SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Quantile Regression for Functional Data

Regressione Quantile per Dati Funzionali

Maria Franco-Villoria and Marian Scott

Abstract Quantile regression allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). We extend quantile regression to the functional case, rewriting the quantile regression model as a generalized additive model where both the functional covariates and the functional coefficients are parametrized in terms of B-splines. Parameter estimation is done using a penalized iterative reweighed least squares (PIRLS) algorithm. We evaluate the performance of the model by means of a simulation study.

Abstract La regressione quantile permette di stimare la relazione fra una variabile risposta e delle covariate considerando un qualsiasi percentile della distribuzione (condizionata alle covariate) della risposta. In questo lavoro si estende la regressione quantile al caso di dati funzionali, riscrivendo il modello di regressione come un modello additivo generalizzato dove sia le covariate funzionali che i coefficienti funzionali vengono parametrizzati attraverso B-splines. La stima dei parametri viene effettuata attraverso un algoritmo iterativo di minimi quadrati pesati. La performance del modello valutata in uno studio di simulazione.

Key words: B-splines, functional coefficient, generalized additive model, PIRLS

1 Introduction

Linear regression has the goal of estimation of the expected value of the response variable and its dependence on any set of explanatory variables. However, there

Maria Franco-Villoria
University of Torino, Italy e-mail: maria.francovilloria@unito.it

Marian Scott
University of Glasgow, UK e-mail: Marian.Scott@glasgow.ac.uk

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

might be situations in which the mean of the distribution is not informative, e.g. if one is interested in the high values of a given variable. Quantile regression [7] allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). As a result, rates of change in the response variable can be estimated for the whole distribution and not only in the mean. Quantile regression is widely used and has been applied in different fields such us finance, medicine or the environment. On the other hand, growing dimensionality of data available has stimulated the development of models for functional data [10], where the observed data are considered as a discrete realization of an underlying smooth function, i.e. a curve. In this work, we extend quantile regression to the functional case. However, the definition of a quantile in a functional data setting is not straightforward given the lack of a distribution function. An interesting proposal to define functional quantiles is that of Lopez-Pintado and Romo [8], who propose to order the curves based on their depth, where the deepest curve would correspond to the median. Quantile regression for functional data is a relatively new area of research that has only been explored in recent years, hence literature available is very limited. Cardot, Cambres e Sarda [1, 2] and Kato [6] have extended functional linear regression models to the case of quantile regression considering functional covariates and scalar response. A non-parametric version was proposed by Dabo-Niang e Laksaci [5], while Crambes, Gannoun and Henchiri [3, 4] use support vector machine methods for fitting quantile regression models where the covariates are functional and the response is scalar.

Regression models for functional data need to be addressed differently depending on whether the response variable is scalar or functional. In section 2 we discuss how the quantile regression coefficients can be estimated when the response variable is scalar, while in Section 3 we present preliminary results from a simulation study. Extrension to the functional response case is briefly discussed in Section 4.

2 The Model

For $\tau \in (0,1)$ fixed, a quantile regression model:

$$Q_Y(\tau|x(t)) = \alpha + \int_T \beta(t)x(t)dt = \alpha + \langle \beta, x \rangle$$

where Y is a scalar response variable, $x(t)$ is a functional covariate, $Q_Y(\tau|x(t))$ is the $100\tau^{th}$ quantile of the distribution of $Y|x(t)$ and $\langle , \rangle$ is the inner product. The parameters α, $\beta(t)$ can be estimated by minimizing the objective function:

$$R(\alpha, \beta(t)) = \sum_{i=1}^{n} \rho_\tau(y_i - (\alpha - \int_T \beta(t)x_i(t)dt))$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the check function and I is an indicator function.

The quantile regression model can be rewritten as a generalized additive model where both the functional covariate and the functional coefficients are parametrized in terms of B-spline basis functions. The objective function to be minimized is a sum of asymmetrically weighted absolute residuals; in the quantile regression literature, linear programming methods are used to estimate the unknown regression parameters. Instead, we approximate the absolute residuals with the squared residuals and adjust the weights accordingly. This way the regression coefficients can be estimated using a penalized iterative reweighed least squares (PIRLS) algorithm.

3 Preliminary results

We evaluate the performance of the estimating algorithm by means of a simulation study, where we consider different sample sizes, two levels of noise and various forms of complexity for the functional coefficient. We evaluate the performance at four different quantiles $q_{0.2}$, $q_{0.5}$, $q_{0.7}$ and $q_{0.9}$. The simulated data are built as

$$y_i^{sim} = \alpha + \int_T \beta(t)x_i(t)dt + \varepsilon_i.$$

The functional covariate $x(t) = \sum_{j=1}^{10} \xi_j B_j(t)$, where $B_j(t)$ are B-spline basis functions evaluated at $t \in T = [0,1] \subset \mathbb{R}$, $j = 1,\ldots,10$ and the spline coefficients $\xi_j \sim N(0,1)$. The random errors ε_i are simulated from a normal distribution $N(q_\tau, \sigma^2)$ with q_τ the $100\tau^{th}$ quantile of the $N(0,\sigma^2)$; values of σ were chosen to ensure a signal to noise ratio of 2 and 4.

To evaluate how well $\beta(t)$ is estimated, we consider two indicators, the distance (L_2 norm) between the simulated and estimated coefficient and the proportion of negative and positive residuals. Results from a preliminary simulation study suggest that the method performs well; when the sample size is small ($n = 50$) distance values range from 0.01 to 0.45 when the functional coefficient is linear and from 0.1 to 1.2 when the functional coefficient is non-linear. Results improve with increasing sample size and the closer we get to the median, as expected. The percentages of positive and negative residuals were very close to the expected $100(1-\tau)\%$ and $100\tau\%$ respectively. Convergence was reached after 4 to 29 iterations.

4 Discussion and Future Work

In this work, we propose a quantile regression model when the covariates are functional and the response is scalar. Preliminary results from a first simulation study suggest good performance for a range of different quantiles. The model can be easily extended to incorporate more covariates keeping the computational cost low thanks to the use of sparse matrix computation.

We are currently working on the case of a quantile regression model where the response is functional too. In this case, the residuals themselves are functional data and working out the weights is not as straightforward as in the scalar response case. A possibility would be to consider the distance from the zero curve as a proxy for the size of each residual, while the sign of the residual could be worked out using some sort of curve ordering technique such as band depth or a more recent proposal based on epigraphs and hypographs [9].

In particular, quantile regression for functional data could prove useful in solving the problem of uncertainty evaluation of a predicted curve, where the 2.5% and 97.5% quantiles could be used to build a functional confidence band.

References

1. Cardot, H., Crambes, C., Sarda, P.: Conditional quantiles with functional covariates: an application to ozone pollution forecasting. In: Antoch, J. (ed.) Compstat 2004 Proceedings, pp. 769776. Physica-Verlag (2004)
2. Cardot, H., Crambes, C., Sarda, P.: Quantile regression when the covariates are functions. Journal of Nonparametric Statistics **17(7)**, 841–856 (2005)
3. Crambes, C., Gannoun, A., Henchiri Y.: Support vector machine quantile regression approach for functional data: Simulation and application studies. Journal of Multivariate Analysis **121**, 50–68 (2013)
4. Crambes, C., Gannoun, A., Henchiri Y.: Modelling functional additive quantile regression using support vector machines approach. Journal of Nonparametric Statistics **26(4)**, 639–668 (2014)
5. Dabo-Niang, S., Laksaci, A.: Nonparametric Quantile Regression Estimation for Functional Dependent Data. Communications in Statistics - Theory and Methods **41(7)**, 1254–1268 (2012)
6. Kato, K.: Estimation in functional linear quantile regression. The Annals of Statistics **40(6)**, 3108–3136 (2012)
7. Koenker, R.: Quantile Regression. Cambridge University Press (2005)
8. Lopez-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Amer. Statist. Assoc. **104(486)**, 718–734 (2009)
9. Martin-Barragan, B., Lillo, R.E., Romo, J.: Functional boxplots based on epigraphs and hypographs. Journal of Applied Statistics **43(6)**, 1088–1103 (2016)
10. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. pringer, Dordrecht (2005)