Structured additive quantile regression with applications to modelling undernutrition and obesity of children

Nora Fenske

München 2012

Structured additive quantile regression with applications to modelling undernutrition and obesity of children

Nora Fenske

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Nora Fenske
am 17.09.2012
in München

Erstgutachter:	Prof. Dr. Ludwig Fahrmeir
Zweitgutachter:	Prof. Dr. Torsten Hothorn
Externer Gutachter:	Prof. Dr. Thomas Kneib
Tag der Disputation:	30.10.2012

Danksagung

Bei Ludwig Fahrmeir möchte ich mich von Herzen bedanken für seine Förderung und sein Vertrauen. Es bedeutete für mich ein großes Glück, zwei Jahre in seinem Team arbeiten zu dürfen. Sowohl seine vielen exzellenten Bücher als auch die persönliche Zusammenarbeit mit ihm haben meinen Blick auf die Statistik und auf wissenschaftliches Arbeiten maßgeblich geprägt.

Genauso herzlich möchte ich mich bei Torsten Hothorn bedanken für sein kontinuierliches Coaching, seine unzähligen motivierenden Worte und seine blitzschnellen Antworten auf alle Fragen. Von ihm lernte ich nicht nur zielführende, manchmal auch unkonventionelle Strategien für die Paper-Optimierung, sondern auch viele Dinge über Statistik und R in gemeinsamen Lehrveranstaltungen.

Thomas Kneib danke ich für seine wertvolle Unterstützung und für die Einladung an die Universität nach Oldenburg, wo ich eine produktive und sehr schöne Woche verbrachte. Sein sanfter Termindruck hat die Veröffentlichung unseres gemeinsamen Papers sicherlich beschleunigt.

Michael Höhle verdanke ich die grundlegende Orientierung im Doktorandenleben, in der Lehre und in der Erarbeitung von Literatur. Ich danke ihm darüber hinaus für sein offenes Ohr bei Fragen und Problemen und besonders für seine geduldige Korrekturarbeit am LISA-Paper.

Bei Andreas Mayr bedanke ich mich für die ergiebige und sympathische Zusammenarbeit an seiner Masterarbeit und an mehreren Papers. Unsere gemeinsame Forschung wird mir als beeindruckendes Beispiel für optimale Teamarbeit in Erinnerung bleiben.

Benjamin Hofner gilt mein Dank für seine ständige und schnelle Assistenz in Boosting-Fragen. Seine Doktorarbeit, die er mir genau im richtigen Moment zuschickte, bildete eine wichtige Grundlage für das Boosting-Kapitel.

Eva Rehfuess möchte ich herzlich danken für die sympathische und strukturierte Zusammenarbeit. Bei der gemeinsamen Analyse von Unterernährung bei Kindern in Indien gewährte sie mir viele spannende Einblicke in die Epidemiologie.

Christina Riedel und Holger Reulen danke ich für die engagierte Bearbeitung ihrer Masterarbeiten, aus der sich entscheidende Denkanstöße für meine Analyse der LISA-Daten ergaben.

Bei Manuel Eugster bedanke ich mich für seine große Hilfsbereitschaft in allen Computer- und R-Fragen und für das Teilen des täglichen Auf und Abs im Büroalltag über mehr als vier Jahre. Diesen Dank möchte ich ebenso an alle anderen Mitarbeiter des Instituts richten, die mich auf dem Weg als Doktorandin begleiteten und für eine angenehme, kurzweilige und sympathische Atmosphäre im Institutsalltag sorgten.

Monika Fenske danke ich von Herzen für ihre rückhaltlose Unterstützung. Ohne sie wäre ich sicher niemals bis hierher gekommen.

Bei Ludwig Bothmann bedanke ich mich herzlich für sein gründliches Korrekturlesen und den intensiven Beistand beim Entstehen dieser Arbeit. Seine Anmerkungen führten oft zu tiefgehenden Diskussionen über strukturiert additive Quantilregression und inspirierten mich immer wieder aufs Neue.

München, September 2012 Nora Fenske

Zusammenfassung

Die Quantilregression erweitert klassische Regressionsmodelle dahingehend, dass nicht nur der bedingte Erwartungswert, sondern die gesamte bedingte Verteilung einer Zielvariablen – ausgedrückt durch Quantile – in Abhängigkeit von Kovariablen modelliert werden kann.

Die vorliegende Arbeit führt die Modellklasse der *strukturiert additiven Quantilregression* ein. Diese Modellklasse kombiniert die Quantilregression mit einem strukturiert additiven Prädiktor, der die flexible Modellierung von zahlreichen Kovariableneffekten ermöglicht. Dieser Prädiktor enthält unter anderem glatte, nicht-lineare Effekte von stetigen Kovariablen und individuenspezifische Effekte, die insbesondere für longitudinale Daten wichtig sind.

Weiterhin gibt die Arbeit einen umfassenden Überblick über existierende Verfahren zur Parameterschätzung in strukturiert additiven Quantilregressionsmodellen, die eingeteilt werden in verteilungsfreie und verteilungsbasierte Schätzverfahren sowie in verwandte Modellklassen. Jedes Verfahren wird systematisch in Bezug auf die vier vorab definierten Kriterien diskutiert, (i) welche Komponenten eines flexiblen Prädiktors geschätzt werden können, (ii) welche Eigenschaften die Schätzer haben, (iii) ob Variablenselektion möglich ist, und (iv) ob es Software für die praktische Umsetzung gibt.

Die hauptsächliche methodische Neuentwicklung der Arbeit ist ein Boosting-Algorithmus, der als alternativer Schätzansatz für strukturiert additive Quantilregression vorgestellt wird. Beim Vergleich dieses innovativen Ansatzes im Hinblick auf die vier Kriterien zeigt sich, dass *Quantil-Boosting* große Vorteile in Bezug auf fast alle Kriterien – insbesondere auf Variablenselektion – mit sich bringt. Einen praktischen Vergleich von Quantil-Boosting mit den existierenden Schätzverfahren liefern anschließend die Ergebnisse mehrerer Simulationsstudien.

Motiviert wird die Entwicklung der strukturiert additiven Quantilregression durch zwei aktuell relevante Anwendungen aus dem Bereich der Epidemiologie: die Untersuchung von Risikofaktoren für Unterernährung bei Kindern in Indien (in einer Querschnittsstudie) sowie für Übergewicht und Adipositas bei Kindern in Deutschland (in einer Geburtskohortenstudie). In beiden Anwendungen werden extreme Quantile der Zielvariablen mit strukturiert additiver Quantilregression modelliert und mit Quantil-Boosting geschätzt. Die Ergebnisse werden ausführlich dargestellt und diskutiert.

Summary

Quantile regression allows to model the complete conditional distribution of a response variable – expressed by its quantiles – depending on covariates, and thereby extends classical regression models which mainly address the conditional mean of a response variable.

The present thesis introduces the generic model class of structured additive quantile regression. This model class combines quantile regression with a structured additive predictor and thereby enables a variety of covariate effects to be flexibly modelled. Among other components, the structured additive predictor comprises smooth non-linear effects of continuous covariates and individual-specific effects which are particularly important in longitudinal data settings.

Furthermore, this thesis gives an extensive overview of existing approaches for parameter estimation in structured additive quantile regression models. These approaches are structured into distribution-free and distribution-based approaches as well as related model classes. Each approach is systematically discussed with regard to the four previously defined criteria, (i) which different components of the generic predictor can be estimated, (ii) which properties can be attributed to the estimators, (iii) if variable selection is possible, and, finally, (iv) if software is available for practical applications.

The main methodological development of this thesis is a boosting algorithm which is presented as an alternative estimation approach for structured additive quantile regression. The discussion of this innovative approach with respect to the four criteria points out that *quantile boosting* involves great advantages regarding almost all criteria – in particular regarding variable selection. In addition, the results of several simulation studies provide a practical comparison of boosting with alternative estimation approaches.

From the beginning of this thesis, the development of structured additive quantile regression is motivated by two relevant applications from the field of epidemiology: the investigation of risk factors for child undernutrition in India (by a cross-sectional study) and for child overweight and obesity in Germany (by a birth cohort study). In both applications, extreme quantiles of the response variables are modelled by structured additive quantile regression and estimated by quantile boosting. The results are described and discussed in detail.

Contents

0 Outline **1**

1 Motivation and research goals **7**
1.1 Basics of linear quantile regression 7
1.2 Usage and typical applications of quantile regression 10
1.3 Research goals of this thesis 14

2 Applications in this thesis **17**
2.1 Undernutrition in developing countries 17
2.2 Overweight and obesity in western countries 29

3 Structured additive quantile regression – model class and estimation **35**
3.1 Generic model class 35
3.2 Estimation approaches – outline and assessment 38
3.3 Distribution-free estimation 39
3.3.1 Classical framework of quantile regression 39
3.3.2 Statistical learning and machine learning approaches 43
3.4 Distribution-based estimation 46
3.4.1 Asymmetric Laplace distribution approaches 46
3.4.2 Flexible Bayesian approaches 49
3.5 Related model classes 51
3.5.1 Expectile regression 52
3.5.2 Gaussian STAR models 54
3.5.3 GAMLSS 56

4 Boosting for structured additive quantile regression **59**
4.1 Algorithm 59
4.2 Base learners 62
4.3 Boosting parameters 70
4.4 Method assessment 73
4.5 Further remarks 76

5 Empirical evaluation of quantile boosting **79**
5.1 Simulation study for linear quantile regression 79
5.2 Simulation study for additive quantile regression 85
5.3 Comparing estimated quantile functions 92
5.4 Quantile boosting for individual-specific effects 95

6 Quantile boosting for child undernutrition in India **99**
6.1 Setup of the analysis 99
6.2 Results 101
6.3 Discussion 117

7 Quantile boosting for child overweight and obesity in Germany **119**
7.1 Setup of the analysis 119
7.2 Results 121
7.3 Discussion 131
7.4 Related own work 132

8 Discussion and outlook **135**
8.1 Summary and contributions of this thesis 135
8.2 Discussion of quantile boosting 136
8.3 Discussion of the application results 137
8.4 Possible directions for future research 138

Bibliography **143**

Chapter 0: Outline

This thesis originated from interdisciplinary work within the Munich Center of Health Sciences (MC-Health). With the aim of state-of-the-art quantitative empirical research in health and health sciences, this project brings together scientists from a wide range of research disciplines, such as epidemiology, medicine, economics, social sciences and statistics, and from different departments at the Ludwigs-Maximilians-Universität München and at the Helmholtz Zentrum München.

From the beginning of this thesis, two applications from the field of biostatistics motivated our research: the analysis of determinants of child undernutrition in developing countries and the analysis of risk factors for overweight and obesity in childhood in western countries. The statistical goal consisted in developing adequate statistical modelling approaches for these applications – we thereby focussed on quantile regression with a flexible predictor – and to explore the relative merits of these approaches regarding both applications.

Therefore, subordinate methodological questions were derived and investigated, resulting in several published manuscripts and manuscripts which are currently under review. These manuscripts build the base for this thesis. However, since they are closely related to each other, their contents are not disjoint. In search of an appropriate structure for this thesis, we aggregated the manuscripts to minimize redundancies and to maximize comprehension. In the following, we give an outline of the resulting structure and summarize the content of the manuscripts.

The index of contents on the previous pages i–ii provides a linear view on the structure, whereas the diagram in Figure 0.1 displays a more content-oriented view on the relationship between chapters. The content is roughly grouped into three grey boxes (model classes, estimation approaches and applications) and shortly described in the following.

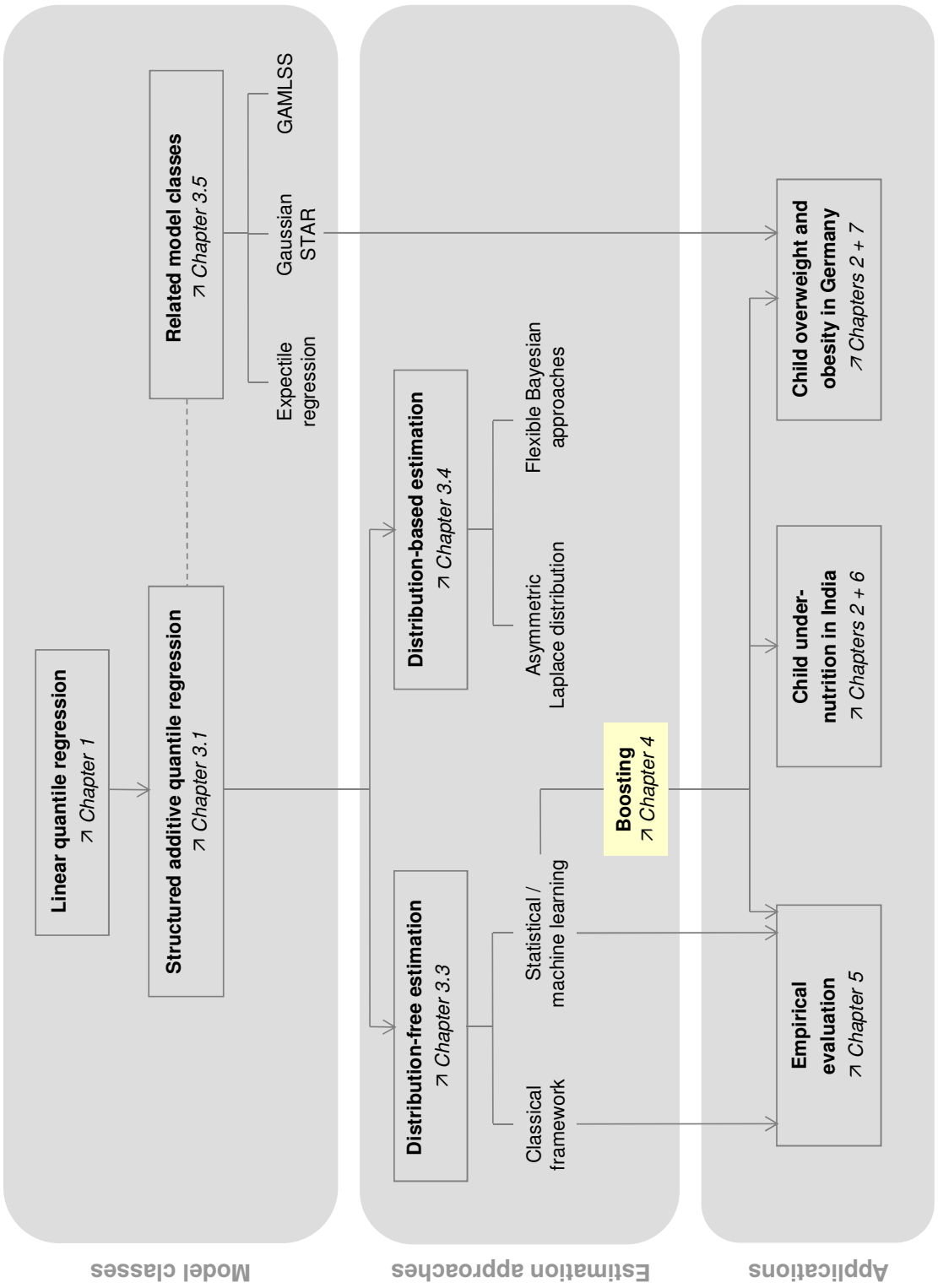


Figure 0.1 Content-oriented overview of the relationships between chapters of this thesis.

Model classes

As can be seen in the upper box of the diagram, we begin with a description of the conventional linear quantile regression model in Chapter 1. Thereby, some of the mathematical notation is introduced, an overview of typical application areas is given and the extension of linear quantile regression to more flexible modelling is motivated.

Structured additive quantile regression is the main model class of this thesis. In Chapter 3.1, we formulate the generic model class in analogy to (Gaussian) structured additive regression models (STAR, Fahrmeir, Kneib, and Lang, 2004) and thereby put quantile regression into the context and notation of modern flexible regression modelling.

Estimation approaches

The middle box of the diagram contains possible estimation approaches for structured additive quantile regression. Apart from the definition of the generic model class in Chapter 3.1, Chapter 3 gives an extensive overview of possible estimation approaches for the presented model class. We properly define criteria for method assessment and comparison in advance (Chapter 3.2).

In Chapter 3.3, we consider distribution-free approaches which do not rely on distributional assumptions for the error terms and aim at direct minimization of the quantile loss criterion. We distinguish between the classical framework of quantile regression, mainly consisting of linear programming algorithms, and computer-intensive statistical learning and machine learning algorithms, such as quantile regression forests or quantile neural networks.

In Chapter 3.4, we describe distribution-based approaches which assume an explicit error distribution, mainly the asymmetric Laplace distribution. We also sketch flexible Bayesian approaches where the error distribution consists of a mixture of Gaussian or other densities and which therefore can be regarded as distribution-based.

Chapter 3.5 treats related model classes to quantile regression which are placed in the upper box of Figure 0.1. These model classes can be applied in similar practice situations in which structured additive quantile regression would be appropriate. We again distinguish between one distribution-free model class (expectile regression) and two distribution-based model classes, that is, Gaussian STAR models and generalized additive models for location, scale and shape (GAMLSS).

Chapter 4 contains the main methodological contribution of this thesis. It presents a component-wise functional gradient descent boosting algorithm as innovative distribution-free estimation approach for structured additive quantile regression. In addition to a detailed description of the estimation of a large variety of effects from the structured additive predictor, properties of the *quantile boosting* algorithm are discussed with regard to the method assessment criteria from Chapter 3. This discussion points out that quantile boosting involves great advantages regarding almost all criteria – in particular regarding variable selection.

Applications

In Chapter 5, several simulation studies empirically evaluate the correctness of the proposed quantile boosting algorithm and compare it to the majority of distribution-free estimation approaches.

The motivating applications are introduced in Chapter 2 to illustrate application context and appropriateness of quantile regression from the beginning of this thesis. Chapter 6 contains the results of applying structured additive quantile regression to investigate determinants of child undernutrition in developing countries by means of a large cross-sectional dataset from India. Finally, Chapter 7 shows the results of a longitudinal quantile regression analysis of risk factors for child overweight and obesity in western countries based on a German birth cohort study called LISA.

Contributing Manuscripts

The present work is mainly based on the following manuscripts:

- *Fenske N, Kneib T, Hothorn T (2011): Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association, 106(494): 494-510.*

  This manuscript introduces quantile boosting, i.e. the model class of structured additive quantile regression combined with a component-wise functional gradient descent boosting algorithm. Quantile boosting is applied and compared to further distribution-free estimation approaches in a simulation study. The approach is illustrated by a first investigation of child undernutrition in India.

  Chapters 3.1, 4, and 5 are mainly based on contents of this manuscript.

- *Fenske N, Burns J, Hothorn T, Rehfuess EA (2012): Understanding child stunting in India: a comprehensive analysis of socio-economic, nutritional and environmental determinants using quantile boosting. American Journal of Clinical Nutrition, to be submitted.*

  This manuscript contains an evidence-based, comprehensive analysis of the various determinants of child undernutrition in India by boosting structured additive quantile regression.

  Chapters 2 and 6 are mainly based on contents of this manuscript.

- *Fenske N, Fahrmeir L, Hothorn T, Rzehak P, Höhle M (2012): Boosting structured additive quantile regression for longitudinal childhood obesity data. International Journal of Biostatistics, submitted.*

  This manuscript investigates boosting estimation for longitudinal quantile regression by focussing on individual-specific effects in the structured additive predictor. The approach is compared to Gaussian STAR models in an analysis of risk factors for overweight and obesity for a German birth cohort study called LISA.

  Chapter 7 and parts of Chapters 4 and 5 are based on contents of this manuscript.

The following manuscripts also contribute to parts of this thesis:

- *Mayr A, Hothorn T, Fenske N (2012): Prediction intervals for future BMI values of individual children – a non-parametric approach by quantile boosting. BMC Medical Research Methodology, 12(6).*

  This manuscript applies quantile boosting to construct prediction intervals for individual BMI values by means of the German LISA birth cohort study.

  Parts of Chapters 1, 2 and 7 are related to contents of this manuscript.

- *Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012): Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. Journal of the Royal Statistical Society, Series C (Applied Statistics), 61(3):403–427.*

  This manuscript introduces a boosting algorithm for the estimation of GAMLSS called gamboostLSS. The gamboostLSS approach is explored in a simulation study and applied to an analysis of data from the Munich rental guide.

  Parts of Chapters 3.5 and 4 are related to contents of this manuscript.

Chapter 1: Motivation and research goals

This chapter gives an introduction of the basic concepts of quantile regression and some of the mathematical notation that will be used throughout the thesis. It also presents typical application areas of quantile regression. Furthermore, the extension of standard linear quantile regression to structured additive quantile regression is motivated and, thereby, the underlying research goals of this thesis are sketched.

1.1 Basics of linear quantile regression

The most popular example for a quantile is probably the median. Empirically, it is defined as the value where 50% of a random sample have smaller values and 50% of the sample have greater values. The extension of this definition to other quantiles is straightforward. The 30% quantile, for example, is the value where 30% of the sample have smaller and 70% have greater values. Regarding several empirical quantiles at the same time can give an impression not only of the location or median of a random sample, but also of further distributional characteristics, such as variance, skewness, and kurtosis. Thus, quantiles go "beyond the mean" and can provide a complete picture of a sample distribution. This is also the basic idea of the boxplot, one of the most common tools to visualize a sample from a continuous variable.

In theory, *quantiles* are defined based on the *cumulative distribution function (cdf)* F_Y of a continuous random variable Y. The $\tau \cdot 100\%$ quantile of Y can be written as a value y_τ where

$$F_Y(y_\tau) = P(Y \leq y_\tau) = \int_{-\infty}^{y_\tau} f(u)\,\mathrm{d}\,u = \tau \qquad \text{for} \qquad \tau \in (0,1)\ .$$

It is only unique if F_Y is strictly monotonic increasing. The boundaries 0 and 1 are not included in the range of τ for reasons of uniqueness. In case that information on an additional random variable X is given, the quantile can similarly be expressed conditional on $X = x$:

$$F_Y(y_\tau(x) \mid X = x) = P(Y \leq y_\tau(x) \mid X = x) = \tau\ .$$

The *quantile function* $Q_Y(\tau \mid X = x)$ is defined as the smallest y where the quantile property is fulfilled if F_Y is not strictly monotonic, i.e.,

$$Q_Y(\tau \mid X = x) = \inf\{y : F_Y(y \mid X = x) \geq \tau\}\ ,$$

and is set to the inverse of the cdf of Y, i.e., $Q_Y(\tau \mid X = x) = F_Y^{-1}(\tau \mid X = x)$, if F_Y is strictly increasing.

Thus, the relationship between quantile function and cdf (for strictly increasing F_Y) can be expressed as

$$F_Y(y_\tau(x) \mid X = x) = \tau \qquad \Leftrightarrow \qquad Q_Y(\tau \mid X = x) = y_\tau(x)\ ,$$

which emphasizes that the quantile function describes $\tau \cdot 100\%$ quantiles of Y depending on covariates x and a quantile parameter $\tau \in (0,1)$.

Note that throughout this thesis, the term *quantile* will be used synonymous with *percentile*.

Quantile regression carries over the idea of going beyond the mean to regression modelling. It is an approach to model the conditional quantile function of a continuous variable of interest Y, denoted as *response variable* in the following, depending on further variables or *covariates* X. In accordance with linear mean regression models, the *linear quantile regression model* can be written as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \varepsilon_{\tau i} \ , \tag{1.1}$$

see, for example, Buchinsky (1998). The index $i = 1, \ldots, n$, denotes the observation, y_i is the response value and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ the given covariate vector for observation i. The quantile-specific linear effects are denoted by $\boldsymbol{\beta}_\tau = (\beta_{\tau 0}, \beta_{\tau 1}, \ldots, \beta_{\tau p})^\top$, and $\tau \in (0, 1)$ indicates a quantile parameter which has to be fixed in advance. The random variable $\varepsilon_{\tau i}$ is assumed to be an unknown error term with cdf $F_{\varepsilon_{\tau i}}$ and density $f_{\varepsilon_{\tau i}}$ depending on quantile parameter τ and observation i.

At first glance, model (1.1) looks like a standard linear regression model which aims at modelling the response's mean depending on covariates. However, the crucial difference between a standard linear regression model with Gaussian errors and quantile regression is the distributional assumption for the error terms. For quantile regression, no specific assumptions are made apart from $\varepsilon_{\tau i}$ and $\varepsilon_{\tau j}$ being independent for $i \neq j$, and

$$\int_{-\infty}^{0} f_{\varepsilon_{\tau i}}(\varepsilon_{\tau i}) \ \mathrm{d}\, \varepsilon_{\tau i} = F_{\varepsilon_{\tau i}}(0) = \tau \ . \tag{1.2}$$

Due to this assumption, model (1.1) aims at describing the quantile function $Q_{Y_i}(\tau|\boldsymbol{x}_i)$ of the response variable Y_i conditional on covariate vector $\boldsymbol{x}_i$ at a given quantile parameter τ. This can be seen after the following steps. First, the cdf of Y_i can be expressed in terms of the cdf of $\varepsilon_{\tau i}$:

$$\begin{aligned} F_{Y_i}(y_\tau \mid \boldsymbol{x}_i) &= P\left(Y_i \leq y_\tau \mid \boldsymbol{x}_i\right) \\ &= P\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \varepsilon_{\tau i} \leq y_\tau \mid \boldsymbol{x}_i\right) \\ &= P\left(\varepsilon_{\tau i} \leq y_\tau - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \mid \boldsymbol{x}_i\right) \quad = F_{\varepsilon_{\tau i}}(y_\tau - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \mid \boldsymbol{x}_i) \ . \end{aligned}$$

Then, the $\tau \cdot 100\%$ quantile of Y can be derived as:

$$\begin{aligned} & & F_{Y_i}(y_\tau \mid \boldsymbol{x}_i) &= \tau \\ \Leftrightarrow & & F_{\varepsilon_{\tau i}}(y_\tau - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \mid \boldsymbol{x}_i) &= \tau \\ \Leftrightarrow & & y_\tau - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau &= F_{\varepsilon_{\tau i}}^{-1}(\tau) \\ \Leftrightarrow & & y_\tau &= \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + F_{\varepsilon_{\tau i}}^{-1}(\tau) \ . \end{aligned}$$

With the assumption in (1.2), it follows that $F_{\varepsilon_{\tau i}}^{-1}(\tau) = 0$, and thus:

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau =: \eta_{\tau i} \ . \tag{1.3}$$

Therefore, the regression parameters $\boldsymbol{\beta}_\tau$ quantify linear relationships between covariates and the quantile function of the response. More specifically, the parameter $\beta_{\tau 1}$, for example, can be interpreted as the change of the conditional quantile function when x_{i1} changes to $x_{i1}+1$, given all other covariates remain constant. The quantile regression *predictor*, which is linear in the simplest case here, is abbreviated with $\eta_{\tau i}$.

The index τ for $\boldsymbol{\beta}_\tau$ points out that the regression parameters can differ for different values of τ. An example for this situation is given by Figure 1.1. Panel (a) shows simulated data from a heteroscedastic data setup as well as true underlying quantile functions for a grid of quantile parameters. It can be observed that x has different linear relationships with the median and other quantiles of the response and that the slope parameters $\beta_{\tau 1}$ depend on τ, see also panel (b). This is a typical data situation where quantile regression would be more appropriate than mean regression.

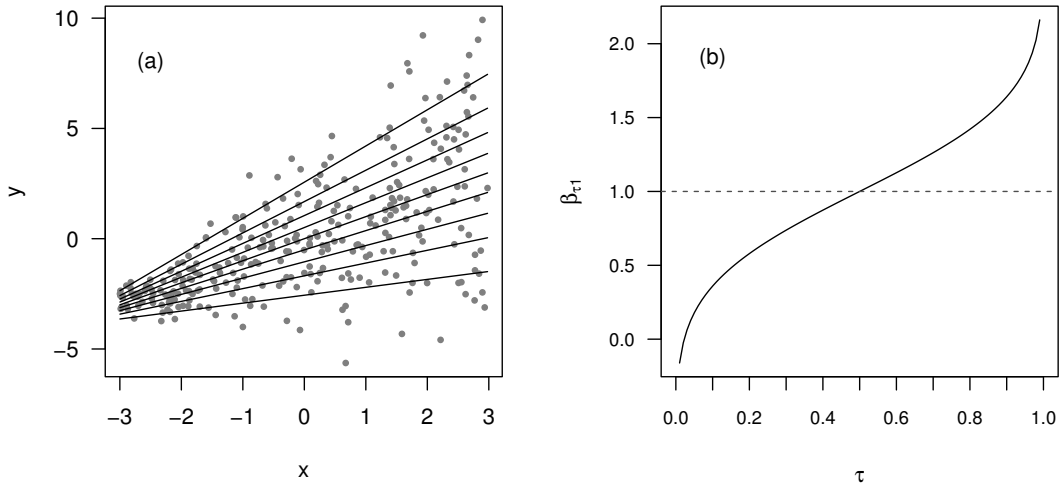


Figure 1.1 Simulation example of a heteroscedastic data setup. Panel (a): Grey points stand for $n = 300$ observations simulated from the model $y_i = x_i + (0.5x_i)\varepsilon_i$ with $x_i \sim \mathcal{U}[-3, 3]$ and $\varepsilon_i \sim \mathcal{N}(0, 4)$. Black lines show the true underlying quantile curves for an equidistant grid of quantile parameters. Panel (b): Functional relationship between τ and true slope parameters $\beta_{\tau 1}$ for simulation setup from panel (a).

However, the parameters $\boldsymbol{\beta}_\tau$ have to be interpreted with care. In general, they cannot be interpreted on an individual-specific level. A person who happens to be at the $\tau \cdot 100\%$ quantile of the response conditional on his/her specific covariate vector would not be at the same quantile any more when his/her covariate vector changes. When knowing $\boldsymbol{\beta}_\tau$, the only conclusion that can be drawn is how the $\tau \cdot 100\%$ quantile of a population with a specific covariate combination differs from the $\tau \cdot 100\%$ quantile of a population with a different covariate combination.

Three additional remarks should be made here. In model (1.1), we directly started from formulating the linear quantile regression model in analogy to standard mean regression models. Originally, the concept of quantile regression traces back to Koenker and Bassett (1978), where regression quantiles were defined by minimizing a weighted sum of absolute deviances:

$$\sum_{i=1}^{n} \rho_\tau(y_i - \eta_{\tau i}) \qquad \text{with } \textit{check function} \qquad \rho_\tau(u) = \begin{cases} u\tau & u \geq 0 \\ u(\tau - 1) & u < 0 \,. \end{cases} \tag{1.4}$$

This definition of quantile regression paves the way for distribution-free estimation approaches which only ask for the specification of a loss function. These approaches will be thoroughly treated in Section 3.3.

Note also that our formulation of quantile regression in (1.1) should emphasize our view that Gaussian mean regression can be regarded as a special case of quantile regression. Every regression model with a specific distributional assumption describes the full conditional distribution – and therefore all quantiles – of the response variable depending on covariates. In case that the covariates are associated with the response's quantiles in a linear way, the resulting quantile regression model is linear as introduced above. However, the relationship between predictor and quantile function is not linear in the general case of regression models with distributional assumptions. This view on quantile regression will be further worked out in Section 3.5.

Regarding introductory literature on quantile regression, the book of Koenker (2005) has established as standard reference since it gives an extensive overview of the status quo of research in the classical framework of quantile regression. A more applied introduction is given by Hao and Naiman (2007) based on various examples from the social sciences. The master's thesis of Fenske (2008) provides a brief and application-oriented introduction to quantile regression in German.

1.2 Usage and typical applications of quantile regression

Over the last years, quantile regression has become a popular statistical method for addressing various research questions. Apart from epidemiological applications treated in this thesis, quantile regression has recently been applied to a large number of different areas, ranging from social and educational sciences (e.g., Hao and Naiman, 2007; Arulampalam *et al.*, 2011) to environmental and ecological sciences (e.g., Cade *et al.*, 2008; Mehtätalo *et al.*, 2008) and problems in economics (e.g., Franck and Nüsch, 2012; Matano and Naticchioni, 2012; Pendakur and Woodcock, 2010).

In general, quantile regression is useful when the shape of the response's distribution depends on covariates, i.e., when the error terms are not *iid*, or when the response does not follow a well-known distribution, e.g., when it is not symmetric or when heavy tails or outliers are present.

The specific usage of quantile regression depends on the goal of the respective analysis. The decisive question that has to be answered prior to each quantile regression analysis is which quantile parameters τ should be considered; and the answer to this question determines the specific usage. In our view, there are two alternative usages: quantile regression for a small number of quantile parameters vs. quantile regression for a large number or a grid of quantile parameters.

In the following, we shortly describe both alternatives with regard to underlying aims and typical corresponding applications.

Alternative 1: Quantile regression for a small number of quantile parameters

- When the area of interest is not the mean of the response but a particular quantile interval, quantile regression can simply be conducted for a small number of quantile parameters from this interval.

  This will be the case for the two motivating applications of this thesis, dealing with undernutrition of children in developing countries and with overweight and obesity of children in western countries. Both datasets will be investigated by quantile regression

for anthropometric measurements depending on child-specific covariates. In case of undernutrition, the area of interest are lower quantiles of height-for-age values rather than the mean, whereas for overweight and obesity, upper quantiles of the body mass index (BMI) are in the focus of the analysis.

The value at risk (VAR) is a typical application from financial and economics research where the interest is directed towards particular quantiles. It is an important measure for quantifying daily risks. Since its definition is directly based on extreme quantiles of risk measures, it seems obvious to use quantile regression for VAR modelling (see Yu *et al.*, 2003, for further references).

- In case that the area of interest is the response's mean but heavy tails or large outliers are present in the sample, median regression can be applied as a special case of this usage alternative. Since the robustness property of the median carries over to median regression (see Koenker, 2005, Chap. 2.3), it is to be preferred to mean regression in the presence of outliers.

- Another situation in which quantile regression just has to be performed for two particular quantile parameters is the construction of prediction intervals, as proposed by Meinshausen (2006).

  To obtain a $(1-\alpha)\cdot 100\%$ prediction interval for a future response value, quantile regression is first performed for the two particular quantile parameters $\tau_1 = \alpha/2$ and $\tau_2 = 1-\alpha/2$. Then, the new covariate observation x_{new} is plugged into the estimated predictor. The resulting estimated quantiles of y_{new} directly denote the borders of a $(1-\alpha)\cdot 100\%$ prediction interval PI for y_{new} as follows:

$$\widehat{\text{PI}}_{1-\alpha}(x_{\text{new}}) = \left[\widehat{Q}_Y\left(\frac{\alpha}{2} \mid X = x_{\text{new}}\right) , \widehat{Q}_Y\left(1-\frac{\alpha}{2} \mid X = x_{\text{new}}\right) \right] .$$

  In Mayr *et al.* (2012c), we applied this usage of quantile regression to construct prediction intervals for future BMI values of individual children based on the German birth cohort study which will be introduced in Section 2.2. Since the BMI distribution in childhood is typically skewed depending on age (see Figures 1.2 and 1.3), quantile regression was more adequate to construct prediction intervals than standard approaches based on mean regression.

Alternative 2: Quantile regression for a large number or a grid of quantile parameters

- The objective of many applications consists of investigating the complete conditional distribution of the response variable depending on covariates. In these situations, it is not sufficient to look at a small number of particular quantile parameters only.

  For example, Gilchrist (2008, p.2) stress that "[...] when modelling regression the whole model should be considered, both deterministic and stochastic terms, and a balanced consideration should be given to the forms of both." He regards the model predictor as deterministic and the error distribution as stochastic component. When quantile regression is conducted for a grid of several quantile parameters $\tau \in (0,1)$ at the same time, it provides a complete picture of the error distribution and, therefore, addresses the stochastic component in the analysis.

A typical application of this usage of quantile regression is the construction of child growth standards. Figures 1.2 and 1.3 exemplarily show the World Health Organization (WHO) growth charts for boys aged 0-5 years and 5-19 years, respectively. Displayed are five quantile curves of the BMI by age. The shape of the quantile curves suggests that the BMI distribution becomes right-skewed beginning somewhere after the age of 6 years. Also, the BMI quantile curves are not linear. For this reason, flexible regression methods are needed to obtain the smooth nonlinear BMI quantile curves shown in Figures 1.2 and 1.3. Borghi *et al.* (2006) presents an extensive review of possible regression methods to obtain such growth charts – including also quantile regression for a large number of quantile parameters.

- The usage of quantile regression for a grid of quantile parameters can also be helpful to detect deviations from an *iid* error distribution. In this context, Koenker (2005) proposes to visualize the regression results by a plot of τ versus β_τ. As an example, Figure 1.1(b) displays the relationship between τ and the true slope parameter $\beta_{\tau 1}$ for the simulation setup from Figure 1.1(a). The shape of the resulting function is not constant and suggests heteroscedasticity in the data. Koenker (2005, p.29) explains how to match further typical patterns of τ versus β_τ with underlying distributional shapes of Y given X.

However, one should be aware that distribution-free estimation of quantile regression is usually performed separately for different quantile parameters. This runs the risk of *quantile crossing* which would, for example, be present when the estimated median is greater than the estimated 60% quantile given the same specific covariate combination. The danger of quantile crossing is in particular given at the boundaries of the covariate space when quantile parameters close to each other are investigated. Quantile crossing at any point of the covariate space can for example be avoided by using distribution-based estimation approaches for quantile regression which directly assume a specific distribution for the stochastic component. High research efforts are also made to develop distribution-free estimation approaches that respect the monotonicity of the quantile function.

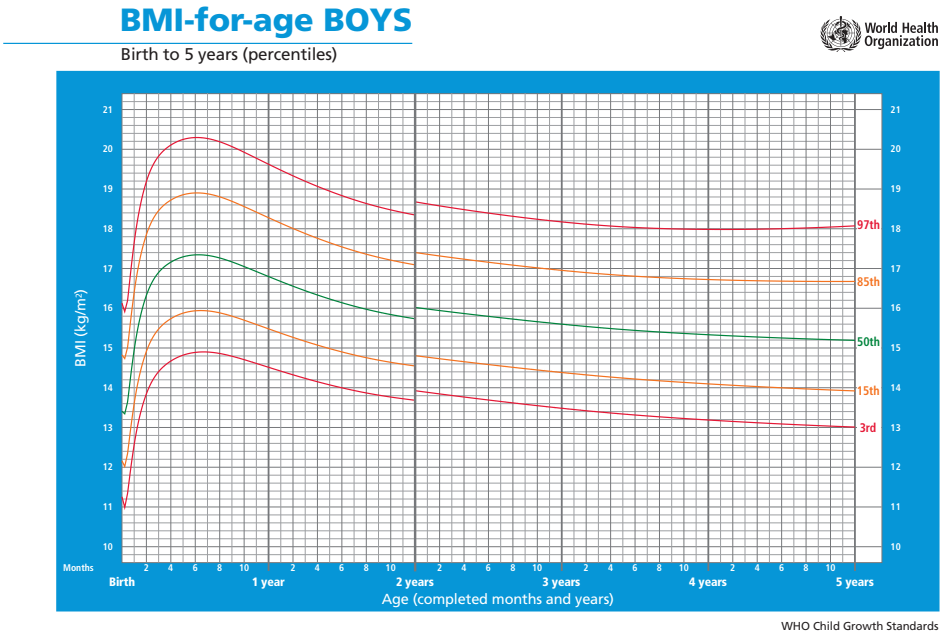


Figure 1.2 BMI-for-age quantile curves for boys aged 0-5 years from the WHO child growth standards. Source: `http://www.who.int/childgrowth/standards/chts_bfa_boys_p/en/index.html`

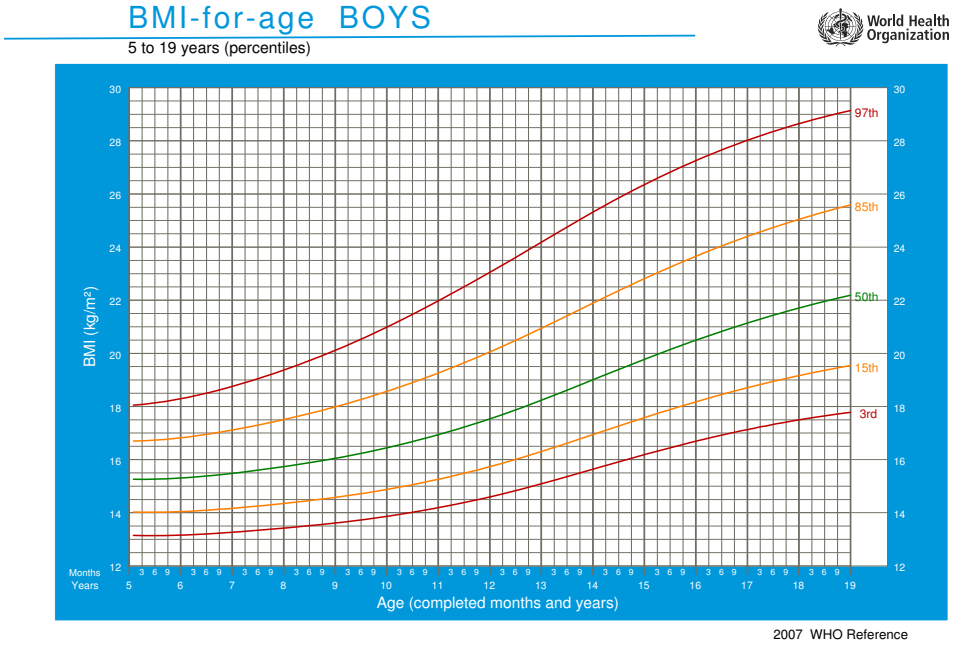


Figure 1.3 BMI-for-age quantile curves for boys aged 5-19 years from the WHO child growth standards. Source: `http://www.who.int/growthref/who2007_bmi_for_age/en/index.html`

1.3 Research goals of this thesis

In practice, the linear quantile regression model (1.1) does not always suffice to adequately express the relationship between covariates and quantile functions of the response variable. For example, Figure 1.4 shows simulated data where the shapes of the quantile curves are nonlinear and even depend on the quantile parameter.

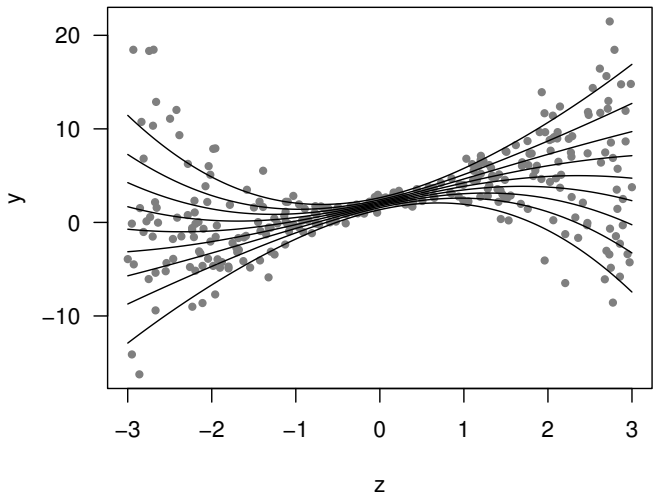


Figure 1.4 Simulation example of a heteroscedastic and nonlinear data setup. Grey points stand for $n = 300$ observations simulated from the model $y_i = 2 + 3\sin(\frac{2}{3}x_i) + x_i^2\varepsilon_i$ with $x_i \sim \mathcal{U}[-3, 3]$ and $\varepsilon_i \sim \mathcal{N}(0, 0.25)$. Black lines show the true underlying quantile curves for an equidistant grid of quantile parameters.

We were faced with a similar data situation in our motivating applications since the relationships between age and anthropometric measurements of children are typically nonlinear, as for example shown for the BMI quantile curves in Figures 1.2 and 1.3.

At the same time, we had to deal with a longitudinal data setup, since our obesity analysis was based on a German birth cohort study with repeated observations per child. Consequently, the statistical task was to account for the unobserved heterogeneity, i.e., correlation between intra-individual observations that is not covered by covariates, which is typically present in longitudinal data.

Two additional challenges were given by the analysis of undernutrition of children in India: First, regional differences were expressed by a spatial covariate with 29 different states of India. Therefore the method also had to consider the unobserved heterogenity arising from the spatial setup since one can assume that observations of children from the same state or from neighbouring states are more similar than observations of non-neighbouring states. Second, our modelling had to reflect that the meaning of feeding variables, such as breastfeeding and complementary feeding, varies with age.

These methodical issues motivated us to define the research goals for this thesis as follows:

- The main research goal was to extend quantile regression to a structured additive predictor comprising a large variety of different effects, such as smooth nonlinear effects to model nonlinear relationships between quantiles of the response and continuous covariates; smooth varying coefficient terms to model (potentially nonlinear) time-varying effects of further covariates; individual-specific effects to account for the longitudinal data structure; and spatial effects to account for potential spatial correlation.
- Consequently the first goal was a comprehensive review of existing estimation approaches for flexible quantile regression in order to explore their potential for estimating structured additive quantile regression models.
- More importantly, we set the goal to develop new approaches for the estimation of structured additive quantile regression. We introduced and investigated boosting as innovative distribution-free estimation algorithm.
- Our final goal was to investigate if applying structured additive quantile regression to our two health applications could lead to new substantive insights.

Chapter 2: Applications in this thesis

In this chapter, we give an overview of data and objectives of our two main applications. In addition, we motivate the appropriateness of structured additive quantile regression for them. Section 2.1 is mainly based on Fenske, Burns, Hothorn, and Rehfuess (2012a), whereas details on Section 2.2 can be found in Fenske, Fahrmeir, Hothorn, Rzehak, and Höhle (2012b).

2.1 Undernutrition in developing countries

Background and epidemiological aim

Child undernutrition is the cause of one third of deaths in children under five and produces serious consequences throughout the life course, including intellectual disability and metabolic and cardiovascular disease (Black *et al.*, 2008; Caulfield *et al.*, 2006; UNICEF *et al.*, 2011). Low height-for-age or *stunting* reflects a failure to reach linear growth potential, and is a key indicator of chronic undernutrition. Globally, 171 million children under five were classified as stunted in 2010 (WHO, 2012), with 90% of this burden occurring in 36 African and Asian countries.

Stunting is the result of a complex interplay of factors. Gaining a better understanding of these factors is critical for identifying entry-points for effective intervention. Thus, the overall epidemiological aim of our study was to undertake a comprehensive, systematic and evidence-based analysis of the multiple determinants of child stunting.

Schematic diagram of determinants

We used the UNICEF childhood undernutrition framework (UNICEF, 1998) as a starting point, since it provides a theoretical basis for system thinking in the area of child undernutrition. Based on extensive literature searches we structured potential risk factors in a schematic diagram of immediate, intermediate and underlying determinants of child stunting, shown by Figure 2.1.

According to the UNICEF framework, we defined sixteen main groups of determinants for stunting and grouped them into non-modifiable factors (child age and sex) and three layers standing for immediate, intermediate and underlying determinants. The top layer of the diagram contains the most important modifiable immediate determinants of stunting, comprising intrauterine growth restriction (IUGR) and inadequate caloric and nutrient intake and uptake. The majority of groups of determinants is located in the middle layer of intermediate determinants, for example household food competition; water, sanitation and hygiene; breastfeeding and complementary feeding practices; indoor air pollution; etc. The bottom layer consists of three groups of underlying determinants, that is maternal, household and regional characteristics. Detailed information on all determinants and corresponding literature can be found in Fenske, Burns, Hothorn, and Rehfuess (2012a).

The complex interplay of determinants is also emphasized by arrows between layers. We assume direct effects of all groups of determinants on stunting, but also indirect effects of determinants through superordinate layers.

Dataset

With an estimated stunting prevalence of 51% and 61 million stunted children, India is the most affected country in the world and, therefore, was chosen as the focus of this study. We used data from the Indian National Family Health Survey (NFHS) for the years 2005/2006 (International Institute for Population Sciences and Macro International, 2007) which corresponds to the Indian version of the well-known Demographic and Health Surveys (DHS). NFHS/ DHS are large-scale, well-established, nationally representative surveys based on a multi-stage cluster sample design that provide high-quality information on the health and nutrition of women and children.

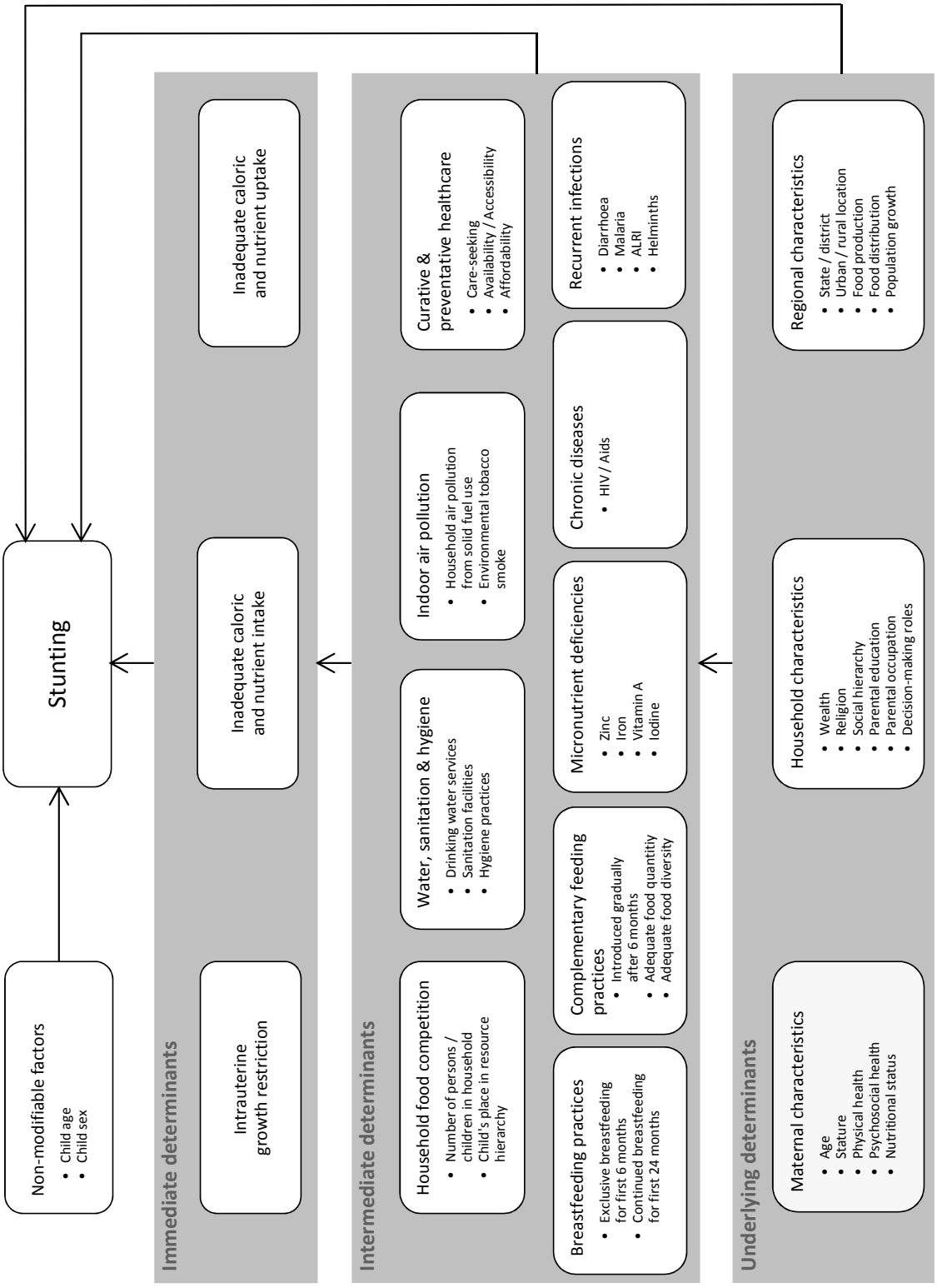


Figure 2.1 Schematic diagram of the multiple determinants of child stunting structured by layer and groups of determinants.

Quantification of stunting

According to the WHO child growth standards (WHO Multicentre Growth Reference Study Group, 2006), we quantified stunting as inadequate height-for-age. More precisely, stunting was measured by a Z-score obtained from a standardization of children's height. The Z-score for child i was computed as

$$Z_i = \frac{\text{height}_i - M(\text{age}_i, \text{sex}_i)}{S(\text{age}_i, \text{sex}_i)} \ , \tag{2.1}$$

with M and S being median and standard deviation of height in the reference group stratified with respect to age and sex. Stunting was quantified as low height-for-age values in our analysis. Figures 2.2 and 2.3 show the corresponding WHO height-for-age reference charts from birth to 5 years for boys and girls, respectively. The reference population consists of exclusively breastfed healthy children born between 1997 and 2003 from comparable affluent backgrounds in different countries. The lowest black curves stand for a Z-score of -3 (obtained from $M-3\cdot S$) and exactly correspond to the 0.1% age- and sex-specific quantile curves of height, whereas the lower red curves for a Z-score of -2 exactly display the 2.3% quantile curves of height in the reference population. The jump discontinuities at the age of two years result from the fact that the growth charts were constructed by means of two separate datasets with children older and younger than two years.

By using Z-scores instead of raw height values, the degree of undernutrition of a child can be assessed without regarding its age and sex. Therefore, Z-scores and binarized versions of them are commonly used in the analysis of child undernutrition. In addition to the Z-score as continuous response variable, we constructed binary variables for being stunted or severely stunted. According to these variables, children with an age- and sex-specific Z-score less than -2 or -3 (i.e., below the lower red or black Z-score curves) were classified as stunted or severely stunted, respectively.

Note that the WHO growth reference curves as shown by Figures 2.2 and 2.3 were obtained by various regression approaches closely related to quantile regression (see Borghi *et al.*, 2006; Cole *et al.*, 2000; Wei *et al.*, 2006, for details), and Z-scores for further anthropometric measures, such as BMI or weight, can be calculated similarly as for height. However, in case of height no age-specific skewness parameter is necessary and the distributional shape reduces to a Gaussian distribution with age- and sex-specific parameters for mean and standard deviation. In case of the BMI, the transformation between BMI values and corresponding Z-scores becomes more involved than in equation (2.1) due to an emerging BMI skewness at the age of 6 (see Section 2.2 and Figure 1.3).

Figure 2.4 displays the observations of children's height depending on age and sex in our final dataset, superimposed by the height-for-age Z-score curves from Figures 2.2 and 2.3. The distributional shape of height is in accordance with the reference distribution. However, beginning around the age of six months, the height distribution of Indian children is clearly below the reference distribution.

Furthermore, Figure 2.5 shows the Z-score values for height-for-age in the final dataset resulting from the transformation described above. One can see that the shape of the Z-score distribution is symmetric and remains stable with age and sex. The Z-score curves from Figure 2.4 are not drawn since they just correspond to constant lines at $-3, -2, 0, 2$ and 3. Instead, the Z-score

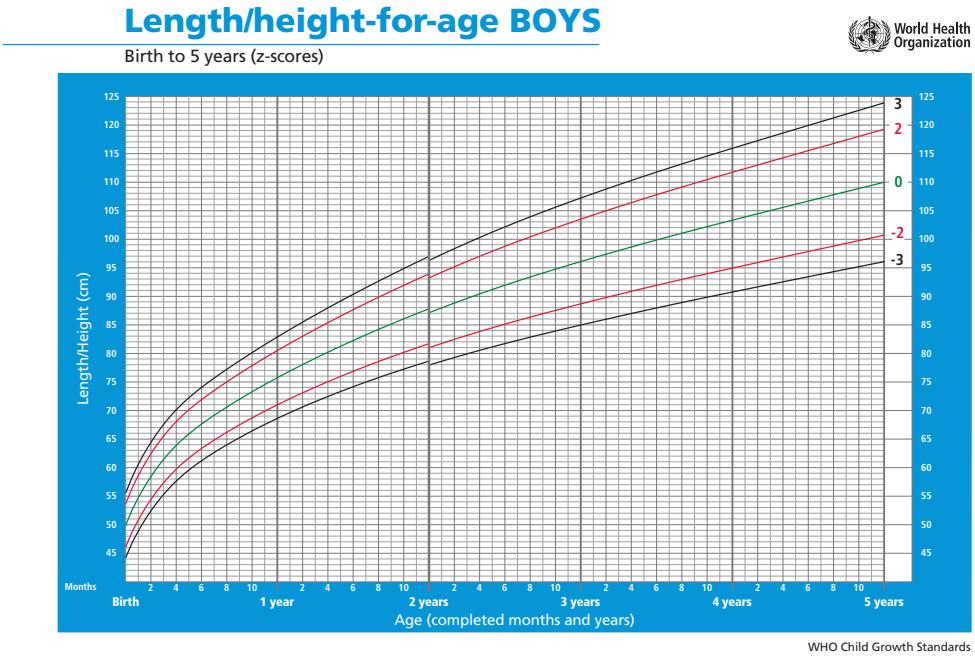


Figure 2.2 Length/Height-for-age Z-score curves for boys aged 0-5 years from the WHO child growth standards. Source: `http://www.who.int/childgrowth/standards/chts_lhfa_boys_z/en/index.html`

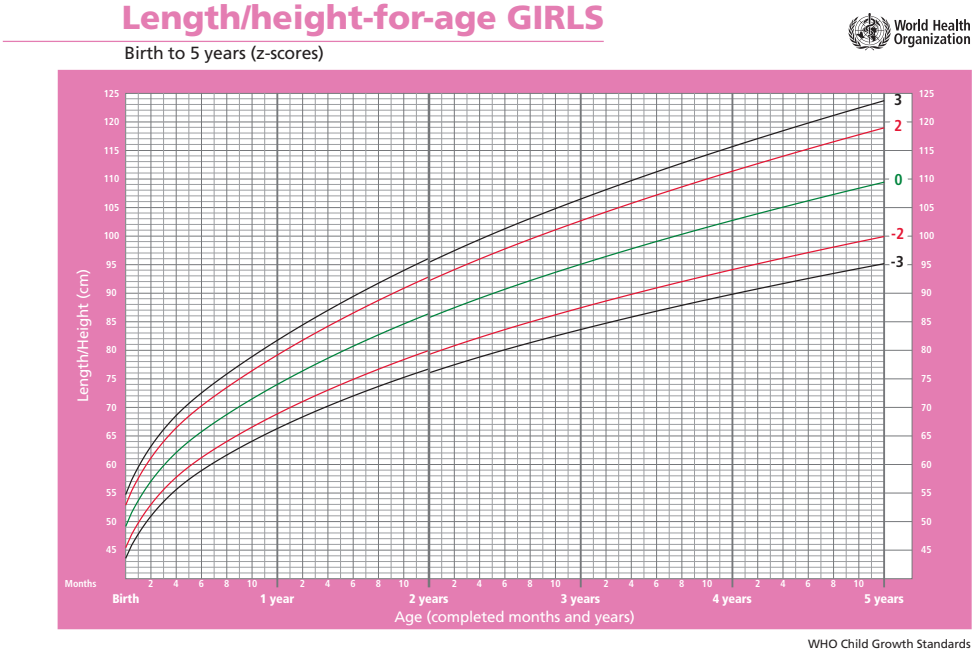


Figure 2.3 Length/Height-for-age Z-score curves for girls aged 0-5 years from the WHO child growth standards. Source: `http://www.who.int/childgrowth/standards/chts_lhfa_girls_z/en/index.html`

observations are superimposed by empirical lower quantile curves (which were estimated by local linear quantile regression; Yu and Jones, 1998). The quantile parameters are chosen in accordance with the later quantile regression analysis.

Thus, Figure 2.5 suggests a negative linear age effect for all Z-score quantiles. The parallel shift of the curves indicate that quantile regression coefficients for age would probably be similar for different quantile parameters. Beginning around the age of 12 months, a huge part of the Indian children have Z-score values smaller than -2 and are therefore classified as stunted.

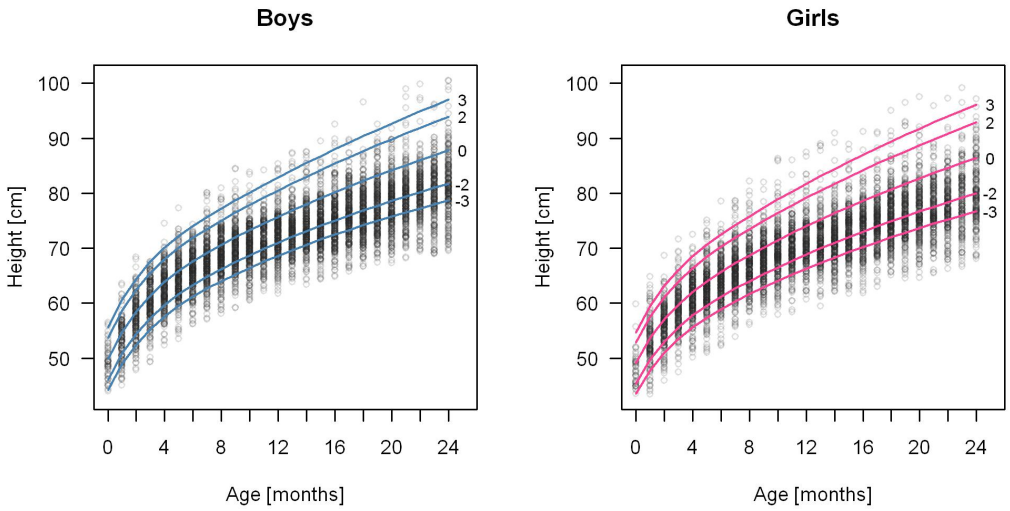


Figure 2.4 Height values (grey points) by age and sex in our final dataset, superimposed by Z-score curves from Figures 2.2 and 2.3.

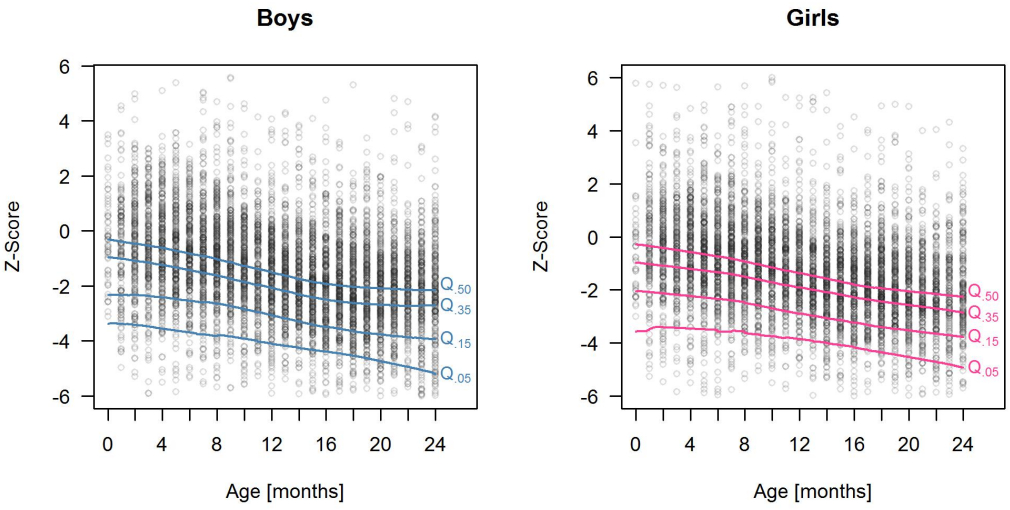


Figure 2.5 Z-score values (grey points) for height-for-age by age and sex in the final dataset, superimposed by empirical lower quantile curves which were estimated by local linear quantile regression (Yu and Jones, 1998).

Covariates

The diagram in Figure 2.1 served as a basis for identifying relevant variables within each group of determinants. Table 2.1 shows all variables and their empirical distributions contained in the final dataset and is arranged by groups of determinants from Figure 2.1.

The dataset contains variables to populate most groups of determinants, but variables from the layer of immediate determinants (intrauterine growth restriction, nutrient intake and uptake) were not available. In the layer of intermediate determinants, we could not assess measures or suitable proxies of the groups of chronic diseases and recurrent infections, since information on infections was only available on disease symptoms two weeks prior to the survey (which were considered unsuitable). For some groups, we could not cover all characteristics of interest, for example in relation to maternal psychosocial health, food production and distribution, hygiene practices, environmental tobacco smoke and zinc.

With regard to the non-modifiable determinant age, our analysis focussed on living children aged 0-24 months, as stunting prevalence progressively increases until it reaches a plateau at around 24 months, see Black *et al.* (2008).

The wealth variable from the group of household characteristics is an established index for economic status and commonly used in DHS surveys. To construct this variable, 33 housing characteristics were aggregated, such as type of toilet facility; type of windows and roofing; presence of electricity, television and radio; possession of a car; etc. The grouping into five categories is done on national level and based on quintiles of an underlying continuous variable (for more details see International Institute for Population Sciences and Macro International, 2007).

Concerning water and sanitation, we prepared the variables according to the guidelines of the WHO/UNICEF joint monitoring programme for water supply and sanitation (see `http://www.wssinfo.org/definitions-methods/watsan-categories`).

Regarding the group of curative and preventative healthcare, we examined various measures (e.g., possession of health card, health facility visit in past three months) but ultimately settled for the number of antenatal visits as a proxy for care during pregnancy and childbirth, and constructed a vaccination index based on vaccinations against measles, polio, tuberculosis (BCG) and diphtheria, pertussis and tetanus (DPT) as a proxy for care during childhood.

We constructed a three-level variable for breastfeeding and two variables for complementary feeding practices. Thereby, food diversity was measured as the number of food groups a child had consumed in the previous 24 hours apart from breast milk, with the eight food groups defined as in the NFHS report comprising food made from grains; food made from roots; food made from beans, peas, lentils, nuts; fruits and vegetables rich in vitamin A; other fruits and vegetables; meat, fish, poultry, eggs; cheese, yoghurt, other milk products. Food quantity was assessed as meal frequency, i.e., the number of times a child received anything to eat other than breast milk in the previous 24 hours. Grouping of both variables was based on empirical frequencies in our dataset in order to obtain sufficiently large group sizes.

Table 2.1 Overview of variables and their empirical distributions contained in the final dataset with N = 12 176 observations, arranged by groups of determinants from Figure 2.1.

Variable	**Values / Description**	**Number**	**Percentage**
Stunting			
Z-score for height-for-age	Mean: -1.37, Median: -1.44, Sd: 1.79, Range: [-6, 6]		
Child is stunted	No	7699	63.2%
	Yes	4477	36.8%
Child is severely stunted	No	10089	82.9%
	Yes	2087	17.1%
Non-modifiable factors			
Child age [months]	Mean: 12.46, Median: 13, Sd: 6.62, Range: [0, 24]		
Child sex	Male	6317	51.9%
	Female	5859	48.1%
Maternal characteristics			
Maternal age [years]	Mean: 25.66, Median: 25, Sd: 5.21, Range: [15, 49]		
Maternal BMI [kg/m^2]	Mean: 20.10, Median: 19.52, Sd: 3.26, Range: [12.04, 40.34]		
Household characteristics			
Household wealth	Poorest	2180	17.9%
	Poorer	2226	18.3%
	Middle	2463	20.2%
	Richer	2726	22.4%
	Richest	2581	21.2%
Religion of household head	Hindu	8683	71.3%
	Muslim	1714	14.1%
	Christian	1232	10.1%
	Sikh	224	1.8%
	(Neo-)Buddhist	137	1.1%
	Other	186	1.5%
Caste/tribe of household head	Scheduled caste	2222	18.2%
	Scheduled tribe	2098	17.2%
	Other backward class	4188	34.4%
	None of them	3668	30.1%
Maternal education [years]	Mean: 5.40, Median: 5, Sd: 5.16, Range: [0, 20]		
Partner's education [years]	Mean: 7.21, Median: 8, Sd: 5.07, Range: [0, 22]		
Partner's occupation	Services	4933	40.5%
	Household & domestic	697	5.7%
	Agriculture	3361	27.6%
	Clerical	1752	14.4%
	Prof./ Tech./ Manag.	497	4.1%
	Did not work	936	7.7%
Mother is currently working	No	9045	74.3%
	Yes	3131	25.7%
Sex of household head	Male	10958	89.8%
	Female	1247	10.2%
Regional characteristics			
State of residence	29 states of India, see Figure 2.6		
Urban/rural location	Urban	4429	36.4%
	Rural	7747	63.6%

Variable	Values / Description	Number	Percentage
Household food competition			
Number of household members	Mean: 6.68, Median: 6, Sd: 3.16, Range: [2, 35]		
Birth order	Mean: 2.64, Median: 2, Sd: 1.82, Range: [1, 14]		
Preceding birth interval [months]	Mean: 26.53, Median: 25, Sd: 25.39, Range: [0, 250]		
Child is twin or multiple birth	No	12037	98.9%
	Yes	139	1.1%
Water, sanitation and hygiene			
Drinking water in household	Unimproved	2164	17.8%
	Improved	6879	56.5%
	Piped	3133	25.7%
Sanitation facility in household	Unimproved	8345	68.5%
	Improved	3831	31.5%
Indoor air pollution			
Type of cooking fuel	Straw/ crop /animal dung	1969	16.2%
	Coal/ charcoal/ wood	6598	54.2%
	Kerosene	388	3.2%
	Gas/ electricity	3221	26.4%
Curative and preventive healthcare			
Vaccination index	None (0)	1093	9.0%
	Low (1-3)	2106	17.3%
	Medium (4-6)	2364	19.4%
	High (7-9)	6613	54.3%
Number of antenatal visits during pregnancy	Mean: 3.91, Median: 3, Sd: 3.44, Range: [0, 26]		
Breastfeeding practices			
Breastfeeding	No breastfeeding	1578	13.0%
	Breastfeeding + complementary feeding	9450	77.6%
	Exclusive breastfeeding	1148	9.4%
Complementary feeding practices			
Food diversity (Number of food groups consumed during last 24 hours other than breast milk)	Low (0-2)	7166	58.9%
	Medium (3-4)	3466	28.5%
	High (5-8)	1544	12.7%
Meal frequency (Number of meals consumed during last 24 hours aside from breast milk)	Low (0-1)	4145	34.0%
	Medium (2-3)	5822	47.8%
	High (4-9)	2209	18.1%
Micronutrient deficiencies			
Child received iron	No	11464	94.2%
	Yes	712	5.8%
Child received vitamin A	No	7724	63.4%
	Yes	4452	36.6%
Iodine-in-salt test result	No iodine	2447	20.1%
	Less than 15 parts per million	2775	22.8%
	15 parts per million or more	6954	57.1%

In our later regression analyses, we had to deal with the fact that meaning and effect of the feeding variables vary with increasing age (Habicht, 2004). For example, exclusive breastfeeding is recommended during the first 6 months and complementary feeding should be gradually introduced afterwards. Figure 2.6 shows the empirical relative frequencies of stunted children in our dataset depending on age and breastfeeding. It can be observed that in the first six months of age, non-breastfed children are the group with greatest stunting proportions, whereas after 14 months breastfed children have greater stunting proportions than non-breastfed children. (The peaks of exclusively breastfed children at the ages of 17, 20, and 23 months are due to very small group sizes.)

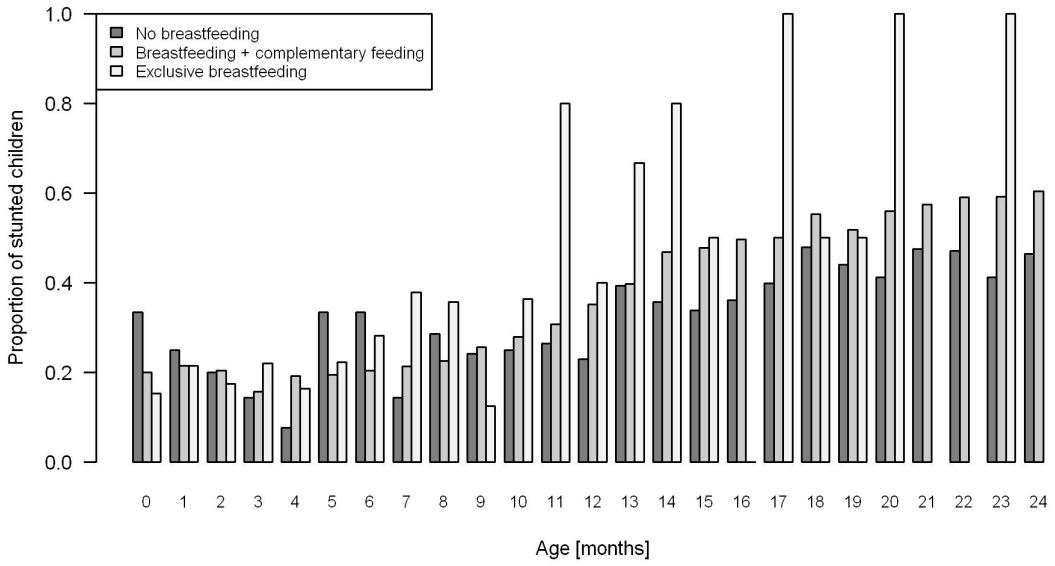


Figure 2.6 Empirical relative frequencies of stunted children depending on age and breastfeeding status.

With respect to the 29 different states of India, Figure 2.7 shows the empirical 35% Z-score quantile by region. Red areas show regions with lowest Z-score quantiles and therefore high stunting prevalences. The aim of our later regression analysis was to explain the spatial differences by other covariates included in the analysis. However, even after adjustment for these covariates, additional spatial correlation might remain which cannot be explained by the covariates. Therefore, it makes sense to assume that observations of children from the same state and from neighbouring states are more similar than observations of children from non-neighbouring states.

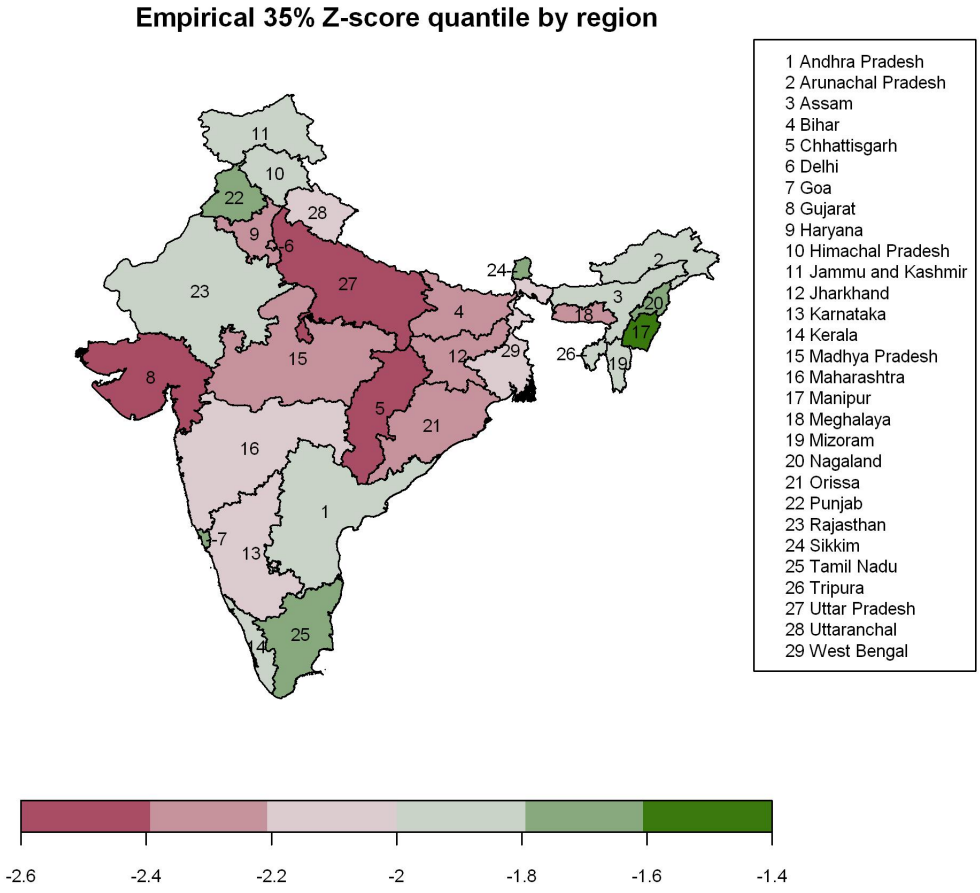


Figure 2.7 Empirical 35% Z-score quantile for height-for-age by 29 states of India.

Missing data handling

With regard to observation numbers and missing data, we pursued a complete case approach and proceeded as follows. Starting from a total of 19 868 living children aged 0-24 months, we excluded 1776 non-de jure residents (as several determinants relate to the household environment), 1053 children that were not the last birth (as detailed nutritional information is only available for the last birth) and another 2779 children due to missing outcome, resulting in a provisional total of 14 260 observations. Further reductions were mainly attributable to seven covariates with 50 or more missing values: caste (640 missing values), partner's occupation (212), partner's education (165), drinking water (50), vaccination index (280), number of antenatal visits (153), vitamin A (450), and iodine (118). Our final dataset comprised 12 176 observations.

Questions of the analysis

As already sketched, the overall epidemiological aim was a comprehensive analysis of the multiple determinants of child stunting. With the data at hand, we formulated more precise questions of the analysis as follows:

- Which variables can be identified as important determinants of child stunting?
- Is it possible to give a range of importance of the determinants?
- Which continuous variables exert their effect in a nonlinear way?
- How can the age-varying meaning of feeding variables (Figure 2.6) be adequately modelled?
- How can be accounted for the correlation between observations of children from the same state within India or from neighbouring states (Figure 2.7)?

Appropriateness of quantile regression

Most analyses of the determinants of undernutrition have used a binary outcome (e.g., stunted vs. not stunted), followed by a binary regression (see, e.g., Kyu *et al.*, 2009; Mishra and Retherford, 2007; Semba *et al.*, 2010). We believe that quantile regression for lower quantiles of the height-for-age Z-score is also a suitable and promising approach to analyze the determinants of stunting and to answer the questions above.

As described in Chapter 1, quantile regression aims at modelling conditional quantiles of the outcome depending on covariates. The underlying dichotomization of the continuous response at a pre-specified quantile parameter corresponds to defining a cut point in binary regression. However, by using quantile regression the coarsening of the outcome – and therefore discard of information – is avoided. Quantile regression is also more flexible than binary regression since no specific distribution for the outcome is assumed.

When undernutrition is the subject-matter of the analysis, lower quantiles of the height-for-age Z-score can be regarded as outcome instead of binary versions. This corresponds to the first type of usage from Chapter 1, where quantile regression is conducted for a small number of quantile parameters. In our analysis, we chose four different values for τ, namely 0.05, 0.15, 0.35, and 0.50. The values 0.35 and 0.15 were derived from the empirical relative frequencies for being stunted (approx. 37%) or severely stunted (approx. 17%) in our dataset, see Table 2.1. The other values were chosen for reasons of model comparison.

In Kandala *et al.* (2001) and Kandala *et al.* (2009), undernutrition was modelled by mean regression with Gaussian errors and a structured additive predictor. Since we are interested in the determinants of undernutrition and not in the average nutritional status, we believe that quantile regression might be more adequate for our purposes. However, since we will also employ our quantile regression models for the median, we can investigate if the association of risk factors on the lower tail of the Z-score distribution differs from their association on the population mean.

Statistical challenges for the quantile regression analysis were the combination of linear, nonlinear, spatial and age-varying effects in the same quantile-specific predictor. In addition, a large number of covariates was present in the dataset, emphasizing the need for variable selection. Altogether, this made the use of advanced quantile regression methods promising.

2.2 Overweight and obesity in western countries

Background and epidemiological aim

Obesity is currently considered almost an epidemic and has spread to children during the last decade (Kosti and Panagiotakos, 2006; Lobstein *et al.*, 2004). Childhood obesity is particularly worrying because once a child has become obese, it will likely remain obese in adulthood (e.g., Freedman *et al.*, 2005). Therefore, obese children are at high risk for severe long-term sequelae of obesity, such as hypertension, heart diseases, and diabetes mellitus. With the objective of developing effective methods of prevention, enormous public health research efforts have been made to investigate determinants of childhood overweight and obesity (Sassi *et al.*, 2009).

Apart from the non-modifiable determinants child age and sex, potential determinants of obesity that have previously been investigated (e.g., Agras and Mascola, 2005; Reilly *et al.*, 2005) include the following (non-exhaustive list):

- Socioeconomic factors: social class and education of parents
- Maternal characteristics: age at birth, BMI
- Intrauterine and perinatal factors: birth weight, maternal smoking during pregnancy, maternal weight gain in pregnancy, gestational age
- Infant feeding and dietary intake: breastfeeding, complementary feeding, food composition, parental control of feeding
- Child characteristics and lifestyle: physical activity, temperament, television viewing, computer activity, sleep
- Genetical disposition of child: ethnicity, parental obesity
- Rapid catch up growth: early adiposity rebound, weight difference in the first years

Based on this prior knowledge of potential determinants and data from a German birth cohort study, the epidemiological aim of our analysis was to investigate the impact of early childhood determinants on obesity throughout life-course. More specifically, we wanted to investigate if the effects of risk factors that were found in the literature are age-constant or age-varying, i.e., if critical age periods can be identified at which these effects emerge.

Dataset

Our analysis was based on data from the recent German birth cohort study called LISA (LISA–plus study group, 1998–2008; Rzehak *et al.*, 2009). The LISA study is a large prospective longitudinal birth cohort study in four German cities (Bad Honnef, Leipzig, München, Wesel), in which 3097 healthy neonates born between 11/1997 and 01/1999 were included. The follow-up time was until the age of ten years, and data was collected through questionnaires at ten time points covering the nine mandatory well-child check-up examinations by a pediatrician at birth and the age of around 2 weeks and 1, 3, 6, 12, 24, 48, and 60 months. For the 10-year (120 months) follow-up, anthropometric measurements were taken by physical examination at the study centres. Thus the maximum number of observations per child was ten.

Originally the LISA study was designed to determine the influence of *L*ife-style factors, environmental exposures and health-related behaviour on the development of the *I*mmune *S*ystem and the incidence of *A*llergic diseases in children. However, since information on anthropometric measurements were available, the LISA study was at the same time suited to

investigate overweight and obesity, even though not all potential determinants of interest were included.

Quantification of overweight and obesity

The body mass index (BMI) is a measure of weight-for-height and has established as the most commonly used anthropometric measure of obesity (WHO Consultation on Obesity, 1999). It is defined as follows:

$$\text{BMI}_i = \frac{\text{weight}_i\ [kg]}{\text{height}_i^2\ [m]}$$

The WHO recommends to classify adults as overweight and obese according to the following scheme:

	BMI	< 18.5	Underweight
$18.5 \leq$	BMI	< 25	Normal weight
$25 \leq$	BMI	< 30	Overweight / Pre-obesity
	BMI	> 30	Obesity

Yet, this scheme cannot be applied for children and adolescents, since height and body composition are substantially changing with child age and sex. Therefore, the classification of children as obese is usually based on reference growth charts for BMI-for-age. Contrary to child stunting, however, to date there is no a widely accepted classification scheme for obesity in childhood.

An example for BMI-for-age reference curves was shown by the cross-national WHO child growth standards in Figures 1.2 and 1.3 on page 13. Due to ethnical differences in body build and body proportions – and thus in the BMI distribution – various growth references have also been developed on national level. In Germany, the curves of Kromeyer-Hauschild *et al.* (2001) are currently the most commonly used reference curves, but come along with certain limitations which were discussed in Schaffrath Rosario *et al.* (2010).

Once the decision on a specific reference chart has been made, it is still not decided which exact cut-off values should be used for the classification of overweight and obesity. There exist different approaches which are all based on age- and sex-specific upper quantile curves of the BMI distribution. Similarly to height, these quantile curves can again be translated to Z-scores not depending on age and sex any more. However, the transformation of raw BMI values to corresponding Z-scores becomes more involved than for height due to an age-specific skewness of the BMI distribution which makes age- and sex-specific skewness parameters necessary.

In order to avoid the decision on a specific reference chart, we decided to directly analyze upper quantile curves of the BMI and chose the 90% quantile for overweight and the 97% quantile for obesity. The adjustment for age and sex was done by including these variables as covariates in the regression model.

To give a first impression of the response in our dataset, Figure 2.8 shows a traceplot for 20 randomly chosen children. It can be observed that the BMI of the majority of children increases until the age of 1 year, decreases afterwards until 6 years and then steadily increases again until the last time point at the age of 10 years.

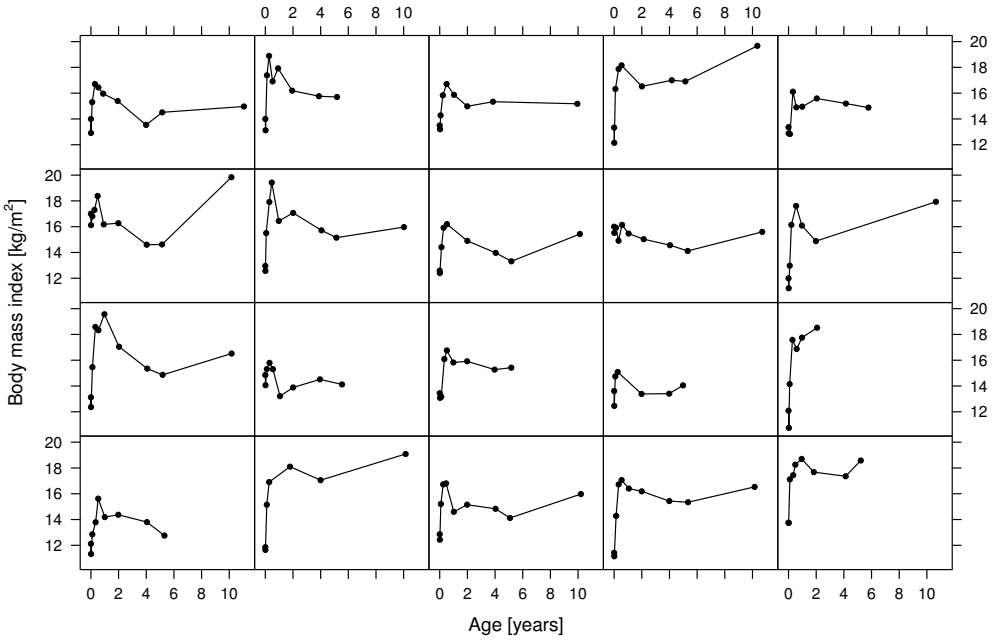


Figure 2.8 Individual BMI patterns by age of 20 randomly chosen children. Every dot denotes a single observation.

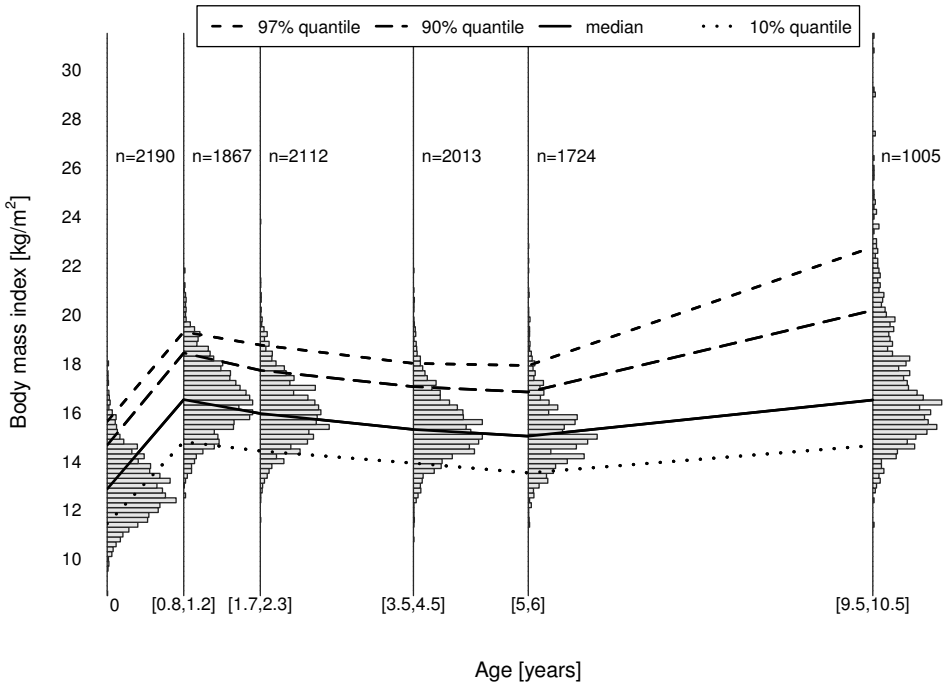


Figure 2.9 Empirical BMI distribution by age in our dataset. Relative frequencies in the histograms were calculated based on all observations within an age interval. The six age intervals are shown on the x-axis, each n denotes the total number of observations in the interval. Vertical lines are drawn at the midpoints of the intervals. Also shown are lines connecting the age-specific empirical 10%, 50%, 90%, and 97% quantiles of the BMI distribution.

In addition, Figure 2.9 displays scale and skewness of the empirical BMI distribution by age. It suggests an age-specific skewness of the BMI distribution in our dataset, beginning after the age of 6 years.

Covariates

Table 2.2 and Table 2.3 give an overview of the continuous and categorical variables, respectively, included in the analysis. Apart from child age and sex, the covariates cover various early childhood risk factors that have been discussed in the literature, such as socio-economic factors (urban/rural location, maternal education), parental overweight (maternal BMI), infant feeding (breastfeeding), and intrauterine and perinatal factors (maternal BMI gain and maternal smoking during pregnancy), which are believed to be associated with rapid postnatal growth of the offspring. All variables except for age and BMI are time-constant.

Table 2.2 Description of continuous variables for complete cases.

Variable	Abbreviation	Unit	Median	Mean	SD	N
Time-varying variables						
BMI	cBMI	kg/m^2	15.36	15.28	2.08	19 819
Age	cAge	Years	0.54	1.86	2.64	19 819
Time-constant variables						
Maternal BMI at pregnancy begin	mBMI	kg/m^2	21.72	22.59	3.76	2226
Maternal BMI gain during pregnancy	mDiffBMI	kg/m^2	4.95	5.12	1.67	2226

Table 2.3 Description of categorical variables. Absolute frequencies N relate to 2226 complete cases.

Covariate	Abbreviation	Categories	Frequency	N
Sex	cSex	0 = Female	47.8%	1064
		1 = Male	52.2%	1162
Study location	cLocation	0 = Rural (Bad Honnef, Wesel)	21.5%	478
		1 = Urban (Leipzig, Munich)	78.5%	1748
Nutrition until the age of 4 months	cBreast	0 = Bottle fed and/or breast fed	41.2%	917
		1 = Breast fed only	58.8%	1309
Maternal smoking during pregnancy	mSmoke	0 = No	85.0%	1899
		1 = Yes	15.0%	327
Maternal highest level of education	mEdu	1 = Certificate of secondary education (CSE) or Hauptschule (lower-level secondary school)	7.0%	157
		2 = Realschule (secondary school)	35.8%	798
		3 = Abitur / Fachabitur (high school diploma)	57.1%	1271

To give a first impression of potential age-varying effects of categorical risk factors, Figure 2.10 shows empirical quantile curves by age and maternal smoking during pregnancy. The differences between quantile curves with and without maternal smoking are almost zero until the age of two years for all three quantile parameters. Then the difference between respective quantile curves with and without maternal smoking continuously increases until the age of ten. At this age, the differences are clearly greater for upper quantiles than for the median.

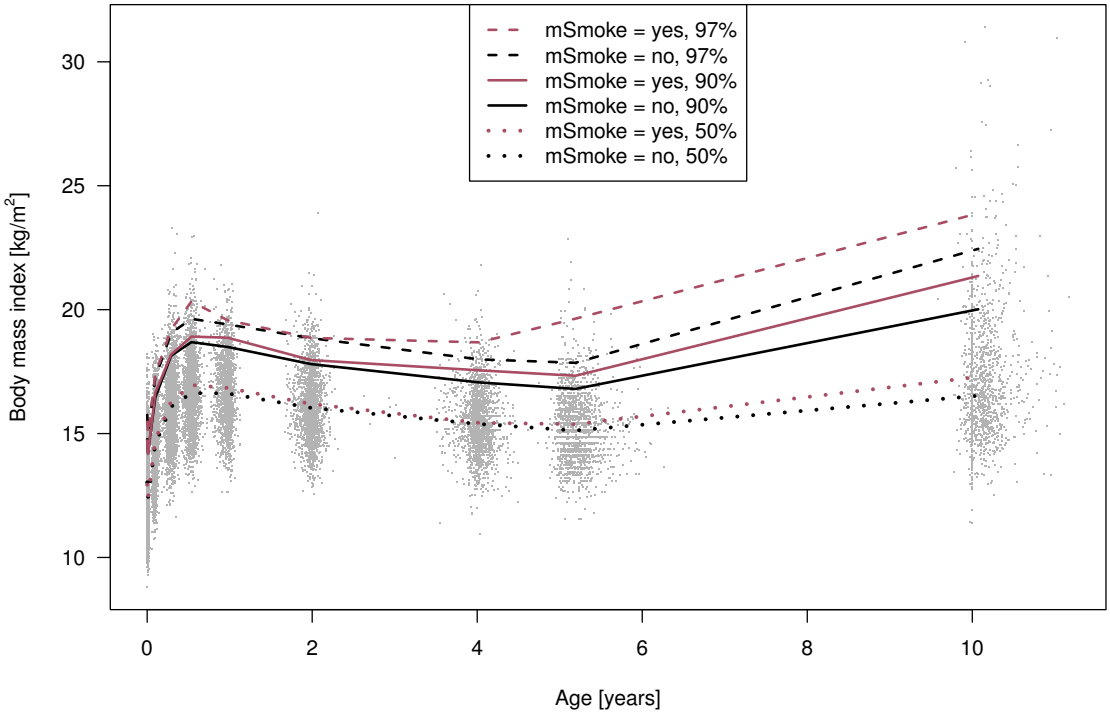


Figure 2.10 Scatterplot of all observations in our LISA dataset (grey points), superimposed by empirical 50%, 90% and 97% BMI quantiles by age and maternal smoking during pregnancy. Black lines represent children of non-smoking mothers, whereas red lines correspond to children of smoking mothers. Quantiles were calculated based on all observations at a specific time point and were connected at age medians of each time point.

Missing data handling

In our analysis, we followed a complete case approach. When an observation of a time-constant covariate was missing, we excluded all observations of the respective child from the analysis. When, on the other hand, a single observation of age or BMI was missing, only this particular observation of the respective child was excluded from the analysis. Altogether, a total of 19 819 observations from 2 226 children were included in the statistical modelling.

The decision for the complete case approach resulted from several analyses with respect to missing data and dropout, suggesting the missing data mechanism to be “missing at random” (see eSupplement of Fenske *et al.*, 2012b).

Questions of the analysis

The main objective of our obesity analysis was to flexibly model nonlinear population age curves of upper BMI quantiles that adequately reflect the shape of the BMI distribution sketched in Figure 2.9, while adjusting for individual-specific age effects and early childhood risk factors that have been discussed in the literature. At the same time, individual-specific life-course patterns of the BMI, as shown by Figure 2.8, should be reflected best as possible.

Furthermore, we wanted to investigate if potential effects of categorical risk factors are constant or varying with age. More precisely, the question was if the age-varying shape of the BMI distribution changes for different levels of the categorical covariates, as suggested by Figure 2.10 for maternal smoking.

Appropriateness of quantile regression

As for the analysis of child undernutrition, a typical statistical approach for analyzing childhood overweight and obesity would be to classify children as obese using reference charts, followed by logistic regression for the resulting binary response (e.g., Reilly *et al.*, 2005; Lamerz *et al.*, 2005). In our investigation, in contrast, we directly model upper BMI quantiles of the study population and thereby avoid possible loss of information implied by reducing the original continuous response BMI to the binary response obesity. Furthermore, binary regression models can be formulated by a threshold approach for a latent continuous variable with a specific (and often symmetric) distributional assumption (see, e.g., Fahrmeir and Tutz, 2001). For example, logit models assume a logistic distribution and probit models assume a standard Gaussian distribution for the latent variable, which are both symmetric. Consequently, age-specific skewness of BMI distributions makes the use of conventional logit and probit models questionable.

For cross-sectional BMI data, quantile regression methods have been used to model a Z-score of the BMI-for-age depending on covariates (Beyerlein *et al.*, 2008, 2010), which was obtained by transforming raw BMI values based on age- and sex-specific reference charts. Here, we directly model raw BMI quantiles and include age and sex as covariates and thereby avoid the decision for a specific reference chart.

The particular statistical challenge of the present analysis was to apply quantile regression with a flexible predictor (which was also the task in the analysis of undernutrition in Section 2.1) and, at the same time, to account for the longitudinal data structure by modelling the intra-individual correlation between repeated observations of the same child.

Chapter 3: Structured additive quantile regression – model class and estimation

In this chapter, we define the model class of structured additive quantile regression – which we abbreviate with *STAQ* in the following – and give an overview of different distribution-free and distribution-based estimation approaches for this model class. We also treat three model classes that are closely related to STAQ models. Currently, none of the estimation approaches outperform another in all respects. Therefore, we discuss the advantages and shortcomings of each approach regarding selected criteria of method assessment. Altogether, this overview chapter should motivate the need to develop further estimation approaches for STAQ models.

3.1 Generic model class

We formulate the model of structured additive quantile regression (STAQ) in accordance with Fahrmeir *et al.* (2004); Kneib *et al.* (2009) and Fenske *et al.* (2011).

Assume that we have data $(y_s, \boldsymbol{z}_s)$ where y_s denotes the continuous response and $\boldsymbol{z}_s$ the vector containing all covariate information for observation s (with generic observation index s). Then, the relationship between quantile function of the true underlying Y_s and a quantile-specific predictor $\eta_{\tau s}$ can be written as:

$$Q_{Y_s}(\tau|\eta_{\tau s}) = \eta_{\tau s}(\boldsymbol{z}_s) \; . \tag{3.1}$$

This notation is similar to the linear quantile regression model (1.3) on page 8 but the predictor $\eta_{\tau s}$ is extended to more flexible model terms here (see below). The underlying assumption on the error terms remains the same as in equation (1.2) on page 8, i.e., $F^{-1}_{\varepsilon_{\tau s}}(\tau) = 0$. To ease notation, we suppress the quantile parameter τ in the following but keep in mind that parameters may be quantile-specific. This might also be the case for the design of the predictor and the set of covariates, in particular when the estimation is performed separately for different quantile parameters.

Letting the quantile parameter τ apart, the generic structured additive predictor can be expressed as

$$\eta_s = \beta_0 + \sum_{d=1}^{D} h_d(\boldsymbol{z}_s) \; , \tag{3.2}$$

where β_0 is an intercept and h_d, $d = 1, \ldots, D$, are generic functions that allow for the inclusion of a large variety of different model components. Each of these functions depend on (usually small) subsets of covariates from $\boldsymbol{z}_s$. In the following description of h_d, let z_{sk} and z_{sl} denote any two univariate elements of the covariate vector, i.e., $\boldsymbol{z}_s = (\ldots, z_{sk}, \ldots, z_{sl}, \ldots)^\top$ for observation s. Depending on the domain of h_d, these univariate covariates z_k and z_l may be continuous or categorical and may include (continuous or categorical) spatial or cluster information.

The following model terms are possible for the generic functions h_d:

- *Linear components:* $h_d(\boldsymbol{z}_s) = \beta_d z_{sk}$

  with linear regression parameter β_d for a categorical or continuous covariate z_k

- *Smooth nonlinear components:* $h_d(\boldsymbol{z}_s) = f_d(z_{sk})$

  with continuous covariate z_k and smooth, potentially nonlinear function f_d that is not specified in advance

- *Varying coefficient terms:* $h_d(\boldsymbol{z}_s) = z_{sk} \cdot f_d(z_{sl})$

  with categorical or continuous covariate z_k and smooth function f_d of a continuous covariate z_l. Thus, the effect of covariate z_l varies smoothly over the domain of z_k according to the function f_d.

- *Bivariate surfaces:* $h_d(\boldsymbol{z}_s) = f_d(z_{sk}, z_{sl})$

  with smooth bivariate function f_d of two continuous covariates z_k and z_l. In case that z_k and z_l denote longitude and latitude of spatially oriented data, the surface corresponds to a spatial effect.

- *Discrete spatial components:* $h_d(\boldsymbol{z}_s) = f_d(z_{sk})$

  with a categorical covariate z_k containing discrete spatial information, e.g., the region within a country, and a function f_d with spatial effects accounting for the neighbourhood structure

- *Cluster-specific components:* $h_d(\boldsymbol{z}_s) = z_{sl} \cdot \left([I(z_{sk} \in G_1), \ldots, I(z_{sk} \in G_K)]^\top \boldsymbol{\gamma}_d\right)$

  with indicator function $I(\cdot)$, categorical or continuous covariate z_l, categorical covariate z_k with K different groups or clusters $G_1, \ldots, G_K$ and a $(K \times 1)$-vector $\boldsymbol{\gamma}_d$ containing cluster-specific parameters. Thus, the effect of z_l differs across groups $G_1, \ldots, G_K$ defined by the grouping factor z_k. An example would be an individual-specific intercept in longitudinal data, where z_l would correspond to the unit vector, z_k to the ID-variable and $\boldsymbol{\gamma}_d$ to the vector of individual-specific (random) intercepts.
  Instead of the above notation with $\boldsymbol{\gamma}_d$, an alternative would be to define a vector $\tilde{\boldsymbol{\gamma}}_d$ containing the cluster-specific parameters per observation. Then the generic function could for example be written as $h_d(\boldsymbol{z}_s) = z_{sl} \cdot \tilde{\gamma}_{sd}$, where the observation index s for $\tilde{\gamma}_{sd}$ underlines the special concept of observation- or individual-specific parameters.

A few remarks regarding our generic notation should be given here. First, the same covariate can of course be included in more than one model component. For a categorical covariate, for example, a linear effect (= main effect) can be estimated together with a varying coefficient term (= interaction effect) according to another continuous covariate.

Second, the generic model notation in Kneib *et al.* (2009) without indices s and d might look simpler at first glance. Here, we explicitly keep the index s to emphasize which covariate values (and parameters) are observation-specific, and we use the generic index d for component-specific unknown parameters and functions that have to be estimated.

Furthermore, except for the cluster-specific components which might contain individual-specific effects, all other components model population effects. Individual-specific effects are only addressed when the cluster variable defines individuals. In classical mixed models for longitudinal data, they are typically assumed to be Gaussian and are therefore called random effects.

Finally one should keep in mind that similar to the predictor being quantile-specific, the interpretation of the various population effects described above is related to the population quantiles of the response, as was illustrated in Chapter 1.

In our applications, we will consider the following two instances of the generic model class:

1. For cross-sectional data, as will be the case in our investigation of child stunting in India, the observation index corresponds to $s = i$ with $i = 1, \ldots, n$, denoting the individual. The structured additive predictor simplifies to

$$\begin{aligned}\eta_i = \beta_0 &+ \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ &+ f_0(t_i) + f_1(z_{i1}) + \ldots + f_q(z_{iq}) \\ &+ v_{i1} \cdot f_{int,1}(t_i) + \ldots + v_{ir} \cdot f_{int,r}(t_i) \\ &+ f_{spat}(w_i) \ ,\end{aligned}$$

   and comprises linear components for p covariates $x_1, \ldots, x_p$, smooth nonlinear components for the time variable t and q continuous covariates $z_1, \ldots, z_q$, varying coefficient terms for r covariates $v_1, \ldots, v_r$ whose effects vary smoothly over time according to a continuous time variable t, and a smooth spatial effect of a spatial categorical covariate w.

   In comparison to the generic model notation above, the different labelling of the covariates here makes it easier to distinguish different types of covariates and components. Note that the same covariate can again be included in more than one component.

2. In case of longitudinal data, as present in our investigation of obesity of children in Germany, the observation index corresponds to $s = (i, j)$ with $i = 1, \ldots, N$, denoting the individual and $j = 1, \ldots, n_i$, standing for the j-th observation of individual i with individual-specific observation numbers n_i. The structured additive predictor can be written as

$$\begin{aligned}\eta_{ij} = \beta_0 &+ \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} \\ &+ f_0(t_{ij}) + f_1(z_{ij1}) + \ldots + f_q(z_{ijq}) \\ &+ v_{i1} \cdot f_{int,1}(t_{ij}) + \ldots + v_{ir} \cdot f_{int,r}(t_{ij}) \\ &+ \gamma_{i1} + \gamma_{i2} \cdot t_{ij} \ ,\end{aligned}$$

   and again contains linear components for p time-constant or time-varying covariates $x_1, \ldots, x_p$, smooth nonlinear components for the time scale t and q continuous covariates $z_1, \ldots, z_q$, time-varying effects of r time-constant covariates $v_1, \ldots, v_r$ with covariate t denoting the time scale, an individual-specific intercept γ_{i1} and an individual-specific slope γ_{i2} for the time scale.

3.2 Estimation approaches – outline and assessment

Due to the growing popularity of quantile regression, a large variety of estimation approaches for this model class have been developed over the last years. Most of them concentrate on specific models and data structures and consider only a few components of the generic predictor introduced in Section 3.1. Thus, it is almost impossible to give an exhaustive overview of all existing estimation approaches and to find a proper and unambiguous classification for them.

Nevertheless, in the remaining chapter we try to give an overview of the most important and frequently used approaches with potential relevance for our work. We gathered recent literature on quantile regression and arrived at the classification into distribution-free estimation approaches, distribution-based approaches and related model classes. Of course this classification is not the only reasonable structure.

Distribution-free approaches on the one hand include all approaches which aim at direct minimization of the quantile loss criterion by linear programming methods; we denote them with *classical framework of quantile regression*. On the other hand, distribution-free approaches comprise recent machine learning and statistical learning procedures that only ask for the specification of a loss function and have been used for quantile regression recently, such as quantile regression forests, quantile neural networks and kernel-based quantile regression with support vector machines. Our own approach of quantile boosting belongs to this category and will be described in detail in Chapter 4.

Distribution-based approaches rely on a specific distributional assumption on the error terms and thereby allow for likelihood-based or Bayesian estimation. We describe typical approaches based on the asymmetric Laplace distribution as well as approaches based on flexible mixture distributions which are well-suited for Bayesian estimation.

Regarding related model classes, we consider expectile regression as distribution-free related model class. As distribution-based related model classes, we describe Gaussian STAR and GAMLSS models with respect to quantile regression.

Currently, none of the estimation approaches outperforms another in all respects. To allow for a comparison with regard to the respective application potential, we define the following criteria for assessment of the estimation approaches:

- **Flexible predictor:** How flexible is the structured additive predictor? Which different components of the generic predictor in (3.2) can be included?
- **Estimator properties and inference:** What can be said about the properties of the estimated parameters, e.g., about bias, consistency and asymptotic distribution? In particular, can (asymptotic) standard errors be obtained and therefore uncertainty about the estimated parameters be quantified? Can quantile crossing occur or is it implicitly prevented by the estimation procedure?
- **Variable selection:** Is variable and model selection possible, i.e., can variables be excluded from the model in order to avoid overfitting of the data and to produce sparse models which contain the most relevant covariates only? Can a range for variable importance be given?
- **Software:** Is software available, preferably as package for the statistical software R (R Development Core Team, 2012)?

For some approaches we state further specific advantages and shortcomings in addition to these criteria, which for example relate to the flexibility of the distributional assumption or longitudinal data.

Note also that we prefer the term *distribution-free approach* to *nonparametric approach* since the latter is frequently used ambiguously. We only use the term *nonparametric* for concepts with parameter estimators that are not of primary interest, for example in the context of flexible Bayesian quantile regression.

3.3 Distribution-free estimation

3.3.1 Classical framework of quantile regression

The classical framework of quantile regression is inextricably linked to Roger Koenker. This can on the one hand be attributed to his recent book (Koenker, 2005) which gives a thorough introduction to quantile regression, and on the other hand to his invaluable contributions on the field of quantile regression since its very beginnings in 1978. In Koenker and Bassett (1978), the τ-th regression quantile was introduced as any solution to the following minimization problem:

$$\min_{\boldsymbol{\beta}_\tau \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau) \; . \tag{3.3}$$

In this criterion, all elements are defined as in the linear quantile regression model (1.3) on page 8 and $\rho_\tau(u)$ denotes the *check function*

$$\rho_\tau(u) = \begin{cases} u \cdot \tau & u \geq 0 \\ u \cdot (\tau - 1) & u < 0 \; , \end{cases} \tag{3.4}$$

which is the suitable loss function for quantile regression displayed by Figure 3.1. For $\tau = 0.5$, the check function is proportional to the absolute value function, i.e., $\rho_{0.5}(u) = 0.5 \cdot |u|$, which is well known for being the suitable loss function for median regression. In case that no covariates are present besides an intercept, minimization of (3.3) leads to the empirical $\tau \cdot 100\%$ quantile of the response as estimator for $\hat{\beta}_{\tau 0}$.

The criterion in (3.3) corresponds to the empirical version of an expected loss criterion. It can be formulated as a set of linear constraints, and therefore its minimization can be conducted by linear programming methods, see Koenker (2005) for an explicit formulation of the problem as linear programs and further references on suitable algorithms. Even though no closed-form solution for $\hat{\boldsymbol{\beta}}_\tau$ can be derived, the resulting quantile regression estimators $\hat{\boldsymbol{\beta}}_\tau$ provide useful properties with regard to equivariance, robustness and asymptotics (see Koenker, 2005, Chap. 2).

In this thesis, we perceive the classical quantile regression framework as consisting of all estimation approaches which aim at direct minimization of the quantile regression loss criterion (also addressing more flexible predictors $\eta_{\tau i}$ instead of the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau$) by linear programming methods. We consider these approaches from the classical framework with respect to our pre-defined criteria in the following.

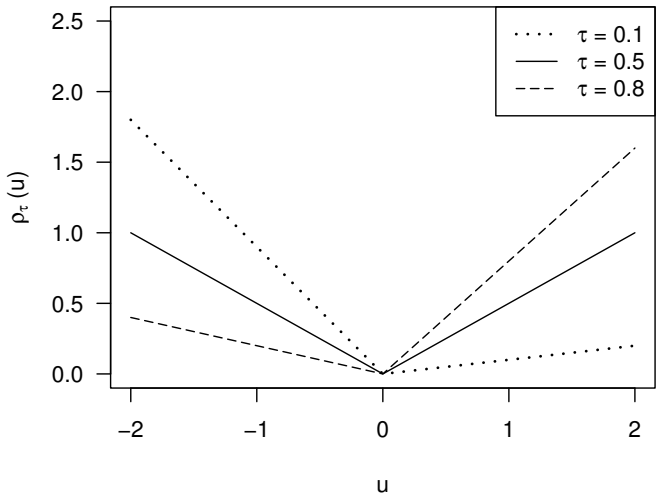


Figure 3.1 Check function: Suitable loss function for quantile regression.

Flexible predictor

Since its original introduction in 1978, various extensions of the linear quantile regression model in (3.3) have been made towards a more flexible predictor. However, to date there has not been a classical estimation approach for the generic predictor presented in (3.2) in Section 3.1, and it is in particular not yet possible to combine smooth nonlinear population effects and individual-specific effects in the same predictor with estimation based on linear programming.

Extensive consideration has been given to additive quantile regression models with nonlinear effects of continuous covariates, resulting in three main concepts based on linear programming: quantile smoothing splines, quantile regression using P-splines and local polynomial quantile regression.

Quantile smoothing splines introduced in Koenker *et al.* (1994) were one of the first attempts to estimate smooth nonlinear functions in additive models. In this approach, the minimization problem in (3.3) is extended by a total variation regularization penalty on the potentially nonlinear functions. For a univariate situation ($q = 1$) with only one continuous covariate z and a smooth functional effect $f_z(\cdot)$ to be estimated, the minimization problem in (3.3) is written as

$$\min_{f_\tau} \left[\sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(z_i)) + \lambda V(f'_\tau) \right] , \tag{3.5}$$

where $V(f'_\tau)$ denotes the total variation of the derivative $f'_\tau : [a, b] \rightarrow \mathbb{R}$ defined as $V(f'_\tau) = \sup \sum_{i=1}^{n-1} |f'_\tau(z_{i+1}) - f'_\tau(z_i)|$ with the $\sup$ taken over all partitions $a \leq z_1 < \ldots < z_n < b$. The tuning parameter $\lambda > 0$ controls the smoothness of the estimated function. Small values of λ lead to wiggly functions, whereas large values of λ lead to smooth functions, with $\lambda \rightarrow \infty$ being the most extreme yielding a linear function for $\hat{f}_\tau(\cdot)$. Koenker *et al.* (1994) showed that the solution can still be obtained by linear programming and that the resulting estimated function is a piecewise linear spline function with knots at the observations. The total variation regularization approach was also applied for bivariate smoothing with penalized triograms in Koenker and Mizera (2004).

Another approach that is closer related to the further work in this thesis is additive quantile regression based on P-splines introduced by Bollaerts *et al.* (2006). In analogy to P-spline estimation for mean regression described in Eilers and Marx (1996), a L_1-norm smoothness penalty based on differences in the coefficients of adjacent B-spline basis functions is added to the quantile minimization criterion, which can in the univariate case be formulated as:

$$\min_{\beta_1,...,\beta_J} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \sum_{j=1}^{J} \beta_j B_j(z_i)) + \lambda \sum_{j=d+1}^{J} |\Delta^d \beta_j| \right] .$$

Here, $B_j(\cdot)$ denote B-spline basis functions of a fixed degree, β_j are the coefficients, λ is again a smoothness parameter and Δ^d are the d-th order differences, that is, $\Delta^d \beta_j = \Delta^1(\Delta^{d-1}\beta_j)$ and $\Delta^1 \beta_j = \beta_j - \beta_{j-1}$. Bollaerts *et al.* (2006) described a linear programming algorithm for the above minimization problem. B-spline basis functions without penalization were also suggested for the estimation of (partially linear) varying coefficient models in Kim (2007) and Wang *et al.* (2009). However, without smoothness penalty term one always has to deal with the question how to determine the number and positions of knots adequately.

The third alternative for estimating nonlinear effects are local polynomial methods, with local linear quantile regression being the simplest case (Yu and Jones, 1998). Thereby, the minimization criterion is multiplied by kernel-based weights:

$$\min_{\beta_{\tau 0},\beta_{\tau 1}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_{\tau 0} - \beta_{\tau 1}(z_i - z)) \cdot K\left(\frac{z_i - z}{h}\right) ,$$

where $K(\cdot)$ is a kernel function with bandwidth h and z is a fixed covariate value for which an estimator $f_\tau(z)$ is sought by estimating $\beta_{\tau 0}$ and $\beta_{\tau 1}$ for a grid of values for z. This problem can be solved by weighted linear quantile regression based on linear programming methods. Since the origins in 1998, a lot of research has been made to extend this approach to more than one continuous covariate in the predictor, with the typical challenges of these nonparametric approaches to avoid the curse of dimensionality and to answer questions on the choice of kernel function and optimal bandwidth. As a result, additive quantile regression models with local kernel-based estimation have been suggested in De Gooijer and Zerom (2003), Yu and Lu (2004), Horowitz and Lee (2005), and Cheng *et al.* (2011). Closely related to our application, a local polynomial kernel-based estimator was recently studied for the construction of reference charts in Li *et al.* (2010).

With regard to individual-specific effects, the first attempt of a quantile regression model for longitudinal data traces back to Koenker (2004), who modelled an individual-specific location shift by adding individual-specific fixed intercepts γ_i for $i = 1, ..., N$ to the linear predictor:

$$Q_{Y_{ij}}(\tau|\boldsymbol{x}_{ij}, \gamma_i) = \gamma_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_\tau .$$

The corresponding minimization problem was extended by a L_1-norm penalty term on the individual-specific intercepts (which was justified by the analogy between random effects and L_2-norm penalization in linear mixed models) and minimized for a grid of quantile parameters $\tau_k, k = 1, ..., K$, simultaneously:

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{n_i} w_k \rho_{\tau_k}(y_i - \gamma_i - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_{\tau_k}) + \lambda \sum_{i=1}^{n} |\gamma_i| . \tag{3.6}$$

This led to shrunken individual-specific (fixed) effects γ_i with the amount of shrinkage controlled by $\lambda > 0$. Because of the L_1-norm penalty term, the minimization problem could still be formulated as a set of linear constraints and solved by linear programming. Since this first model a lot of work has been done to develop suitable concepts for quantile regression with longitudinal data, see, e.g., Karlsson (2007); Galvao and Montes-Rojas (2010). Various approaches for longitudinal quantile regression also rely on the assumption of asymmetric Laplace distributed errors and will be sketched in Section 3.4.1.

Estimator properties and inference

For the linear quantile regression model, Koenker (2005) showed that the parameter estimators are asymptotically unbiased and Gaussian distributed (in addition to the already mentioned results regarding equivariance and robustness). Similar asymptotic results were obtained for models with more flexible predictors, e.g., in Koenker (2004). The asymptotic covariance matrix of the quantile regression estimator $\hat{\boldsymbol{\beta}}_\tau$ can be written as a sandwich matrix which depends on the true error density (see Koenker, 2005, Theorem 4.1, p.120). Therefore, to obtain standard errors for $\hat{\boldsymbol{\beta}}_\tau$ one is faced with the problem of estimating the true error density which somehow erodes the distribution-free character of quantile regression. Kocherginsky *et al.* (2005) compared various different approaches that have been developed for the estimation of the asymptotic covariance matrix, including resampling methods such as the bootstrap, and give recommendations on which estimation method to use in practical situations. It turns out that bootstrap methods give most reliable estimations in almost all situations.

Since the estimation is performed separately for different quantile parameters (except for the longitudinal quantile regression model in (3.6)), quantile crossing is not prevented by the above approaches.

Variable selection

With the aim of variable selection in the quantile regression model, Koenker (2005) proposed to modify the Akaike information criterion (AIC) and the Schwarz information criterion (SIC) by replacing the usual log-likelihood term by the empirical risk. For example, with p model parameters the adapted AIC is

$$\mathsf{AIC}(\tau) = -2 \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau) \right) + 2\, p\,. \tag{3.7}$$

With a redefined version of the degrees of freedom in the second model term, the adapted SIC can also be used for the choice of the smoothing parameter λ in additive quantile regression models. Variable selection with these criteria is certainly a challenge when the number of parameters is large.

Li and Zhu (2008) paved the way for bringing Lasso-like concepts into quantile regression. They imposed a L_1-norm penalty term on the linear quantile regression coefficients and presented a linear programming algorithm for the modified minimization problem which is closely related to the approach for the linear quantile regression problem. This approach can be regarded as an important contribution to variable selection in quantile regression since the L_1-norm penalty not only causes shrinkage of the fitted coefficients towards zero but also forces some of the fitted coefficients to be exactly zero (given sufficiently large smoothness parameters). This variable

exclusion property is for example not given for the L_2-norm penalty typically used in mean regression. Note that the L_1-norm penalty for the linear coefficients can easily be combined with total variation regularization for the nonlinear effects, as has recently been done in Koenker (2011).

Software

The classical framework of quantile regression is implemented in the package `quantreg` (Koenker, 2012) in R (R Development Core Team, 2012). Linear quantile regression can be conducted with the function `rq()`, which amongst others provides various options for estimating the asymptotic covariance matrix. For additive quantile regression with nonlinear effects, an implementation of total variation regularization is available in the function `rqss()`. For quantile regression with longitudinal data, the package `rqpd` (Koenker and Bache, 2012) is currently under development but already available on the central platform for the development of R packages called R-Forge (Theußl and Zeileis, 2009).

3.3.2 Statistical learning and machine learning approaches

Over the last years various computer-intensive estimation procedures which originate from machine learning and statistical learning algorithms have been utilized for quantile regression. These approaches are completely distribution-free but do not always address a particular covariate structure as considered by the STAQ predictor in (3.2) on page 35.

In the following, we shortly describe three approaches that have been suggested to model the quantile function of a response variable depending on a (potentially large) number of covariates, that is quantile regression forests, quantile regression neural networks and kernel-based quantile regression using support vector machines. Since a detailed description of these highly complex and very different concepts would go beyond the scope of this thesis, we here just touch on them and refer to the corresponding literature for further reading.

Note that boosting, which will be treated in detail in Chapter 4, also belongs to the present class of distribution-free, computer-intensive estimation approaches and even allows for a structured additive predictor to be modelled. Boosting can be rated as a *statistical learning* algorithm since it incorporates two competing goals of learning from data: prediction (which is the main goal in the machine learning community) and interpretation (which is an additional main goal in the statistical community).

Quantile regression forests

Chaudhuri and Loh (2002) made one of the first attempts to use tree-based methods for estimating conditional quantile functions. Few years later quantile regression forests were introduced by Meinshausen (2006) as an extension of random forests (Breiman, 2001) to quantile regression. The aim of quantile regression forests is to estimate the cumulative distribution function (cdf) of a response variable conditional on covariates without imposing any structure on their relationship. To achieve this aim, an ensemble of regression trees is grown similar to random forests as follows. First, a large number of bootstrap samples of the training data is drawn. Then, for every single bootstrap sample a random subset of the covariates is drawn and a regression tree is grown. The size of this random subset m_{try} is the only tuning parameter of the algorithm and is typically

selected based on a test dataset. The conditional cdf for a new observation (vector) $X = x$ is estimated by:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) I(y_i \leq y) \ ,$$

where y_i denotes the response, $I(\cdot)$ an indicator function and $w_i(x)$ stand for weights of the original observations $i = 1, \ldots, n$, which are calculated by dropping x down all trees. More specifically, for each single tree the observations which share the leaf with the new x get non-zero, uniformly distributed weights. The resulting weights $w_i(x)$ are the averages over these observation-specific weights from all trees. The quantile function is finally obtained by inverting the cdf. The main difference of quantile regression forests to the original random forests algorithm is that one takes note of all observations in each leaf and not only of the mean, which allows to estimate the whole conditional cdf for a new observation (vector) $X = x$ as described above. In addition, Meinshausen (2006) gives a proof for the consistency of the cdf estimated in this manner.

It is difficult to rate quantile regression forests regarding our criteria for model assessment since they do not address a **flexible predictor** and can rather be seen as black box estimators. Therefore, it is not possible to explicitly quantify the relationships between covariates and response and to obtain **inference** results for single estimators. In random forests **variable selection** is possible by applying variable importance measures, see for example Strobl *et al.* (2008), and it could be a matter of future research how these measures can be adapted for quantile regression forests. The main advantages of quantile regression forests are their applicability for high-dimensional data and their implicit prevention of quantile crossing by estimating the full conditional cdf in one step. Since random forests typically perform well in prediction settings, Meinshausen (2006) suggested to apply quantile regression forests for the construction of prediction intervals for new observations, as was already sketched in alternative 1 for the usage of quantile regression in Section 1.2. **Software** for fitting quantile regression forests is available in the R package `quantregForest` (Meinshausen, 2012).

Quantile regression neural networks

Taylor (2000) introduced quantile regression neural networks as another computer-intensive algorithm which is well suited for prediction and forecasting. The standard approach of artificial neural networks provides a general concept for fitting nonlinear high-dimensional regression models based on the minimization of a loss criterion. Quantile regression is performed when the check function is inserted as a special loss function in the standard algorithm.

Since neural networks rely on gradient-based nonlinear optimization, they theoretically require a loss function which is differentiable everywhere. Due to its kink point at zero this is not fulfilled for the check function, however, and it is not clear if convergence problems might occur when applying the standard optimization algorithm for neural networks (Taylor, 2000). As a solution Cannon (2011) replaced the check function by a differentiable loss function – an approximation which had first been suggested by Chen (2007) – and adapted the quantile regression neural network algorithm of Taylor (2000) accordingly.

Assessing quantile regression neural networks regarding our criteria is as difficult as assessing quantile regression forests since neural networks just provide black box estimators without giving a detailed structure of the effects of single covariates. Therefore neither is a **flexible predictor**

addressed nor can results on **inference** of single parameters and **variable selection** be obtained. The use of quantile neural networks makes sense when predictions or predictive densities are of interest, as demonstrated in the example of Cannon (2011), where daily precipitation amounts were forecasted. However, the danger of quantile crossing is incurred. Concerning **software**, quantile regression neural networks are implemented in the R package `qrnn` (Cannon, 2011).

Kernel-based quantile regression

Another class of completely distribution-free estimation approaches for quantile regression originates from the powerful framework of support vector machines (SVMs). The generic structure of empirical SVMs was described in Christmann and Hable (2012): the aim is to minimize an empirical loss criterion based on a convex loss function between response variable and an unspecified regression function of the covariates. This regression function is assumed to belong to a reproducing kernel Hilbert space (RKHS) and is penalized by a suitable RKHS norm penalty to avoid overfitting and ensure existence.

Takeuchi *et al.* (2006) directly started from the check function as loss function (which is called *pinball loss function* in the machine learning community), whereas Christmann and Hable (2012) formulated the minimization problem of empirical SVMs in a general way and considered the check function as one special instance which leads to quantile regression.

Regarding a **flexible predictor**, no structure is assumed for the covariate predictor and therefore for the relationship between covariates and response in the general formulation of empirical SVMs. However, it is possible to impose a structure by choosing suitable kernel functions for different covariates. For example, Christmann and Hable (2012) considered an additive model with smooth nonlinear functions of continuous covariates. This model covers some components of the generic predictor in (3.2). A crucial tuning parameter of these algorithms is the regularization parameter λ which can for example be chosen by cross-validation. Results on **estimator properties and inference** have recently been obtained by Christmann and Hable (2012) showing consistency of the SVM estimators. In addition, asymptotic confidence sets for the estimators, which can deliver pointwise asymptotical confidence intervals, were derived by Hable (2012). Quantile crossing can occur due to the separate regression fits for different quantile parameters. With regard to **software**, kernel-based quantile regression with SVMs can be fitted by the function `kqr()` from the R package `kernlab` (Karatzoglou *et al.*, 2004).

3.4 Distribution-based estimation

3.4.1 Asymmetric Laplace distribution approaches

The majority of approaches for distribution-based estimation of quantile regression rely on the *asymmetric Laplace (ASL) distribution*. According to Yu and Zhang (2005), a random variable $Y \in \mathbb{R}$ follows an asymmetric Laplace distribution, i.e., $Y \sim \mathcal{ASL}(\mu, \sigma, \tau)$, if the corresponding density function can be expressed as

$$f_Y(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\} ,$$

where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, and $\tau \in (0, 1)$ is a parameter responsible for skewness (and kurtosis) of the density. Figure 3.2 shows the asymmetric Laplace density for different parameter combinations. One can observe that the density becomes left-skewed for $\tau > 0.5$ and right-skewed for $\tau < 0.5$, while it corresponds to the special case of a double-exponentially distributed random variable for $\tau = 0.5$. For increasing values of σ, the variation of Y increases and the tails of the density become heavier.

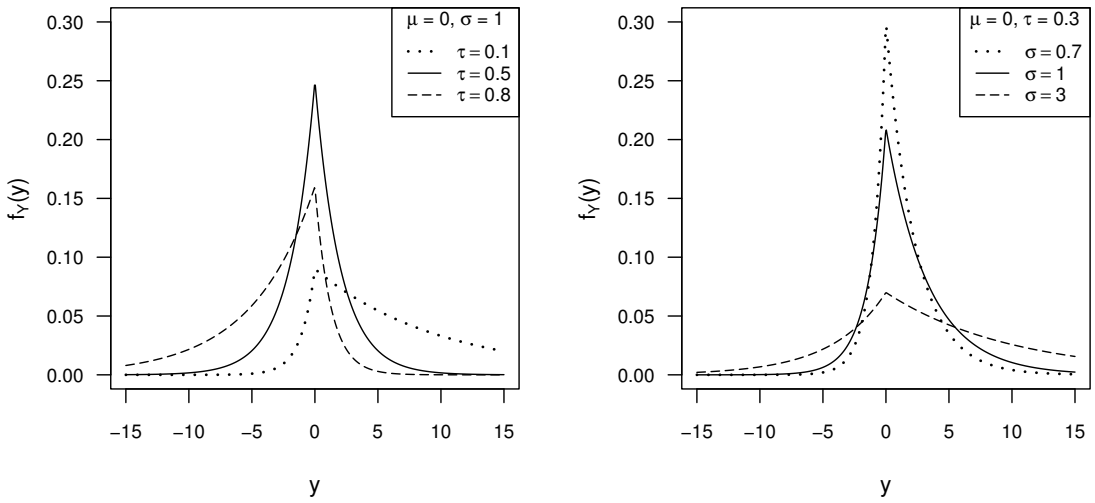


Figure 3.2 Density function of $Y \sim \mathcal{ASL}(\mu, \sigma, \tau)$ depending on skewness parameter τ (left plot) and scale parameter σ (right plot).

Furthermore, the probability mass to the left of μ is exactly τ, i.e., $F_Y(\mu) = P(Y \leq \mu) = \tau$, which means that the location parameter μ corresponds to the $\tau \cdot 100\%$ quantile of Y. This important property of the ASL distribution makes it suitable for quantile regression. Remember that the only assumption that was made for the error distribution in quantile regression models is $F_{\varepsilon_{\tau i}}(0) = \tau$, see equation (1.2) on page 8, which is directly fulfilled for the ASL distribution with $\mu = 0$ and $\tau \in (0, 1)$ being the fixed quantile parameter. Under the assumption of *iid* ASL distributed errors, the likelihood function of the linear quantile regression model can be written as

$$L(\boldsymbol{\beta}_\tau, \sigma|\boldsymbol{y}, \boldsymbol{x}, \tau) \propto \frac{1}{\sigma^n} \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau\right)\right\} .$$

When setting $\sigma = 1$, one can see that maximizing the likelihood function with respect to $\boldsymbol{\beta}_\tau$ is equivalent to minimizing the loss criterion of the classical quantile regression approach in (3.3) on

page 39. Thus, the classical estimator $\hat{\boldsymbol{\beta}}_\tau$ obtained by linear programming can also be regarded as a likelihood-based quantile regression estimator. This corresponds to the analogy of the least squares and the maximum likelihood estimator based on Gaussian error terms in classical linear models.

At first glance this explicit distributional assumption seems to contradict the distribution-free character of quantile regression. Yet, assuming an ASL distribution for the error terms offers several advantages. Most importantly the presence of a likelihood function paves the way for likelihood-based and Bayesian estimation approaches.

Note, however, that the assumption of an ASL distribution should rather be rated as a quasi-likelihood than as a proper likelihood approach since the choice of the distribution is based on the aim of employing quantile regression and not on the specific data character. When assuming an ASL density, the true error density is misspecified in most cases since the distributional shape of the data does not resemble the specific shape of the ASL density (and does not fundamentally change its shape for different quantile parameters τ as is the case for the ASL density).

We discuss the existing approaches based on the ASL distribution regarding our criteria in the following.

Flexible predictor

Bayesian quantile regression using the ASL distribution was suggested for the first time in Yu and Moyeed (2001). This approach addresses the linear quantile regression model and relies on a Metropolis-Hastings algorithm with independent improper uniform priors for the regression parameters $\boldsymbol{\beta}_\tau$. With the similar aim of developing a Bayesian algorithm for the linear quantile regression model, Tsionas (2003) used an alternative representation of the ASL density as a scale mixture of Gaussian densities and proposed an efficient Gibbs sampling algorithm for the estimation. By also estimating the scale parameter σ (through suitable reparametrization), the shape of the underlying ASL density is more flexible compared to the original approach by Yu and Moyeed (2001) with $\sigma = 1$ being fixed.

One of the first extensions towards a more flexible predictor was suggested in the framework of local linear methods. In order to estimate a nonlinear effect, Jones and Yu (2007) applied the ASL likelihood instead of the empirical loss function to improve the original double kernel local linear quantile regression from Yu and Jones (1998).

For longitudinal quantile regression, various models have been proposed with likelihood functions based on the ASL density. Geraci and Bottai (2007) added individual-specific random intercepts to the linear predictor and thereby induced an individual-specific location-shift. For the parameter estimation, they proposed a Monte Carlo EM algorithm based on ASL distributed error terms. Liu and Bottai (2009) further extended this model to individual-specific random slopes and called it quantile mixed effects model to point out the analogy to the classical linear mixed model for longitudinal data. The random effects of this model were assumed to follow a symmetric multivariate Laplace distribution (corresponding to multivariate Gaussian random effects in the linear mixed model). Estimation was again based on a Monte Carlo EM algorithm. Farcomeni (2012) considered a linear mixed quantile regression model with time-varying individual-specific intercepts and assumed them to follow a first-order latent Markov chain. Again, the estimation relied on an EM algorithm with ASL distribution. Also in the context of longitudinal modelling,

Yuan and Yin (2010) considered the linear quantile regression model with individual- and quantile-specific intercepts and slopes and particularly focussed on missing data and dropout. With the assumptions of ASL distributed errors and Gaussian distributed random effects (corresponding to a Gaussian prior or a L_2-norm penalty term), a Gibbs sampler was presented for posterior estimation.

The model which most resembles the STAQ model presented in equation (3.2) on page 35 was suggested in Yue and Rue (2011). The only difference to our predictor is that their predictor contains individual-specific intercepts only – but no slopes – to account for unobserved heterogeneity. The error terms are assumed to follow an ASL distribution for which the representation as scale mixture of Gaussian densities is used to put quantile regression into a well-studied, fully Bayesian framework. Two possible algorithms based on Markov Chain Monte Carlo and on integrated nested Laplace approximation (INLA) are presented. Compared to other algorithms, the INLA algorithm is faster but relies on an approximation of the check function by a loss function with second-order derivatives. The large variety of effects is addressed by appropriate Gaussian-type priors with different forms and degrees of smoothness. The estimation can therefore be embedded in the classical L_2-norm framework.

Estimator properties and inference

From standard likelihood theory it follows that the maximum likelihood estimator is unbiased and follows an asymptotic Gaussian distribution. However, due to the non-differentiability of the likelihood with respect to the parameters, it is not possible to explicitly derive the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_\tau$ through the inverse Fisher information.

For approaches relying on frequentist estimation methods, as the EM algorithms for the longitudinal quantile regression models in Geraci and Bottai (2007), Liu and Bottai (2009) and Farcomeni (2012), bootstrap estimation for the standard errors is mainly used. All observations from the same individual build the basic re-sampling units for the block-wise bootstrap estimation.

For Bayesian methods, one obtains a sample from the posterior distribution and thereby can take the standard deviation of the sample as estimator for the standard error. However, the above mentioned quasi-likelihood character of the ASL distribution calls this proceeding into question. In Yu and Moyeed (2001), Figure 1, one can see that the skewness of the posterior distribution of $\hat{\boldsymbol{\beta}}_\tau$ is different for different values of τ, and is most likely influenced by the skewness of the ASL distribution. Furthermore, Reich *et al.* (2010) showed in a simulation study that confidence intervals obtained by the ASL approach of Yu and Moyeed (2001) achieve only poor coverage rates, in particular for extreme quantile parameters. These results were also confirmed by own simulation studies in the context of a master's thesis (Cieczynski, 2009). Consequently, one should be careful when using Bayesian standard errors for further inference, e.g., for Wald tests on single quantile regression parameters.

Since the estimation is again conducted separately for different quantile parameters, the danger of quantile crossing is not averted.

Variable selection

So far only little has been said about variable selection in connection with ASL distributed errors in literature. In the presence of a likelihood, variable selection can be based on information criteria,

as for example described in Farcomeni (2012). When the scale parameter σ of the ASL distribution is set to one, the AIC is similar to the pseudo-AIC of the classical quantile regression theory in equation (3.7). Additionally, likelihood ratio tests have been proposed (Geraci and Bottai, 2007; Farcomeni, 2012) to test if a single parameter is not equal to zero.

A L_1-norm shrinkage prior, e.g., based on the ASL distribution, has not yet been proposed for the fixed effects based on an ASL likelihood but would probably be a good option for Lasso-type variable selection.

Software

The R package `bayesQR` (Benoit *et al.*, 2011) provides an implementation of the original Bayesian approach by Yu and Moyeed (2001) for linear quantile regression. The linear quantile mixed model introduced in Geraci and Bottai (2007) can be fitted with the package `lqmm` (Geraci, 2012). The flexible quantile regression model from Yue and Rue (2011) can be estimated by the function `inla` from the R package `INLA` (Rue *et al.*, 2009) (not yet available on CRAN, but under `http://www.r-inla.org/`).

3.4.2 Flexible Bayesian approaches

This section shortly describes flexible Bayesian estimation approaches for quantile regression which have increasingly been suggested in literature over the last years, see Taddy and Kottas (2010) for an overview.

These approaches are often referred to as *nonparametric* Bayesian approaches because no explicit distribution is assumed for the error terms but only an infinite or finite mixture of weighted density components. Due to the explicit distributional assumptions on the mixture components, Bayesian estimation can be applied. However, the distribution-free character of quantile regression is conserved since the resulting error density can flexibly adapt to the underlying true shape. The term *nonparametric* can also be justified because the estimated parameters of the flexible error density are not of primary interest. Thus, we also could have placed this section in the chapter of distribution-free estimation approaches.

Note that the term *flexible* Bayesian approach refers to the flexibility of the error density and not to the flexibility of the covariate predictor.

In the following, we sketch two early approaches for Bayesian mixture modelling and shortly discuss them regarding our criteria. In the first approach, Kottas and Krnjajić (2009) started from the usual linear quantile regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + \varepsilon_{\tau i} \qquad \text{with} \quad \varepsilon_{\tau i} \overset{iid}{\sim} f_{\varepsilon_\tau} \ ,$$

and assumed the error terms to be identically distributed across observations while fulfilling the usual quantile constraint $F_{\varepsilon_\tau}(0) = \tau$. They proposed two alternative mixture densities for the errors which were both constructed from a Dirichlet Process (DP) mixture model. The error density resulting from this process can in general be expressed as an infinite mixture density:

$$f_{\varepsilon_\tau}(\varepsilon_{\tau i}|\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k f_{mix,\tau}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau | \boldsymbol{\theta}_k) \ . \tag{3.8}$$

The vector $\boldsymbol{\theta}$ contains all unknown parameters from the mixture density, i.e., the weights π_k and the parameters $\boldsymbol{\theta}_k$ of the k-th density component $f_{mix,\tau}(\cdot|\boldsymbol{\theta}_k)$. The exact hierarchical notation of the present DP mixture model can be found in Kottas and Krnjajić (2009). In short, a realization from a DP prior is a random discrete distribution with an infinite number of point masses. In the stick-breaking representation of DPs, the weights π_k for the point masses arise from a stick-breaking mechanism while the locations of the point masses are drawn from a base distribution G_0. In the present model, the drawn locations correspond to the component-specific parameters $\boldsymbol{\theta}_k$.

Kottas and Krnjajić (2009) considered two different alternatives for the component-specific densities $f_{mix,\tau}(\cdot|\boldsymbol{\theta}_k)$, namely an asymmetric Laplace density and a mixture of two uniform densities. The related MCMC algorithms for the estimation of $\boldsymbol{\beta}_\tau$ and further density parameters were based on well-established posterior simulation algorithms for DP mixtures.

In the first alternative, the k-th density component is an asymmetric Laplace density

$$f_{mix,\tau}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau|\sigma_k) = f_{ASL,\tau}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau|\sigma_k) = \frac{\tau(1-\tau)}{\sigma_k} \exp\left\{-\rho_\tau\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau}{\sigma_k}\right)\right\} ,$$

where the parameter τ is set to the fixed quantile parameter of interest. Thus the skewness of each density component is fixed and the above quantile constraint is not only fulfilled for each single mixture component but also for the final error density. Since the only parameter which can differ between components is σ_k, the shape of the resulting ASL mixture density does however not provide the desired flexibility.

In the second alternative, Kottas and Krnjajić (2009) proposed the k-th density component to be a mixture of two uniform densities:

$$f_{mix,\tau}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau|a_k, b_k) = \frac{\tau}{a_k} \cdot I(-a_k < y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau < 0) + \frac{1-\tau}{b_k} \cdot I(0 \leq y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau < b_k) .$$

The parameters a_k and b_k determine the domain of the density. Similar to the first alternative, by definition each density component – and therefore the final error density – fulfills the quantile constraint. Even though the shape of this density is more flexible than with ASL density components, the flexibility of this approach still remains limited since the assumption of *iid* error terms does not contribute to flexibility across individuals. For this reason, Kottas and Krnjajić (2009) additionally developed an error model which is associated with the covariate information.

A related approach was suggested in Reich *et al.* (2010), who considered the general location-scale model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + (\boldsymbol{x}_i^\top \boldsymbol{\gamma}_\tau)\varepsilon_{\tau i} \qquad \text{with} \quad \varepsilon_{\tau i} \overset{iid}{\sim} f_{\varepsilon_\tau} . \tag{3.9}$$

In this model, the term $\boldsymbol{x}_i^\top \boldsymbol{\gamma}_\tau$ is constrained to be positive for all $\boldsymbol{x}_i$ and the parameter vector $\boldsymbol{\gamma}_\tau$ allows the scale of the response to vary with the covariates $\boldsymbol{x}_i$. Again, the error density f_{ε_τ} is assumed to fulfill the quantile constraint and to follow an infinite mixture as in (3.8). Reich *et al.* (2010) modelled each of the base mixture components by a two-component Gaussian mixture density

$$f_{mix,\tau}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau|\mu_{1k}, \mu_{2k}, \sigma_{1k}^2, \sigma_{2k}^2, q_k) = q_k \phi(\mu_{1k}, \sigma_{1k}^2) + (1-q_k)\phi(\mu_{2k}, \sigma_{2k}^2) ,$$

where $\phi(\mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2, and $q_k \in (0,1)$ are fixed mixture proportions which ensure the quantile constraint to be fulfilled for each mixture

component. The component-specific parameters are not explicitly drawn from a DP prior, but the weights π_k arise from a stick-breaking mechanism. Reich *et al.* (2010) further extended the model to individual-specific random effects with a Gaussian assumption in a conditional and marginal way.

With respect to nonparametric Bayesian quantile regression with a **flexible predictor**, most of the approaches concentrate on the linear quantile regression model and put the focus on flexible modelling of the error assumption instead of the predictor. For some of the estimation approaches, one can imagine that the extension to a more flexible predictor would be straightforward. In particular, individual-specific random effects for longitudinal data could directly be embedded in the Bayesian framework since *all* effects are treated as random and one does not have to differ between fixed and random effects.

Note also that all the above models are fitted separately for different quantile parameters and therefore incur the danger of quantile crossing. Various recent Bayesian approaches, as for example Reich *et al.* (2011) or Reich (2012), considered the quantile process as a whole with the aim of simultaneous inference for all quantiles. In this context, Taddy and Kottas (2010) assumed that the observed data itself (response variable y and covariates x) follow an infinite DP mixture density, not only the error distribution. Since the high-dimensional data density is modelled as a whole, quantile crossing is avoided but the predictor does not provide a particular (additive or flexible) interpretable structure.

As usual in a Bayesian estimation framework, exact and full **inference** of the parameters is straightforward, even with small sample sizes. However, in the first alternative of Kottas and Krnjajić (2009) with the ASL mixture, one should still be careful since the density most likely does not represent the true shape of the errors and one has to assume the quasi-likelihood framework again. Bayesian **variable selection** methods for these approaches have not yet been discussed in literature. **Software** or R packages of the Bayesian approaches are not yet available, but for some of the approaches, e.g., the approach of Reich *et al.* (2010), some code is available on request.

3.5 Related model classes

In this section, we describe three model classes which we perceive as being closely related to quantile regression. They are related to quantile regression since the full response distribution can be derived depending on covariates and therefore also the response's quantile function. As a consequence, these model classes can be applied in similar practice situations in which quantile regression would be appropriate. However, since their original regression target is not the response's quantile function, the latter can (in most cases) not directly be expressed by an additive function of the covariates.

We consider one distribution-free related model class (expectile regression) and two distribution-based model classes (Gaussian STAR models and GAMLSS). Note that quantile modelling is always implied when an explicit distribution is assumed for response or error terms, for example in the case of conventional generalized linear regression models. We demonstrate this fact for the special case of Gaussian error terms in Section 3.5.2.

Similar to the estimation of quantile regression models, various different estimation approaches have been developed for each of the related model classes.

3.5.1 Expectile regression

Expectiles allow to describe the complete distribution of a random variable in a similar way to quantiles. Both, expectiles and quantiles can be regarded as a special case of M-quantiles (Breckling and Chambers, 1988). Therefore, expectile regression can be seen as an important distribution-free pendant to quantile regression. We motivate expectiles by the analogy between expectile and quantile regression in the following and refer to Schnabel and Eilers (2009) for the exact definition of theoretical expectiles of a distribution.

First of all, recall that for a given quantile parameter $\tau \in (0,1)$ the basic aim of quantile regression is to minimize a weighted sum of absolute deviations between response and predictor

$$\min_{\eta_\tau} \sum_{i=1}^{n} w_{i\tau}|y_i - \eta_{i\tau}| \qquad \text{with weights} \quad w_{i\tau} = \begin{cases} \tau & y_i \geq \eta_{i\tau} \\ 1-\tau & y_i < \eta_{i\tau} \end{cases}, \tag{3.10}$$

which is a slightly modified formulation of the classical quantile regression problem in equation (3.3) on page 39. By minimizing the above sum of asymmetrically weighted absolute deviations with respect to the parameters contained in the predictor η_τ, the conditional $\tau \cdot 100\%$ quantile of the response variable Y is modelled depending on covariates.

Following the work of Aigner *et al.* (1976), Newey and Powell (1987) introduced expectile regression in analogy to quantile regression by replacing the weighted absolute deviations by weighted quadratic deviations, as follows:

$$\min_{\eta_\tau} \sum_{i=1}^{n} w_{i\tau}(y_i - \eta_{i\tau})^2$$

with the same weights as defined in (3.10) and the fixed *asymmetry* parameter $\tau \in (0,1)$. Here, minimizing this asymmetric least squares criterion leads to the $\tau \cdot 100\%$ expectile of the response variable conditional on covariates. Setting $\tau = 0.5$ obviously corresponds to the special case of usual least squares estimation and implies mean modelling.

Regarding the relationship between expectiles and quantiles, Jones (1994) pointed out that there is a one-to-one relationship between the expectiles of one distribution and the quantiles of another. For some particular distributions expectiles and quantiles even coincide (see, e.g., Koenker, 2005). However, in the general case it is not straightforward to obtain the $\tau \cdot 100\%$ quantiles from a distribution when its $\theta \cdot 100\%$ expectiles with $\theta \in (0,1)$ are known. Efron (1991) suggested to estimate the conditional $\tau \cdot 100\%$ quantile as the proportion of data which lies below the estimated $\theta \cdot 100\%$ expectile.

One of the main advantages of expectile regression is that its quadratic loss function is continuously differentiable (contrary to the check function). Therefore the parameters can be estimated by an iteratively weighted least squares algorithm, which is for example described in Schnabel and Eilers (2009) or Efron (1991), and the minimization does not rely on complex linear programming algorithms. This is in particular beneficial for extensions to penalized estimation of covariate effects, since well established penalty methods from the L_2-norm framework can easily be applied to expectile regression.

However, the main shortcoming of expectile regression is that the interpretation of expectiles is not as intuitive and straightforward as the interpretation of quantiles. The estimated parameters

of the predictor are interpreted with regard to the response expectiles and apart from the rather heuristic approach in Efron (1991), the transformation of expectiles to the quantile function is not obvious. This interpretability problem might be an unsatisfying issue for practitioners.

All the same, expectile regression also has high potential for becoming a supporting model class for quantile regression. For example, expectile regression can be seen as a good alternative to quantile regression when the complete conditional distribution of a response variable should be modelled. In this case one is rather interested in the comparison of the estimated coefficients for different asymmetry parameters than in the exact interpretation of single parameters.

Flexible predictor

Schnabel and Eilers (2009) combined asymmetrically weighted least squares with P-splines and therefore allowed for the estimation of nonlinear covariate effects in expectile regression models. They proposed several possibilities for choosing the smoothing parameter, for example asymmetric (generalized) cross validation. Sobotka and Kneib (2012) extended the model to a generic structured (geo-)additive predictor and compared asymmetrically weighted least squares estimation with boosting estimation based on the expectile loss function. For longitudinal data Schnabel and Eilers (2009) raised some ideas how to extend the algorithm for the estimation of individual-specific random effects.

Estimator properties and inference

Sobotka *et al.* (2011) studied properties of expectile regression estimators obtained by asymmetrically weighted least squares. These estimators are asymptotically unbiased and Gaussian distributed. Their asymptotic covariance matrix does not depend on the true error density. This result can be seen as another important advantage of expectile regression over quantile regression since the estimation of the covariance matrix does not require the estimation of the true error density and can therefore be made in a consistent way. Consequently, standard errors and asymptotic confidence intervals of the estimators can be obtained, even for nonlinear effects (Sobotka *et al.*, 2011).

Efficiency comparisons in Newey and Powell (1987) furthermore indicate that expectile regression estimators are asymptotically more efficient than quantile regression estimators since they make use of the full information on the difference between response and predictor (and not only of the sign). On the other hand, one should be aware that expectile regression estimators are of course more sensitive to outliers than quantile regression estimators.

Due to the separate estimation for different asymmetry parameters, expectile and quantile crossing are not prevented.

Variable selection

To the best of our knowledge, the issue of variable selection has not yet been studied in literature in the context of expectile regression.

Software

Expectile regression is implemented in the R package `expectreg` (Sobotka *et al.*, 2012). This package not only provides functionality to employ structured additive expectile regression with a large variety of different covariate effects but also allows to calculate the expectiles for common distributions.

3.5.2 Gaussian STAR models

Our definition of the structured additive quantile regression model in equations (3.1) and (3.2) is mainly based on the structured additive (mean) regression model defined in Fahrmeir *et al.* (2004). In the special case of assuming a Gaussian distribution for the error terms, the model equation can be formulated by:

$$y_s = \eta_s(\boldsymbol{z}_s) + \varepsilon_s \qquad \text{where} \quad \varepsilon_s \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2) \; .$$

This notation is similar to the notation in Section 3.1: y_s denotes the response with observation index s and $\eta_s(\boldsymbol{z}_s)$ is a structured additive predictor of covariates $\boldsymbol{z}_s$ as defined in equation (3.2), but is not quantile-specific here. The error terms ε_s are assumed to be *iid* for the sake of simplicity. Of course, it would also be possible to induce an additional correlation structure, for example temporal correlation in the case of longitudinal data.

From the Gaussian assumption it follows that the main regression aim is the response's conditional mean

$$\mathbb{E}(y_s|\eta_s) = \eta_s(\boldsymbol{z}_s)$$

and that the conditional quantile function can be derived as

$$Q_{Y_s}(\tau|\eta_s) = \eta_s(\boldsymbol{z}_s) + q_\tau \sigma_\varepsilon \; ,$$

where q_τ denotes the $\tau \cdot 100\%$ quantile of a standard Gaussian distribution. Thus the quantile function is obtained by shifting the structured additive predictor $\eta_s(\boldsymbol{z}_s)$ by a quantile-specific constant $q_\tau \sigma_\varepsilon$. Gaussian STAR models can therefore be seen as a special case of quantile regression where only the intercept differs for different quantile parameters in order to fulfill the usual constraint for the error density. Consequently, the interpretations of the various linear and nonlinear effects of covariates with respect to the response's mean directly apply to the quantiles.

An important special case of Gaussian STAR models are additive mixed models for longitudinal data, see for example Ruppert *et al.* (2003), which will also used in our application of obesity in Chapter 7. An additive mixed model with linear and nonlinear population effects can be written as

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + f_1(z_{ij1}) + \ldots + f_q(z_{ijq}) + \boldsymbol{v}_{ij}^\top \boldsymbol{\gamma}_i + \varepsilon_{ij} = \eta_{ij}^{(\mu)} + \boldsymbol{v}_{ij}^\top \boldsymbol{\gamma}_i + \varepsilon_{ij} \; , \tag{3.11}$$

with *iid* errors $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and observation index $s = (i, j)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. The population part of the predictor is denoted as $\eta_{ij}^{(\mu)}$ and the individual-specific effects are assumed to be Gaussian distributed $\boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\gamma)$ and independent for different individuals i. By including time-varying covariates such as age in the design vector $\boldsymbol{v}_{ij}$, model (3.11) allows for the estimation of individual-specific random slopes or curves. When only a random intercept γ_{i0} is included in the model equation, an equicorrelation between intra-individual response observations is induced.

Due to the Gaussian error assumption the quantile function of the response can again be derived, but in the additive mixed model we have to distinguish between

$$\begin{aligned} &\text{the conditional quantile function} && Q_{Y_{ij}}(\tau|\eta_{ij}^{(\mu)}, \boldsymbol{\gamma}_i) = \eta_{ij}^{(\mu)} + \boldsymbol{v}_{ij}^\top \boldsymbol{\gamma}_i + q_\tau \sigma_\varepsilon \\ &\text{and the marginal quantile function} && Q_{Y_{ij}}(\tau|\eta_{ij}^{(\mu)}) \quad = \eta_{ij}^{(\mu)} + q_\tau \sqrt{\boldsymbol{v}_{ij}^\top \boldsymbol{\Sigma}_\gamma \boldsymbol{v}_{ij} + \sigma_\varepsilon^2}\,. \end{aligned} \tag{3.12}$$

Thus, for covariates that are just contained in the predictor $\eta_{ij}^{(\mu)}$, the interpretation of covariate effects with respect to the quantile functions remains the same as for the mean. In the case of time-varying covariates included in $\boldsymbol{v}_{ij}$, however, the relationship between covariate and quantile functions becomes more involved since both quantile functions depend on the design of $\boldsymbol{v}_{ij}$. When only a random intercept is included in the model equation, both quantile functions reduce to a simple time-constant shift of the population predictor $\eta_{ij}^{(\mu)}$ as stated above.

Regarding estimation, Gaussian STAR models are a well-studied and established framework for mean regression, and estimation algorithms and software are highly developed; an overview of recent developments in semiparametric regression is given by Ruppert *et al.* (2009). Common approaches for estimating Gaussian STAR models rely on penalized likelihood concepts, full Bayesian inference or a mixed model representation, see Fahrmeir *et al.* (2007) and Ruppert *et al.* (2003) for details. In case of likelihood estimation, smooth functional covariate effects are estimated based on penalized spline functions corresponding to suitable prior densities on the regression parameters from a Bayesian point of view. An overview of the various effect types together with the corresponding penalty matrices is given in Fahrmeir *et al.* (2007).

Flexible predictor

Due to their generic definition, Gaussian STAR models of course allow to model the full variety of different effects of the flexible structured additive predictor. However, one should be aware that they are not adequate for quantile regression if higher moments (variance, skewness, or kurtosis) of the conditional response's distribution depend on covariates, meaning that the *iid* Gaussian error assumption is violated.

Estimator properties and inference

Inference for the estimators is well studied in structured additive mean regression. For Bayesian estimation, exact and full inference for the estimators is straightforward. When penalized likelihood estimation is applied, standard errors for the usual linear coefficients can be derived. However, the asymptotic properties of penalized splines remain one of the main challenges which have been investigated in depth during the last decade (Ruppert *et al.*, 2009). Hypothesis testing on the parameters can be based on likelihood-based tests. Furthermore, Ruppert *et al.* (2003) discussed inference for estimated nonlinear functional effects, and in particular how to obtain pointwise and simultaneous confidence bands for them.

Since the full conditional distribution of the response variable is implicitly modelled by the Gaussian error assumption, all response quantiles are estimated at the same time and, therefore, quantile crossing cannot occur.

Variable selection

According to Fahrmeir *et al.* (2007), there is not much literature and software for model and variable selection in the STAR framework yet. Competing models can be compared based on the Akaike information criterion (with adapted degrees of freedom due to the penalized estimation) or the generalized cross validation criterion. These criteria are also used for smoothing parameter selection. When estimation is based on a mixed model representation of a STAR model, a special likelihood ratio test on the variance parameter can be applied to test if a smooth functional effect of a continuous covariate should be included in a linear or nonlinear way. In case of Bayesian inference, the deviance information criterion (DIC) can be used for model comparison.

Software

Penalized likelihood estimation of STAR models is implemented in the R package `mgcv` (Wood, 2012) providing the functions `gam` to fit generalized additive models and `gamm` to fit generalized additive mixed models for longitudinal data. The selection of smoothing parameters is based on generalized cross validation.

Estimation based on the mixed model representation of STAR models can be done with the R package `amer` (Scheipl, 2011) which we also used for our analysis of the obesity data in Chapter 7. In this approach the smoothing parameter is determined as the ratio between estimated variances of errors and random effects.

Full Bayesian estimation of STAR models can be employed using the software `BayesX` (Belitz *et al.*, 2012). An R interface to this software has recently been made available in the R package `R2BayesX` (Umlauf *et al.*, 2012). Functionality for exploring and visualizing estimation results obtained from `BayesX` is furthermore provided in the R package `BayesX` (Kneib *et al.*, 2011).

3.5.3 GAMLSS

Generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) aim at modelling each of the parameters of a conditional response's distribution based on structured additive covariate predictors. Two main assumptions are made: (i) The response observations y_s are mutually independent, and (ii) they follow a known distribution with density $f_{\text{dens}}(y_s|\boldsymbol{\theta}_s)$ conditional on the observation-specific parameter vector $\boldsymbol{\theta}_s = (\theta_{s1}, \theta_{s2}, \theta_{s3}, \theta_{s4})^\top$ which contains up to four parameters. Even though $\boldsymbol{\theta}$ may include any kind of distribution parameter, its components usually stand for location ($\theta_1 = \mu$), scale ($\theta_2 = \sigma$), skewness ($\theta_3 = \nu$) and kurtosis ($\theta_4 = \varphi$). Each of these distribution parameters θ_k with $k = 1, ..., 4$, is modelled by a separate structured additive predictor

$$g_1(\mu_s) = \eta_s^{(\mu)} \qquad g_2(\sigma_s) = \eta_s^{(\sigma)} \qquad g_3(\nu_s) = \eta_s^{(\nu)} \qquad g_4(\varphi_s) = \eta_s^{(\varphi)} \,,$$

where $g_1(\cdot), \ldots, g_4(\cdot)$ denote suitable monotonic link functions. The structured additive predictors $\eta^{(\mu)}$, $\eta^{(\sigma)}$, $\eta^{(\nu)}$ and $\eta^{(\varphi)}$ can include different sets of covariates and can (theoretically) each be designed as flexible as the STAQ predictor in (3.2).

Regarding the conditional density, it is not restricted to the exponential family. For analyzing BMI data, for example, very flexible distributions based on Box-Cox transformations have proven to be suitable in the past, such as the Box-Cox t distribution (Rigby and Stasinopoulos, 2004), or

the Box-Cox power exponential distribution (BCPE, Rigby and Stasinopoulos, 2006) which has for example been used for the construction of reference charts (see, e.g., Borghi *et al.*, 2006).

With a specific distributional assumption at hand, the quantile function can be derived by inverting the cdf. For example, when a BCPE distribution is assumed for the response variable, the conditional quantile function can be expressed as

$$Q_{Y_s}(\tau \,|\eta_s^{(\mu)}, \eta_s^{(\sigma)}, \eta_s^{(\nu)}, \eta_s^{(\varphi)}) = \begin{cases} \mu_s \, (1 + \sigma_s \nu_s \tilde{q}_\tau)^{1/\nu_s} & \text{if } \nu_s \neq 0 \\ \mu_s \, \exp(\sigma_s \tilde{q}_\tau) & \text{if } \nu_s = 0 \end{cases}, \tag{3.13}$$

where the quantile-specific parameter $\tilde{q}_\tau$ depends on the quantile function of a gamma distributed random variable, see Rigby and Stasinopoulos (2004) for details. To estimate the quantile function, the theoretical parameters can be replaced by their estimates. From this example one can see that the parameters estimated from a GAMLSS cannot be directly interpreted with respect to the response quantile function, as was the case for Gaussian STAR models. On the contrary, the interpretation of parameters for one covariate with respect to the quantile function can be involved since it may be associated with different predictors in different ways. However, one can directly assess the association between covariates and shape parameters of the response.

To sum up, even though GAMLSS implicitly model the quantile function of the response, they cannot be rated as quantile regression models as defined in Chapter 1, since the relationship between predictors and quantile function is not linear.

In the presence of longitudinal data, random individual-specific effects can be included in the predictors similar to the additive mixed model in equation (3.11). However, the quantile function derived from a GAMLSS can only be regarded as being conditional on given realizations of the random effects to conserve the independence assumption between observations y_s; a marginal view on the quantile function is in general not possible.

To estimate the unknown parameters included in the predictors $\eta^{(\theta_k)}$, Rigby and Stasinopoulos (2005) proposed a penalized log-likelihood approach based on two modified versions of the back-fitting algorithm for conventional generalized additive model estimation. Furthermore, a boosting algorithm for GAMLSS estimation has recently been introduced in Mayr *et al.* (2012a).

Flexible predictor

With regard to our first assessment criterion, it can be directly seen that a flexible predictor is inherent to the definition of GAMLSS. Due to their great flexibility in the distributional assumption and in the flexible modelling of each of the different predictors, GAMLSS are of course more flexible than Gaussian STAR models.

In theory, the longitudinal data structure can be modelled by individual-specific random effects, but at present the estimation of GAMLSS with random effects is limited since no software that can estimate individual-specific effects for a large number of clusters or individuals is currently available.

Estimator properties and inference

To the best of our knowledge the asymptotic properties of GAMLSS estimators have not yet been explicitly considered in literature. Standard errors for the estimators can be obtained based on the asymptotics of penalized likelihood inference. Hypothesis testing can be made with likelihood-based tests.

A great advantage of using GAMLSS for quantile modelling is their implicit prevention of quantile crossing due to the direct estimation of the full conditional response distribution.

Variable selection

Variable selection is a key issue of the GAMLSS framework, since up to four parameters of the response distribution (and not only the mean or location) are each associated with a set of covariates. This high degree of flexibility of GAMLSS requires efficient strategies for variable selection in order to select only the most relevant covariates for each distribution parameter. Rigby and Stasinopoulos (2005) proposed to use the Generalized Akaike Information Criteria (GAIC) for variable selection in GAMLSS, which corresponds to the AIC with a general penalty for the degrees of freedom. This approach comes along with several shortcomings – in particular with respect to high-dimensional data – which were discussed in Mayr *et al.* (2012a). As will be worked out in Section 4.4, boosting estimation is provided with an inherent variable selection property; this can be seen as a main advantage of boosting regarding GAMLSS estimation.

Software

The main software for fitting GAMLSS is available in the R package `gamlss` (Stasinopoulos and Rigby, 2007). Moreover, there is a number of additional R packages available providing supplementary functionality, see `www.gamlss.org` for an overview.

For fitting GAMLSS with individual-specific random effects in case of longitudinal data, the function `rc()` from package `gamlss` is currently available but is experimental and not recommended for serious practical usage. The function `gamlssNP()` from package `gamlssMX` allows for a random intercept in the predictor for μ but the random effects distribution is approximated by a Gaussian quadrature with a maximum of ten different values. Since it relies on EM maximization, the estimation is computationally challenging.

Boosting estimation for GAMLSS is implemented in the R package `gamboostLSS` (Hofner, Mayr, Fenske, and Schmid, 2011b).

Chapter 4: Boosting for structured additive quantile regression

This chapter presents a boosting algorithm for estimating structured additive quantile regression models which will often be referred to as *quantile boosting* in the following. In combination with quantile regression this innovative distribution-free estimation approach was first introduced in Fenske, Kneib, and Hothorn (2011) and can be classified among distribution-free statistical learning algorithms sketched in Section 3.3.2.

Sections 4.1, 4.2 and 4.3 describe different aspects of the boosting approach and are mainly based on the above mentioned manuscript (Fenske *et al.*, 2011), on the Ph.D. thesis of Hofner (2011) and on Kneib *et al.* (2009). In Section 4.4, we discuss the boosting algorithm with respect to the method assessment criteria from Section 3.2 and thereby compare it to the other estimation approaches for STAQ models.

4.1 Algorithm

Boosting was first introduced in the machine learning community as a classification algorithm for binary response variables (AdaBoost, see Freund and Schapire, 1996, 1997). Soon afterwards a statistical view of boosting was developed (Friedman *et al.*, 2000) and it was shown that boosting can be interpreted as a gradient descent algorithm in function space (gradient boosting, Friedman, 2001). Thereby *functional gradient descent boosting* was recognized as being suitable for fitting generalized additive regression models; and this was probably one of the key starting points for the growing popularity of boosting as a statistical learning algorithm. Bühlmann and Yu (2003) later introduced *component-wise* functional gradient descent boosting which only selects one component (i.e., covariate) per step and consequently is provided with an inherent variable selection property – note that quantile boosting described in this chapter also belongs to this type of algorithm. Further details on the history of boosting and on the relationship between different boosting algorithms can be found in Hofner (2011), and an excellent overview of state-of-the-art boosting algorithms is given by Bühlmann and Hothorn (2007).

In brief, the main goal of boosting algorithms is to predict a response variable based on a set of covariates. This goal is achieved by combining an *ensemble* of different "weak" statistical models, called *base learners*, to finally get an overall prediction for the response that yields greater prediction accuracy than the results of one single base learner only. A base learner can be any kind of statistical regression tool where the response variable is modelled by one or more covariates, i.e.,

$$\text{covariate(s)} \xrightarrow{\text{base learner}} \text{prediction of the response}$$

and typical examples for base learners are (univariate) linear regression models, classification and regression trees, or penalized regression splines.

Written very generally, the goal of boosting is to find a solution to the expected loss optimization problem

$$\eta^* = \underset{\eta}{\operatorname{argmin}} \, \mathbb{E}[L(y, \eta)] \; , \tag{4.1}$$

where y is the response and η is the predictor of a regression model, while $L(\cdot, \cdot)$ corresponds to the convex loss function that depends on the estimation problem. In practice one only has a sample of observations $(y_i, \boldsymbol{z}_i)$, with $i = 1, \ldots, n$, and the expected loss in (4.1) has to be replaced by the empirical risk

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

In the case of structured additive quantile regression, the appropriate loss function is the check function given by equation (3.4) on page 39, i.e., $L(y_i, \eta_{\tau i}) = \rho_\tau(y_i - \eta_{\tau i})$, and η is the generic structured additive predictor in equation (3.2) on page 35.

Before describing the algorithm in detail, recall the generic notation of the structured additive predictor:

$$\eta_i = \beta_0 + \sum_{d=1}^{D} h_d(\boldsymbol{z}_i) \; . \tag{4.2}$$

Here, we use the observation index $i = 1, \ldots, n$ instead of s used in (3.2) to ease readability. The unknown functions are denoted with $h_d(\boldsymbol{z}_i)$ for $d = 1, \ldots, D$, and $\boldsymbol{z}_i$ is the complete covariate vector for observation i. For some univariate covariate z_l being an element of $\boldsymbol{z}$, two main examples were linear effects $h_d(\boldsymbol{z}_i) = \beta_d z_{il}$ with coefficient β_d to estimate, or nonlinear effects $h_d(\boldsymbol{z}_i) = f_d(z_{il})$ with smooth functions f_d to estimate.

Algorithm 1 on page 61 contains the component-wise functional gradient descent boosting algorithm for estimating the unknown parameters of the functions $h_d(\cdot)$ for $d = 1, \ldots, D$. The algorithmic notation is used in accordance with Hofner (2011).

In this algorithm $\boldsymbol{h}_d = (h_d(\boldsymbol{z}_1), \ldots, h_d(\boldsymbol{z}_n))^\top$ denotes the vector of function evaluations for component $d = 1, \ldots, D$, and a corresponding base learner $\boldsymbol{g}_d = (g_d(\boldsymbol{z}_1), \ldots, g_d(\boldsymbol{z}_n))^\top$ is specified for each component. So far the total number of base learners D is equal to the number of components in the structured additive predictor in (4.2).

The index of the quantile parameter τ is suppressed for ease of notation. All the same one should keep in mind that the unknown functions and their parameters depend on a fixed and pre-specified quantile parameter $\tau \in (0, 1)$. Boosting estimation is performed separately for different quantile parameters.

Algorithm 1: Component-wise functional gradient descent boosting algorithm for structured additive quantile regression

Initialize Set the iteration index $m := 0$ and initialize the additive predictor and the function estimates with suitable starting values, typically the empirical median of the response values $y_1, \ldots, y_n$ as offset, i.e., $\hat{\eta}_i^{[0]} = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_{0.5}(y_i - c)$, and $\hat{\boldsymbol{h}}_d^{[0]} = \mathbf{0}$ for $d = 1, \ldots, D$.

Iterate

i. **Negative gradient** Increase m by 1. Compute the negative gradient residuals of the loss function evaluated at the predictor values $\hat{\eta}_i^{[m-1]}$ of the previous iteration

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

In case of quantile boosting, insert the check function for the loss function and get the following negative gradient residuals:

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

ii. **Estimation** Fit all base learners separately to the negative gradient residuals (see Section 4.2 for details) and obtain estimators $\hat{\boldsymbol{g}}_d^{[m]}$ for each base learner $d = 1, \ldots, D$. Find the best-fitting base learner $\boldsymbol{g}_{d^*}$ that minimizes the L_2 loss

$$d^* = \underset{d}{\operatorname{argmin}} \left[\left(\boldsymbol{u}^{[m]} - \hat{\boldsymbol{g}}_d^{[m]} \right)^\top \left(\boldsymbol{u}^{[m]} - \hat{\boldsymbol{g}}_d^{[m]} \right) \right]$$

with $\boldsymbol{u}^{[m]} = (u_1^{[m]}, \ldots, u_n^{[m]})^\top$ being the vector of gradient residuals of the current iteration.

iii. **Update** Compute the update for the best-fitting base learner

$$\hat{\boldsymbol{h}}_{d^*}^{[m]} = \hat{\boldsymbol{h}}_{d^*}^{[m-1]} + \nu \cdot \hat{\boldsymbol{g}}_{d^*}^{[m]}$$

where $\nu \in (0, 1]$ is a given step length.

Keep all other effects constant, i.e., set $\hat{\boldsymbol{h}}_d^{[m]} = \hat{\boldsymbol{h}}_d^{[m-1]}$ for $d \neq d^*$, and compute the update for the predictor $\hat{\eta}_i^{[m]}$ for $i = 1, \ldots, n$.

Stop Stop if $m = m_{\text{stop}}$ for a given, pre-fixed stopping iteration m_{stop}.

It can be observed that the boosting algorithm has a stepwise and component-wise character. Instead of directly fitting the original observations, the boosting algorithm iteratively fits the gradient of the loss function to the covariates. Due to multiplication of the current estimators $\hat{\boldsymbol{g}}^{[m]}$ with a step length $\nu \in (0, 1]$ in each update, the stepwise increments of the final estimators $\hat{\boldsymbol{h}}^{[m]}$ are small and thus the overall minimum is only slowly approximated. At the same time the additive structure for the resulting model fit is conserved since the final aggregation of the additive predictor and its single components is strictly additive.

The component-wise character of the boosting algorithm arises from the separate fit of the base learners to the gradient residuals. In one step, only the best-fitting base learner is updated. Since a base learner is typically based on only one (or a few) covariates, it is possible that some base learners are never updated during the boosting process and therefore the corresponding covariates are excluded from the model. This builds the basis for the inherent variable selection property of component-wise boosting.

As mentioned above, a base learner can generally be any kind of statistical regression tool that relates covariates to a response variable. In our algorithm here, the response are the negative gradient residuals and the base learners correspond to (penalized) least squares estimators of one covariate. The particular form of a base learner depends on the type of covariate effect that should be estimated. Details are given in Section 4.2.

Quantile regression results from inserting the check function as loss function. The generic boosting algorithm can however be used for a large number of different loss functions and therefore different types of regression models. Typical examples are the L_2-loss function leading to mean regression, or the negative Binomial log-likelihood leading to binary regression. Of course the choice of the loss function should depend on the specific data or estimation problem.

Note that for quantile regression 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 could similarly have chosen $\rho'_\tau(0) = \tau - 1$.

In conclusion, for completely specifying the boosting algorithm, the starting values, the base learners and their degrees of freedom, the step length ν, and the stopping iteration m_{stop} have to be specified. Further details on these parameters will be described in the following sections.

4.2 Base learners

All base learners which are considered in our boosting algorithm are estimated by (penalized) least squares. For each base learner $\boldsymbol{g}_d$ with $d = 1, \ldots, D$, from Algorithm 1, the penalized least squares criterion can be expressed as:

$$\hat{\boldsymbol{\gamma}}_d = \underset{\boldsymbol{\gamma}_d}{\operatorname{argmin}} \left[(\boldsymbol{u} - \boldsymbol{Z}_d\boldsymbol{\gamma}_d)^\top (\boldsymbol{u} - \boldsymbol{Z}_d\boldsymbol{\gamma}_d) + \lambda_d \, \boldsymbol{\gamma}_d^\top \boldsymbol{K}_d \boldsymbol{\gamma}_d \right] , \tag{4.3}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ is the vector of gradient residuals (with iteration index m dropped), $\boldsymbol{Z}_d$ is the design matrix suitable for base learner d, and $\boldsymbol{\gamma}_d$ denotes the vector with all unknown parameters to estimate. The quadratic penalty term additionally contains a suitable penalty matrix $\boldsymbol{K}_d$ and a smoothing parameter $\lambda_d > 0$ which controls the amount of regularization. The special case of unpenalized least squares base learners results from setting $\lambda_d = 0$.

When solving criterion (4.3) with respect to $\boldsymbol{\gamma}_d$, the resulting estimator for the base learner $\boldsymbol{g}_d$ is a penalized least squares estimator:

$$\begin{aligned} \hat{\boldsymbol{g}}_d &= \boldsymbol{Z}_d (\boldsymbol{Z}_d^\top \boldsymbol{Z}_d + \lambda_d \boldsymbol{K}_d)^{-1} \boldsymbol{Z}_d^\top \boldsymbol{u} \\ &= \boldsymbol{Z}_d \, \hat{\boldsymbol{\gamma}}_d \, . \end{aligned} \tag{4.4}$$

The resulting hat matrix $\boldsymbol{S}_d$ which links the estimated to the observed gradient residuals $\hat{\boldsymbol{u}} = \boldsymbol{S}_d \boldsymbol{u}$ is consequently given as $\boldsymbol{S}_d = \boldsymbol{Z}_d (\boldsymbol{Z}_d^\top \boldsymbol{Z}_d + \lambda_d \boldsymbol{K}_d)^{-1} \boldsymbol{Z}_d^\top$.

One can see that all parameters in (4.3) and (4.4) except for $\boldsymbol{u}$ depend on d, i.e., on the particular form of the base learner which is in turn determined by the type of covariate effect that should be estimated.

From the update step in Algorithm 1 it follows that the final boosting estimator for the d-th component can be expressed as

$$\hat{\boldsymbol{h}}_d^{[m_{\text{stop}}]} = \sum_{m=1}^{m_{\text{stop}}} \nu \cdot \hat{\boldsymbol{g}}_d^{[m]} \,,$$

with $\hat{\boldsymbol{g}}_d^{[m]} = \mathbf{0}$ if the d-th base learner was not selected in iteration m. Again $\hat{\boldsymbol{h}}_d = (\hat{h}_d(\boldsymbol{z}_1), \ldots, \hat{h}_d(\boldsymbol{z}_n))^\top$ and $\hat{\boldsymbol{g}}_d = (\hat{g}_d(\boldsymbol{z}_1), \ldots, \hat{g}_d(\boldsymbol{z}_n))^\top$ denote the estimated vectors of unknown function and base learner d, respectively, evaluated at the observations. Thus, the final boosting estimator of component d is just a weighted sum of the fitted base learners for all iterations when the d-th base learner was selected, and the additive structure of the resulting model fit is conserved. From equation (4.4) one can additionally see that the final vector of parameter estimators is also a weighted sum of the parameter estimators of single iterations, i.e.,

$$\hat{\boldsymbol{\gamma}}_d^{[m_{\text{stop}}]} = \sum_{m=1}^{m_{\text{stop}}} \nu \cdot \hat{\boldsymbol{\gamma}}_d^{[m]} \,,$$

with $\hat{\boldsymbol{\gamma}}_d^{[m]} = (\boldsymbol{Z}_d^\top \boldsymbol{Z}_d + \lambda_d \boldsymbol{K}_d)^{-1} \boldsymbol{Z}_d^\top \boldsymbol{u}^{[m]}$ and again $\hat{\boldsymbol{\gamma}}_d^{[m]} = \mathbf{0}$ if the d-th base learner was not selected in iteration m.

In the following, we describe the particular form of the base learners for those components from the structured additive predictor that will be used in our applications: Linear, smooth nonlinear, discrete spatial, and cluster-specific components as well as varying coefficient terms. We refer to Kneib *et al.* (2009) and Hofner (2011) for further alternatives that could similarly be used in connection with STAQ regression, such as bivariate base learners for smooth surfaces, base learners with radial basis functions for smooth spatial effects, or constrained base learners for monotonic effects.

Base learners for linear components

First, we consider the simplest case of base learners with unpenalized linear effects. In the generic model notation, a linear predictor component can be written as $h_d(\boldsymbol{z}_i) = \boldsymbol{\beta}_d^\top \boldsymbol{x}_i$ with coefficient vector $\boldsymbol{\beta}_d$ and covariate vector $\boldsymbol{x}$ being a $(p \times 1)$ subvector of the complete vector of covariates $\boldsymbol{z}$. With an intercept in the linear predictor, the covariate vector for observation i is $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ and at the same time constitutes the i-th row of the $(n \times p)$ design matrix $\boldsymbol{Z}_d = \boldsymbol{X}$ being the usual design matrix of a linear model.

In typical usage examples this linear base learner represents one covariate and the covariate vector $\boldsymbol{x}$ contains: (i) an intercept and one continuous covariate; (ii) an intercept and one dummy variable for a binary covariate; (iii) an intercept and $p - 1$ dummy variables for a categorical covariate with $p > 2$ categories.

Instead of the generic notation with $\boldsymbol{\gamma}_d$, let $\boldsymbol{b}_d$ denote the base learner coefficients corresponding to $\boldsymbol{\beta}_d$. Without penalty, i.e., with $\lambda_d = 0$ in the penalized least squares criterion (4.3), the resulting estimator is just the ordinary least squares estimator

$$\hat{\boldsymbol{b}}_d^{[m]} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}^{[m]} \,,$$

and if this base learner is selected in iteration m of the boosting algorithm, the stepwise update is

$$\hat{\boldsymbol{\beta}}_d^{[m]} = \hat{\boldsymbol{\beta}}_d^{[m-1]} + \nu \cdot \hat{\boldsymbol{b}}_d^{[m]} \,.$$

The inclusion of an intercept in the base learner might at first glance look inconsistent with the generic structured additive predictor in (4.2). There, the generic predictor just contains one global intercept and its single components are written without intercepts. However, in the boosting algorithm an intercept can be included in each base learner, since the final (global) intercept β_0 is calculated as the sum of all intercepts from all base learners.

In some cases it makes sense to omit the intercept in a linear base learner. Then the continuous covariates should be mean-centered before estimation in order to ensure correct estimators (see Hofner, 2011, p.19, for an illustration).

Note that penalization of the linear effects is of course also possible. For example, a ridge penalty is applied when setting $\boldsymbol{K}_d = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix with suitable dimensions. Ridge penalization causes all estimated coefficients to be uniformly shrunken towards zero. This penalty makes sense for a covariate with a large number of categories or a group of several variables in order to make the complexity comparable among different base learners and to equalize selection probabilities. Ridge-type penalization will also be mentioned in the context of base learners for cluster-specific components.

Illustration of quantile boosting with linear base learners

In the following we illustrate the proceeding of the boosting algorithm for the simple case of univariate linear quantile regression by Figure 4.1. First, we drew 400 observations from the following model

$$y_i = 3 + 2x_i + (4x_i) \cdot \varepsilon_i \qquad \text{with} \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 4) \quad \text{and} \quad x_i \overset{iid}{\sim} \mathcal{U}[0, 5] \quad \text{for} \quad i = 1, \ldots, 400 \,.$$

Based on the resulting sample shown in the right column of Figure 4.1, we estimated a linear quantile regression model with $\tau = 0.7$ by boosting. We specified only one linear base learner containing an intercept and the continuous covariate x.

Figure 4.1 illustrates the stepwise approximation of the boosting algorithm to the final estimator $\hat{\boldsymbol{\beta}}^{[m_{\text{stop}}]}$. The two plots in each row refer to the same iteration $m \in \{1, 500, 1000\}$.

The plots in the left column show the current fit of the base learner on gradient level. One can see that the gradient residuals (grey points) can only take two values, namely 0.7 and -0.3. The solid line is the corresponding least squares regression line with parameters $\hat{\boldsymbol{b}}^{[m]}$ while the dashed line displays the linear function with parameters $\nu \cdot \hat{\boldsymbol{b}}^{[m]}$ (with $\nu = 0.1$). Thus, the dashed line corresponds to the current increment added to $\hat{\boldsymbol{\beta}}^{[m-1]}$ in order to obtain $\hat{\boldsymbol{\beta}}^{[m]}$.

The plots in the right column display the situation on response level. Shown are the original observations (grey points), the final quantile regression fit for $\tau = 0.7$ (dashed line) with parameters $\hat{\boldsymbol{\beta}}^{[m_{\text{stop}}]}$ and the current fit (solid line) with parameters $\hat{\boldsymbol{\beta}}^{[m]} = \hat{\boldsymbol{\beta}}^{[m-1]} + \nu \cdot \hat{\boldsymbol{b}}^{[m]}$.

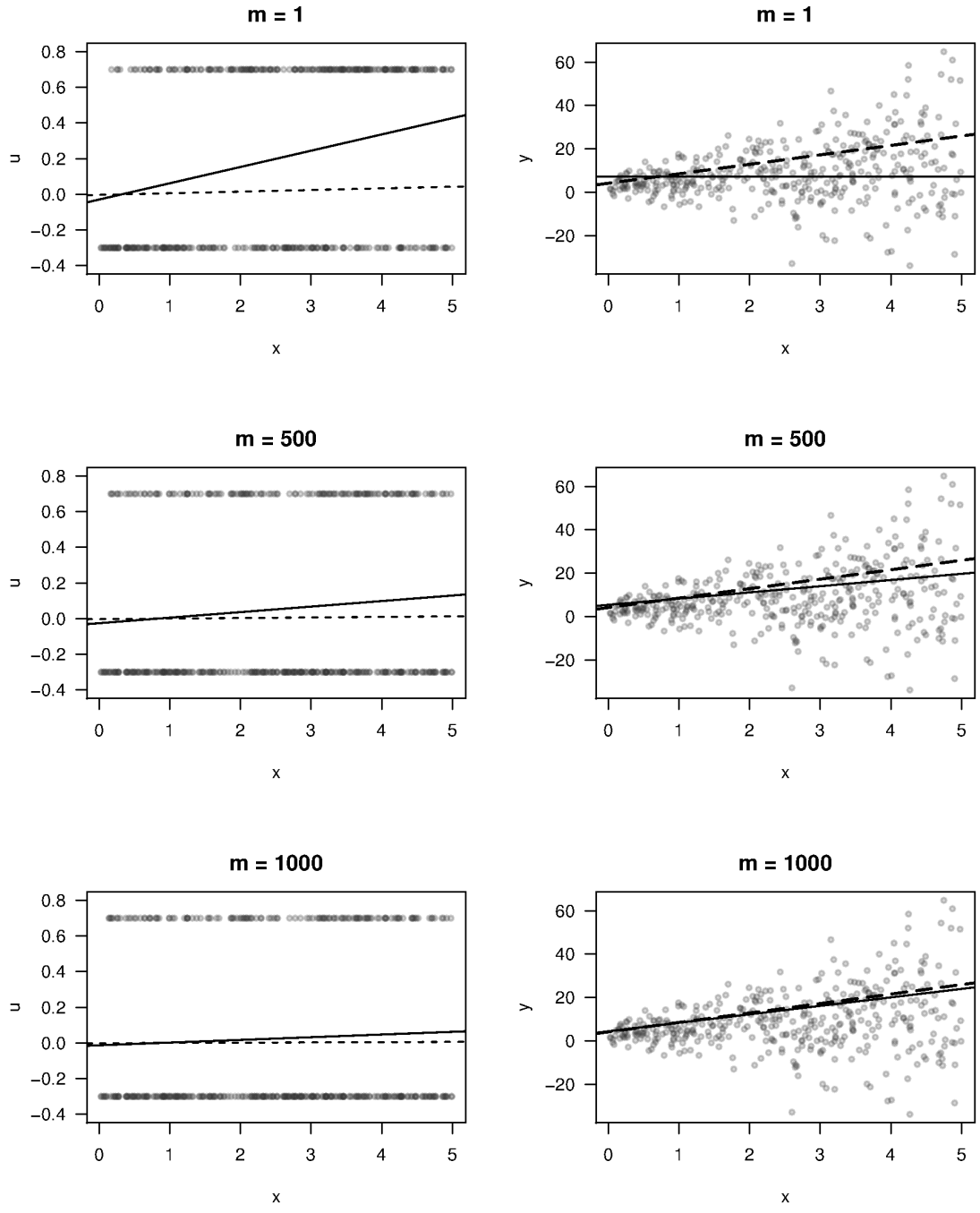


Figure 4.1 Illustration of the boosting algorithm for univariate linear quantile regression with $\tau = 0.7$. *Left column:* Gradient residuals (grey points) together with least squares regression fit with parameters $\hat{\boldsymbol{b}}^{[m]}$ (solid line) and linear function with parameters $\nu \cdot \hat{\boldsymbol{b}}^{[m]}$ (dashed line). *Right column:* Original observations (grey points) together with final quantile regression fit for $\tau = 0.7$ with parameters $\hat{\boldsymbol{\beta}}^{[m_{\text{stop}}]}$ (dashed line) and current regression fit with parameters $\hat{\boldsymbol{\beta}}^{[m]}$ (solid line).

Altogether, one can observe that the stepwise increments of the estimators are very small and almost zero. This can be attributed to the binary character of the gradient residuals and to the small value of $\nu = 0.1$. By using 5-fold cross validation the optimal number of boosting iterations was determined to be $m_{\text{stop}} = 1222$. This demonstrates that even in the simple case of univariate linear quantile regression, a large number of boosting iterations can be necessary to approximate the final quantile regression fit.

One can also observe that with increasing number of iterations roughly 30% of the gradient residuals are equal to 0.7, and the remaining 70% are equal to 0.3. These proportions are determined by the quantile parameter τ, which was chosen to be 0.7 in the present example.

Base learners for smooth nonlinear components

In the generic notation of the structured additive predictor, smooth nonlinear components of a continuous covariate z_l are written as $h_d(\boldsymbol{z}_i) = f_d(z_{il})$ and the task is to estimate the nonlinear function $f_d(z_l)$ in a smooth way. In our boosting algorithm, we use penalized B-splines, i.e., P-splines, that were introduced by Eilers and Marx (1996) and first studied in the boosting framework by Schmid and Hothorn (2008).

For simplicity, we drop the indices d and l in the following and denote the base learner corresponding to $f(z)$ with $g(z)$. A nonlinear function $g(\cdot)$ of a continuous covariate z can be approximated in terms of a moderately sized B-spline basis as follows:

$$g(z) = \sum_{k=1}^{K} \gamma_k B_k(z; \alpha) = \boldsymbol{B}(z)^\top \boldsymbol{\gamma} \ ,$$

where $B_k(z; \alpha)$ is the k-th B-spline basis function of degree α. In vector notation, the above sum can be written as product between design vector $\boldsymbol{B}(z) = (B_1(z; \alpha), \ldots, B_K(z; \alpha))^\top$ and coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^\top$. The basis functions are defined on a grid of equidistant knots and their degree δ can be chosen according to subject-matter knowledge to obtain a function estimate with the desired overall smoothness properties (since a spline of degree δ is $\delta - 1$ times continuously differentiable).

Thus, with observations $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ the unpenalized estimator can be obtained as $\hat{\boldsymbol{g}} = (\hat{g}(z_1), \ldots, \hat{g}(z_n))^\top = \boldsymbol{B}\,\hat{\boldsymbol{\gamma}}$, with $(n \times K)$ design matrix $\boldsymbol{B} = (\boldsymbol{B}(z_1), \ldots, \boldsymbol{B}(z_n))^\top$. Without penalty the coefficients $\hat{\boldsymbol{\gamma}}$ can be estimated by usual least squares, i.e., with a simple linear base learner in the boosting context.

However, for a small number of knots, the question arises how to determine their number and positions adequately, and for a large number of knots one runs the risk of overfitting. Therefore, estimation of the coefficient vector $\boldsymbol{\gamma}$ is based on minimizing a penalized least squares criterion.

In the notation of (4.3) on page 62, the design matrix $\boldsymbol{Z}_d$ is simply the B-spline design matrix $\boldsymbol{B}$ described above and $\boldsymbol{\gamma}_d$ are the spline coefficients. In order to restrict the variability of the function estimate, the squared differences between coefficients of adjacent basis functions are penalized by using the penalty matrix $\boldsymbol{K}_d = \boldsymbol{D}^\top \boldsymbol{D}$ with difference matrices $\boldsymbol{D}$ of order δ (see, e.g., Hofner, 2011, for its exact form). Usually a second order penalty, i.e., $\delta = 2$, is applied which leads to penalization of deviations of the coefficients from a linear function. Altogether this leads to the following particular penalized least squares criterion

$$(\boldsymbol{u} - \boldsymbol{B}\boldsymbol{\gamma}_d)^\top (\boldsymbol{u} - \boldsymbol{B}\boldsymbol{\gamma}_d) + \lambda_d \, \boldsymbol{\gamma}_d^\top \boldsymbol{D}^\top \boldsymbol{D} \boldsymbol{\gamma}_d \ ,$$

where the smoothing parameter λ_d controls the trade-off between smoothness and overfitting of the nonlinear function. Large values for λ_d come along with smoother functions, while the functions are more wiggly for small values of λ_d. However, with boosting estimation λ_d is not treated as a hyperparameter that needs to be optimized (see Section 4.3 for details).

To sum up, for fitting nonlinear functional effects $f_d(z_l)$ with the boosting algorithm, a P-spline base learner $g_d(z_l)$ is used. The B-spline basis for $g_d(z_l)$ carries over to $f_d(z_l)$ and the coefficients γ_d of the basis functions are stepwise updated during the boosting algorithm.

Decomposition of smooth nonlinear components into several base learners

Note that it is also possible to allow the boosting algorithm to differentiate between linear and nonlinear effect of a continuous covariate z, as was proposed in Kneib *et al.* (2009). For this purpose, the complete effect of z_l is decomposed into

$$f_d(z_l) = \beta_{0d} + \beta_{1d} z_l + f_d^{\text{center}}(z_l) \ , \tag{4.5}$$

where $\beta_{0d} + \beta_{1d} z_l$ represents the linear effect of z_l, whereas $f_d^{\text{center}}(z_l)$ stands for the nonlinear deviation of $f_d(z_l)$ from the linear effect. On base learner level, this decomposition can be realized by assigning separate base learners to the linear effect and the nonlinear deviation. With component-wise selection of base learners, the boosting algorithm then decides in a data-driven way whether the linear part in (4.5) is sufficient to describe the effect of z or whether the nonlinear extension should be additionally included in the model.

Technically, for estimating the coefficient vector γ_d, a reparameterization of γ_d is required which can be obtained based on the spectral decomposition of the penalty matrix $\boldsymbol{K}_d$, see Kneib *et al.* (2009) and Fahrmeir *et al.* (2004) for details. It is also possible to consider centering around higher-order polynomials, although the decision between linear and nonlinear effects seems to be most relevant in practice.

When this decomposition is used, the estimation of one component in the additive predictor corresponds to several base learners, and the smoothing parameter λ_d should be chosen so that the complexity of the nonlinear part is comparable to the one of the linear part. This issue was discussed in Kneib *et al.* (2009) and will again be raised in Section 4.3.

Base learners for varying coefficient terms

In the generic notation varying coefficient terms are denoted with $h_d(\boldsymbol{z}_i) = z_{ik} \cdot f_d(z_{il})$ with categorical or continuous covariate z_k and a second continuous covariate z_l, both being elements of $\boldsymbol{z}$. To smoothly estimate the nonlinear function $f_d(z_l)$ only a slight modification of the penalized least squares base learner for nonlinear components is required. To achieve the multiplication of the function evaluations $f_d(z_l)$ with the interaction variable z_k, the design matrix has to be altered to $\boldsymbol{Z}_d = \text{diag}(z_{1k}, \ldots, z_{nk}) \cdot \boldsymbol{B}_l$, where $\boldsymbol{B}_l$ is the B-spline design matrix of the covariate z_l described above. Inserting $\boldsymbol{Z}_d$ into the penalized least squares base learner (4.4) in combination with a difference penalty yields a suitable base-learning procedure for estimating varying coefficients.

Base learners for discrete spatial components

An effect of a covariate z_l with discrete spatial information, for example the region within a country, is denoted by $h_d(\boldsymbol{z}_i) = f_d(z_{il})$ in the generic predictor notation. We describe this discrete spatial effect according to Sobotka and Kneib (2012). In fact, $z_l \in \{1, \ldots, R\}$ is simply a categorical covariate with R possible values and its corresponding $(n \times R)$ design matrix $\boldsymbol{Z}_d = (z_{d,ir})$ just contains binary dummy vectors for each region, more specifically it contains the following elements:

$$z_{d,ir} = \begin{cases} 1 & z_{il} = r \\ 0 & z_{il} \neq r \end{cases} \qquad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad r = 1, \ldots, R\,.$$

The aim is to estimate the coefficient vector $\boldsymbol{\gamma}_d = (\gamma_{d1}, \ldots, \gamma_{dR})^\top$ with region-specific effects γ_{dr} for $r = 1, \ldots, R$, so that the spatial function simplifies to $f_d(z_{il}) = \gamma_{dr} \cdot I(z_{il} = r)$ with indicator function $I(\cdot)$.

Penalized estimation of this effect makes sense to account for spatial autocorrelation. Thereby the effects of neighbouring regions should be more similar to each other than effects of non-neighbouring regions. This can be realized by using the $(R \times R)$ penalty matrix $\boldsymbol{K}_d = (k_{d,rs})$ with the following elements

$$k_{d,rs} = \begin{cases} w_r & r = s \\ -1 & r \neq s, r \sim s \\ 0 & r \neq s, r \nsim s \end{cases} \qquad \text{for} \quad r = 1, \ldots, R \quad \text{and} \quad s = 1, \ldots, R\,,$$

where w_r is the total number of neighbours for region r and $r \sim s$ means that regions r and s are neighbours. As remarked by Sobotka and Kneib (2012), a stochastic interpretation of this penalty is that $\boldsymbol{\gamma}_d$ follows a Gaussian Markov random field.

Base learners for cluster-specific components

Recall that in the generic predictor notation, cluster-specific components were denoted by $h_d(\boldsymbol{z}_i) = z_{il} \cdot \big([I(z_{ik} \in G_1), \ldots, I(z_{ik} \in G_K)]^\top \boldsymbol{\gamma}_d\big)$ with indicator function $I(\cdot)$, a categorical or continuous covariate z_l and a categorical covariate z_k defining K different groups or clusters $G_1, \ldots, G_K$. Accordingly, the $(K \times 1)$-vector $\boldsymbol{\gamma}_d$ contains the cluster-specific parameters for each level of z_k. To estimate the coefficient vector $\boldsymbol{\gamma}_d$, the general structure of the suitable design matrix is:

$$\underset{n \times K}{\boldsymbol{Z}_d} = \underset{n \times n}{\operatorname{diag}(z_{il})} \cdot \underset{n \times K}{\boldsymbol{Z}_k}\,,$$

with $\boldsymbol{Z}_k$ being the standard design matrix of the categorical covariate z_k (consisting of dummy variables for each of the K categories), and the diagonal matrix $\operatorname{diag}(z_{il})$ containing the observed values of z_l on the diagonal. This structure of the design matrix is a combination of the one of varying coefficient terms and of the design matrix of a categorical covariate.

The coefficients are then estimated by ridge penalization, i.e., with setting the penalty matrix $\boldsymbol{K}_d = \boldsymbol{I}_K$ equal to the identity matrix $\boldsymbol{I}_K$ of dimension $K \times K$. This penalty causes all effects to be evenly shrunken towards zero.

Example: Individual-specific effects for longitudinal data

We give an example for cluster-specific components in the particular case of individual-specific effects for longitudinal data in accordance with Fenske *et al.* (2012b). Given longitudinal data with $i = 1, \ldots, N$ individuals and $j = 1, \ldots, n_i$ observations per individual, we consider the following STAQ model

$$Q_{Y_{ij}}(\tau|\cdot) = \eta_{ij}^{(\tau)} + b_{\tau i0} + b_{\tau i1} t_{ij} \ .$$

Here, $\eta_{ij}^{(\tau)}$ denotes a structured additive predictor with population effects while $b_{\tau i0}$ is an individual-specific intercept and $b_{\tau i1}$ is an individual-specific slope for the time-varying covariate t.

Translated to the generic predictor notation from above, the model contains two cluster-specific components – one for the individual-specific intercepts b_{i0} and one for the slopes b_{i1} (with quantile parameter τ dropped). For both components, z_k is the ID variable defining the individuals $i = 1, \ldots, N$. For the individual-specific intercepts z_l simply corresponds to the unit vector whereas for the slopes, z_l is defined by the time-varying variable t.

With boosting estimation the two cluster-specific components b_{i0} and $b_{i1} t_{ij}$ are separated into two different base learners. This allows the algorithm to decide in a data-driven way whether the individual-specific effects should enter the model in order to account for unobserved heterogeneity.

The design matrix $\boldsymbol{Z}_{b_0}$ for fitting individual-specific intercepts is just the design matrix of the categorical ID variable while the design matrix $\boldsymbol{Z}_{b_1}$ for fitting individual-specific slopes links the ID variable to the corresponding observations of the time-varying covariate t, leading to the following structure (with observations ordered by ij):

$$\boldsymbol{Z}_{b_0} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \end{pmatrix} \qquad \boldsymbol{Z}_{b_1} = \begin{pmatrix} t_{11} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ t_{1n_1} & 0 & \cdots & \cdots & 0 \\ 0 & t_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & t_{2n_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \end{pmatrix} .$$

In order to make the estimation of a potentially large number of parameters possible, a ridge penalization is imposed on the estimated effects b_{i0} and b_{i1}. Thereby the penalty matrix is again the $(N \times N)$ identity matrix $\boldsymbol{I}_N$ as described above. Denoting the base learner for the individual-specific intercept b_{i0} with a_{i0} for $i = 1, \ldots, N$, the general penalized least squares criterion from (4.3) for fitting this base learner in iteration m simplifies to

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} (u_{ij}^{[m]} - a_{i0})^2 + \lambda \sum_{i=1}^{N} a_{i0}^2 \ ,$$

where the smoothing parameter λ controls the degree of shrinkage of the individual-specific effects.

Note that the estimation of individual-specific effects with ridge-penalized least squares base learners is a natural concept in analogy to Gaussian random effects in additive mixed models. The quadratic form of the penalty corresponds to the log-density of Gaussian random effects priors from a Bayesian perspective. (This is for example clarified in Appendix A.2 of Hofner, 2011). As will be further pointed out in Section 5.3, the individual-specific effects of a STAQ model can be interpreted in analogy to the conditional view of random effects in additive mixed models.

4.3 Boosting parameters

For a complete specification of the boosting algorithm, the choice of the following parameters is necessary: the starting values for the intercept and function estimates, the base learners and their degrees of freedom $df(\lambda_d)$, the step length ν, and the optimal stopping iteration m_{stop}. While base learners were treated in the preceding section, we discuss the other parameters in the present section.

Note that the only hyperparameter that needs to be tuned within the fitting process is the optimal number of boosting iterations. All other parameters – including the degrees of freedom of smooth nonlinear effects – are fixed in advance.

Starting values $\hat{\boldsymbol{h}}_d^{[0]}$ and $\hat{\eta}^{[0]}$

While it is natural to initialize all function estimates at zero, i.e., $\hat{\boldsymbol{h}}_d^{[0]} = \mathbf{0}$, faster convergence and more reliable results are obtained by defining a fixed offset as a starting value for the additive predictor $\hat{\eta}^{[0]}$. In the context of quantile regression an obvious choice may be the $\tau \cdot 100\%$ sample quantile of the response variable, but our empirical experience suggests that the median is more suitable in general.

This empirical experience can be illustrated by a small simulation example in a heteroscedastic data setup with one covariate, see Figure 4.2 (and Section 5.1 for details on the data structure). For quantile parameters smaller than $\tau = 0.5$, we explored hardly any differences between the resulting optimal m_{stop} and linear quantile regression estimators $\boldsymbol{\beta}_\tau$ depending on the starting values. However, for quantile parameters larger than $\tau = 0.5$ the optimal m_{stop} was considerably increased when taking the $\tau \cdot 100\%$ sample quantile as starting value.

As an example, Figure 4.2 illustrates the stepwise approach of the boosting estimation to the true conditional 90% quantile curves depending on the starting value. One can observe that it takes more iterations until the estimation approximates the true quantile curve when beginning at the 90% sample quantile shown in the left plot of Figure 4.2. On the contrary, the right plot of Figure 4.2 displays that the estimation converges much faster when beginning at the median.

This effect is caused by asymmetric weighting of the observations by the check function during the estimation procedure. When the slope was negative, it would be the other way around and it would take more iterations to approach the quantile regression line for smaller values of τ. Without additional prior knowledge, it makes therefore sense to fix the starting value at the median.

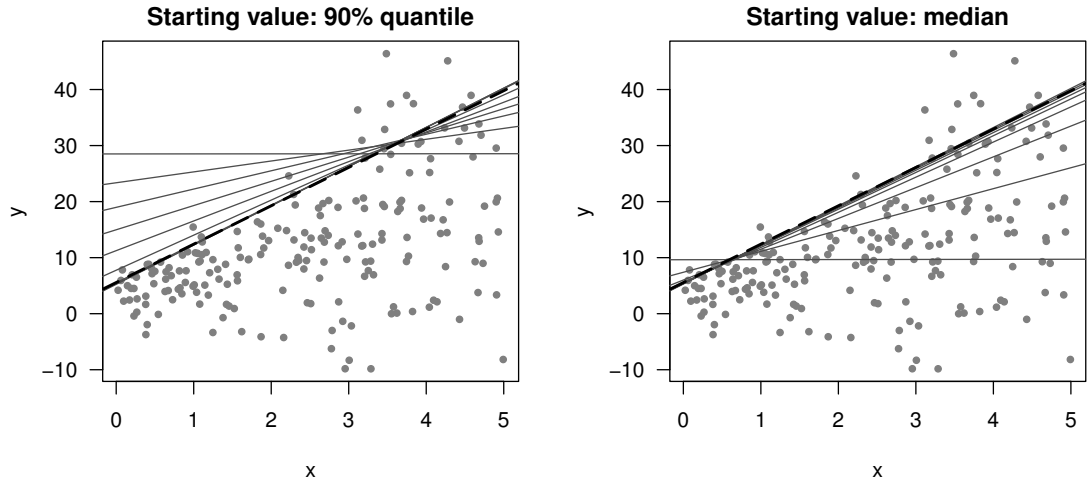


Figure 4.2 Heteroscedastic data example with $n = 200$. Dashed black lines show the true conditional quantile curves for $\tau = 0.9$, grey solid lines illustrate the stepwise boosting fit after each 300 iterations beginning at the horizontal line. *Left plot:* Starting value = 90% quantile (horizontal line); *Right plot:* Starting value = median (horizontal line).

Step length ν

Originally, the step-length factor $\nu \in (0, 1]$ was regarded as an additional tuning parameter of the boosting algorithm and optimized in every step of the boosting algorithm (see, e.g., Friedman, 2001). Later it was established that this parameter is only of minor importance for predictive accuracy of the estimators as long as ν is chosen "sufficiently small" (Bühlmann and Hothorn, 2007).

The step length 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. Thus, when one of these two parameters is fixed, an optimal choice only has to be derived for the remaining one. Since m_{stop} is easier to vary in practice, the step length ν is fixed at a small value, e.g., $\nu = 0.1$, to ensure small steps and therefore weak base learners.

As was illustrated by Figure 4.1, the stepwise increments of the estimators can be very small in case of quantile regression when ν is set to be 0.1 due to the binary character of the gradient residuals. This potentially results in a large number of boosting iterations. To avoid excessive computational effort it can make sense to fix ν at a greater value than 0.1 in the context of quantile regression, e.g., $\nu = 0.2$ or $\nu = 0.4$, as was done in our applications.

Note that multiplying ν with a constant $c > 0$ has the same impact on the estimation result as multiplying the original loss function (and its gradient) with c. For example, the standard loss function for median regression is the absolute value loss, i.e., $L(y, \eta) = |y - \eta|$, while the check function for $\tau = 0.5$ is exactly half of this quantity, i.e., $\rho_{0.5}(y - \eta) = 0.5\,|y - \eta|$. Thus, quantile boosting with $\nu = 0.2$ and $\tau = 0.5$ is equivalent to boosting with the absolute value loss function and $\nu = 0.1$.

Degrees of freedom $df(\lambda_d)$

It is important to note that in the boosting algorithm the smoothing parameters $\lambda_d > 0$ of the penalized least squares base learners $d = 1, \ldots, D$, are not treated as hyperparameters to be optimized. This is one of the main differences of boosting to other penalized model approaches where λ is often the major tuning parameter.

However, when specifying different degrees of freedom for different base learners one would run the risk for a biased selection of base learners. A base learner with greater degrees of freedom, i.e., less penalization, offers greater flexibility than a base learner with smaller degrees of freedom, i.e., more penalization, and therefore has a greater chance to be selected by the boosting algorithm.

To avoid this bias in the base learner selection, Kneib *et al.* (2009) and Hofner *et al.* (2011a) suggested to fix the initial degrees of freedom $df(\lambda_d)$ at the same (small) value for all penalized base learners, for example at $df(\lambda_d) = 1$ for $d = 1, \ldots, D$. This should ensure that the complexity of each base learner is comparable. Since there is a direct relationship between smoothing parameters λ_d and degrees of freedom $df(\lambda_d)$ of a base learner (Bühlmann and Yu, 2003), the smoothing parameters λ_d can be derived by solving the initial equation $df(\lambda_d) = 1$ for λ_d and $d = 1, \ldots, D$ (see Hofner *et al.*, 2011a, Lemma 1 for technical details).

Regarding the degrees of freedom of a base learner, Kneib *et al.* (2009) proposed to use the standard definition from the smoothing literature. According to that, the degrees of freedom of a penalized least squares estimator are defined as the trace of the hat matrix, i.e., $df(\lambda_d) = \mathrm{tr}(\boldsymbol{S}_d)$, with the hat matrix of a base learner resulting from (4.4) on page 62. Soon afterwards Hofner *et al.* (2011a) deduced the alternative $df(\lambda_d) = \mathrm{tr}(2\boldsymbol{S}_d - \boldsymbol{S}_d^\top \boldsymbol{S}_d)$ and demonstrated why applying this definition in the boosting algorithm makes more sense than using the classical one when the aim is an unbiased selection of base learners.

Note that due to the repeated selection of a base learner, in the final model the degree of smoothness of a penalized effect can be of higher order than the one imposed by the initial degrees of freedom (Bühlmann and Hothorn, 2007). In addition, different degrees of smoothness can be obtained for different functional effects as a result of different selection rates of the corresponding base learners.

Regarding nonlinear effects based on P-splines, the degrees of freedom of a smooth nonlinear effect cannot be made arbitrarily small – even for large smoothing parameters λ_d. With a difference penalty of order δ, a $\delta - 1$ polynomial of the nonlinear function always remains unpenalized. For this reason, Kneib *et al.* (2009) suggested to decompose the nonlinear effect into linear part and nonlinear deviation, as was described in (4.5) on page 67. By splitting the complete effect into three base learners for intercept, linear part and nonlinear deviation, the corresponding degrees of freedom of each part can be set to one. In this context, Hofner *et al.* (2011a) advocated that the base learner of a categorical covariate should also be penalized to one degree of freedom.

Number of boosting iterations m_{stop}

The number of boosting iterations m_{stop} is the most important parameter of the boosting algorithm since it controls variable selection and overfitting behaviour of the algorithm, including the amount of shrinkage and smoothness of the estimators.

However, in general the danger of overfitting is relatively small for boosting algorithms when weak base learners with small degrees of freedom and small step lengths are used (Bühlmann and Hothorn, 2007). Stopping the boosting algorithm early enough (*early stopping*) is all the same crucial to induce shrinkage of the estimators towards zero. Shrinkage is desirable since shrunken estimates yield more accurate and stable predictions due to their reduced variance (see, e.g., Hastie *et al.*, 2009). In addition, early stopping is important to employ the inherent variable selection and model choice abilities of boosting (which we will further discuss in Section 4.4).

The optimal number of boosting iterations m_{stop} for STAQ models can be determined by cross-validation techniques, such as k-fold cross-validation, bootstrap or subsampling. With each of these techniques, the data is split into two parts: a training and a test sample. Boosting estimation is then carried out on the training sample with a very large initial number of iterations while the empirical risk is evaluated on the test sample (*out-of-bag risk*) for each boosting iteration. The optimal m_{stop} finally arises as the point of minimal risk of the aggregated empirical out-of-bag risks.

To save computational effort, Mayr *et al.* (2012b) recently proposed a sequential and fully data-driven approach for the search of the optimal m_{stop}. This approach also avoids that the initial number of boosting iterations has to be specified by the user.

4.4 Method assessment

In this section, we discuss properties of quantile boosting with respect to the method assessment criteria from Section 3.2 and thereby compare boosting estimation for STAQ models with the other estimation approaches presented in Chapter 3.

Flexible predictor

With boosting, a particular type of covariate effect is estimated by a particular form of the corresponding base learner. As was shown in Section 4.2, penalized and unpenalized estimation of a variety of different effect types is already possible, and even more possible effect types and their corresponding base learners are described in Hofner (2011) and Kneib *et al.* (2009). Altogether, all components from the STAQ predictor in (3.2) from Chapter 3.1 are completely covered by boosting. Moreover, due to the modular structure of the boosting algorithm with base learners addressing only one or a few covariates, it is straightforward to extend the algorithm to further effect types, as was for example done in Hofner *et al.* (2011c) for effects with monotonicity constraints.

In comparison with other estimation approaches, the combination of smooth nonlinear and individual-specific effects in the STAQ predictor provided by quantile boosting has so far not been possible for other distribution-free estimation approaches. In addition, standard estimation of nonlinear effects (implemented in the R package `quantreg`) is usually conducted by linear programming algorithms and yields piecewise linear functions as estimators (see Section 3.3.1).

By using quantile boosting, 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.

Estimator properties and inference

Boosting with early stopping is a shrinkage method with implicit penalty. As a result, boosting estimators will be biased for finite samples but typically the bias vanishes for increasing sample sizes (Bühlmann and Hothorn, 2007). The number of iterations m_{stop} can be regarded as a smoothing parameter that controls the bias-variance trade-off (Bühlmann and Yu, 2003), and the resulting shrinkage property of boosting estimators is beneficial with respect to prediction accuracy.

Regarding consistency of boosting estimators, Bühlmann and Yu (2003) showed that for a L_2 loss function the optimal minimax rate is achieved by component-wise boosting with smoothing splines as base learners. Zhang and Yu (2005) studied consistency and convergence of boosting with early stopping in general. They showed that models fitted by boosting with early stopping attain the Bayes risk. Unfortunately, their results are not directly applicable for quantile regression since the check function is not twice continuously differentiable with respect to η. Thus, an approximation by means of a continuously differentiable function, as for example given by the expectile loss function (see Section 3.5.1), would have to be applied.

Since boosting just yields point estimators, subsampling strategies, such as the bootstrap, have to be applied to obtain standard errors of the estimators. However, this is no fundamental drawback compared to other estimation approaches for STAQ models since most of the approaches also rely on bootstrap to obtain standard errors.

Similar to the majority of the other estimation approaches, quantile boosting does not prevent quantile crossing since the estimation is performed separately for different quantile parameters.

Variable selection

Boosting with early stopping is accompanied with an inherent and data-driven mechanism for variable selection since only the best-performing covariate is updated in each boosting step. By stopping the algorithm early, less important covariates are not updated and are therefore effectively excluded from the final model.

For example, 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 as those will allow for a better fit to the negative gradient residuals. When the boosting algorithm is stopped after an appropriate number of iterations, spurious non-informative covariates are likely to be not selected.

Thus, boosting combines parameter estimation and variable selection into one single model estimation procedure. When the estimation is additionally conducted on bootstrap samples, not only the variability of the effect estimates is assessed but also the variability of the variable selection process itself.

Boosting also allows for model choice when considering competing modelling possibilities. In this context, the decomposition of a nonlinear functional effect into base learners for linear part and nonlinear deviation is particularly important since the decision on linearity vs. nonlinearity of an effect can be made in a fully data-driven way.

Furthermore, component-wise boosting can be applied in $p >> n$ cases, i.e., for high-dimensional data with more covariates than observations, since a single base learner typically relies on one covariate only and is fitted separately from other base learners. Moreover, problems with multi-collinearity, which in particular arise in high-dimensional data, do not have a negative effect on the estimation accuracy.

Regarding consistency of the variable selection procedure, Bühlmann (2006) studied boosting for linear models with simple linear models as base learners. They pointed out connections to the Lasso and showed that boosting yields consistent estimates for high-dimensional problems. However, there are no similar results available for additive models to the best of our knowledge. For additive models an alternative for a formal variable selection procedure is offered by stability selection (Meinshausen and Bühlmann, 2010) which leads to consistent variable selection and controls of the family-wise error rate.

To sum up, boosting provides a unique framework for variable selection in STAQ models. This can be seen as a major advantage of quantile boosting over other estimation approaches, which in the majority of cases only poorly address variable selection issues.

Software

The R package `mboost` (Hothorn *et al.*, 2010, 2012) provides an excellent implementation of the generic functional gradient descent boosting algorithm presented in Section 4.1, and one can choose between a large variety of different loss functions and base learners.

Quantile regression is applied when specifying the argument `family=QuantReg()` with the two arguments `tau` for the quantile parameter and `qoffset` for the offset quantile. Code examples for estimating STAQ models with `mboost` will be given in Chapters 6.1 and 7.1.

To our knowledge, `mboost` is currently the only software that allows to fit the full variety of different effect types from the structured additive predictor. In comparison to the R package `quantreg`, which has established as a standard tool for fitting linear quantile regression models, more complex models with individual-specific and spatial effects, varying coefficient terms and a larger number of smooth nonlinear effects can be fitted by `mboost`.

4.5 Further remarks

Boosting estimation for related model classes

The generic boosting algorithm from Section 4.1 can also be used for the estimation of related model classes that were treated in Section 3.5 of this thesis, such as Gaussian STAR models, expectile regression and GAMLSS.

Gaussian STAR models can be fitted when inserting the L_2 loss function instead of the check function in the boosting algorithm, leading to the well-studied class of L_2 boosting (Bühlmann and Yu, 2003). Structured additive expectile regression models can be estimated by using the expectile loss function based on weighted quadratic deviations (Sobotka and Kneib, 2012). Both options are implemented in the R package `mboost` and can be realized by specifying `family=GaussReg()` or `family=ExpectReg()`, respectively. Of course, most of the descriptions regarding boosting parameters and properties in this chapter also apply to these alternative loss functions.

For GAMLSS models, a boosting algorithm called gamboostLSS has recently been developed in Mayr, Fenske, Hofner, Kneib, and Schmid (2012a) and was accompanied by an implementation in the R package `gamboostLSS` (Hofner, Mayr, Fenske, and Schmid, 2011b). Boosting estimation is in particular appealing for GAMLSS models because of its inherent variable selection properties. Due to the potentially very large number of parameters and covariate combinations in a GAMLSS, variable selection is of major importance in the GAMLSS framework.

GAMLSS models cannot be fitted by simply inserting an appropriate loss function in the standard boosting algorithm since more than one distributional parameter is modelled by a structured additive predictor. Therefore, an extension of the boosting algorithm (and the package `mboost`) to the gamboostLSS algorithm (and the package `gamboostLSS`) was necessary.

In each iteration of the gamboostLSS algorithm, all distributional parameters are successively updated. More precisely, for each distributional parameter the negative gradient residuals with respect to this parameter (based on the partial derivatives) are computed while the current estimators for the other distributional parameters are inserted as offset values. As for the standard boosting algorithm, base learners are then fitted to the negative gradient residuals and only the best-fitting base learner is updated for each distributional parameter. Regarding the optimal number of stopping iterations, in some cases it makes sense to apply *multi-dimensional stopping* with different stopping iterations for different distributional parameters instead of using the same m_{stop} for all distributional parameters (see Mayr *et al.*, 2012a, for further details).

Finally, note that boosting can also be applied to binary regression models by using the negative Binomial log-likelihood as loss function, corresponding to the argument `family=Binomial()` in the R package `mboost`. Binary regression can also be regarded as a model class related to quantile regression since it is often applied in similar practice situations in which quantile regression would be appropriate. For example, in the context of our applications the standard approach in literature for analyzing overweight or undernutrition is to dichotomize the continuous response variable (BMI or Z-score, respectively) and to employ a logistic regression for this binary response. In Chapter 6, we therefore compare our STAQ approach with structured additive logistic regression for analyzing undernutrition of children in India.

Simultaneous developments of boosting algorithms for quantile regression

Simultaneously to the component-wise boosting algorithm for quantile regression introduced in Fenske *et al.* (2011), two more boosting procedures were published with the similar aim of employing quantile regression.

Kriegler and Berk (2010) also combined boosting with the check function but they used regression trees as base learners and included an additional subsampling step in each iteration of the boosting algorithm – leading to *stochastic* gradient boosting. This procedure was applied to estimate the number of homeless people in small areas in and around Los Angeles. Since underestimation of the number of homeless was considered to be much worse than overestimation, the quantile parameter τ was interestingly used and interpreted as the cost-ratio between under- and overestimation of the response variable.

When large trees are used as base learners, the final model can only be described as a black box and does not allow to quantify the partial influence of single covariates on the response, as provided by component-wise boosting. Even though stumps as base learners, i.e., trees with one split and two terminal nodes, would provide an interpretation on covariate level, the resulting covariate effects are non-smooth step functions only.

Zheng (2012) also put quantile regression into the framework of component-wise boosting but considered the algorithm rather from a machine learning point of view. Base learners were not specified in more detail and simple linear models were taken as base learners in all applications. Moreover, a focus was put on binary response variables and a binary classification scheme was proposed and investigated based on quantile boosting.

Chapter 5: Empirical evaluation of quantile boosting

To evaluate the performance of the quantile boosting algorithm introduced in Chapter 4, we conducted several simulation studies. The main goals of these empirical investigations were

(i) to evaluate the correctness of both the boosting algorithm and its specific implementation in situations in which quantile regression would be appropriate,

(ii) to evaluate the variable selection and model choice properties of quantile boosting in higher-dimensional settings,

(iii) to judge the quality of estimated quantile functions, and

(iv) to get an understanding of individual-specific effects estimated by quantile boosting with ridge-penalized base learners.

For the first goal, we considered linear models (Section 5.1) as well as typical additive model structures with a moderate number of nonlinear effects (Section 5.2). For the second goal we added several nuisance covariates in the additive model settings that do not impact the response but were still considered as candidate covariates during estimation (Section 5.2). For the third goal, we considered a simple univariate setup and compared the estimated quantile functions with the true underlying quantile function directly (Section 5.3). Finally, we simulated longitudinal data settings and estimated quantile boosting with individual-specific effects (Section 5.4).

5.1 Simulation study for linear quantile regression

With the linear simulation setup we wanted to check how quantile boosting works in situations with linear effects on the response's quantile function. In particular, our aim was to compare the performance of quantile boosting with the well-established estimation approach based on linear programming (implemented in the function `rq()` from the R package `quantreg`) that can be regarded as a "gold standard" for linear quantile regression.

Data generating process

We considered the following location-scale-model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, \varepsilon_i \qquad \text{where} \quad \varepsilon_i \overset{iid}{\sim} F_\varepsilon \quad \text{for } i = 1, \ldots, n\,. \tag{5.1}$$

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 F_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

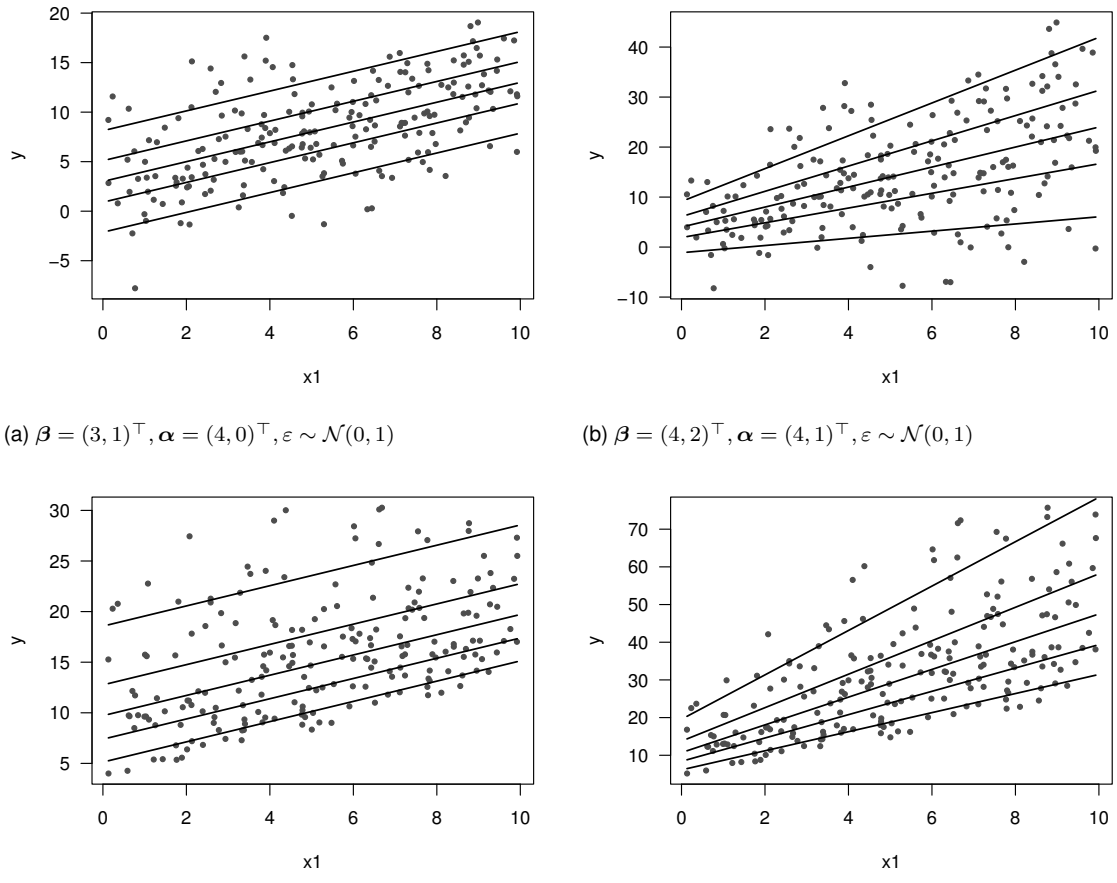
$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, F_\varepsilon^{-1}(\tau) = \boldsymbol{x}_i^\top (\boldsymbol{\beta} + \boldsymbol{\alpha} F_\varepsilon^{-1}(\tau)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau\,.$$

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

Based on the linear model in (5.1), 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 Gaussian distribution $\mathcal{N}(0, 1)$, a t-distribution with 2 degrees of freedom $t(2)$, and a gamma distribution $\mathcal{G}(1, 2)$, where $\mathbb{E}(\varepsilon_i) = \mathbb{V}(\varepsilon_i) = 2$. Figure 5.1 visualizes data examples from the first two setups with one covariate for Gaussian or gamma distributed error terms. Note that $\boldsymbol{\alpha} = (4, 1)^\top$ leads to a heteroscedastic data structure where the quantile curves are no longer parallel shifted as for $\boldsymbol{\alpha} = (4, 0)^\top$.



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

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

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

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

Figure 5.1 Data examples for linear simulation setups with $n = 200$ observations and one covariate in a homoscedastic (left) or heteroscedastic (right) data structure with Gaussian (top) or gamma (bottom) distributed error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

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 quantile boosting (function `glmboost()` from package `mboost`) and by linear programming (function `rq()` from package `quantreg`). For quantile boosting, we fixed the step length at $\nu = 0.1$. The optimal number of iterations m_{stop} was determined 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. Contrary to the additive simulation settings in Section 5.2, we did not consider boosting trees, boosting stumps or quantile regression forests as competitors since these do not assume a linear model and would therefore naturally lead to a degraded fit when being compared to approaches that assume a linear model a priori.

Performance measures

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:

$$\text{Bias}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau jk} - \beta_{\tau j}) \qquad \text{MSE}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau jk} - \beta_{\tau j})^2 \,, \tag{5.2}$$

where $k = 1, \ldots, 100$ indexes the simulation replication and $j = 0, \ldots, p$ the number of covariates. Note that when the mean bias and MSE over all 100 iterations are calculated, those values can be interpreted as Monte Carlo estimators of the true bias and MSE of the nonlinear functions. In case of boosting, we also considered the empirical distribution of m_{stop}.

Performance results

In the following, we will focus on a short summary of the results by just showing some typical examples. Figure 5.2 displays boxplots for the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$ in the heteroscedastic setup with Gaussian 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 with previously reported results and with the fact that boosting estimators are usually shrunken towards zero. This can be traced back to the implicit regularization property of boosting estimation as discussed in Section 4.4.

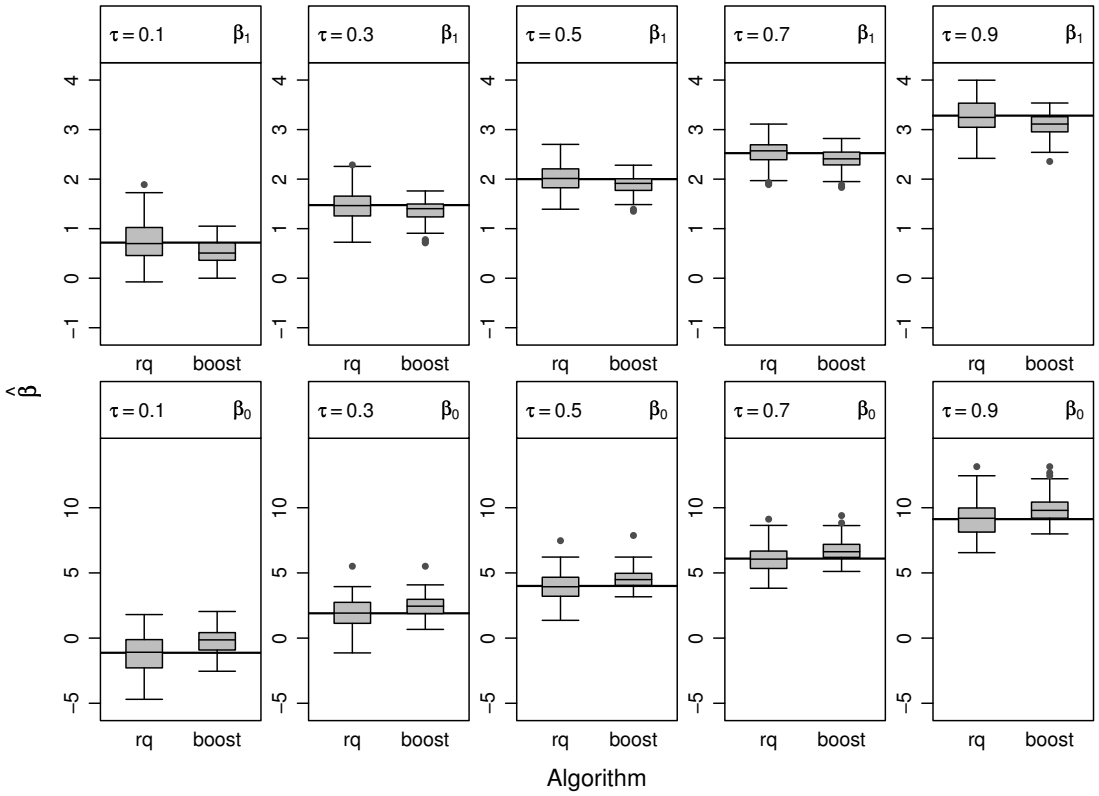


Figure 5.2 Simulation results for heteroscedastic linear setup with one covariate and Gaussian distributed error terms. Boxplots display the empirical distribution of the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$ from 100 replications, depending on quantile τ and estimation algorithm (`rq` for linear programming and `boost` for boosting). Horizontal lines designate true underlying parameters $(\beta_{\tau 0}, \beta_{\tau 1})^\top$.

Regarding the MSE, Table 5.1 shows estimators for setups with one covariate and gamma distributed error terms, obtained according to (5.2). 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.

Table 5.1 Estimated MSE criteria from 100 replications of linear simulation setups with one covariate and gamma distributed error terms. Quantile- and parameter-specific smaller estimators are shown in bold.

	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

In addition, Table 5.2 shows mean m_{stop} criteria for all setups with t-distributed error terms. 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 (30 000 – 70 000) for the multivariable model with six covariates. This again shows that with small step lengths, a large number of boosting iterations may be necessary to approximate the final quantile fit (as discussed in Section 4.3).

Table 5.2 Mean m_{stop} criteria from 100 replications of linear simulation setups with t-distributed error terms.

Setup	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Homoscedastic	10 886	6 415	7 372	4 762	3 589
Heteroscedastic	4 183	8 935	9 133	10 039	6 387
Multivariable	68 055	43 883	40 541	42 317	30 255

We observed similar results for all other simulation setups, i.e., with more covariates or alternative error distributions. Therefore, we conclude that quantile boosting works correctly for linear quantile regression.

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 AIC values as given in equation (3.7) on page 42. Then, the covariate combination with the smallest AIC 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?

Regarding these questions, 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 occurred at upper quantiles in the setup with gamma distributed error terms, but in these cases also the AIC-driven model selection did not yield the correct model. To exemplify these results, Table 5.3 gives a summary for Gaussian distributed error terms and quantile $\tau = 0.7$. It can be observed that the covariates x_5 and x_6 with both no influence on the response, i.e., $\beta_{0.7,5} = \beta_{0.7,6} = 0$, are chosen less frequent and later than all other covariates. However, variable selection by AIC strictly excludes non-significant covariates more often than boosting.

Table 5.3 Summary of variable selection results for $\tau = 0.7$ from linear multivariable simulation setup with Gaussian 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 excluded by boosting
PEA: Proportion of simulations (relating to 100) where covariate was excluded in model with smallest AIC (based on linear programming estimation).

		Int. $\beta_0 = 5.5$	x_1 $\beta_1 = 8.0$	x_2 $\beta_2 = -4.0$	x_3 $\beta_3 = 2.0$	x_4 $\beta_4 = -1.5$	x_5 $\beta_5 = 0$	x_6 $\beta_6 = 0$
`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	0.67	0.79

To sum up, boosting provides useful support in the variable selection process even though there are currently no explicit “hard” criteria available to assess variable importance. Particularly in cases with numerous covariates, boosting has the advantage that it yields information on variable selection within the estimation process, whereas the use of AIC requires multiple model fits.

5.2 Simulation study for additive quantile regression

The contents of this section are mainly based on the empirical evaluation results presented in Section 3 of Fenske, Kneib, and Hothorn (2011).

Data generating process

For the additive simulation settings, we considered the following model:

$$y_i = \beta_0 + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + [\alpha_0 + g_1(z_{i1}) + \ldots + g_q(z_{iq})]\,\varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{iid}{\sim} F_\varepsilon \,. \tag{5.3}$$

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 F_ε not depending on covariates. Choosing all f_j and g_j as linear functions yields the linear model which was addressed in Section 5.1. 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

$$Q_{Y_i}(\tau|\boldsymbol{z}_i) = \beta_0 + f_1(z_{i1}) + \ldots + f_q(z_{iq}) + F_\varepsilon^{-1}(\tau)[\alpha_0 + g_1(z_{i1}) + \ldots + g_q(z_{iq})] \,. \tag{5.4}$$

Based on the additive model in (5.3), we considered 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	$0.7z_{i1}$

and a multivariable setup with $q = 6$:

β_0	$f_1(z_{i1})$	$f_2(z_{i2})$	$f_3(z_{i3})$	$f_4(z_{i4})$	$f_5(z_{i5})$	$f_6(z_{i6})$
2	$1.5\sin(\frac{2}{3}z_{i1})$	$1.5\log(z_{i2} + 1.05)$	$2z_{i3}$	$-2z_{i4}$	0	0

α_0	$g_1(z_{i1})$	$g_2(z_{i2})$	$g_3(z_{i3})$	$g_4(z_{i4})$	$g_5(z_{i5})$	$g_6(z_{i6})$
0.5	$0.5z_{i1}^2$	$0.5z_{i2}$	$0.5z_{i3}$	0	0	0

In the multivariable setup, two covariates (z_1 and z_2) relate nonlinearly to the response, two covariates (z_3 and z_4) have a linear influence on it, and the last two (z_5 and z_6) 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 $\mathrm{Cov}(z_{ik}, z_{il}) = \rho^{|k-l|}$ for possible correlation coefficients $\rho \in \{0, 0.2, 0.5, 0.8\}$. Similar to the linear model simulations, we repeated each setup for different distributions of the error term: a standard Gaussian $\mathcal{N}(0, 1)$, a t distribution with two degrees of freedom $t(2)$, and a gamma distribution $\mathcal{G}(1, 2)$, where $\mathbb{E}(\varepsilon_i) = \mathbb{V}(\varepsilon_i) = 2$. Figure 5.3 shows data examples from $\sin$- and $\log$-setups for Gaussian and gamma distributed error terms.

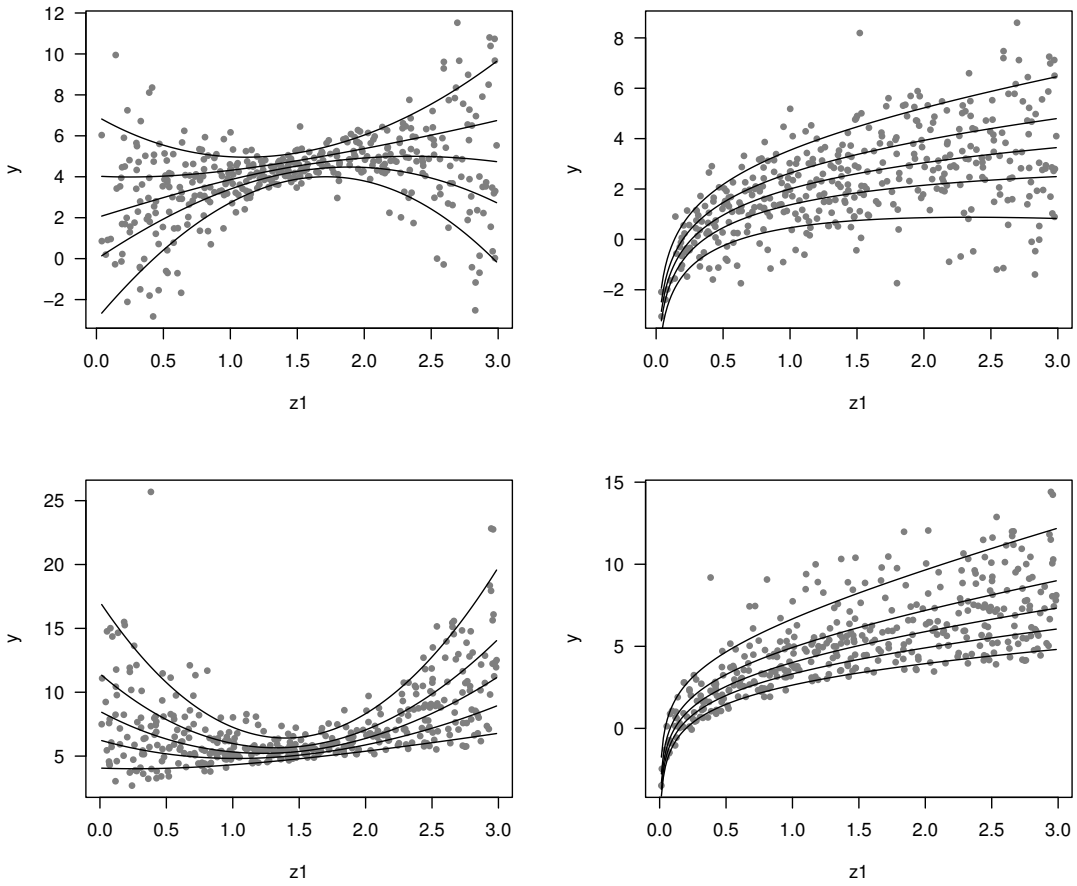


Figure 5.3 Data examples for nonlinear simulation setups with $n = 400$ observations (grey points) and one covariate in the $\sin$-setup (left) or $\log$-setup (right) with standard Gaussian distributed (top) or gamma distributed (bottom) error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

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 optimal 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 *et al.*, 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 (`rqforest`, 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 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 (3.5) on page 40 in the case of `rqss`.

Performance measures

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

$$\text{Risk}(\tau, k) = \frac{1}{1000} \sum_{i=1}^{1000} \rho_\tau(y_i - \hat{y}_{\tau k i}) \qquad \text{for} \quad k = 1, \ldots, 100 \,,$$

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 and observation i. Analogously, quantile- and iteration-specific bias and MSE were estimated as

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

with $y_{\tau i}$ denoting the true underlying τ-th quantile of the response of observation i which can be calculated according to (5.4).

Performance results

To illustrate the results for univariate setups, Figure 5.4 exemplarily displays estimated quantile curves for the $\sin$-setup with standard Gaussian 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.

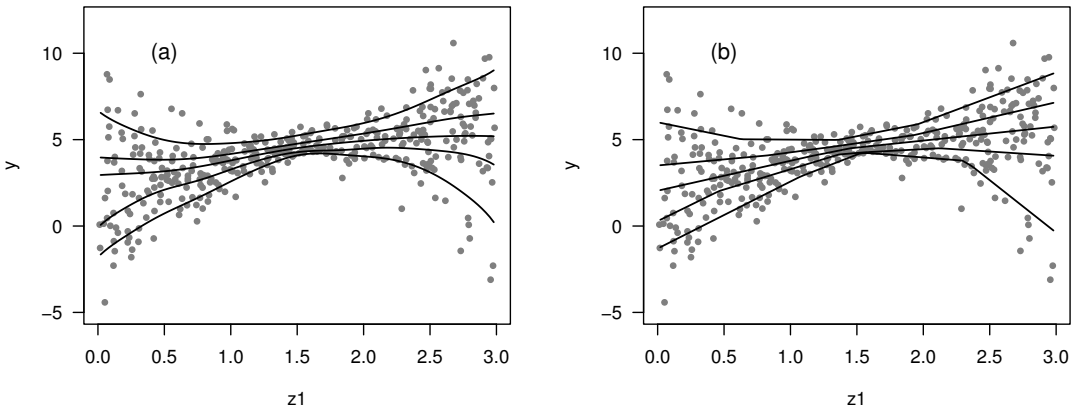


Figure 5.4 One example for resulting estimated quantile curves for the $\sin$-setup with standard Gaussian distributed error terms. Plot (a) displays curves obtained from `gamboost` whereas plot (b) displays curves obtained from `rqss`. True underlying quantile curves are shown in the upper left plot in Figure 5.3.

We show representative results for the multivariable 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 of Fenske *et al.* (2011). Figure 5.5 shows quantile- and algorithm-specific empirical distributions of the resulting performance criteria while Table 5.4 displays the respective means.

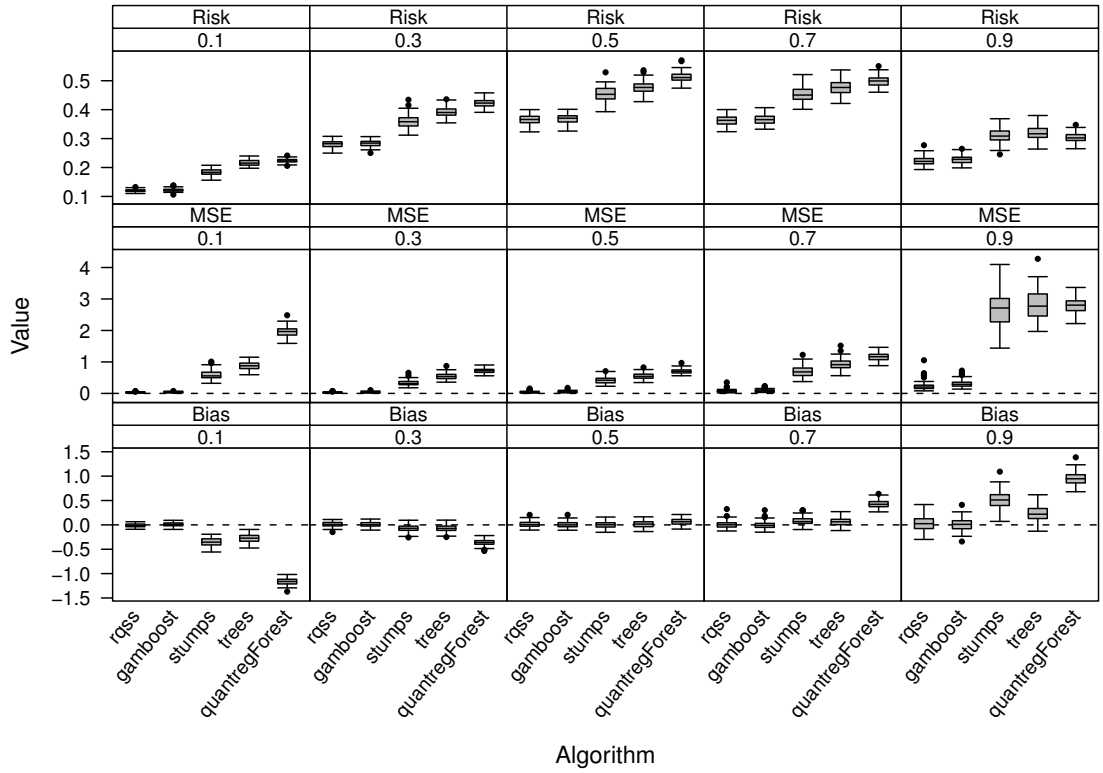


Figure 5.5 Simulation results for the multivariable setup with gamma distributed error terms and a correlation coefficient of 0.5. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

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 both algorithms in comparison with the others can be explained by the specific design of our simulation study. The underlying data structure of our setup corresponds to an additive model without interaction effects which is also assumed for quantile regression estimation with `rqss` and `gamboost`. Therefore, it is hardly surprising that these methods perform better than `stumps`, `trees`, and `rqforest` which in turn work as 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.

Table 5.4 Mean estimated performance criteria from 100 replications from the multivariable setup with gamma distributed error terms and a correlation coefficient of 0.5. Quantile-specific smallest values are printed in bold.

		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
	`rqforest`	0.223	0.422	0.513	0.499	0.303
		0.1	0.3	0.5	0.7	0.9
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
	`rqforest`	1.961	0.720	0.703	1.159	2.782
		0.1	0.3	0.5	0.7	0.9
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
	`rqforest`	-1.162	-0.361	0.064	0.429	0.953

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 non-informative covariates, i.e., $f_k(z_{ik}) = g_k(z_{ik}) \equiv 0$ for $k = 5, \ldots, K$. We considered the three cases $K \in \{6, 16, 20\}$ since the estimation with `rqss` was not possible with more than 20 non-informative covariates.

To exemplify the results, Figure 5.6 displays boxplots of the performance criteria for the higher-dimensional setup with $K = 20$ non-informative 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 regard to variable selection – contrary to tree-based approaches just yielding black boxes.

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 $1, \ldots, 100$. 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 non-informative covariates.

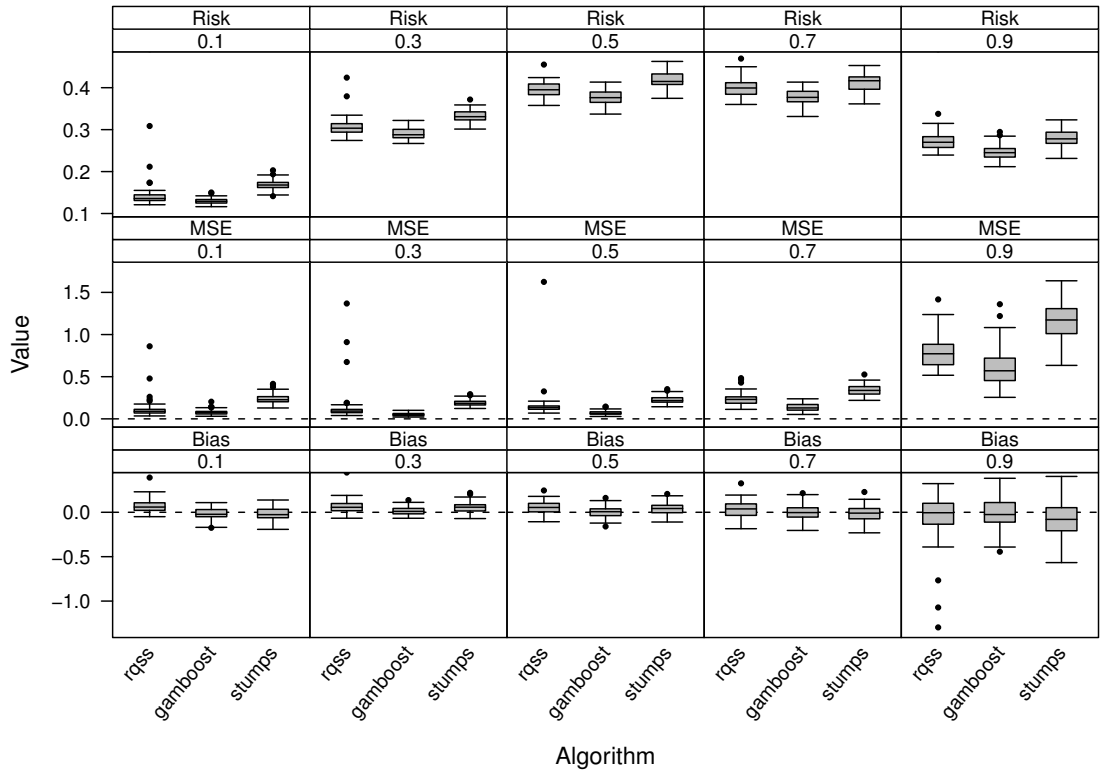


Figure 5.6 Simulation results for the higher-dimensional setup with $K = 20$ non-informative covariates, gamma distributed error terms and a correlation coefficient of 0.5. Boxplots display empirical distributions of the performance criteria from 100 replications, depending on quantile τ and three estimation algorithms.

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 non-informative 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, i.e., `gamboost` and `stumps`, we explored the variable selection results in more detail. Regarding the same higher-dimensional setup as above, Figure 5.7 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 5.8 visualizes the proportion of iterations in which a base learner was selected. Base learners of non-informative covariates $z_5, \ldots, z_{24}$ are compressed in one category. The figures illustrate that for both algorithms and independent of τ, non-informative covariates are selected less frequent and later for the first time during the estimation process than informative covariates $z_1, \ldots, z_4$. These results further substantiate the advantages of boosting with regard to variable selection in high-dimensional setups.

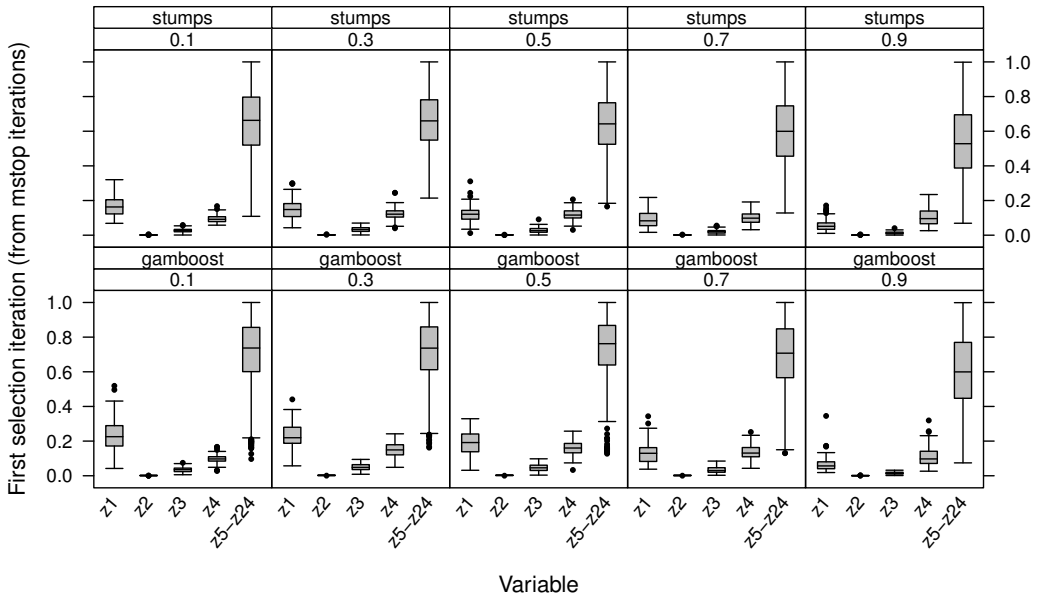


Figure 5.7 Simulation results for the higher-dimensional setup with $K = 20$ non-informative covariates, gamma distributed errors and a correlation coefficient of 0.5. Boxplots display for each base learner $z_1, \ldots, z_{24}$ the empirical distribution of *the first selection iteration* relative to the optimized m_{stop} from 100 simulation replications, depending on τ and estimation algorithm.

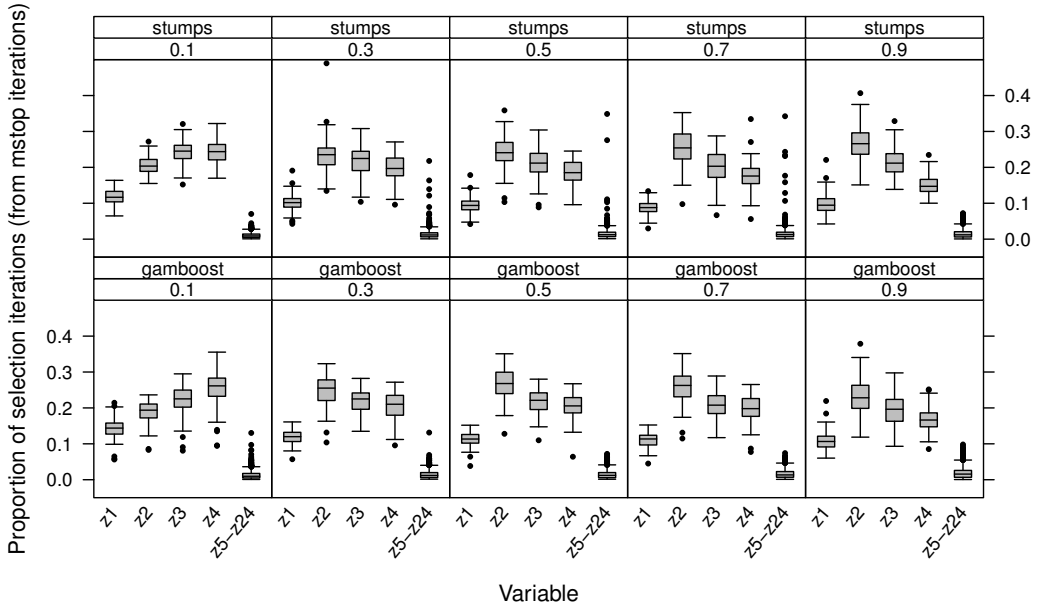


Figure 5.8 Simulation results for the higher-dimensional setup with $K = 20$ non-informative covariates, gamma distributed errors and a correlation coefficient of 0.5. Boxplots display for each base learner $z_1, \ldots, z_{24}$ the empirical distribution of the *proportion of selection iterations* relative to the optimized m_{stop} from 100 simulation replications, depending on τ and estimation algorithm.

5.3 Comparing estimated quantile functions

As an alternative way to evaluate the performance of quantile boosting, we now focus on comparing estimated quantile functions with the true quantile function directly, instead of comparing the out-of-sample empirical risks as in Section 5.2.

Data generating process

For one single covariate $Z \sim \mathcal{U}(0, 1)$ we sampled response values from the conditional distribution

$$Y|Z = z \sim \text{BCPE}(\mu = \sin(2\pi z) + 3, \sigma = \exp(z^2 + 0.1), \nu = 3z, \varphi = \exp(3z + 2)) \,, \tag{5.5}$$

where BCPE refers to a Box-Cox power exponential distribution (Rigby and Stasinopoulos, 2004). In this distribution, μ controls the median, σ the coefficient of variation, ν the skewness, and φ the kurtosis of the response's distribution. Thus, the first four moments of the response variable vary with covariate Z in a smooth nonlinear way, and the corresponding conditional density is depicted in Figure 5.9.

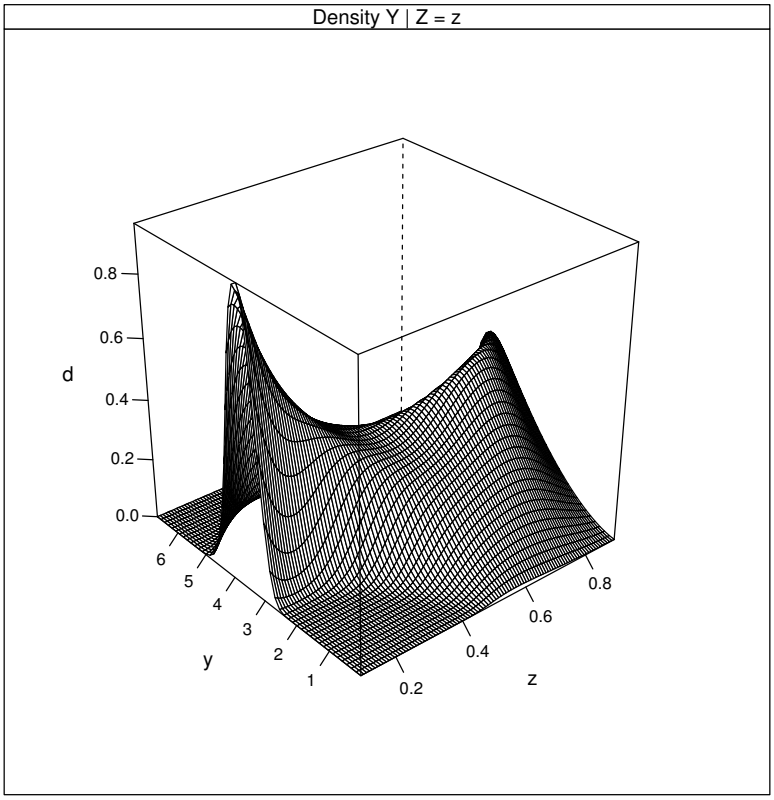


Figure 5.9 Conditional density of $Y|Z = z \sim \text{BCPE}(\mu, \sigma, \nu, \varphi)$ with parameters according to (5.5).

Estimation

Based on a training sample containing $n = 200$ observations and one additional test sample with $n = 100$ observations, we estimated univariate additive quantile regression models with the covariate z, both based on `rqss` and `gamboost` with appropriate hyperparameter tuning.

Performance measures

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 on the grid $[0.1, 0.9]$ (see (3.13) on page 57 for the exact formula). To compare estimated and true quantile functions we calculated the sum of the absolute differences

$$\sum_\tau \sum_z |\hat{f}_\tau(z) - f_\tau(z)| \,,$$

which corresponds to the sum of absolute deviations from the bisecting line in a quantile-quantile plot.

Performance results

Figure 5.10 shows that the quantile functions estimated by `gamboost` approximate the true quantile function better than `rqss`. Of course, this can be attributed 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 5.11 showing the true and estimated quantile functions.

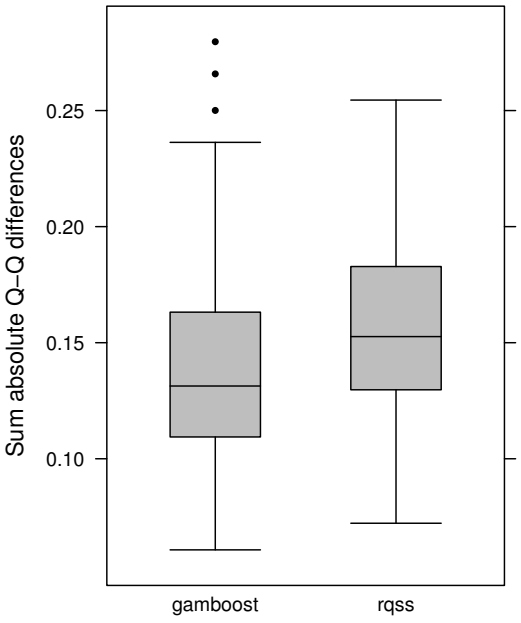


Figure 5.10 Empirical distribution of the sum of absolute deviations between estimated (by `gamboost` and `rqss`) and true quantile function obtained from simulation model (5.5) over a grid of quantile values τ and covariate values z based on 100 simulation replications.

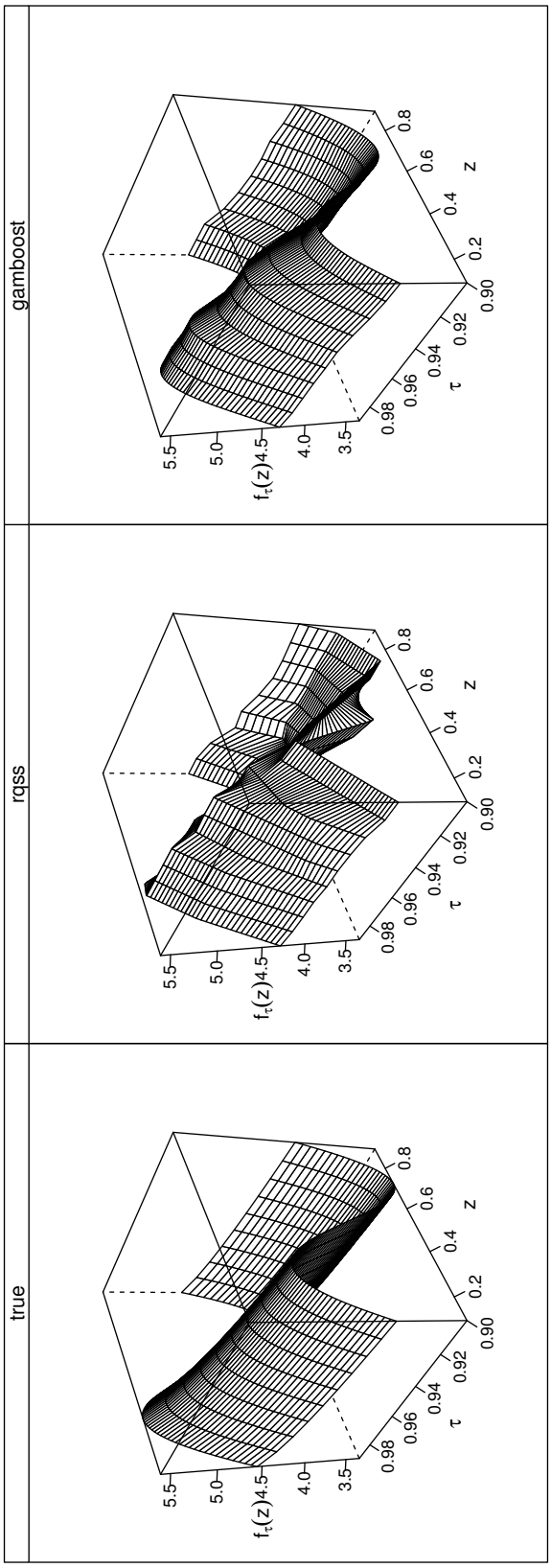


Figure 5.11 Example of true and estimated quantile functions for grids of τ and z values based on the simulation model (5.5). `gamboost` captures the true quantile function better than `rqss`.

5.4 Quantile boosting for individual-specific effects

The aim of the present section is to get a basic understanding of cluster-specific effects estimated by quantile boosting with ridge-penalized base learners (as described in Section 4.2). We are mainly interested in the interpretation and the shrinkage character of these cluster-specific effects and do not evaluate the performance in comparison with other estimation approaches here. In our obesity application with longitudinal data the clusters will correspond to individuals.

Simulation example

We conducted a small simulation example with the simplest case of longitudinal data to illustrate the interpretation of quantile- and individual-specific effects. We simulated longitudinal data from the model $y_{ij} = \beta_0 + b_i + \varepsilon_{ij}$ with $i = 1, \ldots, 10$, individuals and $j = 1, \ldots, 10$, observations per individual, where $b_i \sim \mathcal{N}(0, 4)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$ were independently drawn from Gaussian distributions with different variances. Then, we fitted two STAQ models by quantile boosting for $\tau = 0.50$ and $\tau = 0.75$, containing only one base learner for the individual-specific intercepts $b_{\tau i}$. The estimated population intercept $\hat{\beta}_{\tau 0}$ could then be obtained as sum of the offset and of the mean of the individual-specific intercepts.

Figure 5.12 shows the results for the median in panel (a) and the results for the 75% quantile in panel (b). It can be observed that the estimated individual-specific intercepts differ for different quantiles. Also, the sums of the estimated population effects and individual-specific intercepts are almost equal to the respective individual-specific empirical quantiles given by the boxplots.

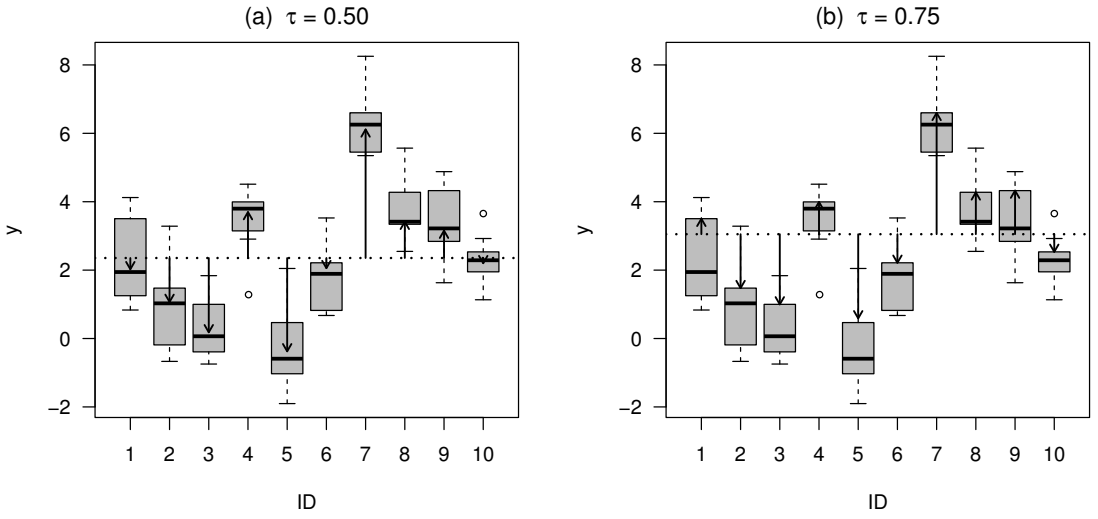


Figure 5.12 Boxplots display empirical distributions of simulated data for 10 individuals with 10 observations each. (a) Results from an STAQ model for $\tau = 0.50$ containing only a population intercept and individual-specific intercepts. (b) Results from a similar STAQ model for $\tau = 0.75$. Dotted horizontal lines correspond to estimated intercepts $\beta_{\tau 0}$, whereas individual vertical arrows stand for estimated individual-specific intercepts as deviations from the population intercept.

Conclusions from Figure 5.12

This simplified example shows that the individual-specific effects of the STAQ model for longitudinal data can be interpreted in accordance with the conditional view on random effects in additive mixed models: Similar to individual-specific conditional means scattering around the population mean in additive mixed models, individual-specific conditional $\tau \cdot 100\%$ quantiles scatter symmetrically around the expected individual-specific conditional $\tau \cdot 100\%$ quantile (which is the population intercept $\beta_{\tau 0}$) in STAQ models. The Gaussian scattering character of the individual-specific quantile effects can be attributed to the quadratic form of the penalty term which from a Bayesian perspective correspond to the log-density of Gaussian random effects priors.

Extensions of the model to covariates and individual-specific slopes would not change this interpretation of individual-specific effects. Similar to the intercepts, individual-specific slopes would scatter in a Gaussian way around the quantile-specific population slope for a given quantile parameter τ. The inclusion of covariates would just change the population part of the predictor which – in our very simple model – only consists of the population intercept β_0.

The illustration in Figure 5.12 additionally points out that the interpretation of population effects estimated by quantile regression is conditional on individual-specific effects. This corresponds to the conditional interpretation of additive mixed models with the conditional quantile function (3.12) presented on page 55 in Section 3.5.2.

Furthermore, we sampled the data with an individual-specific location shift on the response distribution only. However, Figure 5.12 shows that the individual-specific empirical distributions slightly differ regarding their variation. Boosting estimation can account for these individual-specific distributional shapes by estimating different individual-specific effects for different quantiles. One can imagine that these quantile- and individual-specific effects would make even more sense in situations with individual-specific scale or skewness shifts of the response distribution that cannot be explained by further covariates.

Boosting details for individual-specific effects

Even though the above simulation example might seem to oversimplify the situation at first glance, recall that a base learner to fit individual-specific components is exactly based on such a simple model. The effects are estimated by ridge-penalized least squares with the binary negative gradient residuals as response.

If extreme quantile parameters are of interest, one can imagine that a large number of observations per individual is needed to estimate individual-specific quantiles. With, for example, five observations per individual, it is hard for the algorithm to differentiate between the 80-100% quantiles.

With the R package `mboost` (Hothorn *et al.*, 2012), the model for the 75% quantile can be estimated by the following call:

```
model75 <- gamboost(y ~ brandom(id, df = 4), family=QuantReg(tau=0.75),
                        control=boost_control(nu=0.1, mstop=1000))
```

The base learner function for the individual-specific effect is called `brandom()` with the default of four degrees of freedom `df = 4`. The name `brandom()` for this base learner should underline

the similarities to fitting random effects in classical mixed models for longitudinal data. However, the name is probably slightly misleading since the estimated effects can rather be interpreted as individual-specific shrunken fixed effects than as being random.

To illustrate the shrinkage character of the individual-specific effects estimated by boosting, we carried out the same simulation example as above for the 10% quantile with 10 individuals and 10 observations each. Figure 5.13 shows the resulting paths of estimated individual-specific quantiles depending on the number of boosting iterations. It can be observed that the estimated individual-specific quantiles approach the empirical individual-specific quantiles with increasing number of iterations. Stopping the algorithm somewhere before iteration 200 would lead to shrunken estimators.

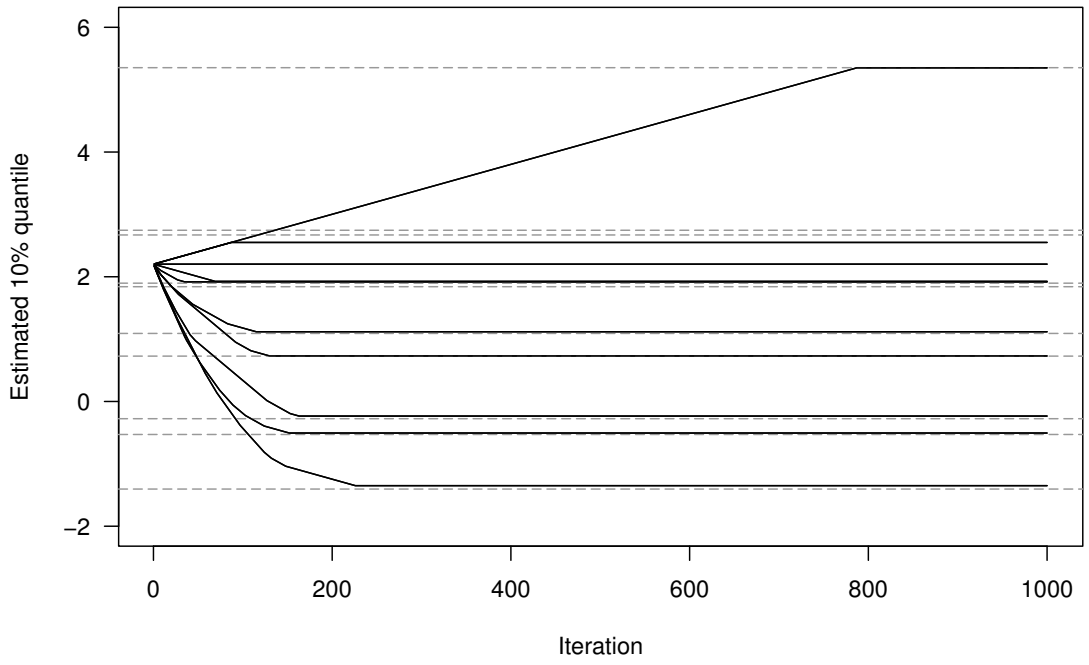


Figure 5.13 Paths of estimated individual-specific 10% quantiles (black solid lines) over 1000 boosting iterations. Dashed grey lines show the empirical individual-specific 10% quantiles.

Figure 5.13 also displays that the paths of several individuals are exactly similar during the first iterations. This clustering is typical for effects estimated by ridge penalization and can be attributed to the grouping property of the ridge estimator (see, for example, Zou and Hastie, 2005). We observed more extreme grouping effects in our obesity application with many children and the quantile parameters 0.90 and 0.97. There, some children had the same individual-specific intercepts during a large number of successive iterations.

Chapter 6: Quantile boosting for child undernutrition in India

In this chapter, we describe our quantile boosting analysis to investigate determinants of child undernutrition in India. Background and dataset of this application were introduced in Section 2.1, while the following contents are mainly based on the manuscript Fenske, Burns, Hothorn, and Rehfuess (2012a); we present additional results here.

6.1 Setup of the analysis

As described in Section 2.1, the main objective of our first application was a comprehensive analysis of the multiple determinants of child stunting based on a large-scale cross-sectional dataset from India. To capture the multi-factorial nature of child stunting, we developed a schematic diagram prior to the analysis in which the various potential risk factors were structured (see Figure 2.1, page 19). This diagram served as a basis for selecting appropriate covariates from the dataset (shown in Table 2.1, page 24).

To assess the impact of the resulting large number of stunting determinants on quantiles of the Z-score, we used the following structured additive predictor:

$$\begin{aligned} Q_{Y_i}(\tau|\cdot) = \eta_{\tau i} = \beta_{\tau 0} + \beta_{\tau 1} x_{i1} + \ldots + \beta_{\tau p} x_{ip} \\ + f_{\tau 0}(\text{age}_i) + f_{\tau 1}(z_{i1}) + \ldots + f_{\tau q}(z_{iq}) \\ + v_{i1} \cdot g_{\tau,1}(\text{age}_i) + \ldots + v_{ir} \cdot g_{\tau,r}(\text{age}_i) \\ + f_{\tau,spat}(u_i) \; . \end{aligned}$$

Thus, for a fixed quantile parameter τ and observation $i = 1, \ldots, n$, the additive predictor $\eta_{\tau i}$ models the conditional quantile function $Q_{Y_i}(\tau|\cdot)$ of the response variable Y_i, being the height-for-age Z-score in our analysis. Note that this predictor is written in accordance with example 1 for the generic structured additive predictor on page 37.

The quantile-specific flexible additive predictor $\eta_{\tau i}$ comprises linear effects $\beta_{\tau 0}, \ldots, \beta_{\tau p}$ for categorical covariates $x_1, \ldots, x_p$, and smooth (potentially) nonlinear functions $f_{\tau 0}, \ldots, f_{\tau q}$ for age and further continuous covariates $z_1, \ldots, z_q$. Also specified are (potentially) nonlinear age-varying effects $g_{\tau,1}, \ldots, g_{\tau,r}$ for different levels of the feeding variables $v_1, \ldots, v_r$, allowing meaning and effect of breastfeeding and complementary feeding to vary with age (according to WHO recommendations; Habicht, 2004). For the categorical variable u, corresponding to 29 Indian states, a discrete spatial effect $f_{\tau,spat}$ is estimated to account for spatial autocorrelation and unobserved heterogeneity.

We specified four different values for the quantile parameter, namely $\tau \in \{0.05, 0.15, 0.35, 0.50\}$. The two values $\tau = 0.15$ and $\tau = 0.35$ were chosen based on the empirical frequencies for stunting (37%) and severe stunting (17%) in our dataset (see Table 2.1) since this choice allows results to be compared across quantile and binary regression models. The value $\tau = 0.05$ was chosen in order to provide a complete picture of covariate effects on the lower tail of the response

distribution. Furthermore, the median ($\tau = 0.50$) allows for a comparison of our results to those from previous mean regression analyses (Kandala *et al.*, 2001, 2009).

Model estimation was undertaken separately for each τ using quantile boosting. More precisely, the above model was realized by the following (shortened) model call for the estimation in R:

```
library(mboost)
boostform <- stunting ~ bols(intercept, intercept=FALSE) +
                        bols(csex, df=1) + bols(ctwin, df=1) +
                        ... +
                        bols(cagec, intercept=FALSE) + bbs(cagec, center=TRUE, df=1) +
                        bols(mbmic, intercept=FALSE) + bbs(mbmic, center=TRUE, df=1) +
                        ... +
                        bmrf(region, bnd=neighbourMat, df=1, center=TRUE) +
                        ... +
                        bols(cagecbreastcf, intercept=FALSE) +
                        bbs(cagecbreastcf, center=TRUE, df=1) +
                        bols(cagecbreastex, intercept=FALSE) +
                        bbs(cagecbreastex, center=TRUE, df=1) +
                        ...

boostmodel <- gamboost(boostform, data = india, family = QuantReg(tau = 0.35),
                       control = boost_control(mstop = 10000, nu=0.2, trace = TRUE))
```

Thus, we used the function `gamboost` from package `mboost` (Hothorn *et al.*, 2012) with option `family=QuantReg()` to estimate separate quantile regression models for four different quantile parameters. As can be seen from the above model call, we defined separate base learners for all covariates as well as a base learner for the intercept. For categorical covariates, such as child sex and twin, base learners were specified using the function `bols()`. Continuous covariates, as for example child age and maternal BMI, were mean-centered before the analysis and split into two base learners: `bols()` with the option `intercept = FALSE` for the linear part and `bbs()` with the option `center = TRUE` for the nonlinear deviation. The smooth spatial effect for the region was estimated by the base learner `bmrf()`, which required a suitable matrix for the neighbourhood structure. The age-varying effects of feeding variables were estimated separately for each level of the categorical feeding variables. Similar to the smooth nonlinear effects of continuous covariates, two separate base learners `bols()` and `bbs()` were specified for each age-varying effect. This decomposition allowed for data-driven decision on linearity vs. nonlinearity of the effects. To make the complexity of the base learners comparable, we set the degrees of freedom of each base learner to one, i.e., $df(\lambda_d) = 1$.

Regarding further parameters of the boosting algorithm, we determined the optimal number of iterations by five-fold cross-validation (code not shown) and set the step length to $\nu = 0.2$. Model estimation was then repeated on 100 bootstrap samples of the dataset to calculate $(1 - \alpha)\%$ bootstrap confidence intervals $[\hat{q}_{j,\alpha/2}, \hat{q}_{j,1-\alpha/2}]$ for the estimators of categorical covariates, where $\hat{q}_{j,\alpha/2}$ denotes the estimated $\alpha/2 \cdot 100\%$ quantile of $\hat{\beta}_{\tau j}$ with $j = 1, \ldots, p$.

We additionally applied a stability selection procedure to allow for formal variable selection (Meinshausen and Bühlmann, 2010) and to support the inherent variable selection property provided by boosting. The idea of this procedure is to control the family-wise error rate, which corresponds to the α or type I error for multiple testing and denotes the probability that at least

one null hypothesis is classified as false negative. In our analysis, we chose a family-wise error rate of 5% and an average number of 15 terms to be expected in a model.

As mentioned in Section 2.1, logistic regression for binarized versions of the height-for-age Z-score is the most commonly used approach in literature to analyze determinants of child stunting. In order to compare the results of our innovative quantile regression approach with this standard approach, we additionally conducted two logistic regression analyses for the binary variables *stunting* and *severe stunting* (see Section 2.1 for details on their construction). In these logistic regression models, we used exactly the same structured additive predictor as in the quantile regression analyses. Estimation was based on boosting with the negative Binomial log-likelihood as loss function. In the R package `mboost`, this could be realized by specifying the argument `family=Binomial()` with the logit link as default.

Note that we did not use other estimation approaches for quantile regression than boosting in our analysis since both our empirical evaluations and preliminary analyses of child stunting in India (see Fenske *et al.*, 2011) showed that quantile boosting outperforms other estimation approaches for STAQ regression in the present application context.

6.2 Results

Results for the 35% Z-score quantile

First, we thoroughly describe the results of quantile regression for the 35% Z-score quantile. All findings on effects of single variables that are described in the following are fully adjusted for other variables.

Table 6.1 presents the effects for categorical covariates and their 95% bootstrap confidence intervals, and summarizes the shape of the estimated functions for continuous variables. An effect of a categorical covariate is rated as 'significant' (and therefore printed in bold) if the 95% bootstrap confidence interval does not contain zero.

Based on this criterion, one can observe that except for indoor air pollution, at least one variable in each of the assessed groups of determinants shows a statistically significant association with the 35% Z-score quantile. Furthermore, the following categorical covariates have at least one significant category compared with the reference category: child sex, household wealth, caste of household head, mother is currently working, child is twin, sanitation facility, vaccination index, vitamin A and iodine. Regarding the interpretation of these effects, for example, the 35% Z-score quantile for children from the richest households is significantly increased by 0.224 compared to children from the poorest households (given all other covariates are equal). Being a twin has a very large significant negative effect of -0.866 which is remarkable as only 1.1% (n=139) of the children in the dataset are twins or multiple births (see Table 2.1). Religion of household head, partner's occupation, sex of household head, urban/rural location, drinking water source, type of cooking fuel, and iron supplementation do not show an effect.

Table 6.1 Estimated effects and 95% bootstrap confidence intervals of quantile boosting models for $\tau = 0.35$ (columns in grey) and $\tau = 0.15$; see Figures 6.1 and 6.2 for detailed results of continuous covariates. Significant effects are shown in bold.

Variable	Values / Description	Quantile regression for 35% quantile		Quantile regression for 15% quantile	
		$\beta_{0.35}$	95% CI($\beta_{0.35}$)	$\beta_{0.15}$	95% CI($\beta_{0.15}$)
Non-modifiable factors					
Child age [months]		**~ Linear, negative**		**~ Linear, negative**	
Child sex	Male	-	-	-	-
	Female	**0.166**	**[0.103, 0.234]**	**0.209**	**[0.130, 0.285]**
Maternal characteristics					
Maternal age [years]		**Nonlinear, inverse U**		**Nonlinear, inverse U**	
Maternal BMI [kg/m^2]		**Nonlinear, positive**		**Nonlinear, positive**	
Household characteristics					
Household wealth	Poorest	-	-	-	-
	Poorer	0.025	[-0.077, 0.110]	0.035	[-0.041, 0.129]
	Middle	0.058	[-0.014, 0.161]	0.001	[-0.067, 0.079]
	Richer	0.089	[-0.016, 0.205]	0.075	[-0.014, 0.207}
	Richest	**0.224**	**[0.069, 0.383]**	**0.214**	**[0.060, 0.367]**
Religion of household head	Hindu	-	-	-	-
	Muslim	0.003	[-0.064, 0.086]	0.003	[-0.075, 0.101]
	Christian	0.034	[-0.023, 0.139]	0.089	[-0.001, 0.222]
	Sikh	0.021	[-0.009, 0.116]	0.068	[-0.001, 0.180]
	(Neo-)Buddhist	0.000	[-0.032, 0.034]	-0.006	[-0.085, 0.066]
	Other	-0.006	[-0.064, 0.028]	-0.030	[-0.132, 0.028]
Caste/tribe of household head	Scheduled caste	-	-	-	-
	Scheduled tribe	**0.088**	**[0.005, 0.224]**	0.037	[-0.060, 0.156]
	Other backward class	**0.112**	**[0.034, 0.214]**	**0.115**	**[0.011, 0.213]**
	None of them	**0.165**	**[0.062, 0.294]**	**0.167**	**[0.049, 0.302]**
Maternal education [years]		**~ Linear, positive**		**~ Linear, positive**	
Partner's education [years]		**~ Linear, positive**		**~ Linear, positive**	
Partner's occupation	Services	-	-	-	-
	Household & domestic	0.035	[-0.021, 0.132]	0.055	[-0.002, 0.179]
	Agriculture	0.028	[-0.031, 0.104]	0.042	[-0.015, 0.136]
	Clerical	0.013	[-0.039, 0.079]	0.005	[-0.059, 0.077]
	Prof./ Tech./ Manag.	0.037	[-0.015, 0.132]	-0.011	[-0.105, 0.069]
	Did not work	0.009	[-0.062, 0.082]	-0.009	[-0.092, 0.049]
Mother is currently working	No	-	-	-	-
	Yes	**-0.078**	**[-0.152, -0.001]**	-0.044	[-0.122, 0.018]
Sex of household head	Male	-	-	-	-
	Female	0.029	[-0.033, 0.124]	0.023	[-0.037, 0.113]
Regional characteristics					
State of residence		**Spatial, see Figures 6.3 and 6.4**		**Spatial**	
Urban/rural location	Urban	-	-	-	-
	Rural	-0.002	[-0.074, 0.071]	0.025	[-0.076, 0.113]
Household food competition					
Number of household members		Nonlinear, inverse U		Nonlinear, inverse U	
Birth order		**Nonlinear, negative**		**Nonlinear, negative**	
Preceding birth interval [months]		**Nonlinear, positive**		Nonlinear, positive	
Child is twin or multiple birth	No	-	-	-	-
	Yes	**-0.866**	**[-1.107, -0.456]**	**-0.890**	**[-1.173, -0.497]**

Variable	Values / Description	Quantile regression for 35% quantile		Quantile regression for 15% quantile	
		$\beta_{0.35}$	95% CI($\beta_{0.35}$)	$\beta_{0.15}$	95% CI($\beta_{0.15}$)
Water, sanitation and hygiene					
Drinking water in household	Unimproved	-	-	-	-
	Improved	-0.026	[-0.093, 0.015]	-0.004	[-0.056, 0.051]
	Piped	-0.007	[-0.078, 0.026]	0.003	[-0.036, 0.043]
Sanitation facility in household	Unimproved	-	-	-	-
	Improved	**0.092**	**[0.041, 0.159]**	**0.114**	**[0.031, 0.227]**
Indoor air pollution					
Type of cooking fuel	Straw/ crop /animal dung	-	-	-	-
	Coal/ charcoal/ wood	-0.040	[-0.090, 0.015]	-0.031	[-0.105, 0.027]
	Kerosene	-0.020	[-0.081, 0.007]	**-0.056**	**[-0.164, -0.001]**
	Gas/ electricity	0.055	[-0.009, 0.170]	**0.076**	**[0.001, 0.179]**
Curative and preventive healthcare					
Vaccination index	None (0)	-	-	-	-
	Low (1-3)	-0.015	[-0.079, 0.033]	0.010	[-0.053, 0.073]
	Medium (4-6)	-0.026	[-0.081, 0.043]	-0.031	[-0.100, 0.033]
	High (7-9)	**0.062**	**[0.004, 0.137]**	**0.080**	**[0.007, 0.175]**
Number of antenatal visits during pregnancy		**Nonlinear, inverse U**		**Nonlinear, inverse U**	
Breastfeeding practices					
Breastfeeding	No breastfeeding	-	-	-	-
	Breastfeeding + complementary feeding	**Nonlinear, negative by age**		**Nonlinear, negative by age**	
	Exclusive breastfeeding	**Nonlinear, negative by age**		**Nonlinear, negative by age**	
Complementary feeding practices					
Food diversity (Number of food groups consumed during last 24 hours aside from breast milk)	Low (0-2)	-	-	-	-
	Medium (3-4)	Constant, positive by age		Constant, positive by age	
	High (5-8)	**~ Linear, positive by age**		**~ Linear, positive by age**	
Meal frequency (Number of meals consumed during last 24 hours aside from breast milk)	Low (0-1)	-	-	-	-
	Medium (2-3)	Constant, zero by age		Constant, zero by age	
	High (4-9)	~ Linear, positive by age		~ Linear, positive by age	
Micronutrient deficiencies					
Child received iron	No	-	-	-	-
	Yes	-0.025	[-0.123, 0.045]	-0.049	[-0.168, 0.035]
Child received vitamin A	No	-	-	-	-
	Yes	**0.076**	**[0.005, 0.140]**	0.046	[0.000, 0.121]
Iodine-in-salt test result	No iodine	-	-	-	-
	Less than 15 parts per million	-0.035	[-0.093, 0.058]	-0.063	[-0.134, 0.014]
	15 parts per million or more	**0.097**	**[0.037, 0.164]**	**0.095**	**[0.036, 0.162]**

Figure 6.1 shows the effects of continuous covariates on the 35% Z-score quantile estimated from the full model and 100 bootstrap iterations. An effect of a continuous covariate is rated as "significant" (and printed in bold in Table 6.1) when the effects are non-zero in all bootstrap samples for at least one interval within the range of the respective covariate. This definition of "significance" is certainly more conservative than considering pointwise bootstrap confidence intervals. However, we prefer the direct visual inspection of the bootstrap estimation results since we think that this best keeps in mind the basis for the decision on significance.

Therefore, with the exception of number of household members, all continuous variables show non-zero effects in all bootstrap samples. Child age shows the largest absolute effect size: the 35% Z-score quantile decreases by almost two units from birth until the age of 24 months.

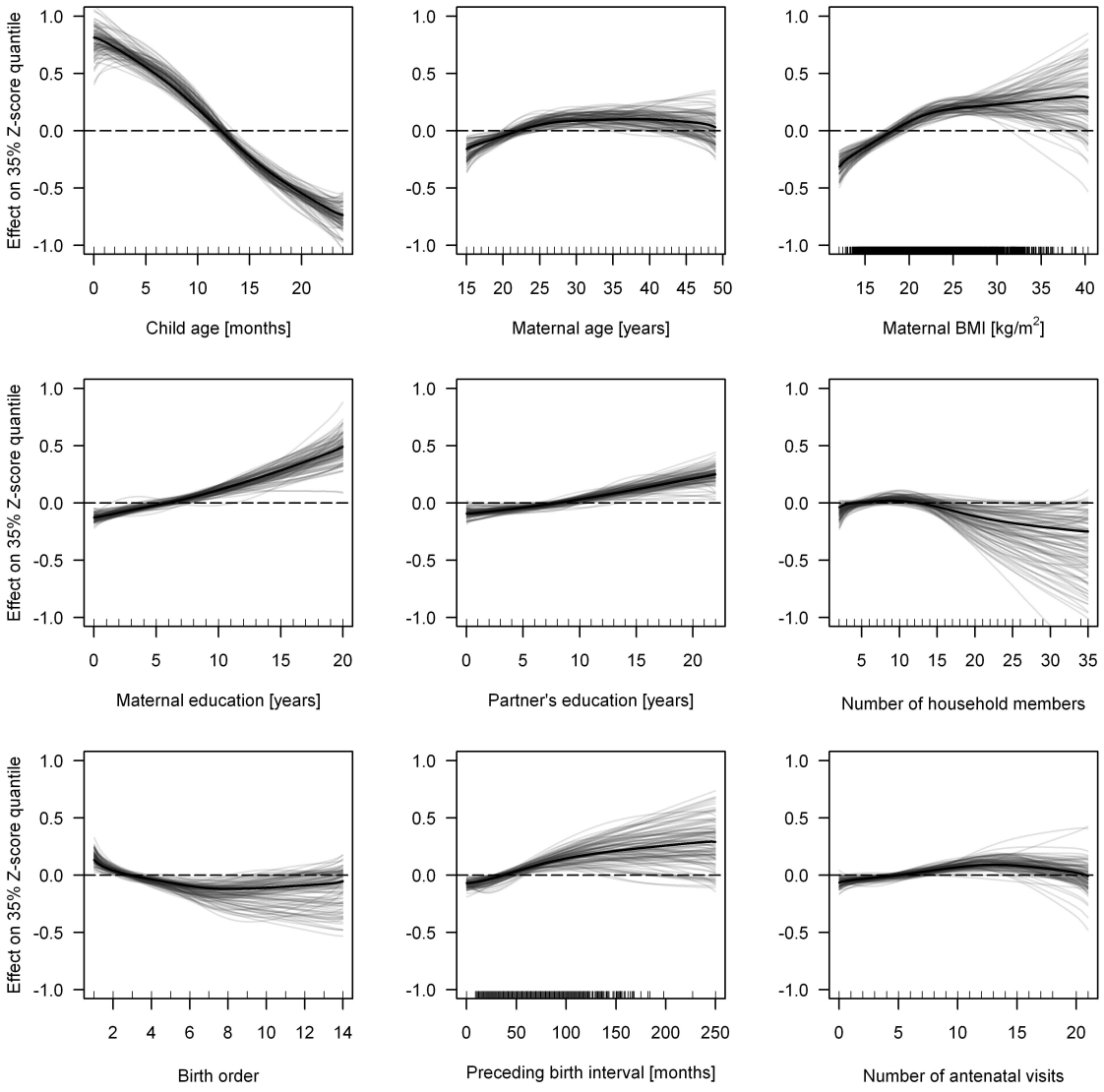


Figure 6.1 Linear or smooth nonlinear effects of continuous covariates from quantile boosting with $\tau = 0.35$ for the full model (black line) and 100 bootstrap iterations (grey lines).

Nonlinear functions are estimated for maternal age and BMI, birth order, preceding birth interval and the number of antenatal visits. The effect of maternal age increases linearly until 30 years, then remains constant and gradually decreases from 45 years. With greater maternal BMI the 35% Z-score quantile increases monotonically, with the slope reducing at 25 kg/m^2. Birth order shows a linearly decreasing effect until the 6th child and then remains approximately constant, while lengthening the interval between births is associated with increased 35% Z-score quantiles up until 100 months. The effect of the number of antenatal visits has a slight inverse U-shape, indicating that low and high numbers of antenatal visits are associated with smaller 35% Z-score quantiles than medium numbers (8-15 visits).

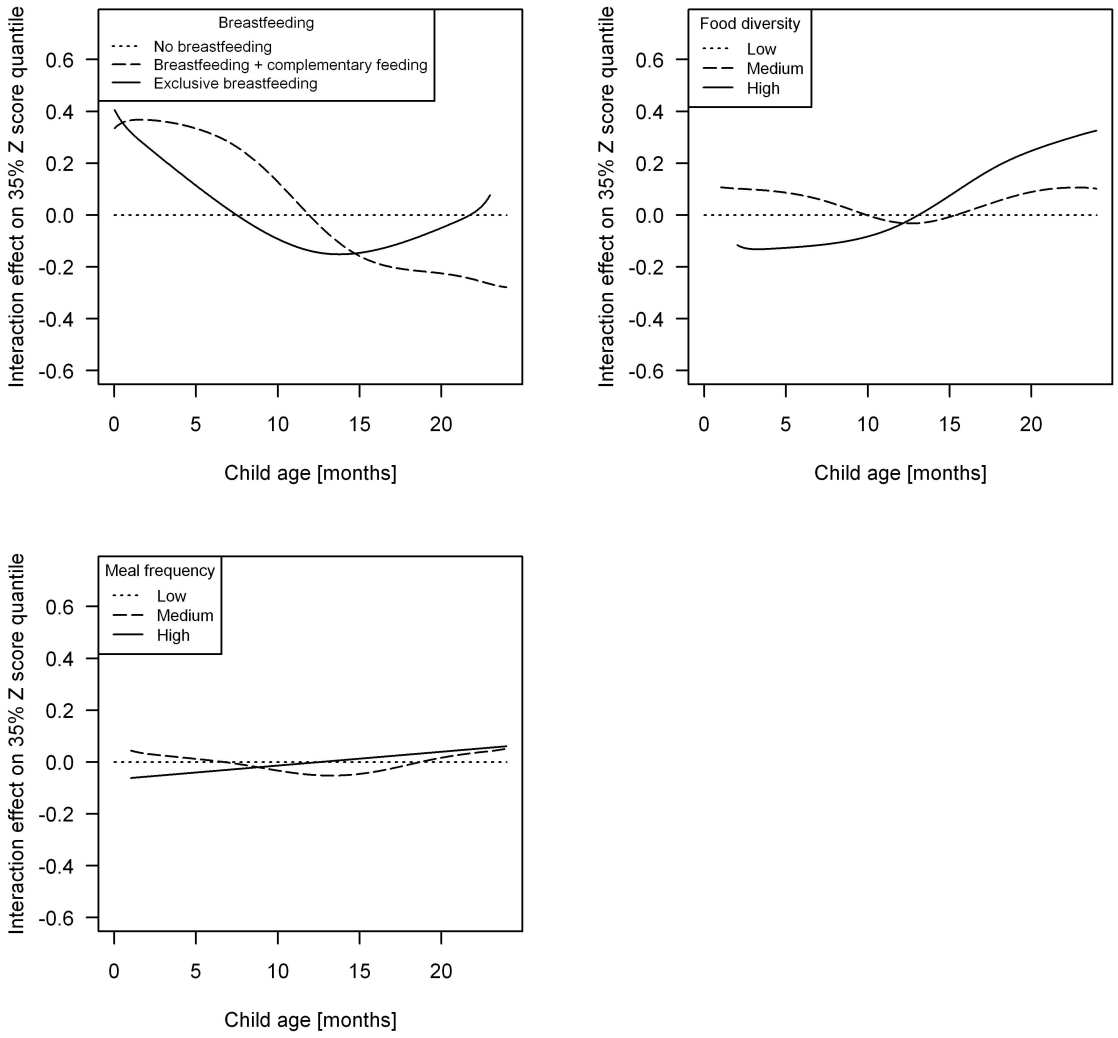


Figure 6.2 Nonlinear age-varying effects of feeding variables estimated by quantile boosting for $\tau = 0.35$ (full model). The dotted horizontal line at zero represents the respective reference category.

Figure 6.2 depicts the age-varying effects of the feeding variables. According to the bootstrap results (not shown) both "exclusive breastfeeding" and "breastfeeding and complementary feeding" differ significantly from the baseline of "no breastfeeding", whereas there is no significant difference between the two breastfeeding categories. The effect of breastfeeding on the 35% Z-score quantile clearly varies with age: breastfeeding exerts a positive effect until 9 months followed by a negative effect beginning at 12 months. Note that the increasing effect of exclusive breastfeeding after 14 months is based on small numbers and shows large variation. Compared to low food diversity, high diversity exerts a negative effect until the age of 12 months, and a positive effect thereafter. This effect can be rated as significant by the bootstrap analysis. Medium food diversity does not differ significantly from the reference category. In relation to meal frequency no significant differences from the baseline of low meal frequency are observed.

Figure 6.3 displays the estimated spatial effect on the 35% Z-score quantile for 29 Indian states of residence. Compared to the empirically observed 35% Z-score quantiles shown in Figure 2.7 (page 27), less pronounced differences in the estimated effects compared to empirical values imply that the further covariates in the model offer a partial explanation for structural differences between Indian states.

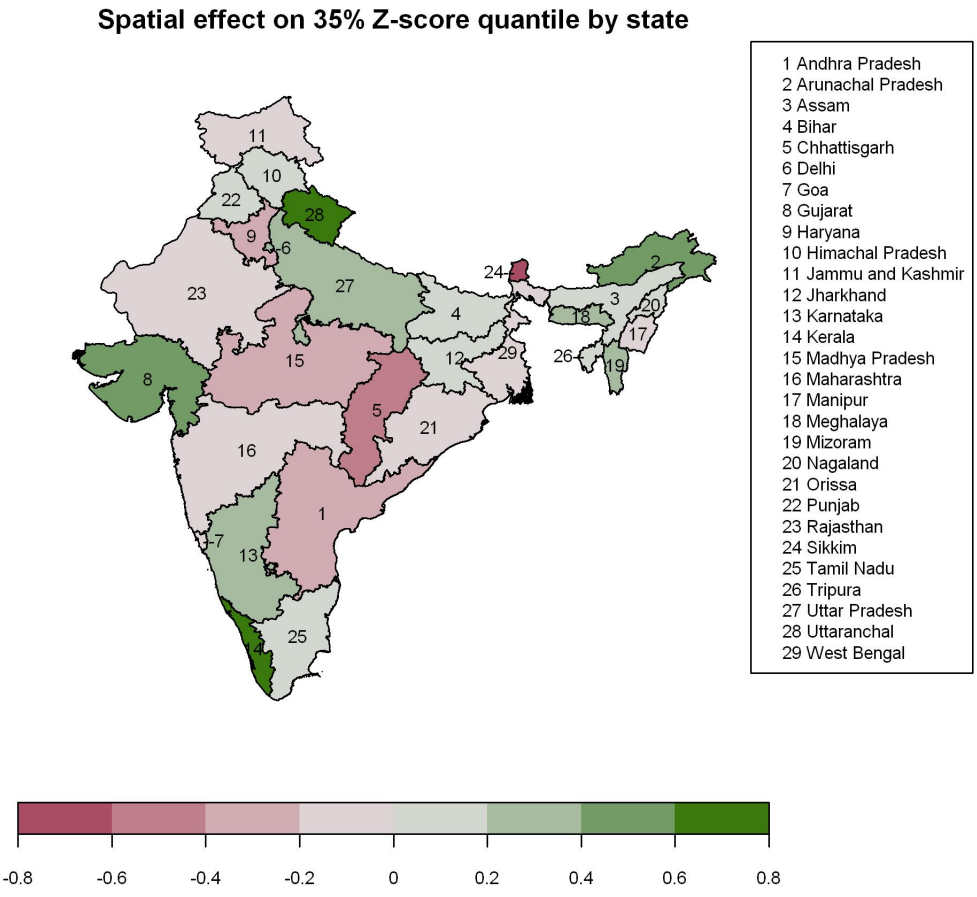


Figure 6.3 Discrete spatial effect on 35% Z-score quantile for 29 Indian states of residence (full model).

Furthermore, Figure 6.4 displays the significance ratings based on the bootstrap confidence intervals corresponding to Figure 6.3. One can see that about 40% of the effects are rated as significant which underlines the importance of including the spatial effect. The effects of three central regions of India (1, 5, 15) on 35% Z-score quantiles are significantly smaller than in other regions. These differences cannot be explained by the other covariates in the model.

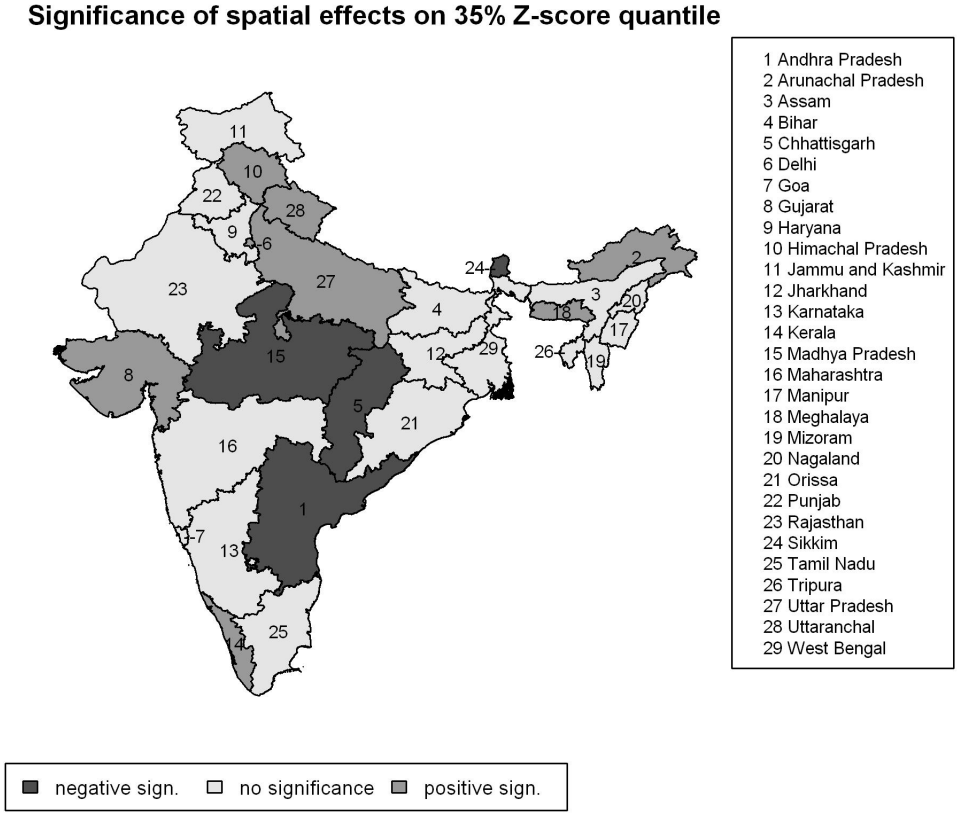


Figure 6.4 Significance of discrete spatial effects on 35% Z-score quantile shown in Figure 6.3 based on bootstrap confidence intervals.

Comparison of quantile regression results for different quantile parameters

Table 6.1 contained the results for the 35% and the 15% Z-score quantiles, whereas Table 6.2 additionally shows the results for the 50% and the 5% Z-score quantile regression analyses. Overall, there are no fundamental differences between the results for the 35% Z-score quantile presented above and those for other quantile parameters.

Regarding the effect significances first, the following covariates exert significant effects on Z-score quantiles across all four quantile parameters: child age, maternal education, partner's education, state of residence, child is twin, iodine, and both levels of breastfeeding by age compared to no breastfeeding. On the contrary, there are two covariates which are each only significant for one quantile parameter: mother is working ($\tau = 0.35$) and cooking fuel ($\tau = 0.15$).

Table 6.2 Estimated effects and 95% bootstrap confidence intervals of quantile boosting models for $\tau = 0.50$ and $\tau = 0.05$. Significant effects are shown in bold.

Variable	Values / Description	Quantile regression for 50% quantile		Quantile regression for 5% quantile	
		$\beta_{0.50}$	95% CI($\beta_{0.50}$)	$\beta_{0.05}$	95% CI($\beta_{0.05}$)
Non-modifiable factors					
Child age [months]		**~ Linear, negative**		**~ Linear, negative**	
Child sex	Male	-	-	-	-
	Female	**0.121**	**[0.059, 0.180]**	0.123	[0.000, 0.256]
Maternal characteristics					
Maternal age [years]		**Nonlinear, inverse U**		Nonlinear, inverse U	
Maternal BMI [kg/m^2]		**Nonlinear, positive**		Linear, positive	
Household characteristics					
Household wealth	Poorest	-	-	-	-
	Poorer	-0.012	[-0.090, 0.086]	-0.026	[-0.129, 0.027]
	Middle	0.050	[-0.039, 0.164]	0.002	[-0.072, 0.090]
	Richer	**0.099**	**[0.012, 0.208]**	-0.005	[-0.089, 0.056}
	Richest	**0.217**	**[0.094, 0.372]**	0.092	[0.000, 0.252]
Religion of household head	Hindu	-	-	-	-
	Muslim	0.031	[-0.042, 0.116]	0.002	[-0.041, 0.079]
	Christian	0.047	[-0.013, 0.144]	0.006	[-0.006, 0.055]
	Sikh	0.010	[-0.022, 0.057]	0.007	[0.000, 0.051]
	(Neo-)Buddhist	-0.010	[-0.067, 0.019]	-0.001	[-0.018, 0.018]
	Other	-0.027	[-0.089, 0.007]	-0.003	[-0.030, 0.006]
Caste/tribe of household head	Scheduled caste	-	-	-	-
	Scheduled tribe	0.053	[-0.015, 0.145]	-0.018	[-0.073, 0.004]
	Other backward class	0.091	[0.000, 0.190]	0.008	[-0.037, 0.059]
	None of them	**0.183**	**[0.090, 0.302]**	0.028	[-0.003, 0.126]
Maternal education [years]		**Nonlinear, positive**		**Nonlinear, positive**	
Partner's education [years]		**~ Linear, positive**		**~ Linear, positive**	
Partner's occupation	Services	-	-	-	-
	Household & domestic	0.034	[-0.025, 0.121]	0.025	[-0.014, 0.125]
	Agriculture	-0.007	[-0.050, 0.050]	0.052	[-0.002, 0.189]
	Clerical	-0.003	[-0.060, 0.079]	0.009	[-0.068, 0.101]
	Prof./ Tech./ Manag.	0.038	[-0.008, 0.130]	0.013	[-0.040, 0.063]
	Did not work	0.006	[-0.058, 0.078]	-0.028	[-0.128, 0.023]
Mother is currently working	No	-	-	-	-
	Yes	-0.055	[-0.123, 0.000]	-0.016	[-0.113, 0.026]
Sex of household head	Male	-	-	-	-
	Female	0.073	[0.000, 0.181]	0.038	[-0.000, 0.173]
Regional characteristics					
State of residence		**Spatial**		**Spatial**	
Urban/rural location	Urban	-	-	-	-
	Rural	0.009	[-0.018, 0.065]	-0.123	[-0.243, 0.118]
Household food competition					
Number of household members		Nonlinear, inverse U		Nonlinear	
Birth order		**Nonlinear, negative**		Linear, negative	
Preceding birth interval [months]		**Nonlinear, positive**		~ Linear, positive	
Child is twin or multiple birth	No	-	-	-	-
	Yes	**-0.642**	**[-0.963, -0.365]**	**-0.407**	**[-0.719, -0.120]**

Variable	Values / Description	Quantile regression for 50% quantile		Quantile regression for 5% quantile	
		$\beta_{0.50}$	95% CI($\beta_{0.50}$)	$\beta_{0.05}$	95% CI($\beta_{0.05}$)
Water, sanitation and hygiene					
Drinking water in household	Unimproved	-	-	-	-
	Improved	-0.026	[-0.101, 0.008]	0.005	[-0.018, 0.074]
	Piped	-0.013	[-0.095, 0.024]	-0.006	[-0.070, 0.006]
Sanitation facility in household	Unimproved	-	-	-	-
	Improved	**0.082**	**[0.016, 0.161]**	0.061	[0.000, 0.239]
Indoor air pollution					
Type of cooking fuel	Straw/ crop /animal dung	-	-	-	-
	Coal/ charcoal/ wood	-0.022	[-0.068, 0.017]	-0.037	[-0.122, 0.000]
	Kerosene	-0.006	[-0.044, 0.012]	-0.011	[-0.065, 0.001]
	Gas/ electricity	0.025	[-0.028, 0.090]	0.016	[-0.006, 0.057]
Curative and preventive healthcare					
Vaccination index	None (0)	-	-	-	-
	Low (1-3)	-0.015	[-0.070, 0.033]	0.008	[-0.043, 0.060]
	Medium (4-6)	-0.002	[-0.054, 0.056]	-0.048	[-0.150, 0.000]
	High (7-9)	0.049	[0.000, 0.123]	0.053	[0.000, 0.129]
Number of antenatal visits during pregnancy		~ Linear, positive		**Nonlinear, inverse U**	
Breastfeeding practices					
Breastfeeding	No breastfeeding	-	-	-	-
	Breastfeeding + complementary feeding	**Nonlinear, negative by age**		**~ Linear, negative by age**	
	Exclusive breastfeeding	**Linear, negative by age**		**Nonlinear, negative by age**	
Complementary feeding practices					
Food diversity (Number of food groups consumed during last 24 hours aside from breast milk)	Low (0-2)	-	-	-	-
	Medium (3-4)	Constant, zero by age		Constant, positive by age	
	High (5-8)	**~ Linear, positive by age**		~ Linear, positive by age	
Meal frequency (Number of meals consumed during last 24 hours aside from breast milk)	Low (0-1)	-	-	-	-
	Medium (2-3)	Constant, zero by age		Constant, zero by age	
	High (4-9)	~ Linear, positive by age		Constant, zero by age	
Micronutrient deficiencies					
Child received iron	No	-	-	-	-
	Yes	-0.013	[-0.091, 0.056]	-0.054	[-0.317, 0.000]
Child received vitamin A	No	-	-	-	-
	Yes	**0.088**	**[0.004, 0.169]**	0.037	[0.000, 0.123]
Iodine-in-salt test result	No iodine	-	-	-	-
	Less than 15 parts per million	-0.015	[-0.075, 0.067]	-0.083	[-0.195, -0.009]
	15 parts per million or more	**0.110**	**[0.038, 0.180]**	**0.143**	**[0.055, 0.251]**

Furthermore, one can observe that several covariates are no longer significant for the 5% quantile compared to the other quantile parameters. This is a typical phenomenon of quantile regression since this method requires a large number of observations to be able to detect covariate effects for extreme quantiles – obviously even more observations than present in our dataset. Therefore, the following covariates show significant effects for the three quantile parameters except for $\tau = 0.05$: child sex, wealth, caste, birth order, sanitation, high food diversity by age.

When regarding the signs and sizes of effects in more detail, one can notice that the differences between quantile parameters are only small, with the signs being similar for all significant effects. Interestingly, despite the small number of twins in the dataset, being a twin shows large significant effects for all quantile parameters. Moreover, the positive effect size of being a female increases from the median to the 15% Z-score quantile.

Likewise, shapes, sizes, and signs of the nonlinear effects are almost similar across all quantile parameters – even for the 5% Z-score quantile, although the effects are often not rated as significant at this quantile. The only difference with regard to linearity vs. nonlinearity is detected for maternal education (linear for $\tau = 0.15$ and $\tau = 0.35$, nonlinear for $\tau = 0.05$ and $\tau = 0.50$). However, a monotonic increase of the effect until 12 years of maternal education is estimated for all quantile parameters with the nonlinearity beginning afterwards. Breastfeeding exerts a significantly nonlinear, age-varying effect with the same shape for all quantile parameters.

Comparison of quantile regression results to those from Kandala *et al.* (2009) and Sobotka *et al.* (2011)

We also compare the results from our quantile regression analyses to the results of two recent studies (Kandala *et al.*, 2009; Sobotka *et al.*, 2011) which we consider as being closely related to our analysis. Both studies also investigated determinants of child stunting by structured additive regression models for the height-for-age Z-score.

In Kandala *et al.* (2009), a Gaussian structured additive regression model was estimated for the mean Z-score based on DHS data from three different African countries. Special emphasis was put on spatial modelling of regional differences between the countries and estimation was realized by a full Bayesian approach. Compared to our analysis, much less covariates were included, with most of them being coded in a completely different way. Nevertheless, we shortly compare the results of this study to the results from our median regression analysis.

Similarities could be found regarding the estimated effects of child sex and age until 24 months (even though age varied between 0-59 months contrary to 0-24 months in our analysis). Also, the shapes of the nonlinear effects for maternal BMI were estimated to be exactly similar. The effects of wealth and birth interval were also significant and positive in both analyses – even though the underlying variables had different scale levels.

With regard to differences, the effect of study location was significant in Kandala *et al.* (2009) with larger Z-score means estimated for urban areas. Also, the household size had a positive and significant effect. These results are contrary to our analysis, where both variables were not rated as significant by any of the models. The two remaining covariates cannot be compared to our results since the variable coding was too different for maternal education, and a variable for single mothers was not included in our analysis.

Sobotka *et al.* (2011) employed structured additive expectile regression with 40 covariates for four asymmetry parameters $(0.05, 0.20, 0.80, 0.95)$ based on the Indian DHS dataset for the year 2001. The underlying methodological aim was to investigate confidence intervals for the parameters of expectile regression. Since the Z-score response ranged between -600 and 600, the estimated effects had to be divided by 100 to be comparable with our results.

Similarities of both analyses were detected with regard to significance and size of the effects of twin and child age. Likewise, both analyses did not rate the effects of maternal work and residence as significant but detected a significant negative effect for birth order, even though this variable was categorical in Sobotka *et al.* (2011).

Regarding main differences between the analyses, the effects of child sex, maternal and partner's education were estimated to have similar signs but were not rated as significant by the expectile regression analysis. Furthermore, the shapes for maternal BMI and age were not estimated in a nonlinear way. Duration of breastfeeding did not show an effect in the expectile regression analysis. We included breastfeeding as a categorical, three-level variable and observed a significant association between breastfeeding by age and Z-score quantiles.

Contrary to our analysis, religion had a significant effect. We attribute this difference to the inclusion of the variable caste in our analysis since we observed in own sensitivity analyses that religion became significant when caste was excluded from our analysis. In India, caste probably captures position in the social hierarchy better than religion.

Further results shown in Sobotka *et al.* (2011) are not really comparable to our analysis since we aggregated the binary variables for household characteristics (electricity, radio, TV, etc.) in one overall variable standing for household wealth. The number of dead children was not included in our analysis.

Comparison of logistic and quantile regression results

When additionally comparing the results of logistic and quantile regression, no fundamental differences can be detected similar as the comparisons between different quantile parameters. As can be seen in Table 6.3, the direction of the effect of variables in quantile regression models is reversed in binary regression models. Also, absolute effect sizes cannot be compared since the interpretation relates to the log-odds ratio for being stunted or severely stunted (contrary to the respective quantiles of the Z-score). As an example for the interpretation, the log-odds ratio for being stunted for girls is estimated to be -0.080 smaller compared to boys, given all other covariates are similar.

Most of the variables being significant for almost all quantile parameters are also significant in binary regression analyses. However, compared to the variables which are significant across all quantile parameters, the age-varying effect of exclusive breastfeeding is no longer significant for stunting and severe stunting. Likewise, birth order, number of antenatal visits and vitamin A supplementation show no significant effect in logistic regression models. Cooking fuel is significant for stunting and severe stunting which was only the case for $\tau = 0.15$. Therefore, cooking with gas/electricity is protective against stunting, while cooking with kerosene emerges as a risk factor for severe stunting. Note also that child sex, twin, wealth, caste, and iodine show larger absolute sizes for severe stunting than for stunting.

Table 6.3 Estimated effects and 95% bootstrap confidence intervals for structured additive logistic regression for stunting and severe stunting estimated by boosting. Significant effects are shown in bold.

Variable	Values / Description	Logistic regression for stunting		Logistic regression for severe stunting	
		$\beta_{stunted}$	95% CI($\beta_{stunted}$)	β_{sevSt}	95% CI(β_{sevSt})
Non-modifiable factors					
Child age [months]		**~ Linear, positive**		**~ Linear, positive**	
Child sex	Male	-	-	-	-
	Female	**-0.080**	**[-0.123, -0.037]**	**-0.120**	**[-0.171, -0.068]**
Maternal characteristics					
Maternal age [years]		**Nonlinear, U-shape**		**Nonlinear, U-shape**	
Maternal BMI [kg/m^2]		**Nonlinear, U-shape**		**Nonlinear, negative**	
Household characteristics					
Household wealth	Poorest	-	-	-	-
	Poorer	0.007	[-0.045, 0.063]	-0.044	[-0.104, 0.026]
	Middle	-0.011	[-0.058, 0.031]	-0.056	[-0.129, -0.002]
	Richer	-0.041	[-0.115, 0.019]	**-0.119**	**[-0.235, -0.030]**
	Richest	**-0.130**	**[-0.244, -0.027]**	**-0.221**	**[-0.353, -0.085]**
Religion of household head	Hindu	-	-	-	-
	Muslim	-0.045	[-0.114, 0.010]	-0.004	[-0.058, 0.059]
	Christian	-0.037	[-0.119, 0.038]	-0.017	[-0.087, 0.033]
	Sikh	-0.046	[-0.124, 0.004]	-0.013	[-0.060, 0.014]
	(Neo-)Buddhist	-0.023	[-0.126, 0.033]	-0.016	[-0.093, 0.020]
	Other	0.041	[-0.002, 0.118]	0.026	[-0.014, 0.103]
Caste/tribe of household head	Scheduled caste	-	-	-	-
	Scheduled tribe	-0.030	[-0.100, 0.021]	-0.038	[-0.120, 0.026]
	Other backward class	**-0.066**	**[-0.126, -0.009]**	**-0.078**	**[-0.132, -0.025]**
	None of them	**-0.112**	**[-0.188, -0.047]**	**-0.134**	**[-0.224, -0.064]**
Maternal education [years]		**~ Linear, negative**		**~ Linear, negative**	
Partner's education [years]		**~ Linear, negative**		**~ Linear, negative**	
Partner's occupation	Services	-	-	-	-
	Household & domestic	-0.030	[-0.090, 0.010]	-0.056	[-0.152, 0.008]
	Agriculture	-0.006	[-0.042, 0.032]	-0.055	[-0.111, -0.012]
	Clerical	-0.011	[-0.047, 0.038]	-0.030	[-0.093, 0.026]
	Prof./ Tech./ Manag.	-0.014	[-0.064, 0.032]	0.016	[-0.030, 0.090]
	Did not work	0.001	[-0.045, 0.049]	0.015	[-0.026, 0.085]
Mother is currently working	No	-	-	-	-
	Yes	0.043	[0.000, 0.086]	0.040	[0.000, 0.093]
Sex of household head	Male	-	-	-	-
	Female	-0.023	[-0.081, 0.003]	-0.006	[-0.067, 0.036]
Regional characteristics					
State of residence		**Spatial**		**Spatial**	
Urban/rural location	Urban	-	-	-	-
	Rural	-0.045	[-0.093, 0.000]	-0.021	[-0.071, 0.000]
Household food competition					
Number of household members		Non-linear, U shape		Non-linear, U shape	
Birth order		Non-linear, positive		~ Linear, positive	
Preceding birth interval [months]		Non-linear, negative		Non-linear, negative	
Child is twin or multiple birth	No	-	-	-	-
	Yes	**0.420**	**[0.251, 0.579]**	**0.566**	**[0.385, 0.750]**

Variable	Values / Description	Logistic regression for stunting		Logistic regression for severe stunting	
		$\beta_{stunted}$	95% CI($\beta_{stunted}$)	β_{sevSt}	95% CI(β_{sevSt})
Water, sanitation and hygiene					
Drinking water in household	Unimproved	-	-	-	-
	Improved	0.019	[-0.005, 0.063]	-0.005	[-0.045, 0.029]
	Piped	0.010	[-0.025, 0.068]	-0.006	[-0.053, 0.019]
Sanitation facility in household	Unimproved	-	-	-	-
	Improved	**-0.057**	**[-0.111, -0.011]**	**-0.049**	**[-0.112, -0.001]**
Indoor air pollution					
Type of cooking fuel	Straw/ crop /animal dung	-	-	-	-
	Coal/ charcoal/ wood	0.014	[-0.018, 0.044]	0.005	[-0.036, 0.055]
	Kerosene	0.018	[-0.028, 0.058]	**0.124**	**[0.019, 0.238]**
	Gas/ electricity	**-0.088**	**[-0.168, -0.015]**	-0.065	[-0.145, 0.001]
Curative and preventive healthcare					
Vaccination index	None (0)	-	-	-	-
	Low (1-3)	-0.005	[-0.074, 0.038]	-0.004	[-0.076, 0.037]
	Medium (4-6)	-0.004	[-0.086, 0.044]	0.006	[-0.099, 0.052]
	High (7-9)	**-0.072**	**[-0.151, -0.013]**	**-0.059**	**[-0.152, -0.005]**
Number of antenatal visits during pregnancy		~ Linear, negative		Non-linear, U shape	
Breastfeeding practices					
Breastfeeding	No breastfeeding	-	-	-	-
	Breastfeeding + complementary feeding	**Nonlinear, positive by age**		**Nonlinear, positive by age**	
	Exclusive breastfeeding	~ Linear, positive by age		~ Linear, positive by age	
Complementary feeding practices					
Food diversity (Number of food groups consumed during last 24 hours aside from breast milk)	Low (0-2)	-	-	-	-
	Medium (3-4)	Constant, zero by age		Constant, negative by age	
	High (5-8)	**~ Linear, negative by age**		**~ Linear, negative by age**	
Meal frequency (Number of meals consumed during last 24 hours aside from breast milk)	Low (0-1)	-	-	-	-
	Medium (2-3)	Constant, zero by age		Constant, zero by age	
	High (4-9)	Constant, zero by age		Constant, zero by age	
Micronutrient deficiencies					
Child received iron	No	-	-	-	-
	Yes	0.022	[-0.016, 0.089]	0.030	[-0.007, 0.138]
Child received vitamin A	No	-	-	-	-
	Yes	-0.036	[-0.077, 0.000]	-0.020	[-0.070, 0.000]
Iodine-in-salt test result	No iodine	-	-	-	-
	Less than 15 parts per million	0.011	[-0.043, 0.044]	0.025	[-0.013, 0.058]
	15 parts per million or more	**-0.056**	**[-0.107, -0.020]**	**-0.066**	**[-0.118, -0.022]**

Figure 6.5 exemplarily displays the effects of continuous covariates on stunting estimated from the full model and 100 bootstrap iterations. Compared to Figure 6.1, one can see that the effect directions are reversed, but the nonlinear shapes are estimated to be approximately similar. For example, maternal BMI displays a U-shape for stunting, meaning that children born to mothers with low and high BMI values are at higher risk.

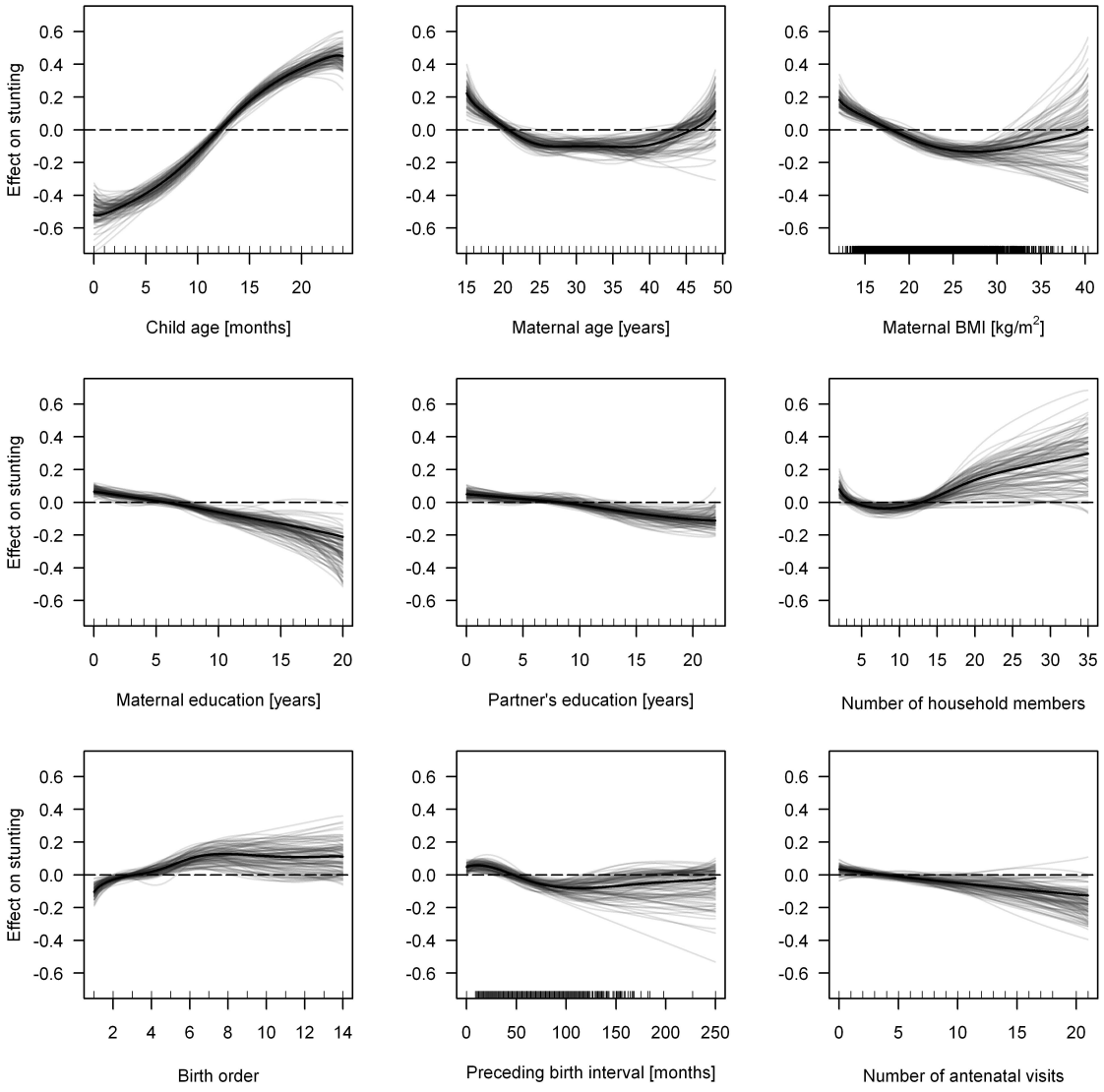


Figure 6.5 Linear or smooth nonlinear effects of continuous covariates from structured additive logistic regression for stunting estimated by boosting for the full model (black line) and 100 bootstrap iterations (grey lines).

Results from the stability selection procedure

Finally, we consider the results from the formal variable selection procedure by stability selection. In our analysis, we chose a family-wise error rate of 5% and an average number of 15 terms to be expected in a model. Table 6.4 contains the results. Selected variables are those which were included in more than 90% of 100 bootstrap models in the first approx. 200 boosting iterations (determined as the boosting iteration in which the 15th model term was included in each bootstrap sample). The underlying rationale for this variable selection is that boosting first selects those variables which achieve a maximum reduction of the variance of the negative gradient residuals (see also Chapter 5.2).

Even though we fixed the number of expected model terms at 15, Table 6.4 shows that only about 10 variables (out of 51 base learners) were selected in each model. Except for indoor air pollution and complementary feeding practices, at least one variable of each of the groups of determinants were selected by stability selection. Looking at Table 6.4 into more detail, both similarities and differences can be detected when comparing selected and significant variables.

Regarding similarities, the following variables were rated as significant by almost all models and were also mostly selected by stability selection: child age, maternal BMI, maternal education, partner's education, sanitation, breastfeeding by age, and iodine. Therefore, strong evidence is provided for an impact of these variables on child stunting. Type of cooking fuel was three times selected by stability selection and also had three times significant effects (even though not in the same models).

With regard to differences, there are several significant variables which were never or only once chosen by stability selection: child sex, wealth, caste, birth order, child is twin, vaccination index, and food diversity by age. This is in particular surprising for the twin variable which showed a great significant effect on all stunting quantiles and binary variables. Probably the variance reduction achieved by this variable during the first boosting iterations is not large enough due to the small number of twins in the dataset.

Concerning the spatial variables, it is interesting that urban/rural location was selected three times by stability selection even though this variable was not rated as significant in any of the regression models. Likewise, a surprising result is that the number of antenatal visits was included in every model by stability selection but only had significant effects in three models.

The following variables neither had significant effects in any of the models nor were selected by stability selection: religion, partner's occupation, sex of household head, number of household members, drinking water in household, meal frequency by age, and iron.

Table 6.4 Results from formal variable selection by stability selection for all regression models. Shown are only selected variables based on a family-wise error rate of 5% and 15 terms to be expected in a model.

Variable	Quantile regression				Logistic regression	
	τ = 0.50	τ = 0.35	τ = 0.15	τ = 0.05	Stunting	Sev. stunting
Non-modifiable factors						
Child age	Nonlinear	Nonlinear	Linear	Linear	Nonlinear	Linear
Maternal characteristics						
Maternal BMI	Nonlinear	Linear			Linear	Linear
Household characteristics						
Maternal education	Linear	Linear	Nonlinear	Nonlinear	Linear	Linear
Partner's education	Linear	Linear	Linear	Linear	Linear	Linear
Regional characteristics						
State of residence	X					
Urban/rural location		X	X	X		
Household food competition						
Child is twin				X		
Water, sanitation and hygiene						
Sanitation facility in household	X	X	X		X	
Indoor air pollution						
Type of cooking fuel		X	X		X	
Curative and preventive healthcare						
Number of antenatal visits	Linear	Nonlinear	Nonlinear	Nonlinear	Linear	Nonlinear
Breastfeeding practices						
Breastfeeding by age	X	X	X	X	X	X
Micronutrient deficiencies						
Iodine-in-salt test result	X	X		X	X	X

6.3 Discussion

Key findings

We employed an evidence-based, systematic approach to identify all likely determinants of child stunting. This extension in breadth and depth of the UNICEF framework (UNICEF, 1998) represents an important intermediate outcome and a basis for further research.

In our analysis, we attempted to quantify the effect of these determinants or their proxies with structured additive quantile and logistic regression models estimated by boosting. The results of this innovative quantile regression approach and the standard logistic regression analyses were largely comparable. This is rather surprising since quantile regression can exploit the full information of the response variable – contrary to logistic regression just relying on a binarized version of the response. We attribute this result to the symmetric distribution of the Z-score response, suggesting that binary regression can compete with quantile regression for analyzing determinants of stunting.

However, our research has demonstrated that several continuous variables (maternal age, maternal BMI, birth order and number of antenatal visits) exert their effect in a nonlinear way. Our prior assumption that the effects of breastfeeding and complementary feeding are age-dependent was realized by including nonlinear age-varying effects in the model.

Regarding further covariates, our research confirms the importance of child age and sex as non-modifiable determinants and highlights greater parental education and greater maternal BMI as major protective factors. Our research also draws attention to twins as a potentially overlooked risk group. The very large significantly negative effect is remarkable, as only 1% of children in the NFHS dataset are twins or multiple births. Strong evidence is also provided for the importance of sanitation and iodine, even though the absolute effect sizes of these variables are rather small.

For several variables, none of the models detected statistically significant effects which contrasts with previous reports. This may be due to the poor quality of the proxy measures we employed or differences in the population distribution of variables. Most importantly, it may reflect the fact that in a more comprehensive model, the effects of some variables are captured by other related variables.

Altogether, for each of the groups of determinants we conceptualized in Figure 2.1, we found at least one variable with a statistically significant effect in all models – except for indoor air pollution, which only shows a significant effect in two of six regression models, but was also selected by stability selection. This emphasizes the broad range of causes of child stunting.

Strengths and limitations

Data quality: Although the NFHS study includes suitable variables for most determinants of stunting, we could not model the impact of all determinants we conceptualized in Figure 2.1. We were unable to populate the groups of determinants chronic diseases and recurrent infections and could only partially assess micronutrient deficiencies, healthcare, maternal or regional characteristics. Similarly, some of the proxies we used in our analysis may not provide an accurate estimate of the underlying concept of interest (e.g., type of cooking fuel as a proxy for indoor air pollution). Consequently, effect sizes for individual variables should be interpreted with caution.

Even though the NFHS is considered a high-quality dataset, large numbers of missing values in selected variables, in particular in the outcome of interest, may have introduced selection bias. Nevertheless, the large-scale, standardized and nationally representative nature of the NFHS, a response rate of eligible women of 94.5% and its coverage of a broad range of health risks makes it ideally-suited for a comprehensive analysis of stunting determinants.

Evidence-based approach: Based on earlier work in this field, a priori reasoning and extensive searches of the literature, we derived a schematic diagram of the multiple determinants of stunting. We believe that this approach to identifying all potential determinants of stunting and to populating as many of these as possible using an existing dataset is novel and takes up recent calls to incorporate systems thinking in epidemiology.

Statistical methods: The innovative approach of structured additive regression models can be seen as one of the main strengths of our analysis. Since we considered a large number of covariates with a variety of different effects in the flexible predictor, boosting was particularly well-suited for the estimation. Boosting combines parameter estimation and variable selection in one estimation step. Therefore, subsequent steps of variable selection were not necessary – contrary to classical likelihood-based regression where, for example, AIC-based model comparisons would have had to be conducted. In addition, we applied stability selection to obtain results from a formal variable selection procedure which provided further evidence for the presence of several effects.

Implications for future research

We believe that structured additive quantile and logistic regression models are both adequate approaches for future investigations of the multi-factorial nature of child stunting, as long as the quantile parameter is not chosen too small. Since we found that several continuous variables exert their effect in a nonlinear way, we would recommend to explicitly consider such nonlinear relationships in future analyses. Further investigations of nonlinear age-varying effects could lead to additional insights.

In our view, it would also be helpful to conduct further (empirical) investigations of the stability selection procedure in connection with boosting in order to explain the differences between variable selection and significance results.

To sum up, our research has demonstrated the importance of a comprehensive and systematic approach to the determinants of child undernutrition. Figure 2.1 may serve as a starting point for furthering the understanding of this system in future analyses.

Chapter 7: Quantile boosting for child overweight and obesity in Germany

This chapter presents the results of a quantile boosting analysis of child overweight and obesity in western countries by means of the LISA study, a recent large-scale German birth cohort study. Background and dataset of this analysis were thoroughly described in Section 2.2 of this thesis. The contents of the present chapter are mainly based on the manuscript Fenske, Fahrmeir, Hothorn, Rzehak, and Höhle (2012b).

7.1 Setup of the analysis

Recall that the main objective of our obesity analysis was to flexibly model nonlinear population age curves of upper BMI quantiles, while adjusting for individual-specific age effects and early childhood risk factors. At the same time, individual-specific BMI life-course patterns should be reflected as best as possible. An additional aim was to investigate if potential effects of categorical risk factors are constant or age-varying.

Based on the LISA data described in Section 2.2, we estimated STAQ models for the 90% and 97% BMI quantiles and – for reasons of model comparison – for the median and the 10% quantile. To answer the above questions, we considered the following predictor for $\tau \in \{0.10, 0.50, 0.90, 0.97\}$:

$$\text{cBMI}_{ij} = \eta_{ij}^{(\tau)} + b_{\tau i0} + b_{\tau i1} \cdot \text{cAge}_{ij} + \varepsilon_{\tau ij} \ , \tag{7.1}$$

where the population part is given by

$$\begin{aligned}
\eta_{ij}^{(\tau)} = \beta_0 &+ f_{\tau\text{cAge}}(\text{cAge}_{ij}) + f_{\tau\text{mBMI}}(\text{mBMI}_i) + f_{\tau\text{mDiffBMI}}\,(\text{mDiffBMI}_i) \\
&+ \beta_{\tau\text{cSex}}\,\text{cSex}_i + \beta_{\tau\text{cLocation}}\,\text{cLocation}_i + \beta_{\tau\text{cBreast}}\,\text{cBreast}_i + \beta_{\tau\text{mSmoke}}\,\text{mSmoke}_i \\
&+ \beta_{\tau\text{mEdu2}}\,\text{mEdu2}_i + \beta_{\tau\text{mEdu3}}\,\text{mEdu3}_i \\
&+ \text{cSex}_i \cdot g_{\tau\text{Male}}(\text{cAge}_{ij}) + \text{cLocation}_i \cdot g_{\tau\text{Urban}}(\text{cAge}_{ij}) + \text{cBreast}_i \cdot g_{\tau\text{Breast}}(\text{cAge}_{ij}) \\
&+ \text{mSmoke}_i \cdot g_{\tau\text{Smoke}}(\text{cAge}_{ij}) + \text{mEdu2}_i \cdot g_{\tau\text{mEdu2}}(\text{cAge}_{ij}) + \text{mEdu3}_i \cdot g_{\tau\text{mEdu3}}(\text{cAge}_{ij}) \ .
\end{aligned}$$

Our model thus contains main effects for the entire set of covariates given in Tables 2.2 and 2.3 on page 32, and age-varying effects of categorical covariates. To account for the longitudinal data structure, individual-specific intercepts $b_{\tau i0}$ and slopes $b_{\tau i1}$ for cAge were included in the model, describing individual-specific deviations from the nonlinear population effect for cAge. Note that the above predictor is composed of the same elements as example 2 for a structured additive predictor on page 37.

For reasons of model comparison, we estimated a Gaussian additive mixed model (AMM) with the same predictor as in (7.1). In longitudinal data settings, this model class is currently the only serious competitor for structured additive quantile regression fitted by boosting since no other estimation approach can handle the full variety of population and individual-specific effects addressed by the predictor in (7.1). As was pointed out in Section 3.5.2, AMMs not only imply

conditional mean modelling, but can also be used for quantile regression since the conditional response distribution is completely determined by the *iid* Gaussian assumption for the error terms.

For estimation with the quantile boosting algorithm, we defined one common base learner for all categorical covariates and separate base learners for all smooth effects. All continuous covariates, including age and age-varying effects, were modelled by separate penalized least squares base learners with df$(\lambda) = 5$. The three-level covariate mEdu was split into two dummy-coded variables relating to the reference category of low maternal education (mEdu=1), but was fitted in the same base learner with main effects of all other categorical covariates and df$(\lambda) = 5$. The individual-specific intercept and slope were separated into two base learners and also fitted by penalized least squares with df$(\lambda) = 5$ to equalize the selection probabilities of different base learners. We did not make use of the decomposition of the smooth effects of continuous covariates into linear part and nonlinear deviation since this was not the primary interest of our analysis.

The number of optimal boosting iterations m_{stop} was chosen by block-wise fivefold cross-validation which resulted in values of roughly 5 000 iterations. We set the step length $\nu = 0.4$ since this results in fewer boosting iterations and therefore lower computational effort than for smaller values of ν, as was illustrated in Section 4.2.

In accordance with the descriptions above, we used the following model calls to estimate STAQ models on the full LISA dataset in R:

```
staqFormula <- cBMI ~ bols(cSex, cLocation, cBreast, mSmoke, mEdu, df=5) +
                      bbs(cAgeC, df=5) + bbs(mDiffBMIC, df=5) +
                      bbs(mBMIC, df=5) + bbs(ageSexInt, df=5) +
                      bbs(ageLocationInt, df=5) + bbs(ageBfInt, df=5) +
                      bbs(ageSmokeInt, df=5) + bbs(ageMedu2Int, df=5) +
                      bbs(ageMedu3Int, df=5) + brandom(cID, df=5) +
                      brandom(cID, by=cAgeC, df=5)

staq90 <- gamboost(staqFormula, data=lisaLong,
                   family=QuantReg(tau = 0.90),
                   control=boost_control(mstop=5000, nu=0.4))
```

Similar to the analysis of undernutrition, we used the function `gamboost` from package `mboost` (Hothorn *et al.*, 2012) with option `family=QuantReg()`. Base learners for estimating individual-specific effects were specified with `brandom()`, while smooth nonlinear effects and age-varying effects of categorical covariates were estimated by using the base learner function `bbs()`. The variables `cAgeC`, `mDiffBMIC` and `mBMIC` are mean-centered versions of the `cAge`, `mDiffBMI` and `mBMI`. All variables ending on `Int` are mean-centered interaction variables between age and different levels of the categorical covariates which were defined to model age-varying effects. For fitting AMMs, we used the function `amer` from package `amer` (Scheipl, 2011).

Since boosting does not directly provide standard errors for the estimated effects, we additionally conducted a block-wise bootstrap analysis. We obtained one single bootstrap sample by randomly choosing 2226 children with replacement at the first stage. To conserve the longitudinal data structure, all observations corresponding to the chosen children were included in the bootstrap

sample at the second stage. In this way, we generated a total of 50 different bootstrap samples and used each sample to fit STAQ models and AMMs as described above.

To formally compare the results from AMMs and STAQ models, we additionally constructed 50 out-of-bag samples with children that were not contained in the respective bootstrap samples. These out-of-bag samples were used to calculate the empirical risks based on the check function for the four different quantile parameters and two different model classes. To obtain an estimated predictor $\hat{q}_{\tau ij}$ for a child in an out-of-bag sample, we set its individual-specific effects to zero in order to obtain the empirical risk for model comparison.

7.2 Results

The resulting smooth nonlinear population effects of age on BMI quantiles are shown in Figure 7.1. Overall, the shape of the age effect remains stable over the bootstrap iterations for all models and confirms the first impression from the descriptive analysis in Figure 2.9 on page 31. In sparse data regions, i.e., between the ages of 6 and 10 years, the variation of the effects is larger than in regions with more observations. The effects for the AMM and STAQ median look roughly similar. For upper quantiles, the age effect strongly increases beginning after the age of 6 years.

Furthermore, Figure 7.2 shows estimated age-specific quantile curves from the two model classes. To make the effects comparable, we concentrate on the population quantile functions conditional on individual-specific effects as given in (3.12) on page 55. Thus, AMM curves for upper quantiles were obtained by a parallel shift of the mean curve, whereas for STAQ models, all quantile curves were modelled separately. The resulting curves are estimated to be roughly similar until the age of 6 years. At the age of 10 years, the 90% BMI curve estimated by STAQ regression is above the 97% curve estimated by AMMs, whereas the 10% STAQ curve is below the AMM median.

We additionally compared the model fits by block-wise bootstrap and calculated the empirical quantile-specific risks based on the check function in the 50 out-of-bag bootstrap samples, as described in Section 7.1. Figure 7.3 shows that there are no fundamental differences between the empirical risks for the median, but that STAQ models clearly outperform AMMs for other quantiles. This result is in accordance with Figure 7.2, which together demonstrates that STAQ models are more appropriate for handling the age-specific skewness of the BMI distribution than AMMs.

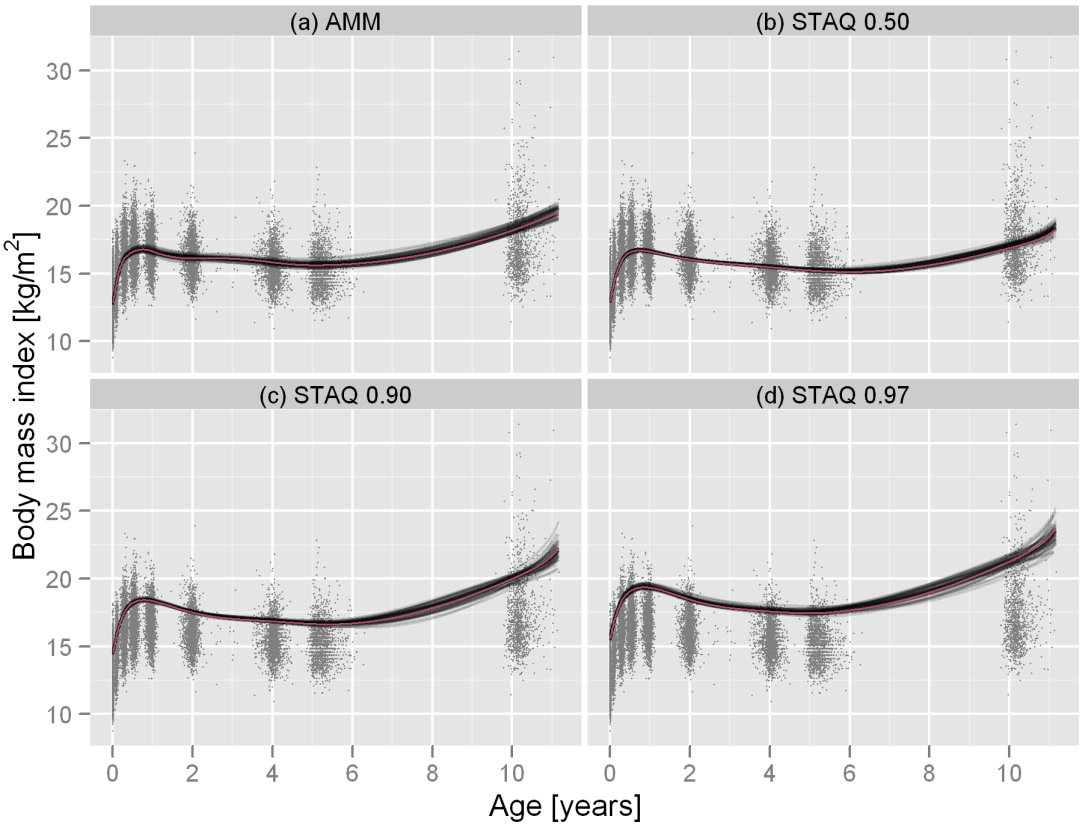


Figure 7.1 Estimated nonlinear BMI quantile curves by age resulting from (a) AMM and from STAQ models for (b) $\tau = 0.50$, (c) $\tau = 0.90$, and (d) $\tau = 0.97$. Shown are BMI observations (grey points) with estimated curves (grey lines) for 50 different bootstrap samples; the superimposed red line represents the estimated nonlinear quantile curve of the respective model on the full dataset. Quantile curves are adjusted for a fixed combination of time-constant covariates (mean for continuous covariates, reference category for categorical covariates).

To assess the uncertainty for individual-specific mean predictions, we constructed individual-specific 90% prediction intervals based on estimated 5% and 95% BMI quantile curves. This procedure was based on the suggestion in Meinshausen (2006) and was made in analogy to Mayr *et al.* (2012c). Figure 7.4 shows individual-specific BMI quantile curves depending on age for 12 randomly chosen children estimated by AMMs. The dashed quantile curves, corresponding to the interval limits, are parallel shifts of the mean curves. The symmetric shape and the distance between the curves remains the same for all children. The offset differences between children can be attributed to the child-specific intercept and covariate combination; the shape differences can be attributed to the child-specific slopes. One can see that the mean curve reproduces the true BMI pattern (in grey) in most cases.

For STAQ models, individual-specific BMI quantile curves are shown in Figure 7.5. Since the three quantile curves are estimated independently from each other, the quantile curves are no longer parallel shifts of the mean/median curves, as was the case for AMMs. The interval widths vary notably between children since the individual-specific intercepts estimated by STAQ models differ for the different quantiles. The upper quantile curves, in particular, seem to account better for the increasing skewness of the BMI distribution with increasing age than the ones of AMMs.

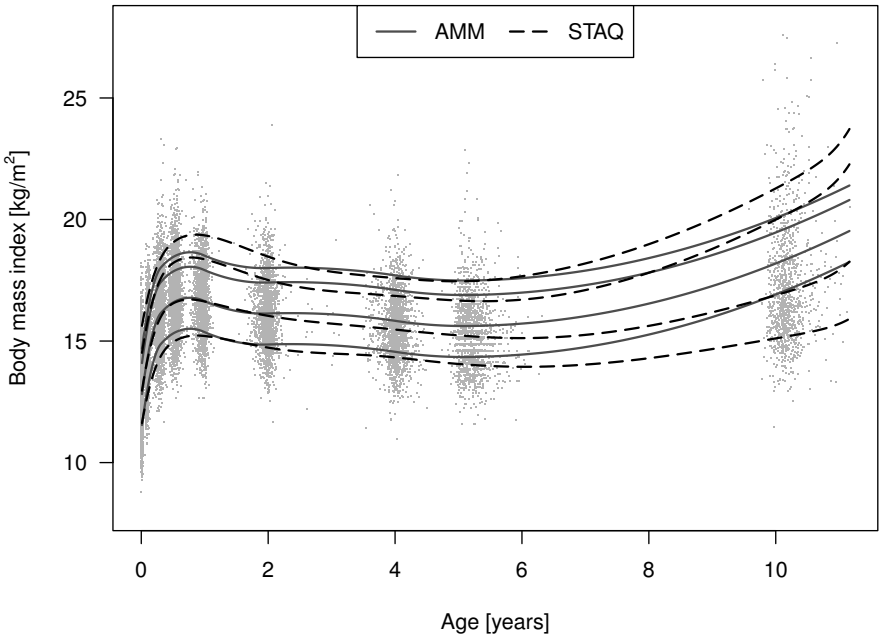


Figure 7.2 Comparison of AMM (solid lines) and STAQ (dashed lines) results for the four quantiles $\tau = 0.10$, $\tau = 0.50$, $\tau = 0.90$, and $\tau = 0.97$. Shown are BMI observations (grey points) with estimated nonlinear quantile curves depending on age and a fixed combination of time-constant covariates (mean for continuous covariates, reference category for categorical covariates).

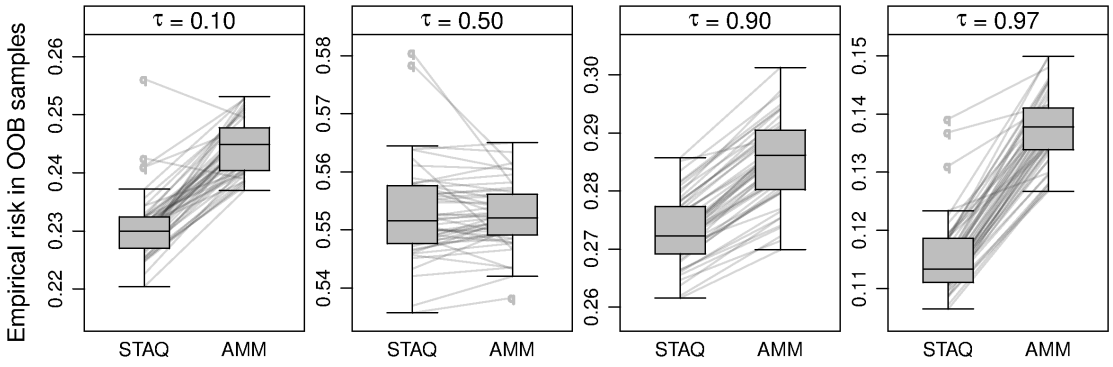


Figure 7.3 Risk comparison of STAQ and AMM for the four quantiles $\tau = 0.10$, $\tau = 0.50$, $\tau = 0.90$, and $\tau = 0.97$. Boxplots show empirical distributions of the empirical risks calculated on 50 out-of-bag (OOB) bootstrap samples. Results for one out-of-bag sample are connected by grey lines.

However, there are almost no individual-specific slope differences. Note that this is contrary to AMMs which capture the BMI skewness at the age of 10 years by the individual-specific slopes (see Figure 7.4). On population level, however, AMMs do not succeed to adequately model the BMI skewness (see Figure 7.2).

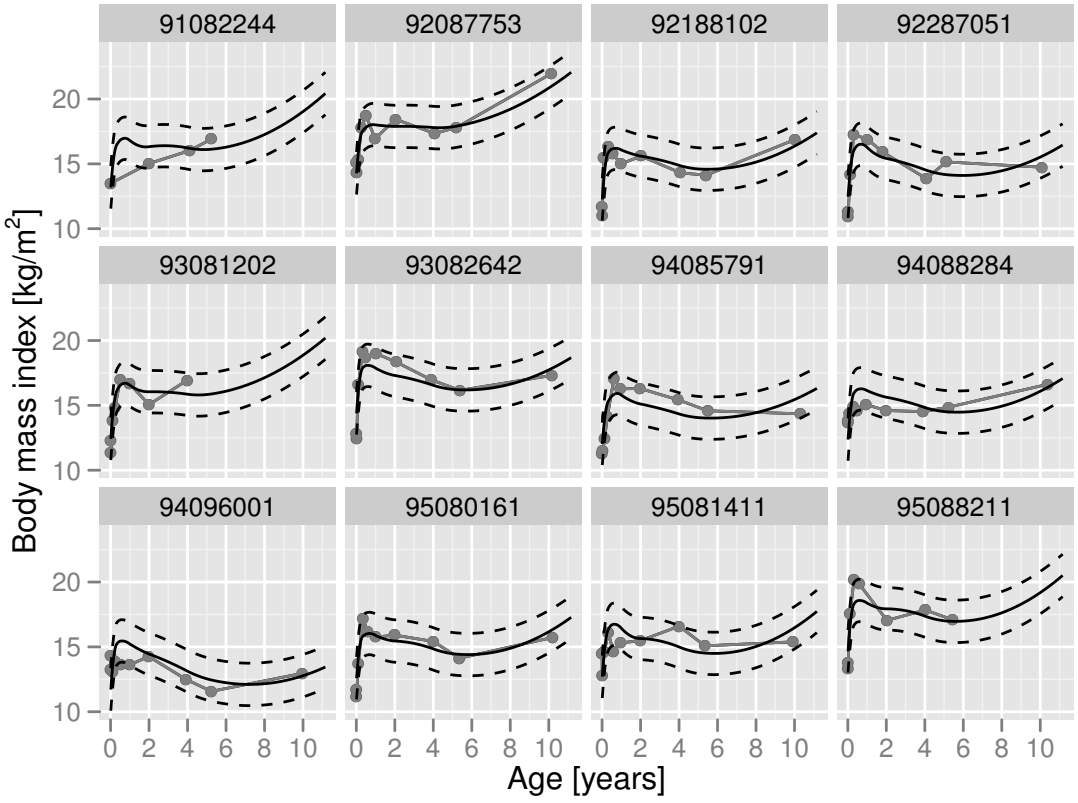


Figure 7.4 Individual-specific BMI quantile curves estimated by AMMs for 12 randomly chosen children. Solid black lines show the estimated mean, while the dashed lines show the estimated 5% and 95% quantile curves, respectively. Observed BMI values are displayed by grey line-connected points.

The results of smooth nonlinear effects for covariates other than age are not shown but briefly described here. With regard to the effect of maternal BMI, the shape of all BMI quantile curves looks roughly similar. The effect increases with increasing maternal BMI and remains constant from maternal BMI values around 30 kg/m^2. The slope of the 97% BMI quantile is estimated to be larger than that of other quantiles. The effect of maternal BMI gain during pregnancy is estimated to be almost linear and slightly increasing throughout all models, which suggests that larger maternal BMI gains during pregnancy result in larger BMI values of children.

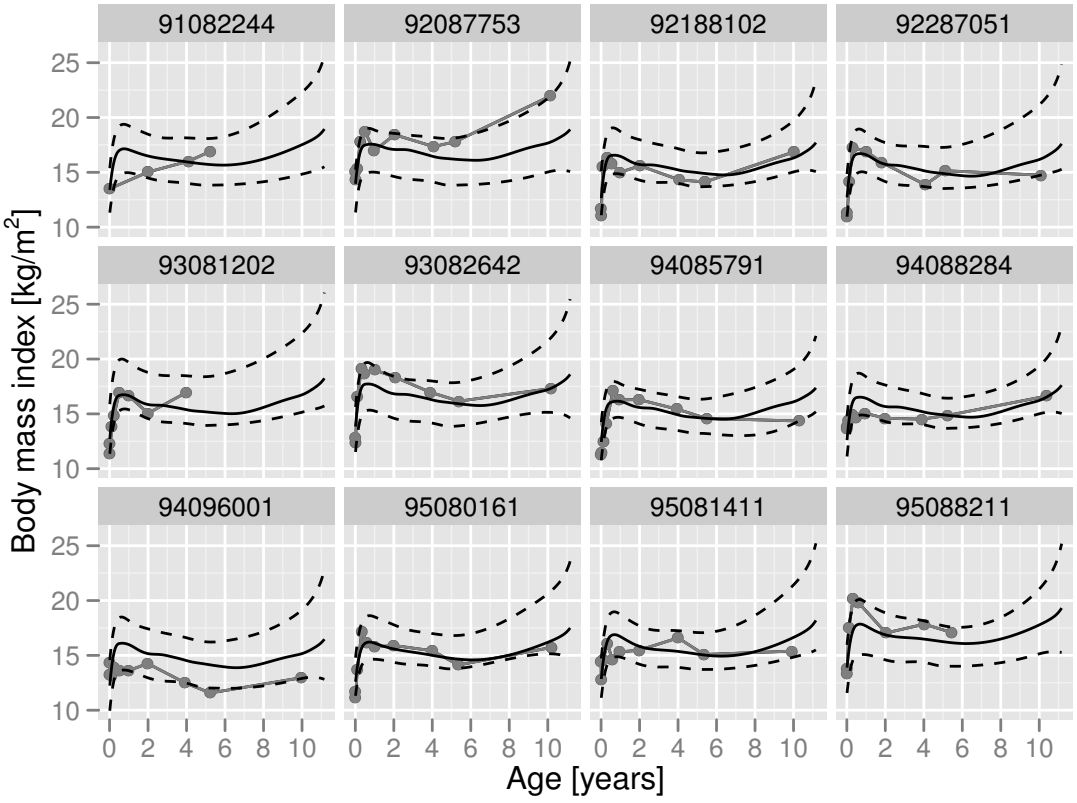


Figure 7.5 Individual-specific BMI quantile curves estimated by STAQ models for 12 randomly chosen children. Solid black lines show the estimated median, while the dashed lines show the estimated 5% and 95% quantile curves, respectively. Observed BMI values are displayed by grey line-connected points.

Regarding age-varying effects, Figure 7.6 exemplarily displays estimated age-varying effects for high compared to low maternal education. The effect of high maternal education is estimated to be almost zero for the BMI median and the 10% quantile. Yet, estimated upper BMI quantiles are smaller for children whose mothers have achieved a high school diploma (compared children of mothers with “CSE or Hauptschule”). These effects are not present at birth and do not emerge before the age of around 5 years. Then they show a continuing decrease until the age of 10 years. One should be cautious to attribute this effect on maternal education only, since maternal education is closely related to the socio-economic status and can also be seen as a proxy of further life style factors.

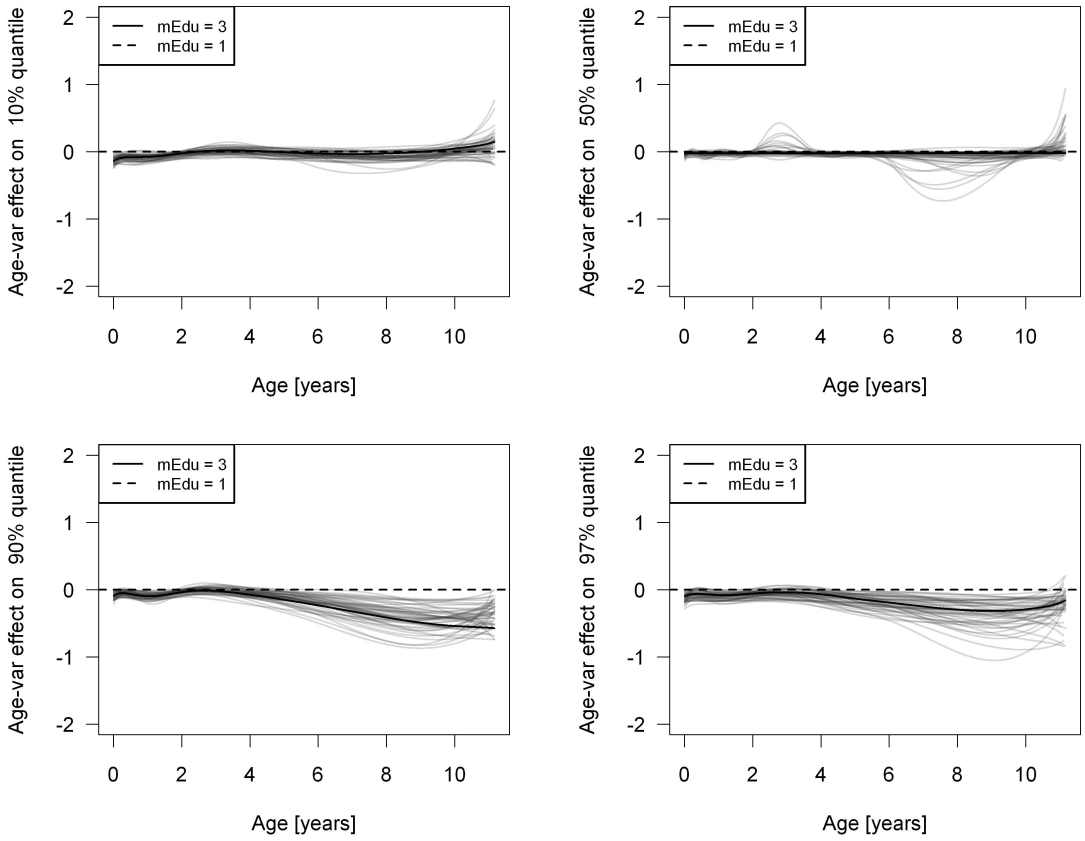


Figure 7.6 Estimated age-varying effects for high maternal education (mEdu=3) compared to low maternal education (mEdu=1) resulting from STAQ models for all four quantile parameters. Age-varying effects are shown by solid black lines (full model) and grey lines (50 different bootstrap samples). Dashed line at zero corresponds to the reference category.

The results for further age-varying effects are shown by Figure 7.7. Concerning sex, conditional BMI quantiles for boys are estimated to be larger than those for girls, and the effect size is clearly varying with age. The results for study location suggest that upper BMI quantiles of children living in urban areas are smaller than those of children from rural areas. Yet, this effect is not present for the median and the 10% quantile and not before the age of 7 years. Both age-varying and main effects of breastfeeding are estimated to be almost zero for all quantiles. Maternal smoking during pregnancy exerts a slightly positive effect between three and six years and no clear effect afterwards, but the effect clearly varies with age. Age-varying effects of maternal education refer to the reference category of low maternal education (mEdu=1, “CSE or Hauptschule”). The age-varying effect of high maternal education was already shown and discussed in Figure 7.6. Medium maternal education (mEdu=2, “secondary school or Realschule”) does not show an age-varying effect.

To sum up, such age-varying effects as observed for sex, study location, maternal smoking, and education can only be detected with STAQ models, since they are only present for upper BMI quantiles.

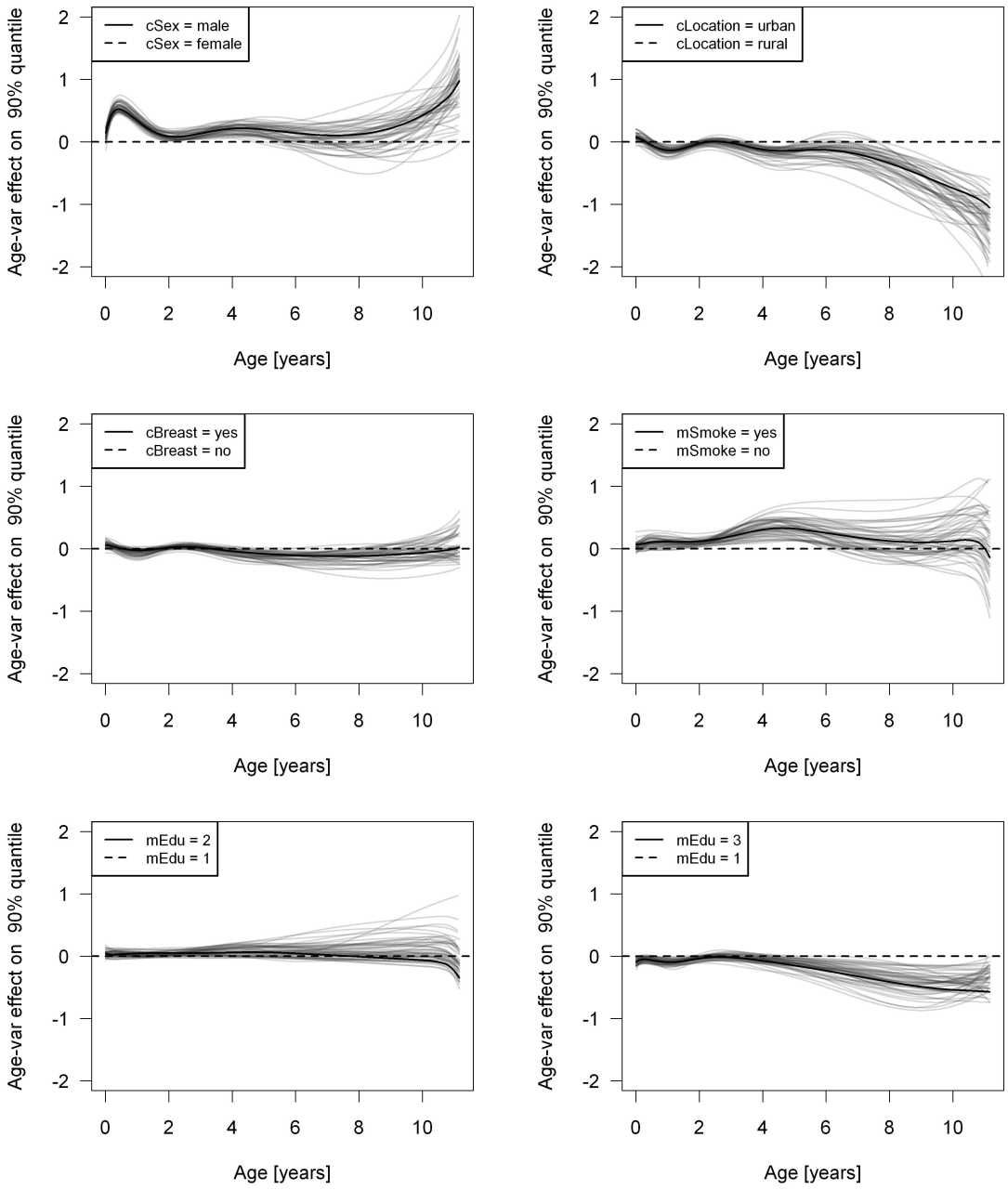


Figure 7.7 Estimated age-varying effects for all categorical covariates resulting from STAQ models for $\tau = 0.90$. Age-varying effects are shown by solid black lines (full model) and grey lines (50 different bootstrap samples). Dashed line at zero corresponds to the respective reference category.

Regarding the results for individual-specific effects, Figure 7.8 shows empirical kernel densities of the estimated child-specific effects from AMM and STAQ model median models. As discussed in Section 5.4, the shapes of the densities seem to be approximately Gaussian. However, the effects estimated by STAQ models are considerably shrunken towards zero compared to the effects estimated by AMM. This can be explained by the inherent shrinkage property of boosting. In particular, the individual-specific slopes are estimated to be almost zero.

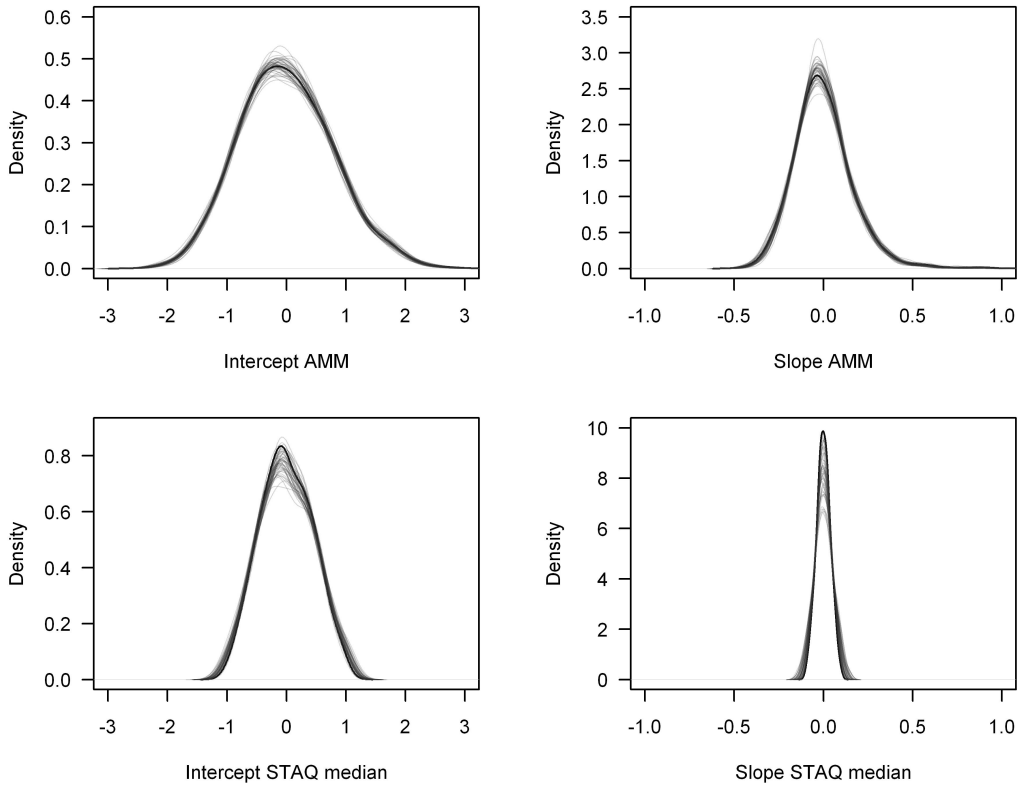


Figure 7.8 Comparison of the empirical densities (obtained by kernel density estimation) of individual-specific intercepts and slopes from AMM and STAQ median models on the full dataset (black line) and on 50 bootstrap samples (grey lines).

Furthermore, Table 7.1 contains the empirical correlation coefficients between estimated individual-specific effects for all models. There are large positive correlations between individual-specific intercepts and rather large correlations between individual-specific slopes. Intercepts and slopes show almost no respective correlations.

Table 7.1 Empirical correlation coefficients between individual-specific intercepts and slopes estimated by the AMM and STAQ models on the full LISA dataset.

		Intercept					Slope				
		AMM	STAQ 10%	STAQ 50%	STAQ 90%	STAQ 97%	AMM	STAQ 10%	STAQ 50%	STAQ 90%	STAQ 97%
Intercept	AMM	1.00	0.74	0.93	0.83	0.64	0.14	0.02	0.03	0.04	0.02
	STAQ 10%	0.74	1.00	0.70	0.41	0.26	0.04	0.01	0.01	-0.01	-0.03
	STAQ 50%	0.93	0.70	1.00	0.71	0.48	0.08	0.03	0.02	0.00	-0.02
	STAQ 90%	0.83	0.41	0.71	1.00	0.79	0.14	0.02	0.04	0.06	0.01
	STAQ 97%	0.64	0.26	0.48	0.79	1.00	0.16	0.02	0.02	0.09	0.05
Slope	AMM	0.14	0.04	0.08	0.14	0.16	1.00	0.51	0.79	0.75	0.58
	STAQ 10%	0.02	0.01	0.03	0.02	0.02	0.51	1.00	0.50	0.18	0.08
	STAQ 50%	0.03	0.01	0.02	0.04	0.02	0.79	0.50	1.00	0.50	0.29
	STAQ 90%	0.04	-0.01	0.00	0.06	0.09	0.75	0.18	0.50	1.00	0.69
	STAQ 97%	0.02	-0.03	-0.02	0.01	0.05	0.58	0.08	0.29	0.69	1.00

To complete the picture, Figure 7.9 shows the boosting paths of estimated individual-specific intercepts and slopes for 200 randomly selected children based on the full models for three different quantile parameters. For all quantiles, one can observe the typical grouping effect of effects estimated by ridge penalization (as described in Section 5.4). The grouping seems to be most pronounced for the intercepts at non-median quantiles. For the median, the individual-specific effects scatter symmetrically around zero. For the other quantiles, however, the distributions of the intercepts are skewed. We attribute this phenomenon to the asymmetric loss function together with the underlying binary character of the negative gradient residuals.

For the 90% quantile, for example, the negative gradient residuals can only take the two values 0.9 and -0.1. The individual-specific intercepts are estimated by least squares related to zero (without intercept). Therefore, the absolute size of negative values cannot be as large as the one of positive values. With additional shrinkage induced by multiplication with the step length parameter ν, the absolute size of negative increments becomes even smaller. Thus, it would take a large number of iterations to obtain a symmetric distribution. For the 10% quantile, it is exactly the other way around and positive increments can only be very small.

Regarding the individual-specific slopes, one can observe that they are not selected before the 2000th iteration in any of the models. With knowledge on the variable selection properties of boosting in mind, this indicates that individual-specific slopes are not as important as intercepts and other covariates in the analysis. The slopes probably capture some of the remaining error variability.

Finally, note that in a simultaneous analysis with respect to missing data, we created an imputed version of the LISA dataset using several missing data imputation methods. Then, we repeated all statistical analyses with the imputed data. Overall, we obtained very similar results as from the complete case approach.

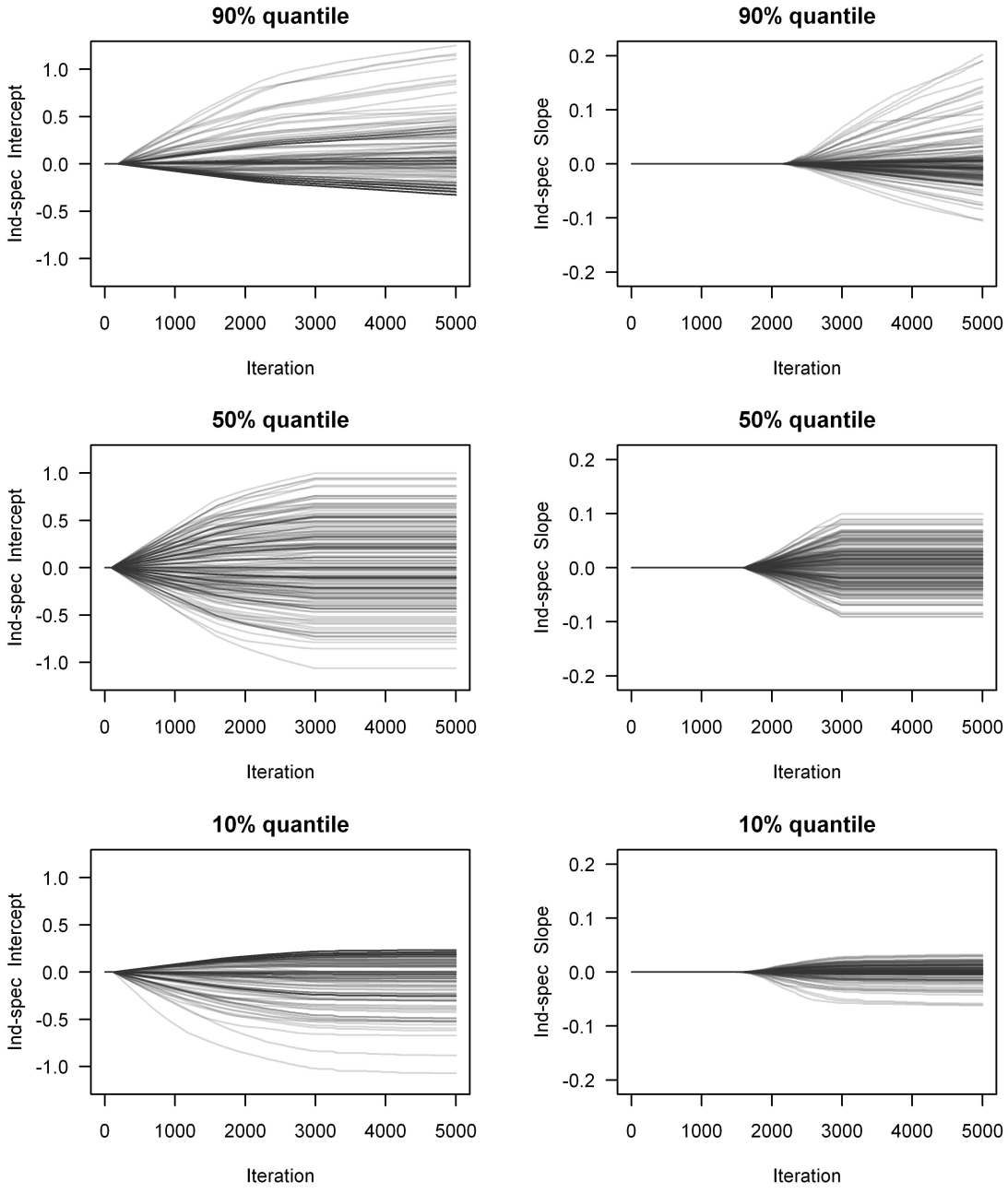


Figure 7.9 Paths of estimated individual-specific intercepts (left column) and individual-specific slopes (right column) for 200 randomly selected children, depending on quantile parameter $\tau \in \{0.10, 0.50, 0.90\}$ and 5000 boosting iterations from the respective models on the full dataset. The degree of grey intensity depends on the number of overlapping paths.

7.3 Discussion

Key findings

The comparison of STAQ models with classical Gaussian AMMs suggested that STAQ models can better handle the age-specific BMI skewness and are thus more adequate than AMMs when the interest is directed towards overweight and obesity.

By using quantile regression, we obtained similar results as Beyerlein *et al.* (2008, 2010) with regard to time-constant risk factors: Apart from age, other risk factors also exert their effect in a different way on upper quantiles of the BMI distribution than on the mean. In our analysis here, we could additionally assess individual-specific BMI patterns during life course and age-varying effects.

The results of the smooth age-varying effects of categorical covariates were particularly interesting. For several variables, quantile curves were estimated to be similar until a certain age period at which the age-varying effects emerged. For example, the effect of high maternal education did not become apparent before the age of around 5 years. Then the effect size slowly increased until it was most pronounced at the age of 10 years. To the best of our knowledge, such age-varying effects have not yet been investigated in the obesity literature.

Strengths and limitations

By applying STAQ regression for upper quantiles of the BMI distribution, it was possible to adequately and flexibly model the age-specific skewness of the BMI distribution while adjusting for other risk factors and individual-specific effects. Since we analyzed raw BMI values directly, our analysis did not require reference charts to construct a binary or Z-score response which are modelled in usual regression approaches for overweight and obesity.

We applied boosting estimation which is currently the only approach that can estimate individual-specific and smooth nonlinear effects as well as varying coefficient terms in the same STAQ predictor. By including individual-specific effects, the model accounted for the temporal correlation between repeated measurements. Varying coefficient terms enabled several age-varying effects of time-constant risk factors to be detected.

In addition, our life-course approach allowed to adequately model individual-specific BMI patterns, as illustrated by Figure 7.5. The figure shows individual-specific 90% BMI prediction intervals which were constructed in accordance with the approaches in Meinshausen (2006) and Mayr *et al.* (2012a). First, two separate models for the 5% and 95% BMI quantiles were estimated and then, the individual-specific quantile predictions were taken as interval limits.

An inherent limitation of boosting estimation is that subsampling strategies have to be applied in order to obtain standard errors. This can be computationally challenging and time-consuming – in particular when a large dataset with many individuals is analysed, as was the case here.

Furthermore, it is probably not yet satisfying that the distribution of the individual-specific effects at upper quantiles was skewed at the optimal stopping iteration, and not Gaussian as theoretically expected. The strong shrinkage and grouping of the individual-specific effects can be attributed to the small degrees of freedom that are conceded to the corresponding base learner in order to

equalize selection probabilities. Due to the small number of covariates in our analysis, we did not pursue to study variable selection results in more detail.

Implications for future research

We believe that quantile boosting for STAQ models is a promising approach for further investigating risk factors for overweight and obesity, since analyzing upper quantiles of the BMI distribution is more adequate than analyzing the mean for this purpose. Furthermore, quantile regression does not incur information coarsening in the same way as usual binary regression approaches.

From an epidemiological point of view, further investigation of the age-varying effects could lead to additional insights and could, for example, explain differences between findings on the impact of risk factors in previous studies relying on different age populations. It would also be interesting to re-run the analysis with data from an additional time point (which is currently collected) since the BMI skewness becomes even more pronounced for older children.

Another relevant research question would be to investigate if children with similar BMI life-course patterns can be clustered into groups. In a previous analysis of the LISA study, Heinzl *et al.* (2012) clustered BMI patterns of children until the age of 6 years based on extensions of AMMs. Analogue statistical methodology would have to be developed for quantile regression.

Regarding the quantile boosting algorithm, the grouping of paths of the individual-specific effects at extreme quantiles should be further investigated.

Finally, it might have seemed obvious that quantile regression would be more adequate than Gaussian mean regression in our analysis. However, we would like to stress that the use of Gaussian AMMs for quantile modelling makes sense when the response distribution is homoscedastic and approximately Gaussian conditional on covariates. In a previous analysis of the LISA study (Fenske *et al.*, 2008), for example, observations were only available until the age of 6 years. AMMs could compete with the more complex quantile regression models since the age-specific BMI skewness is not present until the age of 6 years (see Figure 2.9). In such cases, we would recommend to use the well-studied framework of AMMs for longitudinal quantile regression instead of applying more complex quantile regression strategies.

7.4 Related own work

This section gives a short overview of further work related to the topic of child overweight and obesity with own contributions.

In a recent master's thesis, Reulen (2011) conducted a life-course analysis of risk factors for child overweight and obesity similar to the one presented here. With the particular aim of investigating age-varying effects of time-constant risk factors, STAQ models were estimated by quantile boosting. The analysis relied on a pooled dataset with data from four different longitudinal studies. As in our analysis, age-varying effects were detected for child sex, maternal smoking during pregnancy, and maternal education, and these effects were even more pronounced than in our analysis. These greater absolute sizes can be explained by the larger number of observations in the pooled dataset.

In another recent master's thesis, Riedel (2011) investigated nonlinear effects of risk factors for child overweight and obesity based on a cross-sectional dataset. The analysis setup followed Beyerlein *et al.* (2010) where the effects of continuous covariates had simply been included in a linear way. The main question of the master's thesis was to investigate whether assuming linear effects as in Beyerlein *et al.* (2010) was adequate or too restrictive, i.e., if the effects of continuous covariates should be modelled in a nonlinear way. Therefore, STAQ models were estimated by quantile boosting with the decomposition of smooth nonlinear effects into linear part and nonlinear deviation as described in equation (4.5) on page 67. The results indicated that nonlinear modelling is only necessary for the effect of maternal BMI on upper BMI quantiles. The other continuous covariates (weight gain until the age of around 2 years and birth weight) seemed to be adequately modelled by linear effects for all quantiles.

In Mayr *et al.* (2012c), we studied the construction of prediction intervals (PIs) with quantile boosting based on the LISA study. Due to the skewness of the BMI distribution, the construction of PIs for future BMI values with standard (symmetric) approaches was problematic – in the similar way as using AMMs for BMI modelling. With quantile boosting, the borders of a PI are directly estimated based on an interpretable predictor structure. Thereby, distributional assumptions are avoided and the interval borders are not just parallel shifts of the mean curves (as shown in Figure 7.5). The analysis in Mayr *et al.* (2012c) showed interesting results. For example, estimated PIs for children of mothers who smoked during pregnancy were larger than the PIs for other children. We additionally pointed out the general concept of conditional coverage – in contrast to sample coverage which is usually applied – to prove the accuracy of prediction intervals.

Chapter 8: Discussion and outlook

8.1 Summary and contributions of this thesis

The present thesis introduced the generic model class of structured additive quantile regression (STAQ). In this model class, quantile regression was combined with a structured additive predictor and, thereby, flexible modelling of a variety of different covariate effects was made possible.

To estimate the parameters of STAQ models, a broad overview of existing state-of-the-art estimation approaches was given. Each approach was classified into one of three categories (distribution-free approaches, distribution-based approaches, or related model classes) and was systematically discussed with respect to four previously defined criteria.

We believe that this systematic overview is an important contribution to quantile regression research since we are not aware of any other comparable work. We structured a wide range of estimation approaches for quantile regression and in particular compared distribution-based and distribution-free concepts – in contrast to most of the literature which concentrates on approaches within one of our categories only. For example, the book of Koenker (2005) thoroughly treats one important group of approaches which we denoted by "classical framework" of quantile regression. The book of Hao and Naiman (2007) gives an application-oriented introduction to quantile regression but also focusses on the classical framework. Due to their compact character, most research papers cite only a few approaches related to their new developments. Hence, we think that our overview and the suggested structure can be helpful to classify novel estimation approaches and to spot areas with further research potential.

The main methodological contribution of this thesis is *quantile boosting* – a boosting algorithm that was introduced as alternative approach to estimate STAQ models (based on Fenske, Kneib, and Hothorn, 2011).

We thoroughly discussed quantile boosting with regard to our predefined criteria and evaluated the algorithm in several simulation studies. We concluded that quantile boosting provides great advantages over the other existing estimation approaches regarding the flexible predictor, variable selection in high-dimensional settings and software (see Section 8.2 for details on advantages and limitations of quantile boosting).

Furthermore, we investigated whether the use of STAQ models – in particular in combination with quantile boosting – could lead to new substantial insights in two relevant applications from the field of epidemiology. We conducted a comprehensive analysis of risk factors for child undernutrition in India as well as a life-course analysis of risk factors for child overweight and obesity in Germany. Since the distributions of anthropometric measurements in childhood are typically skewed, STAQ regression was a priori assumed to be well-suited for these applications. We believe that the results of both analyses contribute to subject-matter knowledge on risk factors for undernutrition and obesity (see Section 8.3 for a more detailed discussion).

8.2 Discussion of quantile boosting

This section briefly describes the advantages and limitations of quantile boosting compared to other estimation approaches for STAQ regression. It is based on Section 4.4, which contains a thorough discussion of the properties of quantile boosting.

Advantages of quantile boosting

Boosting with early stopping is a shrinkage method that yields sparse models. Component-wise boosting thus combines two general aims of statistical learning: prediction (by shrinkage of effect estimates) and interpretation (by splitting the predictor into univariate base learners). For quantile regression, component-wise boosting is in particular appealing because the minimization problem relying on a sum of weighted absolute deviations is solved by methods from the well-studied L_2-norm framework.

Quantile boosting offers advantages over the other presented estimation approaches for STAQ models with respect to almost all four criteria for model assessment. Regarding the flexible predictor, quantile boosting can estimate a large variety of different effects in the same STAQ predictor. In particular the combination of smooth nonlinear and individual-specific effects has so far not been possible for other distribution-free estimation approaches. With respect to nonlinear effects, quantile boosting allows for the data-driven determination of the amount of smoothness of nonlinear effects – contrary to estimation by total variation regularization yielding piecewise linear functions.

One of the major advantages of quantile boosting over the other estimation approaches relate to the inherent variable selection and model choice properties. Since parameter estimation and variable selection are performed in one single estimation step, boosting is particularly favourable for high-dimensional predictors and can even be applied in settings with (much) more covariates than observations. Furthermore, boosting can decide on linearity vs. nonlinearity of an effect in a fully data-driven way within the fitting process.

The superiority of boosting in high-dimensional settings was also confirmed by our simulation studies. In low-dimensional simulation setups, quantile boosting performed on par with estimation approaches from the classical framework. In higher-dimensional simulation setups, however, quantile boosting clearly outperformed all other approaches.

Compared to software for other estimation approaches, the R package `mboost` is currently the only software that allows to fit the full variety of different effect types from the structured additive predictor. In comparison with the (standard) R package `quantreg`, more complex models with a larger number of smooth nonlinear effects can be fitted.

Limitations of quantile boosting

The probably most problematic limitation of boosting is the lack of standard errors for the parameter estimators. Since boosting just yields point estimators, subsampling strategies, such as the bootstrap, have to be applied to obtain standard errors. However, this should not be rated as a fundamental drawback compared to other estimation approaches. In practice, most of them also rely on bootstrap to obtain standard errors since the asymptotic covariance matrix of the estimators depends on the true (unknown) error density.

Another limitation of quantile boosting is that quantile crossing is not prevented. Similar to the majority of the other estimation approaches, the estimation is performed separately for different quantile parameters and, therefore, quantile crossing can occur. From the approaches considered in this thesis, quantile crossing is only prevented with Gaussian STAR models and GAMLSS due to their direct modelling of the full conditional response distribution.

Finally, boosting estimation can be computationally expensive and time-consuming for large datasets. Even though the R package `mboost` provides a well-equipped and user-friendly software, some basic understanding of the (relatively complex) boosting algorithm is required to correctly handle and interpret the estimation results. Therefore, statistical modelling with boosting might be a challenging task for practitioners.

8.3 Discussion of the application results

In this section, we briefly discuss if applying STAQ models together with quantile boosting led to new substantial insights regarding our applications. More detailed discussions of the respective results can be found in Sections 6.3 and 7.3.

Undernutrition of children in India (Section 2.1 and Chapter 6)

The results of our quantile regression and logistic regression models with the same flexible predictor were largely comparable with respect to the size and significance of estimated effects and the shape of nonlinear effects. We attribute these findings to the symmetric distribution of the Z-score response, suggesting that standard logistic regression can compete with quantile regression for analyzing determinants of child stunting.

However, our analysis showed that several continuous covariates exert their effect on stunting in a nonlinear way. By additionally including age-varying effects for breastfeeding and complementary feeding in the model predictor, we could also confirm the age-dependent impact of these variables on child stunting and thereby investigate the corresponding WHO recommendations. We conclude that a flexible predictor should be considered in future regression analyses of potential determinants of child stunting. As demonstrated by our analysis, the estimation of such models can for example be realized by boosting.

Our analysis was also innovative with respect to the selection of covariates for the analysis. We developed a schematic diagram of the multiple determinants of stunting and selected the covariates for our later regression analyses based on this diagram.

As a result, a large number of covariates was included in our analyses. Since boosting combines parameter estimation and variable selection in one estimation step, subsequent steps of variable selection were not necessary – contrary to classical likelihood-based regression where, for example, AIC-based model comparisons would have had to be conducted.

Finally, for all but one of the groups of determinants we conceptualized in our diagram, we found at least one variable with a significant effect in all quantile and logistic regression analyses. This emphasizes the broad range of causes of child stunting and suggests many potential entry-points for intervention.

Overweight and obesity of children in Germany (Section 2.2 and Chapter 7)

Our analysis showed that STAQ models can better handle the age-specific BMI skewness beginning at the age of seven than additive mixed models with Gaussian errors. Since we were interested in analyzing overweight and obesity, quantile regression for upper BMI quantiles was more adequate than mean regression. Regarding estimation, quantile boosting is currently the only approach that allows to fit longitudinal STAQ models with many individuals, as present for the LISA data. Hence, we combined an adequate statistical model with the currently only possible estimation method in our obesity analysis.

With regard to risk factors for overweight and obesity, we obtained novel results for the smooth age-varying effects of categorical covariates. To the best of our knowledge, such age-varying effects have not yet been investigated in the obesity literature. The results for individual-specific BMI patterns were also interesting and we plan further investigation of these findings.

From a methodological point of view, the comparison of STAQ and GAMLSS models for analyzing obesity would also have been interesting. However, the currently available software does not allow to estimate longitudinal GAMLSS models with a large number of individuals. The alternative of boosting estimation based on the gamboostLSS algorithm (Mayr *et al.*, 2012a) has not yet been sufficiently investigated for estimating longitudinal GAMLSS models.

In summary, we believe that STAQ models based on quantile boosting are a promising approach for longitudinal quantile regression, not only for modelling overweight and obesity, but for all applications where the shape of a response variable changes over time depending on covariates.

8.4 Possible directions for future research

In the following, we sketch several possible entry-points for future research on estimation approaches for STAQ models and related model classes. We thereby focus on alternatives for the classical framework (which aims at direct minimization of the quantile regression loss criterion).

Boosting framework

Beginning with boosting algorithms, we think that further empirical investigations to compare STAQ and GAMLSS models estimated by boosting would be useful since these approaches have not yet been compared in practice. In our view, GAMLSS has high potential for becoming a competing model class to quantile regression, in particular in combination with boosting estimation and its beneficial variable selection properties. Thus, we think that empirical comparisons between quantile boosting and the gamboostLSS approach (Mayr *et al.*, 2012a) could lead to additional insights, especially when flexible predictors in longitudinal data settings are considered.

In addition, empirical evaluations of the stability selection procedure of Meinshausen and Bühlmann (2010) would be interesting since its performance in connection with boosting has not yet been studied. Stability selection controls the family-wise error rate, i.e., the probability to falsely include base learners in the model, and offers a formal variable selection procedure that can simply be applied to the final model after boosting estimation. Moreover, it does not require the calculation of the final degrees of freedom. Our results from the analysis of child stunting in India also indicate that stability selection is a promising procedure for formal variable selection subsequent to boosting.

Expectile regression

In this thesis, we considered expectile regression as related model class of quantile regression (see Section 3.5.1). The main advantage of expectile regression over the classical framework of quantile regression is that its quadratic loss function is continuously differentiable and, therefore, estimation can be performed within the well-studied L_2 framework. This is somehow similar to boosting which tackles the quantile regression minimization problem of weighted absolute differences by penalized least squares base learners.

However, expectile regression offers the additional advantage that the asymptotic covariance matrix of the estimators does not depend on the true error density. For this reason, standard errors can be easily obtained with expectile regression.

Even though expectiles are not provided with an intuitive interpretation (as given for quantiles), we believe that expectile regression is very promising for becoming a supporting or even competing model class for quantile regression. In our view, empirical comparisons of structured additive expectile and quantile regression, for example in simulation studies and data applications, can contribute to further figure out similarities and differences of both approaches.

Simultaneous modelling of the complete response distribution

In recent years, several approaches have been developed which aim at simultaneous modelling of the complete response distribution conditional on covariates. As a distribution-based example for these approaches, we considered GAMLSS models in this thesis (see Section 3.5.3). Another example is given by conditional transformation models that are completely distribution-free and were recently developed by Hothorn *et al.* (2012). Approaches which aim at simultaneous inference for all response quantiles have also been developed based on flexible Bayesian estimation. In Reich *et al.* (2011) and Reich (2012), for example, the (stochastic) quantile process was considered and modelled as a whole.

Although most of these models do not provide an interpretable relationship between covariates and quantile function, the implicit prevention of quantile crossing makes them particularly appealing for quantile modelling. Therefore, we believe that these approaches can be good alternatives to quantile regression in appropriate application scenarios, and further investigation of these models seems to be promising.

Flexible Bayesian quantile regression

Finally, we believe that further developments of flexible Bayesian estimation approaches offer additional research potential. Embedding quantile regression in a full Bayesian framework would help to overcome problems regarding the asymptotic covariance matrix of quantile regression estimators since credibility intervals and standard errors for the parameter estimates are directly available with Bayesian estimation.

Another great advantage of Bayesian algorithms relates to longitudinal data. From a Bayesian point of view, all model parameters are treated as random and no difference is made between “fixed” and “random” effects. Therefore, the extension of a Bayesian algorithm to random effects for longitudinal data is natural and often straightforward when choosing adequate prior distributions. This makes Bayesian approaches particularly promising since appropriate methods for longitudinal quantile regression are urgently needed.

In our view, existing Bayesian approaches based on the asymmetric Laplace distribution (see, for example, Yu and Moyeed, 2001; Tsionas, 2003; Yue and Rue, 2011) do not provide a solution to the problem of parameter inference. Due to the quasi-likelihood character of the error density, these approaches in most cases lead to misspecified standard errors, as for example demonstrated in Reich *et al.* (2010).

The main challenges for developing a suitable Bayesian approach for quantile regression are (i) to specify a flexible error density that "imitates" the distribution-free character of quantile regression and at the same time is appropriate for MCMC sampling, and (ii) to handle the stochastic constraint $F_{\varepsilon_{\tau i}}(0) = \tau$ on the cdf of the error distribution.

To be more concrete, we shortly sketch our own ideas for a flexible Bayesian approach in the following. In analogy to the approach in Reich *et al.* (2010), which was shortly sketched in Section 3.4.2 (see model (3.9)), our idea is to consider the following location-scale model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\gamma})\, \varepsilon_i \qquad \text{with} \quad \varepsilon_i \overset{iid}{\sim} F_\varepsilon \;, \tag{8.1}$$

where $\boldsymbol{x}_i^\top \boldsymbol{\gamma}$ is constrained to be positive for all $\boldsymbol{x}_i$. The error terms ε_i are assumed to be *iid* with cdf F_ε and density f_ε. For a given quantile parameter $\tau \in (0, 1)$, model (8.1) is identical to the following quantile regression model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau + (\boldsymbol{x}_i^\top \boldsymbol{\gamma})\, \varepsilon_{\tau i} \qquad \text{with} \quad \varepsilon_{\tau i} \overset{iid}{\sim} F_{\varepsilon_\tau} \;. \tag{8.2}$$

Here, $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta} + z_\tau \boldsymbol{\gamma})$ denote quantile-specific coefficients and $\varepsilon_{\tau i} = \varepsilon_i - z_\tau$ stand for error terms corresponding to the original errors ε_i shifted by the $\tau \cdot 100\%$ error quantile $z_\tau = F_\varepsilon^{-1}(\tau)$. Thus, the quantile-specific errors $\varepsilon_{\tau i}$ are also *iid* distributed with cdf F_{ε_τ}, which is just a location-shifted version of F_ε and by definition fulfills the quantile constraint $F_{\varepsilon_\tau}(0) = \tau$.

We express the error density f_ε in model (8.1) by a finite mixture of Gaussian densities

$$f_\varepsilon(\varepsilon) = \sum_{k=1}^{K} w_k \phi_k(\varepsilon | \mu_k, \sigma_0^2) \;,$$

with K being the number of mixture components, weights w_k with $\sum_{k=1}^K w_k = 1$ and $w_k > 0$, and Gaussian mixture densities $\phi_k(\cdot|\mu_k, \sigma_0^2)$ for components $k = 1, \ldots, K$ with mean μ_k and fixed variance σ_0^2.

We estimate the error density f_ε by a penalized Gaussian mixture as described in Komárek and Lesaffre (2008). The number of density components is chosen very large, e.g., $K = 20$, and the corresponding means μ_k are fixed on a fine equidistant grid so that $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$. After a suitable transformation of the weights w_k to avoid constraints, the weight parameters are estimated in a penalized way so that weights of neighbouring density components are more similar than weights of non-neighbouring components. Estimation is then realized by MCMC sampling.

Note that this concept for estimating the error distribution corresponds to estimating nonlinear effects based on P-splines. As remarked by Komárek and Lesaffre (2008), this model can be seen as a limiting case of classical B-spline smoothing by Eilers and Marx (1996) since Gaussian densities result when the degree of B-spline basis functions tends to ∞.

Altogether, this estimation procedure considerably simplifies the approach suggested in Reich *et al.* (2010) for the same location-scale model. Since the models (8.1) and (8.2) are equivalent, we can concentrate on the estimation of the parameters for model (8.1) and afterwards simply calculate the estimators $\hat{\boldsymbol{\beta}}_\tau = \hat{\boldsymbol{\beta}} + \hat{z}_\tau \hat{\boldsymbol{\gamma}}$ and their standard errors for model (8.2).

Although this approach might seem to be limited since it only addresses a location-scale model, we think that it is very promising for flexible quantile regression. The model specifies the complete conditional response distribution and covers heteroscedastic data settings with non-standard error densities. Furthermore, the estimation concept is intuitive and much easier than, for example, the one proposed by Reich *et al.* (2010) which tries to handle the quantile constraint in a complicated way. Extensions of the model to a structured additive predictor would be straightforward since the different types of effects could be realized in a similar way as in the different components of the Bayesian STAR model in Fahrmeir *et al.* (2004).

We plan to work out our approach in more detail in the next future and to compare the estimation results with those from quantile boosting.

Bibliography

Agras WS, Mascola AJ (2005). "Risk factors for childhood overweight." *Current Opinion in Pediatrics*, **17**, 648–652.

Aigner DJ, Amemiya T, Poirier DJ (1976). "On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function." *International Economic Review*, **17**(2), 377–396.

Arulampalam W, Naylor RA, Smith J (2011). "Am I missing something? The effects of absence from class on student performance." *Economics of Education Review*, **31**(4), 363–375. doi: 10.1016/j.econedurev.2011.12.002.

Belitz C, Brezger A, Kneib T, Lang S, Umlauf N (2012). *BayesX – Bayesian inference in structured additive regression models*. Version 2.1, URL `http://www.stat.uni-muenchen.de/ bayesx`.

Benoit DF, Al-Hamzawi R, Yu K, van den Poel D (2011). *bayesQR: Bayesian quantile regression*. R package version 1.3, URL `http://CRAN.R-project.org/package=bayesQR`.

Beyerlein A, Fahrmeir L, Mansmann U, Toschke AM (2008). "Alternative regression models to assess increase in childhood BMI." *BMC Medical Research Methodology*, **8**, 59. doi: 10.1186/1471-2288-8-59.

Beyerlein A, Toschke AM, von Kries R (2010). "Risk factors for childhood overweight: shift of the mean body mass index and shift of the upper percentiles: results from a cross-sectional study." *International Journal of Obesity*, **34**(4), 642–648. doi:10.1038/ijo.2009.301.

Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, Mathers C, Rivera J (2008). "Maternal and child undernutrition: global and regional exposures and health consequences." *The Lancet*, **371**(9608), 243–260. doi:10.1016/S0140-6736(07)61690-0.

Bollaerts K, Eilers PHC, Aerts M (2006). "Quantile regression with monotonicity restrictions using P-splines and the L_1-norm." *Statistical Modelling*, **6**, 189–207. doi: 10.1191/1471082X06st118oa.

Borghi E, de Onis M, Garza C, van den Broeck J, Frongillo EA, Grummer-Strawn L, van Buuren S, Pan H, Molinari L, Martorell R, Onyango AW, Martines JC, WHO Multicentre Growth Reference Study Group (2006). "Construction of the World Health Organization child growth standards: selection of methods for attained growth curves." *Statistics in Medicine*, **25**, 247–265. doi: 10.1002/sim.2227.

Breckling J, Chambers R (1988). "M-quantiles." *Biometrika*, **75**(4), 761–771.

Breiman L (2001). "Random forests." *Machine Learning*, **45**(1), 5–32. doi: 10.1023/A:1010933404324.

Buchinsky M (1998). "Recent advances in quantile regression models: a practical guideline for empirical research." *The Journal of Human Resources*, **33**(1), 88–126. doi:10.2307/146316.

Bühlmann P (2006). "Boosting for high-dimensional linear models." *The Annals of Statistics*, **34**, 559–583. doi:10.1214/009053606000000092.

Bühlmann P, Hothorn T (2007). "Boosting algorithms: regularization, prediction and model fitting (with discussion)." *Statistical Science*, **22**, 477–522. doi:10.1214/07-STS242.

Bühlmann P, Yu B (2003). "Boosting with the L_2 loss: regression and classification." *Journal of the American Statistical Association*, **98**(462), 324–338. doi:10.1198/016214503000125.

Cade BS, Terrell JW, Porath MT (2008). "Estimating fish body condition with quantile regression." *North American Journal of Fisheries Management*, **28**(2), 349–359. doi:10.1577/M07-048.1.

Cannon AJ (2011). "Quantile regression neural networks: implementation in R and application to precipitation downscaling." *Computers & Geosciences*, **37**(9), 1277–1284. doi:10.1016/j.cageo.2010.07.005.

Caulfield LE, Richard SA, Rivera JA, Musgrove P, Black RE (2006). "Stunting, wasting and micronutrient deficiency disorders." In DT Jamison, JG Breman, ARM et al (eds.), "Disease control priorities in developing countries," pp. 551–568. Oxford University Press and the World Bank, New York, 2nd edition.

Chaudhuri P, Loh WY (2002). "Nonparametric estimation of conditional quantiles using quantile regression trees." *Bernoulli*, **8**(5), 561–576.

Chen C (2007). "A finite smoothing algorithm for quantile regression." *Journal of Computational and Graphical Statistics*, **16**(1), 136–164. doi:10.1198/106186007X180336.

Cheng Y, de Gooijer JG, Zerom D (2011). "Efficient estimation of an additive quantile regression model." *Scandinavian Journal of Statistics*, **38**(1), 46–62. doi:10.1111/j.1467-9469.2010.00706.x.

Christmann A, Hable R (2012). "Consistency of support vector machines using additive kernels for additive models." *Computational Statistics & Data Analysis*, **56**(4), 854–873. doi:10.1016/j.csda.2011.04.006.

Cieczynski S (2009). *Bayesianische Quantilregression*. Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München. In German.

Cole T, Bellizzi M, Flegal K, Dietz W (2000). "Establishing a standard definition for child overweight and obesity worldwide: international survey." *British Medical Journal*, **320**, 1240–1245. doi:10.1136/bmj.320.7244.1240.

De Gooijer JG, Zerom D (2003). "On additive conditional quantiles with high-dimensional covariates." *Journal of the American Statistical Association*, **98**(461), 135–146. doi:10.1198/016214503388619166.

Efron B (1991). "Regression percentiles using asymmetric squared loss." *Statistica Sinica*, **1**(1), 93–125.

Eilers PHC, Marx BD (1996). "Flexible smoothing with B-splines and penalties." *Statistical Science*, **11**(2), 89–121. doi:10.1214/ss/1038425655.

Fahrmeir L, Kneib T, Lang S (2004). "Penalized structured additive regression for space-time data: a Bayesian perspective." *Statistica Sinica*, **14**, 731–761.

Fahrmeir L, Kneib T, Lang S (2007). *Regression – Modelle, Methoden und Anwendungen*. Serie: Statistik und ihre Anwendungen. Springer.

Fahrmeir L, Tutz G (2001). *Multivariate statistical modelling based on generalized linear models*. Springer Series in Statistics. Springer, 2nd edition.

Farcomeni A (2012). "Quantile regression for longitudinal data based on latent Markov subject-specific parameters." *Statistics and Computing*, **22**(1), 141–152. doi:10.1007/s11222-010-9213-0.

Fenske N (2008). *Flexible Longitudinaldaten-Regression mit Anwendungen auf Adipositas*. Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München. In German.

Fenske N, Burns J, Hothorn T, Rehfuess EA (2012a). "Understanding child stunting in India: a comprehensive analysis of socio-economic, nutritional and environmental determinants using quantile boosting." *American Journal of Clinical Nutrition*. To be submitted.

Fenske N, Fahrmeir L, Hothorn T, Rzehak P, Höhle M (2012b). "Boosting structured additive quantile regression for longitudinal childhood obesity data." *International Journal of Biostatistics*. Submitted.

Fenske N, Fahrmeir L, Rzehak P, Höhle M (2008). "Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data." *Technical Report 38*, Institut für Statistik, Ludwig-Maximilians-Universität München. URL `http://epub.ub.uni-muenchen.de/6260/`.

Fenske N, Kneib T, Hothorn T (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.

Franck E, Nüsch S (2012). "Talent and/or popularity: What does it take to be a superstar?" *Economic Inquiry*, **50**(1), 202–216. doi:10.1111/j.1465-7295.2010.00360.x.

Freedman DS, Khan LK, Serdula MS, Dietz WH, Srinivasan SR, Berenson GS (2005). "The relation of childhood BMI to adult adiposity: the Bogalusa heart study." *Pediatrics*, **115**, 22–27. doi:10.1542/peds.2004-0220.

Freund Y, Schapire R (1996). "Experiments with a new boosting algorithm." In "Proceedings of the Thirteenth International Conference on Machine Learning Theory," San Francisco: Morgan Kaufmann Publishers Inc.

Freund Y, Schapire R (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, **55**(1), 119–139. doi:10.1006/jcss.1997.1504.

Friedman J (2001). "Greedy function approximation: a gradient boosting machine." *The Annals of Statistics*, **29**(5), 1189–1232. doi:10.1214/aos/1013203451.

Friedman J, Hastie T, Tibshirani R (2000). "Additive logistic regression: a statistical view of boosting." *The Annals of Statistics*, **28**. doi:10.1214/aos/1016218223.

Galvao AF, Montes-Rojas GV (2010). "Penalized quantile regression for dynamic panel data." *Journal of Statistical Planning and Inference*, **140**(11), 3476–3497. doi: 10.1016/j.jspi.2010.05.008.

Geraci M (2012). *lqmm: Linear quantile mixed models*. R package version 1.01, URL `http://CRAN.R-project.org/package=lqmm`.

Geraci M, Bottai M (2007). "Quantile regression for longitudinal data using the asymmetric Laplace distribution." *Biostatistics*, **8**(1), 140–154. doi:10.1093/biostatistics/kxj039.

Gilchrist W (2008). "Regression revisited." *International Statistical Review*, **76**(3), 401–418. doi: 10.1111/j.1751-5823.2008.00053.x.

Habicht JP (2004). "Expert consultation on the optimal duration of exclusive breastfeeding: the process, recommendations, and challenges for the future." *Advances in Experimental Medicine and Biology*, **554**, 79–87.

Hable R (2012). "Asymptotic confidence sets for general nonparametric regression and classification by regularized kernel methods." *Technical report*, arXiv.org. URL `http://arxiv.org/abs/1203.4354`.

Hao L, Naiman DQ (2007). *Quantile regression*. Number 07-149 in Quantitative Applications in the Social Sciences. SAGE Publications.

Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer New York, 2 edition.

Heinzl F, Fahrmeir L, Kneib T (2012). "Additive mixed models with Dirichlet process mixture and P-spline priors." *AStA Advances in Statistical Analysis*, **96**(1), 47–68. doi:10.1007/s10182-011-0161-6.

Hofner B (2011). *Boosting in structured additive models*. Verlag Dr. Hut, München. Ph.D. thesis.

Hofner B, Hothorn T, Kneib T, Schmid M (2011a). "A framework for unbiased model selection based on boosting." *Journal of Computational and Graphical Statistics*, **20**(4), 956–971. doi: 10.1198/jcgs.2011.09220.

Hofner B, Mayr A, Fenske N, Schmid M (2011b). *gamboostLSS: Boosting methods for GAMLSS models*. R package version 1.0-3, URL `http://CRAN.R-project.org/package=gamboostLSS`.

Hofner B, Müller J, Hothorn T (2011c). "Monotonicity-constrained species distribution models." *Ecology*, **92**, 1895–1901. doi:10.1890/10-2276.1.

Horowitz JL, Lee S (2005). "Nonparametric estimation of an additive quantile regression model." *Journal of the American Statistical Association*, **100**(472), 1238–1249. doi: 10.1198/016214505000000583.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2010). "Model-based boosting 2.0." *Journal of Machine Learning Research*, **11**, 1851–1855.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2012). *mboost: Model-Based Boosting*. R package version 2.1-2, URL `http://CRAN.R-project.org/package=mboost`.

Hothorn T, Kneib T, Bühlmann P (2012). "Conditional transformation models." *Technical report*, arXiv.org. URL http://arxiv.org/abs/1201.5786.

International Institute for Population Sciences, Macro International (2007). "National Family Health Survey (NFHS-3), 2005–06, India: Volume I." URL http://www.nfhsindia.org/about.shtml.

Jones M (1994). "Expectiles and M-quantiles are quantiles." *Statistics & Probability Letters*, **20**(2), 149–153. doi:10.1016/0167-7152(94)90031-0.

Jones MC, Yu K (2007). "Improved double kernel local linear quantile regression." *Statistical Modelling*, **7**(4), 377–389. doi:10.1177/1471082X0700700407.

Kandala NB, Fahrmeir L, Klasen S, Priebe J (2009). "Geo-additive models of childhood undernutrition in three sub-Saharan African countries." *Population, Space and Place*, **15**(5), 461–473. doi:10.1002/psp.524.

Kandala NB, Lang S, Klasen S, Fahrmeir L (2001). "Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries." *Technical Report 245*, SFB 386, Ludwig-Maximilians-Universität München. URL http://epub.ub.uni-muenchen.de/1626/.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). "kernlab – An S4 package for kernel methods in R." *Journal of Statistical Software*, **11**(9), 1–20.

Karlsson A (2007). "Nonlinear quantile regression estimation of longitudinal data." *Communications in Statistics - Simulation and Computation*, **37**(1), 114–131. doi:10.1080/03610910701723963.

Kim MO (2007). "Quantile regression with varying coefficients." *The Annals of Statistics*, **35**(1), 92–108. doi:10.1214/009053606000000966.

Kneib T, Heinzl F, Brezger A, Bove DS (2011). *BayesX: R utilities accompanying the software package BayesX*. R package version 0.2-5, URL http://CRAN.R-project.org/package=BayesX.

Kneib T, Hothorn T, Tutz G (2009). "Variable selection and model choice in geoadditive regression models." *Biometrics*, **65**(2), 626–634. doi:10.1111/j.1541-0420.2008.01112.x.

Kocherginsky M, He X, Mu Y (2005). "Practical confidence intervals for regression quantiles." *Journal of Computational and Graphical Statistics*, **14**(1), 41–55. doi:10.1198/106186005X27563.

Koenker R (2004). "Quantile regression for longitudinal data." *Journal of Multivariate Statistics*, **91**, 74–89. doi:10.1016/j.jmva.2004.05.006.

Koenker R (2005). *Quantile regression*. Economic Society Monographs. Cambridge University Press.

Koenker R (2011). "Additive models for quantile regression: model selection and confidence bandaids." *Brazilian Journal of Probability and Statistics*, **25**(3), 239–262. doi:10.1214/10-BJPS131.

Koenker R (2012). *quantreg: Quantile regression*. R package version 4.90, URL `http://CRAN.R-project.org/package=quantreg`.

Koenker R, Bache SH (2012). *rqpd: Regression quantiles for panel data*. R package version 0.5, URL `http://rqpd.r-forge.r-project.org/`.

Koenker R, Bassett G (1978). "Regression quantiles." *Econometrica*, **46**(1), 33–50. doi: 10.2307/1913643.

Koenker R, Mizera I (2004). "Penalized triograms: total variation regularization for bivariate smoothing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 145–163. doi:10.1111/j.1467-9868.2004.00437.x.

Koenker R, Ng P, Portnoy S (1994). "Quantile smoothing splines." *Biometrika*, **81**(4), 673–680. doi:10.1093/biomet/81.4.673.

Komárek A, Lesaffre E (2008). "Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions." *Journal of the American Statistical Association*, **103**(482), 523–533. doi:10.1198/016214507000000563.

Kosti RI, Panagiotakos DB (2006). "The epidemic of obesity in children and adolescents in the world." *Central European Journal of Public Health*, **14**(4), 151–159.

Kottas A, Krnjajić M (2009). "Bayesian semiparametric modelling in quantile regression." *Scandinavian Journal of Statistics*, **36**(2), 297–319. doi:10.1111/j.1467-9469.2008.00626.x.

Kriegler B, 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**(3), 1234–1255. doi:10.1214/10-AOAS328.

Kromeyer-Hauschild K, Wabitsch M, Kunze D, Geller F, Geiß HC, Hesse V, von Hippel A, Jaeger U, Johnsen D, Korte W, Menner K, Müller G, Müller JM, Niemann-Pilatus A, Remer T, Schaefer F, Wittchen HU, Zabransky S, Zellner K, Ziegler A, Hebebrand J (2001). "Perzentile für den Body-Mass-Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben." *Monatsschrift Kinderheilkunde*, **149**(8), 807–818.

Kyu HH, Georgiades K, Boyle MH (2009). "Maternal smoking, biofuel smoke exposure and child height-for-age in seven developing countries." *International Journal of Epidemiology*, **38**(5), 1342–1350. doi:10.1093/ije/dyp253.

Lamerz A, Kuepper-Nybelen J, Wehle C, Bruning N, Trost-Brinkhues G, Brenner H, Hebebrand J, Herpertz-Dahlmann B (2005). "Social class, parental education, and obesity prevalence in a study of six-year-old children in Germany." *International Journal of Obesity*, **29**, 373–380. doi:10.1038/sj.ijo.0802914.

Li Y, Graubard BI, Korn EL (2010). "Application of nonparametric quantile regression to body mass index percentile curves from survey data." *Statistics in Medicine*, **29**(5), 558–572. doi: 10.1002/sim.3810.

Li Y, Zhu J (2008). "L_1-norm quantile regression." *Journal of Computational and Graphical Statistics*, **17**(1), 163–185. doi:10.1198/106186008X289155.

LISA–plus study group (1998–2008). URL http://www.helmholtz-muenchen.de/epi/arbeitsgruppen/umweltepidemiologie/projects-projekte/lisa-plus/index.html.

Liu Y, Bottai M (2009). "Mixed-effects models for conditional quantiles with longitudinal data." *The International Journal of Biostatistics*, **5**(1), 28. doi:10.2202/1557-4679.1186.

Lobstein T, Bauer L, Uauy R (2004). "Obesity in children and young people: a crisis in public health." *Obesity Reviews*, **5**(Suppl.1), 4–85.

Matano A, Naticchioni P (2012). "Wage distribution and the spatial sorting of workers." *Journal of Economic Geography*, **12**(2), 379–408. doi:10.1093/jeg/lbr013.

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012a). "Generalized additive models for location scale and shape for high-dimensional data – a flexible approach based on boosting." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **61**(3), 403–427. doi:10.1111/j.1467-9876.2011.01033.x.

Mayr A, Hofner B, Schmid M (2012b). "The importance of knowing when to stop – a sequential stopping rule for component-wise gradient boosting." *Methods of Information in Medicine*, **51**(2), 178–186. doi:10.3414/ME11-02-0030.

Mayr A, Hothorn T, Fenske N (2012c). "Prediction intervals for future BMI values of individual children – a non-parametric approach by quantile boosting." *BMC Medical Research Methodology*, **12**(6). doi:10.1186/1471-2288-12-6.

Mehtätalo L, Gregoire TG, Burkhart HE (2008). "Comparing strategies for modeling tree diameter percentiles from remeasured plots." *Environmetrics*, **19**(5), 529–548. doi:10.1002/env.896.

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

Meinshausen N (2012). *quantregForest: Quantile regression forests*. R package version 0.2-3, URL http://CRAN.R-project.org/package=quantregForest.

Meinshausen N, Bühlmann P (2010). "Stability selection." *Journal of the Royal Statistical Society, Series B*, **72**(4), 417–473. doi:10.1111/j.1467-9868.2010.00740.x.

Mishra V, Retherford RD (2007). "Does biofuel smoke contribute to anaemia and stunting in early childhood?" *International Journal of Epidemiology*, **36**(1), 117–129. doi:10.1093/ije/dyl234.

Newey WK, Powell JL (1987). "Asymmetric least squares estimation and testing." *Econometrica*, **55**(4), 819–847.

Pendakur K, Woodcock S (2010). "Glass ceilings or glass doors? Wage disparity within and between firms." *Journal of Business and Economic Statistics*, **28**(1), 181–189. doi:10.1198/jbes.2009.08124.

R Development Core Team (2012). *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/.

Reich BJ (2012). "Spatiotemporal quantile regression for detecting distributional changes in environmental processes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(4), 535–553. doi:10.1111/j.1467-9876.2011.01025.x.

Reich BJ, Bondell HD, Wang H (2010). "Flexible Bayesian quantile regression for independent and clustered data." *Biostatistics*, **11**, 337–352. doi:10.1093/biostatistics/kxp049.

Reich BJ, Fuentes M, Dunson DB (2011). "Bayesian spatial quantile regression." *Journal of the American Statistical Association*, **106**(493), 6–20. doi:10.1198/jasa.2010.ap09237.

Reilly JJ, Armstrong J, Dorosty AR, Emmett PM, Ness A, Rogers I, Steer C, Sherriff A (2005). "Early life risk factors for obesity in childhood: cohort study." *British Medical Journal*, **330**, 1357–1363. doi:10.1136/bmj.38470.670903.E0.

Reulen H (2011). *Wann wirken Risiken? Modellierung longitudinaler Gewichtsverläufe im Kindesalter mit flexibler Quantilregression*. Masterarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München. In German.

Riedel C (2011). *Nichtlineare Quantilregressionseffekte in Daten zu Risikofaktoren für Übergewicht bei Kindern*. Masterarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München. In German.

Rigby RA, Stasinopoulos DM (2004). "Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution." *Statistics in Medicine*, **23**(19), 3053–3076. doi:10.1002/sim.1861.

Rigby RA, Stasinopoulos DM (2005). "Generalized additive models for location, scale and shape." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Rigby RA, Stasinopoulos DM (2006). "Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis." *Statistical Modelling*, **6**(3), 209–229. doi:10.1191/1471082X06st122oa.

Rue H, Martino S, Lindgren F (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured (geo-)additive models using Integrated Nested Laplace Approximaxion*. R package, URL `http://www.r-inla.org/download`.

Ruppert D, Wand M, Carroll R (2003). *Semiparametric regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ruppert D, Wand M, Carroll R (2009). "Semiparametric regression during 2003–2007." *Electronic Journal of Statistics*, **3**, 1193–1256. doi:10.1214/09-EJS525.

Rzehak P, Sausenthaler S, Koletzko S, Bauer CP, Schaaf B, von Berg A, Berdel D, Borte M, Herbarth O, Kraemer U, Fenske N, Wichmann HE, Heinrich J (2009). "Period-specific growth, overweight and modification by breastfeeding in the GINI and LISA birth cohorts up to age 6 years." *European Journal of Epidemiology*, **24**, 449–467. doi:10.1007/s10654-009-9356-5.

Sassi F, Devaux M, Cecchini M, Rusticelli E (2009). "The obesity epidemic: analysis of past and projected future trends in selected OECD countries." *Technical Report 45*. doi:10.1787/225215402672. OECD Health Working Papers.

Schaffrath Rosario A, Kurth BM, Stolzenberg H, Ellert U (2010). "Body mass index percentiles for children and adolescents in Germany based on a nationally representative sample (KiGGS 2003-2006)." *European Journal of Clinical Nutrition*, **64**, 341–349. doi:10.1038/ejcn.2010.8.

Scheipl F (2011). *amer: Additive mixed models with lme4*. R package version 0.6.10, URL http://CRAN.R-project.org/package=amer.

Schmid M, Hothorn T (2008). "Boosting additive models using component-wise P-splines." *Computational Statistics & Data Analysis*, **53**(2), 298–311. doi:10.1016/j.csda.2008.09.009.

Schnabel SK, Eilers PH (2009). "Optimal expectile smoothing." *Computational Statistics & Data Analysis*, **53**(12), 4168–4177. doi:10.1016/j.csda.2009.05.002.

Semba RD, de Pee S, Sun K, Bloem MW, Raju VK (2010). "The role of expanded coverage of the national vitamin A program in preventing morbidity and mortality among preschool children in India." *The Journal of Nutrition*, **140**(1), 208S–212S. doi:10.3945/jn.109.110700.

Sobotka F, Kauermann G, Schulze Waltrup L, Kneib T (2011). "On confidence intervals for semiparametric expectile regression." *Statistics and Computing*, pp. 1–14. doi:10.1007/s11222-011-9297-1.

Sobotka F, Kneib T (2012). "Geoadditive expectile regression." *Computational Statistics and Data Analysis*, **56**(4), 755–767. doi:10.1016/j.csda.2010.11.015.

Sobotka F, Schnabel S, Schulze Waltrup L, with contributions from Eilers P, Kneib T and Kauermann G (2012). *expectreg: Expectile and quantile regression*. R package version 0.35, URL http://CRAN.R-project.org/package=expectreg.

Stasinopoulos DM, Rigby RA (2007). "Generalized additive models for location scale and shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7).

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008). "Conditional variable importance for random forests." *BMC Bioinformatics*, **9**(307), 1471–2105. doi:10.1186/1471-2105-9-307.

Taddy MA, Kottas A (2010). "A Bayesian nonparametric approach to inference for quantile regression." *Journal of Business & Economic Statistics*, **28**(3), 357–369. doi:10.1198/jbes.2009.07331.

Takeuchi I, Le QV, Sears TD, Smola AJ (2006). "Nonparametric quantile estimation." *Journal of Machine Learning Research*, **7**, 1231–1264.

Taylor JW (2000). "A quantile regression neural network approach to estimating the conditional density of multiperiod returns." *Journal of Forecasting*, **19**(4), 299–311. doi:10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V.

Theußl S, Zeileis A (2009). "Collaborative software development using R-Forge." *The R Journal*, **1**(1), 9–15.

Tsionas EG (2003). "Bayesian quantile inference." *Journal of Statistical Computation and Simulation*, **73**(9), 659–674. doi:10.1080/0094965031000064463.

Umlauf N, Kneib T, Lang S, Zeileis A (2012). *R2BayesX: Estimate structured additive regression models with BayesX*. R package version 0.1-1, URL http://CRAN.R-project.org/package=R2BayesX.

UNICEF (1998). *The state of the World's children 1998: focus on nutrition*. Oxford University Press, New York.

UNICEF, WHO, World Bank, UN DESA/Population Division (2011). *Levels and trends in child mortality: report 2011*. Oxford University Press, New York.

Wang HJ, Zhu Z, Zhou J (2009). "Quantile regression in partially linear varying coefficient models." *The Annals of Statistics*, **37**(6B), 3841–3866. doi:10.1214/09-AOS695.

Wei Y, Pere A, Koenker R, He XM (2006). "Quantile regression methods for reference growth charts." *Statistics in Medicine*, **25**(8), 1369–1382. doi:10.1002/sim.2271.

WHO (2012). "Global database on child growth and malnutrition." URL `http://www.who.int/nutgrowthdb/estimates/en/`.

WHO Consultation on Obesity (1999). "Obesity: preventing and managing the global epidemic: report of a WHO consultation." *Technical Report 894*, Geneva, Switzerland. WHO Technical Report Series.

WHO Multicentre Growth Reference Study Group (2006). *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization, Geneva, Switzerland.

Wood S (2012). *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.7-20, URL `http://CRAN.R-project.org/package=mgcv`.

Yu K, Jones MC (1998). "Local linear quantile regression." *Journal of the American Statistical Asoociation*, **93**(441), 228–237.

Yu K, Lu Z (2004). "Local linear additive quantile regression." *Scandinavian Journal of Statistics*, **31**(3), 333–346. doi:10.1111/j.1467-9469.2004.03_035.x.

Yu K, Lu Z, Stander J (2003). "Quantile regression: applications and current research areas." *The Statistician*, **52**(3), 331–350. doi:10.1111/1467-9884.00363.

Yu K, Moyeed RA (2001). "Bayesian quantile regression." *Statistics and Probability Letters*, **54**, 437–447. doi:10.1016/S0167-7152(01)00124-9.

Yu K, Zhang J (2005). "A three-parameter asymmetric Laplace distribution and its extension." *Communications in Statistics – Theory and Methods*, **34**(9), 1867–1879. doi:10.1080/03610920500199018.

Yuan Y, Yin G (2010). "Bayesian quantile regression for longitudinal studies with nonignorable missing data." *Biometrics*, **66**(1), 105–114. doi:10.1111/j.1541-0420.2009.01269.x.

Yue YR, Rue H (2011). "Bayesian inference for additive mixed quantile regression models." *Computational Statistics & Data Analysis*, **55**(1), 84–96. doi:10.1016/j.csda.2010.05.006.

Zhang T, Yu B (2005). "Boosting with early stopping: convergence and consistency." *The Annals of Statistics*, **33**, 1538–1579. doi:10.1214/009053605000000255.

Zheng S (2012). "QBoost: predicting quantiles with boosting for regression and binary classification." *Expert Systems with Applications*, **39**(2), 1687–1697. doi:10.1016/j.eswa.2011.06.060.

Zou H, Hastie T (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

Eidesstattliche Versicherung

(Gemäß Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne unerlaubte Beihilfe angefertigt und keine anderen als die angegebenen Hilfsmittel verwendet habe.

München, den 17. September 2012 Nora Fenske