional average distribution of the degree mark. In case of CS:Male, the OLS coefficient provides a positive effect of this regressor on the degree mark, but Figure 7 shows that this effect decreases moving from lower to higher quantiles, becoming negative after the conditional median. The opposite happens for the CS:CourseAttendanceRegular regressor, which is negative on average but increasing and positive after the median. In fairness, the interpretation of all the coefficients of the group interaction model for each quantile of interest becomes unfeasible. An innovative approach to overcoming this limit is proposed in the next section. It aims to take into account the effect played by the group membership on the degree mark for each regressor but without penalizing the informative ability of the obtained results.

4) *Multilevel model*

Finally, multilevel models are a natural solution given the hierarchical structure of the analysed data: students, level 1 units, are enrolled in different Schools, level 2 units.

Following common practice, the null model (from now Model 0) with no independent variable is estimated. It is useful for obtaining estimates of the residuals and intercept variance when only the clustering by School is considered. It is a reference model for discussion and comparisons with more complex models. Results from Model 0, in particular AIC and BIC values along with intercept and residual standard deviations, are given in the first column of Table 4. A second step consists of estimating the random intercept model (from now Model 1) that contains varying intercepts but constant slopes across the level 2 units. It assumes that the Schools have different averages of the response variable

Table 3. OLS (first column) and QR coefficients (from the second to the last column) for five distinct conditional quantiles: $\theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ using the Schools as dummy regressors in the model. Significant coefficients at α = 0.10 are shown in bold. The effect played by each School varies in the lowest part of the degree mark distribution

	OLS	θ=0.10	θ=0.25	θ=0.50	θ=0.75	θ=0.90
CS	**100.5**	**93.72**	**97.81**	**100.2**	**103.08**	**105.26**
EC	**96.65**	**88.03**	**91.44**	**96.49**	**101.25**	**103.55**
ED	**100**	**93.52**	**97.59**	**100.18**	**102.62**	**104.94**
LA	**100.64**	**93.86**	**98.02**	**100.37**	**103.41**	**105.43**
LW	**95.01**	**85.41**	**90.03**	**95.34**	**99.46**	**102.72**
PS	**98.69**	**90.21**	**95.24**	**99.43**	**101.72**	**103.57**
Male	**-2.53**	**-4.03**	-2.06	**-3.18**	**-2.51**	**-1.19**
Residence in Marche region	0.23	-0.9	0.79	**0.39**	0.48	0.09
Residence outside Marche	**-2.12**	-3.07	-1.22	**-2.37**	-0.89	-0.71
Regular course attendance	0.89	2.14	0.13	0.63	0.42	**0.80**
Working student	0.13	0.69	0.96	0.53	0.07	-0.15
Years to get a degree	**-0.55**	-0.38	**-0.72**	**-0.59**	-0.24	-0.17
Diploma marks	**0.09**	**0.10**	**0.12**	**0.11**	**0.07**	**0.05**

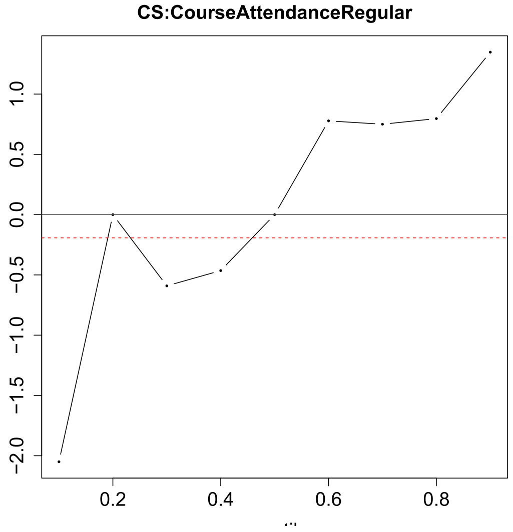

Figure 8. OLS and QR coefficients for the CS:CourseAttendanceRegular regressor.

Table 4. Performance measures and random effect in Model 0 *and Model* 1*. The random intercept model (Model* 1*) provides intercept and residual standard errors lower than the null model, in which only the intercept is considered*

		Model 0	Model 1
	AIC	4189.41	4008.93
	BIC	4202.98	4058.67
Random effects	Intercept StdDev	2.93	2.04
	Residual StdDev	5.16	4.78

but the regressors play a constant effect across the level 2 units. AIC and BIC values can be used to compare the two random models, where smaller values reflect better model fit. Table 4 shows that Model 1 provides a better fit to the data: the within-School variation (Residual StdDev in Table 4) and the variation in the intercept across the Schools (Intercept StdDev in Table 4) decrease after introducing the regressors into the model.

The estimated coefficients are consistent with the expected effect of the regressors on the students' performance. In particular, for Model 1 the estimates are (significant coefficient for $\alpha = 0.10$ in bold): intercept=**97.15**, male=**-1.99**, residence in Marche region=0.08, residence outside Marche=**-2.69**, regular course attendance=**1.37**, foreign experience=**1.36**, working student=-0.23, years to get a degree=**-0.59**, diploma mark=**0.14**.

It is worth highlighting the numerical problem founded in estimating a random slope model, i.e. inserting a random coefficient to model variation across the School levels. Because the considered model includes several regressors, the optimization algorithm based on the likelihood function does not converge. This is a widespread problem affecting the possibility of completely exploiting the potentialities of multilevel models.

5. THE QUANTILE REGRESSION PROPOSED APPROACH

This section introduces a procedure aiming to detect group effects through a QR approach. Despite the use of different models for each School, the proposed procedure estimates the group dependence structure using the whole sample, and thus it does not require ad hoc tools for comparing the models. Contrary to the use of a dummy variable for each School level, the proposal is able to capture differing impacts of the regressors in the groups. Unlike the model enclosing the group variable and all the interactions, it is parsimonious in the number of coefficients, meanwhile mimicking a multilevel approach with the association of a particular quantile model for each group. The approach is structured in the three steps detailed below: identification of the best model for each group, estimation of the group dependence structure and test of the differences among groups.

1) *Identification of the best model for each group*

Exploring the whole conditional distribution of the response through QR offers a different perspective on the dependence structure linking the response with the considered regressors, as shown before in Figures 7 and 8. Taking into account the group variable, the School, our approach starts with the association of a representative quantile with each group. Whereas the group variable is relevant for describing the data, such quantiles should be different, hence also determining differences in the dependence structure among groups. If instead the quantiles are similar, the group variable will not play a relevant role in describing data.

The conditional quantiles representative of each group are determined by computing the rank percentiles of each statistical unit with respect to the response variable and then averaging them by groups. The obtained means are considered representative of the groups. The choice of the proper location index should be dictated by the shape of the distribution of the rank percentiles in the groups. The motivation for the use of the percentile ranks is further examined in Section 6 through a working example on synthetic data.

Figure 9 depicts the case of the analysed data: the plot is divided into six panels referring to the six Schools, with each point representing the percentile rank of the corresponding units in the marginal degree mark distribution. The six stars depict the group means computed through arithmetic average and are considered the quantiles representative of each group. In particular, we have: CS: 0.70, EC: 0.40, ED: 0.72, LA: 0.69, LW: 0.35, PS: 0.51. For the analysed data the use of the median as location summary provides very similar results.

In the following, we denote with $\theta_g^{best}, g = 1, \ldots, m$, the quantiles representative of each group. The identified best

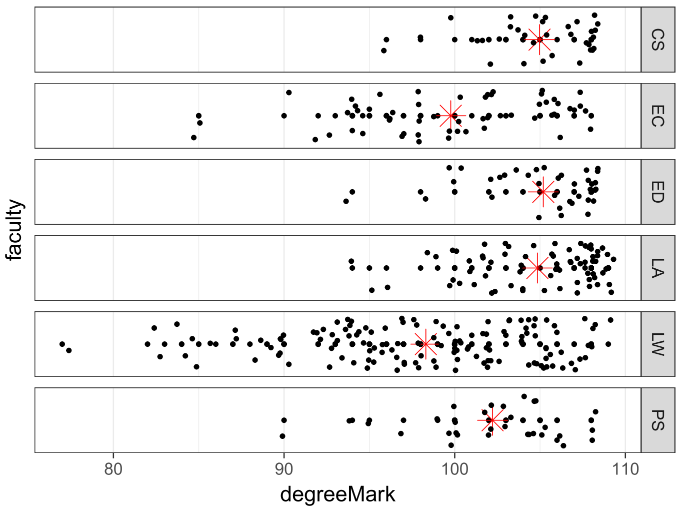

Figure 9. Percentile rank representation of the degree mark according to School.

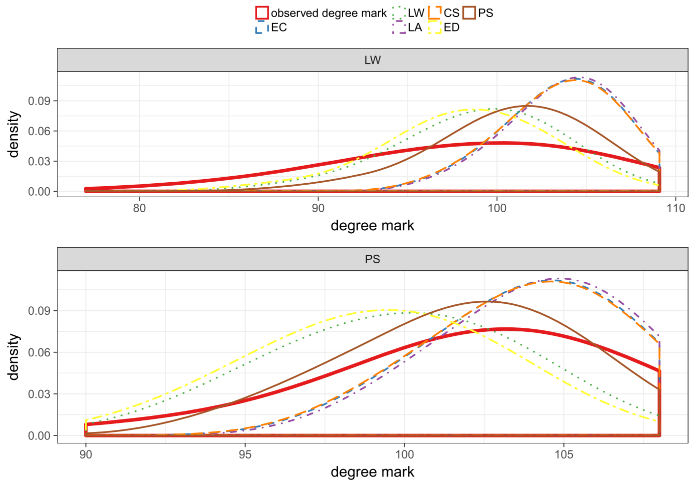

Figure 10. Degree mark distributions for the LW (upper part) and PS (lower part) students. Each panel depicts the observed degree mark (thick line) and the estimated curves obtained through the best models assigned to each School. Observed and estimated densities converge if the latter are obtained using the proper best model.

quantiles characterize the groups, meaning that, for example, the dependence structure for EC students is best represented by a QR model with $\theta = 0.40$, i.e. the features of EC students mainly affect the 40th conditional percentile of the degree mark. All the above remarks about the peculiarities of the Schools are fully confirmed by the percentile rank means as the best quantiles for LA, CS and ED are almost equal.

The comparison among the θ_g^{best} values provides information about the presence of group differences and peculiarities. Namely, two groups characterized by different θ_g^{best} values could have typical dependence structures. However, as different θ_g^{best} values do not necessarily imply a different impact of the regressors on the dependent variable, the next step allows us to focus on the group dependence structures.

2) *Estimation of the group dependence structure*

In the second step, QR is carried out on the whole sample using the m quantiles θ_g^{best} assigned to the m groups in the previous step. The generic element $\hat{\beta}_j\left(\theta_g^{best}\right)$ of the estimated coefficient matrix $\hat{\mathbf{B}}\left(\theta^{best}\right)_{[p\times m]}$ provides the effect of the jth regressor in the gth group. The coefficient matrix consists of m column vectors, one for each considered conditional quantile, i.e. for each group. The inspection of such a matrix allows detection of the group dependence structure. Tools to test interquantile differences [6] are available to evaluate the statistical significance of the differences among the coefficients related to each group.

The results for the considered data are shown in Table 5, where each column refers to a θ_g^{best}. Each School is indeed characterized through the θ_g^{best} assigned according to its percentile rank mean. Significant coefficients at $\alpha = 0.10$ are shown in bold for each covariate (rows of the table). The QR coefficients highlight the differences among the groups. In particular, such differences can be identified in terms of the intensity of the values of each regressor. For example, the effect on the degree mark of living in the Marche region is always positive but is stronger in Schools such as EC and LW. An additional check of the effectiveness of the θ_g^{best} for each group is possible through inspection of the models associated to the groups in terms of predicted values. A trivial casting out nines can be carried out, comparing the predictions for the units of a given group using the best model for the group with the ones provided by a model associated to another group. Results should indeed worsen insomuch as the two groups differentiate with respect to the associated best quantiles. For example, Figure 10 (upper part) shows, for LW students, the observed response variable (thick line) and the estimated densities obtained using the best models associated with each School. It is evident that the best model assigned to the LW School (with $\theta = 0.35$) provides the density closer to the observed distribution (dotted line). On the contrary, if the LW degree mark is estimated through the best model assigned to the LA/ED Schools (with $\theta = 0.0.69$ and $\theta = 0.0.72$), then the observed and estimated densities move away. It is, of course, obvious to expect that when referring to PS students, the best choice to estimate the degree mark relies on the model assigned to the PS School, as shown in the lowest part of Figure 10.

3) *Test of the differences among groups*

In the final step, the evaluation of the statistical significance of the differences among the coefficients related to each group can be carried out by exploiting the classical inferential tools available in the quantile regression framework [1, 10]. It is important to highlight that group coefficients can be compared because they have been estimated on the

Table 5. Group effects estimated on the real data example (significant coefficients at $\alpha = 0.10$ in bold). The coefficients measure the dependence structure inside each group

	CS θ=0.70	EC θ=0.40	ED θ=0.72	LA θ=0.69	LW θ=0.35	PS θ=0.51
Intercept	**103.26**	**100.97**	**103.47**	**102.90**	**102.61**	**104.12**
gender-Male	**-3.39**	**-4.81**	**-2.54**	**-3.59**	**-5.08**	**-4.14**
place of residence-Marche region	**0.97**	1.70	**0.14**	**0.87**	**2.06**	**1.31**
place of residence-outside Marche	-1.03	-2.67	-0.99	-1.07	-2.29	-1.35
course attendance-regular	**1.60**	1.15	**1.37**	**1.89**	0.75	**1.91**
working student	-0.60	0.07	-0.61	-0.65	0.00	-0.03
years to get a degree	**-0.42**	**-0.98**	**-0.38**	-0.42	**-1.16**	**-0.89**
diploma mark	**0.05**	**0.07**	**0.05**	0.05	**0.06**	0.05

Table 6. P-values derived from testing differences on each slope coefficient (first through second to last column) and on the whole model (last column) obtained considering all the possible pairwise comparisons between Schools

	Male	Marche	outside Marche	regular student	working student	years to get a degree	diploma mark	joint test
CS vs EC	0.136	0.300	0.226	0.527	0.386	0.027	0.469	0.062
CS vs ED	0.010	0.195	0.932	0.120	0.986	0.478	0.724	0.060
CS vs LA	0.242	0.421	0.683	0.008	0.796	0.939	0.610	0.008
CS vs LW	0.077	0.135	0.175	0.224	0.431	0.002	0.779	0.005
CS vs PS	0.366	0.542	0.789	0.589	0.347	0.017	0.793	0.238
EC vs ED	0.022	0.428	0.206	0.764	0.385	0.018	0.436	0.013
EC vs LA	0.208	0.246	0.244	0.294	0.380	0.029	0.508	0.063
EC vs LW	0.497	0.311	0.612	0.278	0.846	0.161	0.402	0.697
EC vs PS	0.306	0.430	0.168	0.134	0.846	0.603	0.189	0.245
ED vs LA	0.005	0.149	0.867	0.001	0.854	0.535	0.566	0.000
ED vs LW	0.011	0.209	0.198	0.388	0.432	0.001	0.734	0.001
ED vs PS	0.069	0.763	0.749	0.334	0.345	0.010	0.859	0.049
LA vs LW	0.128	0.108	0.194	0.101	0.417	0.002	0.827	0.005
LA vs PS	0.508	0.437	0.821	0.984	0.354	0.020	0.732	0.313
LW vs PS	0.199	0.198	0.365	0.042	0.960	0.163	0.542	0.145

whole sample, unlike an approach estimating separate models for each group. This step can be conducted using one of the classical tests proposed in [12] aimed at evaluating the significance of the differences among coefficients pertaining to different quantiles. The most common test statistic is a variant of the Wald test, which is also able to provide a joint test on all slope parameters. If we take into account pairwise comparisons, $\binom{6}{2} = 15$ possible comparisons are possible for the analysed real data example. They are presented on the rows of Table 6, in terms of the p–values deriving from testing differences on each slope coefficient (first though second to last column) and on the whole model (last column). Several significant differences come to light, both on the whole model and for couples of coefficients. The same analysis could be carried out comparing more than two groups using a similar test statistic.

6. FURTHER CONSIDERATIONS THROUGH SYNTHETIC DATA

This section uncovers the main features of the proposed approach through synthetic data: when the group dependence structure is known, it is indeed possible to discuss properties, strengths and weaknesses of the proposal with respect to its ability to discover heterogeneity among units.

In Subsection 6.1 a working example is introduced to show, step by step, if and how the group dependence structure identified by the proposed method matches with the generated data. A Monte Carlo simulation study is provided in Subsection 6.2 to deal with different patterns of group dependence structure.

6.1 A working example

A simple synthetic dataset is generated according to a specified group dependence structure: the data refers to two groups of different sizes ($n_1 = 30$; $n_2 = 70$) and with different dependence structures. Details are provided in Table 7.

By stacking data pertaining to the two groups, a dependent variable $\mathbf{y}$ and a regressor $\mathbf{x}$ are derived:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \tag{2}$$

Figure 11 shows the distribution of the dependent vari-

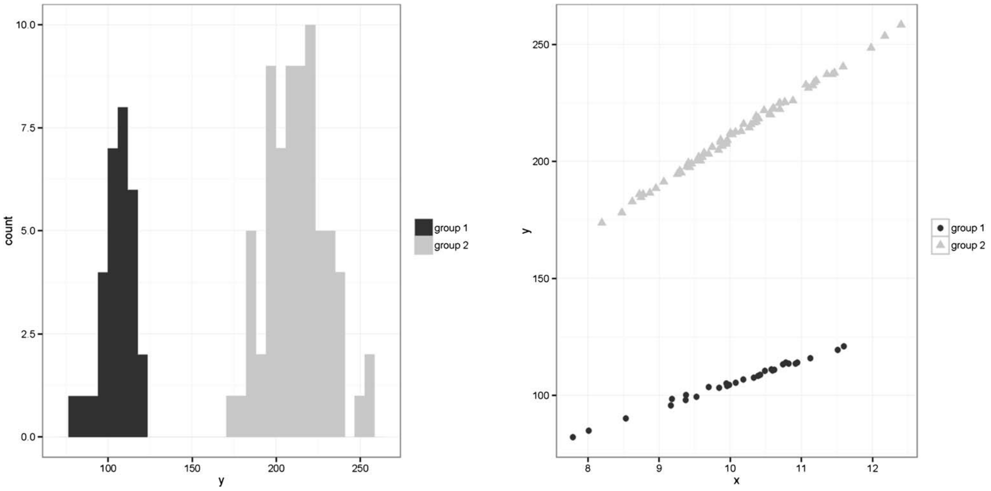

Figure 11. Histogram of the dependent variable (left) and scatter plot of the two variables (right) for the synthetic dataset. The group membership is highlighted using different grey levels.

Table 7. Structure of the two groups composing the synthetic dataset. The groups are well separated with respect to the involved features (see also Figure 11)

	group 1	group 2
sample size	$n_1 = 30$	$n_2 = 70$
regressor	$\mathbf{x}_1 \sim N(10;1)$	$\mathbf{x}_2 \sim N(10;1)$
error	$\mathbf{e}_1 \sim N(0;1)$	$\mathbf{e}_2 \sim N(0;1)$
response variable	$\mathbf{y}_1$=5+10$\mathbf{x}_1$+$\mathbf{e}$	$\mathbf{y}_2$=10+20$\mathbf{x}_2$+$\mathbf{e}$

Table 8. Group effects estimated on the synthetic dataset. The estimates properly detect the structure of the two groups (see Table 7)

	$\theta = 0.145$	$\theta = 0.640$
	group 1	group 2
intercept	4.813	8.399
x	10.001	20.143

able (left-hand side) and the scatter plot of the two variables (right-hand side) distinguishing the role of the two groups. The group structure is evident both from the univariate distribution of the $\mathbf{y}$ variable and from the scatter plot depicting the relationship between the two variables.

The best model for each group can be obtained by analysing the percentile rank distribution of the dependent variable in each group (left-hand side in Figure 12). In both cases, the distribution looks quite symmetric, thus suggesting that the best model for the two groups can be identified using the percentile rank means (g1: 0.145, g2: 0.640). In order to validate such percentile rank means as representative quantiles for each group, the whole conditional dependence structure can be exploited in detail. In particular, the quantile regression model $Q_\theta(\hat{\mathbf{y}}|\mathbf{x}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}$ is estimated using the whole quantile process [1, 10] and the best model for each unit is identified by choosing the quantile that returns the estimated value of the dependent variable closest to the observed value. It is worth recalling that the quantile process fits quantile regression models for the entire range of quantile levels from 0 to 1. Therefore, it estimates the entire probability distribution of a response variable conditional on its covariates. From a practical point of view, for each unit, several possible models are available, one for each possible solution identified through the quantile process. Among such models, the best model in terms of prediction is assigned to each unit. The right-hand panel in Figure 12 compares the observed response variable (left-hand side) with its estimates: one provided by OLS (middle) and one representing the best estimated vector $\hat{\mathbf{y}}\left(\theta^{best}\right)$ (right-hand side). From the figure, the added value in considering the whole quantile process is evident in that it almost perfectly reconstructs the dependent variable.

Using the conditional quantiles that identify the best model for each unit along with the group membership, it is possible to derive the best model for each group. Such models are indeed obtained by averaging the best quantiles assigned to the units belonging to the group. In the case of the analysed data, the results are practically equivalent to the values obtained by exploiting the percentile ranks, as previously introduced. This also allows easier interpretation of the best quantiles associated to the involved groups.

For the synthetic dataset, the comparison of the QR coefficients associated to the best quantiles (Table 8) compared to the values determining the structure of the two groups (Table 7), shows that the proposed approach is able to correctly estimate the dependence structure in each group. A separate analysis of each group would lead to equivalent results, but the added value of the proposed approach relies

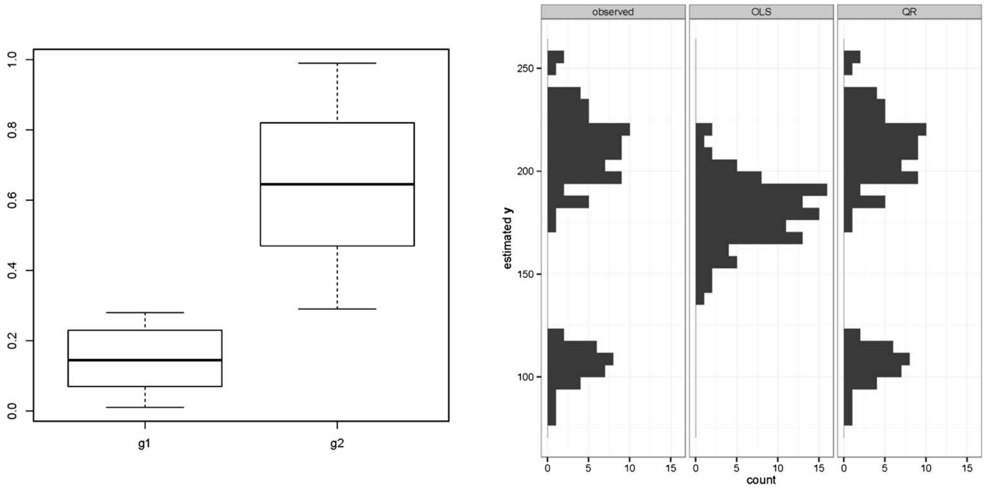

Figure 12. Boxplots of the percentile ranks of the dependent variable in the two groups (left). Distribution of the dependent variable and the OLS and QR estimated dependent variable (right).

on the possibility of comparing groups through coefficients obtained on the whole sample. Moreover, well-known tools for statistical comparisons among coefficients related to different groups [6] can be used.

The presence of such a difference between the two groups simplifies interpretation of the results. The value $\theta = 0.145$, which characterizes group 1, indicates that the corresponding units are in the lower tail of the dependent variable distribution. The same happens for group 2 but with $\theta = 0.640$.

Both the test of equality of distinct slopes and the joint test of equality of slopes confirm that coefficients associated to the two groups are statistically different.

6.2 A simulation study

The working example described in the previous section permits appreciation of how the group dependence structure is identified and how to interpret the best quantiles associated to each group. However, the results are specific to the generated dataset and to its simple structure. There are indeed several issues that should be taken into account to provide a wider discussion on the features of the proposed method. Such issues have been considered for the simulation study presented in the next subsection.

6.2.1 Description of the study

The simulation study aims to explore the robustness of the method with respect to the degree and type of overlapping among the groups; the cardinality of each group (equal or unbalanced); and the sample size.

This paper focuses on the case of one regressor and two groups. Evaluations related to the introduction of more than one regressor and more than two groups are postponed to a future work.

To explore the effect played by the *degree and type of overlapping between the groups*, a set of scenarios is generated. Figure 13 shows the scatter plot of the two variables for each considered scenario, distinguishing units belonging to each group by symbols and grey levels. Each row of the scatter plot matrix refers to a class of scenarios:

Case 1: parallel group structures;
Case 2: group structures crossing outside the considered range of the regressor;
Case 3: group structures crossing inside the considered range of the regressor.

The columns represent instead the different degree of overlapping among the groups, distinguishing three increasing levels denoted as a, b and c.

The dependence structures associated to each of the nine considered scenarios are detailed in Table 9, where β_0 and β_1 represent, respectively, the intercept and the slope of each model. For example, the scatter plot in the upper left part of Figure 13 (Case 1a) refers to a model very similar to that described in Subsection 6.1. Each scenario is generated considering a regressor $\mathbf{x}_1 \sim N(10; 1)$ for group 1, a regressor $\mathbf{x}_2 \sim N(10; 1)$ for group 2, and in both cases an error $\mathbf{e} \sim N(0; 1)$. Data pertaining to the two groups are stacked, thus obtaining a unique dependent variable and a unique regressor observed on n units, as shown for the working example in the previous subsection. The effect played by the *cardinality of each group* is explored by hypothesizing equal group sizes ($n_1 = n_2 = 70$) or unbalanced group sizes ($n_1 = 30$; $n_2 = 70$). As the sample size could have an additional effect on the ability of the method to discover the group dependence structure, 10 different sample sizes are considered: from 100 to 1,000 with step equal to 100.

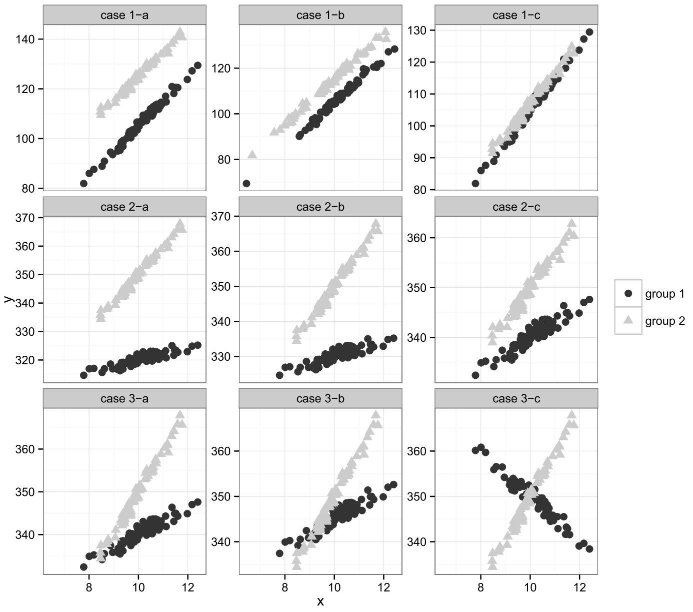

Figure 13. Scatter plots for each scenario of the simulation study. Different type (rows) and degree (columns) of overlapping among groups are taken into account.

Table 9. Coefficient values related to the considered scenarios

	Group 1		Group2	
	β_0	β_1	β_0	β_1
Case 1a	5	10	25	10
Case 1b	5	10	15	10
Case 1c	5	10	7	10
Case 2a	300	2	250	10
Case 2b	310	2	250	10
Case 2c	310	3	280	7
Case 3a	310	3	250	10
Case 3b	315	3	250	10
Case 3c	400	-5	250	10

Summarizing, the simulation design is composed by:

- nine scenarios corresponding to different types (Case 1, Case 2, Case 3) and degrees (a, b, c) of overlapping between the groups;
- 1,000 replications of the data corresponding to each scenario, both for the case of equal groups ($n_1 = n_2 = 70$) and for the unbalanced case ($n_1 = 30; n_2 = 70$); and
- 1,000 replications of the data corresponding to each scenario for each considered sample size (n from 100 to 1,000 with step equal to 100) considering unbalanced groups ($n_1 = 30\% \times n$; $n_2 = 70\% \times n$). This choice was preferred to the balanced case because it represents a more difficult case.

6.2.2 Main results

The approach described in Section 5 is applied to each scenario. The best model for each group has been computed as the mean of the percentile ranks of the units belonging to each group. The final estimates are obtained by performing a QR on the whole sample considering only the two quantiles representing the groups. Finally, the estimates obtained on the 1,000 replications are averaged and reported in Table 10 for the balanced case and in Table 11 for the unbalanced case.

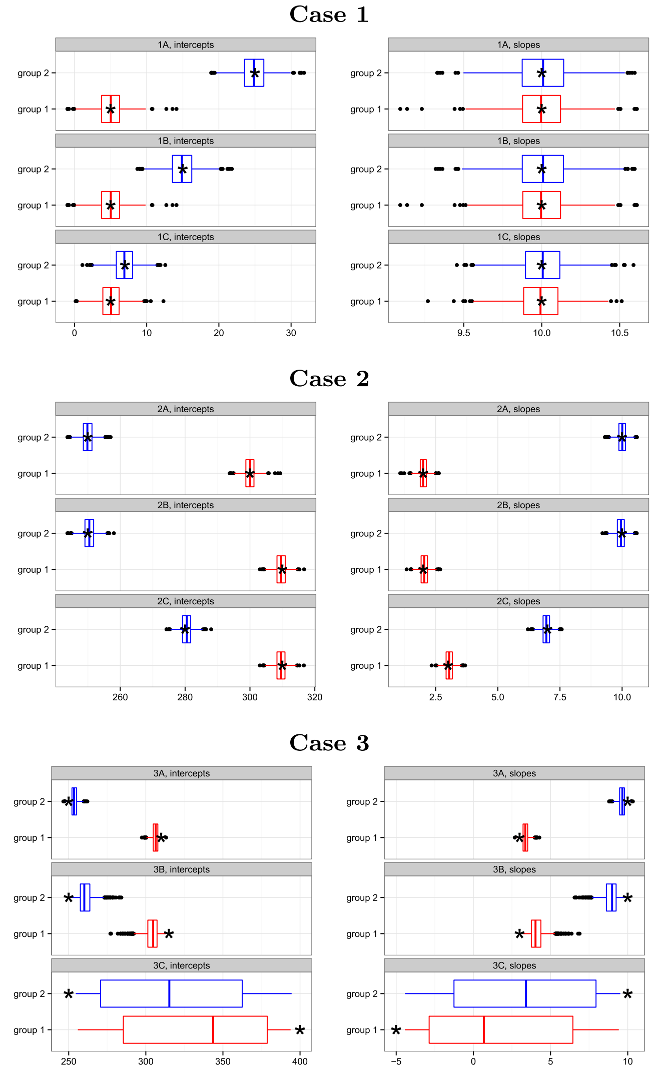

Figure 14. Monte Carlo distributions of the estimates computed on the 1,000 replications for the intercepts (left-hand side) and the slopes (right-hand side) ($n_1 = n_2$=70). Stars and segments inside the boxes represent, respectively, the true coefficients and the Monte Carlo estimates. The method is able to identify the group effects in most of the scenarios.

The comparison of the original coefficients with the estimates shows how the method is able to correctly capture the dependence structure in each group. Moreover, the variability and the distribution of the estimates complement the evaluation.

Table 10. Group QR estimates related to the 9 considered scenarios (standard deviations of the estimates are given in parentheses) and corresponding coefficient values ($n_1 = n_2$=70). The method is able to identify the group effects in most of the scenarios

		Group 1		Group2	
		β_0	β_1	β_0	β_1
Case 1a	estimate	4.99 (*1.93*)	10.00 (*0.19*)	24.92 (*2.08*)	10.00 (*0.21*)
	coefficient	5.00	10.00	25.00	10.00
Case 1b	estimate	4.99(*1.92*)	10.00 (*0.19*)	14.96 (*2.07*)	10.00 (*0.21*)
	coefficient	5.00	10.00	15.00	10.00
Case 1c	estimate	5.03 (*1.66*)	10.00 (*0.17*)	6.95 (*1.75*)	10.00 (*0.18*)
	coefficient	5.00	10.00	7.00	10.00
Case 2a	estimate	299.99 (*1.93*)	2.00 (*0.19*)	249.94 (*2.08*)	10.00 (*0.21*)
	coefficient	300.00	2.00	250.00	10.00
Case 2b	estimate	309.56 (*1.91*)	2.04 (*0.19*)	250.45 (*2.07*)	9.95 (*0.21*)
	coefficient	310.00	2.00	250.00	10.00
Case 2c	estimate	309.54 (*1.87*)	3.05 (*0.19*)	280.52 (*1.97*)	6.95 (*0.20*)
	coefficient	310.00	3.00	280.00	7.00
Case 3a	estimate	306.22 (*2.39*)	3.38 (*0.24*)	253.68 (*2.36*)	9.63 (*0.24*)
	coefficient	310.00	3.00	250.00	10.00
Case 3b	estimate	303.74 (*5.20*)	4.13 (*0.53*)	261.08 (*5.17*)	8.88 (*0.52*)
	coefficient	315.00	3.00	250.00	10.00
Case 3c	estimate	332.77 (*46.25*)	1.72 (*4.63*)	317.31 (*45.82*)	3.27 (*4.58*)
	coefficient	400.00	-5.00	250.00	10.00

Table 11. Group QR estimates related to the 9 considered scenarios (standard deviations of the estimates are given in parentheses) and corresponding coefficient values (n_1=30; n_2=70). The method is able to identify the group effects in most of the scenarios

		Group 1		Group2	
		β_0	β_1	β_0	β_1
Case 1a	estimate	5.01 (*3.33*)	10.00 (*0.33*)	24.92 (*1.78*)	10.00 (*0.18*)
	coefficient	5.00	10.00	25.00	10.00
Case 1b	estimate	5.01(*3.34*)	10.00 (*0.33*)	14.92 (*1.78*)	10.00 (*0.18*)
	coefficient	5.00	10.00	15.00	10.00
Case 1c	estimate	5.07 (*2.32*)	9.99 (*0.23*)	6.96 (*1.65*)	10.00 (*0.16*)
	coefficient	5.00	10.00	7.00	10.00
Case 2a	estimate	299.97 (*3.33*)	2.00 (*0.33*)	249.92 (*1.78*)	10.00 (*0.18*)
	coefficient	300.00	2.00	250.00	10.00
Case 2b	estimate	308.81 (*3.50*)	2.12 (*0.35*)	250.17 (*1.77*)	9.98 (*0.18*)
	coefficient	310.00	2.00	250.00	10.00
Case 2c	estimate	308.82 (*3.15*)	3.12 (*0.32*)	280.20 (*1.72*)	6.98 (*0.17*)
	coefficient	310.00	3.00	280.00	7.00
Case 3a	estimate	300.41 (*6.06*)	3.96 (*0.62*)	251.60 (*1.86*)	9.83 (*0.19*)
	coefficient	310.00	3.00	250.00	10.00
Case 3b	estimate	281.37 (*10.42*)	6.38 (*1.11*)	253.51 (*2.00*)	9.64 (*0.20*)
	coefficient	315.00	3.00	250.00	10.00
Case 3c	estimate	266.91 (*13.20*)	8.30 (*1.31*)	254.85 (*2.28*)	9.51 (*0.23*)
	coefficient	400.00	-5.00	250.00	10.00

Let us consider the case of *equal cardinality of each group* ($n_1 = n_2 = 70$). Table 10 shows for each scenario (rows of the table) the original coefficients and the average of the Monte Carlo estimates for both the intercept and the slope of each model (the standard errors of the estimates are given in parentheses).

The Monte Carlo distributions of the estimates are reported in Figure 14: each boxplot represents the distribution for the intercepts (left-hand side) and the slopes (right-hand side). Stars inside the boxplots represent the true coefficients.

The first issue worth mentioning is the ability of the method to detect the two dependence structures in all the simulated scenarios with the exception of case 3c. The results also begin to degrade in case 3b, as expected since the crossing between the group structure starts to become

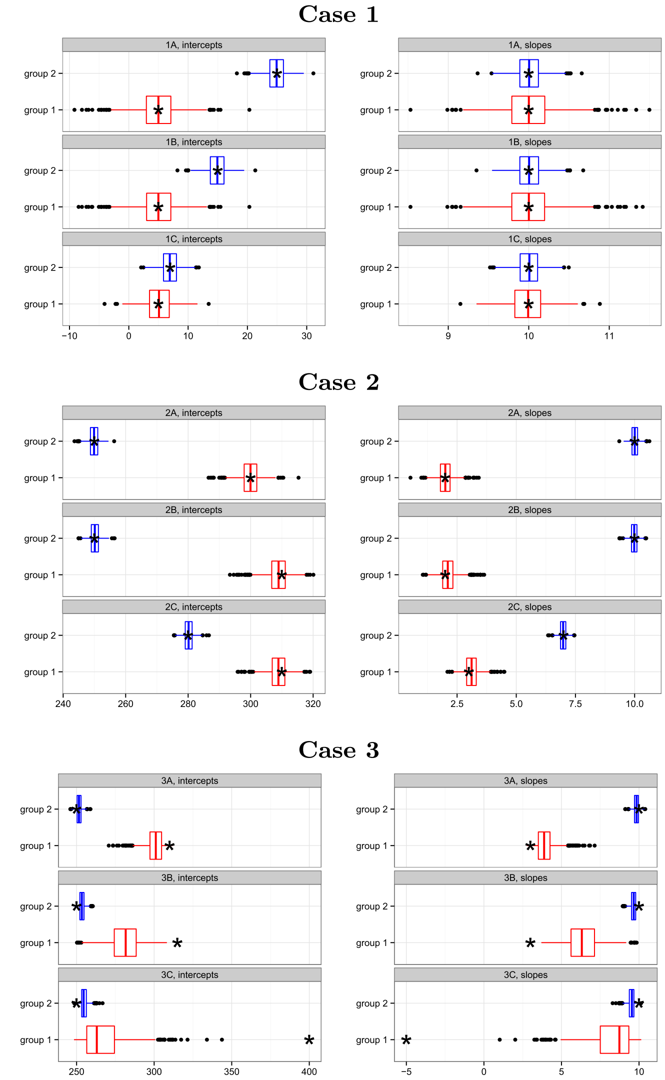

Figure 15. Monte Carlo distributions of the estimates computed on the 1,000 replications for the intercepts (left-hand side) and the slopes (right-hand side) (n_1=30; n_2=70). Stars and segments inside the boxes represent, respectively, the true coefficients and the Monte Carlo estimates. The method is able to identify the group effects in most of the scenarios.

more pronounced. The crossing considered in case 3c is not detectable using linear QR.

For Case 1 and Case 2, the estimates are instead unbiased and the degree of overlapping (a, b or c) seems not to affect the ability of the method to identify the true coefficients.

Also, in the case of *unbalanced cardinality of the groups* (Figure 15 and Table 11), the ability of the method to identify the right coefficients for each group and for each model still holds for Case 1 and Case 2. It is worth mentioning the higher variability of the Monte Carlo estimates as shown by the standard deviations in Table 11.

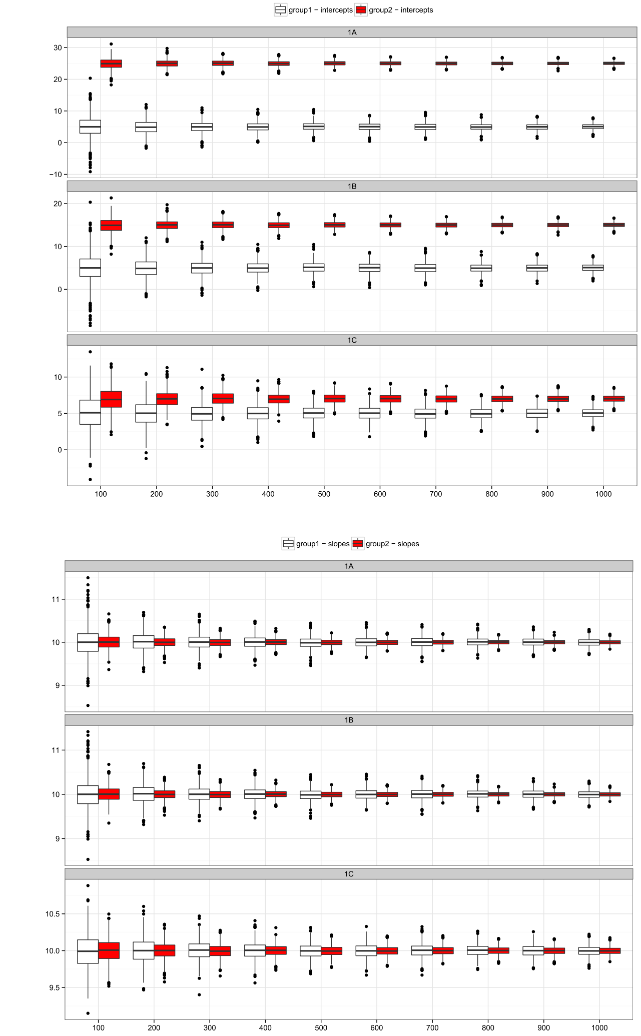

Figure 16. Monte Carlo distributions of the intercept (upper graph) and slope (lower graph) estimates, varying the sample size from 100 to 1,000 with step equal to 100 - Case 1 (n_1=30% of n; n_2=70% of n). The variability of the estimates falls as the sample size increases.

The effect of the *sample size* for the 10 considered sizes is shown in Figures 16, 17 and 18, depicting for each of the three considered cases (1, 2 and 3) the distribution of the intercepts (upper graphs) and the slopes (lower graphs) according to the different degrees of overlapping among the groups (a, b and c). The white boxplots refer to group 1 ($n_1 = 30\% \times n$), and the black ones represent group 2 ($n_2 = 70\% \times n$). As expected, in all the simulated scenarios, the variability of the Monte Carlo distributions falls moving from the left boxplot (sample size equal to 100) to the right one (sample size equal to 1,000). Moreover, the variability of the estimates of the smaller group is always greater than the variability of the other group. These results confirm the

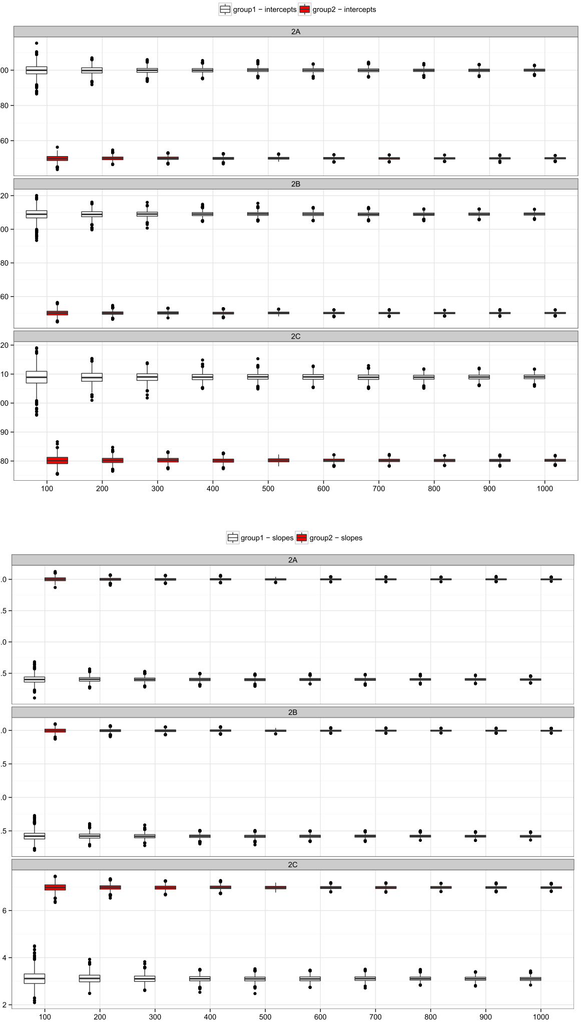

Figure 17. Monte Carlo distributions of the intercept (upper graph) and slope (lower graph) estimates, varying the sample size from 100 to 1,000 with step equal to 100 - Case 2 (n_1=30% of n; n_2=70% of n). The variability of the estimates falls as the sample size increases.

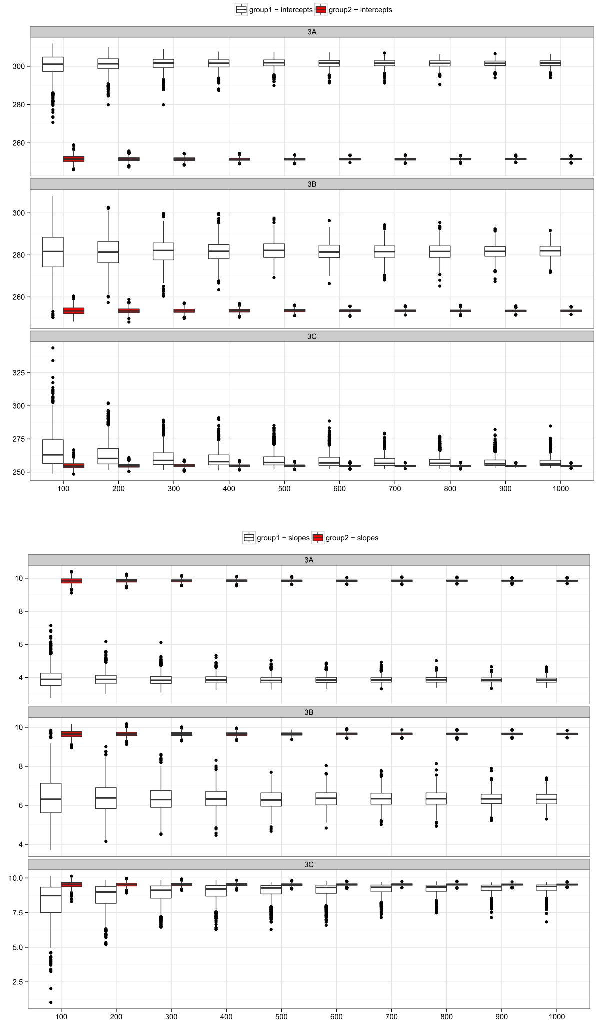

Figure 18. Monte Carlo distributions of the intercept (upper graph) and slope (lower graph) estimates, varying the sample size from 100 to 1,000 with step equal to 100 - Case 3 (n_1=30% of n; n_2=70% of n). The variability of the estimates falls as the sample size increases.

consistency of the proposed approach. Finally, the effect of the sample size seems to be the same for the different degrees of overlapping.

7. CONCLUDING REMARKS AND FURTHER DEVELOPMENTS

The approach introduced in this paper exploits quantile regression to evaluate if and how group membership affects the relationship between a response variable and a set of regressors. The effect of the group membership is identified through assigning to each group the quantile best representing its impact on the dependent variable.

The working example on synthetic data has shown the ability of the proposed method to distinguish the different dependence structures characterizing the groups. The synthetic dataset illustrated the interpretation of the results. The presented simulation study has allowed to appreciate the robustness of the method with respect to the dependence structure and the degree of separation between the groups in the case of simple regression models when data are stratified into two groups. The analysis in the case of more than two groups, with different distributions of the variables and multiple regression models, will be examined in a future simulation study.

The method has been shown in action on a real data application, aiming to evaluate the effect of several students' features on University outcome. The application allows us to highlight the following distinguishing features of the proposed method:

- estimation of the group dependence structure: for each group, a set of regressor coefficients measures the specificity of the internal dependence structure;
- clarity of the final results: the coefficients associated to each group follow the same interpretation of any linear model; furthermore, the best quantile assigned to each group synthesizes the location of the response conditional distribution on which the group exerts the main effect;
- availability of classical inferential procedures for testing differences among the group, since the group effects are identified using the whole sample.

The procedure estimates the best quantiles through the mean of rank percentiles of the observed data according to the group membership. Albeit the best quantiles are data driven, and this could represent a potential limitation, in real applications accurate information about the order statistics of the population are quite rare. To deal with such an issue, the procedure could be enhanced introducing a preliminary study for assessing the stability of the estimated best quantiles. Resampling methods are valuable tools at this end.

A further issue worth of future work concerns the extension of the proposal to manage longitudinal data where the role of the grouping variable is played by the time.

Received 15 April 2016

REFERENCES

[1] Davino C., Furno M., Vistocco D. (2013). *Quantile Regression: Theory and Applications*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Inc. MR3236027

[2] Davino C., Vistocco D. (2007). The evaluation of University educational processes: a quantile regression approach. *Statistica*, 3, 267–278. MR2664981

[3] Gelman A., Hill J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

[4] Geraci M., Bottai M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24, 461–479. MR3192268

[5] Goldstein H. (2011). *Multilevel Statistical Models*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Inc.

[6] Gould W.W. (1997). sg70: Interquantile and simultaneous-quantile regression. *Stata Technical Bulletin*, 38. 14–22. Reprinted in *Stata Technical Bulletin Reprints*, 7, 167–176. College Station, TX: Stata Press.

[7] Gujarati D.N. (2003). *Basic Econometrics*. New York: McGraw–Hill, International Edition.

[8] Jaccard J., Turrisi R. (2003). *Interaction Effects in Multiple Regression*. Series: Quantitative Applications in the Social Sciences Volume 72. Second edition. Thousand Oaks, CA: SAGE Publications, Inc.

[9] Koenker R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Data Analysis*, 91, 74–89. MR2083905

[10] Koenker R. (2005). *Quantile Regression*. Econometric Society Monographs No. 38. New York: Cambridge University Press. MR2268657

[11] Koenker R., Basset G.W. (1978). Regression quantiles. *Econometrica*, 46, 33–50. MR0474644

[12] Koenker R., Basset G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1). MR0640165

[13] Koenker R., D'Orey W.V. (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3), 383–393.

[14] Lamarche C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics*, 157, 396–498. MR2661611

[15] Parzen M.I., Wei L., Ying Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 81, 341–350. MR1294895

[16] Raudenbush S.W., Bryk A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Second Edition. Thousand Oaks, CA: Sage Publications.

[17] Snijders T.A.B., Bosker R.J. (2012). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. Second Edition. London: Sage Publications. MR3137621

Cristina Davino
Department of Economics and Statistics
University of Naples Federico II
Italy
E-mail address: cristina.davino@unina.it

Domenico Vistocco
Department of Economics and Law
University of Cassino and Southern Lazio
Italy
E-mail address: vistocco@unicas.it