



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Nora Fenske, Ludwig Fahrmeir, Peter Rzehak, Michael Höhle

# Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data

Technical Report Number 038, 2008  
Department of Statistics  
University of Munich

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



# Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data

Nora Fenske<sup>1,\*</sup>, Ludwig Fahrmeir<sup>1</sup>, Peter Rzehak<sup>2,3,4</sup>, Michael Höhle<sup>1,3</sup>

<sup>1</sup> Department of Statistics, Ludwig-Maximilians-Universität München, Germany

<sup>2</sup> Institute of Epidemiology, HelmholtzZentrum München – German Research Center for Environmental Health, Neuherberg, Germany

<sup>3</sup> Munich Center of Health Sciences, Ludwig-Maximilians-Universität München, Germany

<sup>4</sup> Department of Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-Universität München, Germany

---

## Abstract

This article compares and discusses three different statistical methods for investigating risk factors for overweight and obesity in early childhood by means of the LISA study, a recent German birth cohort study with 3097 children. Since the definition of overweight and obesity is typically based on upper quantiles (90% and 97%) of the age specific body mass index (BMI) distribution, our aim was to model the influence of risk factors and age on these quantiles while as far as possible taking the longitudinal data structure into account. The following statistical regression models were chosen: additive mixed models, generalized additive models for location, scale and shape (GAMLSS), and distribution free quantile regression models. The methods were compared empirically by cross-validation and for the data at hand no model could be rated superior. Motivated by previous studies we explored whether there is an age-specific skewness of the BMI distribution. The investigated data does not suggest such an effect, even after adjusting for risk factors. Concerning risk factors, our results mainly confirm results obtained in previous studies. From a methodological point of view, we conclude that GAMLSS and distribution free quantile regression are promising approaches for longitudinal quantile regression, requiring, however, further extensions to fully account for longitudinal data structures.

**Key words:** GAMLSS, quantile regression, longitudinal data, mixed models, body mass index, obesity, overweight

---

## 1 Introduction

Childhood obesity has become an area of public health focus since its prevalence has increased continuously in industrialized countries during the last decades (Kosti and Panagiotakos (2006), Lobstein and Frelut (2003), Lobstein et al. (2004)). With the objective of prevention it is therefore essential to detect the main risk factors of obesity.

---

\* Adress of correspondence: Nora Fenske, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr.33, 80539 München, Germany; E-mail: nora.fenske@stat.uni-muenchen.de

The *epidemiological aim* of the present analysis was to investigate whether the LISA study, a recent German birth cohort study, could confirm risk factors, like parental overweight, socioeconomic factors or breastfeeding, that were found in previous cohort studies such as Agras and Mascola (2005) and Haslam and James (2005). We decided to realize this aim by directly assessing the impact of risk factors on upper BMI quantiles of the study population. This differs from the usual approach of classifying children as obese using reference charts, which are based on age specific upper quantiles of the body mass index (BMI) distribution in reference populations (Borghetti et al. (2006)), followed by a logistic regression for the resulting binary response. Modeling quantiles directly avoids possible loss of information implied by reducing the original continuous response BMI to the binary response obesity. Furthermore, logit (and probit) models assume a specific symmetric distribution for the original continuous response variable. Possible skewness of BMI distributions makes the use of conventional logit (and probit) models questionable. As a consequence, our interest was to determine the influence of covariates on upper BMI quantiles of the study population, such as the 90% BMI quantile for overweight and the 97% BMI quantile for obesity, not on quantiles derived from reference charts being based on a population of different nature and characteristics than the one investigated. In particular, we wanted to explore the hypothesis that the shape of the BMI distribution changes with a child's age.

Our *statistical aim* was thus to consider appropriate statistical regression methods for the quantile modeling task given by the epidemiological questions and to evaluate their relative merits for longitudinal quantile regression. Choice of methods was done with the following criteria in mind: To what extent does the method take the longitudinal data structure of the LISA data into account? Does the method allow flexible quantile modeling of the response variable and can non-linear or time-varying covariate effects be easily integrated? Is the method capable of handling an age-specific skewness of the BMI distribution, as described in Cole et al. (2000)?

Hence, we decided to apply and to compare the following three regression approaches, described in more detail in Section 2.2: Additive mixed models for longitudinal data, see e.g. Verbeke and Molenberghs (1997) or Fitzmaurice et al. (2004), generalized additive mixed models of location scale and shape (GAMLSS, Rigby and Stasinopoulos (2005)), which have already been applied to cross-sectional BMI modeling in Rigby and Stasinopoulos (2004), and distribution free quantile regression (Koenker (2005)), which has recently been propagated for a cross-sectional study on adult body size in Terry et al. (2007). In order to perform a quantitative comparison between the three approaches we used cross-validation with a criterion that is suitable for quantile modeling.

The rest of this article is organized as follows. Section 2 describes the study and, in concise form, the regression approaches used in our analyses. Section 3 contains and describes the results. From a methodological point of view, we conclude that all three models led to similar results for the data at hand, and based on cross-validation none of them could be rated superior. Generally, we consider GAMLSS and distribution free quantile regression as useful and promising approaches for longitudinal quantile regression. In conclusion, we point out the need to extend these models to fully take into account longitudinal data structures in the discussion section.

## 2 Materials and Methods

### 2.1 Study design and study population

The LISA study is a prospective longitudinal birth cohort study in four cities of Germany (Bad Honnef, Leipzig, Munich, Wesel). 3097 healthy neonates born between 11/1997 and 01/1999 were included in the study. Follow-up time was until the age of six by questionnaires in connection with the nine mandatory examinations (well-baby check-up "U-Untersuchungen") at birth and around the age of 2 weeks, 1, 3, 6, 12, 24, 48 and 60 months. Thus, the maximum number of observations per child was nine.

Originally the study was designed to determine the influence of life-style factors, environmental exposures and health-related behaviour on the development of the immune system and the incidence of allergic diseases in children. However, the LISA study is at the same time suited to investigate the effect of covariates on the body mass index.

Table 1 as well as Table 2 give an overview over the covariates that were included in the analysis. They cover most of the risk factors that are discussed in the mentioned literature such as parental overweight (maternal BMI), socioeconomic factors (sex, area and maternal education), physical activity (hours spent TV watching and computer playing as well as hours spent outdoor), nutrition (breastfeeding) and rapid growth factors as discussed in Ong et al. (2000) and Toschke et al. (2004). The latter is equivalent to weight gain until the age of two years, maternal weight gain during pregnancy and maternal prenatal smoking. All covariates except for age are considered as being time constant.

Covariate	Unit	Median	Mean	Sd	N
Age	Years	0.47	1.27	1.70	
Weight gain until the age of 2 years	<i>kg</i>	8.85	8.92	1.29	2508
Hours spent outdoor at the age of 4 years	Hours per day	3.50	3.50	1.17	2419
Maternal BMI at pregnancy begin	<i>kg/m<sup>2</sup></i>	21.72	22.65	3.92	2972
Maternal BMI gain during pregnancy	<i>kg/m<sup>2</sup></i>	4.96	5.17	1.71	2908

Table 1: Description of the continuous covariates. Age is a time dependent covariate with a total of 23175 observations (on average 7.6 and maximum 9 observations per child).

Missing data problems were handled by using only complete cases for the statistical modeling. Hence, if an observation of a time constant covariate was missing, all observations of the respective child were excluded from the analysis. If on the other hand a single observation of age or BMI was missing, this particular observation was excluded from the analysis. Finally a total of 17316 observations from 2043 children were included in the statistical modeling.

Covariate	Categories	Frequency	N
Sex	0 = female	48.8%	1511
	1 = male	51.2%	1586
Nutrition until the age of 4 months	0 = bottle-feed and/or breastfeeding	48.6%	1506
	1 = breastfeeding only	51.4%	1591
Hours spent TV watching and computer playing at 4 years	1 = less than 1 hour	55.2%	1709
	2 = 1 up to 2 hours	21.6%	669
	3 = more than 3 hours	1.3%	41
	NA = missing value	21.9%	678
Maternal smoking during pregnancy	0 = no	78.8%	2441
	1 = yes	15.8%	490
	NA = missing value	5.3%	166
Maternal highest level of education	1 = No degree	0.6%	18
	2 = Hauptschule (CSE)	5.3%	165
	3 = Realschule (secondary school)	28.5%	884
	4 = Fachabitur (subject specific high school diploma)	13.9%	430
	5 = Abitur (high school diploma)	29.2%	904
	NA = missing value	22.5%	696
Area	0 = Rural (Bad Honnef, Wesel)	21.1%	654
	1 = Urban (Leipzig, Munich)	78.9%	2443

Table 2: Description of the discrete covariates. Percentages are related to all 3097 children.

## 2.2 Statistical analysis

From a statistical point of view the question was to model the quantiles of a response variable depending on covariates by taking into account the longitudinal data structure as far as possible. Thus, we chose three methods to model the relationship between upper BMI quantiles and covariates in the LISA data: Additive mixed models for longitudinal data, generalized additive models of location, scale and shape (GAMLSS) and distribution free quantile regression. The selected models do not fulfill all selection criteria given in the introduction, particularly the only model that accounts for a longitudinal data structure is the additive mixed model. Conceptually, the other models allow the inclusion of random effects, which would honor the longitudinal structure, but currently a computationally feasible implementation does not yet exist.

All methods will be briefly described below using the following notation:

- Indices  $i = 1, \dots, N$  for individuals,  $j = 1, \dots, n_i$  for intra-individual measurements
- Total number of observations:  $n = \sum_{i=1}^N n_i$
- Response variable  $y_{ij}$  (BMI) and design vector  $\mathbf{x}_{ij}$  of selected covariates for individual  $i$  at  $j$ th observation
- Quantile function for the  $\tau \cdot 100\%$  quantile of the response variable  $Y$  conditional upon a given covariate vector  $\mathbf{x}$ :  $Q_Y(\tau|\mathbf{x}) = F_Y^{-1}(\tau|\mathbf{x})$  where  $F_Y^{-1}(\tau|\mathbf{x})$  is the inverse of the distribution function of  $Y|\mathbf{x}$

### Additive mixed models (AMM)

Linear and additive mixed models are a common statistical approach to model relationships between covariates and the conditional expectation of a response variable in longitudinal data (see e.g. Verbeke and Molenberghs (1997) or Fitzmaurice et al. (2004)). In the present analysis the model can be expressed as

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + h_1(v_{ij1}) + \dots + h_m(v_{ijm}) + b_i + \varepsilon_{ij} = \eta_{ij}^{(\mu)} + b_i + \varepsilon_{ij} \quad (1)$$

with independent error terms  $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ . The linear part  $\mathbf{x}'_{ij}\boldsymbol{\beta}$  of the predictor  $\eta_{ij}^{(\mu)}$  contains the usual fixed effects of risk factors  $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$  while the unknown functions  $h_l(\cdot), l = 1, \dots, m$ , model possible nonlinear effects of continuous covariates  $\mathbf{v}_{ij} = (v_{ij1}, \dots, v_{ijm})'$  such as age, weight gain until second birthday, maternal BMI and maternal BMI gain during pregnancy. The additional individual random intercepts  $b_i \sim \mathcal{N}(0, \sigma_b^2)$  are mutually independent and independent from the error terms  $\varepsilon_{ij}$ . By means of the random intercept as well as flexible modeling of the time scale (here: age) the model accounts for the longitudinal data structure. In our case, the covariate vector  $\mathbf{x}_{ij}$  contains the time-constant discrete risk factors from Table 2. Labeling the predictor  $\eta_{ij}$  with  $\mu$  arises from the implicit property that AMMs model the conditional expectation of the response variable (but not conditional variance or skewness).

The resulting conditional and marginal quantile functions of the response variable are given as

$$Q_{Y_{ij}}(\tau | \eta_{ij}^{(\mu)}, b_i) = \eta_{ij}^{(\mu)} + b_i + z_\tau \sigma_\varepsilon \quad (2)$$

$$Q_{Y_{ij}}(\tau | \eta_{ij}^{(\mu)}) = \eta_{ij}^{(\mu)} + z_\tau (\sigma_\varepsilon + \sigma_b) \quad (3)$$

where  $z_\tau$  is the  $\tau$ -100%-quantile of a standard normal distribution. Thus, AMM allow the estimation of quantiles of the response variable's distribution based on the assumption of a normal distributed response variable given the covariates. In case of additional random effects, i.e. random slopes for time-dependent covariates, the expression for the marginal quantile in (3) becomes more involved because of the variance component now depending on covariates.

### GAMLSS

Generalized linear models for location, scale and shape (Rigby and Stasinopoulos (2005)) aim at modeling the parameters of a response variable's distribution. Therefore two main assumptions are made: (1) The response variable follows a known distribution with density  $f(y_{ij} | \boldsymbol{\theta}_{ij})$  conditional on the parameter vector  $\boldsymbol{\theta}_{ij}$ . (2) The observations  $y_{ij}$  are mutually independent given the parameter vectors  $\boldsymbol{\theta}_{ij}$ .

For the data at hand we chose the Box-Cox power exponential (BCPE) distribution since it is very flexible and was already used for the modeling of BMI data (Borghi et al. (2006), Rigby and Stasinopoulos (2004)). Hence,  $Y_{ij} > 0$  (BMI) is assumed to be a random variable with BCPE distribution and parameter vector  $\boldsymbol{\theta}_{ij} = (\mu_{ij}, \sigma_{ij}, \nu_{ij}, \varphi_{ij})$ . These parameters denote location  $\mu_{ij} > 0$ , scale  $\sigma_{ij} > 0$ , skewness  $\nu_{ij} \in \mathbb{R}$  and kurtosis  $\varphi_{ij} > 0$  and are each modeled by separate predictors as follows:

$$\mu_{ij} = \eta_{ij}^{(\mu)}, \quad \sigma_{ij} = \exp(\eta_{ij}^{(\sigma)}), \quad \nu_{ij} = \eta_{ij}^{(\nu)}, \quad \varphi_{ij} = \exp(\eta_{ij}^{(\varphi)}). \quad (4)$$

The predictors  $\eta_{ij}^{(\cdot)}$  are all composed of fixed and nonlinear effects in analogy to (1), hence

$$\eta_{ij}^{(\cdot)} = \mathbf{x}'_{ij}\boldsymbol{\beta} + h_{.1}(v_{.1ij}) + \dots + h_{.m}(v_{.mij}) \quad (5)$$

where the design vectors  $\mathbf{x}'_{ij}$  and  $\mathbf{v}'_{ij}$  as well as parameters  $\beta$ . and nonlinear effects  $h_l(\cdot), l = 1, \dots, m$ , are predictor specific.

GAMLSS provide a large flexibility since they allow the use of a wide range of different distributions for the response together with very flexible predictors. Theoretically the longitudinal data structure can be modeled by random effects (Rigby and Stasinopoulos (2005)), but at present no computationally feasible implementation for large sample sizes and complex models exists. Hence, no random intercepts  $b_i$  are included in the model formula, but a thorough and flexible modeling of the time effect (here: age) to some extent also accounts for the temporal correlation occurring in longitudinal data, see Fahrmeir and Kneib (2008) for a discussion on the relation between trend and correlation.

As the conditional distribution of the response variable is assumed to be the BCPE-distribution, the quantiles can be expressed as

$$Q_{Y_{ij}}(\tau | \eta_{ij}^{(\mu)}, \eta_{ij}^{(\sigma)}, \eta_{ij}^{(\nu)}, \eta_{ij}^{(\varphi)}) = \begin{cases} \mu_{ij} (1 + \sigma_{ij} \nu_{ij} q_\tau)^{1/\nu_{ij}} & \text{if } \nu_{ij} \neq 0 \\ \mu_{ij} \exp(\sigma_{ij} q_\tau) & \text{if } \nu_{ij} = 0 \end{cases} . \quad (6)$$

To estimate the quantile function the theoretical parameters can be replaced by their estimates. The quantile specific parameter  $q_\tau$  is given as

$$q_\tau = \begin{cases} -c [ 2 F_S^{-1}(1 - 2\tau) ]^{1/\varphi_{ij}} & \text{if } \tau \leq 0.5 \\ c [ 2 F_S^{-1}(2\tau - 1) ]^{1/\varphi_{ij}} & \text{if } \tau > 0.5 \end{cases} \quad (7)$$

with  $c$  a specific constant and  $F_S^{-1}$  the quantile function of a gamma distributed random variable  $S$ , see Rigby and Stasinopoulos (2004) for details.

### Distribution free quantile regression

Distribution free quantile regression models are directed at modeling fixed, specific  $\tau$ -100% quantiles of a response variable in a completely nonparametric way, i.e. without assuming any parametric assumption for the response variable. These models are thoroughly treated in Koenker (2005).

For the present analysis a quantile regression model for a fixed value of  $\tau$  can be expressed as

$$y_{ij} = \eta_{ij}^{(\tau)} + \varepsilon_{\tau ij} = \mathbf{x}'_{ij} \beta_\tau + h_{\tau 1}(v_{\tau 1 ij}) + \dots + h_{\tau m}(v_{\tau m ij}) + \varepsilon_{\tau ij} \quad (8)$$

$$\text{with } \int_{-\infty}^0 f(\varepsilon_{\tau ij}) d\varepsilon_{\tau ij} = F_{\varepsilon_{\tau ij}}(0) = \tau . \quad (9)$$

Here, the error terms  $\varepsilon_{\tau ij}$  are assumed to be mutually independent, but without any specific distributional assumption apart from the restriction given in (9). Moreover the fixed effects  $\beta_\tau$  and the functions  $h_{\tau l}(\cdot), l = 1, \dots, m$  are quantile specific. Due to the condition  $F_{\varepsilon_{\tau ij}}(0) = \tau$  for the error terms, it follows that this model implies a quantile modeling of the response variable and hence

$$Q_{Y_{ij}}(\tau | \eta_{ij}^{(\tau)}) = \eta_{ij}^{(\tau)} . \quad (10)$$

An alternative presentation of distribution free quantile regression is via the minimization criterion

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \rho_{\tau}(y_{ij} - \eta_{ij}^{(\tau)}) \rightarrow \min \quad (11)$$

where  $\rho_{\tau}(u) = u(\tau - I(u < 0))$  is the check function with  $I(\cdot)$  being an indicator function. For  $\tau = 0.5$  the check function is proportional to the absolute value function, i.e.  $\rho_{0.5}(u) = 0.5|u|$ . Minimization of (11) can be done by linear programming and leads to the  $\tau \cdot 100\%$  quantiles of the response variable (Koenker (2005)). Thus, the check function is the appropriate loss function for quantile regression problems regarded from a decision theoretical point of view.

### Nonlinear effects

All models permit the estimation of nonlinear covariate effects  $h(\cdot)$  via polynomial spline functions. In our work we used B-spline basis functions with fixed knots and without penalty (Ruppert et al. (2003), Fahrmeir and Tutz (2001)). Since they are defined recursively it is not really possible to present them using a closed form expression opposite to the more intuitive truncated power series basis. Nevertheless we chose B-splines because they are computationally more robust (Eilers and Marx (1996)). The evaluations of the basis functions at the fixed knots, which were positioned quantile based, can be included in the linear design matrices so that model estimation is straightforward.

### Model estimation

All model estimation was performed in R – a free software environment for statistical computing and graphics (R Development Core Team (2008)) by means of the following model specific functions: `lme` from package `nlme` for additive mixed models (Pinheiro et al. (2007)), `gamlss` from package `gamlss` (Stasinopoulos et al. (2008)) and `rq` from package `quantreg` for distribution free quantile regression (Koenker (2008)).

For each of the three models a full model was estimated with main effects for all available covariates and an interaction effect between age and breastfeeding. The degrees of freedom for the nonlinear B-spline covariate effects were adapted by minimizing a model specific criterion, i.e. Generalized Akaike’s Information Criterion for AMM and GAMLSS given as

$$\text{GAIC}(k) = -2 \cdot \text{loglikelihood}(\hat{\theta}) + k \cdot p \quad (12)$$

where  $\hat{\theta}$  denotes the ML estimator of the parameter vector  $\theta$  with length  $p$ . In case of quantile regression, Koenker’s pseudo GAIC was used which means that the term  $\text{loglikelihood}(\hat{\theta})$  in (12) is replaced by the minimization criterion in (11), see Koenker (2005). For alle criteria a penalty parameter of  $k = 3$  was chosen. Finally, model fits were checked by residual diagnostics.

### Model comparison via cross-validation

Fitted models were compared by 5-fold cross-validation to identify the most appropriate model for the data at hand (Hastie et al. (2001)). Thus, the complete set of individuals was randomly divided into a partition of 5 parts having approximately equal size. Thereafter a model  $\mathcal{M}$  was considered as better than another at a specific quantile  $\tau$  if it had a smaller cross-validation criterion, as given by:

$$CV(\mathcal{M}, \tau) = \frac{1}{n} \sum_{i=1}^N \sum_{j=1}^{n_i} \rho_{\tau}(y_{ij} - \hat{y}_{\mathcal{M}, \tau, ij}) \quad (13)$$

In the above,  $\rho_\tau(u)$  is the check function as defined in (11) and  $\hat{y}_{\mathcal{M},\tau,ij}$  is the predicted BMI from model  $\mathcal{M}$  for quantile  $\tau$  and observation  $ij$ , where  $\mathcal{M}$  was estimated based on the 4 parts of the data not containing the individual  $i$ . The 5-fold cross-validation process was repeated several times in order to avoid artifacts that are based on a single data division. Hence, the final cross-validation criterion was calculated as the mean of the individual values:  $\bar{CV}(\mathcal{M}, \tau) = \frac{1}{s} \sum_{k=1}^s CV_k(\mathcal{M}, \tau)$ .

### 3 Results

#### 3.1 Descriptive statistics

Figure 1 gives a first impression of individual BMI patterns in the LISA data by showing a trace plot for 20 randomly chosen children. The figure shows that the body mass index of the majority increases until the age of 1 year, afterwards it decreases slightly until the age of 6 years.

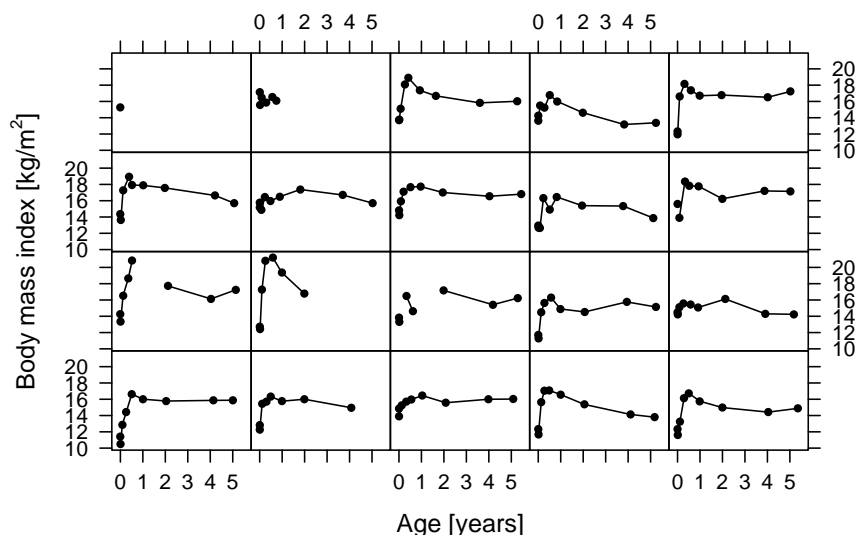


Figure 1: Individual BMI patterns by age for 20 randomly chosen children. Every dot denotes a single observation. If the dots are not line-connected, there is at least one observation missing in-between.

Figure 2 shows a scatterplot of all available BMI observations ( $n=23175$ ) by age. Here the point clusters indicate the 9 different dates of the standard examinations whose exact observation time is subject to some random variation. In addition to the single observations, Figure 2 contains visualizations of nonparametric curves for the upper BMI quantiles by age, neglecting other covariate information. These curves are estimated by a local linear quantile regression technique (Yu and Jones (1998)) and give a first impression of the relationship between BMI and age for the two selected quantiles, moreover they confirm the BMI-age relationship suggested by Figure 1. However, this view neglects the longitudinal structure of the LISA data. The single BMI outlier of  $35 \text{ kg}/\text{m}^2$  was excluded from all further analyses.

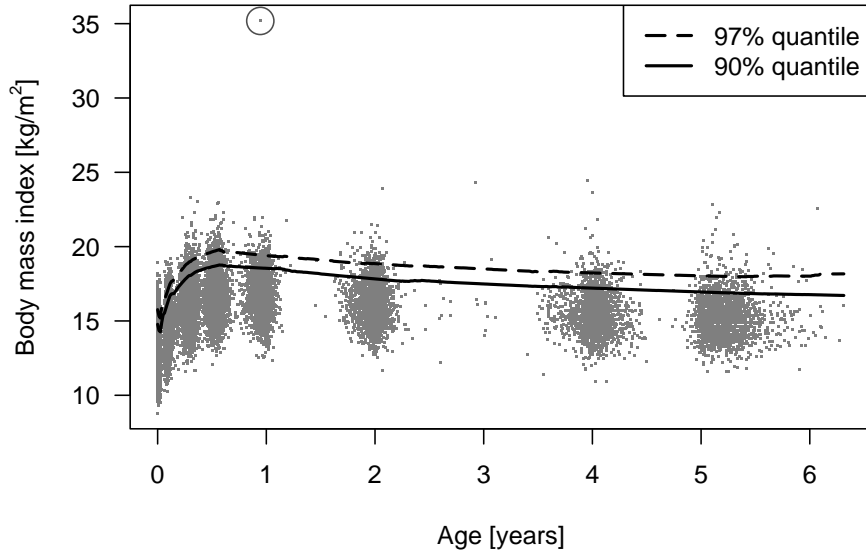


Figure 2: BMI observations by age with curves for upper quantiles estimated by local linear quantile regression,  $n = 23175$ . The single outlier is highlighted by a circle.

Figure 3 displays scale and skewness of the empirical BMI distribution by age in months for the first 12 months. Thus, the relationship between age and skewness of the empirical BMI distribution can be inspected, which is an important tool for checking the distributional assumptions in later modeling, i.e. a normal distribution conditional upon covariates for the mixed effects model. No additional covariate information is considered in these plots.

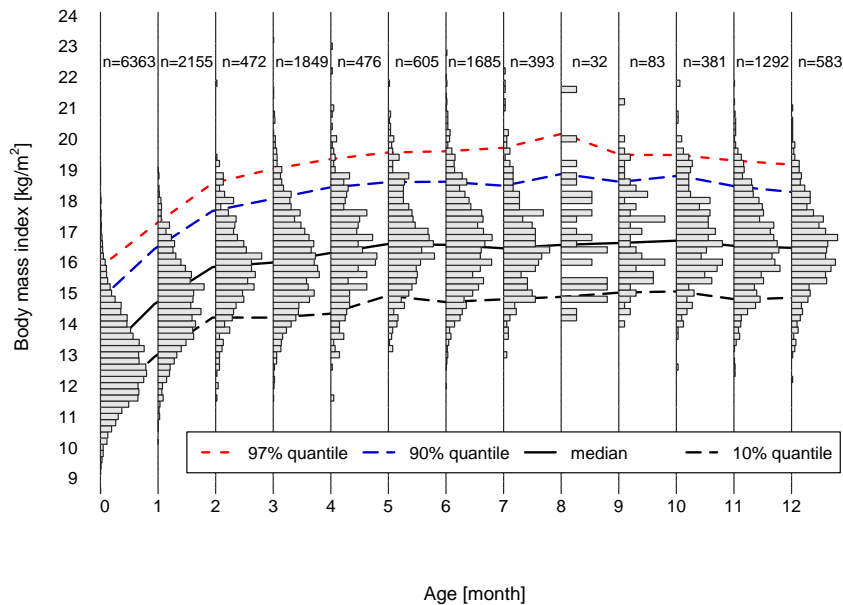


Figure 3: Monthly histograms for the BMI distribution by age in months until the age of 12 months. Also shown are lines for the empirical 10%, 50%, 90% and 97% quantiles.

The question of the present analysis was whether the BMI quantile curves in Figure 2 and the shape of the BMI distribution in Figure 3 change depending on other covariates than age, e.g. breastfeeding. Therefore, Figure 4 shows quantile curves depending on breastfeeding, again estimated by local linear quantile regression. This plot suggests an effect of breastfeeding on the upper BMI quantiles with increased age. Note that it is only the upper quantiles which appear to differ for the two groups. Such a situation, where only certain quantiles appear to be affected, requires a more specific quantile modeling than provided by AMMs. The modeling in Section 3.2 will reveal, whether this effect still exists when other covariates are included in the analysis.

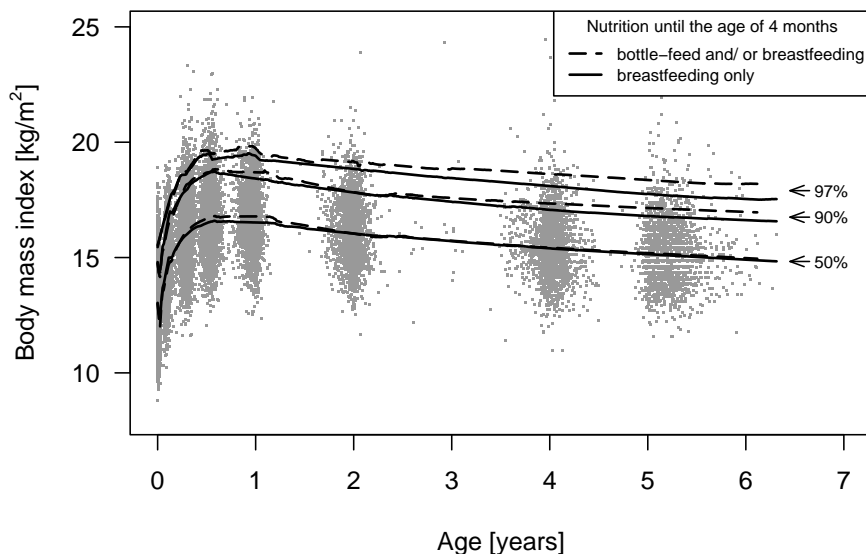


Figure 4: BMI observations by age with curves for median, 90% and 97% BMI quantiles depending on breastfeeding during the first 4 months. Curve estimation was conducted by local linear quantile regression.

### 3.2 Model estimation

All models were estimated with main effects for the entire set of covariates given in Table 1 and 2, and an additional interaction effect between age and breastfeeding. This interaction was motivated by a special interest in the effect of breastfeeding on upper BMI quantiles in obesity research. Table 3 displays the results of these model fits and compares the estimated significance levels and effect tendencies. The corresponding  $p$ -values were calculated by Wald and, in case of quantile regression, Wald-like tests.

The AMM contains only one predictor  $\eta^{(\mu)}$  while the GAMLSS with BCPE distribution is composed of four different parameter specific predictors. For the purpose of model comparison the predictor  $\eta^{(\mu)}$  contains all available covariates whereas the other predictors in GAMLSS are reduced as much as possible by means of the GAIC criterion. In quantile regression the predictors are each quantile specific, therefore we chose three different values of  $\tau$ : the median for reasons of model comparison as well as the 90% and 97% BMI quantiles since overweight and obesity are defined by these reference values by default. In AMM results, the standard deviation

Covariate	AMM $\eta^{(\mu)}$	GAMLSS				Quantile regression		
		$\eta^{(\mu)} = \mu$	$\eta^{(\sigma)} = \log(\sigma)$	$\eta^{(\nu)} = \nu$	$\eta^{(\varphi)} = \log(\varphi)$	$\eta^{(\tau=0.5)}$	$\eta^{(\tau=0.9)}$	$\eta^{(\tau=0.97)}$
Age	bs(df=10)***	bs(df=10)***	bs(df=9)***	bs(df=9)***	(-)**	bs(df=10)***	bs(df=10)***	bs(df=10)***
Age*Breastfeeding	bs(df=10)***	bs(df=10)***				bs(df=10)*	bs(df=10)*	bs(df=10)*
Weight gain 2 <sup>nd</sup> year	(+)***	(+)***	(+)***	(+)***		bs(df=4)***	(+)***	(+)***
Maternal BMI	bs(df=3)***	bs(df=3)***	bs(df=3)***	(-)**		bs(df=4)***	bs(df=3)***	bs(df=4)***
Maternal BMI gain during pregnancy	(+)***	(+)***				(+)***	(+)***	bs(df=4)***
Sex_female	(-)*	(-) <sup>n.s.</sup>	(+)***	(+)***		(-)**	(-) <sup>n.s.</sup>	(-) <sup>n.s.</sup>
Breastfeeding_1	(+)**	(+)**				(+) <sup>n.s.</sup>	(+) <sup>n.s.</sup>	(+) <sup>n.s.</sup>
TV_PC_2	(+) <sup>n.s.</sup>	(+) <sup>n.s.</sup>		(+) <sup>n.s.</sup>		(+) <sup>n.s.</sup>	(+)*	(+) <sup>n.s.</sup>
TV_PC_3	(+) <sup>n.s.</sup>	(+)*		(-)*		(+) <sup>n.s.</sup>	(+)***	(+)***
Hours spent outdoor	(+)*	(+)***				(+)**	(+)*	(+) <sup>n.s.</sup>
Maternal smoke_1	(-) <sup>n.s.</sup>	(-) <sup>n.s.</sup>	(+)*			(-) <sup>n.s.</sup>	(+) <sup>n.s.</sup>	(+) <sup>n.s.</sup>
Maternal edu_1	(+) <sup>n.s.</sup>	(-) <sup>n.s.</sup>	(+) <sup>n.s.</sup>			(+) <sup>n.s.</sup>	(+)***	(-) <sup>n.s.</sup>
Maternal edu_2	(+)**	(+)***	(-) <sup>n.s.</sup>			(+)***	(+)*	(+)***
Maternal edu_3	(+)***	(+)***	(+)**			(+)***	(+)***	(+)***
Maternal edu_4	(+)***	(+)***	(+) <sup>n.s.</sup>			(+)***	(+)***	(+)***
Area_urban	(+)*	(+)***				(+)*	(+)**	(+) <sup>n.s.</sup>

Table 3: Comparison between model fits: reported are significance levels and covariate effect tendencies. For the discrete covariates all effects (except for maternal education) and significances refer to the lowest category, the effects and their significances of maternal education refer to the highest category, as given in Table 2.

Legend:

(+) Response parameter increases with increasing covariate value

(-) Response parameter decreases with increasing covariate value

bs(df=x) Nonlinear effect modeled as B-spline with x degrees of freedom

\* Effect significance on level  $p < 0.05$

\*\* Effect significance on level  $p < 0.01$

\*\*\* Effect significance on level  $p < 0.001$

<sup>n.s.</sup> Effect not significant, i.e.  $p \geq 0.05$

of the random intercept was estimated as 0.705, so that the individual specific intercepts vary between 7.21 and 11.06  $kg/m^2$ .

How can these estimated effects of AMM and GAMLSS be interpreted with respect to the quantiles? Since the conditional and marginal quantile function in AMM can be obtained by a simple time-constant shift of the predictor (see (3)), the estimated parameters for AMM apply not only for the conditional expectation but also for the quantiles in the selected model. As far as GAMLSS are concerned, the translation of the effects to the quantiles is not straightforward, but from (6) it follows that the effect signs from the predictor  $\eta^{(\mu)}$  also hold for the quantiles.

Table 3 displays that fits from the three different models have a tendency to agree on covariate effects and significances. Highly significant covariates (i.e.  $p < 0.001$ ) for overweight and obesity in the LISA cohort are age, weight gain until second birthday, maternal BMI, maternal BMI gain during pregnancy, maternal education and the interaction term between age and breastfeeding. The covariates for sex, physical activity and prenatal smoking show in most cases no influence on the considered BMI quantiles. Moreover, there are no fundamental differences between the model fits, in particular the results of quantile regression are quite similar for the three investigated quantiles.

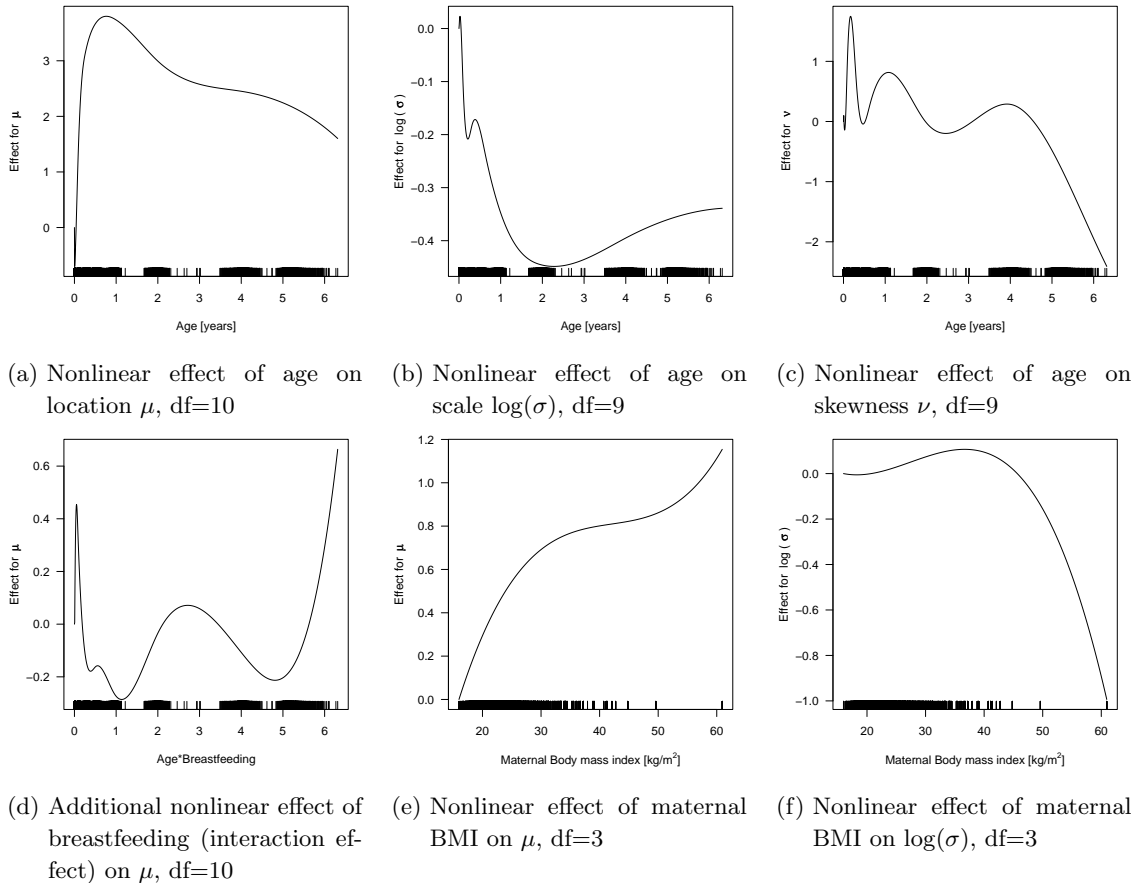


Figure 5: Estimated nonlinear covariate effects by the GAMLSS with assumed BCPE distribution

The functional shapes of the nonlinear effects are exemplified in Figure 5 by the fitted GAMLSS for age and maternal BMI. All interpretations are made qualitatively based on the functional

shapes – the absolute value of the effect is not immediately interpretable. In case of age, the shape for  $\mu$  in Figure 5(a) corresponds to the shape already seen in Figure 2. According to Figure 5(b) the nonlinear effect for  $\log(\sigma)$  indicates that there is maximal BMI variation at birth which decreases until the age of 2 years, afterwards the BMI variation increases slightly again. The functional shape of the nonlinear effect for the skewness parameter  $\nu$  in Figure 5(c) suggests that the distributional shape of children’s BMI is rather symmetric until the age of 4 and becomes slightly right skewed with increasing age. Moreover, Figure 5(d) shows the estimated interaction effect between age and breastfeeding for  $\mu$  and displays the additional nonlinear effect of breastfed children in comparison to Figure 5(a). This effect seems to be rather wiggly and its interpretation questionable.

In case of maternal BMI in Figure 5(e) the nonlinear effect on  $\mu$  is increasing until a maternal BMI of 35 after which it remains almost constant. The nonlinear effect of the maternal BMI on  $\log(\sigma)$  indicates that the children’s BMI variation decreases from a BMI value of 35 on (see Figure 5(f)). As can be seen from the rug plot on the x-axis, the nonlinearity of these effects is based on a few data points only.

For the other model fits, i.e. AMM and distribution free quantile regression, the nonlinear effects show more or less the same functional shapes as in Figure 5. As far as the other two continuous covariates are concerned, i.e. maternal BMI gain during pregnancy and children’s weight gain until the age of 2 years, they are modeled as linear in almost all models except for the quantile regression fits. Even in these models there are only a few degrees of freedom fitted so that these effects are almost linear and hence not visualized here.

Finally, an advantage of GAMLSS is that the parameters are all modeled separately and therefore the model provides an extensive overview over the distributional effects conditional upon covariates. However, a disadvantage of GAMLSS is that the interpretation of effects becomes involved in case of complicated three or four parameter distributions. Besides, confidence intervals would be useful in these interpretations, but they are not immediately available in a GAMLSS.

In order to visualize the nonlinear, time-varying effect of age on BMI quantiles after adjusting for remaining covariate effects, a covariate combination of the time constant covariates was fixed and then, for each model, individual BMI quantiles were estimated depending on age and the specific covariate combination. This combination should represent a mean risk for overweight and obesity so that the mean was chosen for all continuous covariates and the most frequent category for all discrete covariates. Figure 6 show the corresponding curves for the median and upper BMI quantiles. In case of the median, the estimated curves are almost identical for all models. However, the estimated AMM curves for the upper quantiles are located above the other curves starting at the age of 1. This may be due to the larger flexibility of GAMLSS and quantile regression compared to AMM in estimating other quantiles beyond the mean, because the AMM curves for the upper quantiles are obtained by a parallel shift of the median curve (see (3)).

Concerning the interaction effect between age and breastfeeding, Figure 7 as well as Figure 8 display the BMI quantile curves depending on breastfeeding estimated by AMM and quantile regression, respectively. In case of AMM, only slight differences between the estimated curves occur in sparse data regions. Note that this result corresponds to the one expected after

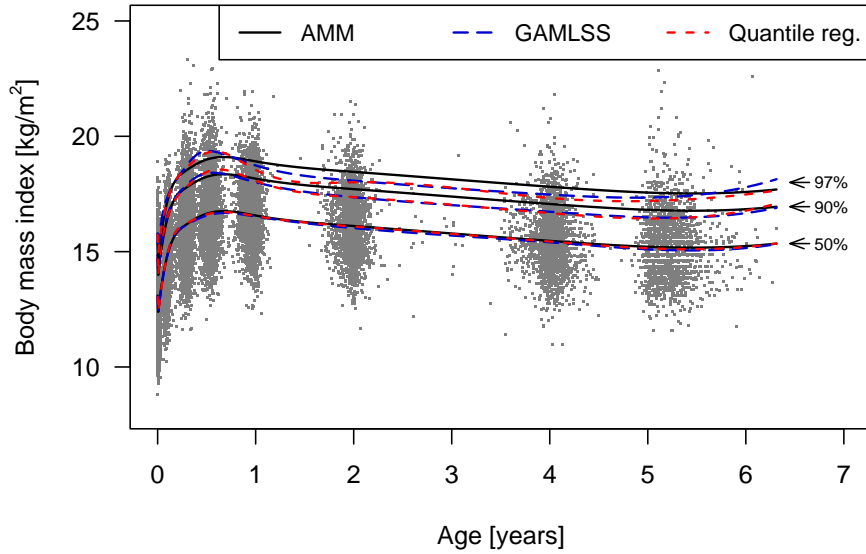


Figure 6: Model specific estimated BMI quantile curves for median, 90% and 97% BMI quantiles depending on age and a fixed covariate combination of time-constant covariates.

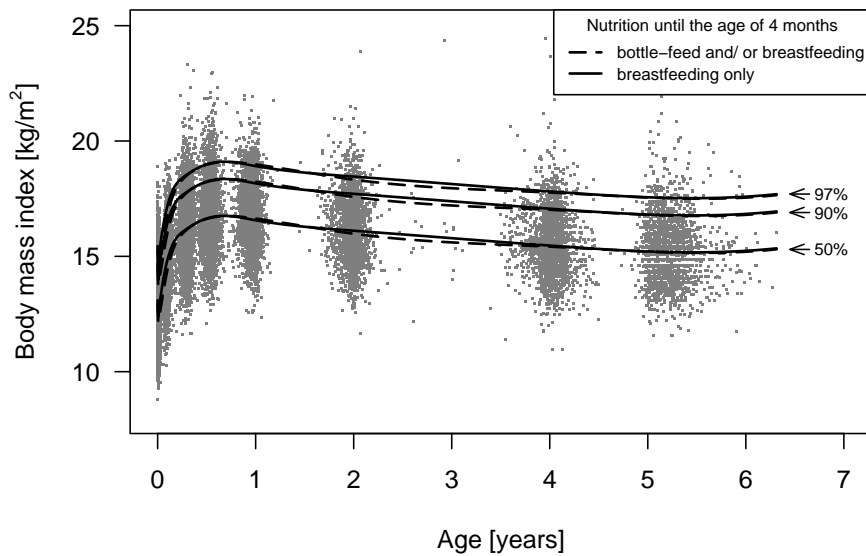


Figure 7: BMI quantile curves estimated by AMM for median, 90% and 97% BMI quantiles depending on age, breastfeeding and a fixed covariate combination of all other time-constant covariates, as stated in Table 1 and 2.

regarding Figure 4. For quantile regression the differences between the curves of breastfed and non-breastfed children are more obvious.

In addition to the visual comparison, the model fits were compared by 5-fold cross-validation, as described in Section 2.2. Table 4 shows the results of this cross-validation which was carried out ten times. There are hardly any differences between the mean cross-validation criteria until the third decimal place which means that for the data at hand no model can be rated as being superior.

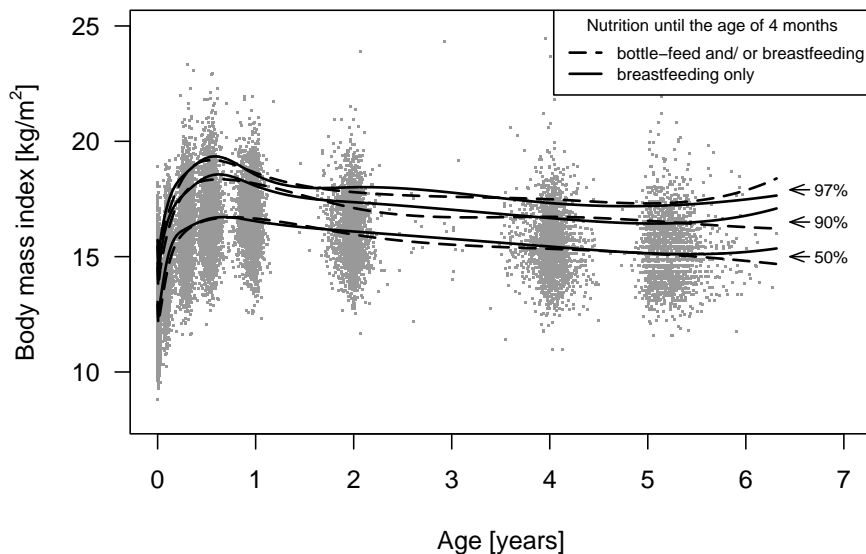


Figure 8: BMI quantile curves estimated by quantile regression for median, 90% and 97% BMI quantiles depending on age, breastfeeding and a fixed covariate combination of all other time-constant covariates, as stated in Table 1 and 2.

Quantile	AMM	GAMLSS	Quantile regression
$\tau=0.50$	0.49297 (se = $2.57 \cdot 10^{-4}$ )	0.49253 (se = $1.95 \cdot 10^{-4}$ )	0.49404 (se = $1.59 \cdot 10^{-4}$ )
$\tau=0.90$	0.23109 (se = $2.53 \cdot 10^{-4}$ )	0.22871 (se = $3.34 \cdot 10^{-4}$ )	0.22830 (se = $3.15 \cdot 10^{-4}$ )
$\tau=0.97$	0.09472 (se = $1.79 \cdot 10^{-4}$ )	0.09363 (se = $3.35 \cdot 10^{-4}$ )	0.09500 (se = $5.15 \cdot 10^{-4}$ )

Table 4: Means  $\bar{C}\bar{V}(\mathcal{M}, \tau)$  und standard errors  $se$  of the cross-validation criteria depending on model  $\mathcal{M}$  and quantile  $\tau$  after repeated 5-fold cross-validation

## 4 Discussion

Two views on the obtained results are of interest. From a methodological perspective, all considered models led to similar results so that none of them could be rated superior. Although Figure 6 suggests a slight inferiority of AMM compared to the other models in describing the upper quantiles, cross-validation did not support this for the data at hand. This result seems all the more remarkable as the current state of the art of GAMLSS and quantile regression methods does not fully account for the longitudinal data structure.

Concerning the epidemiological perspective, our results mainly confirm the risk factors that were found in previous studies, such as maternal overweight, rapid growth and maternal weight gain during pregnancy. However, there are some variations compared to results in Agras and Mascola (2005) or Reilly et al. (2005) that have to be interpreted and discussed with care.

After adjusting for the entire set of risk factors given in Tables 1 and 2, such as maternal BMI or maternal education, our quantile modeling of the available data did not show evidence for a

negative effect of breastfeeding on overweight or obesity in early childhood.

The covariates that stand for physical activity did not have a significant effect on BMI. In case of the covariate on TV and computer consumption, this might be due to an insufficient number of categories in the questionnaire that do not allow enough differentiation between children. The covariate on hours spent outdoor was considered as an auxiliary variable expressing children's physical activity. Apparently its lack of significance might be more due to its lack of representation for physical activity than this being a statement for physical activity.

The covariate on maternal education showed a negative effect on the BMI which means that with a higher level of maternal education the BMI quantiles of children decrease. However, no significant difference between the lowest and the highest categories could be detected. An explanation for this nonsignificance could be the sparsely frequented lowest category. In case of the covariate area, its effect was estimated as significant and positive which means that BMI quantiles for children that grow up in urban areas are higher than for children that grow up in rural areas. In contrary, the covariate for maternal smoking during pregnancy was not rated as significant from neither of the models.

Regarding the relationship between age and BMI, our results for the LISA study give some indication, but does not provide clear evidence, for previous hypotheses that the shape, in particular variability and skewness, of the BMI distribution changes significantly with child's age. For the analysis based on GAMLSS there is some indication about such effects, reported in Table 3 and visible in Figure 5(b) and Figure 5(c). On the other hand, looking at Figure 6, there is only a minor indication for such a change starting at the age of six. We can speculate that age-specific skewness and variability may become significant for children older than six years.

For this reason, we consider regression methods such as GAMLSS and distribution-free quantile regression as useful approaches for exploring such changes in the shape of a response variable over time, in particular in (future) studies with older children, teenagers and young adults in the sample. However, there is further methodological research needed to extend these models to fully take into account specific issues arising in longitudinal data structures: First, incorporation of individual-specific random effects or trajectories as well as nonparametric modeling of linear, time-varying effects in a similar fashion as in Gaussian semiparametric mixed models, and second, refined modeling of residual temporal correlation. We plan such extensions in future work based on penalized quasi-likelihood and semiparametric Bayesian concepts.

## 5 Acknowledgements

We are grateful to Dr. Joachim Heinrich and Prof. Dr. Dr. Heinz-Erich Wichmann from the Institute of Epidemiology, HelmholtzZentrum München (German Research Center for Environmental Health) for providing the data, in this connection we also thank the LISA-plus study group (2008) for their work. The LISA-plus study was funded by grants of the German Federal Ministry for Education, Science, Research and Technology (Grant No. 01 EG 9705/2 and 01 EG 9732) and the 6-years follow-up of the LISA-plus study was funded by the German Federal Ministry of Environment (IUF, FKS 20462296). Furthermore, we thank for financial support from the LMUinnovativ project "Munich Center of Health Sciences".

## References

- Agras, W. and A. Mascola (2005). Risk factors for childhood overweight. *Current Opinion in Pediatrics* 17, 648–652.
- Borghini, E., M. de Onis, C. Garza, J. van den Broeck, E. Frongillo, L. Grummer-Strawn, S. van Buuren, H. Pan, L. Molinari, R. Martorell, A. Onyango, and J. Martines (2006). Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Statistics in Medicine* 25, 247–265. For the WHO Multicentre Growth Reference Study Group.
- Cole, T., M. Bellizzi, K. Flegal, and W. Dietz (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal* 320, 1240–1245.
- Eilers, P. and B. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2), 89–121.
- Fahrmeir, L. and T. Kneib (2008). On the Identification of Trend and Correlation in Temporal and Spatial Regression. In Shalab and C. Heumann (Eds.), *Recent Advances in Linear Models and Related Areas*. Springer.
- Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2 ed.). Springer Series in Statistics. Springer.
- Fitzmaurice, G., N. Laird, and J. Ware (2004). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley.
- Haslam, D. and W. James (2005). Obesity. *The Lancet* 366, 1197–1209.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning – Data Mining, Inference and Predictions*. Springer Series in Statistics. Springer.
- Koenker, R. (2005). *Quantile Regression*. Economic Society Monographs. Cambridge University Press.
- Koenker, R. (2008). *quantreg: Quantile Regression*. R package version 4.17.
- Kosti, R. and D. Panagiotakos (2006). The epidemic of obesity in children and adolescents in the world. *Central European Journal of Public Health* 14(4), 151–159.
- LISA-plus study group (1998–2008). Information about the study is available at <http://www.helmholtz-muenchen.de/epi/arbeitsgruppen/umweltepidemiologie/projects-projekte/lisa-plus/index.html>.
- Lobstein, T., L. Bauer, and R. Uauy (2004). Obesity in children and young people: a crisis in public health. *Obesity Reviews* 5 (Suppl.1), 4–85.
- Lobstein, T. and M.-L. Frelut (2003). Prevalence of overweight among children in Europe. *Obesity Reviews* 4, 195–200.
- Ong, K., M. Ahmed, P. Emmett, M. Preece, and D. Dunger (2000). Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *British Medical Journal* 320, 967–971.

- Pinheiro, J., D. Bates, S. DebRoy, and D. Sarkar (2007). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-86.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Reilly, J., J. Armstrong, A. Dorosty, P. Emmett, A. Ness, I. Rogers, C. Steer, and A. Sherriff (2005). Early life risk factors for obesity in childhood: cohort study. *British Medical Journal* 330, 1357–1363.
- Rigby, R. and D. Stasinopoulos (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine* 23, 3053–3076.
- Rigby, R. and D. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics* 54(3), 507–554.
- Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Stasinopoulos, M., B. Rigby, and contributions from C. Akantziliotou (2008). *gamlss: Generalized Additive Models for Location Scale and Shape*. R package version 1.8-0.
- Terry, M., Y. Wei, and D. Esserman (2007). Maternal, Birth, and Early-Life Influences on Adult Body Size in Women. *American Journal of Epidemiology* 166(1), 5–13.
- Toschke, A., V. Grote, B. Koletzko, and R. von Kries (2004). Identifying children at high risk for overweight at school entry by weight gain during the first 2 years. *Archives of Pediatrics & Adolescent Medicine* 158(5), 449–452.
- Verbeke, G. and G. Molenberghs (1997). *Linear Mixed Models for Longitudinal Data* (2 ed.). Springer Series in Statistics. Springer.
- Yu, K. and M. Jones (1998). Local Linear Quantile Regression. *Journal of the American Statistical Association* 93(441), 228–237.