Comparative study of flexible quantile regression techniques: constructing reference curves in Pediatrics

Martínez-Silva, I.$^{(1)}$**; Roca-Pardiñas, J.**$^{(2)}$**; Cadarso-Suárez C.**$^{(1)}$**;**
Leis Trabazo, R.$^{(3)}$**; Tojo-Sierra, R.**$^{(3)}$

(1) Biostatistics Unit-Department of Statistics and Operative Research-University of Santiago de Compostela
(2) Statistics Department-University of Vigo
(3) Pediatrics Departament-Santiago de Compostela University Hospital Complex

`isabelmaria.martinez@usc.es, roca@uvigo.es, carmen.cadarso@usc.es`

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIVERSIDADE DE VIGO

Introduction

In general, quantile regression (QR) is a regression model to asses the relationships between covariate vector, x, and quantile curves of a response variable, y, for a given value x.

Fixed $\tau \in (0,1)$, the additive quantile regression model takes the form:

$$y = \beta_{0,\tau} + \sum_{j=1}^{p} h_{\tau j}(x_j) + \varepsilon_\tau, \quad \varepsilon_\tau \sim F_\tau \texttt{ where } F_\tau(0|x) = \tau \texttt{ and } (x_1, ..., x_p) = \boldsymbol{x}$$

and the resulting quantile function has a nonlinear predictor structure, given by:

$$Q_y(\tau|x) = \beta_{0,\tau} + \sum_{j=1}^{p} h_{\tau j}(x_j)$$

where $Q_y(\tau|x) = inf\{y : F_y(y|x) \geq \tau\}$

Accordingly, one of the main goals of this work has been to perform a simulation study to compare statistically different additive quantile regression approaches. Also, all the reviewed techniques (implemented in RDevelopmentCoreTeam2010) were used to construct the overall sex- and height-specific reference curves of anthropometric measures in real data.

Quantile Regression Techniques

In the literature there are different statistical methodologies propossing flexible quantile regression models. In this work, the following techniques were reviewed:

- Koenker and Basset technique (Koenker and Bassett, 1978; Koenker, Ng and Portnoy, 1994):
  - the $\tau-$th conditional centile, takes the form:

$$Q_y(\tau|\boldsymbol{x}) = \sum_{j=1}^{p} h_{\tau j}(x_j)$$

  - cubic regression splines are used (de Boor, 2001)
  - implemented in the `quantreg` package
- Lambda Mean Standard deviation (LMS) method (Cole, 1988)
  - the smooth curve for the $\tau-$th centile is giving by:

$$Q_y(\tau|\boldsymbol{x}) = M(\boldsymbol{x})[1 + L(\boldsymbol{x})S(\boldsymbol{x})z_\tau]^{1/L(\boldsymbol{x})}$$

  - based on the power transformation family Box-Cox (Box and Cox, 1964)
  - the model is represented as a Vector Generalized Additive Model (Yee and Wild, 1996)
  - smoothing splines (Hastie and Tibshirani, 1990) are used
  - implemented in the `VGAM` package
- Generalized Additive Models for Location, Scale and Shape (GAMLSS) methodology (Rigby and Stasinopoulos, 2005)
  - given a normal response variable, y, the $\tau-$th centile curve is expressed as follows:

$$Q_y(\tau|\boldsymbol{x}) = \sum_{j=1}^{p} f_{\tau j}(x_j) + exp\left(\sum_{j=1}^{p} g_{\tau j}(x_j)\right) z_\tau = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})z_\tau$$

  - B-splines regression (de Boor, 2001) are used
  - implemented in the `gamlss` package
- Boosting algorithms (Fenske, Kneib and Hothorn, 2009)
  - `base learner` are used to estimate the smooth functions $h_{\tau j}$. The $\tau-$th following centile curve is obtained:

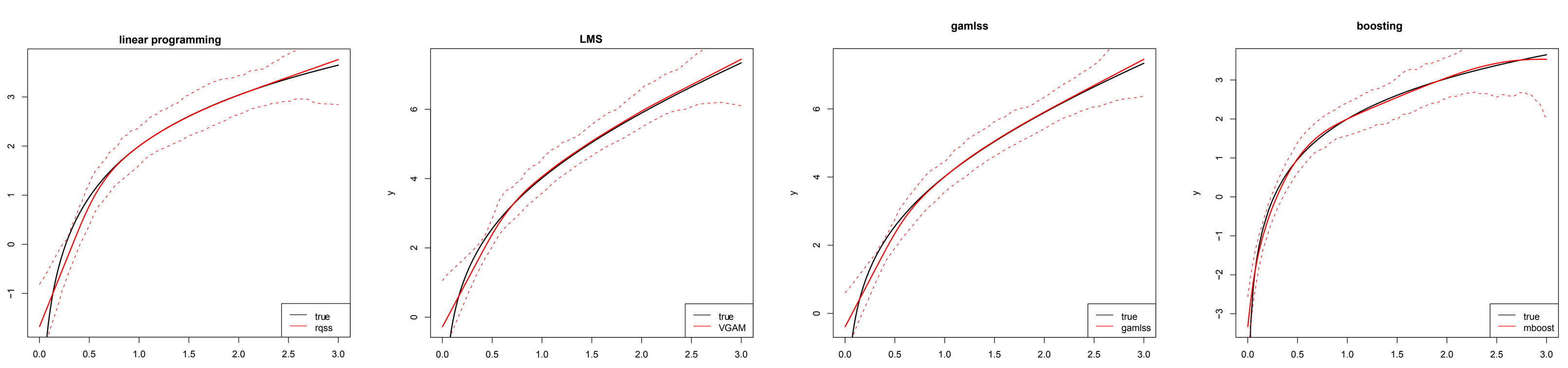
$$Q_y(\tau|\boldsymbol{x}) = \sum_{j=1}^{p} h_{\tau j}(x_j)$$

  - P-splines regression (Schmid and Hothorn, 2008) is used
  - implemented in the `mboost` package

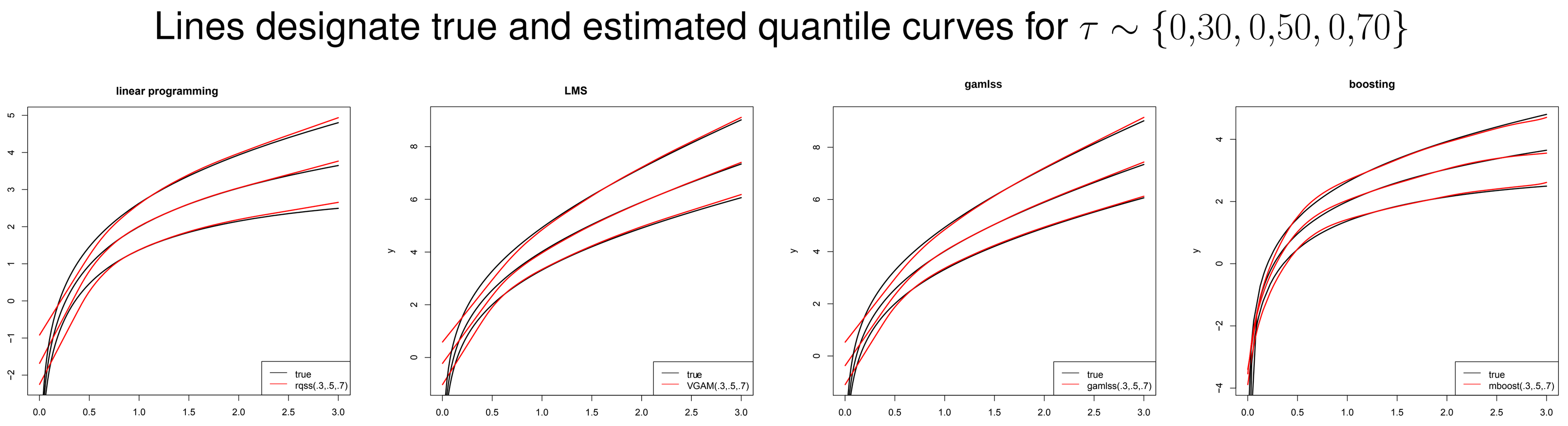
Simulation Study

We considered an additive model with non linear terms for location and scale, as follows:
Equation: $y = 2 + 1.5\log(X) + (0.7 + 0.5X) \cdot \varepsilon$ Error $\varepsilon \sim N(0,1)$ Sample of $n = 200$ observations

$95\,\%$ Confidence Bands for the median



Lines designate true and estimated quantile curves for $\tau \sim \{0{,}30, 0{,}50, 0{,}70\}$



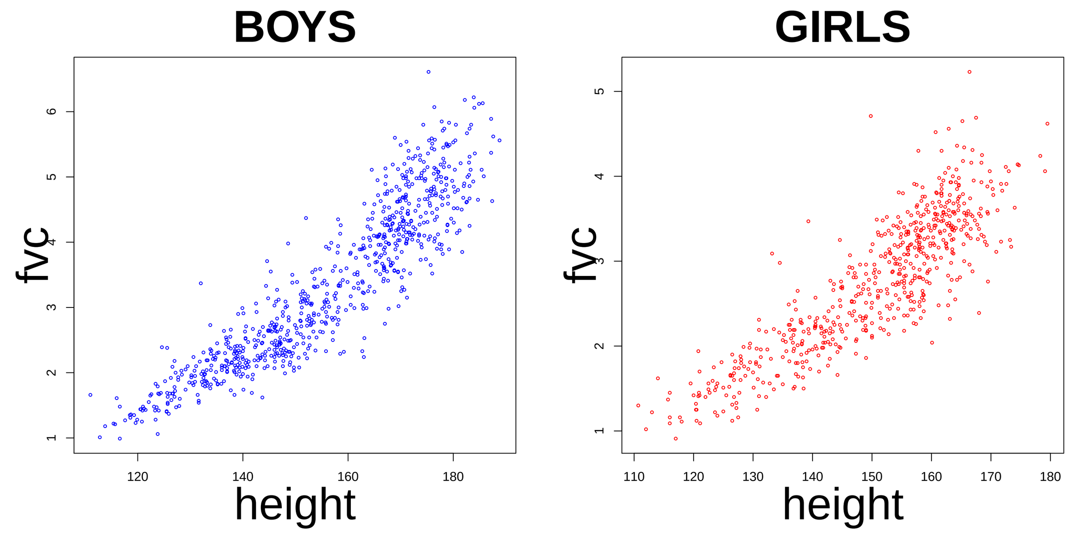
Application in Pediatrics

Sample Description

Key features:

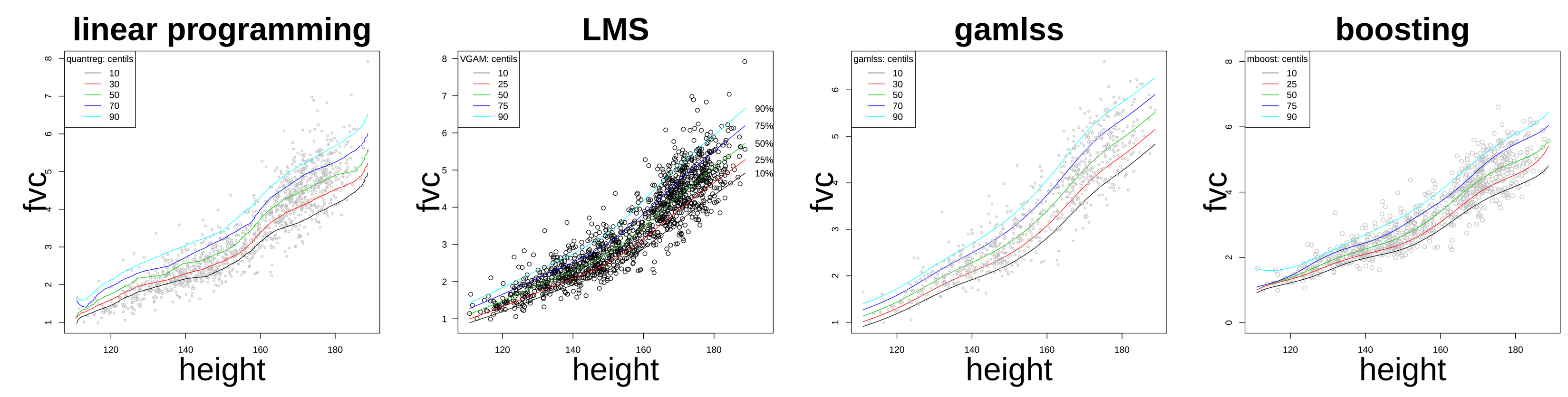
- 2395 healthy school-aged individuals
- composed by 1201 boys and 1194 girls
- ages between 6 and 18 years

Variable studied: forced vital capacity (fvc) depending on height (cm.) and sex

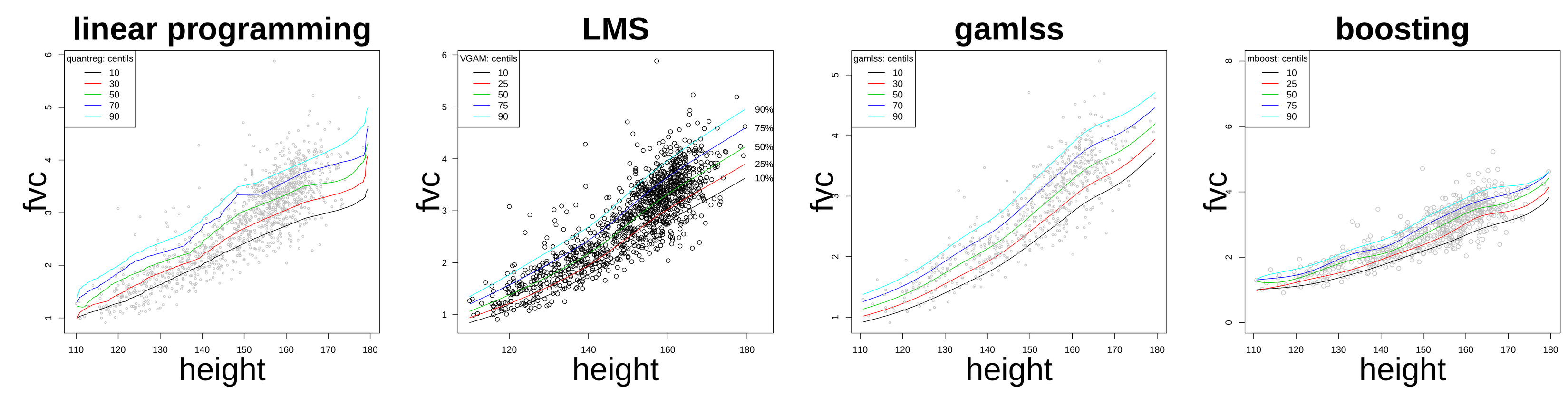


Representations of the estimated quantile curves for τ by sex

Representations for $\tau \sim \{0{,}10, 0{,}25, 0{,}50, 0{,}75, 0{,}90\}$:



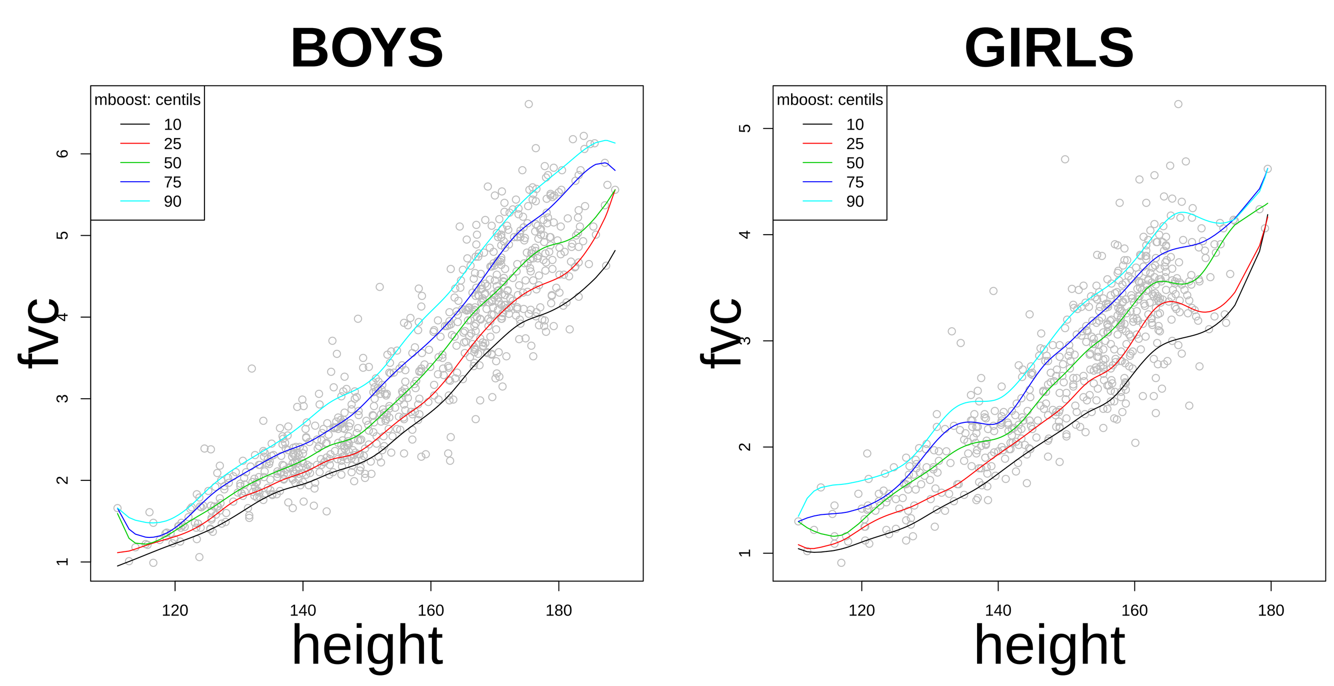
Relationship between fvc and height for boys.



Relationship between fvc and height for girls.

Discussion

- While automatic criteria for selecting smoothing parameter are provided in the boosting technique, there is no automatic selection neither for linear programming, LMS technique nor gamlss methodology.
- In comparison to linear programming, boosting a) can handle a larger number of non linear covariate effects; b) parameter estimation and variable selection are executed in one single estimation step.
- LMS methodology, presented as a Vector Generalized Additive Models, needs positive response variable.
- Real data:
  - is necessary a flexible model to describe the relationship between fvc and height
  - the quantile curves are different by sex
- The quantile curves are independently estimated by linear programming and boosting. That produces crossing quantile curves problems, as shown in these figures:



Acknowledgements

This work was partially supportted by grants MTM2008-01603 (Spanish Ministry of Science and Innovation), INCITE08PXIB208113PR and INCITE09PXIB228166PR (Xunta de Galicia)

References

1. Cole T. J. (1988). Using the LMS method to measure skewness in the NCHS and dutch national height standards. *Ann. Hum. Biol.*, 16:407-419.
2. de Boor C. A. (2001). *A practical guide to splines (Rev. Edn)(Rev. Edn)*. New York: Springer
3. Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89-121.
4. Fenske N., Kneib T. and Hothorn T. (2009). Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Departament of statistics. University of Munich. Technical Report Number 052, 2009*.
5. González Barcala F.J., Cadarso Suárez C., Valdés Cuadrado L., Leis R., Cabanas R. and Tojo R. (2008). Valores de referencia de función respiratoria en niños y adolescentes (6-18 años) de Galicia. *Arch. Bronconeumol.*, 44(6):295-302
6. Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.
7. Koenker R. and Bassett G. (1978). Regression quantiles. *Econometrica*, 46:33-50.
8. Koenker R., Ng P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*,81:673-680.
9. Rigby R. and Stasinopoulos D. M. (2005). Generalized additive models for location, scale and shape. *Appl. Statist.*, 54:507-554.
10. Yee T. W. and Wild C. J. (1996). Vector Generalized Additive Models. *Journal of Royal Statistical Society, Series B*, 58(3):481-493.