Open access • Journal Article • DOI:10.1198/JASA.2011.AP09272

Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression — Source link

Nora Fenske, Thomas Kneib, Torsten Hothorn

Published on: 01 Jan 2011 - Journal of the American Statistical Association (Taylor & Francis)

Topics: Quantile regression, Binomial regression, Regression analysis, Quantile and Linear regression

Nora Fenske, Thomas Kneib & Torsten Hothorn

Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression

Technical Report Number 052, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression

Nora Fenske
Ludwig-Maximilians-Universität
München

Thomas Kneib
Carl-von-Ossietzky-Universität
Oldenburg

Torsten Hothorn
Ludwig-Maximilians-Universität München

Abstract

Ordinary linear and generalized linear regression models relate the mean of a response variable to a linear combination of covariate effects and, as a consequence, focus on average properties of the response. Analyzing childhood malnutrition in developing or transition countries based on such a regression model implies that the estimated effects describe the average nutritional status. However, it is of even larger interest to analyze quantiles of the response distribution such as the 5% or 10% quantile that relate to the risk of children for extreme malnutrition. In this paper, we analyze data on childhood malnutrition collected in the 2005/2006 India Demographic and Health Survey based on a semiparametric extension of quantile regression models where nonlinear effects are included in the model equation, leading to additive quantile regression. The variable selection and model choice problems associated with estimating an additive quantile regression model are addressed by a novel boosting approach. Based on this rather general class of statistical learning procedures for empirical risk minimization, we develop, evaluate and apply a boosting algorithm for quantile regression. Our proposal allows for data-driven determination of the amount of smoothness required for the nonlinear effects and combines model selection with an automatic variable selection property. The results of our empirical evaluation suggest that boosting is an appropriate tool for estimation in linear and additive quantile regression models and helps to identify yet unknown risk factors for childhood malnutrition.

Keywords: functional gradient boosting, penalized splines, additive models, variable selection, model choice.

1. Introduction

The reduction of malnutrition and in particular childhood malnutrition is among the United Nations Millennium Development Goals, aiming at halving the proportion of people suffering from hunger until 2015. Therefore, a better understanding of risk factors for malnutrition is of utmost importance. Malnutrition can be measured in terms of a score that compares the nutritional status of children in the population of interest with the nutritional status in a reference population. While previous analyses of childhood malnutrition, see for example Kandala, Lang, Klasen, and Fahrmeir (2001) or Kandala, Fahrmeir, Klasen, and Priebe (2008), have focused on mean regression for this score, statistical analyses of the lower quantiles are, to the best of our knowledge, still missing. Such analyses can add important additional information since they allow to identify risk factors for severe malnutrition, expressed by the 5% or 10% quantiles of the score, in contrast to mean regression that describes the expected nutritional status.

In this paper, we focus on analyzing risk factors for childhood malnutrition, using data collected in the 2005/2006 Demographic and Health Survey (DHS, `www.measuredhs.com`) for India. We apply quantile modeling for estimating the influence of potential risk factors on the lower quantiles of the conditional distribution of a malnutrition score. Previous analyses of the mean nutritional status (Kandala *et al.* 2001, 2008) revealed nonlinear effects of important risk factors, such as child's or mother's age or body mass index of the mother. Consequently, appropriate modeling has to take such flexible smooth effects into account. Additive quantile regression models allow for semiparametric predictors including linear and nonlinear effects. Our analysis is based on 21 potentially important risk factors but aims at deriving a possibly sparse and interpretable model. Therefore, we are faced with a variable selection and model choice problem and our final model should only consist of relevant risk factors modeled at appropriate complexity.

State-of-the-art procedures for additive quantile regression, as introduced later on, lack adequate variable and model selection properties. We therefore develop an alternative estimation procedure for additive quantile regression in the empirical risk minimization framework based on a boosting approach. This technique has successfully been used to address variable and model selection in other regression contexts, e.g., in Friedman, Hastie, and Tibshirani (2000); Bühlmann and Yu (2003), or Bühlmann and Hothorn (2007).

A completely distribution free approach that directly addresses quantile modeling is given by quantile regression, which is thoroughly treated in Koenker (2005). The simple linear quantile regression model can be written as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \varepsilon_{\tau i} \qquad \varepsilon_{\tau i} \sim H_{\tau i} \quad \text{subject to} \quad H_{\tau i}(0) = \tau \ , \tag{1}$$

see Buchinsky (1998). Here, the index $i = 1, \ldots, n$, denotes the individual, y_i and $\boldsymbol{x}_i$ stand for response variable and covariate vector (including an intercept) for individual i, respectively. The quantile specific linear effects are given by $\boldsymbol{\beta}_\tau$ and $\tau \in (0, 1)$ indicates a fixed and known quantile. The random variable $\varepsilon_{\tau i}$ is assumed to be an unknown error term with cumulative distribution function $H_{\tau i}$, on which no specific distributional assumptions are made apart from the restriction in (1), which implies that the distribution function at 0 is τ. Due to this restriction, it follows that the model aims at describing the quantile function $Q_{Y_i}(\tau|\boldsymbol{x}_i)$ of the continuous response variable Y_i conditional on covariate vector $\boldsymbol{x}_i$ at a given quantile τ, more specifically

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = H_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \ , \tag{2}$$

where H_{Y_i} is the cumulative distribution function of Y_i. Note that, in principle, every ordinary mean regression, like linear or additive models, imply quantile modeling of the response variable because the distributional assumptions on the conditional response also determine its conditional quantiles. Although regression models like generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos 2005), allow to introduce additional flexibility, they typically do not result in easily interpretable expressions for the quantiles, since they are based on specifying distinct distributional parameters.

An alternative, common representation of linear quantile regression can be achieved via the following minimization problem:

$$\underset{\boldsymbol{\beta}_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau) \qquad \text{where} \qquad \rho_\tau(u) = \begin{cases} u\tau & u \geq 0 \\ u(\tau - 1) & u < 0. \end{cases} \tag{3}$$

For $\tau = 0.5$, the so called 'check function' $\rho_\tau(u)$ is proportional to the absolute value function, i.e., $\rho_{0.5}(u) = 0.5|u|$. The minimization problem in (3) can be formulated as a set of linear constraints, therefore the estimation of $\boldsymbol{\beta}_\tau$ can be conducted by linear programming and leads to the $\tau \cdot 100\%$ quantiles of the response variable, see Koenker (2005). Thus, the check function is the appropriate loss function for quantile regression problems regarded from a decision theoretical point of view.

However, in cases where nonlinear relationships between covariates and quantiles of the response variable occur, more flexibility is needed. Such nonlinear effects have been found in mean regressions for malnutrition, for example for the effect of the child's age or the body mass index of the mother. Of course, it seems plausible to expect similar nonlinear patterns when turning the attention to quantile modeling for malnutrition. To account for possible nonlinearities, the above model can be extended to additive quantile regression models which allow for the inclusion of nonlinear covariate effects. The corresponding quantile function is

$$Q_{Y_i}(\tau|\boldsymbol{x}_i, \boldsymbol{z}_i) = \eta_{\tau i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \sum_{j=1}^{q} f_{\tau j}(z_{ij}) \ , \tag{4}$$

where the predictor $\eta_{\tau i}$ is composed of a linear term $\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau$ including an intercept and a sum of nonlinear terms, where $f_{\tau j}$, for $j = 1, \ldots, q$, denote smooth functions of continuous covariates z_{ij} which are assumed to relate in a nonlinear way to the response's quantile function and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^\top$. Thereby, the underlying assumption on the error term remains the same as in (1).

The estimation of (4) is possible in an easy manner by using spline functions for the nonlinear terms, e.g., B-spline basis functions, with a fixed and relatively small number of knots at fixed positions. Since the evaluations of the selected basis functions are known, they can be included in the linear design matrices and thus, the additive model can be estimated by linear programming algorithms from linear quantile regression. However, in this case the question arises how to determine the number and positions of knots adequately. To avoid an arbitrary choice of these parameters, penalty methods, such as quantile smoothing splines treated in Koenker, Ng, and Portnoy (1994), are in use. For a univariate situation with only one continuous covariate z_i ($q = 1$) the minimization problem in (3) is extended by a penalty term to

$$\underset{f_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(z_i)) - \lambda V(f'_\tau) \ . \tag{5}$$

Here, $V(f'_\tau)$ denotes the total variation of the derivative f'_τ which is defined as $V(f'_\tau) = \sup \sum_{i=1}^{n} |f'_\tau(z_{i+1}) - f'_\tau(z_i)|$ and λ is a tuning parameter that controls the smoothness of the estimated function. Therefore, this approach is also called 'total variation regularization'. For continuously differentiable f'_τ, the total variation can be written as $V(f'_\tau) = \int |f''_\tau(z)| dz$, i.e., as the L_1-norm of f''_τ. This points out the link to penalty approaches in mean regression where the penalty term consists of the L_2-norm of f''_τ. In classical quantile regression, the L_2-norm is less suitable since it inhibits the use of linear programming to determine the optimal estimate. Koenker *et al.* (1994) show that the solution to (5) can still be obtained by linear programming when considering a somewhat larger function space comprising also functions with derivative existing only almost everywhere. Within this function space, the minimizer of (5) is a piecewise linear spline function with knots at the observations z_i, for further details see Koenker *et al.* (1994); Koenker (2005). An implementation of this technique is available in the function `rqss()` of the **quantreg** package (Koenker 2008) in R (R Development Core Team 2008).

An alternative approach for estimating additive quantile regression models based on local polynomial estimation with nice asymptotic properties has been suggested by Horowitz and Lee (2005). Also, other versions for the penalty term in (5) are imaginable, e.g., an L_1-norm as in Li and Zhu (2008); Wang and Leng (2007) or a Reproducing Kernel Hilbert Space (RKHS) norm as explored in Takeuchi, Le, Sears, and Smola (2006) and Li, Liu, and Zhu (2007). By using the RKHS norm, Takeuchi *et al.* (2006) obtain remarkable results, particularly with regard to prevention of quantile crossing. However, the estimation of nonlinear effects by piecewise linear splines might seem somewhat limited if smoother curves are of interest. Moreover, the choice of λ is crucial for the shape of the estimated functions, but currently there is no algorithm implemented to select λ automatically at least for additive models with several nonlinear functions.

A practically important issue arising in any regression context is model and variable selection. Thereby questions concerning the design and inclusion of covariate effects are of interest: How should continuous covariates be included in the model, in linear or nonlinear form? Which covariates and interaction effects are relevant and necessary in order to describe the response variable adequately? Is it possible to identify and to rank the covariates according to their importance? For usual regression models, the Akaike Information Criterion (AIC) is one approach to answer this kind of questions. For linear quantile regression, Koenker (2005) suggests an adapted AIC (aAIC) where the likelihood is replaced by the empirical risk in (3) as follows:

$$\text{aAIC}(\tau) = -2 \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau) \right) + 2\, p \tag{6}$$

Here, p denotes the number of parameters estimated by the model. Alternative criteria have been proposed in the literature which are also based on the replacement of the likelihood by the empirical risk, see e.g., Cade, Noon, and Flather (2005). In order to trade off between model fit and the number of parameters, this ‘ad hoc’ replacement seems to make sense; on the other hand the question arises how it can be justified theoretically. In case of additive quantile regression models, the AIC criterion in (6) can be modified to:

$$\text{aAIC}(\lambda) = -2 \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{f}_{\tau,\lambda}(z_i)) \right) + \frac{1}{n} p_\lambda \ , \tag{7}$$

where p_λ are the effective degrees of freedom, which can be interpreted as the number of ‘active’ knots in the resulting piecewise linear spline, see Koenker and Mizera (2004). Although the specification of (7) (and analogous criteria) was originally motivated by the bandwidth choice for λ in (5), it can also be used for model and variable selection.

In this article, we propose boosting as an alternative estimation method for linear and additive quantile regression models by combining the models described above with the boosting algorithms for additive models described in Kneib, Hothorn, and Tutz (2009). In brief, boosting is an optimization algorithm that aims at minimizing an expected loss criterion by stepwise updating an estimator according to the steepest gradient descent of the loss criterion. In order to find the stepwise maxima, base-learners are used, i.e., simple regression models fitting the negative gradient by (penalized) least-squares. For quantile regression, the check function $\rho_\tau(\cdot)$ is employed as appropriate loss function.

With the objective of quantile regression, Kriegler and Berk (2007) also combine boosting with the loss function $\rho_\tau(\cdot)$, but they use regression trees as base-learners in contrary to the additive modeling approach described here. Therefore, when using larger trees as base-learners the final model can only be described as a 'black box' and does not easily allow to quantify the partial influence of the single covariates on the response, as provided by our approach. Stumps as base-learners lead to non-smooth step functions for each of the covariates. In a similar way, Meinshausen (2006) introduces a machine-learning algorithm that permits quantile regression by linking random forests to the check function. This leads again to a black box which is justified by focusing rather on constructing prediction intervals for new observations than on quantifying the influence of covariates on the response.

In summary, the advantages offered by our boosting approach are the following: (i) Additive quantile regression estimation is embedded in the well studied class of boosting algorithms for empirical risk minimization. (ii) Estimation of additive quantile regression is usually conducted by linear programming algorithms. In case of additive models with a nonlinear predictor this yields piecewise linear functions as estimators for the nonlinear effects. By using a boosting algorithm, the flexibility in estimating the nonlinear effects is considerably increased, since the specification of differentiability of the nonlinear effects remains part of the model specification and is not determined by the estimation method itself. (iii) In comparison to the currently available software for additive quantile regression, more complex models with a larger number of nonlinear effects can be fitted using our approach. (iv) The variable and model selection process is implicitly supported when using boosting for model estimation. In particular, parameter estimation and variable selection are combined into one single model estimation procedure. (v) Finally, standard boosting software can be used for estimating quantile regression models.

The remainder of this article is structured as follows: Section 2 introduces a functional gradient descent boosting algorithm as an alternative for estimation in linear and additive quantile regression models. Section 3 presents the results of an empirical simulation study to compare usual linear programming algorithms and boosting for estimation in quantile regression, also with regard to model selection. In Section 4, our methods are applied to and evaluated on the India childhood malnutrition data set. Section 6 contains concluding remarks.

2. Quantile Regression by Boosting

Functional gradient boosting as discussed extensively in Friedman (2001) and Bühlmann and Hothorn (2007) is a functional gradient descent algorithm that aims at finding the solution to the optimization problem

$$\eta^* = \operatorname*{argmin}_{\eta} \mathbb{E}[L(y, \eta)] \tag{8}$$

where η is the predictor of a regression model and $L(\cdot,\cdot)$ corresponds to the loss function that represents the estimation problem. For practical purposes, the expectation in (8) has to be replaced by the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i,\eta_i).$$

In case of additive quantile regression, the appropriate loss function is given by the check function introduced in the decision theoretical justification of quantile modeling, i.e., $L(y,\eta) = \rho_\tau(y-\eta)$. The regression model, on the other hand, is specified by the general additive predictor in (4). To facilitate description of the boosting algorithm, we will suppress dependence of regression effects on the quantile τ in the following.

Different types of base-learning procedures are of course required for linear and nonlinear effects. Let $\boldsymbol{\beta}$ be decomposed into disjoint sets of parameter vectors $\boldsymbol{\beta}_l$ such that $\boldsymbol{\beta} = (\boldsymbol{\beta}_l, l = 1,\dots,L)$ (possibly after appropriate re-indexing) and let $\boldsymbol{X}_l$ denote the corresponding design matrix. Each of the coefficient vectors $\boldsymbol{\beta}_l$ relates to a block of covariates that shall be attributed to a joint base-learning procedure. For example, all binary indicator variables representing a categorical covariate will typically be subsumed into a vector $\boldsymbol{\beta}_l$ with one single base-learner. Another example are polynomials of a covariate, where also several regression coefficients may be combined into a single base-learner. Still, in most cases $\boldsymbol{\beta}_l$ will simply correspond to a single regression coefficient forming the effect of a single covariate component of the vector $\boldsymbol{x}$. The base-learner assigned to a vector $\boldsymbol{\beta}_l$ will be denoted as $\boldsymbol{b}_l$ in the following. Similarly, the base-learner for the vector of function evaluations $\boldsymbol{f}_j = (f_j(z_{1j}),\dots,f_j(z_{nj}))^\top$ will be denoted as $\boldsymbol{g}_j$.

A componentwise boosting algorithm for additive quantile regression models is then given as follows:

[i.] Initialize all parameter blocks $\boldsymbol{\beta}_l$ and vectors of function evaluations $\boldsymbol{f}_j$ with suitable starting values $\hat{\boldsymbol{\beta}}_l^{[0]}$ and $\hat{\boldsymbol{f}}_j^{[0]}$. Choose a maximum number of iterations m_{stop} and set the iteration index to $m = 1$.

[ii.] Compute the negative gradients of the empirical risk

$$u_i = -\left.\frac{\partial}{\partial\eta} L(y_i,\eta)\right|_{\eta=\hat{\eta}_i^{[m-1]}},\ i = 1,\dots,n,$$

that will serve as working responses for the base-learning procedures. Inserting the check function for the loss function yields the negative gradients

$$u_i = \rho'_\tau(y_i - \hat{\eta}_i^{[m-1]}) = \begin{cases} \tau & y_i - \hat{\eta}_i^{[m-1]} > 0 \\ 0 & y_i - \hat{\eta}_i^{[m-1]} = 0 \\ \tau - 1 & y_i - \hat{\eta}_i^{[m-1]} < 0. \end{cases}$$

[iii.] Fit all base-learning procedures to the negative gradients to obtain estimates $\hat{\boldsymbol{b}}_l^{[m]}$ and $\hat{\boldsymbol{g}}_j^{[m]}$ and find the best-fitting base-learning procedure, i.e., the one that minimizes the L_2 loss

$$(\boldsymbol{u} - \hat{\boldsymbol{u}})^\top(\boldsymbol{u} - \hat{\boldsymbol{u}})$$

inserting either $\boldsymbol{X}_l\hat{\boldsymbol{b}}_l^{[m]}$ or $\hat{\boldsymbol{g}}_j^{[m]}$ for $\hat{\boldsymbol{u}}$.

[iv.] If the best-fitting base-learner is the linear effect with index l^*, update the corresponding coefficient vector as

$$\hat{\boldsymbol{\beta}}_{l^*}^{[m]} = \hat{\boldsymbol{\beta}}_{l^*}^{[m-1]} + \nu \hat{\boldsymbol{b}}_{l^*}^{[m]}$$

where $\nu \in (0,1]$ is a given step size, and keep all other effects constant, i.e.,

$$\hat{\boldsymbol{\beta}}_{l}^{[m]} = \hat{\boldsymbol{\beta}}_{l}^{[m-1]},\ l \neq l^* \quad \text{and} \quad \hat{\boldsymbol{f}}_{j}^{[m]} = \hat{\boldsymbol{f}}_{j}^{[m-1]},\ j = 1, \ldots, q.$$

Correspondingly, if the best-fitting base-learner is the nonlinear effect with index j^*, update the vector of function evaluations as

$$\hat{\boldsymbol{f}}_{j^*}^{[m]} = \hat{\boldsymbol{f}}_{j^*}^{[m-1]} + \nu \hat{\boldsymbol{g}}_{j^*}^{[m]}$$

and keep all other effects constant, i.e.,

$$\hat{\boldsymbol{\beta}}_{l}^{[m]} = \hat{\boldsymbol{\beta}}_{l}^{[m-1]},\ l = 1, \ldots, L, \quad \text{and} \quad \hat{\boldsymbol{f}}_{j}^{[m]} = \hat{\boldsymbol{f}}_{j}^{[m-1]},\ j \neq j^*.$$

[v.] Unless $m = m_{\text{stop}}$ increase m by one and go back to [ii.].

Note that there is some ambiguity in defining the gradient since the check function is not differentiable in zero. In practice, this case will only occur with zero probability (for continuous responses), so there is no conceptual difficulty and the gradient could similarly be defined as $\rho'_\tau(0) = \tau$ (as in Meinshausen 2006) or as $\rho'_\tau(0) = \tau - 1$.

To complete the specification of the componentwise boosting algorithm for additive quantile regression, the starting values, the base-learning procedures, the number of boosting iterations m_{stop} and the step length factor ν have to be chosen. While it is natural to initialize all effects at zero, it turns out that faster convergence and more reliable results are obtained by defining a fixed offset as a starting value for the intercept. An obvious choice may be the τ-th sample quantile of the response variable but our empirical experience suggests that the median is more suitable in general, as will be illustrated in an example in Section 3.

Concerning the base-learning procedures, least-squares base-learners are a natural choice for the parametric effects, i.e.,

$$\hat{\boldsymbol{b}}_{l}^{[m]} = (\boldsymbol{X}_l^\top \boldsymbol{X}_l)^{-1} \boldsymbol{X}_l^\top \boldsymbol{u}.$$

For nonlinear effects, we consider penalized spline base-learners that can be cast in the framework of penalized least-squares estimation. Penalized splines can be motivated from simple scatterplot smoothing for inferring a non-linear relationship $u_i = g_j(z_{ij}) + \epsilon_i$ from data (u_i, z_{ij}), $i = 1, \ldots, n$. Following Eilers and Marx (1996), we approximate the function $g_j(z_j)$ in terms of a moderately sized B-spline basis, i.e.,

$$g_j(x) = \sum_{k=1}^{K} \gamma_{jk} B_k(z_j)$$

where $B_k(z_j)$ are B-splines of degree D defined upon a set of equidistant knots. The degree D can be chosen by the user according to subject-matter knowledge to obtain a function estimate with the desired overall smoothness properties since a spline of degree D is $D-1$

times continuously differentiable. Estimation of the spline coefficients $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jK})^\top$ is based on minimizing the penalized least squares criterion

$$\underset{\boldsymbol{\gamma}_j}{\operatorname{argmin}}\, (\boldsymbol{u} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j)^\top(\boldsymbol{u} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j) + \lambda_j\boldsymbol{\gamma}_j^\top\boldsymbol{K}\boldsymbol{\gamma}_j \tag{9}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ is the vector of responses and $\boldsymbol{Z}_j$ is the corresponding B-spline design matrix. The penalty term augmented to the least squares fit criterion consists of a smoothing parameter λ_j as in (5) that trades off fit against smoothness and a penalty matrix $\boldsymbol{K}$ that penalizes variability in the function estimate. Eilers and Marx (1996) suggest a simple approximation to the typical integrated squared derivative penalties that is based on squared differences within the sequence of coefficients $\boldsymbol{\gamma}_j$ and leads to the penalty matrix $\boldsymbol{K} = \boldsymbol{D}^\top\boldsymbol{D}$ where $\boldsymbol{D}$ is a difference matrix, usually of second order to approximate the second derivative. Solving (9) yields the penalized least squares estimate

$$\hat{\boldsymbol{g}}_j^{[m]} = \boldsymbol{Z}_j(\boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \lambda_j\boldsymbol{K})^{-1}\boldsymbol{Z}_j^\top\boldsymbol{u}$$

that defines the base-learning procedure for a non-linear effect $f_j(z_j)$ (Schmid and Hothorn 2008).

The step length factor ν and the optimal number of boosting iterations trade off each other with smaller step lengths resulting in more boosting iterations and vice versa. Therefore we can safely fix one of them and derive an optimal choice only for the remaining quantity. Since the number of boosting iterations is easier to vary in practice, we fix the step length at $\nu = 0.1$ to obtain relatively small steps of the boosting algorithm. The optimal number of boosting iterations m_{stop} should then be chosen to obtain a model that generalizes well to new data. In the presence of test data, m_{stop} can therefore be determined by evaluating the empirical risk on the test data as a function of the boosting iterations and by choosing the point of minimal risk on the test data.

Stopping the boosting algorithm early enough is also crucial to employ the inherent variable selection and model choice abilities of boosting. Suppose that a large number of covariates is available in a particular application. Then the boosting algorithm will start by picking the most influential ones first since those will allow for a better fit to the negative gradients. When the boosting algorithm is stopped after an appropriate number of iterations, spurious noninformative covariates are likely to be not selected and therefore effectively drop from the model equation. In addition, both the first iteration when a base-learner has been selected and the relative frequency of selections within the total m_{stop} boosting iterations may serve as rough guides characterising the importance of an effect. When considering competing modeling possibilities, such as linear and nonlinear base-learners for the same covariate, boosting also enables model choice. Note also that the componentwise boosting approach with separate base-learners for the different effects allows to set up candidate models that may even contain more covariates than observations.

3. Empirical Evaluation

In order to evaluate the performance of the algorithm introduced in Section 2, we conducted a simulation study. In particular, we wanted to explore three partial questions: (Q1) How does boosting estimation work in situations with linear effects on the response's quantile function, i.e., when linear quantile regression is appropriate? (Q2) How does boosting estimation work in situations with nonlinear covariate effects on the response's quantile function, i.e., when additive quantile regression is appropriate? (Q3) Is boosting estimation capable of selecting the covariates with influence on the response's quantile function correctly?

Although the linear simulation setup is not aimed at additive regression models directly, we think that it is indispensable in order to get an idea how our algorithm works for linear quantile regression – also in consideration of the fact that the additive model in (4) consists of both, a linear and a nonlinear term. In addition, this setup gives us the opportunity to compare boosting estimation with linear programming which can be seen as the current 'gold standard' for quantile regression estimation.

3.1. Linear Quantile Regression

Model. To investigate the question (Q1) for the linear simulation setup, we considered the following location-scale-model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, \varepsilon_i \qquad \text{where} \quad \varepsilon_i \overset{iid}{\sim} H \quad \text{for } i = 1, \ldots, n \tag{10}$$

Here, the location as well as the scale of the response y_i depend in linear form on a covariate vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ and an error term ε_i with distribution function H_ε not depending on covariates. The coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ affects the response's location while $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ affects its scale. The resulting quantile function has a linear predictor structure and can be written as

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, H^{-1}(\tau) = \boldsymbol{x}_i^\top (\boldsymbol{\beta} + \boldsymbol{\alpha} H^{-1}(\tau)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \ .$$

Hence, quantile specific coefficients as in (1) can be determined as $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + \boldsymbol{\alpha} H^{-1}(\tau)$.

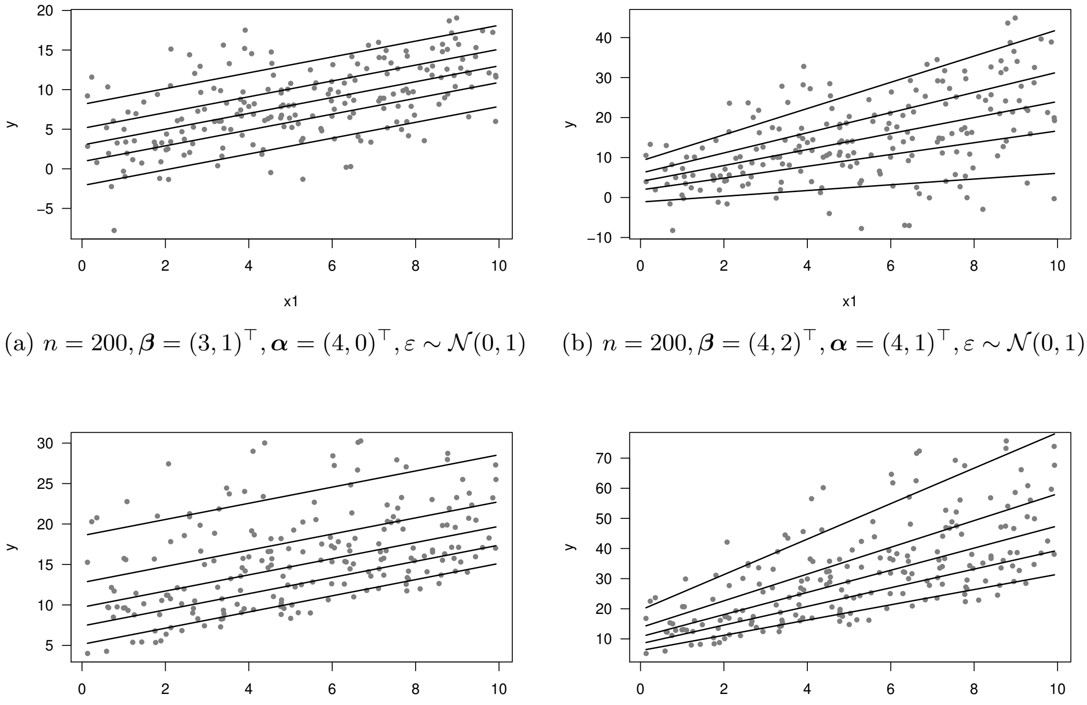
Based on the linear model in (10), we draw 100 datasets with the following parameter combinations:

- Homoscedastic setup: $n = 200, \boldsymbol{\beta} = (3,1)^\top, \boldsymbol{\alpha} = (4,0)^\top$
- Heteroscedastic setup: $n = 200, \boldsymbol{\beta} = (4,2)^\top, \boldsymbol{\alpha} = (4,1)^\top$
- Multivariable setup: $n = 500, \boldsymbol{\beta} = (5,8,-5,2,-2,0,0)^\top, \boldsymbol{\alpha} = (1,0,2,0,1,0,0)^\top$

All required covariates were independently drawn from a continuously uniform distribution $\mathcal{U}[0,10]$. We repeated all setups for three different distributions of the error terms: a standard normal distribution, a t-distribution with 2 degrees of freedom and a gamma distribution, where $\mathbb{E}(\varepsilon_i) = \mathbb{V}(\varepsilon_i) = 2$. Figure 1 visualizes data examples from the first two setups with one covariate for normal or gamma distributed error terms. Note that $\boldsymbol{\alpha} = (4,1)$ leads to a heteroscedastic data structure where the quantile curves are no longer parallel shifted as for $\boldsymbol{\alpha} = (4,0)$.

Estimation. For each of the generated datasets, we estimated the parameter vector $\boldsymbol{\beta}_\tau$ for a fixed quantile grid on $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ by our algorithm (function `glmboost()` from package **mboost**) and by linear programming (function `rq()` from package **quantreg**). In the boosting case, we fixed the step length at $\nu = 0.1$ and determined the optimal number of boosting iterations m_{stop} by evaluating the empirical risk on a test dataset with 1000 observations drawn from the respective simulation setup and by choosing the point of minimal risk on the test data.

As already mentioned in Section 2, we decided to take the median as starting value for the intercept instead of the τ-th sample quantile of the response variable. This decision was based on the following empirical results. For quantiles smaller than $\tau = 0.5$, we explored



(a) $n = 200, \boldsymbol{\beta} = (3, 1)^\top, \boldsymbol{\alpha} = (4, 0)^\top, \varepsilon \sim \mathcal{N}(0, 1)$ (b) $n = 200, \boldsymbol{\beta} = (4, 2)^\top, \boldsymbol{\alpha} = (4, 1)^\top, \varepsilon \sim \mathcal{N}(0, 1)$

(c) $n = 200, \boldsymbol{\beta} = (3, 1)^\top, \boldsymbol{\alpha} = (4, 0)^\top, \varepsilon \sim \mathcal{G}(1, 2)$ (d) $n = 200, \boldsymbol{\beta} = (4, 2)^\top, \boldsymbol{\alpha} = (4, 1)^\top, \varepsilon \sim \mathcal{G}(1, 2)$

Figure 1: Data examples for linear simulation setups with one covariate in a homoscedastic (left) or heteroscedastic (right) data structure with normal (top) or gamma (bottom) distributed error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

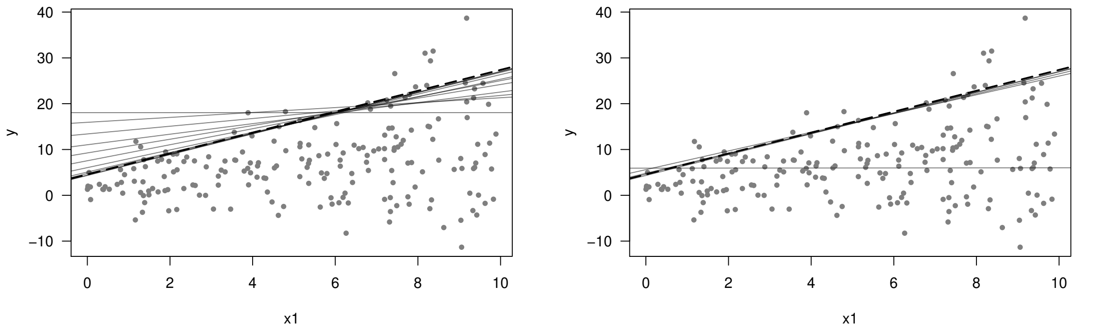
hardly any differences between resulting m_{stop} criteria and estimators for $\boldsymbol{\beta}_\tau$ depending on the starting values. However, for quantiles larger than $\tau = 0.5$ the m_{stop} criterion was dramatically increased when taking the τ-th sample quantile as starting value. As an example, Figure 2 illustrates the stepwise approach of the boosting estimation to the true underlying 90% quantile curves depending on the starting value. Note that it takes considerably more iterations until the estimation approaches the true quantile curve when beginning at the 0.9-th sample quantile, shown in Figure 2(a). On the contrary, Figure 2(b) displays that the estimation converges much faster when beginning at the median.

Performance results. In order to evaluate and to compare estimation results of the two considered algorithms, we estimated Bias and MSE for each quantile specific parameter $(\beta_{\tau 0}, \beta_{\tau 1}, \ldots, \beta_{\tau p})^\top$ by the following formulae:

$$\text{Bias}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau j,k} - \beta_{\tau j}) \ , \qquad \text{MSE}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau j,k} - \beta_{\tau j})^2 \ , \tag{11}$$

where $j = 0, \ldots, p$ indexes the number of covariates and $k = 1, \ldots, K$ the simulation replications. In case of boosting, we also considered the m_{stop} criteria.

In the following, we will focus on a short summary of the results by just showing some typical examples. Figure 3 displays boxplots for the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$



(a) Starting value = 90% quantile (horizontal line), $m_{\text{stop}} = 18474$

(b) Starting value = median (horizontal line), $m_{\text{stop}} = 7513$

Figure 2: Data example with parameters $n = 200$, $\boldsymbol{\beta} = (2, 1)^\top$ and $\boldsymbol{\alpha} = (2, 1)^\top$ and normal distributed error terms. Dashed black lines show the true underlying quantile curve for $\tau = 0.9$, grey lines illustrate the stepwise boosting fit after each 2000 iterations beginning at the horizontal line.

in the heteroscedastic setup with normal distributed error terms. Note that estimators resulting from linear programming (`rq`) are less biased but have a larger variance than those resulting from boosting (`boost`). This is consistent to previously reported results and to the fact that boosting estimators are usually shrunken towards zero, which can be traced back to the implicit regularization property of boosting estimation (Bühlmann and Hothorn 2007).

Regarding the MSE, Table 1 shows estimators for setups with one covariate and gamma distributed error terms, obtained according to (11). For the slope estimator $\hat{\beta}_{\tau 1}$, boosting achieves smaller MSE estimators on almost the whole quantile grid. Concerning the intercept estimator $\hat{\beta}_{\tau 0}$, boosting performs better in the homoscedastic setup while linear programming obtains better results in the heteroscedastic setup.

In addition, the optimal number of boosting iterations, determined by means of test data, ranges roughly between 3000 and 10000 in cases with one covariate and is considerably increased (30000 – 70000) for the multivariable model with six covariates.

Table 1: Estimated MSE criteria from 100 replications of linear simulation setups with one covariate and gamma distributed error terms. Shown in bold are quantile and parameter specific smaller estimators.

	Homoscedastic setup				Heteroscedastic setup			
	MSE($\beta_{\tau 0}$)		MSE($\beta_{\tau 1}$)		MSE($\beta_{\tau 0}$)		MSE($\beta_{\tau 1}$)	
τ	rq	boost	rq	boost	rq	boost	rq	boost
0.1	**0.328**	0.350	0.010	**0.008**	**0.762**	1.007	0.050	**0.038**
0.3	0.676	**0.582**	0.016	**0.012**	**1.417**	1.475	0.063	**0.052**
0.5	0.732	**0.685**	0.020	**0.015**	**1.627**	1.962	0.099	**0.074**
0.7	1.751	**1.595**	0.048	**0.040**	4.168	**4.165**	0.229	**0.157**
0.9	4.983	**2.992**	0.129	**0.066**	**10.404**	17.971	**0.618**	0.657

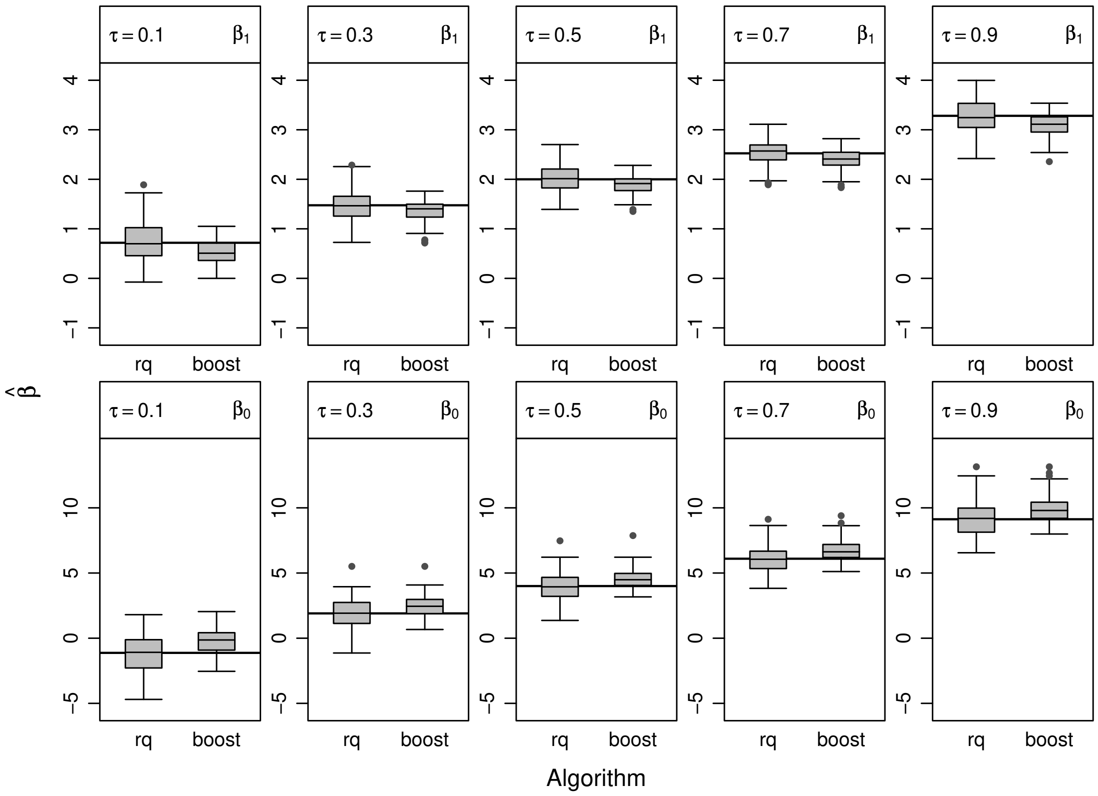


Figure 3: Simulation results for heteroscedastic linear setup with one covariate and normal distributed error terms. Boxplots display the empirical distribution of the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$ from 100 replications, depending on quantile τ and estimation algorithm (`rq` for linear programming and `boost` for boosting). Horizontal lines designate true underlying parameters $(\beta_{\tau 0}, \beta_{\tau 1})^\top$.

Even if not plotted here, we observed similar results for all other simulation setups, i.e., with more covariates or alternative error distributions. Therefore, we conclude that boosting estimation is competitive to linear programming estimation in situations with linear effects on the response's quantile function, i.e., when linear quantile regression is appropriate.

Variable selection results. Concerning model and variable selection, we wanted to explore whether the algorithms are able to extract the right covariates in the multivariable setup. In case of linear programming, models for all different covariate combinations were estimated followed by a calculation of aAIC values. Then, the covariate combination with the smallest aAIC value was chosen. In case of boosting, we answered the following three questions: Which covariate was not chosen at all during the boosting estimation? When was a covariate chosen for the first time? In how many iterations was a covariate chosen? In this regard, we observed the following results: The more important a covariate was (measured in terms of $|\beta_\tau|$), the earlier it was chosen for the first time and the more often it was chosen during the estimation process, and this independent of τ. In the majority of cases, only covariates with $\beta_\tau = 0$ were not chosen at all. Some problems

Table 2: Summary of variable selection results for $\tau = 0.7$ from linear multivariable simulation setup with normal distributed error terms. β coefficients are quantile specific for $\tau = 0.7$. MPI: Mean proportion of iterations (relating to m_{stop}), where covariate was chosen; MFI: Mean first iteration (relating to m_{stop}), where covariate was chosen; PEB: Proportion of simulations (relating to 100), where covariate was not chosen by boosting; PEA: Proportion of simulations (relating to 100), where covariate was excluded in model with smallest aAIC.

		Int.	x_1	x_2	x_3	x_4	x_5	x_6
		$\beta_0 = 5.5$	$\beta_1 = 8.0$	$\beta_2 = -4.0$	$\beta_3 = 2.0$	$\beta_4 = -1.5$	$\beta_5 = 0$	$\beta_6 = 0$
`boost`	MPI	0.284	0.266	0.134	0.170	0.084	0.036	0.035
	MFI	0.323	0.000	0.027	0.191	0.129	0.430	0.428
	PEB	0	0	0	0	0	0.11	0.16
`rq`	PEA	0	0	0	0	0	0.33	0.21

occured at upper quantiles in the setup with gamma distributed error terms, but in these cases also the aAIC driven model selection did not lead to the correct model. To exemplify these results, Table 2 gives a summary for normal distributed error terms and quantile $\tau = 0.7$. It shows that the covariates x_5 and x_6 with both $\beta_{0.7,5} = \beta_{0.7,6} = 0$, i.e., no influence on the response, are chosen fewer and later than all other covariates. Compared to variable selection by aAIC, this criterion excludes non significant covariates more often than boosting.

To sum up, boosting provides useful support in the variable selection process, even though there is currently no explicit criterion available. Particularly in cases with numerous covariates, boosting has the advantage that it yields to variable selection information within the estimation process, whereas the use of aAIC requires multiple model fits.

3.2. Additive Quantile Regression

Model. Analogous to the linear simulation setup in (10) we considered an additive model with nonlinear terms for location and scale, as follows:

$$y_i = \beta_0 + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + [\alpha_0 + g_1(z_{i1}) + \ldots + g_q(z_{iq})]\,\varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{iid}{\sim} H \qquad (12)$$

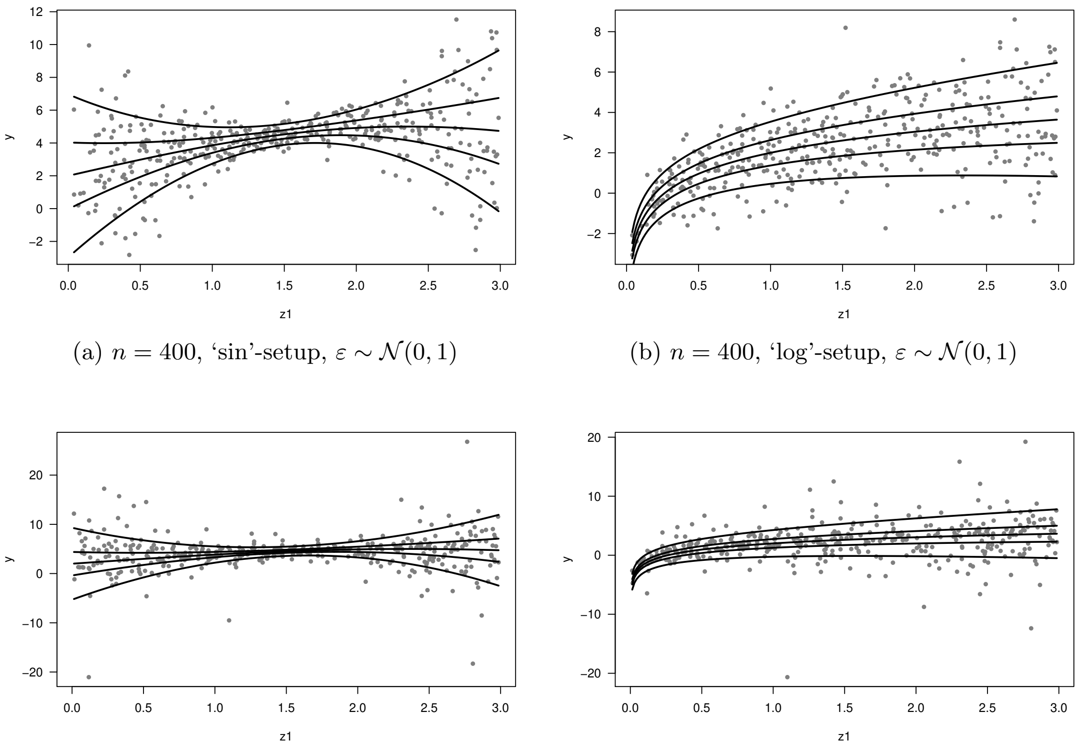
In this model, location and scale of the response can depend in nonlinear form on covariates $z_{i1}, \ldots, z_{iq}$. Choosing all f and g as linear functions yields the linear model as in (10). If some functions f and g are zero, the associated covariates have no influence on the response. The resulting quantile function has a nonlinear predictor structure, as given by

$$Q_{Y_i}(\tau|\boldsymbol{z}_i) = \beta_0 + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + H^{-1}(\tau)[\alpha_0 + g_1(z_{i1}) + \ldots + g_q(z_{iq})]\ .$$

Note that it is not possible to explicitly determine quantile specific coefficients here, like $\boldsymbol{\beta}_\tau$ in the linear simulation.

Based on the additive model in (12), we draw 100 datasets with the following parameter combinations:

'sin'-setup: $q = 1$ $\quad \beta_0 = 2$ $\quad \alpha_0 = 0.5$ $\quad f_1(z_{i1}) = 3\sin(\frac{2}{3}z_{i1})$ $\quad g_1(z_{i1}) = 1.5(z_{i1} - 1.5)^2$
'log'-setup: $q = 1$ $\quad \beta_0 = 2$ $\quad \alpha_0 = 0.7$ $\quad f_1(z_{i1}) = 1.5\log(z_{i1})$ $\quad g_1(z_{i1}) = 0.5z_{i1}$
multivariable setup: $q = 6$ $\quad \beta_0 = 2$ $\quad \alpha_0 = 0.7$
$f_1(z_{i1}) = 3\sin(\frac{2}{3}z_{i1})$ $\quad f_2(z_{i2}) = 1.5\log(z_{i2})$ $\quad f_3(z_{i3}) = 2$ $\quad f_4(z_{i4}) = -2$ $\quad f_5(z_{i5}) = f_6(z_{i6}) = 0$
$g_1(z_{i1}) = 1.5(z_{i1} - 1.5)^2$ $\quad g_2(z_{i2}) = g_3(z_{i3}) = 0.5$ $\quad g_4(z_{i4}) = g_5(z_{i5}) = g_6(z_{i6}) = 0$



(a) $n = 400$, 'sin'-setup, $\varepsilon \sim \mathcal{N}(0, 1)$

(b) $n = 400$, 'log'-setup, $\varepsilon \sim \mathcal{N}(0, 1)$

(c) $n = 400$, 'sin'-setup, $\varepsilon \sim t(2)$

(d) $n = 400$, 'log'-setup, $\varepsilon \sim t(2)$

Figure 4: Data examples for nonlinear simulation setups with one covariate in the 'sin'-setup (left) or 'log'-setup (right) with standard normal (top) or $t(2)$ (bottom) distributed error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

We fixed the number of observations to 400 for all setups. Covariates z_i were independently drawn from a uniform distribution $\mathcal{U}[0, 3]$. As in the linear simulation, we repeated all setups for a standard normal, a gamma and a t distribution. In the multivariable setup, two covariates relate in a nonlinear way to the response, two have a linear influence on it and the last two have no influence at all. Figure 4 shows data examples from 'sin'-and 'log'-setups for normal and t distributed error terms. Due to its heavy tail property, the t-distribution leads to some extreme outliers.

Estimation. For each of the generated datasets, we estimated nonlinear effects for all covariates for a fixed quantile grid on $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ by our algorithm (function `gamboost()` from package **mboost**) and by linear programming with a total variation regularization approach (function `rqss()` from package **quantreg**). In case of boosting, we used cubic penalized spline base-learners with second order difference penalty, 20 inner knots and three degrees of freedom, and fixed the step length at $\nu = 0.1$. In accordance with the linear simulations, the optimal number of boosting iterations m_{stop} was chosen by means of a test dataset with 1000 observations. This test data was also used to determine covariate specific smoothing parameters $\lambda_1, \ldots, \lambda_q$ for the penalty terms for `rqss`, as given in (5), at the point of minimal risk on the test data.

Performance results. The evaluation of the estimation results was also based on these test datasets. In every simulation step, models were estimated on training data while

Table 3: Estimated evaluation criteria from 100 replications of nonlinear 'log' simulation setup with gamma distributed error terms. Shown in bold are quantile specific better results.

	Risk		Bias		MSE	
τ	rqss	boost	rqss	boost	rqss	boost
0.1	0.248	**0.245**	**0.003**	0.019	0.059	**0.048**
0.3	0.593	**0.590**	0.028	**0.023**	0.080	**0.071**
0.5	0.772	**0.769**	0.028	**0.015**	0.113	**0.097**
0.7	0.761	**0.758**	0.026	**-0.016**	0.177	**0.149**
0.9	0.454	**0.451**	**0.062**	-0.104	0.392	**0.281**

evaluation was conducted on the test data. Therefore we considered the empirical risk as well as a sort of Bias and MSE for nonlinear functions. The resulting quantile specific empirical risk is given by

$$\text{Risk}(\tau) = \frac{1}{100}\sum_{k=1}^{100}\text{Risk}_{\tau k} \quad \text{where} \quad \text{Risk}_{\tau k} = \frac{1}{1000}\sum_{j=1}^{1000}\rho_\tau(y_j - \hat{y}_{\tau j,k}) \ .$$

Here, y_j stands for the response of observation j in the test data, while $\hat{y}_{\tau j,k}$ denotes the estimated response value at quantile τ for observation j and iteration k. Thus, the final empirical risk is determined as the mean of the single risks from 100 replications. Analogously the quantile specific Bias and MSE were estimated as means of the single criteria, which can directly combined to

$$\text{Bias}(\tau) = \frac{1}{100\cdot 1000}\sum_{k=1}^{100}\sum_{j=1}^{1000}(\hat{y}_{\tau j,k} - y_{\tau j}) \qquad \text{MSE}(\tau) = \frac{1}{100\cdot 1000}\sum_{k=1}^{100}\sum_{j=1}^{1000}(\hat{y}_{\tau j,k} - y_{\tau j})^2 \ ,$$

where $y_{\tau j}$ is the true underlying τ-th quantile of the response. Note that Bias and MSE as defined above can be interpreted as monte carlo estimators of the true Bias and MSE of the nonlinear functions.

Just as for the linear simulation results, we will shortly summarize the results and give some typical examples. Table 3 shows the described performance criteria for the 'log' setup with gamma distributed error terms.

At first glance it appears that boosting performs better than total variation regularization, but a closer look, which is provided by Figure 5, shows that the performance results are mostly located in the same range for both algorithms. Similar results were obtained for the other situations.

In order to illustrate the estimation results, Figure 6 displays the estimated quantile curves for the 'sin'-setup with normal distributed error terms. Even if the estimated curves obtained by boosting seem to be smoother than the piecewise linear curves obtained by total variation regularization, there are hardly any differences between the performance results.

Regarding mean m_{stop} criteria from boosting, they are considerably smaller than those from the linear simulation. Even in case of the multivariable setup, less than 5000 iterations are needed in the majority of simulations. This might be due to a larger flexibility of the

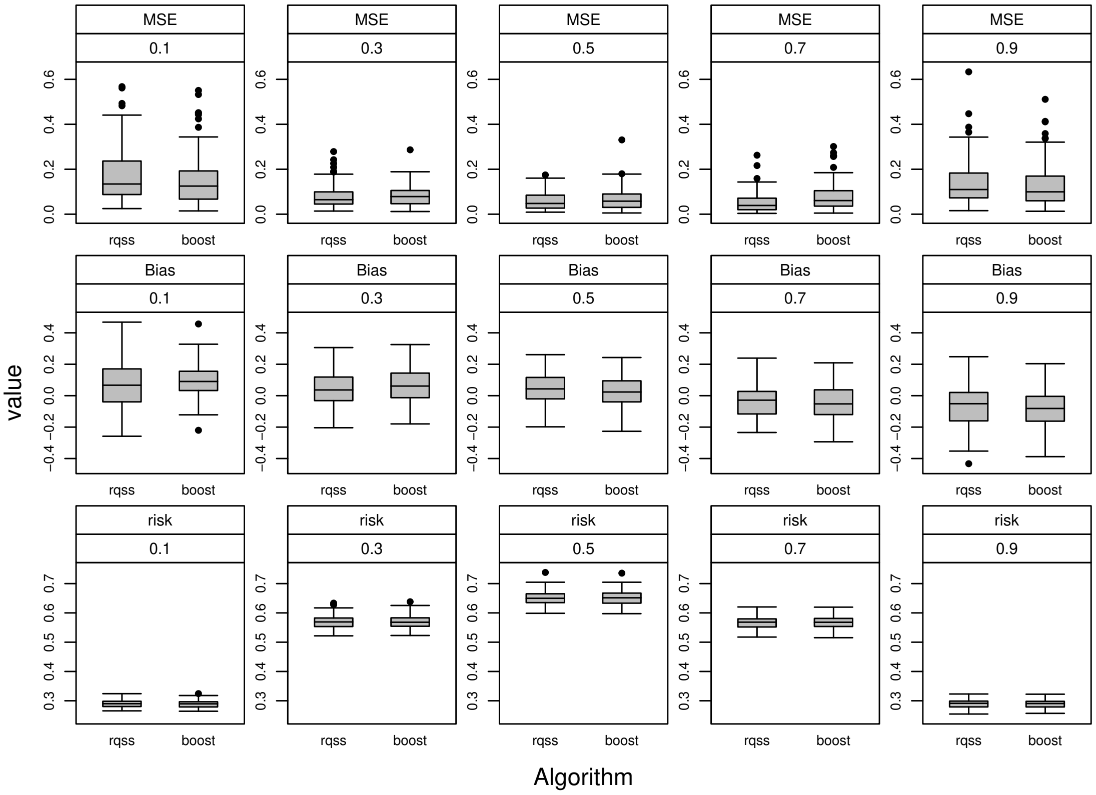
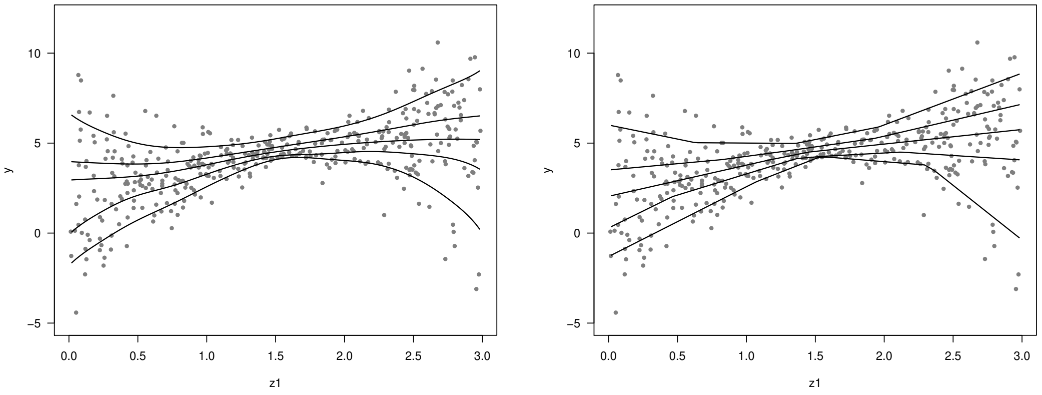


Figure 5: Simulation results for nonlinear 'log'-setup with gamma distributed error terms. Boxplots display the empirical distribution of the estimated criteria Risk, Bias and MSE from 100 replications, depending on quantile τ and estimation algorithm (`rq` for linear programming and `boost` for boosting).

used base-learners which permit nonlinearities and can adapt much faster to the respective data situation than linear functions.

Variable selection results. Concerning variable selection in the multivariable setup, we observed analogous results as for the linear simulation. Covariates without influence were chosen less often during the estimation procedure and considerably later for the first time. Therefore, we refrain from describing these results in detail. However, we could not compare our results to those obtained by total variation regularization followed by calculating aAIC because of severe estimation problems with the function `rqss()` from package **quantreg** in R.

In summary, we conclude that boosting and total variation regularization lead to comparable performance results for additive quantile regression. However, by using our boosting approach, the flexibility in estimating the nonlinear effects is considerably increased, since the specification of differentiability remains part of the model specification and is not determined by the estimation method itself. Comparing the currently avaible software for both algorithms, boosting can handle a larger number of nonlinear covariate effects.



(a) Boosting estimation

(b) Total variation regularization

Figure 6: Example for estimated quantile curves for ‘sin’-setup and standard normal distributed error terms. True underlying quantile curves are shown in Figure 4(a).

4. Childhood Malnutrition in India

Malnutrition of wide parts of the population, and in particular childhood malnutrition, is one of the most urgent problems in developing and transition countries. In order to provide information not only on the nutritional status but on health and population trends in general, MEASURE Demographic and Health Surveys (DHS) conduct nationally representative surveys on fertility, family planning, maternal and child health, as well as child survival, HIV/AIDS, malaria, and nutrition. The resulting data, more than 200 surveys in 75 countries so far, are available free of charge for research purposes.

Childhood malnutrition is usually measured in terms of a score Z that compares an anthropometric characteristic of the child to values from a reference population, i.e.,

$$Z_i = \frac{\mathrm{AC}_i - m}{s}$$

where AC denotes the anthropometric characteristic of interest while m and s correspond to median and standard deviation in the reference population (stratified with respect to age, gender, and some further covariates). While weight might be considered as the most obvious indicator for malnutrition, we will focus on stunting, i.e., insufficient height for age, in the following. Stunting provides a measure of chronic malnutrition, while insufficient weight for age might result from either acute or chronic malnutrition. Note that the score Z, despite its name, is not assumed to be normally distributed. Typically, it will not even be symmetric or have mean zero, since it is used to assess the nutritional status in a malnourished population with respect to a reference population. The score is not standardized and ranges from -600 to 600 in our data set.

Most previous analyses like the ones in Kandala *et al.* (2001, 2008) have analyzed malnutrition based on regression models for the expectation of the malnutrition score, yielding an implicit focus on average nutritional status. However, when we are interested in severe malnutrition, regression models for quantiles such as the 5% or the 10% quantile may be much more interesting and natural.

Table 4: Variables in the childhood malnutrition data set.

Variable	Explanation
Z	Score for stunting (continuous)
cage	age of the child in months (continuous)
cfeed	duration of breastfeeding in months (continuous)
csex	gender of the child (categorical: male, female)
ctwin	indicator for twin children (categorical: single birth, twin)
cbord	number of the child in the birth order (categorical: 1,2,3,4,5)
mbmi	body mass index of the mother (continuous)
mage	age of the mother in years (continuous)
medu	years of education of the mother (continuous)
medupart	years of education of the mother‘s partner (continuous)
munem	employment status of the mother (categorical: employed, unemployed)
mreli	religion of the mother (categorical: christian, hindu, muslim, sikh, other)
resid	place of residence (categorical: rural, urban)
nodead	number of dead children (categorical: 0,1,2,3)
wealth	wealth index (categorical: poorest, poorer, middle, richer, richest)
electricity	household has electricity supply (categorical: yes, no)
radio	household has a radio (categorical: yes, no)
tv	household has a television (categorical: yes, no)
fridge	household has a refrigerator (categorical: yes, no)
bicycle	household has a bicycle (categorical: yes, no)
mcycle	household has a motorcycle (categorical: yes, no)
car	household has a car (categorical: yes, no)

In the following, we will present an analysis on childhood malnutrition in India based on DHS data from 2005/06 since India is one of the fastest growing economies and the second-most populated country in the world. From the original data set obtained from www.measuredhs.com, we extracted a number of covariates that are deemed to be important determinants of childhood malnutrition, see Table 4 for an overview and short descriptions. Based on this set of covariates, we specified a candidate model where all continuous covariates are included with possibly nonlinear effects based on cubic penalized spline base-learners with second order difference penalty, 20 inner knots and five degrees of freedom. All categorical covariates were assigned least-squares base-learners where dummies corresponding to different levels of the same covariate are combined into one single base-learner. This yields the quantile-specific model equation

$$\begin{aligned} Z_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + f_{\tau 1}(\text{cage}_i) + f_{\tau 2}(\text{cfeed}_i) + f_{\tau 3}(\text{mbmi}_i) + f_{\tau 4}(\text{mage}_i) \\ &\quad + f_{\tau 5}(\text{medu}_i) + f_{\tau 6}(\text{medupart}_i) + \varepsilon_{\tau i}. \end{aligned}$$

We considered three different quantiles, namely 5%, 10%, and 50%, to compare effects on severe malnutrition as well as effects on average nutrition measured in terms of the median. After plausibility checks and deletion of observations with missing values, we obtained a data set with 37,623 observations. Two thirds of the data were used for estimation, while the remaining third (12,541 observations) were used to determine the optimal stopping iteration of the boosting algorithm. To be more specific, we evaluated the out-of-sample risk on one third of the data and chose m_{stop} as the iteration index where the out-of-sample

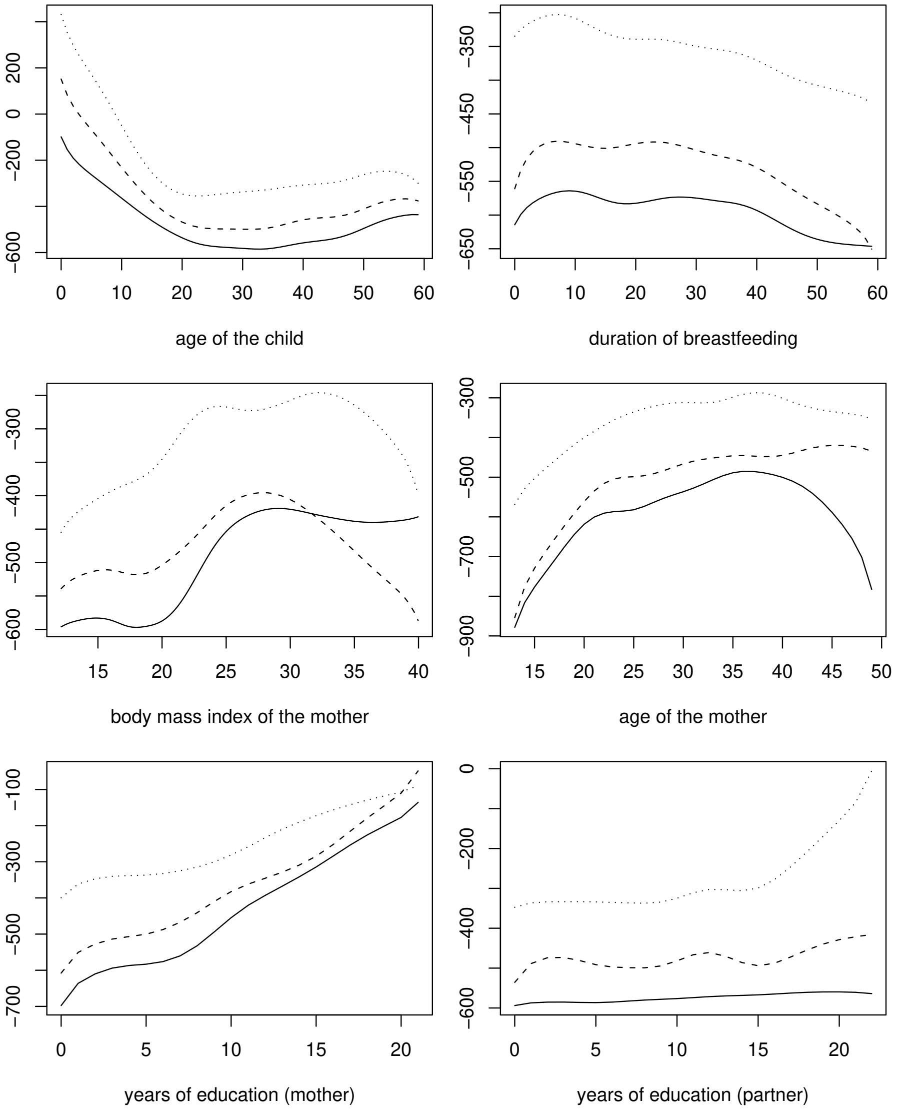


Figure 7: Estimated nonparametric effects for the 50% quantile (dotted line), the 10% quantile (dashed line) and the 5% quantile (solid line) of the stunting score. All effects are adjusted for the overall quantile levels.

risk was minimized. The minimum was achieved after 107,754 (5% quantile), 84,966 (10% quantile) and 41,702 (50% quantile) iterations.

Estimated nonlinear effects for the three selected quantiles are visualized in Figure 7. Note that all effects are centered around the average effect of all further continuous covariates and that the reference category has been inserted for categorical covariates to make the levels comparable. In most cases, we find the expected relation between the three estimated effect curves insofar as the median effect has a higher level than the 10% quantile which itself has a higher level than the 5% quantile. However, there are also some rare cases where quantile crossing occurs, namely for the effect of the mother‘s body mass index and (to a much lesser extent) for years of education and duration of breastfeeding. The underlying reason is that separate models have been fitted for the three quantiles and

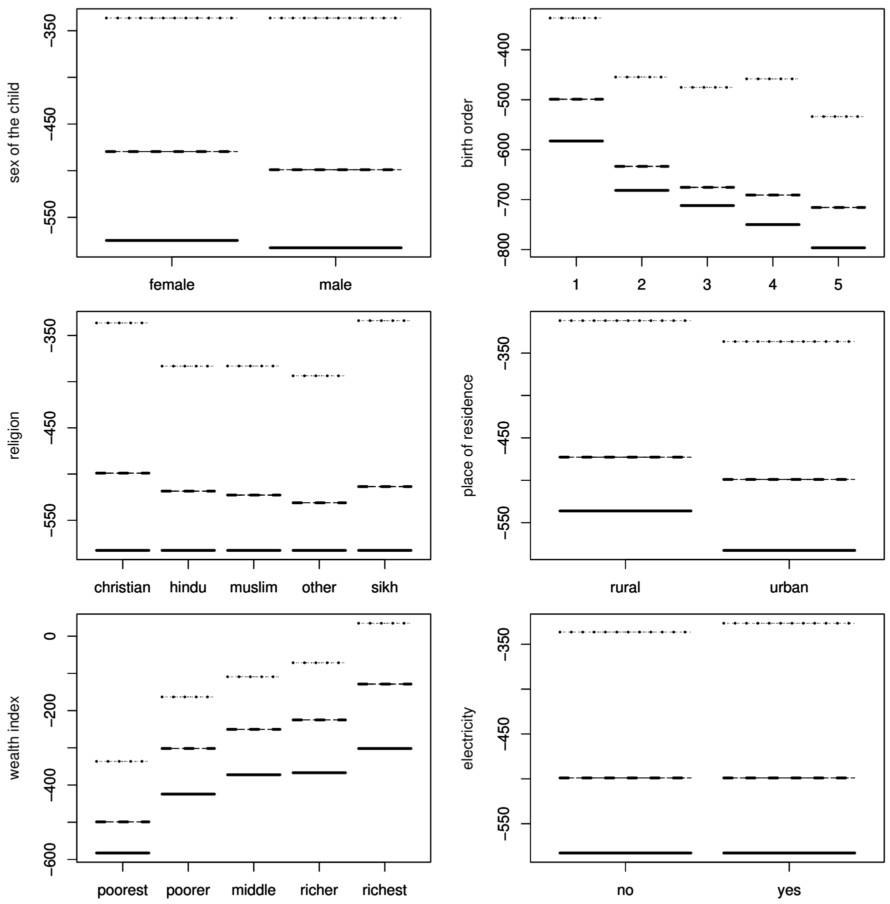


Figure 8: Selected estimated effects of categorical covariates for the 50% quantile (dotted line), the 10% quantile (dashed line) and the 5% quantile (solid line) of the stunting score. All effects are adjusted for the overall quantile levels.

therefore no ordering restriction can be imposed on the estimated curves. Note that we have checked the distribution of the covariates and that the crossing of quantiles does not seem to be related to sparse data in this area of the covariate domain.

The effect of the age of the child shows a strong decrease within the first months which stabilizes at an almost constant level after 20 months. This qualitative behaviour is consistently observed for all quantiles and has also been found in regression models for the expectation. Still, the decrease in the first 20 months is more strongly expressed for the median regression than for the 5% quantile regression curve. Another interesting effect is found for the age of the mother. It has been hypothesized (see for example Kandala *et al.* (2001)) that this effect should be close to an inverse u-shape which we find for the 5% quantile while the 10% quantile and the median effects rise steadily over the domain of observed ages. Obviously, these differences could not have been identified with a usual regression model for the expectation.

Figure 8 shows the effects of some selected categorical covariates. Again, all effects have been adjusted for the overall quantile level to make them comparable across the three quantile regression models. The effects of gender and the presence of electricity supply in a household seem to have no influence on any of the three quantiles, but interesting

Table 5: Variable selection information for the childhood malnutrition data. FI: First iteration (relative to m_{stop}) when a variable was chosen. PI: Proportion of iterations (relative to m_{stop}) when a variable was chosen.

	$\tau = 0.05$		$\tau = 0.1$		$\tau = 0.5$	
Variable	FI	PI	FI	PI	FI	PI
cage	0.034	0.161	0.017	0.272	0.001	0.204
cfeed	0.273	0.084	0.125	0.069	0.020	0.174
csex	0.275	0.017	0.212	0.019	0.213	0.007
ctwin	0.328	0.035	0.128	0.025	0.063	0.012
cbord	0.061	0.092	0.057	0.071	0.032	0.046
mbmi	0.070	0.077	0.057	0.064	0.013	0.054
mage	0.106	0.161	0.082	0.092	0.035	0.175
medu	0.000	0.097	0.000	0.091	0.000	0.065
medupart	0.070	0.081	0.026	0.137	0.017	0.122
munem	0.277	0.021	0.302	0.009	0.303	0.002
mreli	0.786	0.006	0.212	0.012	0.064	0.013
resid	0.275	0.035	0.228	0.021	0.097	0.014
nodead	0.216	0.023	0.108	0.029	0.069	0.020
wealth	0.000	0.040	0.002	0.030	0.000	0.031
electricity	0.811	0.006	0.711	0.000	0.052	0.009
radio	NA	0.000	0.826	0.003	NA	0.000
tv	NA	0.000	NA	0.000	0.236	0.003
fridge	NA	0.000	0.948	0.000	0.036	0.011
bicycle	0.061	0.044	0.074	0.034	0.302	0.008
mcycle	NA	0.000	0.610	0.002	0.045	0.016
car	0.255	0.019	0.100	0.018	0.047	0.013

differences can be found for other covariates. For example, the effect of position in the birth order does not show a clear pattern for the median regression, apart from a sharp drop between position one and two. In contrast, both lower quantiles show a steady decrease of the nutritional status for increasing birth order. For religion, we find the reverse behaviour: The effect on the two lower quantiles seems to be negligible while at least some effect is observed on the median. The effect of the wealth index reflects the expected relation that richer households indicate a better nutritional status of the child. This effect is found consistently over the three quantiles with most differences being related to the shift between the quantiles.

Table 5 collects information on the inherent variable selection of boosting-based quantile regression. Apart from some of the asset variables, all covariates are included in all three regression models in at least one iteration. The very early inclusion of the covariates wealth and education of the mother indicates their relevance, again irrespective of the chosen quantile. The proportion of iterations reflect the importance of age of the child and age of the mother, which also corresponds to the strongly nonlinear effect observed for these two variables. Note that equal contributions of all variables would lead to proportions of $1/21 = 0.0476$.

5. Computational Details

For the implementation of our boosting algorithm, already available standard software for boosting was slightly extended. The described methodology is implemented in the R add-on package **mboost** (Hothorn, Bühlmann, Kneib, Schmid, and Hofner 2009). Linear or additive quantile regression can easily be performed by using the standard functions `glmboost()` or `gamboost()`, respectively, in combination with the argument `family=QuantReg()` allowing for specification of τ and an offset quantile.

In order to make the results of our data analyses reproducible, an electronic supplement to this paper contains all necessary R commands to prepare and analyze the Indian malnutrition data (provided that the original data set has been obtained from www.measuredhs.com). If package **mboost** is installed, simply type

```
R> source(system.file("India_quantiles.R", package = "mboost"))
```

to reproduce our analyses.

6. Discussion

Motivated by the analysis of risk factors for childhood malnutrition in India, we developed a boosting algorithm for estimation in additive quantile regression models in this paper. The data our investigation is based on were collected in the 2005/06 Demographic and Health Survey and contained numerous covariates as well as a malnutrition score serving as response. By using its lower quantiles instead of just the mean or median as regression objective—i.e., by using quantile regression—it was possible to identify risk factors for severe malnutrition and not only for the population average of the score. Before applying our boosting algorithm to the India data, we conducted an empirical simulation study in order to explore if and how the method works, also in comparison with currently used algorithms for quantile regression estimation which are based on linear programming.

The results of this empirical evaluation suggest that boosting estimation is competitive to linear programming, both for linear and additive quantile regression models. With regard to variable selection, boosting provides useful support in the following way. When a covariate is not chosen at all during the estimation process, this indicates no (or only weak) influence on the response. For the other covariates being chosen, further information about their importance can be obtained by checking how often they are chosen, and in which iteration they are chosen for the first time.

The application of additive quantile regression to the India data led to interesting results which could not have been obtained with a usual mean regression model. For the age of the mother, an inverse u-shape effect was detected for the 5% quantile but not for the median. Concerning the categorical covariate birth order, the estimated effects for the lower quantiles were considerably different than for the median. We conclude that quantile regression models are an appropriate tool for exploring risk factors for childhood malnutrition. However, since all quantile curves are estimated independent from each other, these models may also involve problems with quantile crossing, as observed for the body mass index of the mother.

In comparison to total variation regularization, boosting estimation for additive quantile regression offers three main advantages. First, boosting enables data-driven determination of the amount of smoothness required for the nonlinear effects and does not necessarily lead to piecewise linear functions. Second, comparing the currently available software for both algorithms, boosting can handle a larger number of nonlinear covariate effects. Third, parameter estimation and variable selection are executed in one single estimation step which is particularly favorable for high-dimensional predictors. Both estimation algorithms require the specification of a hyper parameter, the optimal number of iterations m_{stop} in case of boosting and the smoothness parameter λ in case of total variation regularization, which is usually done by splitting the original dataset into training and test data.

Apart from malnutrion, quantile modeling is also of interest in applications where the quantiles depend on covariates in a different way than the mean, with the simplest form being heteroscedastic data. Other typical areas of application for quantile modeling are the construction of reference charts in epidemiology (e.g., Wei 2008), the analysis of quantiles of gene expression through probe level measurements (Wang and He 2007), or the analysis of the value at risk in financial econometrics, see Yu, Lu, and Stander (2003) for further examples. Our approach helps to overcome the variable selection and model choice problems, especially when the primary aim is to fit a sparse quantile regression model based on a moderate or high number of potentially useful covariates.

Extensions of the boosting algorithm for random and spatial effects seem feasible by including these effects in the predictors of the base-learners in a way similar to the approach developed by Kneib *et al.* (2009). The same methodology can then be used to allow for time-varying effects, an application also studied by Cai and Xu (2008). Using similar techniques, future research will focus on quantile regression for longitudinal data to account for more complex data structures.

Acknowledgements

We thank Jan Priebe for his assistance in the preparation of the India data and Ludwig Fahrmeir for motivating the investigation of boosting for quantile regression. This paper has been written while Thomas Kneib was visiting the University of Göttingen in the winter term 2008/09.

References

Buchinsky M (1998). "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research." *The Journal of Human Resources*, **33**(1), 88–126.

Bühlmann P, Hothorn T (2007). "Boosting Algorithms: Regularization, Prediction and Model Fitting (with Discussion)." *Statistical Science*, **22**(4), 477–505.

Bühlmann P, Yu B (2003). "Boosting with the L_2 Loss: Regression and Classification." *Journal of the American Statistical Association*, **98**, 324–339.

Cade B, Noon B, Flather C (2005). "Quantile Regression Reveals Hidden Bias and Uncertainty in Habitat Models." *Ecology*, **86**(3), 786–800.

Cai Z, Xu X (2008). "Nonparametric Quantile Estimations for Dynamic Smooth Coefficient Models." *Journal of the American Statistical Association*, **103**(484), 1595–1608.

Eilers P, Marx B (1996). "Flexible Smoothing with B-Splines and Penalties." *Statistical Science*, **11**(2), 89–121.

Friedman J (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, **29**(5), 1189–1232.

Friedman J, Hastie T, Tibshirani R (2000). "Additive Logistic Regression: A Statistical View of Boosting (with Discussion)." *The Annals of Statistics*, **28**, 337–407.

Horowitz J, Lee S (2005). "Nonparametric Estimation of an Additive Quantile Regression Model." *Journal of the American Statistical Association*, **100**(472), 1238–1249.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2009). *Model-Based Boosting*. R package version 1.1-0, URL http://CRAN.R-project.org/package=mboost.

Kandala NB, Fahrmeir L, Klasen S, Priebe J (2008). "Geo-Additive Models of Childhood Undernutrition in Three Sub-Saharan African Countries." *Population, Space and Place*. To appear.

Kandala NB, Lang S, Klasen S, Fahrmeir L (2001). "Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of Undernutrition in Two African Countries." *Research in Official Statistics*, **4**(1), 81–99.

Kneib T, Hothorn T, Tutz G (2009). "Variable Selection and Model Choice in Geoadditive Regression Models." *Biometrics*. To appear.

Koenker R (2005). *Quantile Regression*. Economic Society Monographs. Cambridge University Press, New York.

Koenker R (2008). *quantreg: Quantile Regression*. R package version 4.26, URL http://www.r-project.org/package=quantreg.

Koenker R, Mizera I (2004). "Penalized Triograms: Total Variation Regularization for Bivariate Smoothing." *Journal of the Royal Statistical Society: Series B*, **66**(1), 145–163.

Koenker R, Ng P, Portnoy S (1994). "Quantile Smoothing Splines." *Biometrika*, **81**(4), 673–680.

Kriegler B, Berk R (2007). "Boosting the Quantile Distribution: A Cost-Sensitive Statistical Learning Procedure." *Technical report*, Department of Statistics, UCLA. URL http://www.crim.upenn.edu/faculty/papers/berk/quantile2.pdf.

Li Y, Liu Y, Zhu J (2007). "Quantile Regression in Reproducing Kernel Hilbert Spaces." *Journal of the American Statistical Association*, **102**(477), 255–268.

Li Y, Zhu J (2008). "L_1-Norm Quantile Regression." *Journal of Computational and Graphical Statistics*, **17**(1), 163–185.

Meinshausen N (2006). "Quantile Regression Forests." *Journal of Machine Learning Research*, **7**, 983–999.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rigby R, Stasinopoulos D (2005). "Generalized Additive Models for Location, Scale and Shape." *Applied Statistics*, **54**(3), 507–554.

Schmid M, Hothorn T (2008). "Boosting Additive Models using Component-wise P-splines as Base-learners." *Computational Statistics & Data Analysis*, **53**(2), 298–311.

Takeuchi I, Le Q, Sears T, Smola A (2006). "Nonparametric Quantile Estimation." *Journal of Machine Learning Research*, **7**, 1231–1264.

Wang H, He X (2007). "Detecting Differential Expressions in GeneChip Microarray Studies: A Quantile Approach." *Journal of the American Statistical Association*, **102**(477), 104–112.

Wang H, Leng C (2007). "Unified LASSO Estimation by Least Squares Approximation." *Journal of the American Statistical Association*, **102**(479), 1039–1048.

Wei Y (2008). "An Approach to Multivariate Covariate-dependent Quantile Contours with Application to Bivariate Conditional Growth Charts." *Journal of the American Statistical Association*, **103**(481), 397–409.

Yu K, Lu Z, Stander J (2003). "Quantile Regression: Applications and Current Research Areas." *The Statistician*, **52**(3), 331–350.

Affiliation:

Nora Fenske
Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstraße 33
DE-80539 München, Germany
E-mail: Nora.Fenske@stat.uni-muenchen.de
URL: http://www.stat.uni-muenchen.de/~fenske/

Thomas Kneib
Fakultät V
Institut für Mathematik
Carl-von-Ossietzky-Universität Oldenburg
DE-26111 Oldenburg i.O.

Torsten Hothorn
Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstraße 33
DE-80539 München, Germany