Quantile Regression Coefficients Modeling: a Penalized Approach

Modelli Quantili Parametrici: un Approccio Penalizzato

Gianluca Sottile, Paolo Frumento and Matteo Bottai

Abstract Modeling quantile regression coefficients functions permits describing the coefficients of a quantile regression model as parametric functions of the order of the quantile. This approach has numerous advantages over standard quantile regression, in which different quantiles are estimated one at the time: it facilitates estimation and inference, improves the interpretation of the results, and is statistically efficient. On the other hand, it poses new challenges in terms of model selection. We describe a penalized approach that can be used to identify a parsimonious model that can fit the data well. We describe the method, and analyze the dataset that motivated the present paper. The proposed approach is implemented in the `qrcmNP` package in R.
Abstract *I coefficienti di una regressione quantilica sono funzioni iniettive dell'ordine del quantile. L'approccio standard è quello di stimare i quantili uno alla volta. Un metodo alternativo è quello di esprimere la forma funzionale dei coefficienti usando un modello parametrico. Questo approccio ha numerosi vantaggi: semplifica le procedure di stima e inferenza, migliora l'interpretazione dei risultati, e risulta statisticamente efficiente. Al tempo stesso, pone nuove sfide in termini di selezione del modello. La nostra proposta è quella di usare un metodo penalizzato che permetta di identificare un modello parsimonioso che rappresenti correttamente la funzione quantilica. In questo articolo descriviamo il metodo, e analizziamo il dataset che ha motivato il lavoro. L'approccio proposto è stato implementato nel pacchetto R `qrcmNP`.*

Key words: Lasso penalty, Penalized integrated loss minimization, Penalized quantile regression coefficients modeling, Inspiratory capacity

Gianluca Sottile
University of Palermo, Department of Economic, Business and Statistical sciences, Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Paolo Frumento and Matteo Bottai
Karolinska Institutet, Institute of Environmental Medicine, Unit of Biostatistics, Stockholm, Sweden, e-mail: paolo.frumento@ki.se, matteo.bottai@ki.se

1 Introduction

Quantiles fully describe the conditional distribution of a response variable given covariates. Quantile regression (QR; [6]) and its generalizations (e.g., [3]) are the standard tools for quantile modeling. In QR, the conditional quantile function is usually written as

$$Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}(p), \tag{1}$$

where $\boldsymbol{x}$ is a q-dimensional vector of covariates, and $\boldsymbol{\beta}(p)$ is a vector of unknown coefficients describing the relationship between $\boldsymbol{x}$ and the p-th quantile of the response variable, $p \in (0,1)$. In standard quantile regression, different quantiles are estimated one at the time. When a grid of quantiles is computed, e.g., $p = 0.01, 0.02, \ldots, 0.99$, results can only be summarized graphically. The estimated coefficients are generally non-smooth functions of p and may suffer from high volatility, which can hinder their interpretability.

Recently, [5] suggested modeling the quantile regression coefficient functions, $\boldsymbol{\beta}(p)$, by using parametric functions. Model (1) is reformulated as follows:

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\beta}(p \mid \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ is a vector of model parameters. This approach is referred to as *quantile regression coefficients modeling* (QRCM) and permits modeling the entire quantile function, while keeping the quantile regression structure expressed by equation (1). Consider, for example, describing $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$ by k-th degree polynomial functions:

$$\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1} p + \ldots + \theta_{jk} p^k, j = 1, \ldots, q.$$

Each covariate has $(k+1)$ associated parameters, for a total of $q \times (k+1)$ model coefficients. When either q or k are large, estimation may become difficult and the model may be poorly identified, causing the variability to grow out of control.

Among different approaches discussed in literature, the least absolute shrinkage and selection operator (LASSO; [9]) is the most used method to perform model selection. This procedure requires selecting a tuning parameter. In the literature, traditional criteria include cross-validation (CV), Akaike's information criterion (AIC), and Bayesian information criterion (BIC).

Numerous papers (e.g., [1, 10]) have investigated the estimation of penalized quantile regression models in high-dimensional setting using the L_1-norm of the coefficients, denoted by L_1-QR [1, 8]. These approaches, however, focus on model selection when estimating one quantile at a time. Generally, this is inefficient and makes it difficult to interpret the results, because some coefficients could be only significant at some quantiles.

We propose applying the L_1-penalty to the integrated loss function described by [5], which is minimized to estimate the unknown parameter $\boldsymbol{\theta}$ in model (2). We refer to this procedure as *penalized quantile regression coefficients modeling* (QRCMPEN).

The paper is structured as follows. We introduce a penalized estimator in Section 2, and propose criteria to select the tuning parameter in Section 3. Section 4 concludes the paper with the analysis of the dataset that motivated this research.

2 The estimator

We assume that model (2) holds, and adopt the following parametrization: $\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$, where $\boldsymbol{b}(p) = [b_1(p), \dots, b_k(p)]^{\mathrm{T}}$ is a set of k known functions of p, and $\boldsymbol{\theta}$ is a $q \times k$ matrix with entries θ_{jh} such that $\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j1}b_1(p) + \ldots + \theta_{jk}b_k(p)$, $j = 1, \dots, q$. The conditional quantile function is then rewritten as $Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\theta}\boldsymbol{b}(p)$. The choice of the vector $\boldsymbol{b}(p)$, is something arbitrary when the model is not known in advance, indeed, an intuitive approach could be to choose functions as flexible as possible. Moreover, as discussed by [5], the values of $\boldsymbol{b}(0)$ and $\boldsymbol{b}(1)$ should reflect the assumptions about the support of the outcome, and the interpretation of parameters may be highly dependent on the model specification. Although all outcomes are bounded in practice, including unbounded functions facilitates modeling the tails of the distribution. As shown by [5], estimation is carried out by minimizing

$$\overline{L}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\mathrm{d}p, \tag{3}$$

where $L(\boldsymbol{\beta}(p))$ is the loss function of standard quantile regression given by $L = \sum_{i=1}^{n} (p - I(y_i \le \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(p)))(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(p))$. This estimation procedure is referred to as *integrated loss minimization* (ILM), and implemented in the `qrcm` package in R.

This modeling approach is very flexible, and usually provides a good fit of the data. However, it tends to generate large models, causing overparametrization and loss of efficiency. To implement an automatic procedure for model selection, we modify the loss function (3) by introducing a L_1-norm penalizing factor:

$$\overline{L}^{(\lambda)}_{\mathrm{PEN}}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta})) + \lambda \sum_{j=1}^{q}\sum_{h=1}^{k} \mid \theta_{jh} \mid \mathrm{d}p, \tag{4}$$

where $\lambda \ge 0$. We refer to this estimation approach as *penalized integrated loss minimization* (PILM). To minimize $L^{(\lambda)}_{\mathrm{PEN}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we use a pathwise coordinate descent algorithm [4]. The described PILM estimator is implemented in the `qrcmNP` package in R.

3 Tuning parameter selection

With a given set of data, the true model is not known. Having adequate criteria for model selection is therefore crucial. In penalized regression, the tuning parameter

λ balances the trade-off between goodness of fit and efficiency. We denote by $\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}^{(\lambda)}$ the estimator of $\boldsymbol{\theta}$ obtained by minimizing (4) at a given value of λ. AIC- and BIC-type selectors are grid-search criteria that minimize $\text{Dev}^{(\lambda)} + c_n \cdot \text{df}^{(\lambda)}$, where $\text{Dev}^{(\lambda)}$ is the explained deviance of the model (a measure of goodness-of-fit defined below) corresponding to $\widehat{\boldsymbol{\theta}}$, c_n is a constant that could depend on the sample size n, and $\text{df}^{(\lambda)}$ reflects the number of nonzero elements of $\widehat{\boldsymbol{\theta}}$. To improve efficiency and computation, we propose standardizing both $\boldsymbol{x}$ and $\boldsymbol{b}(p)$. Following [7], we define $\text{Dev}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}})$, i.e., the logarithm of the minimized loss function given by (4). The AIC and BIC criteria are given by

$$\text{AIC}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}}) + n^{-1}\text{df}^{(\lambda)}, \tag{5}$$

$$\text{BIC}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}}) + (2n)^{-1}\log(n)\text{df}^{(\lambda)} C_n, \tag{6}$$

where C_n is some positive constant, that diverges to infinity as n increase. The value $C_n = 1$ corresponds to the ordinary BIC.

4 Variables selection for inspiratory capacity

We applied the PILM estimator to a subset (n = 2201) of the data analyzed in [2]. The data refer to a study carried out in 1988-1991 in Northern Italy, and included 1063 males and 1138 females. The study aimed to estimate percentiles of inspiratory capacity (IC), a measure of lungs function. The following nine predictors were available: age, height, body mass index (BMI), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma.

We model the intercept using a linear combination of $\log(p)$ and $\log(1-p)$, that together define the quantile function of the asymmetric Logistic distribution, a very flexible model used to describe possibly skewed random variables with heavy tails, while the coefficients associated with the covariates were described by a shifted Legendre polynomial up to third degree, inclusive of an intercept. The maximal model had $3 + 4 \times 9 = 39$ parameters. We used AIC and BIC to assess model fit. As shown by simulations, AIC criterion tends to select overparametrized models, while BIC criterion is more parsimonious with a higher ability to discard irrelevant covariates ($\boldsymbol{x}$) and basis functions ($\boldsymbol{b}(p)$). Results are reported in Table 1.

Table 1 Model selection based on AIC and BIC criteria. We report the number of parameters, the number of selected covariates, the optimal λ value, the value of the minimized loss function, and the p-value of a Kolmogorov-Smirnov goodness-of-fit test.

Criterion	n. of parameters	n. of covariates	λ	Loss	P-value KS
AIC	31/36	7/9	20.79	293.31	.77
BIC	19/36	4/9	60.47	294.01	.53

We used the model selected by BIC and estimated it again using unpenalized QRCM. The model is represented graphically in Figure 1. Because we were mostly interested in the low quantiles of IC, in Table 2 we only report the estimated quantile regression coefficients, $\widehat{\boldsymbol{\beta}}(p) = \boldsymbol{\beta}(p \mid \widehat{\boldsymbol{\theta}})$, at $p = 0.01$, $p = 0.05$, and $p = 0.50$. Age, height, BMI and sex were statistically significant. Figure 1 shows the regression coefficient functions for all covariates over the interval $p \in (0,1)$. Age had a negative effect at all quantiles, and the associated coefficient function showed an increasing linear trend. Per each one-year increase in age, the 1st and 5th percentile of IC decreased by about 0.014 and 0.013 liters respectively, while its median decreased by about 0.01 liters. Height and BMI both had a positive effect. Quantile regression coefficients were increasing, but had a non linear trend. For each one-centimeter increase in height, quantiles below the median increased by approximately 0.03 liters. For each unit increase of BMI, IC increased by 0.033 and 0.037 at the 1st and 5th percentile, respectively, and by 0.056 at the median. The coefficient function associated with the indicator of male gender was positive and increasing. This indicated that the distribution of IC in males was shifted towards upper values and had a longer right tail than that of females.

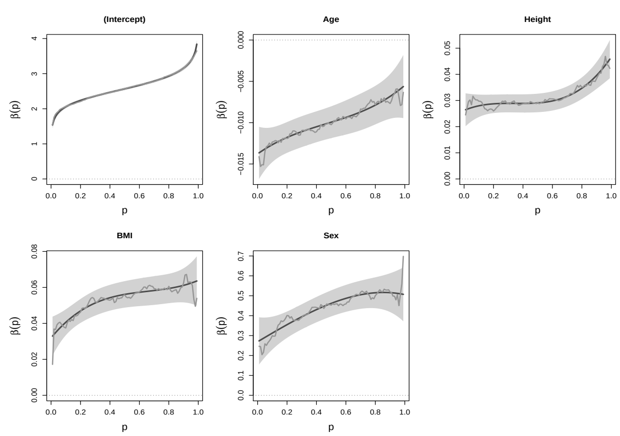

Fig. 1 ILM estimates of $\boldsymbol{\beta}(p)$ under the model selected by BIC. Confidence bands are displayed as shaded areas. The broken lines connect the coefficients of ordinary quantile regression estimated at a grid of quantiles. The dashed line indicates the zero.

Finally, comparing our proposal with standard penalized quantile regression [1, 8] we could observe that it is inefficient and makes it difficult to interpret results, as already mentioned in the introduction. Indeed, different variables are discarded for each percentile, i.e., sex and wheeze for $p = 0.01$, none for $p = 0.05$ and, smoke, occupational exposure, cough and wheeze for $p = 0.50$.

Table 2 Estimated quantile regression coefficients at $p = 0.01$, $p = 0.05$ and $p = 0.50$, obtained from the model selected by BIC. Estimated standard errors in brackets. The asterisk (*) denotes significance less than 0.05.

	$p = 0.01$	$p = 0.05$	$p = 0.50$
Intercept	1.539(0.049)*	1.893(0.029)*	2.567(0.018)*
Age	−0.014(0.002)*	−0.013(0.001)*	−0.010(0.001)*
Height	0.026(0.003)*	0.027(0.003)*	0.029(0.002)*
BMI	0.033(0.005)*	0.037(0.004)*	0.056(0.004)*
Male	0.273(0.061)*	0.291(0.051)*	0.462(0.033)*

5 Discussion

We described a penalized approach that can be applied to the QRCM framework introduced by [5]. Modeling the conditional quantile function parametrically can be more efficient than estimating quantiles one at a time, as in ordinary quantile regression. Moreover, it permits performing model selection directly on the parameters that describe conditional quantiles, instead of proceeding quantile-by-quantile, as the penalized methods for quantile regression proposed so far.

Using this approach has the disadvantage that, as each covariate has multiple associated parameters, the number of model coefficients tends to be large. The PILM estimator demonstrated to select the correct model with a high probability. A computationally efficient algorithm is implemented in the `qrcmNP` package in R.

References

1. Belloni, A., Chernozhukov, V.: L1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics **39**(1), 82–130 (2011)
2. Bottai, M., Pistelli, F., Di Pede, F., Baldacci, S., Simoni, M., Maio, S., Carrozzi, L., Viegi, G.: Percentiles of inspiratory capacity in healthy nonsmokers: a pilot study. Respiration **82**(3), 254–262 (2011)
3. Chaudhuri, P.: Global nonparametric estimation of conditional quantile functions and their derivatives. Journal of multivariate analysis **39**(2), 246–269 (1991)
4. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010)
5. Frumento, P., Bottai, M.: Parametric modeling of quantile regression coefficient functions. Biometrics **72**(1), 74–84 (2016)
6. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: journal of the Econometric Society pp. 33–50 (1978)
7. Lee, E., Noh, H., Park, B.: Model selection via bayesian information criterion for quantile regression models. Journal of the American Statistical Association **109**, 216–229 (2014)
8. Li, Y., Zhu, J.: L1-norm quantile regression. Journal of Computational and Graphical Statistics **17**(1), 163–185 (2008)
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1996)
10. Wu, Y., Liu, Y.: Variable selection in quantile regression. Statistica Sinica pp. 801–817 (2009)