Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

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

Nora Fenske, Thomas Kneib & Torsten Hothorn

To cite this article: Nora Fenske, Thomas Kneib & Torsten Hothorn (2011) Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression, Journal of the American Statistical Association, 106:494, 494-510, DOI: 10.1198/jasa.2011.ap09272

To link to this article: https://doi.org/10.1198/jasa.2011.ap09272

Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

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

Nora FENSKE, Thomas KNEIB, and Torsten HOTHORN

We investigated the risk factors for childhood malnutrition in India based on the 2005/2006 Demographic and Health Survey by applying a novel estimation technique for additive quantile regression. 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. The use of such a regression model for analyzing childhood malnutrition in developing or transition countries implies that the estimated effects describe the average nutritional status. However, it is of even greater interest to analyze quantiles of the response distribution, such as the 5% or 10% quantile, which relate to the risk of extreme malnutrition. Our investigation is based on a semiparametric extension of quantile regression models where different types of nonlinear effects are included in the model equation, leading to additive quantile regression. We addressed the variable selection and model choice problems associated with estimating such an additive quantile regression model using a novel boosting approach. Our proposal allows for data-driven determination of the amount of smoothness required for the nonlinear effects and combines model choice with an automatic variable selection property. In an empirical evaluation, we compared our boosting approach with state-of-the-art methods for additive quantile regression. The results suggest that boosting is an appropriate tool for estimation and variable selection in additive quantile regression models and helps to identify yet unknown risk factors for childhood malnutrition. This article has supplementary material online.

KEY WORDS: Additive models; Functional gradient boosting; Model choice; Penalized splines; Stunting; Variable selection.

1. INTRODUCTION

The reduction of malnutrition and in particular childhood malnutrition is among the United Nations Millennium Development Goals, which aims at halving the proportion of people suffering from hunger by 2015. Childhood malnutrition is one of the most urgent public health problems in developing and transition countries since it not only affects children's growth directly but also has severe long-term consequences. Caulfield et al. (2004) estimated that malnutrition is an underlying cause for about 53% of child deaths worldwide. Therefore, a better understanding of malnutrition risk factors is of utmost importance.

Here we focus on analyzing risk factors for chronic childhood malnutrition in India, one of the fastest growing economies and the second-most populated country in the world. Our investigation is based on India's 2005/2006 Demographic and Health Survey (DHS). The final report of this survey (NFHS 2007) provides results on the nutritional status of children in India and emphasizes the severity of this issue. Based on the preceding DHS from 1998/1999, numerous investigations focusing on different aspects of malnutrition risk factors have been carried out: Som, Pal, and Bharati (2007) and Rajaram, Zottarelli, and Sunil (2007) explore individual and household factors of childhood malnutrition while Bharati, Pal, and Bharati (2008) and Nair (2007) consider the regional structure of malnutrition in India. For a summary of childhood nutrition in India in the 1990s, see Tarozzi and Mahajan (2007).

Childhood malnutrition is usually 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. The nutritional status is expressed by anthropometric characteristics, that is, height for age; in cases of chronic childhood malnutrition, the reduced growth rate in human development is termed stunted growth or stunting. In previous analyses of stunting, for example, by Som, Pal, and Bharati (2007), Rajaram, Zottarelli, and Sunil (2007), Mishra and Retherford (2007), and Ackerson and Subramanian (2008), children are classified as stunted based on this score, followed by a logistic regression for the resulting binary response. Other recent analyses (see for example Kandala et al. 2001, 2009), have focused on mean regression for the stunting score. In contrast to these approaches, we apply quantile regression for estimating the influence of potential risk factors on the lower quantiles of the conditional distribution of the stunting score. Stunting is thereby expressed by the 5% or 10% quantiles of the score, in contrast to mean regression, which describes the average nutritional status. Fenske et al. (2008) proceeded similarly for modeling overweight and obesity in developed countries, but for malnutrition, statistical analyses of lower quantiles are, to the best of our knowledge, still lacking. Such analyses can add important additional information since they avoid possible loss of information implied by a classification and directly concentrate on the malnutrition part of the score distribution.

Nora Fenske is Research Associate (E-mail: nora.fenske@stat.uni-muenchen.de) and Torsten Hothorn is Professor (E-mail: torsten.hothorn@stat.uni-muenchen.de), Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 München, Germany. Thomas Kneib is Professor, Institut für Mathematik, Carl von Ossietzy Universität Oldenburg, 26111 Oldenburg, Germany (E-mail: thomas.kneib@uni-oldenburg.de). The authors thank two anonymous reviewers, an associate editor, and the editor for valuable comments, which led to a substantial improvement in our manuscript. The authors are indebted to Roger Koenker for fruitful discussions on fitting additive quantile regression with `rqss()` and the analysis of the malnutrition data and to Karen Brune for improving the language. Furthermore, the authors thank Jan Priebe for assisting in the preparation of the India data and Ludwig Fahrmeir for motivating the investigation of boosting for quantile regression. This article was written while Thomas Kneib was a visiting professor at the Georg-August-Universität Göttingen. Nora Fenske and Torsten Hothorn received support by the Munich Center of Health Sciences (MC-Health). Thomas Kneib and Torsten Hothorn acknowledge financial support by the Deutsche Forschungsgemeinschaft (grants KN 922/4-1 and HO 3241/1-3).

Journal of the American Statistical Association
June 2011, Vol. 106, No. 494, Applications and Case Studies
DOI: 10.1198/jasa.2011.ap09272

Also, previous analyses of the mean nutritional status (Kandala et al. 2001, 2009) revealed nonlinear effects of important risk factors, such as the child's or mother's age or the body mass index of the mother. Of course, it seems plausible to expect similar nonlinear patterns when focusing on quantile modeling for malnutrition. Consequently, appropriate modeling has to take such flexible smooth effects into account. Additive quantile regression models allow for such semiparametric predictors including linear and nonlinear effects. Moreover, an inclusion of varying coefficient terms will help to detect specific interaction terms, such as the presence of gender differences in linear as well as nonlinear effects.

Our analysis is based on 21 potentially important risk factors but aims at deriving a parsimonious and interpretable model consisting only of relevant risk factors modeled at appropriate complexity. Hence, we are faced with a variable selection and model choice problem, and therefore, the design and inclusion of covariate effects can be questioned: How should continuous covariates be included in the model: in linear or nonlinear form? Which covariates are relevant and necessary for describing the response variable adequately? Is it possible to identify and to rank the covariates according to their importance? Are there gender-specific differences in some of the effects?

State-of-the-art procedures for additive quantile regression as discussed in Section 2.1 currently lack adequate variable selection and model choice properties, in particular in the extended class of additive models comprising varying coefficients which we are considering. We therefore developed an alternative estimation procedure for additive quantile regression in the empirical risk minimization framework based on a boosting approach. We combine quantile regression models, thoroughly treated in Koenker (2005), with boosting algorithms for additive models described by Kneib, Hothorn, and Tutz (2009). In brief, boosting is an optimization algorithm that aims at minimizing an expected loss criterion by stepwise updating of an estimator according to the steepest gradient descent of the loss criterion. To find the stepwise maxima, base learners are used, that is, simple regression models fitting the negative gradient by (penalized) least squares. For quantile regression, the check function (introduced later) is employed as the appropriate loss function. Boosting has successfully been used to address variable selection and model choice in other contexts, for example, in Friedman, Hastie, and Tibshirani (2000), Bühlmann and Yu (2003), and Bühlmann and Hothorn (2007).

With the objective of quantile regression, Kriegler and Berk (2010) also combine boosting with the check function, but they use regression trees as base learners in contrast to the additive modeling approach described here. Therefore, if larger trees are used as base learners, the final model can be described as a "black box" only and does not easily allow quantification of the partial influence of the single covariates on the response, as provided by our approach. Stumps as base learners lead to nonsmooth 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 on constructing prediction intervals for new observations rather than on quantifying the influence of covariates on the response.

The advantages offered by our boosting approach for detecting risk factors for stunting are the following: (i) The variable selection and model choice process is implicitly supported when boosting is used for model estimation. In particular, parameter estimation and variable selection are combined into one single model estimation procedure. With suitably defined additive predictors, it is possible to choose between a linear effect and corresponding nonlinear deviations for a specific covariate, thus facilitating practical and interesting model choice procedures. (ii) Estimation of additive quantile regression is usually conducted by linear programming algorithms. In the 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 as well as varying coefficient terms can be fitted using our approach. (iv) Additive quantile regression estimation is embedded in the well-studied class of boosting algorithms for empirical risk minimization. Therefore, standard boosting software can be used for estimating quantile regression models. (v) Inferences about the estimated model can be obtained by subsampling replications and stability selection (Meinshausen and Bühlmann 2010).

In the following, Section 2.1 describes state-of-the-art linear and additive quantile regression models as well as corresponding estimation techniques. Section 2.2 introduces a functional gradient descent boosting algorithm as an alternative for estimation in additive quantile regression models. Section 3 presents the results of an empirical investigation to comparing state-of-the-art methods and boosting for estimation in additive quantile regression. In Section 4, we present risk factors and their relationship to childhood malnutrition as obtained from the tailored quantile boosting procedure.

2. METHODOLOGY

2.1 Additive Quantile Regression

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

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

see Buchinsky (1998). Here, the index $i = 1, \ldots, n$, denotes the individual and y_i and $\mathbf{x}_i$ 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. Owing to this restriction it follows that the model aims at describing the quantile function $Q_{Y_i}(\tau|\mathbf{x}_i)$ of the continuous response variable

Y_i conditional on covariate vector $\mathbf{x}_i$ at a given quantile τ, and more specifically

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

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, implies quantile modeling of the response variable because the distributional assumptions on the conditional response also determine its conditional quantiles. Although regression models, such as generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos 2005), enable additional flexibility to be introduced, they typically do not result in easily interpretable expressions for the quantiles because they are based on specifying distinct distributional parameters.

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

$$\arg\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\tau),$$

$$\text{where } \rho_\tau(u) = \begin{cases} u\tau, & u \geq 0 \\ u(\tau - 1), & u < 0. \end{cases} \tag{3}$$

For $\tau = 0.5$, the "check function" $\rho_\tau(u)$ is proportional to the absolute value function, that is, $\rho_{0.5}(u) \propto |u|$. 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. To account for nonlinearities, the above model can be extended to additive quantile regression models which allow for the inclusion of nonlinear covariate effects. The quantile function is then given by

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

where the structured additive predictor $\eta_{\tau i}$ is composed of a linear term $\mathbf{x}_i^\top \boldsymbol{\beta}_\tau$ including an intercept and a sum of nonlinear terms. The functions $f_{\tau j}$, for $j = 1, \ldots, q$, denote generic functions of covariates $\mathbf{z}_i$ for the ith observation. In addition to the covariates $\mathbf{x}$ with linear effects, the vector $\mathbf{z} = (z_1, \ldots, z_r)$ may contain additional variables, for example grouping factors. Although the principal structure of (4) looks like a simple additive model, the generic dependence of $f_{\tau j}$ on the complete covariate vector $\mathbf{z}$ allows for the inclusion of a variety of different model terms such as (i) a nonlinear effect of some univariate continuous covariate z_k where $f_{\tau j}(\mathbf{z}) = f_\tau(z_k)$ is a smooth function of z_k, (ii) a varying coefficient term where $f_{\tau j}(\mathbf{z}) = z_{k'} f_\tau(z_k)$, that is, the effect of covariate $z_{k'}$ varies smoothly over the domain of z_k according to some function f_τ, (iii) a bivariate surface of two continuous covariates z_k and $z_{k'}$ (e.g., longitude and latitude in spatially oriented data) where $f_{\tau j}(\mathbf{z}) = f_\tau(z_k, z_{k'})$, and (iv) cluster-specific effects where $f_{\tau j}(\mathbf{z}) = (z_{k'} I\{z_k \in A_1\}, \ldots, z_{k'} I\{z_k \in A_K\})^\top \boldsymbol{\gamma}_j$, that is, the effect of $z_{k'}$ differs across groups $A_1, \ldots, A_K$ defined by the grouping factor z_k; see Kneib, Hothorn, and Tutz (2009) and Fahrmeir, Kneib, and Lang (2004) for similar generic model specifications. In our application, the focus will be on a special instance of the generic additive model comprising nonlinear and varying coefficient terms. When the predictor structure is modified to (4), the underlying assumptions on the error term remain the same as in (1).

The estimation of purely additive models comprising only nonlinear effects of continuous covariates is easily possible by using spline functions for these terms, for example, 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 design matrices and thus, the additive model can be estimated by linear programming algorithms for 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 used. For a univariate situation with only one continuous covariate z ($q = 1$), the minimization problem in (3) is extended by a penalty term to

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

Here, $V(f'_\tau)$ denotes the total variation of the derivative $f'_\tau : [a, b] \to \mathbb{R}$, which is defined as $V(f'_\tau) = \sup \sum_{i=1}^{n-1} |f'_\tau(z_{i+1}) - f'_\tau(z_i)|$, where the sup is taken over all partitions $a \leq z_1 < \cdots < z_n < b$, 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 $V(f'_\tau) = \int |f''_\tau(z)|\, dz$, that is, 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, Ng, and Portnoy (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 derivatives 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, Ng, and Portnoy (1994) and Koenker (2005). An implementation of this technique is available in the function `rqss()` of the package **quantreg** (Koenker 2010b) in R (R Development Core Team 2010).

An alternative approach for estimating additive quantile regression models based on local polynomial estimation with good asymptotic properties has been suggested by Horowitz and Lee (2005). Also, other versions for the penalty term in (5) are imaginable, for example, 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 et al. (2006) and Li, Liu, and Zhu (2007). By using the RKHS norm, Takeuchi et al. (2006) obtain remarkable results, particularly with regard to the 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.

For the multivariable situation ($q > 1$), the above approach requires an appropriate choice of the smoothing parameters λ_j, $j = 1, \dots, q$, for each nonlinear effect separately. Without an automatic procedure for simultaneous tuning parameter selection at hand, fitting additive models with total variation regularization to more than three covariates is computationally burdensome or even impossible for a moderate number of covariates. These computational challenges and corresponding practical problems in package **quantreg** have been recently discussed by Koenker (2010a).

Variable selection and model choice based on Akaike-type (AIC) and Schwarz-type (SIC) information criteria adapted to quantile regression (see, e.g., Koenker, Ng, and Portnoy 1994; Koenker 2005) requires that all possible models are fitted, and all subsets of covariates and all possible choices of the q tuning parameters have to be taken into account. This renders AIC-based or SIC-based model choice and variable selection in additive quantile regression models fitted via total variation regularization almost impossible, even for problems with a moderate number of covariates. Therefore, in the following, we concentrate on boosting which is an estimation procedure well known for its superb variable selection and model choice properties. This estimation procedure will also allow us to consider the generic additive model comprising more than only nonlinear effects of continuous covariates.

2.2 Estimating Additive 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

$$\eta^* = \underset{\eta}{\arg\min}\, \mathbb{E}[L(y, \eta)], \tag{6}$$

where η is the predictor of a regression model, for example, an additive predictor as specified in (4), and $L(\cdot, \cdot)$ corresponds to the loss function that represents the estimation problem. For practical purposes, the expectation in (6) has to be replaced by the empirical risk $n^{-1} \sum_{i=1}^{n} L(y_i, \eta_i)$.

In the case of additive quantile regression, the appropriate loss function is given by the check function introduced in the decision theoretical justification of quantile modeling, that is, $L(y, \eta) = \rho_\tau(y - \eta)$. 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 reindexing) and let $\mathbf{X}_l$ denote the corresponding design matrices. 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. Other examples 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 $\mathbf{x}$. The base learner assigned to a vector $\boldsymbol{\beta}_l$ will be denoted as $\mathbf{b}_l$ in the following. Similarly, the base learner for the vector of function evaluations $\mathbf{f}_j = (f_j(\mathbf{z}_1), \dots, f_j(\mathbf{z}_n))^\top$ for one of the generic nonlinear effects will be denoted as $\mathbf{g}_j$.

A component-wise 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 $\mathbf{f}_j$ with suitable starting values $\hat{\boldsymbol{\beta}}_l^{[0]}$ and $\hat{\mathbf{f}}_j^{[0]}$. 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

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

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

$$u_i = \rho_\tau'\big(y_i - \hat{\eta}_i^{[m-1]}\big) = \begin{cases} \tau, & y_i - \hat{\eta}_i^{[m-1]} \geq 0 \\ \tau - 1, & y_i - \hat{\eta}_i^{[m-1]} < 0. \end{cases}$$

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

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

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

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

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

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

$$\hat{\boldsymbol{\beta}}_l^{[m]} = \hat{\boldsymbol{\beta}}_l^{[m-1]}, \qquad l \neq l^*, \qquad \text{and}$$
$$\hat{\mathbf{f}}_j^{[m]} = \hat{\mathbf{f}}_j^{[m-1]}, \qquad j = 1, \dots, q.$$

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

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

and keep all other effects constant, that is,

$$\hat{\boldsymbol{\beta}}_l^{[m]} = \hat{\boldsymbol{\beta}}_l^{[m-1]}, \qquad l = 1, \dots, L, \qquad \text{and}$$
$$\hat{\mathbf{f}}_j^{[m]} = \hat{\mathbf{f}}_j^{[m-1]}, \qquad j \neq j^*.$$

[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): therefore, there is no conceptual difficulty. We decided to choose the gradient as $\rho'_\tau(0) = \tau$ (as in Meinshausen 2006), but it could similarly be defined as $\rho'_\tau(0) = \tau - 1$.

To complete the specification of the component-wise 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, 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 τth sample quantile of the response variable, but our empirical experience suggests that the median is more suitable in general, as illustrated in an example in eSupplement B.

With regard to the base-learning procedures, least-squares base learners are a natural choice for the parametric effects, that is, $\hat{\mathbf{b}}_l^{[m]} = (\mathbf{X}_l^\top \mathbf{X}_l)^{-1}\mathbf{X}_l^\top \mathbf{u}$. For nonlinear effects, we consider penalized least-squares (PLS) base learners $\hat{\mathbf{g}}_j^{[m]} = \mathbf{Z}_j(\mathbf{Z}_j^\top \mathbf{Z}_j + \lambda_j \mathbf{K})^{-1}\mathbf{Z}_j^\top \mathbf{u}$ based on suitably defined design and penalty matrices $\mathbf{Z}_j$ and $\mathbf{K}$ corresponding to the jth base learner and model term. Each base learner comprises a smoothing parameter λ_j that trades off fit against smoothness. However, it is important to note that, in the boosting approach, these smoothing parameters are not treated as hyperparameters to be estimated but are determined from prespecified degrees of freedom (Bühlmann and Yu 2003).

For a nonlinear effect of some continuous covariate z, a suitable PLS base learner can be obtained from penalized spline smoothing in a scatterplot smoothing setup for inferring the nonlinear relationship $u = g_j(z) + \varepsilon$ from data $(u, z)_i$, $i = 1, \dots, n$ (Schmid and Hothorn 2008). First, we approximate the function $g_j(z)$ in terms of a moderately sized B-spline basis, that is,

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

where $B_k(z)$ 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 $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})^\top$ is based on minimizing the penalized least-squares criterion

$$\arg\min_{\boldsymbol{\gamma}_j} (\mathbf{u} - \mathbf{Z}_j\boldsymbol{\gamma}_j)^\top(\mathbf{u} - \mathbf{Z}_j\boldsymbol{\gamma}_j) + \lambda_j \boldsymbol{\gamma}_j^\top \mathbf{K}\boldsymbol{\gamma}_j, \tag{7}$$

where $\mathbf{u} = (u_1, \dots, u_n)^\top$ is the vector of responses and $\mathbf{Z}_j$ is the corresponding B-spline design matrix. The penalty matrix $\mathbf{K}$ should be chosen such that it penalizes variability in the function estimate. Eilers and Marx (1996) suggest an 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 $\mathbf{K} = \mathbf{D}^\top\mathbf{D}$, where $\mathbf{D}$ is a difference matrix, usually of second order to approximate the second derivative. To allow the boosting algorithm to differentiate between linear and nonlinear effects of a continuous covariate z, it is often useful to decompose the complete effect of z into

$$f_j(z) = \beta_{0j} + z\beta_{1j} + f_j^{\text{center}}(z), \tag{8}$$

where $\beta_{0j} + z\beta_{1j}$ represents the linear effect of z, whereas $f_j^{\text{center}}(z)$ represents the nonlinear deviation of $f_j(z)$ from this effect. By assigning separate base learners to the linear effect and the nonlinear deviation, the boosting algorithm allows one to decide in a data-driven way whether the linear part in (8) is sufficient to describe the effect of z or whether the nonlinear extension is really required. It is also possible to consider centering around higher-order polynomials (Kneib, Hothorn, and Tutz 2009), although the decision between linear and nonlinear effects seems to be most relevant in practice. In the following, we will simply assume that the linear effects of all continuous covariates are subsumed in the linear part of the predictor and that all nonlinear effects are centered around the linear effects. Technically, this requires a reparameterization of the parameter vector $\boldsymbol{\gamma}_j$, which can be obtained based on the spectral decomposition of the penalty matrix; see Kneib, Hothorn, and Tutz (2009) and Fahrmeir, Kneib, and Lang (2004) for details. Kneib, Hothorn, and Tutz (2009) also discuss appropriate choices for the smoothing parameter λ_j that yield penalized least-squares base-learning procedures that are comparable in complexity to parametric base learners with one free parameter.

For varying coefficient terms $f_{\tau j}(\mathbf{z}) = z_{k'} f_\tau(z_k)$ only a slight modification of the PLS base learner for nonlinear effects is required. To achieve the multiplication of the function evaluations $f_\tau(z_k)$ with the interaction variable $z_{k'}$, the design matrix has to be altered to $\mathbf{Z}_j = \operatorname{diag}(z_{k'1}, \dots, z_{k'n})\mathbf{Z}_j^*$, where $\mathbf{Z}_j^*$ is the design matrix corresponding to a spline approximation of $f_\tau(z_k)$. Inserting $\mathbf{Z}_j$ into the PLS base learner in combination with a difference penalty yields a base-learning procedure for varying coefficients. Similarly, further extensions can be cast in the general form of PLS base learners; see Kneib, Hothorn, and Tutz (2009) for details.

The step-length factor ν and the optimal number of boosting iterations m_{stop} 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 m_{stop} is easier to vary in practice, we fix the step length at $\nu = 0.1$ to obtain relatively small steps of the boosting algorithm. 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 validation 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. 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 component-wise boosting approach with separate base learners for the different effects allows candidate models to be set up that may even contain more model terms than observations.

From a theoretical point of view, the Bayes consistency and consistency of variable selection should be discussed. Note that boosting with early stopping is a shrinkage method with implicit penalty. Therefore, boosting estimates will be biased for finite samples, but typically the bias vanishes for increasing sample sizes. For a quadratic loss function, Bühlmann and Yu (2003) showed that the optimal minimax rate is achieved by component-wise boosting with smoothing splines as base learners. Under rather weak assumptions, Zhang and Yu (2005) showed that models fitted using a boosting algorithm with early stopping attain the Bayes risk. Unfortunately, the results are not directly applicable here since the check function is not twice continuously differentiable with respect to η, and an approximation by means of a continuously differentiable function would have to be applied. An alternative provide expectiles that consider an asymmetrically weighted least-squares loss function (Schnabel and Eilers 2009), leading to a similar characterization of the conditional distribution as with quantile regression but with the advantage of a continuously differentiable loss function.

With regard to consistent variable selection, Bühlmann (2006) studied boosting for linear models, that is, with simple linear models as base learners showing that the procedure yields consistent estimates for high-dimensional problems. There are no similar results available for additive models to the best of our knowledge. We therefore apply a stability selection procedure of Meinshausen and Bühlmann (2010), which leads to consistent variable selection and control of the family-wise error rate.

3. EMPIRICAL EVALUATIONS

The main goals of the empirical investigations presented here were (i) to evaluate the correctness of both the boosting algorithm and its specific implementation on which our subsequent analysis of childhood malnutrition is based on, (ii) to evaluate the variable selection and model choice properties in higher-dimensional settings, and (iii) to judge the quality of estimated quantile functions. For the first goal, we considered typical additive model structures with a moderate number of nonlinear effects while for the second goal we added several nuisance covariates that actually do not impact the response but were still considered as candidate covariates during estimation (Section 3.1). Finally, we compared the estimated quantile functions directly with the true underlying quantile function in a simple univariate setup (Section 3.2).

3.1 Comparing Empirical Risks

Basic Model. The basic model specification for the empirical investigations is

$$y_i = \beta_0 + f_1(z_{i1}) + \cdots + f_q(z_{iq}) + [\alpha_0 + g_1(z_{i1}) + \cdots + g_q(z_{iq})]\varepsilon_i, \qquad \text{where } \varepsilon_i \stackrel{\text{iid}}{\sim} H. \quad (9)$$

Here, the location and the scale of the response y_i can depend in nonlinear form on covariates $z_{i1}, \ldots, z_{iq}$ and an error term ε_i with distribution function H not depending on covariates. Choosing all f_j and g_j as linear functions yields a linear model: see eSupplement B. If functions f_j and g_j are zero, the associated covariates have no influence on the response. The resulting quantile function has a nonlinear predictor structure and can be written as

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

Based on the additive model in (9), we considered the following two univariable setups:

$q = 1$	β_0	$f_1(z_{i1})$	α_0	$g_1(z_{i1})$
'sin'-setup:	2	$1.5\sin(\frac{2}{3}z_{i1})$	0.5	$1.5z_{i1}^2$
'log'-setup:	2	$1.5\log(z_{i1} + 1.05)$	1	$0.7z_{i1}$

and a multivariable setup with $q = 6$:

β_0	2	α_0	0.5
$f_1(z_{i1})$	$1.5\sin(\frac{2}{3}z_{i1})$	$g_1(z_{i1})$	$0.5z_{i1}^2$
$f_2(z_{i2})$	$1.5\log(z_{i2} + 1.05)$	$g_2(z_{i2})$	$0.5z_{i2}$
$f_3(z_{i3})$	$2z_{i3}$	$g_3(z_{i3})$	$0.5z_{i3}$
$f_4(z_{i4})$	$-2z_{i4}$	$g_4(z_{i4})$	0
$f_5(z_{i5})$	0	$g_5(z_{i5})$	0
$f_6(z_{i6})$	0	$g_6(z_{i6})$	0

In the multivariable setup, two covariates relate nonlinearly to the response, two have a linear influence on it, and the last two have no influence at all.

For generating datasets based on these variable setups, covariates z_i were drawn from a uniform distribution $\mathcal{U}[0, 1]$ with a Toeplitz-structured covariance matrix, leading to $\text{cov}(z_{ik}, z_{il}) = \rho^{|k-l|}$ for possible correlation coefficients $\rho \in \{0, 0.2, 0.5, 0.8\}$. Also, we repeated each setup for different distributions of the error term: a standard normal, a t distribution with two degrees of freedom, and a gamma distribution, where $\mathbb{E}(\varepsilon_i) = \mathbb{V}(\varepsilon_i) = 2$. Figure 1 shows data examples from 'sin'- and 'log'-setups for normal- and gamma-distributed error terms.

For each parameter combination consisting of a specific variable setup, correlation coefficient, and error distribution, we generated three independent datasets: A validation dataset consisting of 200 observations to select optional tuning parameters, a training dataset with 400 observations for model estimation, and a test dataset with 1000 observations to evaluate the performance of each algorithm.

Estimation. We estimated additive quantile regression models for each parameter setup with potential nonlinear effects for all covariates on a fixed quantile grid with $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We used five different estimation algorithms: additive quantile boosting (`gamboost`), total variation regularization (`rqss`, Koenker, Ng, and Portnoy 1994), boosting with stump base learners (`stumps`, Kriegler and Berk 2010), boosting with higher-order tree base learners (`trees`, Kriegler and Berk 2010), and quantile regression forests (`quantregForest`, Meinshausen 2006). In the case of `gamboost`, we used cubic-penalized spline base learners with a second-order difference penalty, 20 inner knots, five degrees of freedom, and fixed the step length at $\nu = 0.1$. The

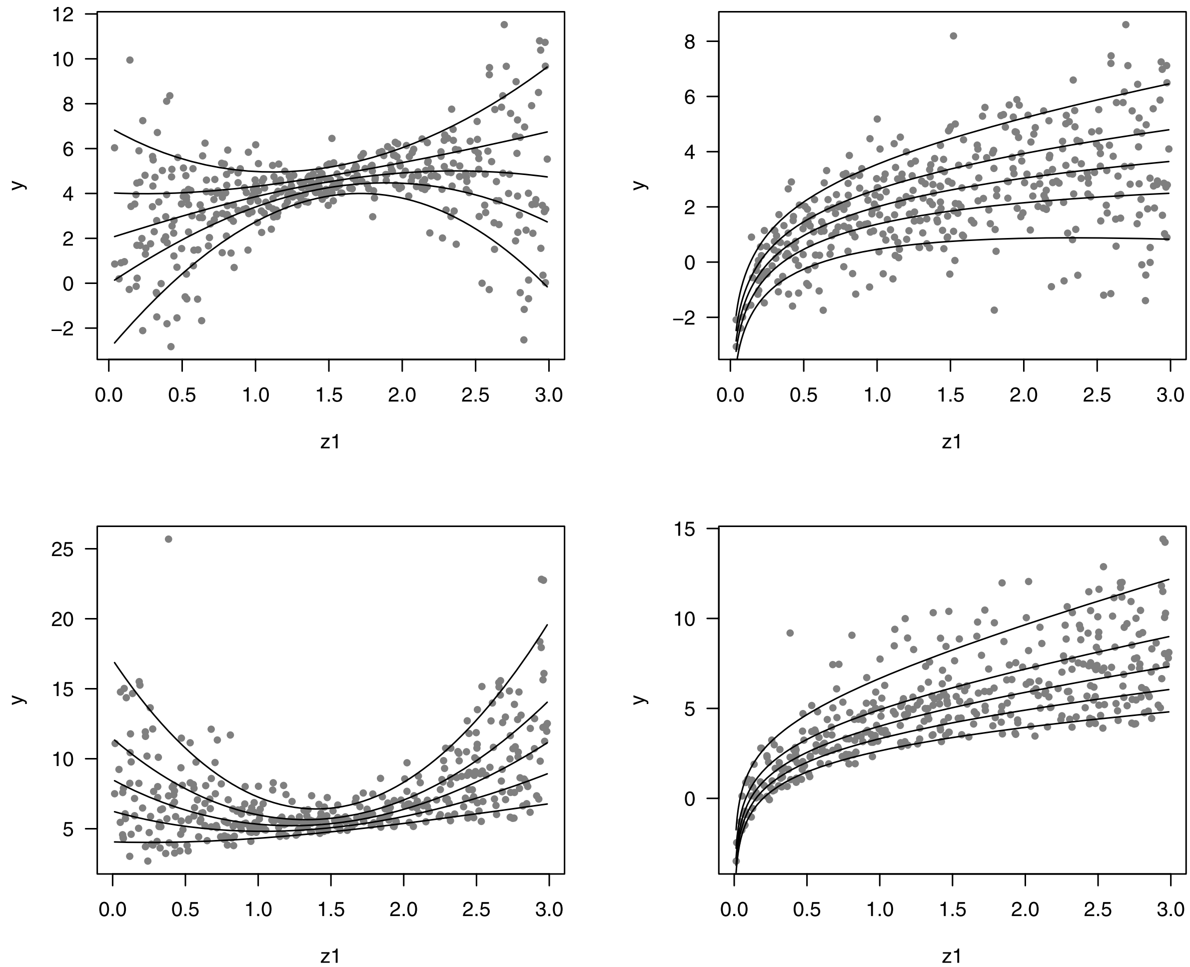

Figure 1. Data examples for nonlinear simulation setups with $n = 400$ data points and one covariate in the 'sin'-setup (left) or 'log'-setup (right) with standard normal-distributed (top) or gamma-distributed (bottom) error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

validation dataset was used to determine the optimal number of iterations m_{stop} for all boosting algorithms as well as covariate-specific smoothing parameters $\lambda_1, \ldots, \lambda_q$, as given in (5), in the case of rqss.

Performance Results. To evaluate the performance results, data generation and estimation was repeated 100 times for each parameter setup and quantile. As performance criteria, we considered empirical risk, bias, and mean-squared error (MSE). We defined the quantile- and iteration-specific empirical risk as

$$\text{Risk}(\tau, k) = \frac{1}{1000} \sum_{i=1}^{1000} \rho_\tau (y_i - \hat{y}_{\tau k i}),$$

where y_i stands for the response of observation i on the test dataset and $\hat{y}_{\tau k i}$ denotes the estimated response value at quantile τ for iteration k, $k = 1, \ldots, 100$, and observation i. Analogously, quantile- and iteration-specific bias and MSE were estimated as

$$\text{Bias}(\tau, k) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{y}_{\tau k i} - y_{\tau i}),$$

$$\text{MSE}(\tau, k) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{y}_{\tau k i} - y_{\tau i})^2,$$

where $y_{\tau i}$ denotes the true underlying τth quantile of the response of observation i.

To illustrate the results for univariate setups, Figure 2 exemplarily displays estimated quantile curves for the 'sin'-setup with standard normal-distributed error terms. Visual inspection reveals hardly any differences between the smooth curves obtained by gamboost and the piecewise linear curves from rqss. This result is also confirmed when the respective performance criteria are compared.

We show representative results for the multivariable setup for the simulation setup with gamma-distributed error terms and a correlation coefficient of $\rho = 0.5$. Other multivariable setups lead to similar results, which can be viewed in eSupplement A. Figure 3 shows quantile- and algorithm-specific empirical distributions of the resulting performance criteria while Table 1 displays the respective means.

In comparison with other algorithms, rqss and gamboost show the best performance results. rqss achieves lowest empirical risk and MSE values whereas gamboost attains the smallest bias for all quantiles. However, the differences between rqss and gamboost results seem to be rather slight. The clear superiority of these algorithms in comparison with the others can be explained by the specific design of our simulation

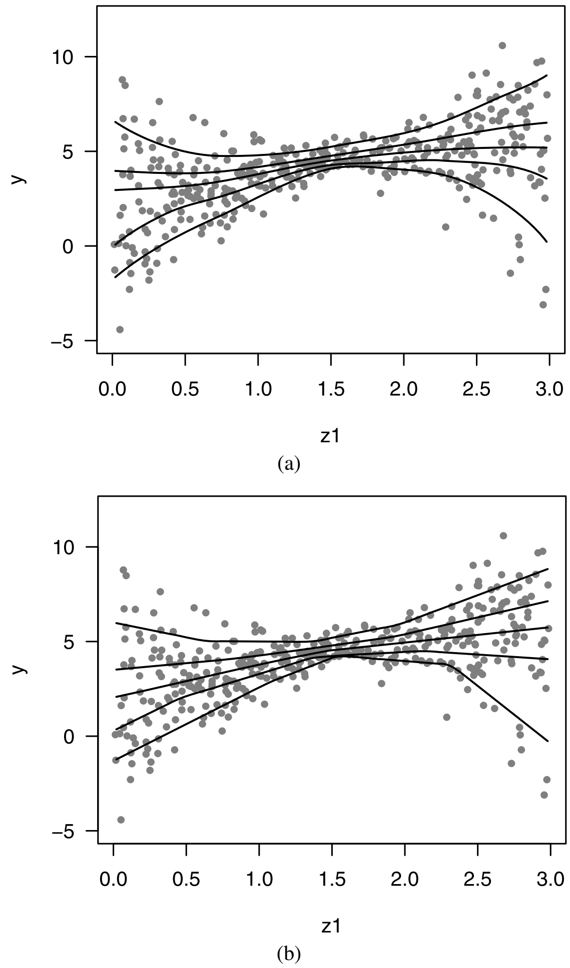

Figure 2. One example for resulting estimated quantile curves for the ‘sin’-setup with standard normal-distributed error terms. Plot (a) displays curves obtained from `gamboost` whereas plot (b) displays curves obtained from `rqss`. True underlying quantile curves are shown in the upper left plot in Figure 1.

study. The underlying data structure of our setup corresponds to an additive model without interaction effects which is also assumed when estimating quantile regression with `rqss` and `gamboost`. Therefore, it is hardly surprising that these methods perform better than `stumps`, `trees`, and `quantregForest` which in turn work as a black boxes and do not assume any specific predictor structure.

In summary, our boosting approach performed on par with the well-established total variation regularization algorithm and clearly outperformed tree-based approaches. We further addressed the special case of linear quantile regression with a separate linear simulation setup described in eSupplement B. In this setup, the results for linear programming and boosting were very similar, which indicates that our boosting approach to linear quantile regression works correctly.

Variable Selection Results. To investigate the performance of the algorithms in higher-dimensional setups, we generated data from a setup with the first four covariate effects being similar to the multivariable setup, but with higher numbers of noninformative covariates, that is, $f_k(z_{ik}) = g_k(z_{ik}) \equiv 0$ for $k = 5, \dots, K$. We considered the three cases $K \in \{6, 16, 20\}$, since the estimation with `rqss` was not possible with more than 20 noninformative covariates.

To exemplify the results, Figure 4 displays boxplots of the performance criteria for the higher-dimensional setup with $K = 20$ noninformative covariates, gamma-distributed error terms and a correlation coefficient of $\rho = 0.5$. We focus on the three algorithms `rqss`, `gamboost` and `stumps` since their results can be interpreted with regards to variable selection — contrary to tree-based approaches just yielding a black box.

The results show that `gamboost` outperforms `rqss` with regard to risk and MSE. The risk difference between `gamboost` and `rqss` is rated as significant at the 5% level by a linear mixed model with risk as response variable, fixed covariate effects for quantile and algorithm, and a random intercept for the datasets $1, \dots, 100$. Regarding also the results for $K = 6$ and $K = 16$ in the same setup, we could observe that absolute risk and MSE differences increase with an increasing number of noninformative covariates. According to previously published results on boosting estimation in high-dimensional setups (see Bühlmann and Yu 2003; Bühlmann 2006; Bühlmann and Hothorn 2007, among others), we expect the advantages of `gamboost` over `rqss` to be even more pronounced if more noninformative covariates are included in the data. However, studying this setup further was rendered impossible since the current implementation of `rqss` could not be used to fit models with more than 25 covariates.

For boosting based algorithms, that is, `gamboost` and `stumps`, we explored the variable selection results in more detail. Regarding the same higher-dimensional setup as above, Figure 5 shows for each base learner the empirical distribution of the first selection iteration relative to the optimized m_{stop} from 100 simulation replications. In the same way, Figure 6 visualizes the proportion of iterations in which a base learner was selected. Base learners of noninformative covariates $z_5, \dots, z_{24}$ are compressed in one category. The figures illustrate that for both algorithms and independent of τ, noninformative covariates are selected less frequent and later for the first time during the estimation process than informative covariates $z_1, \dots, z_4$. These results further substantiate the advantages of boosting with regards to variable selection in high-dimensional setups.

3.2 Comparing Estimated Quantile Functions

As an alternative way to judge the performance of quantile regression models, we now focus on comparing estimated quantile functions directly with the true quantile function instead of comparing the out-of-sample empirical risk. For one single covariate $Z \sim U(0, 1)$ we sampled response values from the conditional distribution

$$Y|Z = z \sim \text{BCPE}\big(\mu = \sin(2\pi z) + 3, \sigma = \exp(z^2 + 0.1), \nu = 3z, \varphi = \exp(3z + 2)\big),$$

where BCPE refers to a Box–Cox Power Exponential distribution (Rigby and Stasinopoulos 2004). Here, μ controls the median, σ the coefficient of variation, ν the skewness, and φ the kurtosis of the response’ distribution. Thus, the first four moments vary with covariate Z in a smooth nonlinear way (the conditional density is depicted in Figure 14 of eSupplement A). Based on a learning sample containing $n = 200$ observations

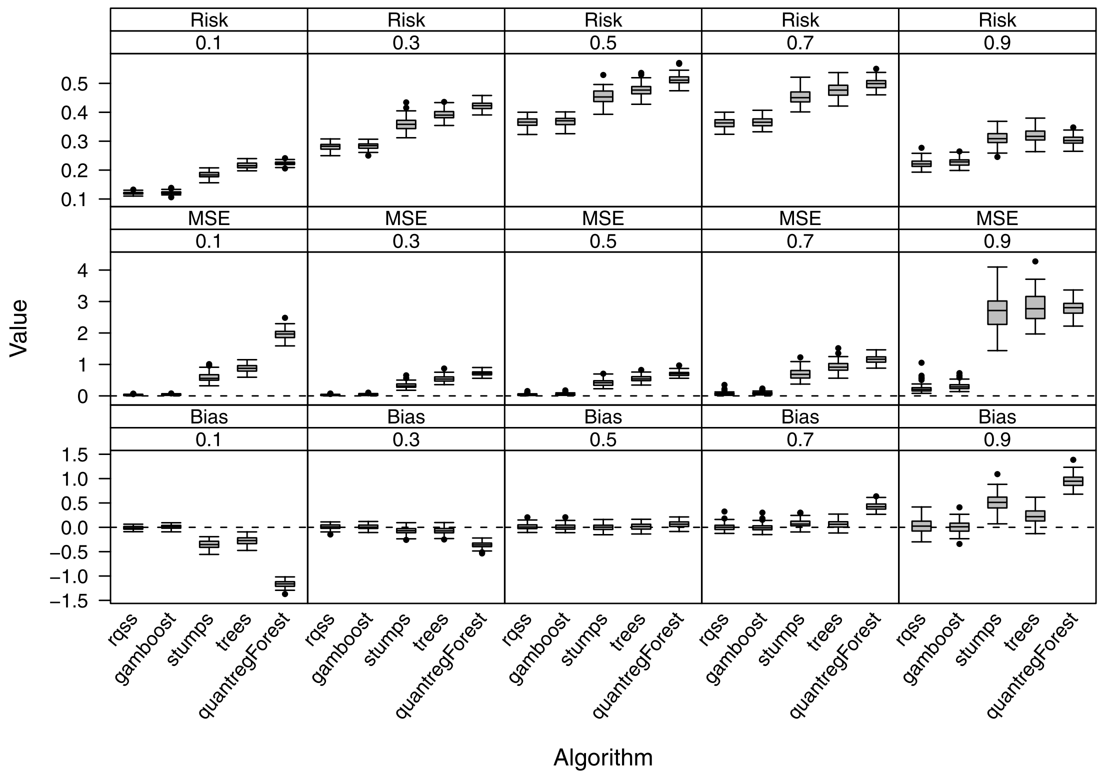

Figure 3. Simulation results for the multivariable setup with gamma-distributed error terms and a correlation coefficient of 0.5. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

and one additional validation sample of $n = 100$ both `rqss` and `gamboost` were fitted (including hyperparameter tuning). For each of 100 simulated datasets, we compared the estimated quantile functions $\hat{Q}_Y(\tau|z) = \hat{f}_\tau(z)$ for `gamboost` and `rqss` with the true quantile function $Q_Y(\tau|z) = f_\tau(z)$ obtained from the BCPE distribution for $\tau \in \{0.90, 0.91, \ldots, 0.99\}$ and 100 equidistant z values $\in [0.1, 0.9]$. As a means for comparing the estimated and true quantile functions we chose the sum of the absolute differences $\sum_\tau \sum_z |\hat{f}_\tau(z) - f_\tau(z)|$ which corresponds to the absolute deviations from the bisecting line in a quantile-quantile plot.

Figure 7 clearly shows that the quantile functions estimated by `gamboost` resemble the true quantile function better than `rqss`. Of course, this is due to the ability of `gamboost`to adapt to the smoothness of the nonlinear effects. In contrast, `rqss` has to approximate a smooth curve by piecewise linear splines. The nature of this phenomenon is illustrated in Figure 13 of eSupplement A.

Table 1. Mean estimated performance criteria from 100 replications from the multivariable setup with gamma-distributed error terms and a correlation coefficient of 0.5

		Quantile				
Criterion	Algorithm	0.1	0.3	0.5	0.7	0.9
Risk	`rqss`	0.120	0.281	0.366	0.363	0.222
	`gamboost`	0.122	0.283	0.368	0.366	0.228
	`stumps`	0.183	0.360	0.454	0.453	0.311
	`trees`	0.216	0.391	0.477	0.478	0.320
	`quantregForest`	0.223	0.422	0.513	0.499	0.303
Bias	`rqss`	−0.012	0.010	0.012	0.007	0.025
	`gamboost`	0.010	0.009	0.003	−0.003	0.002
	`stumps`	−0.355	−0.072	0.000	0.078	0.506
	`trees`	−0.275	−0.070	0.013	0.054	0.226
	`quantregForest`	−1.162	−0.361	0.064	0.429	0.953
MSE	`rqss`	0.030	0.027	0.040	0.072	0.229
	`gamboost`	0.040	0.039	0.053	0.090	0.306
	`stumps`	0.586	0.330	0.413	0.698	2.707
	`trees`	0.879	0.538	0.552	0.927	2.811
	`quantregForest`	1.961	0.720	0.703	1.159	2.782

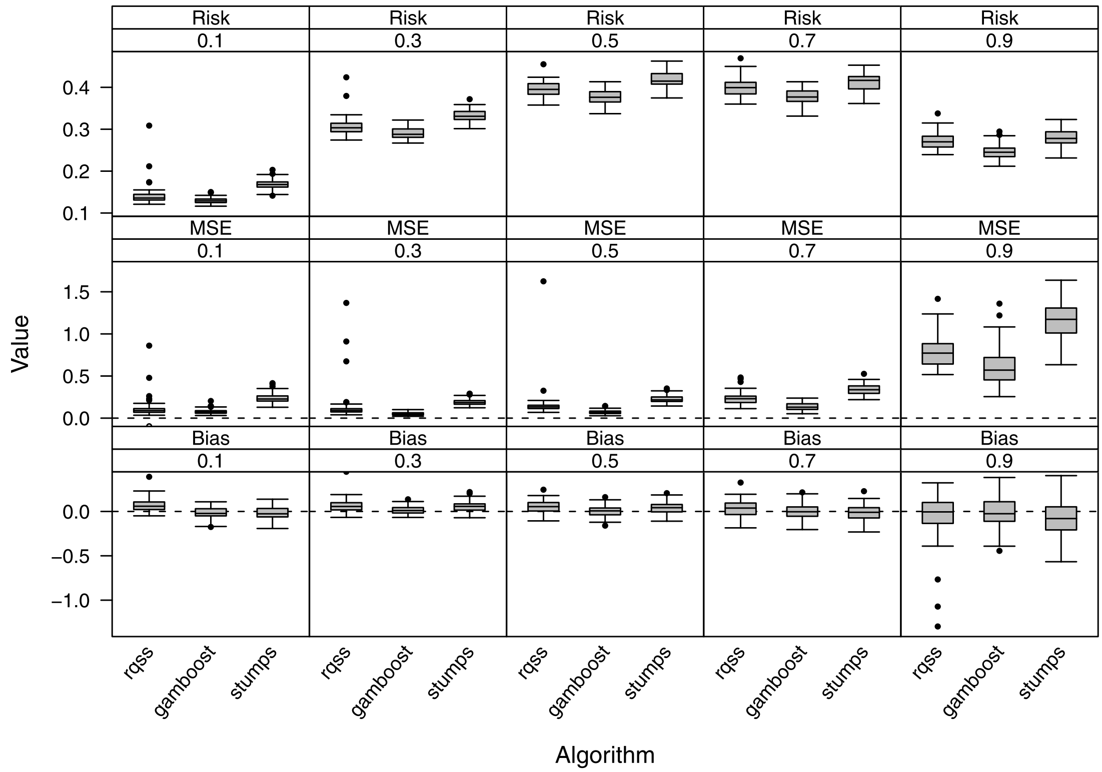

Figure 4. Simulation results for the higher-dimensional setup with $K = 20$ noninformative covariates, gamma-distributed error terms and a correlation coefficient of 0.5. Boxplots display empirical distributions of the performance criteria from 100 replications, depending on quantile τ and three estimation algorithms.

4. ANALYZING CHILDHOOD MALNUTRITION IN INDIA

Childhood malnutrition is one of the most urgent problems in developing and transition countries. To provide information not only on the nutritional status but also on health and population trends in general, Demographic and Health Surveys (DHS) conduct nationally representative surveys on fertility, family planning, maternal and child health, as well as child survival,

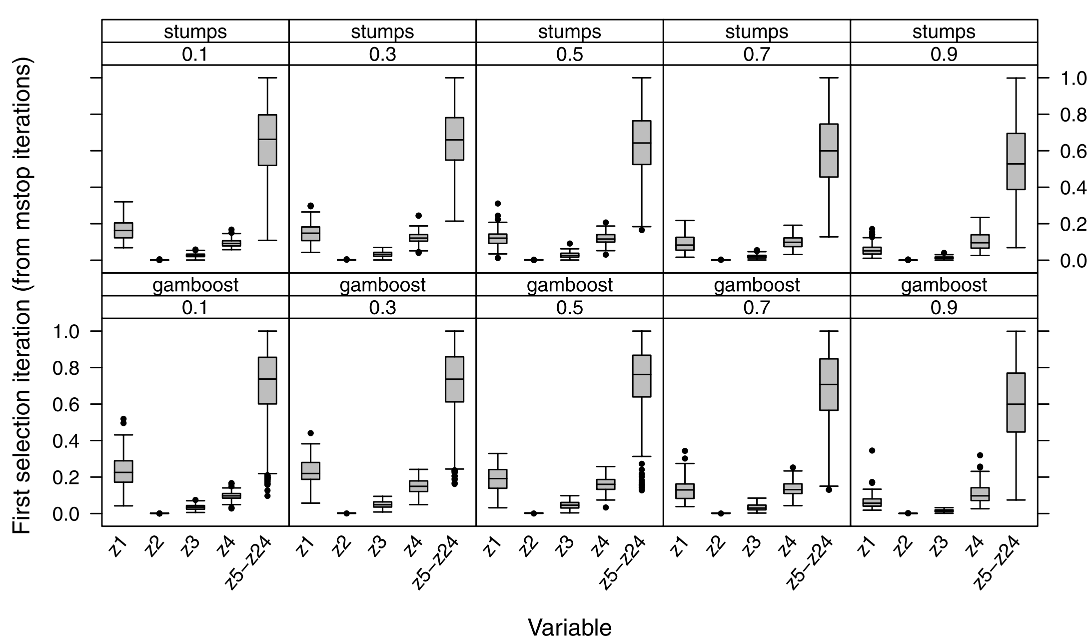

Figure 5. Simulation results for the higher-dimensional setup with $K = 20$ noninformative covariates, gamma-distributed error terms and a correlation coefficient of 0.5. Boxplots display for each base learner $z_1, \ldots, z_{24}$ the empirical distribution of the first selection iteration relative to the optimized m_{stop} from 100 simulation replications, depending on quantile τ and estimation algorithm.

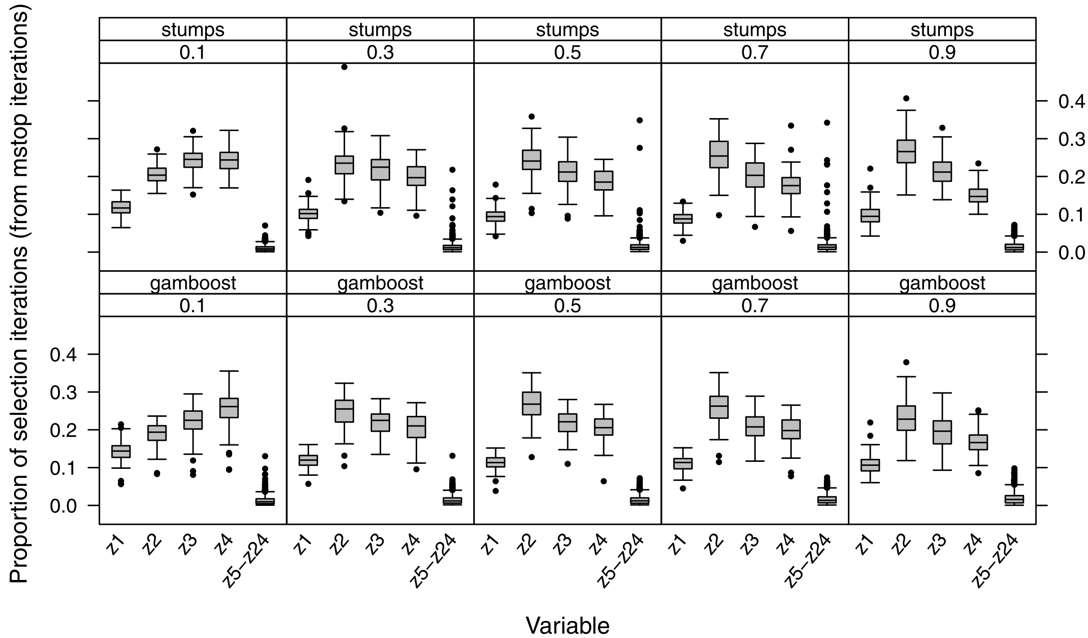

Figure 6. Simulation results for the higher-dimensional setup with $K = 20$ noninformative covariates, gamma-distributed error terms and a correlation coefficient of 0.5. Boxplots display for each base learner $z_1, \ldots, z_{24}$ the empirical distribution of the proportion of selection iterations relative to the optimized m_{stop} from 100 simulation replications, depending on quantile τ and estimation algorithm.

HIV/AIDS, malaria, and nutrition. The resulting data—from more than 200 surveys in 75 countries so far—are available for research purposes at *www.measuredhs.com*.

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

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

where AC denotes the anthropometric characteristic of interest and m and s correspond to median and (a robust estimate

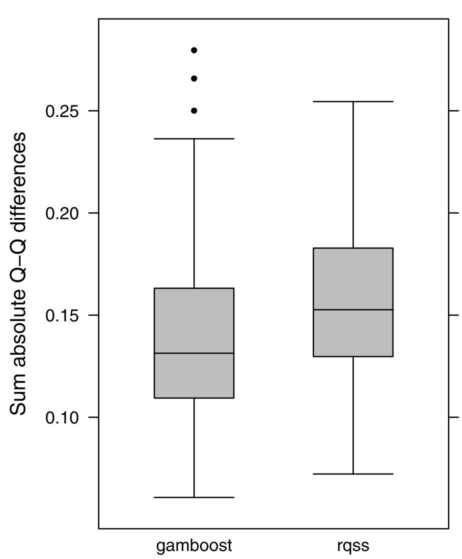

Figure 7. Sum of absolute differences between estimated (by `gamboost` and `rqss`) and true quantile function corresponding to the BCPE distribution over a grid of τ and z values based on 100 simulation replications.

for the) 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, that is, insufficient height for age, in the following. Stunting provides a measure of chronic malnutrition, whereas insufficient weight for age might result from either acute or chronic malnutrition. Note that the Z score, despite its name, is not assumed to be normal. 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.

Previous analyses, like the ones presented by Kandala et al. (2001, 2009) have focused on modeling the expectation of the malnutrition score, yielding regression results for the average nutritional status. However, if we are interested in severe malnutrition, regression models for quantiles such as the 5% or the 10% quantile, may be much more interesting and informative.

In the following, we will present an analysis on childhood malnutrition in India based on DHS data from 2005/2006. From the original dataset, we extracted a number of covariates that are deemed to be important determinants of childhood malnutrition, see Table 2 for an overview and short descriptions.

Based on this set of covariates, we specified five different candidate models. In the first additive model estimated by our boosting approach, all continuous covariates are included with possibly nonlinear effects based on cubic-penalized spline base learners with a second-order difference penalty and 20 inner knots. The nonlinear effects are centered around the linear effects as in (8) to allow linear and nonlinear effects to be separated. The nonlinear deviation effect is assigned one degree of freedom to determine an appropriate smoothing parameter for the base learner. All categorical covariates were assigned least-squares base learners, where dummies corresponding to different levels of the same covariate are combined into one single

Table 2. Variables in the childhood malnutrition dataset

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	Position 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)

base learner. This yields the quantile-specific model equation

$$\begin{aligned} Z_i = {} & (\text{cage}, \text{cfeed}, \text{csex}, \ldots, \text{car})_i^\top \boldsymbol{\beta}_\tau \\ & + f_{\tau 1}(\text{cage}_i) + f_{\tau 2}(\text{cfeed}_i) + f_{\tau 3}(\text{mbmi}_i) \\ & + f_{\tau 4}(\text{mage}_i) + f_{\tau 5}(\text{medu}_i) + f_{\tau 6}(\text{medupart}_i) \\ & + \varepsilon_{\tau i}. \end{aligned} \tag{10}$$

As an extension of Model (10), we considered the same type of model but with all effects being gender specific. For the parametric effects, this yields usual interaction terms while for the nonlinear terms, we end up with varying coefficient terms with gender as interaction variable. This varying coefficient model (VCM) was also estimated by our boosting approach. As benchmark for the predictive performance of our models, we further considered a simple additive model similar to (10) only including nonlinear effects for the continuous covariates. This allowed us to use total variation regularization for model estimation since with the currently available software for this approach neither VCMs nor separation between linear and nonlinear contributions of covariate effects are feasible. In addition, we estimated models based on boosting trees and boosting stumps since these allow for a very flexible model structure. On the other hand, interpretation of covariate effects is more difficult in the tree-based models while additive models yield a structured model fit. Comparing the predictive ability between tree-based models and additive models will allow us to check whether the simplified model structure imposed in additive models deteriorates or improves predictions.

We considered three different quantiles, namely 5%, 10%, and 50%, to compare effects on severe chronic malnutrition as well as effects on its average, measured in terms of the median. After plausibility checks and deletion of observations with missing values, we obtained a dataset with 37,623 observations. This dataset was randomly split into three parts: one third was used for estimation, one third was employed as a test sample to determine the tuning parameters (the stopping iteration m_{stop} in case of boosting algorithms and the nine smoothing parameters $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$ in case of total variation regularization), and one third served as an evaluation sample for the predictive performance. To allow for a proper uncertainty assessment in the results, 50 different splits were performed on the data.

Figure 8 displays the cross-validated empirical risk for the 50 samples obtained from the evaluation parts of the splits; see eSupplement D on how this risk was determined exactly. For most of the splits, we found that the simple additive model provides the best fit while the fit deteriorates considerably when gender-specific effects are included in the varying coefficients models. The models estimated with total variation regularization are also associated with good predictive performance, although the tuned smoothing parameters and therefore the empirical risk are very sensitive to the starting values (see Figures 30 and 31 from eSupplement D for an illustration). The fact that additive models seem to outperform tree-based models is an indication contradicting strong interaction effects in the data.

Figure 9 visualizes the boosting estimates for the nonlinear effects obtained in the additive model and a selected set of covariates. The effect of the child's age was estimated quite consistently across the 50 samples for any quantile. Specifically, uncertainty in the age effect seems to be mostly associated with the level but not with the functional form of the effect. When we compare the results for the 5% quantile and the median, we found that the age effect approximates a U-shape for severe malnutrition whereas the median steadily decreases until an age of about 20 months, followed by an almost constant age effect. This indicates a moderate improvement in the stunting score after a certain age when severe malnutrition is considered and no such improvement in the average level. All other effects

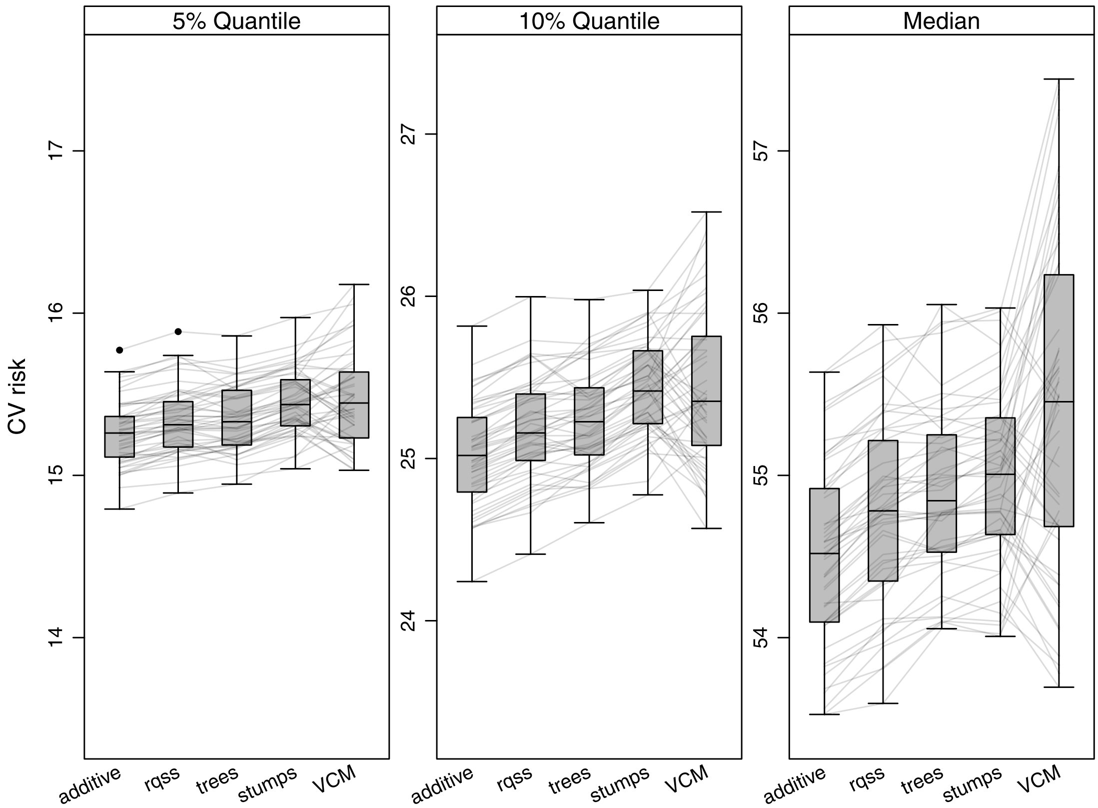

Figure 8. Boxplots display empirical distributions of the cross-validated empirical risks for the evaluation parts of the 50 data splits. Results for one split are connected by gray lines.

are associated with a much stronger uncertainty and less clear patterns across the 50 samples. For example, the effect of the duration of breastfeeding considerably varies in the functional form, in particular for the median. Here we also find a moderate decrease in the stunting score, while there seems to be almost no effect on the 5% quantile. Similarly, the age of the mother has a somewhat nonlinear effect on the median, but is closer to a linear effect for the 5% quantile.

Results for some selected effects of categorical covariates obtained in the additive model are shown in Figure 10. We found a clear indication, consistent across all quantiles, that a better nutritional status is associated with richer families. Similarly, a better nutritional status is associated with single births as compared to twin births, which seems plausible since in a twin birth two children are competing for fixed resources. Note that the twin effect is actually much stronger for the median, where the estimated effects are quite expressed in every replication, which leads to almost separated levels for twin and single births. For the gender effect, we found a moderate preference for female children, which was more prominent for the 5% quantile. Complete results for all covariate effects obtained in the additive boosting approach can be found in eSupplement C.

For illustrative purposes, we also show one selected effect from the varying coefficient model, although we have found that it does not have the same predictive ability as the additive model. Figure 11 shows the gender-specific age effect, and reveals qualitatively similar differences between the 5% quantile and the median for both boys and girls. However, some interesting differences become apparent when one looks more closely at the precise shape of the functions. For the 5% quantile, the improvement in the stunting score after and at the age of about 30 months is less prominent for girls. For boys, in contrast, the age effect on the 5% quantile is almost symmetric about 30 months. Similarly, for boys, there is a moderate improvement after an age of 20 months in the median effect for boys, but this effect is much less expressed, or even absent, for girls.

Finally, we addressed the inference problem in the additive model more formally based on the stability selection procedure proposed by Meinshausen and Bühlmann (2010), which allows the family-wise error rate in model choice to be controlled similarly as usual significance tests control the Type I error. The basic idea is to fix an average number of terms to be expected in a model and an upper bound for the family-wise error rate. The stability selection procedure then provides a threshold for the relative frequency of a model term to be among the first terms selected in the boosting procedure. We chose an average number of ten terms and a family-wise error rate of 5% in the following, but the results were almost insensitive with respect to changes in the expected number of terms. For an additive model and the 5% quantile, a nonlinear effect of the child's age and linear effects of the birth order, the wealth indicator, and the educational level of both the mother and the mother's partner had a "significant" impact on the nutritional status. For the additive median model, also linear effects of the duration of breastfeeding and the body mass index of the mother were found. When varying coefficient models were considered, exactly the same effects for the 5% quantile were found as with the additive model. In particular, there was no gender-specific

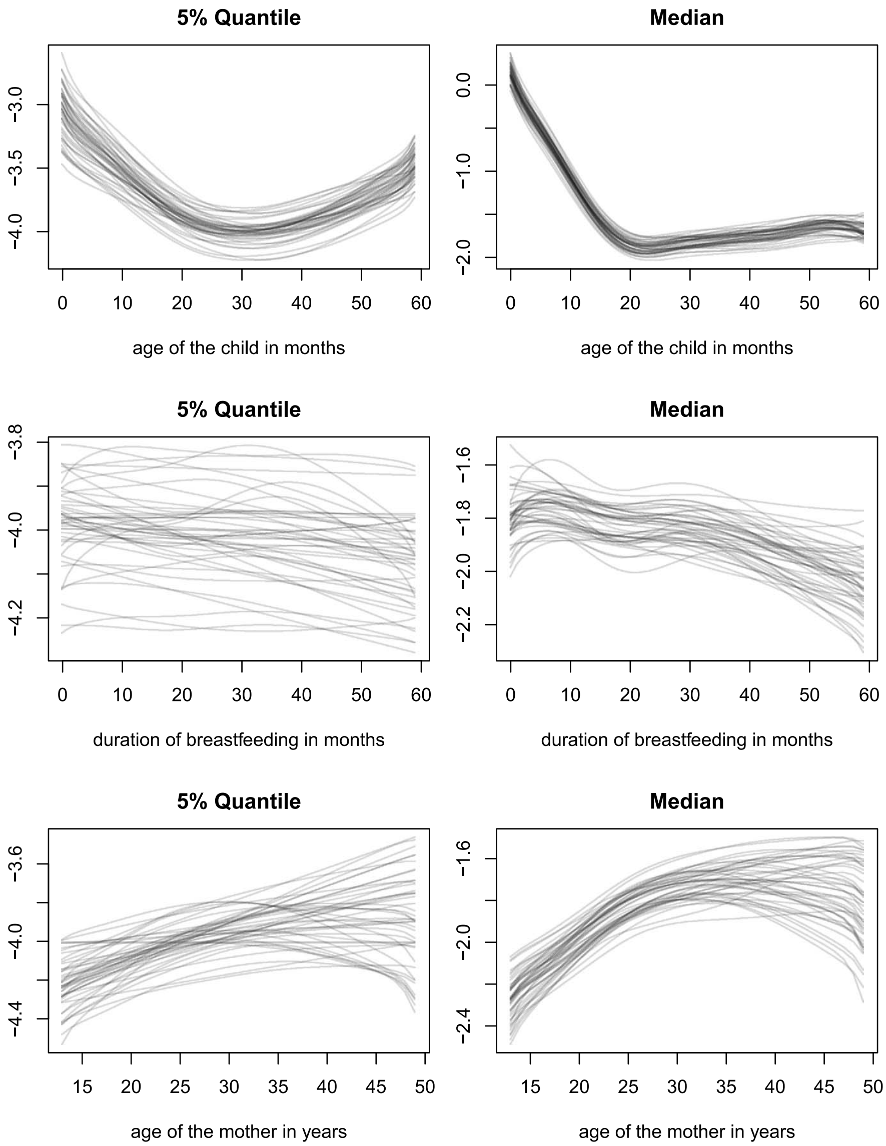

Figure 9. Selected nonlinear effects from the additive model. Gray lines show results from the 50 samples. All effects are adjusted for the overall quantile levels.

effect, which provides further evidence for the better predictive performance of the additive model. For the median, the effect of the birth order drops from the model when varying coefficients were considered, but again no gender-specific effect appears to be significant according to stability selection. When the number of expected effects was increased, a linear age-by-gender interaction was included both for the median and the 5% quantile in the varying coefficient model, which provides an indication for a moderate difference in age effects. In addition, the bicycle indicator appears in the additive 5% quantile model as one of the asset-related variables.

Removing variables from the additive model estimated by total variation regularization would require a formal variable selection procedure, for example, based on AIC-type or SIC-type criteria—as shortly discussed in Section 2.1. We refrained from a further variable selection step for this model since the currently available implementation of `rqss` does not provide an automatic procedure for variable selection.

5. DISCUSSION

Motivated by the analysis of risk factors for childhood malnutrition in India, we developed a novel boosting algorithm for estimation in additive quantile regression models. Our investigation was based on data collected in the 2005/2006 DHS and contained numerous covariates as well as a score for chronic malnutrition serving as the response. By using lower quantiles instead of just the mean or median as the regression objective, that is, by using quantile regression, it was possible to identify risk factors for severe malnutrition and not only for the population average of the score. 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, in particular in combination with the novel stability selection approach. The most important covariate in all models was the age of the child, which also followed a clear nonlinear pattern. The precise form of this age-specific variation, however, was quite different for the median and extreme quan-

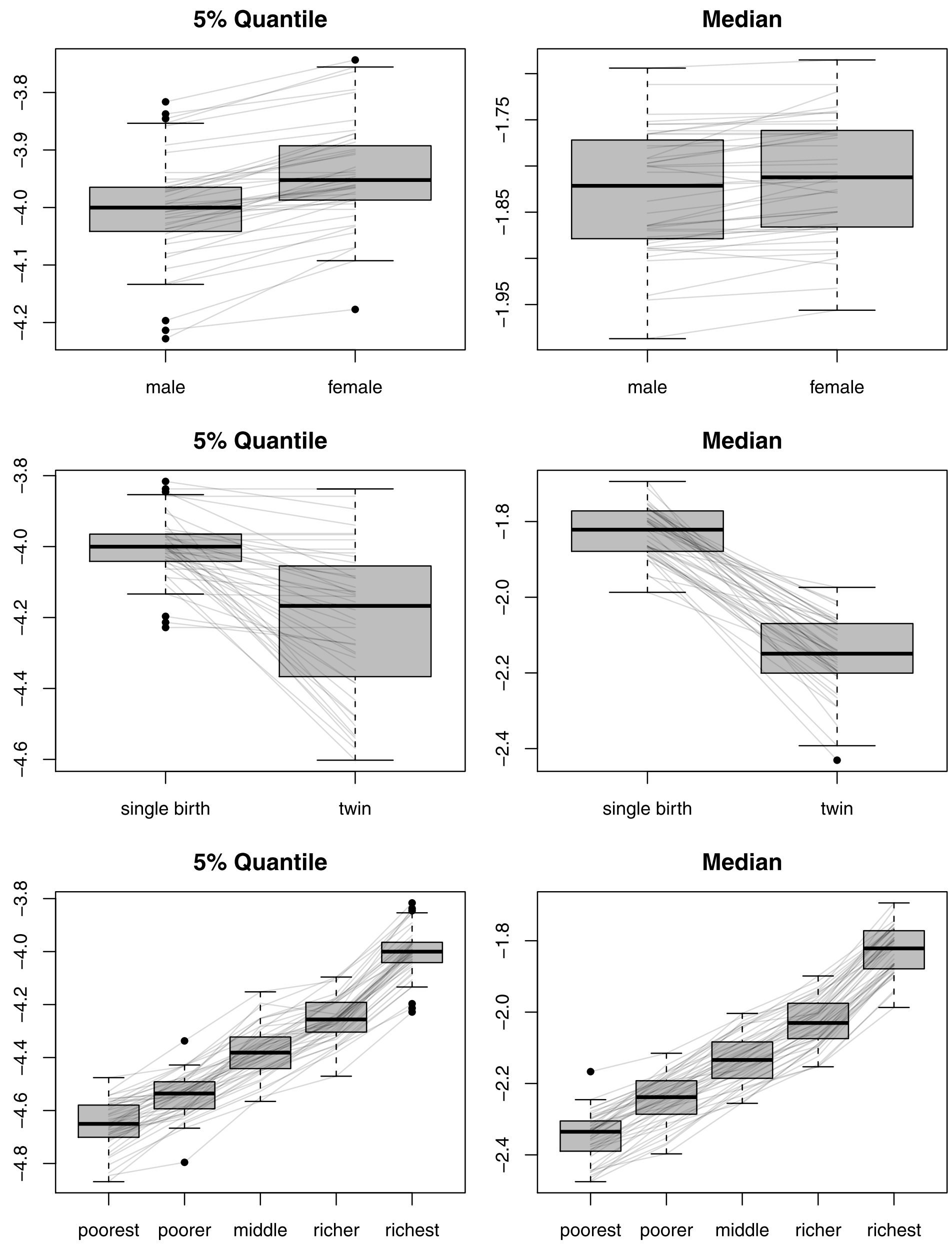

Figure 10. Selected estimated effects of categorical covariates from the additive model. Gray lines show results from the 50 samples. All effects are adjusted for the overall quantile levels.

tiles. Other important covariates identified by stability selection were related to the education of the mother and the partner of the mother, the position of the child in the birth order and the general economic situation of the family.

Compared to total variation regularization, the boosting estimation for additive quantile regression offers the following 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, considering the currently available software for both algorithms, boosting can handle a much larger number of nonlinear covariate effects. Third, parameter estimation and variable selection are executed in one single estimation step, which is particularly favorable for higher-dimensional predictors (as illustrated by the empirical evaluation). Both estimation algorithms require the specification of tuning parameters. In the case of boosting, only the optimal number of iterations m_{stop} has to be tuned, whereas in the case of total variation regularization, the q smoothing parameters $\lambda_1, \ldots, \lambda_q$ have to be chosen carefully. Furthermore, the model choice procedure inherent in our approach helps to differentiate between simple linear effects and more complicated nonlinear effects. A simulation experiment comparing total variation regularization and boosting for additive quantile regression showed that both methods perform on par in situations where the classical procedure was shown to be appropriate. In a higher-dimensional simulation setup and for the India malnutrition analysis, however, our boosting approach outperformed all other methods including total variation regularization.

Meinshausen (2006) and Kriegler and Berk (2010) adapted the popular random forest and tree-based boosting algorithm for the estimation of quantile regression. Both approaches also

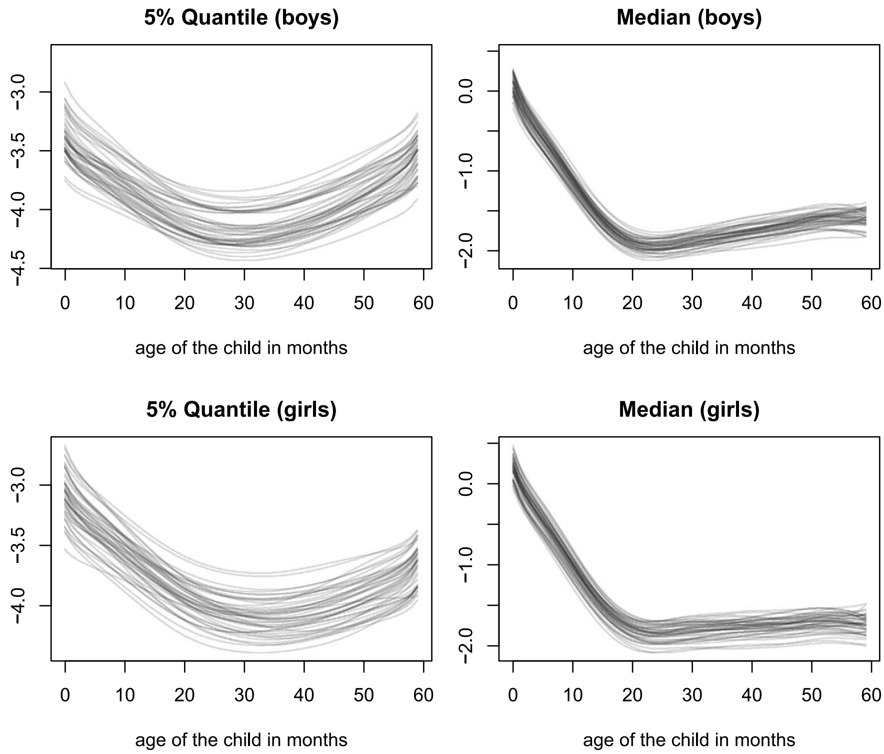

Figure 11. Estimated gender-specific age effects from the varying coefficient model. Gray lines show results from the 50 samples. All effects are adjusted for the overall quantile levels.

minimize the empirical risk defined by the check function. For these machine-learning approaches, the regression predictor is implicitly defined as a complex sum of trees and is therefore difficult to interpret. However, the models for childhood malnutrition considered in this article are based on a well-defined additive quantile regression model, where the effects of each covariate have a clear statistical interpretation. When the true data-generating process follows such a "simple" model, the machine-learning approaches are clearly outperformed by additive quantile regression models (fitted via total variation regularization or boosting), as shown in Section 3.1.

Extensions of the boosting algorithm for random and spatial effects are feasible if these effects are included in the predictors of the base learners (see Section 2.2). Kriegler and Berk (2010) analyzed quantiles of the number of homeless people in Los Angeles. In this application, observations are clearly spatially structured, and it would be interesting to assess the amount of unobserved heterogeneity by including a spatial model term (either a bivariate surface or a model term taking the neighborhood structure into account; see Kneib, Hothorn, and Tutz 2009). In contrast to unstructured regression trees as base learners, such an assessment is easily possible within our proposed framework. 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 models accounting for the longitudinal and spatial structure of childhood malnutrition measurements. In summary, quantile regression models including varying coefficients, random, time-varying, and spatial effects or otherwise penalized model terms, can be fitted using the approach and software presented in this article.

Apart from malnutrition, quantile modeling is also of interest in applications where the quantiles depend on covariates differently than the mean does, 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.

COMPUTATIONAL DETAILS

For the implementation of our boosting algorithm, we slightly extended already available standard software for boosting. The described methodology is implemented in the R add-on package **mboost** (Hothorn et al. 2011, 2010). 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()`, which allows τ and an offset quantile to be specified.

Other approaches for estimating additive quantile regression that were used here are available in the following functions in R (R Development Core Team 2010): Estimation by total variation regularization is possible by using the function `rqss()` from package **quantreg** (Koenker 2010b). Boosting of stumps and regression trees as described in Kriegler and Berk (2010) is available via the function `blackboost()` from package

mboost. Quantile regression forests can be estimated with the function `quantregForest()` from package **quantregForest** (Meinshausen 2007).

To assure the reproducibility of the results of our data analyses, we include an electronic supplement to this article that contains all necessary R commands to prepare and analyze the Indian malnutrition data (provided that the original dataset was obtained from *www.measuredhs.com*). If package **mboost** (version 2.0-10 or higher) is installed, simply type

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

to reproduce our analyses.

SUPPLEMENTARY MATERIALS

Results: We provide additional results from simulation experiments and data analyses, and the complete source code of the simulation experiments. (eSupplement.pdf)

[Received April 2009. Revised February 2011.]

REFERENCES

Ackerson, L. K., and Subramanian, S. (2008), "Domestic Violence and Chronic Malnutrition Among Women and Children in India," *American Journal of Epidemiology*, 167, 1188–1196. [494]

Bharati, S., Pal, M., and Bharati, P. (2008), "Determinants of Nutritional Status of Pre-School Children in India," *Journal of Biosocial Science*, 40, 801–814. [494]

Buchinsky, M. (1998), "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *The Journal of Human Resources*, 33, 88–126. [495]

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34, 559–583. [499,501]

Bühlmann, P., and Hothorn, T. (2007), "Boosting Algorithms: Regularization, Prediction and Model Fitting" (with discussion), *Statistical Science*, 22, 477–505. [495,497,501]

Bühlmann, P., and Yu, B. (2003), "Boosting With the L_2 Loss: Regression and Classification," *Journal of the American Statistical Association*, 98, 324–339. [495,498,499,501]

Cai, Z., and Xu, X. (2008), "Nonparametric Quantile Estimations for Dynamic Smooth Coefficient Models," *Journal of the American Statistical Association*, 103, 1595–1608. [509]

Caulfield, L., de Onis, M., Blossner, M., and Black, R. (2004), "Undernutrition as an Underlying Cause of Child Deaths Associated With Diarrhea, Pneumonia, Malaria and Measles," *American Journal of Clinical Nutrition*, 80, 193–198. [494]

Eilers, P., and Marx, B. (1996), "Flexible Smoothing With B–Splines and Penalties," *Statistical Science*, 11, 89–121. [498]

Fahrmeir, L., Kneib, T., and Lang, S. (2004), "Penalized Structured Additive Regression for Space–Time Data: A Bayesian Perspective," *Statistica Sinica*, 14, 731–761. [496,498]

Fenske, N., Fahrmeir, L., Rzehak, P., and Höhle, M. (2008), "Detection of Risk Factors for Obesity in Early Childhood With Quantile Regression Methods for Longitudinal Data," Technical Report 38, Ludwig-Maximilians-Universität München. Available at *http://epub.ub.uni-muenchen.de/6260/*. [494]

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [497]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting" (with discussion), *The Annals of Statistics*, 28, 337–407. [495]

Horowitz, J., and Lee, S. (2005), "Nonparametric Estimation of an Additive Quantile Regression Model," *Journal of the American Statistical Association*, 100, 1238–1249. [496]

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010), "Model-Based Boosting 2.0," *Journal of Machine Learning Research*, 11, 1851–1855. [509]

——— (2011), "mboost: Model-Based Boosting," R package version 2.0-10, available at *http://CRAN.R-project.org/package=mboost*. [509]

Kandala, N.-B., Fahrmeir, L., Klasen, S., and Priebe, J. (2009), "Geo-Additive Models of Childhood Undernutrition in Three Sub-Saharan African Countries," *Population, Space and Place*, 15, 461–473. [494,495,504]

Kandala, N. B., Lang, S., Klasen, S., and Fahrmeir, L. (2001), "Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of Undernutrition in Two African Countries," *Research in Official Statistics*, 4, 81–99. [494,495,504]

Kneib, T., Hothorn, T., and Tutz, G. (2009), "Variable Selection and Model Choice in Geoadditive Regression Models," *Biometrics*, 65, 626–634. [495,496,498,509]

Koenker, R. (2005), *Quantile Regression. Economic Society Monographs*, New York: Cambridge University Press. [495-497]

——— (2010a), "Additive Models for Quantile Regression: An Analysis of Risk Factors for Malnutrition in India," in *Advances in Social Science Research Using R*, ed. H. D. Vinod, New York: Springer-Verlag. [497]

——— (2010b), "quantreg: Quantile Regression," R package version 4.50, available at *http://CRAN.R-project.org/package=quantreg*. [496,509]

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680. [496,497,499]

Kriegler, B., and Berk, R. (2010), "Small Area Estimation of the Homeless in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting," *Annals of Applied Statistics*, 4, 1234–1255. [495,499,508,509]

Li, Y., and Zhu, J. (2008), "L_1-Norm Quantile Regression," *Journal of Computational and Graphical Statistics*, 17, 163–185. [496]

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile Regression in Reproducing Kernel Hilbert Spaces," *Journal of the American Statistical Association*, 102, 255–268. [496]

Meinshausen, N. (2006), "Quantile Regression Forests," *Journal of Machine Learning Research*, 7, 983–999. [495,498,499,508]

——— (2007), "quantregForest: Quantile Regression Forests," R package version 0.2-2, available at *http://CRAN.R-project.org/package=quantregForest*. [510]

Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection," *Journal of the Royal Statistical Society, Ser. B*, 72, 417–473. [495,499,506]

Mishra, V., and Retherford, R. (2007), "Does Biofuel Smoke Contribute to Anaemia and Stunting in Early Childhood?" *International Journal of Epidemiology*, 36, 117–129. [494]

Nair, K. (2007), "Malnourishment Among Children in India: A Regional Analysis," *Economic and Political Weekly*, 42, available at *http://epw.in*. [494]

NFHS (2007), "National Family Health Survey (NFHS-3), 2005-06: India," International Institute for Population Sciences (IIPS) and Macro International Inc. Calverton, I, available at *http://www.measuredhs.com/pubs*. [494]

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org*, ISBN 3-900051-07-0. [496,509]

Rajaram, S., Zottarelli, L., and Sunil, T. (2007), "Individual, Household, Programme and Community Effects on Childhood Malnutrition in Rural India," *Maternal and Child Nutrition*, 3, 129–140. [494]

Rigby, R. A., and Stasinopoulos, D. M. (2004), "Smooth Centile Curves for Skew and Kurtotic Data Modelled Using the Box–Cox Power Exponential Distribution," *Statistics in Medicine*, 23, 3053–3076. [501]

——— (2005), "Generalized Additive Models for Location, Scale and Shape," *Applied Statistics*, 54, 507–554. [496]

Schmid, M., and Hothorn, T. (2008), "Boosting Additive Models Using Component-Wise P–Splines as Base-Learners," *Computational Statistics & Data Analysis*, 53, 298–311. [498]

Schnabel, S., and Eilers, P. (2009), "Optimal Expectile Smoothing," *Computational Statistics & Data Analysis*, 53, 4168–4177. [499]

Som, S., Pal, M., and Bharati, P. (2007), "Role of Individual and Household Level Factors on Stunting: A Comparative Study in Three Indian States," *Annals of Epidemiology*, 34, 632–646. [494]

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006), "Nonparametric Quantile Estimation," *Journal of Machine Learning Research*, 7, 1231–1264. [496]

Tarozzi, A., and Mahajan, A. (2007), "Child Nutrition in India in the Nineties," *Economic Development and Cultural Change*, 55, 441–486. [494]

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

Wang, H., and Leng, C. (2007), "Unified LASSO Estimation by Least Squares Approximation," *Journal of the American Statistical Association*, 102, 1039–1048. [496]

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, 397–409. [509]

Yu, K., Lu, Z., and Stander, J. (2003), "Quantile Regression: Applications and Current Research Areas," *The Statistician*, 52, 331–350. [509]

Zhang, T., and Yu, B. (2005), "Boosting With Early Stopping: Convergence and Consistency," *The Annals of Statistics*, 33, 1538–1579. [499]