Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018
Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors
Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018

Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors

Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA Press

ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference : Naples, 24-26 October 2018 / editors Stefania Capecchi, Francesca Di Iorio, Rosaria Simone. – Napoli : FedOAPress, 2018. – (Scuola di Scienze Umane e Sociali. Quaderni ; 11).

Accesso alla versione elettronica:
http://www.fedoabooks.unina.it

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
ISSN Collana: 2499-4774

Comitato scientifico
Enrica Amaturo (Università di Napoli Federico II), Simona Balbi (Università di Napoli Federico II), Antonio Blandini (Università di Napoli Federico II), Alessandra Bulgarelli (Università di Napoli Federico II), Adele Caldarelli (Università di Napoli Federico II), Aurelio Cernigliaro (Università di Napoli Federico II), Lucio De Giovanni (Università di Napoli Federico II), Roberto Delle Donne (Università di Napoli Federico II), Arturo De Vivo (Università di Napoli Federico II), Oliver Janz (Freie Universität, Berlin), Tullio Jappelli (Università di Napoli Federico II), Paola Moreno (Université de Liége), Edoardo Massimilla (Università di Napoli Federico II), José Gonzàlez Monteagudo (Universidad de Sevilla), Enrica Morlicchio (Università di Napoli Federico II), Marco Musella (Università di Napoli Federico II), Gianfranco Pecchinenda (Università di Napoli Federico II), Maria Laura Pesce (Università di Napoli Federico II), Domenico Piccolo (Università di Napoli Federico II), Mario Rusciano (Università di Napoli Federico II), Mauro Sciarelli (Università di Napoli Federico II), Roberto Serpieri (Università di Napoli Federico II), Christopher Smith (British School at Rome), Francesca Stroffolini (Università di Napoli Federico II), Giuseppe Tesauro (Corte Costituzionale)

© 2018 FedOAPress - Federico II Open Access University Press

Università degli Studi di Napoli Federico II
Centro di Ateneo per le Biblioteche "Roberto Pettorino"
Piazza Bellini 59-60
80138 Napoli, Italy
http://www.fedoapress.unina.it/

Published in Italy

Gli E-Book di FedOAPress sono pubblicati con licenza
Creative Commons Attribution 4.0 International

Exploring synergy between CUB models and quantile regression: a comparative analysis through continuousized data

Cristina Davino*, Rosaria Simone**, Domenico Vistocco***

Abstract: The paper investigates a parallel between CUB models and quantile regression through an illustrative case study on rating data. While CUB models have been proposed for modeling ordinal variables, quantile regression is mostly convenient for quantitative responses. The goal is to advance a comprehensive approach in which discrete ordinal outcomes on one hand and their continuousized version on the other coexist so to take advantage of two modern modeling frameworks.

Keywords: CUB models, Quantile Regression, Continuousized data.

1. Introduction and Motivation

The generalization of empirical findings from average is one of the factors that generates the common sense of diffidence about Statistics in the layman. It is efficiently described in the flaw of averages: “plan based on the assumptions that average conditions will occur are usually wrong” (Savage, 2002). The focus on the mean is a widespread approach even among insiders, since most applied Statistics is related to the estimation of average effects. The sentence that introduces regression in the book of Mosteller and Tukey (1977) is a clear invitation for insiders to go beyond the mean: “Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions”. The if and how the insiders have welcomed (and will welcome) this invitation can help to dissipate layman’s mistrust in statistical tools. The flaw of averages is becoming increasingly important in recent times because of the huge data dimension and of the complexity of the relationships among the data itself (Aguilar, 2018).

*University of Naples Federico II, cristina.davino@unina.it
**University of Naples Federico II, rosaria.simone@unina.it
***University of Cassino and Southern Lazio, vistocco@unicas.it

In this framework, Quantile Regression (QR), which was introduced as far back as 1978 (Koenker and Basset 1978), can be revitalised and regarded as one of the most modern and challenging methods in the era of big data. QR is based on the estimation of a set of conditional quantiles of a response variable as a function of a set of covariates. The method allows to verify if the effect played by the regressors varies on the low, middle and upper parts of the dependent variable thus suggesting different interpretation paths and revealing a scale and/or shape effect (Davino et al. 2014). If on one hand QR can be considered complementary to classical ordinary least squared regression (OLS), on the other hand the method represents a proper and suitable option when the homoschedastic assumption of the classical regression model cannot be satisfied, if the dependent variable has a skewed distribution or in presence of outliers. Nevertheless QR and the implied interpretation are not always suitable for ordinal data analysis, especially in cases where the response is on a finite discrete support and with a low number of possible answers. This is very common in survey data where the number of categories typically ranges from 5 up to 10 and thus a straightforward quantile modeling can raise some issues being not always greatly informative. In this respect, a manyfold perspective can be adopted with CUB models (D'Elia and Piccolo, 2005). The main feature of this class of models is the parsimonious yet flexible specification of both perceptual and decisional aspects of the rating process as a mixture of *feeling* and *uncertainty* directly on the measurement scale (Piccolo *et al.*, 2018). Thus, both QR and CUB models are appealing statistical frameworks for the analysis of evaluation–type data, for continuous and ordinal responses respectively. This contribution aims to investigate the connection between the two approaches. A combined analysis of CUB models and QR can be pursued if continuous variables are collected and then discretized, or conversely if genuine ordinal outcomes are *continuosized.* We opt here for the latter strategy, exploiting a solution proposed by Tamhane et al. (2002). In particular, let R be an ordinal variable collected on a rating scale coded with integers $1, \dots, m$ ($m > 3$). If n_j is the observed cell count for $R = j$, then continousized data in $[0, 1]$ can be obtained by uniformly spreading such observation in the interval $(\frac{j-1}{m}, \frac{j}{m}]$ to be then rescaled in the interval $[1, m]$. The approach can be easily adapted in case categories are not equally spaced. In the following a brief introduction of the two methods, CUB and QR, along with remarks on their possible integrated use will be provided through an illustrative case study on rating data.

2. *CUB and QR in a nutshell*

For quantitative variables measuring latent traits like happiness, social behaviors, self-evaluations, and so on, it is often preferable to pursue a discretization to summarize the phenomenon into ordered categories. Since in these cases it is of primary importance to understand the psychological mechanism driving the response process, the framework of CUB models offers advantageous interpretation of results by allowing a combined modeling of perceptual and decisional aspects of the choice. The rationale of this class of models is that each respondent has a propensity to provide a deliberate answer which is unavoidably mixed with the indeterminacy produced by the discretization of the latent trait. As an extreme circumstance, such indeterminacy collapses to a random choice. Thus, if R_i is the rating response given by the i–th subject and collected on a rating scale coded with integers $1, \dots, m$ ($m > 3$), then a two-component mixture is specified between a shifted binomial and a discrete uniform distribution:

$$Pr\big(R_i = r \mid \pi_i, \xi_i\big) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} \, + \, (1-\pi_i)\frac{1}{m},$$

$$logit(\pi_i) = \beta^{'} y_i, \qquad logit(\xi_i) = \gamma^{'} w_i,$$

where y_i, w_i are subjects' covariates specified to identify response profiles. The parameter ξ_i is referred to as the *feeling* parameter since $1 - \xi_i$ measures the preference of a category over the lower ones in a sequence of pairwise comparisons among categories. The mixing weight π_i, instead, is called the *uncertainty* parameter since $1 - \pi_i$ measures the overall uncertainty of the respondent's assessment: then, in particular, the larger it is, the higher the overall heterogeneity in the response distribution. ML estimation of parameter can be implemented by running the EM algorithm, and significance of variables' effects can be checked according to Wald test (Piccolo, 2006). Quantile regression has been instead proposed to model the whole conditional distribution of a response y given a set of p covariates $\mathbf{X}$, data observed on n units. Although models to deal with binary, nominal and categorical responses recently appeared in literature, QR is mostly used in case of numerical responses. In this paper we restrict our consideration to the case of linear effects. In such a case, QR estimates separate linear models for different quantiles $\theta \in [0, 1]$:

$$y_i(\theta) = \beta_0(\theta) + \mathbf{x}_i^\top \beta(\theta) + \epsilon_i, \tag{1}$$

such that $P(\epsilon_{i\theta} \leq 0) = \theta$ and $i = 1, \ldots, n$. The separate models are interpretable in terms of regression models for the quantiles of the response. The conditional distribution of the response can be estimated using a dense set of conditional quantiles. QR is distribution free since it does not pose any parametric assumption for the error (and hence response) distribution. The coefficients are commonly estimated through a variant of the simplex algorithm, while interior–point methods are especially suitable to deal with large scale problems (Koenker, 2005). Alternative estimators have been recently proposed exploiting the asymmetric Laplace distribution as a convenient model for the error distribution, thus allowing to embed QR in the likelihood framework and to extend it in a bayesian approach (Furno, Vistocco, 2018). As regards inference, QR estimators are asymptotically normal distributed with different forms of the covariance matrix depending on the model assumptions; resampling methods being a valid and widespread option.

3. The case study on relational goods and leisure time

The combined analysis between CUB models and QR will be discussed on the basis of a self-evaluation of the family making ends meet collected during a survey at University of Naples Federico II in December 2014. The purpose of the survey was to carry out an observational study on relational goods and activities for leisure time. Questionnaires were filled by students who were in turn asked to administer it also to acquaintances of theirs, according to a snowball sampling scheme. Every participant was asked to evaluate family end meet (from now, *EndsMeet*) on a 10-point Likert scale, ranging from 1 = 'never, at all' to 10 = 'always, a lot'. In the end, a sample of $n = 2181$ observations is considered. The goal is investigating the effects that the following covariates have on *EndsMeet*: *Child* and *Residence*, two dichotomous factors respectively with level 1 if there is any child aged less than 12 in the family and if the respondent lives in Naples or in its province. The solution proposed by Tamhane et al. (2002) has been used to transform the ordinal variable *EndsMeet* into continuousized data, so to use it as response variable in the QR model. Figure 2 (left-hand side) shows the observed frequency distribution of *EndsMeet*, with the kernel density of the corresponding continuousized data superimposed. The distribution of continuousized *EndsMeet* in the categories of the two covariates is shown in the right-hand part of Figure 2. The distribution of the response variable appears asymmetric in the group of families with at least one child less than 12 years old. It is worth of mention that just 20% of the interviewed belongs to this category and that almost 74% lives in Naples or in its province. The complete dataset with detailed description of all the collected variables is loaded within the R package `CUB` (Iannario *et al.*, 2018), which has been used for CUB models, tests and validation; for quantile regression, the R package `quantreg` (Koenker, 2018) has been used.

3.1. A parallel between CUB and QR results

The simplest QR model with a dichotomous regressor can help in testing the synergy between QR and CUB. Table 1 (first block of rows) reports the OLS and QR coefficients at five chosen quantiles, θ=[0.1, 0.25, 0.5, 0.75, 0.9] in a model with only *Child* as regressor. Either the OLS and the QR coefficients are significant with p-values less than 0.001 (standard errors have been estimated through resampling methods). but QR integrates results provided by classical regression. For example, having children less than 12 years old negatively impacts on the capability to get end of the month but this effect is higher on the lowest part of the distribution (at the 10% percentile is almost twice the average effect) and it becomes negligible and not significant on the

Table 1. OLS and QR estimated parameters for the two considered models

		OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
Child	$\hat{\beta}_0$	6.33	3.08	5.01	6.57	7.96	9.11
	$\hat{\beta}_1$	-0.45	-0.88	-0.68	-0.52	-0.27	-0.10
Residence	$\hat{\beta}_0$	0.60	0.22	0.45	0.64	0.79	0.91
	$\hat{\beta}_1$	-0.03	-0.02	-0.02	-0.04	-0.04	-0.03

highest part of the distribution (estimating a much more dense of quantiles, it results that the lowest slope is equal to -1.14 and the highest to 0.006). Thus, there is evidence for heterogeneity of effects of the regressor along the measurements scale. This claim is fully supported by inspecting CUB regression fit to the ordinal data ($BIC = 9689.38$):

$$logit(1-\hat{\pi}_i) = \underset{(0.099)}{0.100} + \underset{(0.256)}{0.687}\, Child_i, \quad logit(1-\hat{\xi}_i) = \underset{(0.040)}{0.694} - \underset{(0.114)}{0.255}\, Child_i.$$

As a result, responses are more heterogeneous in case there is a child aged less than 12 years in the family (uncertainty importance in the sense of weight for the uniform distribution increases from $1 - \hat{\pi}_0 = 0.529$ to $1 - \hat{\pi}_1 = 0.697$ when switching from $Child = 0$ to $Child = 1$, whereas perceived easiness in making ends meet (as measured by $1 - \hat{\xi}_i$) decreases from 0.668 to 0.617.

A further investigation of the synergy deriving from a conjoint use of QR and CUB is realised using the second regressor, *Residence*, which is dichotomous too but with a different impact on the response variable. Indeed, it affects only the location of the distribution being statistically significant only for the feeling component (as evident also from the right panel of Figure 3):

$$1 - \hat{\pi} = \underset{(0.022)}{0.557}, \quad logit(1 - \hat{\xi}_i) = \underset{(0.084)}{0.873} - \underset{(0.094)}{0.281}\, Residence_i$$

Specifically, being resident in the metropolitan area of Naples decreases (perceived) easiness in making ends meet. The constant uncertainty level given *Residence* gains insight when looking at QR results on the continuousized response: the impact of living in Naples or in its province is negative but almost constant along the distribution (see second block of rows in Table 1). More-

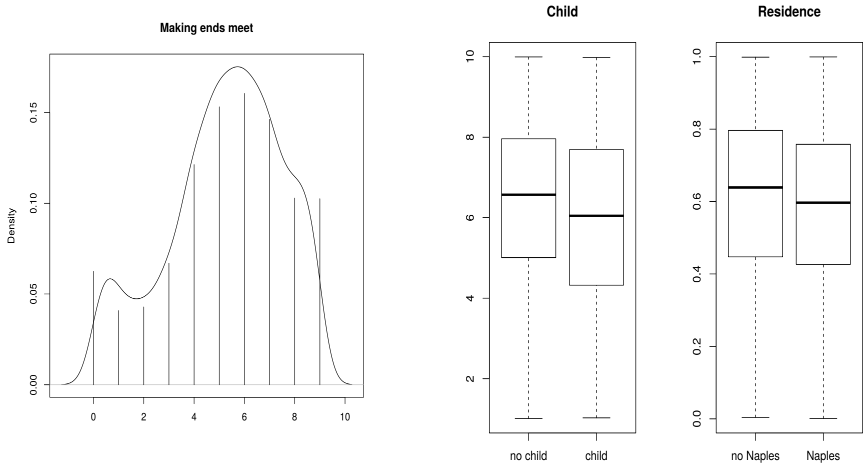

Figure 1. Rating data and continuousized version for the rating: Do you easily make ends meet? (left). Boxplot for the continuousized version given Child and Residence (right))

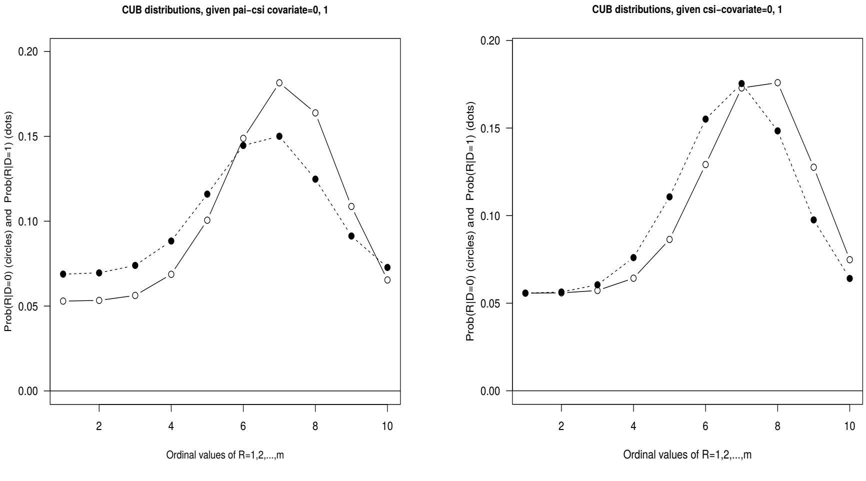

Figure 2. Conditional CUB distributions given $Child$ (left-hand side) and $Residence$(right-hand side))

over, this effect is related to very high standard errors in the lowest part of the distribution.

4. Conclusions and future research

The paper has advanced a comparative application of quantile regression methods for quantitative responses and CUB models for rating data: continuousized data allows to switch from one setting to the other with the goal of understanding mutual advantages, analogies and differences of the two approaches. Particular emphasis has been given to the interpretation of uncertainty and heterogeneity of regressors' effects. In this vein, future research developments can be outlined by simulation studies and more challenging empirical evidence.

References

Aguilar S. J. (2018) Learning Analytics: at the Nexus of Big Data, Digital Innovation, and Social Justice in Education, *TechTrends*, 62, 37-45.

D'Elia A., Piccolo D. (2005) A mixture model for preference data analysis, *Computational Statistics and Data Analysis*, 49, 917-934.

Davino C., Furno M., Vistocco D. (2014) *Quantile Regression: Theory and Applications*, John Wiley & Sons.

Furno M., Vistocco D. (2018) *Quantile Regression: Estimation and Simulation*, John Wiley & Sons.

Iannario M., Piccolo D., Simone R. (2018) CUB: A Class of Mixture Models for Ordinal Data, R package version 1.1.2. https://CRAN.R-project.org/package=CUB.

Koenker R.W., Basset G. (1978) Regression quantiles, *Econometrica*, 46, 33-50.

Koenker, R. (2005) *Quantile Regression*, Econometric Society Monographs, Cambridge: Cambridge University Press.

Koenker R. (2018) quantreg: Quantile Regression. R package version 5.35. https://CRAN.R-project.org/package=quantreg

Mosteller F., Tukey J. (1977) *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison–Wesley.

Piccolo D. (2006) Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33-78.

Piccolo D., Simone R., Iannario M. (2018) Cumulative and CUB models for rating data: a comparative analysis, *International Statistical Review*, forthcoming.

Savage, S. (2002) The flaw of averages, *Harvard Business Review*, 80, 20-22.

Tamhane A., Ankemanman B., Yang Y. (2002) The Beta distribution as a latent response model for ordinal data (I): Estimation of location and dispersion parameters, *Journal of Statistical Computation and Simulation*, 72, 473 - 494.