SIS - CLADAG

CLADAG 2015

10° Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula, October 8-10, 2015

BOOK OF ABSTRACTS

Editors:
Francesco Mola, Claudio Conversano

university press
ricerche sociali

Università degli Studi di Cagliari

Fondazione
Banco di Sardegna

CUEC

CLADAG 2015

10th Scientific Meeting of the
Classification and Data Analysis Group
of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula, October 8-10, 2015

BOOK OF ABSTRACTS

Editors:

Francesco Mola,
Claudio Conversano

CUEC
editrice

CUEC Editrice
by Sardegna Novamedia Soc. Coop.
Via Basilicata n. 57/59
09127 Cagliari, Italy
Tel. & Fax (+39) 070 271573
www.cuec.eu
info@cuec.eu

ISBN: 978-88-8467-949-9
Euro 6,90

eBook Design: Giovanni Caprioli www.servizi-per-editoria.it • info@servizi-per-editoria.it

Table of Contents

Likelihood criterion [*Marco Bertoletti, Nial Friel and Riccardo Rastelli*]

Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood [*Jean-Patrick Baudry*]

On the different ways to compute the Integrated Completed Likelihood criterion [*Gilles Celeux*]

- Exploring relationships between blocks of variables [*Organizer and Chair: Giorgio Russolillo*]

Weighted Multiblock Clustering [*Ndéye Niang, Mory Ouattara*]

Thematic Model Exploration through Multiple Co-Structure maximisation: Method and Software [*Xavier Bry, Thomas Verron*]

A New Component-based Approach of Regularisation for Multivariate Generalised Linear Regression [*Catherine Trottier, Xavier Bry, Frederic Mortier, Guillaume Cornu*]

SOLICITED SESSION

- Advances in Density-based clustering [*Organizer and Chair: Francesca Greselin*]

A Nonparametric Clustering method for Image Segmentation [*Giovanna Menardi*]

Robust Clustering for Heterogenous Skew Data [*Luis A.*

Posterior predictive model checks for assessing the goodness of fit of Bayesian multidimensional IRT models [*Mariagiulia Matteucci, Stefania Mignani*]

International tourism in Italy: a Bayesian Network approach [*Federica Cugnata, Giovanni Perucca*]

Clustering upper level units in multilevel models for ordinal data [*Leonardo Grilli, Agnese Panzera, Carla Rampichini*]

- Functional data analysis for environmental data [*Organizer and Chair: Tonio Di Battista*]

Clustering Spatially dependent Functional Data: a method based on the concept of spatial dispersion function of a curve [*Elvira Romano, Antonio Balzanella, Rosanna Verde*]

Two case studies on object oriented spatial statistics [*Piercesare Secchi, Simone Vantini, Valeria Vitelli*]

Inference on functional biodiversity tools [*Tonio Di Battista, Francesca Fortuna, Fabrizio Maturo*]

- Advances in quantile regression [*Organizer and Chair: Cristina Davino*]

M-quantile regression: diagnostics and parametric representation of the model [*Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati, Nikos Tzavidis*]

M-quantile regression: diagnostics and parametric representation of the model

Annamaria Bianchi[1], *Enrico Fabrizi*[2], *Nicola Salvati*[3] *and Nikos Tzavidis*[4]

[1] DSAEMQ, Università degli Studi di Bergamo, (e-mail: annamaria.bianchi@unibg.it)

[2] DISES, Università Cattolica del S. Cuore, (e-mail: Enrico.Fabrizi@unicatt.it)

[3] DEM, Università di Pisa, (e-mail: nicola.salvati@unipi.it)

[4] University of Southampton, (e-mail: n.tzavidis@soton.ac.uk)

Abstract: M-quantile regression generalizes both quantile and expectile regression using M-estimation ideas. This paper covers several topics related to estimation, model assessment and hypothesis testing that were so far neglected in the many articles about M-quantile regression methods that appeared in recent years.

Keywords: M-quantile regression, small area estimation, cluster test, likelihood ratio-type test.

1 Introduction

Quantile regression (Koenker and Bassett, 1978) represents a useful generalization of median regression whenever the interest is not limited to the description of the average relationship between a target variable and its predictors, but it encompasses the local behaviour of this relationship at different quantiles of the distribution. Similarly, expectile regression (Newey and Powell, 1987) generalizes mean regression by modelling expectiles of the dependent variable distribution.

Breckling & Chambers, 1988 introduce M-quantile (MQ) regression that uses the ideas of M-regression (Huber, 1973) to model the relationship between the dependent variable and its predictors for various M-quantiles of the distribution. Depending on the choice of the loss function, M-quantiles may reduce to ordinary quantiles and expectiles.

Chambers & Tzavidis, 2006 apply M-quantile regression to small area estimation. The distinguishing features of their approach to this application are its being robust and distribution free. Since then, a number of papers on M- quantile regression applied to small areas has been published (Bianchi and Salvati, 2014, Chambers *et al.*, 2014, Fabrizi *et al.*, 2014). Nonetheless little attention has been paid so far to the assessment of goodness-of-fit and other inferential procedures.

The main objective of this paper is to fill this gap by introducing a pseudo-R^2 diagnostic, likelihood-ratio and Wald type tests for linear hypotheses on the regression parameters. We also introduce a parametric representation of the model, which allows us to introduce an estimator of the scale parameter and, in the case of the Huber proposal 2 influence function, of the loss function's tuning constant. Finally, we introduce a test for the presence of cluster effects in the data. The finite sample properties of the proposed methods are assessed by means of simulation studies.

2 M-quantile regression

Let y be a random variable and $\mathbf{x}$ a p-dimensional random vector with first component $x_1 = 1$. The observable data $\{(\mathbf{x}_i, y_i), i = 1,\dots, n\}$ are assumed to be a random sample of size n drawn from the population; thus, they are independent and identically distributed random variables. Assuming a linear model, for any $\tau \in (0, 1)$, the M-quantile of order τ of y_i given $\mathbf{x}_i$ is defined by $MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\beta_\tau$. $\beta_\tau \in \Theta \subset \mathbb{R}^p$ is the solution to the population minimization problem

$$\min_{\beta \in \Theta} E\left[\rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \beta}{\sigma_\tau}\right)\right],$$

where $\rho_\tau(u) = | \tau - I(u < 0) | \rho(u)$, $\rho(\cdot)$ is a continuously differentiable loss function, and σ_τ is a scale parameter that characterizes the distribution of $\varepsilon_{\tau i} = y_i - \mathbf{x}_i^T\beta_\tau$. Since p is continuously differentiable and convex, the MQ regression coefficient estimator $\hat{\beta}_\tau$ can be obtained as the solution of the system of equations

$$\textstyle\sum_{i=1}^n \psi_\tau\left(\frac{y_i - \mathbf{x}_i^T \beta}{\hat{\sigma}_\tau}\right) \mathbf{x}_i = \mathbf{0},$$

where $\psi_\tau(u) = d\rho_\tau(u)/du = | \tau - I(u < 0) | \psi(u)$, with $\psi(u) = d\rho(u)/du$, and $\sigma\hat{}_\tau$ is an estimator of σ_τ. An iterative method is needed here to obtain a solution.

3 Diagnostics

The MQ-model can be partitioned as $MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}^T{}_{i1}\beta_{\tau 1} + \mathbf{x}^T{}_{i2}\beta_{\tau 2}$, where $\beta_\tau = (\beta^T{}_{\tau 1}, \beta^T{}_{\tau 2})^T$, $\beta_{\tau 1}$ is a $(p - k) \times 1$ vector and $\beta_{\tau 2}$ is a $k \times 1$ $(0 < k < p)$ vector. Interest focuses on tests of the null hypothesis $H_0 : \beta_{\tau 2} = 0$. Let $\hat{\beta}_\tau$ denote the MQ-estimator of the full model and let $\tilde{\beta}_\tau = (\tilde{\beta}^T{}_{\tau 1}, 0^T)^T$ denote the MQ-estimator under H_0.

When $\beta_{\tau 1}$ restricts to include only the intercept, a goodness-of-fit measure may be defined as

$$R^2_\rho(\tau) = 1 - \textstyle\sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \hat{\beta}_\tau}{\hat{\sigma}_\tau}\right) / \sum_{i=1}^n \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \tilde{\beta}_\tau}{\hat{\sigma}_\tau}\right).$$

This measure is the natural analog of the well-known R^2 used in ordinary least squares. It varies between 0 and 1 and it represents a local measure of goodness-of-fit for a specified MQ regression (at a fixed τ).

Under assumptions given in Bianchi & Salvati, 2015, we provide (asymptotically consistent) likelihood ratio-type and Wald test statistics for testing H_0.

4 Parametric representation

We introduce a useful parametrization for M-quantile regression. Given a loss function ρ_τ, we can define a random variable with density function

$$f_\tau(y;\mu_\tau,\sigma_\tau) = \frac{1}{\sigma_\tau B_\tau} exp\left[-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right], \quad -\infty < y < +\infty, \tag{1}$$

where

$$B_\tau = \int_{-\infty}^{+\infty} exp[-\rho_\tau(y)]\,dy < +\infty.$$

We show that μ_τ represents the M-quantile of the distribution and that under distribution (1) and when $\mu_{\tau i} = MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\beta_\tau$, the estimates of the unknown regression parameters β_τ and the scale σ_τ may be obtained by maximising the corresponding log-likelihood function. In the case of the Huber proposal 2 loss function, the density f_τ is called *Asymmetric Least Information* distribution (ALI). The ALI distribution depends on the tuning constant *c*. We propose to interpret *c* as a parameter of the density f_τ and estimate *c* together with β_τ and σ_τ by maximising the log-likelihood function. A similar approach can be applied to special other cases associated to different loss functions and tuning constants.

5 LR-type test for the presence of clustering effects

Cluster effects are typical of the analysis of hierarchical data resulting from sampling in two or more stages. The observable data can then be denoted as $\{(\mathbf{x}_{ij}, y_{ij}), i = 1,\dots, n_j, j = 1,\dots, d\}$, where d is the number of the primary units drawn from the population and $n_j > 0$ is the number of secondary (individual) units drawn from each primary unit. In this context, individual level covariates can fail to account for the specificity of units belonging to the same primary unit. Linear mixed models are a popular tool for the analysis of this type of data, in which the clustering is modelled by means of random effects.

Here we propose a new approach based on M-quantile regression and the characterization of clustering of observations first introduced by Chambers & Tzavidis, 2006. The definition of group-specific MQ-coefficients $\tau = (\tau_1,\dots, \tau_D)^T$ differs from that proposed by Chambers & Tzavidis, 2006. Within each group, τ_j is assumed to uniquely solve

$$\min_{\tau} E\left[\rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T\beta_\tau}{\sigma}\right) | j\right], \qquad (2)$$

i.e. τ_j is the minimizer of the specified objective function within group j.

For MQ models, testing for the presence of clustering effects is equivalent to testing whether the group-specific MQ-coefficients are all equal, that is H_0: $\tau_j = 0.5\ \forall j = 1,\dots, D$ against $H_A : \tau_j \neq 0.5$ for at least one j. The value 0.5 corresponds to the global minimizer. The vector of estimated MQ-coefficients $\hat{\tau} = (\hat{\tau}_1,\dots, \hat{\tau}_d)^T$ is defined as the empirical counterpart of (2). Based on this estimator, we propose an asymptotic test for testing H_0 based on the χ^2 distribution.

References

Bianchi, A., & Salvati, N. 2015. Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators. *Commun. Stat. Theory.*, 44, 2016-2429.

Breckling, J., & Chambers, R. 1988. M-quantiles. *Biometrika.*, 75, 761-771.

Chambers, R., Chandra H. Salvati N., & Tzavidis, N. 2014. Outlier robust small area estimation. *J. Roy. Stat. Soc. B.*, 76, 47-69.

Chambers, R., & Tzavidis, N. 2006. M-quantile Models for Small Area Estimation. *Biometrika.*, 93, 255-268.

Fabrizi, E., Salvati N. Giusti C. Tzavidis N. 2014. Mapping average equivalized income using robust small area methods. *Pap. Reg. Sci.*

Huber, RJ. 1973. Robust regression: Asymptotics, conjectures and monte carlo. *Ann. Statist.*, 1, 799-821.

Koenker, R., & Bassett, G. 1978. Regression quantiles. *Econometrica.*, 46, 33-50.

Newey, W.K., & Powell, J.L. 1987. Asymmetric least squares estimation and testing. *Econometrica.*, 55, 819-847.