HANDLING HETEROGENEITY AMONG UNITS IN QUANTILE REGRESSION

Cristina Davino [1] and Domenico Vistocco[2]

[1] Department of Political Sciences, Communications and Intern. Relations, University of Macerata, (e-mail: cristina.davino@unimc.it)

[2] Department of Economics and Law, University of Cassino, (e-mail: vistocco@unicas.it)

KEYWORDS: Quantile regression, group effects, cluster analysis.

1 Introduction

In many real data applications statistical units belong to different groups and a statistical model should be tailored to incorporate and exploit this heterogeneity among units. This is also the case of the analysis of the relationship between a response variable and a set of regressors that cannot be carried out by neglecting the membership of the units to the different groups. Several approaches have been proposed in the literature to analyze group effects in a dependence model (the use of dummy variables to denote group membership or multilevel models among the others). All of them share the aim to inspect how the group structure affects the impact of the regressors on the dependent variable, without providing details on the dependence structure inside the groups. Moreover, they are tailored for the estimation of the average effects.

To estimate group effects at different points of the response conditional distribution, Davino & Vistocco, 2008 proposed to exploit quantile regression (QR) (Koenker & Basset, 1978) (Davino *et al.*, 2013), a method that is able to model the entire conditional distribution of a response variable. This paper discusses strengths and properties of such proposal through a simulation study.

2 A quantile approach to handle heterogeneity among units

Let us consider a data structure composed of a dependent variable vector $\mathbf{y}_{[n]}$ and a matrix $\mathbf{X}_{[n\times p]}$ of regressors, where n denotes the number of units and p the number of regressors. Each unit belongs to one of m groups ($g = 1,\ldots,m$). Group effects can be modeled in the framework of QR. At this regard, the first step consists in estimating the global dependence structure through a QR model, $Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\beta}(\theta)$, where $0 < \theta < 1$ and $Q_\theta(.|.)$ is the conditional quantile function for the θ^{th} quantile. Although different functional forms can be used, the paper will refer to linear regression models.

The previous step provides a coefficient matrix $\hat{\Theta}_{[p\times k]}$ with a generic element that can be interpreted as the rate of change in the θ^{th} quantile of the conditional distribution of the dependent variable per unit change in the value of the j^{th} regressor holding down constant the other regressors. The value of k is the number of estimated conditional quantiles. Although it is theoretically possible to extract infinite quantiles, a finite number is numerically distinct in practice, the so-called quantile process (Koenker & D'Orey, 1988).

In the second step, the coefficients matrix $\hat{\Theta}$ and the regressors data matrix $\mathbf{X}$ are used to estimate the conditional distribution matrix of the response variable $\hat{\mathbf{Y}}_{[n\times k]}$ whose generic element is the estimate of the response variable in correspondence to the i^{th} unit according to the θ^{th} quantile. Among such estimates, the quantile providing the estimate closer to the observed value determines the best model for each unit and is denoted as the best quantile.

The best estimated vector, $\hat{\mathbf{y}}_{\theta}^{best}$, is then defined by extracting from the $\hat{\mathbf{Y}}$ matrix the values corresponding to the best quantile assigned to each unit. To identify the best model for each group - θ_g^{best} - the best quantiles assigned to the units belonging to each group are summarised through a proper location index such as the median which is more in line with the spirit of QR.

In the last step, QR is again executed on the total sample using only the m quantiles θ_g^{best} assigned to the m groups in the previous step. The aim of this final step is to detect the group dependence structure through the inspection of a single matrix $\hat{\Theta}_{[p\times m]}^{best}$ whose generic element $\hat{\beta}_{j\theta_g^{best}}$ provides the effect of the j^{th} regressor in the g^{th} group represented by the quantile θ_g^{best}.

As the estimation process of the group dependence structure involves the whole sample, the comparison among models obtained from different samples can be easily realised. Furthermore, the approach provides a characterization of each group through the identification of a conditional quantile that mostly represents the group.

3 A simulation study: description and main results

The main features of the proposed approach and its capability to discover group effects are described through a simulation study. The aim is to explore the robustness of the method with respect to the degree and type of overlapping among the groups, the model complexity (simple or multiple QR), the sample size, the cardinality of each group (equal or unbalanced) and the number of groups. For the sake of brevity, this paper only shows results related to the case of one regressor, two groups, a sample size equal to 100 and equal number of observations in each group. A set of scenarios is generated varying the degree and the type of overlapping between the groups. Figure 1 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 (parallel group structures - case 1, group structures crossing outside the considered range of the regressor - case 2 and group structures crossing inside the considered range of the regressor - case 3). 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 9 considered scenarios are detailed in Table 1 (columns coefficient) where β_0 and β_1 represents respectively the intercept and the slope of each model. For example the scatter plot in the upper left part of Figure 1 refers to the following models:

$$\mathbf{y}_1=5+2\mathbf{x}_1+\mathbf{e} \text{ and } \mathbf{y}_2=25+10\mathbf{x}_2+\mathbf{e}$$

where $\mathbf{x}_1 \sim N(10;1)$, $\mathbf{x}_2 \sim N(10;1)$ and $\mathbf{e} \sim N(0;1)$. Data pertaining to the two groups are stacked obtaining a unique dependent variable and a unique regressor observed on 100 units. The approach described in the previous

Table 1. *Coefficients and estimates related to the 9 considered scenarios.*

		case 1a		case 1b		case 1c	
		coefficient	estimate	coefficient	estimate	coefficient	estimate
Group 1	β_0	5.0	3.7	310.0	308.7	5.0	5.7
	β_1	2.0	2.1	10.0	10.1	3.0	3.1
Group 2	β_0	25.0	30.4	280.0	285.4	15.0	17.5
	β_1	10.0	9.5	10.0	9.5	7.0	6.5
		case 2a		case 2b		case 2c	
		coefficient	estimate	coefficient	estimate	coefficient	estimate
Group 1	β_0	310.0	308.0	5.0	4.8	315.0	310.7
	β_1	10.0	9.9	3.0	3.2	10.0	10.0
Group 2	β_0	250.0	258.5	7.0	9.8	250.0	262.6
	β_1	10.0	9.7	10.0	9.1	10.0	9.7
		case 3a		case 3b		case 3c	
		coefficient	estimate	coefficient	estimate	coefficient	estimate
Group 1	β_0	300.0	298.7	400.0	372.7	310.0	308.7
	β_1	3.0	3.4	2.0	2.1	-5.0	-2.3
Group 2	β_0	250.0	255.4	250.0	362.5	250.0	254.8
	β_1	10.0	8.7	10.0	9.51	10.0	-1.4

section is carried out on each scenario. A dense grid of quantiles is exploited to estimate the global dependence structure and, after the identification of the best model for each unit, the best model for each group has been computed as the median of the best quantiles assigned to the units belonging to each group. Table 1 (columns estimate) shows results provided by the proposed

procedure for each considered scenario. The estimates are obtained performing a QR on the whole sample considering only the two quantiles representing the groups. The comparison of the original coefficients with the estimates allows to identify how the method is able to correctly capture the dependence structure in each group. The results show that the proposed approach is able to detect heterogeneity among units in all the simulated scenarios with the exception of case 3c. The results begin to degrade also in case 3b. In both the cases, the data structure is not very suitable for QR features. Showed results are related to a single dataset for each scenario. The complete simulation study will include a Monte Carlo approach and the empirical distributions of the coefficients derived from the replications will be analysed.

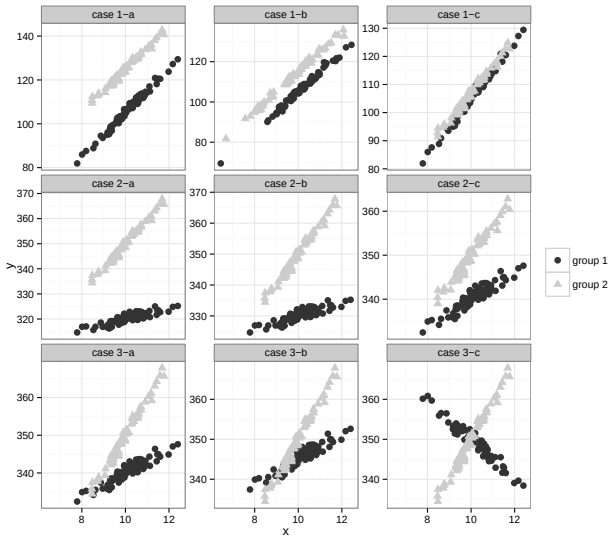

Figure 1. *Scatter plots for each scenario of the simulation study.*

References

DAVINO, C., & VISTOCCO, D. 2008. Quantile regression for the evaluation of student satisfaction. *Italian Journal of Applied Statistics*, **20**, 179–196.

DAVINO, C., FURNO, M., & VISTOCCO, D. 2013. *Quantile Regression: Theory and Applications*. New York: Wiley.

KOENKER, R, & BASSET, G.W. 1978. Regression Quantiles. *Econometrica*, **46**, 33–50.

KOENKER, R, & D'OREY, W.V. 1988. Algorithm AS 229: Computing Regression Quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **36**, 383–393.