

Page Proof Instructions and Queries

Journal Title: SMI

Article Number: 825523

Greetings, and thank you for publishing with SAGE. We have prepared this page proof for your review. Please respond to each of the below queries by digitally marking this PDF using Adobe Reader (free at <http://www.adobe.com/products/reader.html>).

Please use *only* the circled tools to indicate your requests and responses, as edits via other tools/methods are not compatible with our software. To ask a question or request a formatting change (such as italics), please click  the tool and then choose “Text Callout.” To access the necessary tools, choose “Comment” from the right-side menu.



No.	Query
	Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
	Please confirm that the Funding and Conflict of Interest statements are accurate.
	Please ensure that you have obtained and enclosed all necessary permissions for the reproduction of artistic works (e.g. illustrations, photographs, charts, maps, other visual material, etc.) not owned by yourself. Please refer to your publishing agreement for further information.
	Please note that this proof represents your final opportunity to review your article prior to publication, so please do send all of your changes now.
AQ: 1	Please check and confirm whether the affiliations have been correctly amended. Edited Affiliations: 1Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy. 2Institute of Environmental Medicine, Unit of Biostatistics, Karolinska Institutet, Stockholm, Sweden. Address for correspondence: Gianluca Sottile, Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Ed. 13, Palermo 90128, Italy. E-mail: gianluca.sottile@unipa.it
AQ: 2	Please provide a valid figure caption for Figure 3.

AQ: 3	We have inserted the publisher detail (cities) in the following references as per our online search. Please confirm that our changes are correct. Abramowitz, M. and Stegun, I. (1964) Edited Sentence: Abramowitz M and Stegun I (1964) Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables. Vol. 55. Chelmsford, MA: Courier Corporation. Bühlmann, P. and Van De Geer, S. (2011) Edited Reference: Bühlmann P and Van De Geer S (2011) Statistics for High-dimensional Data: Methods, Theory and Applications. Berlin: Springer Science & Business Media.
AQ: 4	We have inserted the year of publication in the reference 'Belloni, A. and Chernozhukov, V. Least squares' and its corresponding citation as per our online search. Please confirm that our change is correct. Edited Reference: Belloni A and Chernozhukov V (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli, 19, 521–547.
AQ: 5	We have inserted the volume numbers in the following references as per our online search. Please confirm that our changes are correct. Beran, R. (1982) Edited Reference: Beran R (1982) Estimated sampling distributions: The bootstrap and competitors. The Annals of Statistics, 10, 212–225. Koenker, R. and Bassett, J. (1978) Edited Reference: Koenker R and Bassett J (1978) Regression quantiles. Econometrica, 46, 33–50. Wu, Y. and Liu, Y. (2009) Edited Reference: Wu Y and Liu Y (2009) Variable selection in quantile regression. Statistica Sinica, 19, 801–817.
AQ: 6	We have updated the fourth author's name in the reference 'Lee, J., Sun, D., Sun, Y., and JE, T. (2016)' as per our online search. Please confirm that our change is correct. Edited Reference: Lee J, Sun D, SunY andTaylor JE (2016) Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44, 907–927.

A penalized approach to covariate selection through quantile regression coefficient models

AQ: 1 Gianluca Sottile¹, Paolo Frumento², Marcello Chiodi¹ and Matteo Bottai²

¹Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy.

²Institute of Environmental Medicine, Unit of Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Abstract: The coefficients of a quantile regression model are one-to-one functions of the order of the quantile. In standard quantile regression (QR), different quantiles are estimated one at a time. Another possibility is to model the coefficient functions parametrically, an approach that is referred to as *quantile regression coefficients modeling* (QRCM). Compared with standard QR, the QRCM approach facilitates estimation, inference and interpretation of the results, and generates more efficient estimators. We designed a penalized method that can address the selection of covariates in this particular modelling framework. Unlike standard penalized quantile regression estimators, in which model selection is quantile-specific, our approach permits using information on all quantiles simultaneously. We describe the estimator, provide simulation results and analyse the data that motivated the present article. The proposed approach is implemented in the `qrcmNP` package in R.

Key words: Inspiratory capacity, Lasso penalty, penalized integrated loss minimization (PILM), penalized quantile regression coefficients modelling (QRCM_p), tuning parameter selection

Received September 2018; revised November 2018; accepted December 2018

1 Introduction

Conditional quantile functions fully describe the distribution of a response variable-given covariates. Quantile regression (QR; Koenker and Bassett, 1978) and its generalizations (e.g., Chaudhuri, 1991) are popular methods to model conditional quantiles. The regression model is typically written as

$$Q(p | \mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}(p), \quad (1.1)$$

where $\mathbf{x}$ is a q -dimensional vector of covariates, and $\boldsymbol{\beta}(p)$ is a vector of unknown coefficients describing the relationship between $\mathbf{x}$ and the p th quantile of the response variable, $p \in (0, 1)$.

Address for correspondence: Gianluca Sottile, Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Ed. 13, Palermo 90128, Italy.
E-mail: gianluca.sottile@unipa.it

2 Gianluca Sottile et al.

Standard quantile regression (QR) methods can estimate different quantiles one at a time. For varying p , $\beta(p)$ is treated as an infinite-dimensional parameter. When a grid of quantiles is estimated, for example, $p = 0.01, 0.02, \dots, 0.99$, results are usually summarized graphically. The estimated coefficients are generally non-smooth functions of p and may suffer from high volatility, especially in the tails.

Recently, Frumento and Bottai (2016, 2017) suggested modelling the quantile regression coefficient functions, $\beta(p)$, as parametric functions of the order of the quantile:

$$Q(p | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\beta}(p | \boldsymbol{\theta}), \quad (1.2)$$

where $\boldsymbol{\theta}$ is a vector of model parameters. This approach is referred to as *quantile regression coefficients modeling* (QRCM) and permits modelling the quantile function parametrically, while keeping the quantile regression structure expressed by equation (1.1). Compared with standard QR, this method presents numerous advantages, which include simpler estimation and inference (owing to a smooth objective function), improved efficiency, and better interpretation of the results. Standard QR can be seen as a special case of QRCM in which $\boldsymbol{\beta}(p | \boldsymbol{\theta})$ is described by a ‘very large’ number of parameters.

In the absence of strong parametric assumptions on the shape of $\boldsymbol{\beta}(p | \boldsymbol{\theta})$, a meaningful approach is that of using some flexible model. Consider, for example, describing $\boldsymbol{\beta}(p | \boldsymbol{\theta})$ by a k th degree polynomial function:

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

Each covariate has $(k + 1)$ associated parameters, for a total of $q \times (k + 1)$ model coefficients. As q and k increase, estimation may become difficult and sampling variability may grow out of control. The dimension of the model can be reduced by selecting the most informative covariates. The j th covariate is removed from the model by letting $\theta_j = 0$ that forces $\beta_j(p | \boldsymbol{\theta}) = 0$. The goal of this article is to describe a procedure to determine the ‘optimal’ subset of covariates.

A popular approach to perform model selection is to use some form of penalized regression, in which a tuning parameter controls the trade-off between bias and variance. Well-known procedures include the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), least angle regression (LARS; Efron et al., 2004) and the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001). In literature, traditional criteria to select the tuning parameter include cross-validation (CV), generalized cross-validation (GCV; Craven and Wahba, 1979), Akaike’s information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978) and generalized information criterion (GIC; Nishii, 1984; Zhang et al., 2010).

Numerous articles (e.g., Belloni and Chernozhukov, 2011; Wang et al., 2012; Wu and Liu, 2009; Zheng et al., 2013) have investigated the estimation of penalized quantile regression (QR_p) models in high-dimensional settings. The penalization is usually given by the l_1 -norm of the coefficients, denoted by l_1 -QR (Belloni and Chernozhukov, 2011; Li and Zhu, 2008). These approaches, however, focus on model

selection when estimating one quantile at a time. This is generally inefficient and makes it difficult to interpret the results, as illustrated by the motivating example presented in Section 2. The main advantage of the *QRCM* framework is that it enables performing model selection directly on the parameters of the conditional quantile function, using information on all quantiles. We propose applying the l_1 -penalty to the integrated loss function described by Frumento and Bottai (2016), which is minimized to estimate the unknown parameter vector θ in model (1.2). We refer to this procedure as *penalized quantile regression coefficients modeling* (*QRCM_p*).

The article is structured as follows. In Section 2, we illustrate a motivating example. In Section 3, we introduce a penalized estimator. In Section 4, we present simulation results, and in Section 5, we briefly discuss computation. Section 6 concludes the article with the analysis of the dataset that motivated this research.

2 Motivating example

The motivation for this article comes from a study on the association between pulmonary inspiratory capacity (IC), a measure of lung's volume and the following risk factors: age, height, body mass index (BMI), a binary indicator of smoking, and indicators for sex, occupation exposure, cough, wheezing and asthma. Extreme quantiles of IC, and in particular very low quantiles, can be used to identify problems and inform therapy strategies. We first estimated 13 percentiles (0.01, 0.05, 0.10, 0.20, ..., 0.80, 0.90, 0.95, 0.99) with standard penalized quantile regression (*QR_p*), as implemented in the R package *rqPen* of Sherwood and Maidman (2017). Each variable was standardized to have zero mean and unit standard deviation. Results are summarized in Table 1.

The coefficients associated with strong predictors, such as age, height and BMI, were consistently included at nearly all quantiles. Other coefficients, however, were only included at some quantiles, making it difficult to interpret the results. For

Table 1 Variable selection using standard penalized quantile regression

	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
Intercept	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Height	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BMI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓
Male	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Smoker	✓	✓	-	✓	-	-	-	✓	-	✓	✓	✓	✓
Non-exposure	✓	✓	-	-	-	-	-	-	✓	-	✓	✓	✓
Cough	✓	✓	-	✓	✓	✓	-	✓	-	-	✓	✓	-
Wheezing	-	✓	✓	✓	✓	✓	-	-	-	-	✓	✓	✓
Asthma	✓	✓	✓	-	-	-	✓	-	-	-	✓	-	✓

Note: The symbol '✓' indicates that the covariate was included in the model. We estimated a grid of percentiles, namely $p = (0.01, 0.05, 0.10, 0.20, \dots, 0.80, 0.90, 0.95, 0.99)$, and utilized *bic* for model selection.

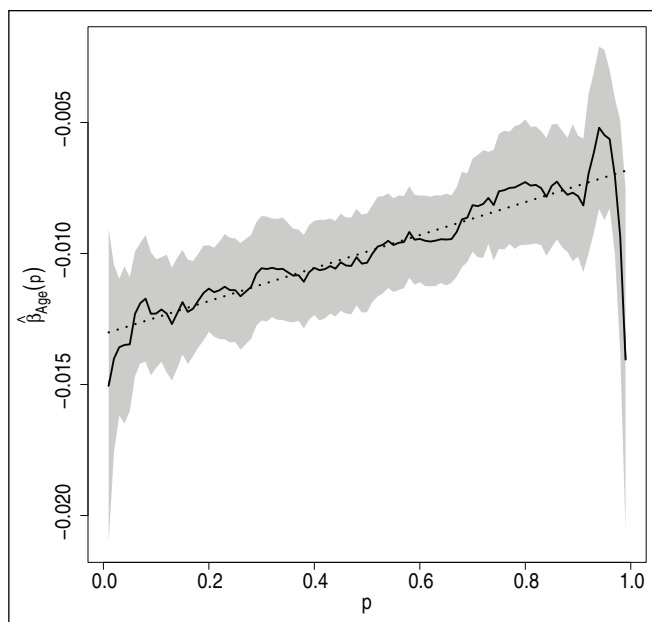


Figure 1 Estimated quantile regression coefficients associated with age

Note: The shaded area represents 95% point-wise confidence intervals. The dashed line represents a hypothetical linear trend, estimated as in Frumento and Bottai (2016).

example, the coefficient of smoking was excluded at $p = 0.5$ and $p = 0.7$, but included at $p = 0.6$ and $p = 0.8$. The observed volatility is probably accounted for by random variation and data sparsity, which represents a main source of sampling error in the tails of the distribution.

The idea behind $\mathcal{QRCM}$ is to define a more parsimonious representation of $\beta(p)$. Consider, for example, Figure 1, showing the estimated regression coefficients associated with age. The coefficient function appears to be adequately described by a linear function,

$$\beta_{\text{age}}(p \mid \theta) = \theta_0 + \theta_1 p.$$

Working with a parametric quantile function enables performing model selection using information on all quantiles. As shown in the rest of the article, this has important consequences on the performance of penalized regression methods.

3 The estimator

Throughout the article, we assume model (1.2) to hold and adopt the following linear parametrization:

$$\beta(p \mid \theta) = \theta b(p), \quad (3.1)$$

where $\mathbf{b}(p) = [b_1(p), \dots, b_k(p)]^T$ is a set of k known functions of p , and $\boldsymbol{\theta}$ is a $q \times k$ matrix with entries θ_{jh} such that $\beta_j(p | \boldsymbol{\theta}) = \theta_{j1}b_1(p) + \dots + \theta_{jk}b_k(p)$, $j = 1, \dots, q$. The conditional quantile function is rewritten as

$$Q(p | \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta} \mathbf{b}(p). \quad (3.2)$$

This modelling approach was used by Frumento and Bottai (2016), Frumento and Bottai (2017) and Yang et al. (2017), and is very flexible and computationally convenient. As shown by Frumento and Bottai (2016), estimation is carried out by minimizing

$$\bar{L}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p | \boldsymbol{\theta})) dp, \quad (3.3)$$

where $L(\boldsymbol{\beta}(p))$ is the loss function of standard QR given by

$$L(\boldsymbol{\beta}(p)) = \sum_{i=1}^n (p - I(y_i \leq \mathbf{x}_i^T \boldsymbol{\beta}(p)))(y_i - \mathbf{x}_i^T \boldsymbol{\beta}(p)).$$

This estimation procedure is referred to as *integrated loss minimization* (ILM) and implemented in the `qrqm` package in R (Frumento, 2017).

To implement an automatic procedure for model selection, we modify the loss function (3.3) by introducing a l_1 -norm penalizing factor:

$$\bar{L}_{PEN}^{(\lambda)}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p | \boldsymbol{\theta})) dp + \lambda \sum_{j=1}^q \sum_{h=1}^k |\theta_{jh}|, \quad (3.4)$$

where $\lambda \geq 0$ is a tuning parameter. We refer to this estimation approach as *penalized integrated loss minimization* (PILM). The LASSO minimizes $\bar{L}(\boldsymbol{\theta})$ subject to $\sum_{j=1}^q \sum_{h=1}^k |\theta_{jh}| \leq s$, where s is a user-defined bound to the model parameters. A one-to-one correspondence between s and λ can be established. In particular, if $\hat{\boldsymbol{\theta}}^{(\lambda)}$ minimizes (3.4), then it also satisfies $\sum_{j=1}^q \sum_{h=1}^k |\hat{\theta}_{jh}^{(\lambda)}| = s$. To ensure that all model parameters have the same order of magnitude, both $\mathbf{x}$ and $\mathbf{b}(p)$ are assumed to be standardized, as suggested for example in Tibshirani (1996). To minimize $\bar{L}_{PEN}^{(\lambda)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and λ , we use a path-wise coordinate descent algorithm (Friedman et al., 2010) combined with standard Newton–Raphson algorithm. The described PILM estimator is implemented in the `qrqmNP` (Sottile, 2018) package in R.

3.1 Tuning parameter selection

As for all penalized regression methods, our estimator requires defining a tuning parameter that affects the degree of shrinking to be applied to the model coefficients. Traditional selection criteria, such as AIC-type and BIC-type selectors, are grid-search

criteria that minimize

$$\text{Dev}^{(\lambda)} + c_n \cdot \text{df}^{(\lambda)}, \quad (3.5)$$

where $\text{Dev}^{(\lambda)}$ is the explained deviance of the model (a measure of goodness-of-fit defined in the following), c_n is a scaling factor that could depend on the sample size and $\text{df}^{(\lambda)}$ reflects the number of nonzero coefficients.

We follow the definitions proposed by Schwarz (1978), Lee et al. (2014) and Zheng and Peng (2017), and define

$$\text{Dev}^{(\lambda)} = \log \bar{L}_{PEN}^{(\lambda)}(\hat{\theta}),$$

the logarithm of the minimized loss function given by (3.4), in which both $\mathbf{x}$ and $\mathbf{b}(p)$ are assumed to be standardized, and where $\hat{\theta} := \hat{\theta}^{(\lambda)}$ is the estimator of θ obtained by minimizing (3.4) at a given value of λ . The AIC and BIC criteria are given by

$$AIC^{(\lambda)} = \log \bar{L}_{PEN}^{(\lambda)}(\hat{\theta}) + 2\text{df}^{(\lambda)}n^{-1}, \quad (3.6)$$

$$BIC^{(\lambda)} = \log \bar{L}_{PEN}^{(\lambda)}(\hat{\theta}) + \log(n)\text{df}^{(\lambda)}n^{-1}. \quad (3.7)$$

3.2 Inference

Inference on penalized regression estimator is not straightforward, and the majority of the existing literature refers to generalized linear models. Tibshirani (1996) proposed using ridge-type approximation of the absolute value function; Osborne et al. (2000) derived a covariance matrix ensuring that the estimated standard errors are positive for all nonzero coefficients; Beran (1982) and Kyung et al. (2010) showed that bootstrap is inconsistent. Recently, Meinshausen and Bühlmann (2012) discussed a multisplit method to assess statistical significance and construct conservative p -values for high-dimensional problems; Van de Geer et al. (2014), Zhang and Zhang (2014) and Javanmard and Montanari (2014) proposed a desparsified or debiased LASSO; Lockhart et al. (2014) discussed the covariace test for hypothesis testing problems; and Lee et al. (2016) developed a general approach to perform inference after model selection.

Other authors suggested to use LASSO regression as a screening procedure. For example, Belloni and Chernozhukov (2013) and Belloni et al. (2014) studied post-model selection estimators that apply ordinary least squares to the model selected by the first-step penalized estimator. Following the same logic, we suggest using the proposed penalized approach as an exploratory analysis aiming to reveal the important associations between the regressors and the response. Standard errors are then estimated by fitting unpenalized regression using only the selected covariates.

4 Simulations

To evaluate the finite-sample properties of the proposed estimator, we conducted a simulation study with four different simulation scenarios. For simplicity, we only considered the case $n > q$. Note, however, that the classical regularity conditions of LASSO regression (Sun et al., 2015; Bühlmann and Van De Geer, 2011; Zhao and Yu, 2006) must be satisfied even in low-dimensional models. In particular, the average value of the nonzero coefficients should be ‘large enough’, that is, $\min_{j \in \mathcal{A}} |\int_0^1 \beta_j(p | \theta^0) dp| \geq c\sigma\sqrt{2\log(q)}$, where θ^0 is the true parameter, $\mathcal{A}$ is the set of true nonzero regression coefficients, $c > 1$ and σ is a scale parameter that serves as a normalizing constant.

We considered the quantile function defined by

$$Q(p | \mathbf{x}, \boldsymbol{\theta}) = \beta_0(p | \boldsymbol{\theta}) + \beta_1(p | \boldsymbol{\theta})x_1 + \cdots + \beta_q(p | \boldsymbol{\theta})x_q,$$

in which $\mathbf{X} = (x_1, x_2, \dots, x_q) \sim \mathcal{N}_{0,1}^+(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathcal{N}_{0,1}^+$ identifies the truncated normal distribution. In simulations 1 and 1*, we defined $\boldsymbol{\Sigma} = \mathcal{I}$; in simulations 2 and 2*, we used a correlation structure based on the Toeplitz matrix, defined by $\Sigma_{ij} = 0.5^{|i-j|}$, $i, j = 1, \dots, q$.

In all simulations, we used $q = 15$ covariates including an intercept, and defined

$$\beta_0(p) = 1 + z(p), \quad \beta_1(p) = -1 + 2 \log p, \quad \beta_2(p) = 2 + 2p,$$

$$\beta_3(p) = 2 + 2\sqrt{p}, \quad \beta_4(p) = 3 + z(p), \quad \beta_5 = \cdots = \beta_{15} = 0,$$

where $z(p)$ denotes the quantile function of a standard normal distribution.

The four scenarios differ by the correlation structure of $\mathbf{X}$, and by the definition of $b(\cdot)$ used in the modelling phase. In details, in simulations 1 and 2, we defined

$$\beta_0(p | \boldsymbol{\theta}) = \theta_{00} + \theta_{01}z(p); \tag{4.1}$$

$$\beta_j(p | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1} \log p, \quad j = 1, 11, 12, 13; \tag{4.2}$$

$$\beta_j(p | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}p, \quad j = 2, 8, 9, 10; \tag{4.3}$$

$$\beta_j(p | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}\sqrt{p}, \quad j = 3, 14, 15; \tag{4.4}$$

$$\beta_j(p | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}z(p), \quad j = 4, 5, 6, 7. \tag{4.5}$$

In simulations 1* and 2*, we used

$$\beta_0(p | \boldsymbol{\theta}) = \theta_{00} + \theta_{01}z(p); \tag{4.6}$$

$$\beta_j(p | \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}b_1(p) + \theta_{j2}b_2(p) + \theta_{j3}b_3(p), \quad j = 1, \dots, 15, \tag{4.7}$$

where $\{b_1(p), b_2(p), b_3(p)\}$ form the basis of a 3rd-degree shifted Legendre polynomial (e.g., Abramowitz and Stegun, 1964), an orthogonal polynomial in $(0, 1)$. In simulations 1 and 2, the fitted model included the true model as a special case. Instead, in simulations 1* and 2*, the specification of $b(p)$ did not correspond to

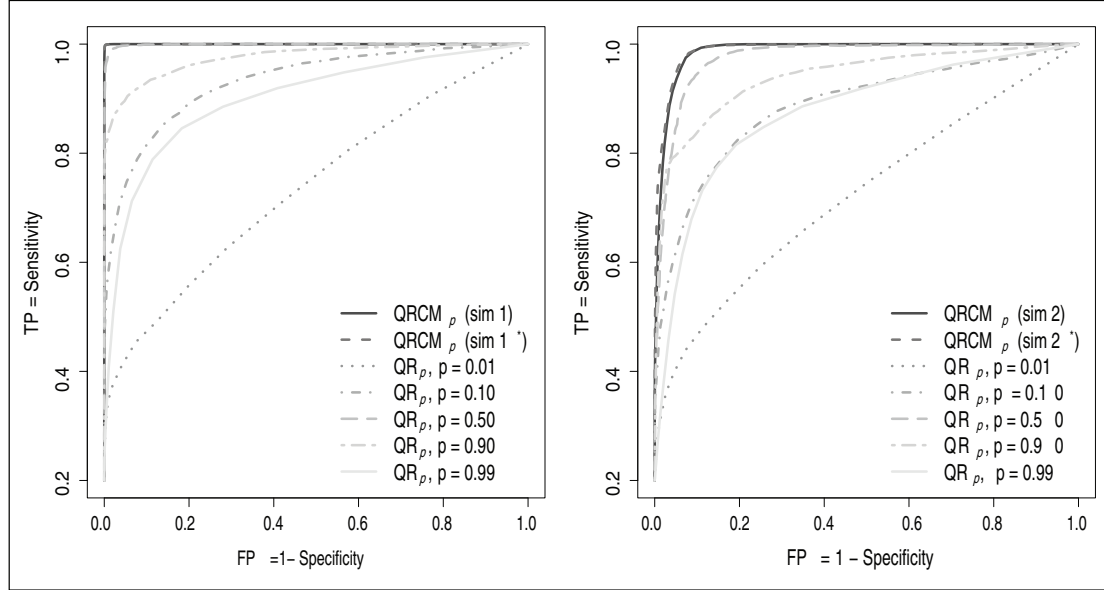


Figure 2 ROC curves based on simulations 1 and 1* (left panel), and 2 and 2* (right panel)

Note: In simulations 1 and 2, we used the true definition of $b(p)$, while in simulations 1* and 2* we used a shifted Legendre polynomial.

the true model. This allowed assessing the performance of the described estimator when the true data-generating process is not known and a flexible parametrization is used to approximate $\beta(p \mid \theta)$.

For each scenario, we generated $B = 1000$ simulated datasets of $n = 500$ observations. We applied the described QRCM_p method, using an evenly spaced sequence of 100 $\log \lambda$ -values, and selected the ‘best’ model according to AIC and BIC separately. We compared our estimator with a standard QR_p model. To have a term of comparison, we also fitted the true model using unpenalized QRCM .

We defined ‘true positive’ (TP), the situation in which an important covariate is correctly selected for inclusion, and ‘true negative’ (TN), the case in which an irrelevant covariate is correctly removed from the model. False positives (FP) and false negatives (FN) are defined accordingly. For varying λ , these quantities can be used to construct ROC curves. Figure 2 shows the ROC curves generated by simulations 1 and 1*, on the left, and those obtained from simulations 2 and 2*, on the right. In the figures, we also report the ROC curves that are obtained by applying standard QR_p at quantiles $p = (0.01, 0.10, 0.50, 0.90, 0.99)$. Results show that the proposed QRCM_p method performs well, irrespective of whether $b(p)$ corresponds to the true basis or is defined by a flexible model. Moreover, the described method outperforms standard QR_p , in which model selection is carried out in a quantile-by-quantile fashion.

Figure 3 summarizes the distribution of the number of covariates selected by AIC and BIC, respectively. Results suggest that BIC can be used to select consistently

Table 2 Simulation results

	$p = 0.01$	$p = 0.10$	$p = 0.50$	$p = 0.90$	$p = 0.99$
QRCM	0.031(.006)	0.077(.010)	0.065(.005)	0.014(.002)	0.005(.002)
QRCM _p -AIC (sim 1)	0.027(.006)	0.069(.010)	0.059(.005)	0.012(.002)	0.003(.002)
QRCM _p -BIC (sim 1)	0.024(.005)	0.064(.010)	0.056(.005)	0.010(.002)	0.002(.001)
QRCM _p -AIC (sim 1*)	0.038(.010)	0.066(.011)	0.060(.006)	0.012(.002)	0.003(.002)
QRCM _p -BIC (sim 1*)	0.032(.009)	0.060(.010)	0.057(.006)	0.011(.002)	0.002(.001)
QR _p -AIC	0.047(.012)	0.080(.013)	0.063(.006)	0.014(.003)	0.008(.003)
QR _p -BIC	0.044(.012)	0.074(.013)	0.060(.006)	0.012(.002)	0.007(.003)
QRCM	0.016(.004)	0.045(.007)	0.047(.005)	0.009(.001)	0.002(.001)
QRCM _p -AIC (sim 2)	0.014(.004)	0.041(.007)	0.043(.005)	0.008(.001)	0.001(.001)
QRCM _p -BIC (sim 2)	0.013(.003)	0.038(.007)	0.042(.005)	0.007(.001)	0.001(.000)
QRCM _p -AIC (sim 2*)	0.022(.006)	0.038(.007)	0.045(.006)	0.008(.001)	0.002(.001)
QRCM _p -BIC (sim 2*)	0.017(.006)	0.034(.007)	0.044(.005)	0.007(.001)	0.001(.001)
QR _p -AIC	0.029(.008)	0.048(.010)	0.047(.006)	0.009(.002)	0.004(.002)
QR _p -BIC	0.026(.009)	0.042(.009)	0.044(.006)	0.008(.001)	0.003(.002)

Note: For each scenario, we report the measure of MSE described in (4.8), with standard errors in brackets. In the table, QRCM denotes the unpenalized estimator in which the model is correctly specified and only the nonzero parameters are computed; QRCM_p denotes the penalized estimator in which the best tuning parameter is selected using the criteria proposed in Section 3.1; and QR_p denotes standard penalized quantile regression in which the best tuning parameter is selected by AIC and BIC, as proposed by Lee et al. (2014).

the true set of important covariates, while AIC tends to select overparametrized models.

To assign a numerical value to the performance of each estimator, we defined the following aggregated measure of mean squared error:

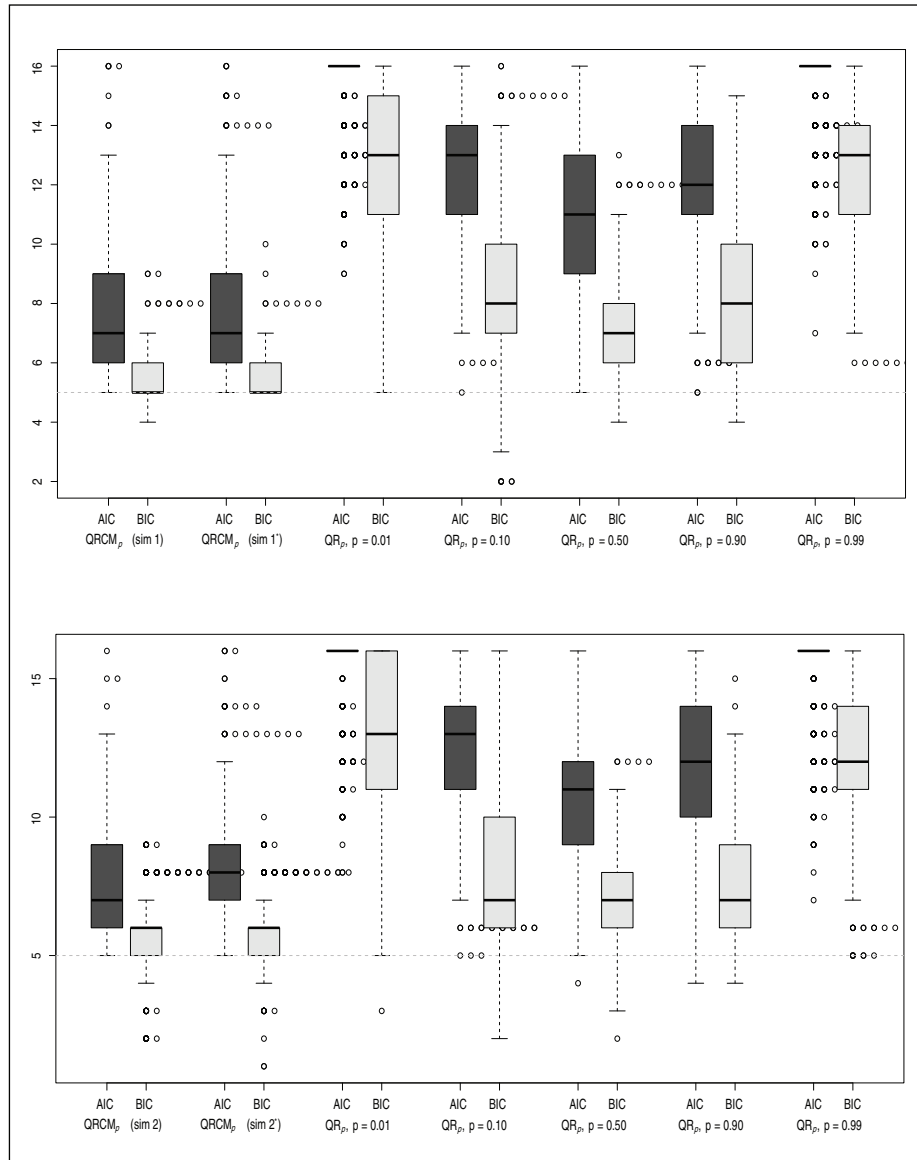
$$\text{MSE}_F = \mathbb{E}_b \left\{ \mathbb{E}_i \left[\left(F(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(p) \mid \mathbf{x}_i) - p \right)^2 \right] \right\}, \quad (4.8)$$

where $F(\cdot \mid \mathbf{x}_i)$ is the true value of the conditional distribution function and $\hat{\boldsymbol{\beta}}(p)$ indicates the estimated quantile regression coefficients, which can be computed either using standard QR_p, or using QRCM_p; in the latter case, $\hat{\boldsymbol{\beta}}(p) = \hat{\boldsymbol{\theta}}\mathbf{b}(p)$. Note that, by definition, $F(\mathbf{x}_i^\top \boldsymbol{\beta}(p) \mid \mathbf{x}_i) = p$. The operators $\mathbb{E}_i$ and $\mathbb{E}_b$ are used to indicate averaging across observations and across datasets, respectively.

We evaluated MSE_F at a grid of quantiles defined by $p^* = (0.01, 0.10, 0.50, 0.90, 0.99)$. Results are summarized in Table 2. Also in this case, the QRCM_p estimators outperformed standard QR_p, irrespective of whether the true basis or a flexible polynomial model was used.

5 Computation

The described *PILM* estimator is implemented in the `qrcmNP` package in R. Although computation is relatively easy, there are important issues. The procedure is described



AQ: 2 Figure 3

Note: Each boxplot in the figure represents the distribution of the number of covariates selected by AIC and BIC. The top panel refers to simulations 1 and 1*; the bottom panel refers to simulations 2 and 2*. In simulations 1 and 2, the true specification of $b(p)$ was used to fit the model. In simulation 1* and 2*, a shifted Legendre polynomial was used. In each figure, we also illustrate the performance of QR_p at selected quantiles. Dotted grey lines indicate the true number of nonzero regression coefficients.

in Algorithm 1. The objective function includes an integral that, in principle, can be evaluated numerically. Under model (3.1), however, $\bar{L}(\boldsymbol{\theta})$ and its gradient admit a closed-form expression, which is described in details in Frumento and Bottai (2016).

Algorithm 1 Pseudocode for our algorithm

Input: The model matrix $\mathbf{x}$, the response vector $\mathbf{y}$, a symbolic description of $\mathbf{b}(p)$

Initialize

Evaluate numerically $\mathbf{b}(p)$

Standardize $\mathbf{x}$ and $\mathbf{b}(p)$ to have zero mean and unit norm

Select the initial values $\hat{\boldsymbol{\theta}}^{(0)}$

Compute the unpenalized gradient vector $S(\hat{\boldsymbol{\theta}}^{(0)}) = \nabla_{\boldsymbol{\theta}} \bar{L}(\hat{\boldsymbol{\theta}}^{(0)})$

Define $\lambda_{\max} = \max |S(\hat{\boldsymbol{\theta}}^{(0)})|$ and $\lambda_{\min} = \epsilon \lambda_{\max}$, with $0 < \epsilon < 1$

Compute a sequence of K values of λ between $\lambda_{\min}$ and $\lambda_{\max}$, evenly spaced on the log scale

For $t = 1$ to K **do:**

Determine those covariates $\mathbf{x}_j$ such that $|S(\hat{\boldsymbol{\theta}}^{(t-1)})_j| \geq \lambda_t$

Solve $\nabla_{\boldsymbol{\theta}} \bar{L}_{PEN}^{(\lambda_t)}(\boldsymbol{\theta}) = \mathbf{0}$ using Newton–Raphson

Compute the unpenalized gradient vector $S(\hat{\boldsymbol{\theta}}^{(t)}) = \nabla_{\boldsymbol{\theta}} \bar{L}(\hat{\boldsymbol{\theta}}^{(t)})$

End for

Output: The sequence of penalized parameter vectors $\hat{\boldsymbol{\theta}}^{\lambda}$

5.1 qrcmNP package

The `qrcmNP` package contains a main function `piqr` that implements model fitting using path-wise coordinate descent and Newton–Raphson algorithm to minimize $L_{PEN}^{(\lambda)}$ at the selected values of the tuning parameter λ , as described in Algorithm 1. The function is defined as

```
piqr(formula, formula.p = ~ slp(p, 3), weights, data, s,
     nlambdas = 100, lambda.min.ratio = ifelse(nobs < nvars, 0.01,
     1e-04), lambda, tol = 1e-06, maxit = 100, display = TRUE)),
```

where the main arguments are: `formula`, a symbolic description of the quantile regression model; `formula.p`, that describes how quantile regression coefficients depend on p , the order of the quantile; `s`, an optional 0/1 matrix that permits excluding some model coefficients; `nlambdas` and `lambda.min.ratio`, that correspond to K and ϵ in Algorithm 1; and `lambda`, a user-defined sequence of values of λ .

Table 3 Model selection based on different criteria

Criterion	λ^*	n. of parameters	n. of covariates	Loss	P-value KS
AIC	19.25	31/39	8/10	293.31	0.77
BIC	53.56	19/39	5/10	294.01	0.53

Note: Optimal λ values, number of parameters, number of selected covariates, minimized loss function, and p -value of a Kolmogorov-Smirnov (KS) goodness-of-fit test of the model selected by AIC and BIC. The maximal model had 39 parameters. AIC and BIC selected 31 and 19 parameters, respectively, corresponding to 8 and 5 nonzero regression coefficients. The model selected by BIC is summarized in Table 4 and represented graphically in Figure 5.

Summary measures, variable selection procedures, predictions and graphical tools are provided in a variety of auxiliary functions, which include `summary.piqr()`, `gof.piqr()`, `predict.piqr()` and `plot.piqr()`.

6 Variable selection for inspiratory capacity

We applied the proposed QRCM_p method to a subset ($n = 2\,201$) of the data analysed in Bottai et al. (2011). The data refer to a study carried out in 1988–1991 in Northern Italy, and include 1 063 males and 1 138 females. The study aimed to estimate percentiles of IC, a measure of lungs function. The following nine predictors were available: age, height, BMI, sex and indicators for current smoking, occupational exposure, cough, wheezing and asthma.

We used $b(p) = [1, \log(p), \log(1 - p)]^T$ to model the intercept, 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. Several covariates were discarded. Table 3 and Figure 4 illustrate the results of our analysis and report additional diagnostic tools, including the p -value of the Kolmogorov–Smirnov goodness-of-fit procedure proposed by Frumento and Bottai (2016), in which the fitted values of the cumulative distribution function are compared with a $U(0, 1)$ distribution.

We used the model selected by BIC and estimated it again using unpenalized QRCM , as described in Section 3.2. The model is summarized in Table 4, and represented graphically in Figure 5, while selected quantiles are reported in Table 5. Because we were mostly interested in the low quantiles of IC, we only report the estimated quantile regression coefficients, $\hat{\beta}(p) = \beta(p | \hat{\theta})$, at $p = 0.01$, $p = 0.05$ and $p = 0.50$. Age, height, BMI and sex were statistically significant. Figure 5 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 litres respectively, while its median decreased by about 0.01 litres. Height and BMI both had a positive effect. Quantile regression coefficients were increasing with a non-linear

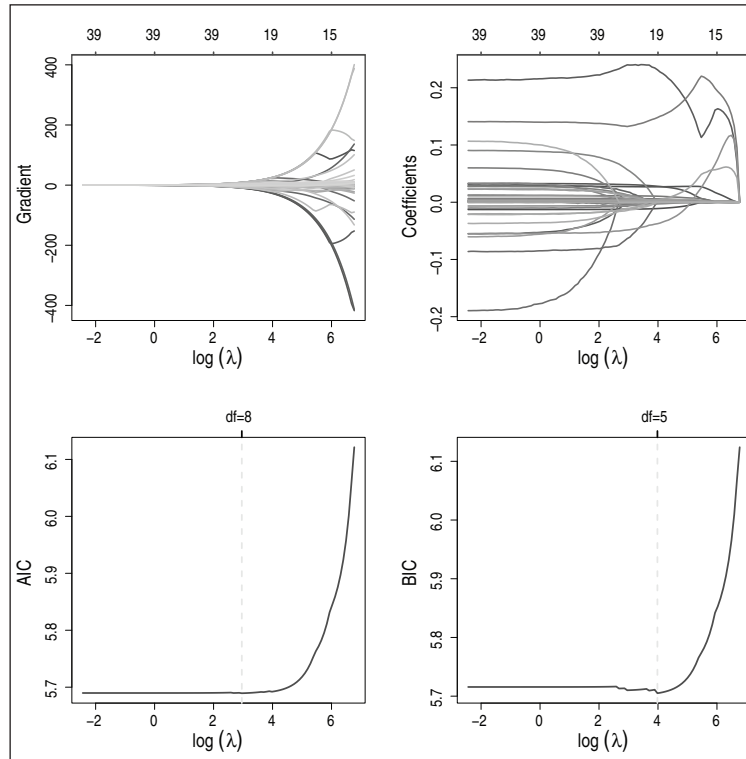


Figure 4 Gradient plot and coefficient profile plot versus $\log(\lambda)$ (top panels); AIC and BIC curves versus $\log(\lambda)$ (bottom panels), for the inspiratory capacity data

trend. For each one-centimeter increase in height, quantiles below the median increased by approximately 0.03 litres. 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.

7 Discussion

We described a penalized approach that can be applied to the *QRCM* framework introduced by Frumento and Bottai (2016). Modelling 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 the conditional quantile function. The proposed method demonstrated to be an efficient tool to recover the structure of the true model with a high probability. A computationally efficient algorithm is implemented in the *qrcmNP* package in R.

Table 4 Summary of the model selected by BIC. Unpenalized QRCM estimates of θ and asymptotic standard errors (in brackets) based on the model selected by BIC

	1	$\log(p)$	$\log(1 - p)$	$\text{SLP}(p, 3)[1]$	$\text{SLP}(p, 3)[2]$	$\text{SLP}(p, 3)[3]$	P -value
Intercept	2.515(.002)*	0.212(.015)*	-0.288(.017)	-	-	-	.000*
Age	-0.014(.002)*	-	-	0.004(.001)*	0.000(.001)	0.001(.001)	.000*
Height	0.026(.003)*	-	-	0.007(.002)*	0.005(.002)*	0.003(.001)*	.000*
BMI	0.032(.006)*	-	-	0.013(.003)*	-0.005(.003)	0.003(.003)	.000*
Male	0.269(.063)*	-	-	0.128(.038)*	-0.050(.032)	-0.010(.022)	.000*
Smoker	-	-	-	-	-	-	-
Exposure	-	-	-	-	-	-	-
Cough	-	-	-	-	-	-	-
Wheezing	-	-	-	-	-	-	-
Asthma	-	-	-	-	-	-	-
P -value	.000*	.000*	.000*		.000*		

Note: In the table, $\text{SLP}(p, 3)$ denotes a 3rd-degree shifted Legendre polynomial. The corresponding graphical representation is proposed in Figure 5. The bottom row contains the p -values for the null hypothesis that the corresponding column of θ is 0 and represents the significance of the components of $b(p)$. The last column is defined analogously and can be interpreted as the significance of a test for a null effect of covariates. The asterisk (*) denotes significance less than 0.05

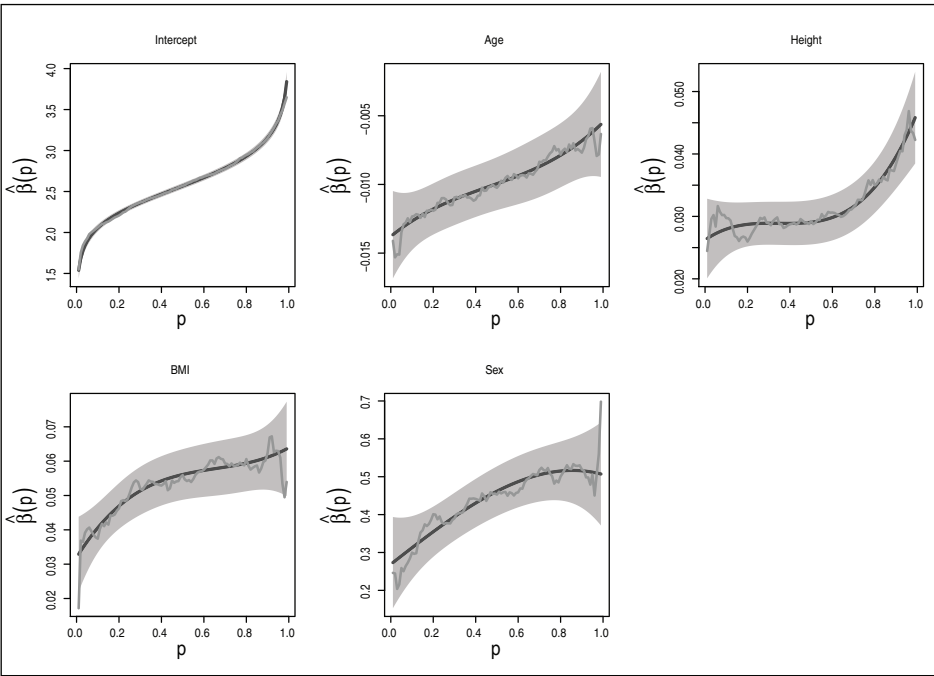


Figure 5 Unpenalized QRCM estimates of $\beta(p)$ under the model selected by BIC (see Table 4)

Note: 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.

Table 5 Estimated quantile regression coefficients at $p = 0.01$, $p = 0.05$ and $p = 0.50$, obtained from the model selected by BIC

	$p = 0.01$	$p = 0.05$	$p = 0.50$
Intercept	1.539(.049)*	1.893(.029)*	2.567(.018)*
Age	−0.014(.002)*	−0.013(.001)*	−0.010(.001)*
Height	0.026(.003)*	0.027(.003)*	0.029(.002)*
BMI	0.033(.005)*	0.037(.004)*	0.056(.004)*
Male	0.273(.061)*	0.291(.051)*	0.462(.033)*
Smoker	-	-	-
Non-exposure	-	-	-
Cough	-	-	-
Wheezing	-	-	-
Asthma	-	-	-

Note: Estimated standard errors in brackets. The asterisk (*) denotes significance less than 0.05.

Supplementary materials

Supplementary materials for this article containing R scripts to reproduce simulations and application, and an artificial dataset with similar characteristics as the real data are available from <http://www.statmod.org/smij/archive.html>. The `qrcmNP` (Sottile, 2018) R package is available on CRAN.

Acknowledgements

We are very grateful to Dr Giovanni Viegi for sharing the lung function data used in the present article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

References

- AQ: 3 Abramowitz M and Stegun I (1964) *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Vol. 55. Chelmsford, MA: Courier Corporation.
- Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–23.
- Belloni A and Chernozhukov V (2013) AQ: 4 Least squares after model selection in

- high-dimensional sparse models. *Bernoulli*, **19**, 521–47.
- (2011) L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, **39**, 82–130.
- Belloni A, Chernozhukov V and Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81**, 608–50.
- AQ: 5 Beran R (1982) Estimated sampling distributions: The bootstrap and competitors. *The Annals of Statistics*, **10**, 212–25.
- Bottai M, Pistelli F, Di Pede F, Baldacci S, Simoni M, Maio S, Carrozzi L and Viegi G (2011) Percentiles of inspiratory capacity in healthy nonsmokers: A pilot study. *Respiration*, **82**, 254–62.
- Bühlmann P and Van De Geer S (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin: Springer Science & Business Media.
- Chaudhuri P (1991) Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, **39**, 246–69.
- Craven P and Wahba G (1979) Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.
- Efron B, Hastie T, Johnstone I and Tibshirani R (2004) Least angle regression. *Annals of Statistics*, **32**, 407–99.
- Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–60.
- Friedman J, Hastie T and Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.
- Frumento P (2017) qrcm: Quantile regression coefficients modeling. *R package version 2.1*. URL <https://CRAN.R-project.org/package=qrcm> (last accessed 11 January 2019).
- Frumento P and Bottai M (2016) Parametric modeling of quantile regression coefficient functions. *Biometrics*, **72**, 74–84.
- (2017) Parametric modeling of quantile regression coefficient functions with censored and truncated data. *Biometrics*, **73**, 1179–88.
- Javanmard A and Montanari M (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, **15**, 2869–2909.
- Koenker R and Bassett J (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
- Kyung M, Gilly J, Ghoshz M and Casella G (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5**, 1–44.
- Lee E, Noh H and Park B (2014) Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, **109**, 216–29.
- Lee J, Sun D, Sun Y and Taylor JE (2016) Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, **44**, 907–27. AQ: 6
- Li Y and Zhu J (2008) L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, **17**, 163–85.
- Lockhart R, Taylor J, Tibshirani R and Tibshirani R (2014) A significance test for the lasso. *The Annals of Statistics*, **42**, 413–68.
- Meinshausen N and Bühlmann P (2012) P-values for high-dimensional regression. *Journal of the American Statistical Association*, **104**, 1671–81.
- Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758–65.
- Osborne M, Presnell B and Turlach B (2000) On the lasso and its dual. *Journal of Computational Graphical Statistics*, **9**, 319–37.
- Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–64.
- Sherwood B and Maidman A (2017) rqPen: Penalized quantile regression. *R package version 2.0*. URL <https://CRAN.R-project.org/package=rqPen> (last accessed 11 January 2019).
- Sottile G (2018) qrcmNP: Nonlinear and penalized parametric modeling of quantile regression coefficient functions. *R package version 0.1.2*. URL <https://CRAN.R-project.org/package=qrcmNP> (last accessed 11 January 2019).

- Sun W, Reich B, Tony C, Guindani M and Schwartzman A (2015) False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B*, **77**, 59–83.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–88.
- Van de Geer S, Bühlmann S, Ritov Y and Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.
- Wang L, Wu Y and Li R (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, **107**, 214–22.
- Wu Y and Liu Y (2009) Variable selection in quantile regression. *Statistica Sinica*, **19**, 801–17.
- Yang C, Chen Y and Chang H (2017) Composite marginal quantile regression analysis for longitudinal adolescent body mass index data. *Statistics in Medicine*, **36**, 3380–97.
- Zhang C and Zhang S (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, **76**, 217–42.
- Zhang Y, Li R and Tsai C (2010) Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, **105**, 312–23.
- Zhao P and Yu B (2006, November) On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–63.
- Zheng Q and Peng L (2017) Consistent model identification of varying coefficient quantile regression with BIC tuning parameter selection. *Communications in Statistics: Theory and Methods*, **46**, 1031–49.
- Zheng Q, Gallagher C and Kulasekera K (2013) Adaptive penalized quantile regression for high dimensional data. *Journal of Statistical Planning and Inference*, **143**, 1029–38.