Metadata of the chapter that will be visualized online

Chapter Title	Quantile Regression for Clustering and Modeling Data	
Copyright Year	2015	
Copyright Holder	Springer International Publishing Switzerland	
Author	Family Name	**Davino**
	Particle	
	Given Name	**Cristina**
	Suffix	
	Email	cristina.davino@unimc.it
Author	Family Name	**Vistocco**
	Particle	
	Given Name	**Domenico**
	Suffix	
	Email	vistocco@unicas.it
Abstract	This paper aims to propose an innovative approach to identify a typology in a quantile regression model. Quantile regression is a regression technique that allows to focus on the effects that a set of explanatory variables has on the entire conditional distribution of a dependent variable. The proposal concerns the use of multivariate techniques to simultaneously cluster and model data and it is illustrated using an empirical analysis. This analysis regards the impact of student features on the university outcome, measured by the degree mark~~, is evaluated.~~ The analysis is based on the idea that the dependence structure could be different for units belonging to different groups.	
Keywords (separated by "-")	Cluster analysis - Quantile regression - Unsupervised learning	

Quantile Regression for Clustering and Modeling Data

AQ1 **Cristina Davino and Domenico Vistocco**

Abstract This paper aims to propose an innovative approach to identify a typology in a quantile regression model. Quantile regression is a regression technique that allows to focus on the effects that a set of explanatory variables has on the entire conditional distribution of a dependent variable. The proposal concerns the use of multivariate techniques to simultaneously cluster and model data and it is illustrated using an empirical analysis. This analysis regards the impact of student features on the university outcome, measured by the degree mark, is evaluated. The analysis is based on the idea that the dependence structure could be different for units belonging to different groups.

AQ2 **Keywords** Cluster analysis • Quantile regression • Unsupervised learning

1 Introduction

In many regression problems, the estimation of a single set of coefficients provides a misrepresentation of the true dependence structure if units belong to different groups. The solution to this issue becomes more difficult to achieve when group membership is not a priori known. A simplistic solution would consist in clustering units and later estimating different models for each group. This solution however would not permit to identify the impact of the groups on the dependent variable and it would require tools for comparing of the models estimated on different samples.

The aim of this paper is to propose an innovative approach for simultaneously clustering and modeling data. It is based on the conjoint use of multivariate methods and quantile regression to identify a typology in a dependence model.

C. Davino (✉)
Department of Political Sciences, University of Macerata, Communications and Intern. Relations, 62100 Macerata, Italy
e-mail: cristina.davino@unimc.it

D. Vistocco
Department of Economics and Law, University of Cassino, 03043 Cassino (FR), Italy
e-mail: vistocco@unicas.it

I. Morlini et al. (eds.), *Advances in Statistical Models for Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-17377-1_10

Quantile regression, as introduced by Koenker and Basset [14], is an extension of the classical estimation of the conditional mean to the estimation of a set of conditional quantiles. It offers a complete view of a response variable providing a method for modeling the rates of changes at multiple points (conditional quantiles) of its conditional distribution [2, 13].

The rest of the paper is an extension of the supervised approach proposed by the authors Davino and Vistocco [3, 4]. The use of an unsupervised approach to classify units in the dependence model characterizes this proposal, whereas the former proposals exploit a priori defined groups.

In literature, a quite widespread approach to simultaneously identify a partition of the data and the related model is represented by clusterwise linear regression. The method is based on the hypothesis that there exist a finite number of unknown classes and each class is characterized by a different linear regression model. The literature concerning clusterwise linear regression is quite wide, from the starting works of Spath [19, 20] to the maximum likelihood approach of DeSarbo and Cron [5], until ~~to more~~ recent proposals ~~among which~~ [8, 11, 24]. Sharing the main goals of clusterwise linear regression, the method proposed in this paper exploits quantile regression and hierarchical clustering to model and partition data identifying a different dependence structure for each detected group. To pursue such aims, the method assigns a separate quantile model best representing each group. However, the different models are estimated on the total sample, making easier the comparisons among the group coefficients. The use of quantile regression allows us to study the dependence exploring the whole conditional distribution of the dependent variable unlike the clusterwise linear regression that focuses on the conditional mean. Furthermore, it offers well-known advantages with respect to robustness issues. Moreover, our approach attempts to overcome the main drawback of the original clusterwise linear regression, the a priori setting of the number of groups, while hierarchical clustering provides a data driven criterion to partition the sample. A proper comparison with clusterwise linear regression would require a wide simulation study that takes into account also the sensitivity of clusterwise linear regression solutions to the tuning parameters (e.g., the initial partition and the number of required groups) and it will be therefore subject of a specific paper.

The paper is organized as follows. Section 2 presents the dataset used to apply the proposed approach: it concerns the evaluation of the effectiveness of the university educational process. In ~~Sect.~~ 3 the methodology is described together with results deriving from the empirical analysis: students are grouped according to the relationship between the degree mark and their features. Some concluding remarks and future work directions are reported in ~~Sect.~~ 4.

2 A Dataset on Student University Outcome

The proposed approach is described in the following sections through an empirical analysis aiming at evaluating if and how the student features (socio-demographic and university experience attributes) affect the outcome of the university career, measured through the degree mark. As stated above, the underlying idea is that this effect can be very different for students belonging to different groups. Such groups are detected according to the relationship between the degree mark and the student features. The typology identification is embedded in a quantile regression model, and it is thus able to exploit the whole conditioned distribution of the degree mark.

The analysis is carried out on a random sample of 685 students who graduated from the University of Macerata [3], 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 degree mark is measured on a discrete scale ranging between 66 and 110, with the "cum laude" mark coded as 110. 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), foreign experience (yes, no), working condition (full-time student, working student), number of years to obtain a degree, diploma mark.

The density plot of the response variable (Fig. 1) shows the presence of a strong right skewness, further supporting the recourse to the dependence analysis outside the classical regression framework.

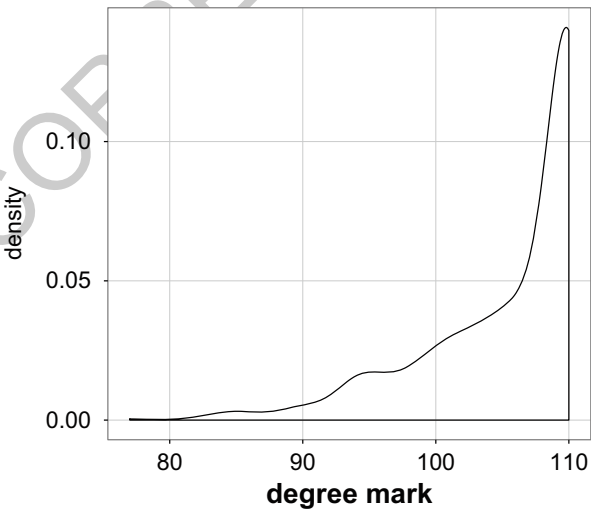

Fig. 1 Degree mark density

3 The Proposed Approach: Methodology and Results

The proposed unsupervised learning procedure is based on the joint use of hierarchical clustering and quantile regression. The approach is structured in the following four steps

1. Estimation of the global dependence structure
2. Identification of the best model for each unit
3. Identification of a typology
4. Estimation of the group dependence structure

Following sections detail the meaning of each step showing them in action on the student university outcome study.

3.1 Estimation of the Global Dependence Structure

In the first step, a quantile regression (QR) model is estimated on the whole sample:

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

where $0 < \theta < 1$ denotes the θth conditional quantile, $Q_\theta(.|.)$ is the corresponding conditional quantile function, $\mathbf{y}_{[n]}$ is the dependent variable (degree mark in the application) and $\mathbf{X}_{[n\times p]}$ is the matrix of the explanatory variables (students features in the application), n denoting the number of units and p the number of regressors.

Using a grid of k conditional quantiles, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the model provides a coefficient matrix $\hat{\Theta}_{[p\times k]}$ with a generic element that can be interpreted as the rate of change in the θth quantile of the conditional distribution of the dependent variable per unit change in the value of the jth regressor. The value of k is therefore the number of estimated conditional quantiles. A fairly accurate approximation of the whole quantile process [15] can be obtained using a dense grid of equally spaced quantiles in the unit interval (0; 1) [2].

In Fig. 2, QR coefficients, obtained using a selected grid of quantiles ($\theta = [0.1, 0.25, 0.5, 0.75, 0.9]$), are graphically represented 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 for $\alpha = 0.1$ [17]. The solid lines parallel to the horizontal axis correspond to OLS coefficients, and the related confidence intervals are represented using dashed lines for $\alpha = 0.1$.

The graphical representation allows to visually catch the different effect of the student characteristics on the degree mark. Gender and residence during university education have a great influence on the lowest quantiles of the distribution: males and residents outside the Marche region show negative coefficients. A foreign

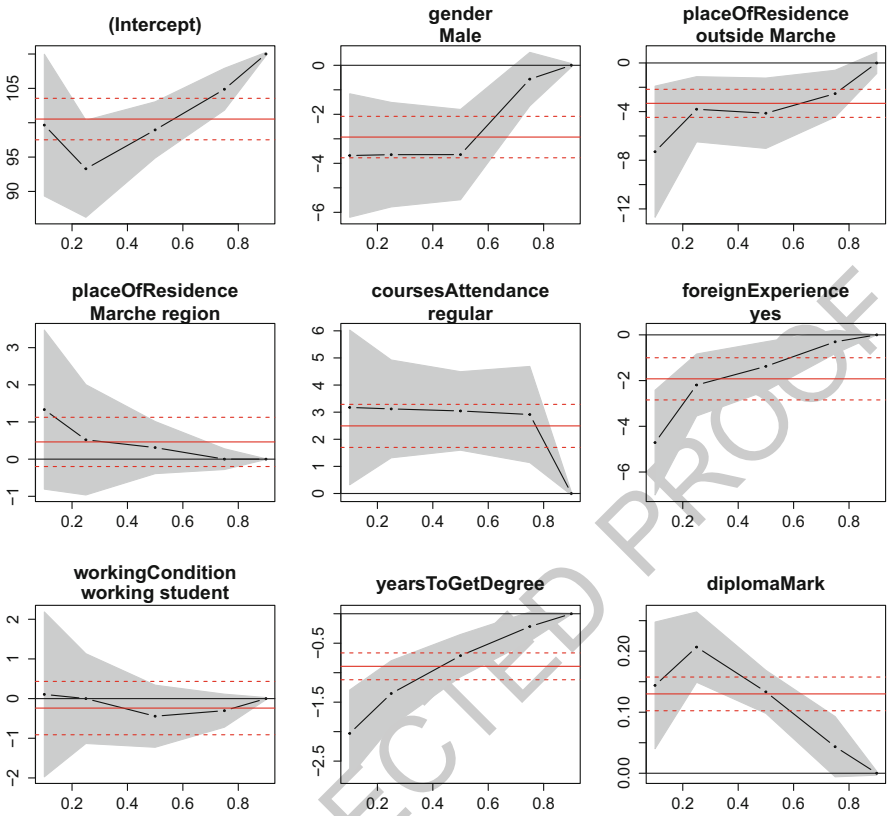

Fig. 2 OLS and QR coefficients and related confidence intervals

experience negatively influences the degree mark. This effect becomes null in the higher part of the distribution pointing out that very good students are not influenced by their university experiences abroad. Working students are less likely to get high degree marks but the QR results show how their impact is almost negligible. All the coefficients of the variable numbers of years to get a degree are negative, particularly for the lowest quantiles. Finally, the diploma mark always has a positive effect, but its value is very low for successful students.

3.2 *Identification of the Best Model for Each Unit*

In the second step, the coefficient matrix $\hat{\Theta}_{[p\times k]}$ and the regressor data matrix $\mathbf{X}$ are used to estimate the conditional distribution matrix of the response variable: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\Theta}$. The generic element of the $\hat{\mathbf{Y}}_{[n\times k]}$ matrix is the estimate of the response variable in correspondence of the ith units according to the θth quantile.

The best model for each unit i is identified by the quantile able to better estimate the response variable and it is denoted as the *best quantile*:

$$\hat{\theta}_i^{\text{best}} : \text{argmin}_{\theta=1,\ldots,k} |y_i - \hat{y}_i(\theta)| \qquad (i = 1, \ldots, n) \tag{2}$$

The best quantile, $\hat{\theta}_i^{\text{best}}$, is therefore obtained by minimizing the difference between the observed and the estimated values.

From the $\hat{\mathbf{Y}}$ matrix it is then possible to extract the best estimated vector, $\hat{\mathbf{y}}_\theta^{\text{best}}$, identifying for each unit the estimated value corresponding to the assigned best quantile. Such vector provides both an accurate approximation of the response variable and embeds information on the dependence structure relating the response variable with the regressors.

Figure 3 reproduces the histograms of the dependent variable (left panel) and the estimated dependent variable using OLS (middle panel) or the proposed QR approach (right panel). For some considerations on the added value provided by considering $\hat{\mathbf{y}}_\theta^{\text{best}}$ instead of the classical OLS predicted values, the interest reader is referred to Davino and Vistocco [4].

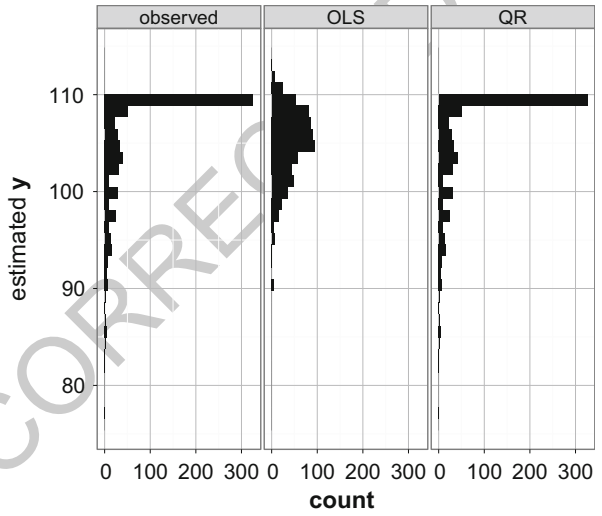

Fig. 3 Distribution of the dependent variable (*left panel*) and of the estimated dependent variable using OLS (*middle panel*) or the proposed QR approach (*right panel*)

3.3 *Identification of a Typology*

The third step of the proposed strategy aims to identify a typology on the basis of the QR results obtained in the previous step.

Units are grouped according to the best quantile they have been assigned because it can be considered as an indicator of a similar dependence structure. Working on the $\hat{\boldsymbol{\theta}}^{\text{best}}$ vector, it is possible to group units in clusters. The simplest criterion is based on its categorization. Albeit automatic methods (e.g. [21] or [18] rules) are available in literature, a certain degree of subjectivity remains. In the present paper a multivariate approach is proposed performing a hierarchical clustering [6, 10] on the estimated $\hat{\mathbf{Y}}$ matrix in order to classify units sharing similar patterns for the predicted values for all the considered quantiles.

Several criteria have been proposed in the literature to select the "best" partition by optimizing some cluster validity indexes (see, e.g., [12, 16, 23]). The seminal work of Milligan and Cooper [16] describes above 30 internal criterion measures coming from a wide variety of fields. More recently, other proposals combine the use of a cluster validity index with a searching strategy for exploring the extended hierarchy housed in a dendrogram [9] or exploit permutation tests in order to automatically detect a partition [1]. A competing method (GAP), proposed by Tibshirani et al. [22], permits to estimate the number of clusters starting from the output of any clustering algorithm. It is based on a Monte Carlo approach to derive the reference distribution of a test statistic and it requires as input the different partitions among which the optimal one has to be selected. Using the GAP statistics, the best partition is obtained by cutting the dendrogram in four groups (Fig. 4, left-hand side) with 318, 144, 154, and 64 observations, respectively.

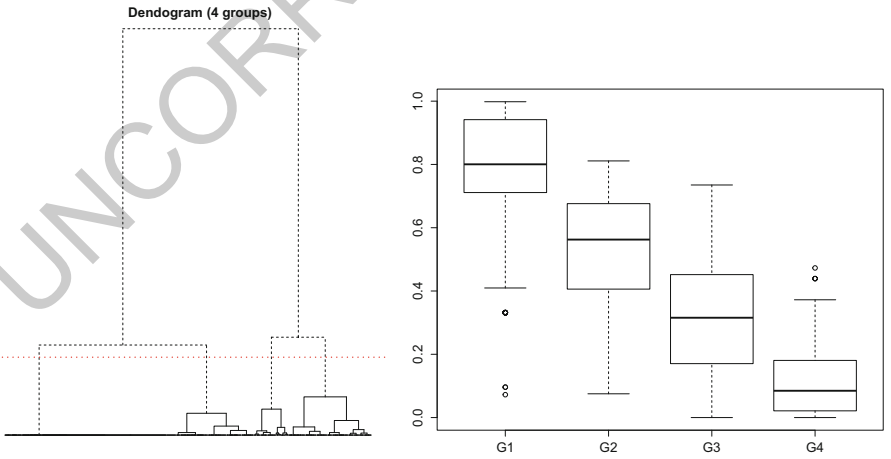

Fig. 4 Dendrogram and best partition (*left-hand side*) and distribution of the best quantiles in the identified groups (*right-hand side*)

In order to tailor a proper dependence structures as derived in the next step, a reference quantile is then associated to each group. The choice of the reference quantile is based on a synthesis measure of the distribution of the group best quantiles. Figure 4 (right-hand side) reveals a certain degree of skewness in the distribution of the best quantiles for each group. The $\hat{\boldsymbol{\theta}}^{\text{best}}$ median value of each group can then be considered as a robust reference quantile: $G1 = 0.800$, $G2 = 0.562$, $G3 = 0.315$, $G4 = 0.085$. The obtained reference quantiles are clearly distinct signaling an impact on different locations of the degree mark distribution played by the features of students belonging to each group, as it will be shown in the next section.

3.4 Estimation of the Group Dependence Structure

The four reference quantiles previously defined are used to estimate the group dependence structure. In particular, a QR model is carried out on the whole sample estimating the following four θ values: 0.800 ($G1$), 0.562 ($G2$), 0.315 ($G3$), 0.085 ($G4$). Estimating a model on the whole sample allows us to easily compare the group dependence structure through the evaluation of the statistical significance of the differences among the coefficients. In the QR framework such a comparison is based on the classical tools to test interquantile differences [7].

The QR results for the different groups are shown in Table 1 reporting the covariates on the rows and the groups on the columns; significant coefficients at $\alpha = 0.10$ are shown in bold. Reading the table by columns details information on the features mainly affecting each group, while a comparison of a specific coefficient among the different groups is provided by a row-wise inspection of the table. The reference quantiles play a crucial role in interpreting the results. For example, the effect on the degree mark of living outside Marche is negative for all the groups, but it is stronger for students belonging to group $G4$. On the other hand, as group 4

Table 1 Group effects estimates (in bold significant coefficients at $\alpha = 0.10$)

Variable	G1	G2	G3	G4
	$\theta = 0.800$	$\theta = 0.563$	$\theta = 0.315$	$\theta = 0.085$
Intercept	**109.00**	**98.87**	**97.17**	**96.51**
Gender (male)	0.00	**−2.87**	**−3.55**	**−3.23**
Place of residence (outside Marche)	**−2.00**	**−2.62**	**−4.61**	**−5.89**
Place of residence (Marche region)	0.00	0.12	1.11	1.27
Courses attendance (regular)	1.00	**3.12**	**3.28**	**3.14**
Foreign experience (yes)	0.00	−0.85	−2.00	**−5.46**
Working student	0.00	−0.75	0.00	0.26
Years to get a degree	0.00	**−0.50**	**−1.28**	**−1.82**
Diploma mark	0.00	**0.12**	**0.17**	**0.16**

is characterized by a reference quantile equal to 0.085, the negative effect of living outside Marche reduces the degree mark of 6.58 marks for student with a low degree mark, i.e. for the 8.7 % of students with the lowest marks. Moving toward the center of the conditional distribution, the effect is still negative but with a substantial numerical decrease. A foreign experience negatively influences the degree mark. This effect becomes null in group $G1$ pointing out that very good students are less influenced by their university experiences abroad.

It is worth to mention the peculiarities of $G1$ describing the effect of the covariates on the best performer students ($\theta = 0.800$). Most of the regressors do not play any effect on the 80th conditional percentile of the degree mark, which is a sign that the highest performances are related to other student features not included in the analysis.

To further highlight the potentialities of the proposed approach, it is useful to compare the observed and the estimated response values. In particular, if the best model for a given group is used to predict the response variable for the units belonging to another group, results worsen as much as the groups differentiate with respect to the best quantiles as shown in Fig. 5. The figure is structured in

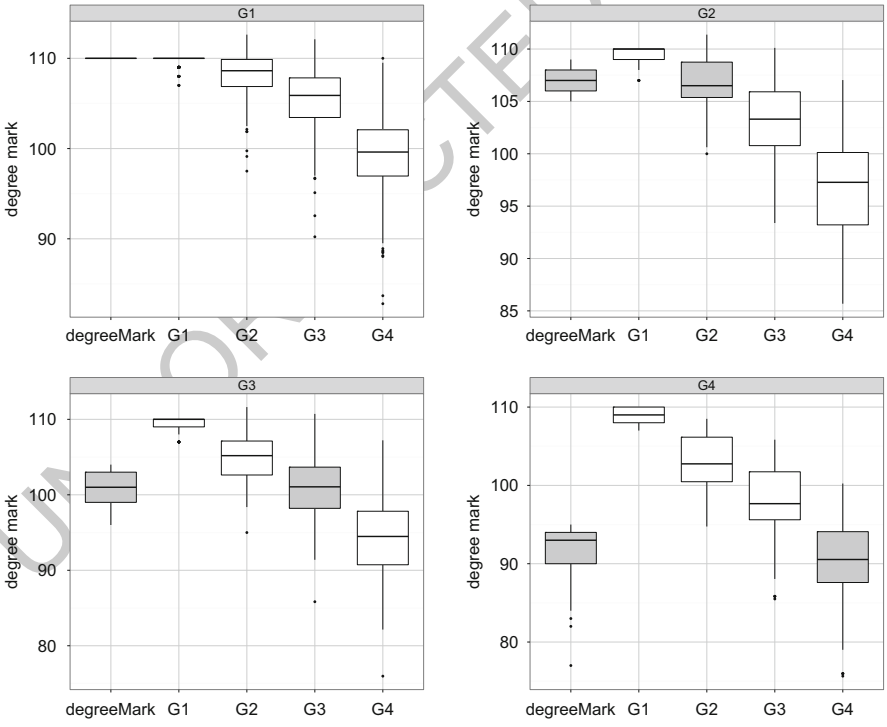

Fig. 5 Observed response distribution compared with the estimated distributions using the reference quantile of each group. Each panel depicts the degree mark of the students belonging to the group represented in the *grey label*. The two matching boxplots (observed and estimated) are colored in *grey*

four panels, one for each group. For example, the left-top panel refers to students belonging to group $G1$. From the third step of the proposed strategy (Sect. 3.3), the reference quantile for such a group is equal to 0.800. The left-most boxplot depicts the observed degree mark distribution in $G1$; the others show the estimated degree mark distributions obtained exploiting the reference quantiles associated to each group. It is evident how the estimated degree mark matches the observed degree mark using the model based on the reference quantile equal to 0.800. The above mentioned peculiarity of $G1$ (highest performance students) explains the flattened shape of the two matching boxplots. From the analysis of the other panels, it is evident how, in each panel, the observed matches with the estimated distribution obtained using the reference quantile associated to the specific group. To facilitate the reading, the two matching boxplots (observed and estimated) are colored in grey.

4 Concluding Remarks

The proposed approach provides a clustering of units according to the conditioned distribution of the dependent variable estimated through QR. It can represent a valuable tool to cluster units taking into account the dependence structure in the data. The underlying idea leading our approach relies on the expected observation that the dependence structure is affected by the features of the involved units. Obtaining a partition starting from the QR conditional distributions allows to group units where the effect played by the covariates on the dependent variable is similar.

The main strengths of the proposed approach are represented by the use of the whole sample to estimate the group dependence structures and the association of each group with a specific conditional quantile. The former point enables to easily test the statistical significance of the differences among the group. The latter provides a characterization of each group through the identification of a reference quantile describing the impact on the specific location of the dependent variable played by the features of the units belonging to the considered group. Finally, as the approach is embedded in the regression framework, the interpretation of the results can exploit the well-known rules of any linear model.

Avenues for further developments concern: *(i)* a proper comparison with the clusterwise linear regression, with whom we share the goal to simultaneously identify a partition of the data and the related model; *(ii)* a testing of the robustness of the method with respect to the number of groups, the distribution of the variables involved in the model and the model complexity through a simulation study.

References

AQ5

1. Bruzzese, D., Vistocco D.: DESPOTA: DEndrogram slicing through a pemutation test approach. J. Classif. (~~accepted for publication~~)
2. Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Applications. Hoboken, NJ, Wiley (2013)
3. Davino, C., Vistocco, D.: The evaluation of University educational processes: a quantile regression approach. Statistica **3**, 267–278 (2007)
4. Davino, C., Vistocco, D.: Quantile regression for the evaluation of student satisfaction. Italian J. Appl. Stat. **20**, 179–196 (2008)
5. DeSarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. J. Classif. **5**, 249–282 (1988)
6. Gordon, A.D.: Classification: Methods for the Exploratory Analysis of Multivariate Data. Chapman & Hall, London (1981)
7. Gould W.W.: sg70: Interquantile and simultaneous-quantile regression. Stata Tech. Bull. **38**, 14–22. Reprinted in Stata Tech. Bull. Reprints **7**, 167–176 College Station, TX: Stata Press (1997)
8. Grun, B., Leisch, F.: Fitting finite mixtures of linear regression models with varying & fixed effects in R. In: Rizzi, A., Vichi, M. (eds.) In: COMPSTAT 2006 - Proceedings in Computational Statistics, vol. 853–860. Physica Verlag, Heidelberg, Germany (2006)
9. Gurrutxaga, I., Albisua, I., Arbelaitz, O., Martn, J. I., Muguerza, J., Prez, J.M., Perona, I.: SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. Pattern Recogn. **43**(10), 3364–3373 (2010)
10. Hartigan, J. A.: Clustering Algorithms. Wiley, New York (1975)
11. Hennig, C.: Identifiability of models for clusterwise linear regression. J. Classif. **17**, 273–296 (2000)
12. Kim, M., Ramakrishna, R.S.: New indices for cluster validity assessment. Pattern Recogn. Lett. **26**(15), 2353–2363 (2005)
13. Koenker, R.: Quantile Regression. Econometric Society Monographs. Cambridge University Press, Cambridge (2005)
14. Koenker, R., Basset, G.W.: Regression quantiles. Econometrica **46**, 33–50 (1978)
15. Koenker, R., Dorey, W.V.: Algorithm AS 229: computing regression quantiles. J. R. Stat. Soc. Ser. C (Appl. Stat.) **36**(3), 383–393 (1987)
16. Milligan, G.W.: A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika **46**(2), 187–199 (1981)
17. Parzen, M.I., Wei, L., Ying, Z.: A resampling method based on pivotal estimating functions. Biometrika **81**, 341–350 (1994)
18. Scott, D.W.: On optimal and data-based histograms. Biometrika **66**, 605–610 (1979)
19. Spath, H.: Algorithm 39: clusterwise linear-regression. Computing **22**, 367–373 (1979)
20. Spath, H.: Correction to algorithm 39: clusterwise linear-regression. Computing **26**(3), 275 (1981)
21. Sturges, H.A.: The choice of a class interval. J. Am. Stat. Assoc. **21**, 6566 (1926)
22. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. B **83**(2), 411–423 (2001)
23. Wu, K.-L., Yang, M.-S., Hsieh, J.-N.: Robust cluster validity indexes. Pattern Recogn. **42**(11), 2541–2550 (2009)
24. Zhen, Z., Yan, L., Nan, K.: Clusterwise linear regression with the least sum of absolute deviations - An MIP approach. Int. J. Oper. Res. **9**(3), 162–172 (2012)

AUTHOR QUERIES

AQ1. Please note that the first author has been treated as corresponding author. Please check.
AQ2. Please check the sentence “This analysis regards...”.
AQ3. Please check the end part of the sentence “The literature concerning...”.
AQ4. The term “University” has been lower cased throughout. Please check throughout and update as required.
AQ5. Please update Ref. [1].