Additive Quantile Regression for the Analysis of Childhood Malnutrition

Thomas Kneib

Department of Mathematics
Carl von Ossietzky University Oldenburg

Nora Fenske & Torsten Hothorn

Department of Statistics
Ludwig-Maximilians-University Munich

CARL VON OSSIETZKY universität OLDENBURG
FAKULTÄT V
MATHEMATIK UND NATURWISSENSCHAFTEN

Göttingen, 2.7.2009

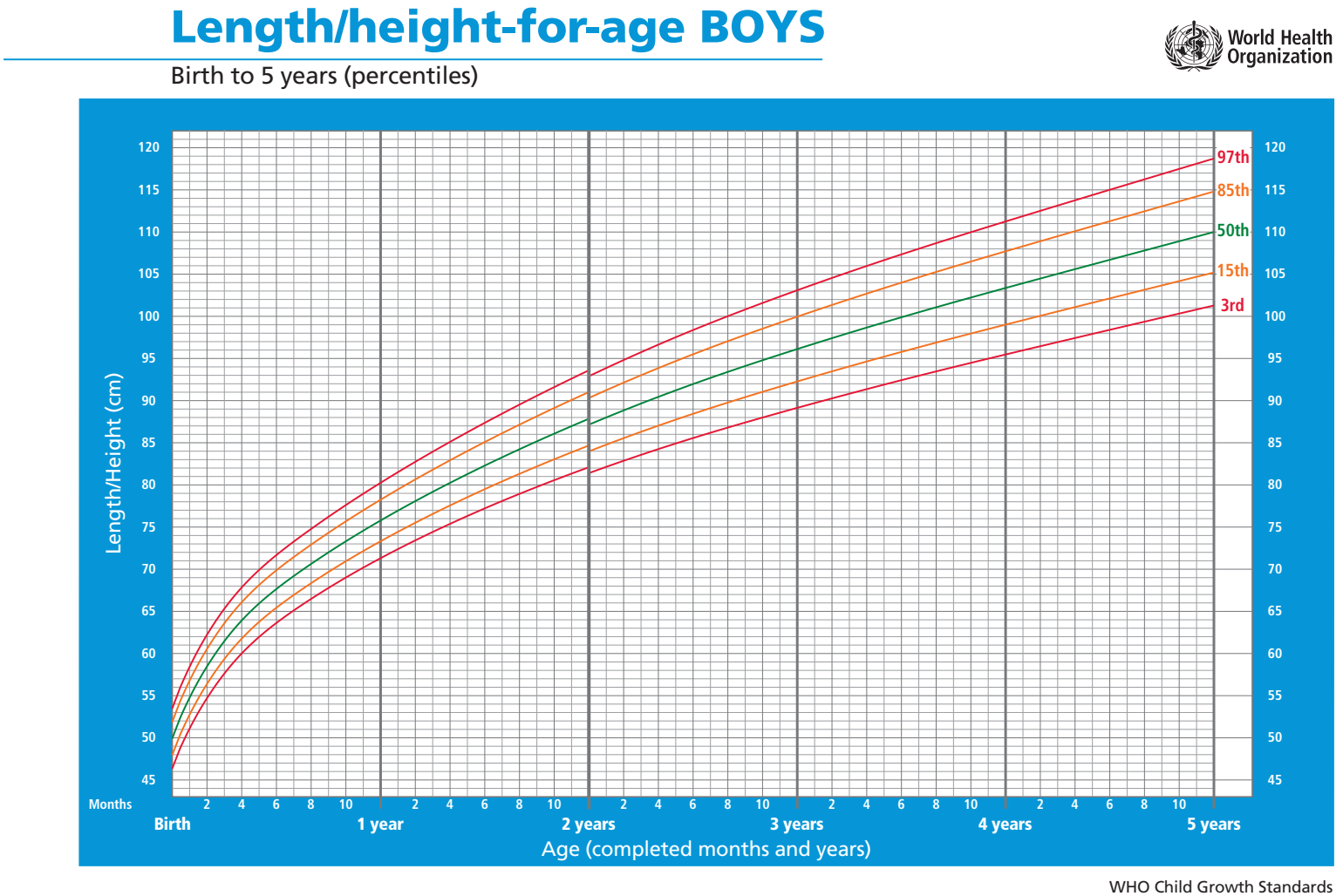
Data and Scientific Question

- Childhood malnutrition is commonly assessed by Z-scores formed from an appropriate anthropometric indicator AI relative to a reference population:

$$Z_i = \frac{\mathrm{AI}_i - \mu}{\sigma}$$

  where μ and σ refer to median and standard deviation in the reference population.

- Chronic undernutrition (stunting) is measured by insufficient height for age.
- Children are classified as stunted based on lower quantiles from reference charts such as the WHO Child Growth Standards.



Source: `http://www.who.int/childgrowth/standards`

- We used data from the 2005/06 India Demographic and Health Survey (`http://www.measuredhs.com`).
- Nationally representative cross-sectional study on fertility, family planning, maternal and child health, as well as child survival, HIV/AIDS, and nutrition.
- Information on 37.632 children is available (after excluding observations with missing information).
- Possible determinants of childhood malnutrition:

Child-specific factors:	age, gender, duration of breastfeeding, . . .
Maternal factors:	age, body mass index, years of education, employment status, . . .
Household factors:	place of residence, electricity, radio, tv, . . .

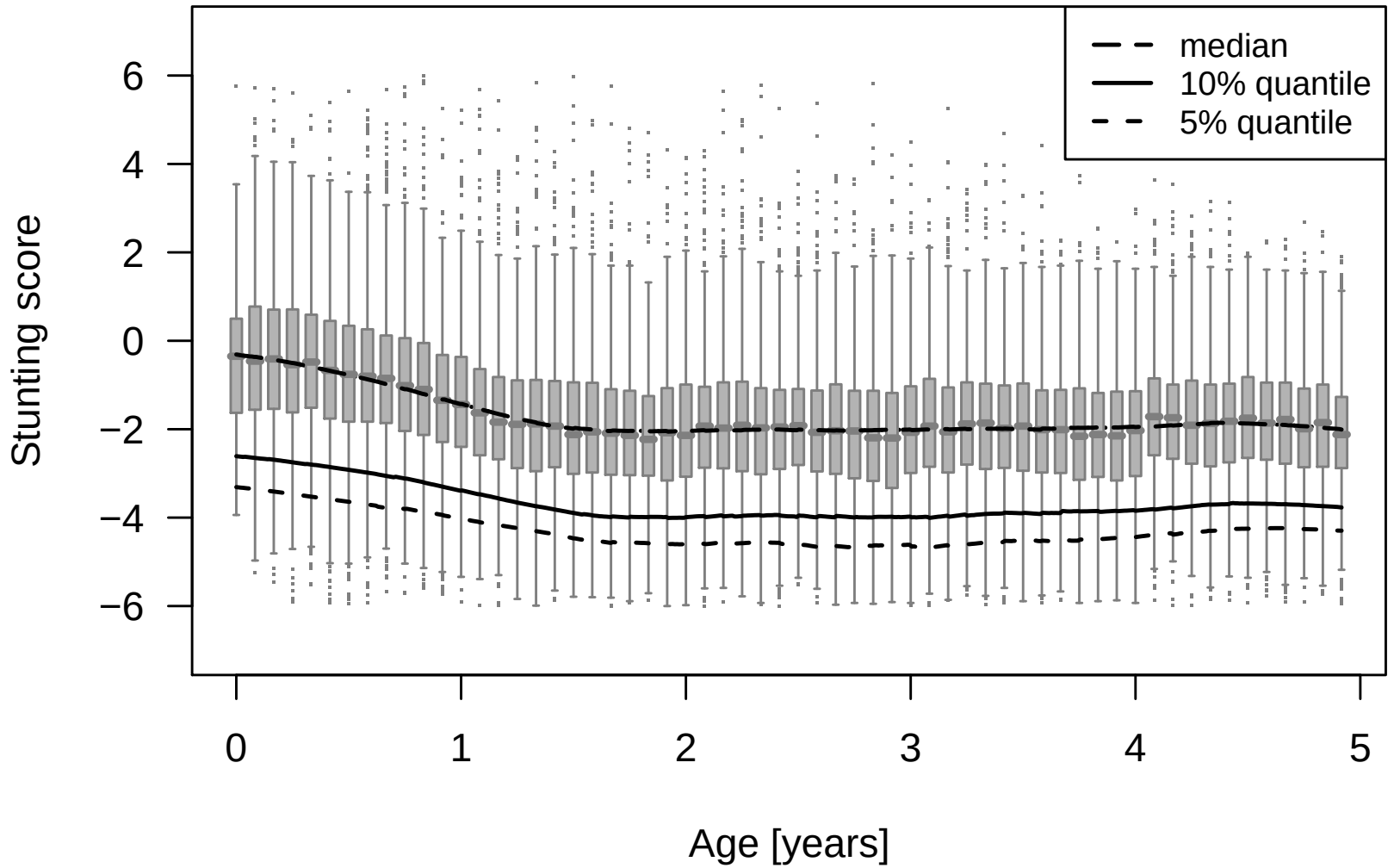
(21 risk factors in total).

- Regression models aim at quantifying the impact of covariates on undernutrition where the Z-score forms the response.
- Most common approach: Direct regression of the Z-score on covariates

$$Z = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2).$$

- Difficulties:
  - All effects are assumed to be linear while effects of continuous covariates may be suspected to be nonlinear.
  - The direct regression model explains the expectation of Z, i.e. it focusses on the average nutritional status.
  - Restrictive assumptions on the error terms ε.

  $\Rightarrow$ Additive quantile regression models.


median
10% quantile
5% quantile
Stunting score
6
4
2
0
−2
−4
−6
0
1
2
3
4
5
Age [years]

Additive Quantile Regression Models

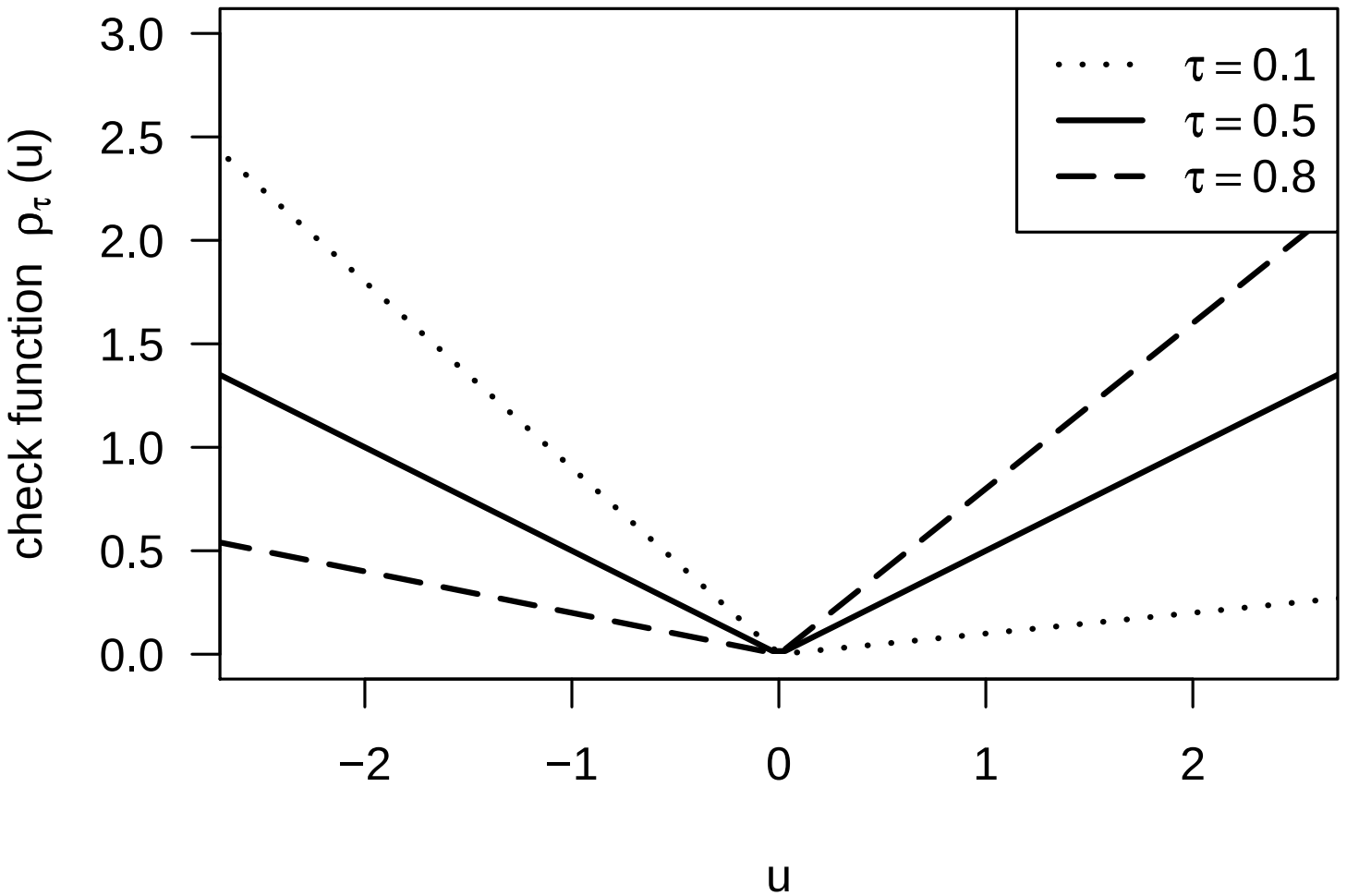
- Quantile regression aims at describing conditional quantiles in terms of covariates instead of the mean.
- We will consider 1%, 5% and 50% quantiles corresponding to severe and average malnutrition.
- Formulation in terms of a loss function:

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Z_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau)$$

where

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

and τ denotes the quantile of interest.


3.0
2.5
2.0
1.5
1.0
0.5
0.0
check function $\rho_\tau(u)$
−2
−1
0
1
2
u
$\tau = 0.1$
$\tau = 0.5$
$\tau = 0.8$

- Equivalent formulation as a regression problem:

$$Z_i = \boldsymbol{x}_i' \boldsymbol{\beta}_\tau + \varepsilon_{\tau i}$$

where $\varepsilon_{\tau i}$ are independent error terms subject to

$$F_{\varepsilon_{\tau i}}(\tau) = 0$$

with the cumulative distribution function $F_{\varepsilon_{\tau i}}(\tau)$.

- Important:
  - No explicit distributional assumption on the errors.
  - Errors are not assumed to be identically distributed.

  $\Rightarrow$ Quantile regression allows for heteroscedasticity.

- Additive quantile regression models extend the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$ with nonlinear effects of continuous covariates:

$$Z_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \sum_{j=1}^{q} f_{\tau j}(z_{ij}) + \varepsilon_{\tau i}.$$

- The functions $f_{\tau j}(z_j)$ shall be estimated nonparametrically with only qualitative assumptions about their smoothness.

Statistical Inference

- Boosting is a simple but versatile iterative stepwise gradient descent algorithm.
- General aim: Finding the solution to the minimisation problem

$$\eta^* = \underset{\eta}{\operatorname{argmin}} \, \mathbb{E}[L(Z, \eta)]$$

  where L is a loss function and η denotes the predictor of a regression model.
- In practice: Estimation of η^* by minimizing the empirical loss

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

- Minimisation is achieved by iteratively fitting simple base-learning procedures to updated residuals

$$u_i^{[m]} = -\frac{\partial}{\partial \eta} L(Z_i, \eta)\big|_{\eta=\hat{\eta}_i^{[m-1]}}, \; i = 1, \ldots, n$$

(negative gradient of the loss function).

- Componentwise boosting: Restrict gradient descent to directions induced by the covariate effects and iteratively move along the steepest descent.
- For quantile regression, the loss function is given by the check function, i.e.

$$L(Z_i, \eta_i) = \rho_\tau(Z_i - \eta_i)$$

and

$$u_i^{[m]} = -\rho_\tau'(Z_i - \hat{\eta}_i^{[m-1]}) = \begin{cases} \tau & \text{if} \quad Z_i - \hat{\eta}_i^{[m-1]} > 0 \\ 0 & \text{if} \quad Z_i - \hat{\eta}_i^{[m-1]} = 0 \\ \tau - 1 & \text{if} \quad Z_i - \hat{\eta}_i^{[m-1]} < 0. \end{cases}$$

- Least squares base-learner for parametric effects $\boldsymbol{\beta}_l$:

$$\hat{\boldsymbol{\beta}}_l^{[m]} = (\boldsymbol{X}_l'\boldsymbol{X}_l)^{-1}\boldsymbol{X}_l'\boldsymbol{u}^{[m]}.$$

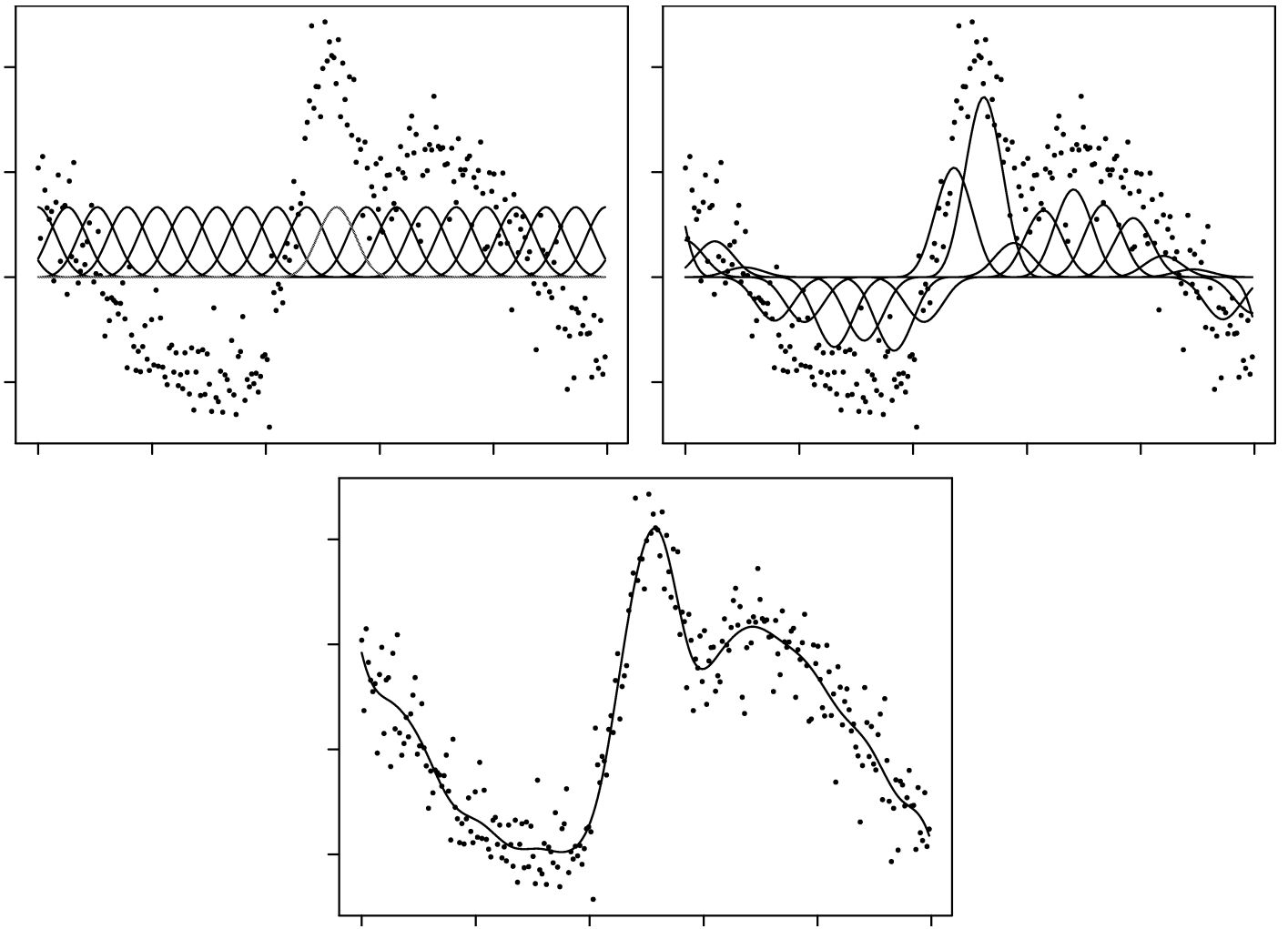
- Penalised spline base-learner for nonparametric effects $\boldsymbol{f}_j$:

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

  where

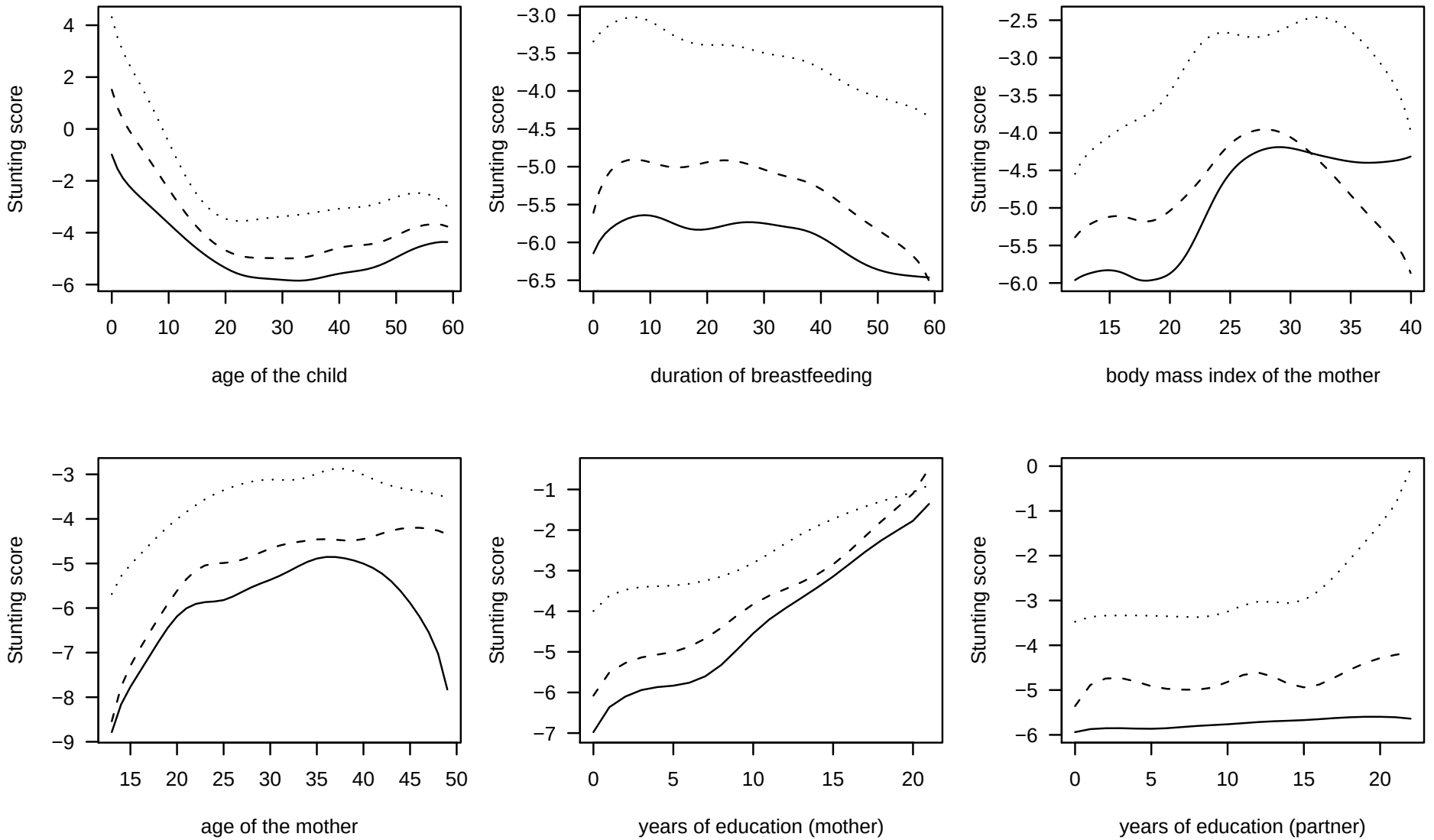
  - $\boldsymbol{Z}_j$ is a design matrix formed from a B-spline basis,
  - $\boldsymbol{K}_j$ is a difference penalty that ensures smoothness of the estimated curve, and
  - λ_j is a smoothing parameter chosen such that the base-learner has five degrees of freedom.

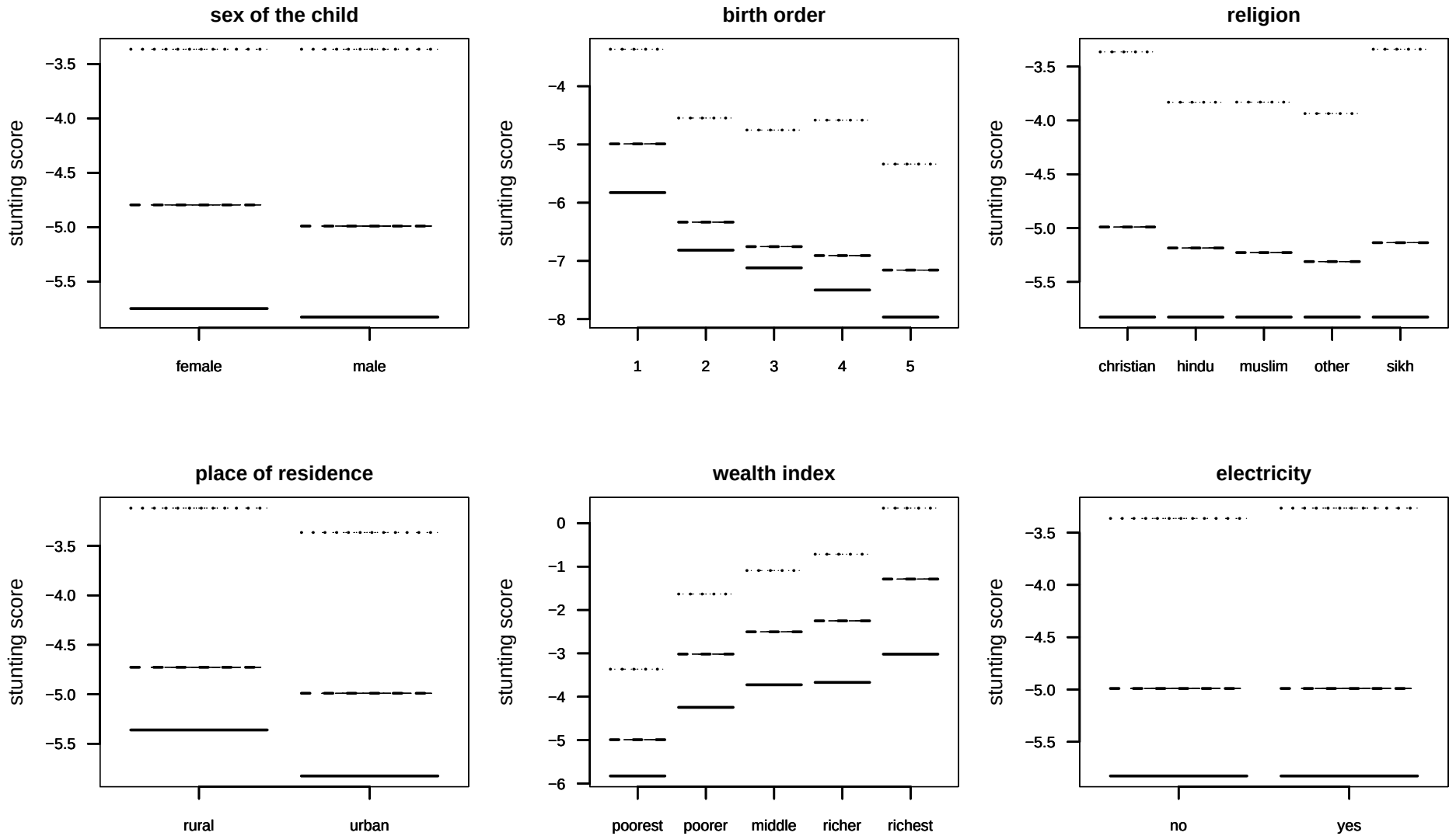
- Schematic representation of a penalised spline fit:



- Crucial tuning paramater: Number of boosting iterations m_{stop}.
- Early stopping implements variable selection property.
- Choice of m_{stop} based on cross-validation or booststrap.
- In our application, a third of the data were used as a validation sample to determine m_{stop}.

Results


Stunting score
age of the child
duration of breastfeeding
body mass index of the mother
age of the mother
years of education (mother)
years of education (partner)


sex of the child
stunting score
−3.5
−4.0
−4.5
−5.0
−5.5
female
male
birth order
stunting score
−4
−5
−6
−7
−8
1
2
3
4
5
religion
stunting score
−3.5
−4.0
−4.5
−5.0
−5.5
christian
hindu
muslim
other
sikh
place of residence
stunting score
−3.5
−4.0
−4.5
−5.0
−5.5
rural
urban
wealth index
stunting score
0
−1
−2
−3
−4
−5
−6
poorest
poorer
middle
richer
richest
electricity
stunting score
−3.5
−4.0
−4.5
−5.0
−5.5
no
yes

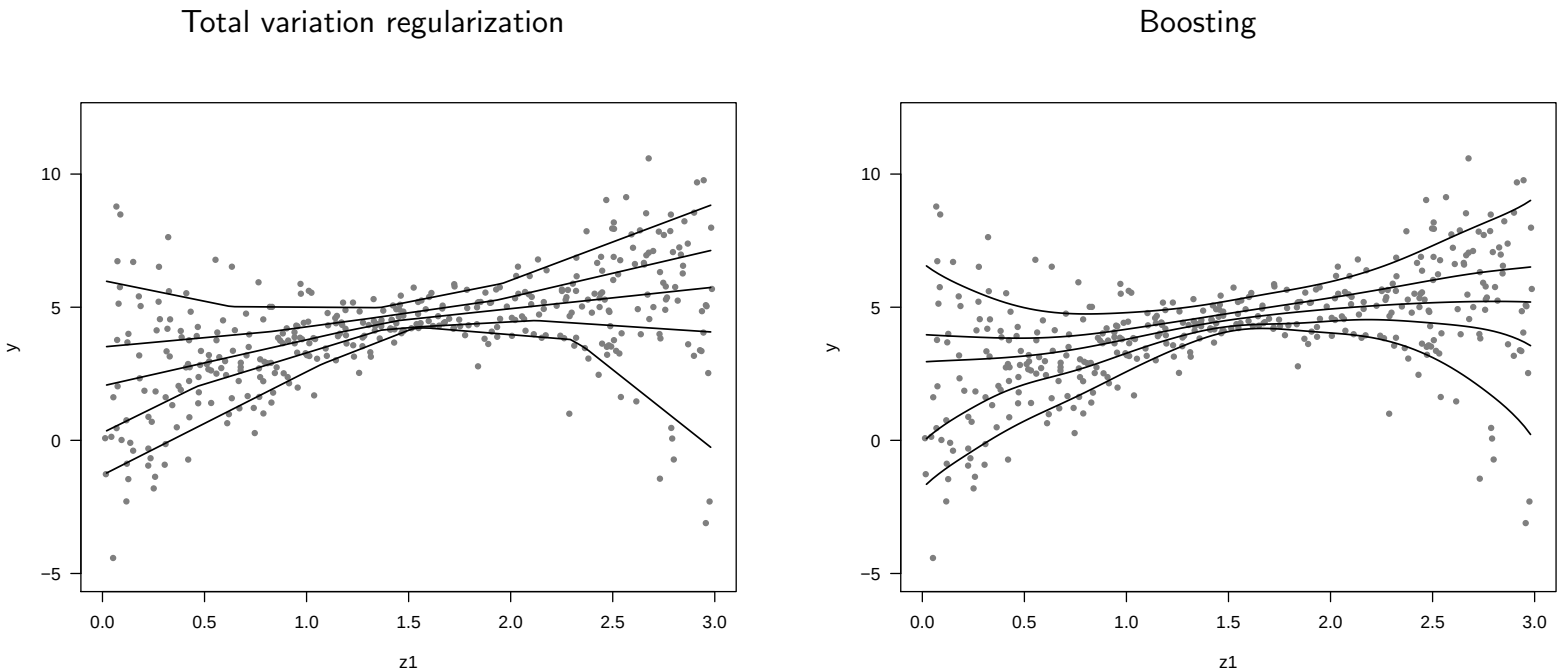
	$\tau = 0.05$		$\tau = 0.1$		$\tau = 0.5$	
Variable	FI	PI	FI	PI	FI	PI
cage	0.034	0.161	0.017	0.272	0.001	0.204
cfeed	0.273	0.084	0.125	0.069	0.020	0.174
csex	0.275	0.017	0.212	0.019	0.213	0.007
ctwin	0.328	0.035	0.128	0.025	0.063	0.012
cbord	0.061	0.092	0.057	0.071	0.032	0.046
mbmi	0.070	0.077	0.057	0.064	0.013	0.054
mage	0.106	0.161	0.082	0.092	0.035	0.175
medu	0.000	0.097	0.000	0.091	0.000	0.065
medupart	0.070	0.081	0.026	0.137	0.017	0.122
munem	0.277	0.021	0.302	0.009	0.303	0.002
mreli	0.786	0.006	0.212	0.012	0.064	0.013
resid	0.275	0.035	0.228	0.021	0.097	0.014
nodead	0.216	0.023	0.108	0.029	0.069	0.020
wealth	0.000	0.040	0.002	0.030	0.000	0.031
electricity	0.811	0.006	0.711	0.000	0.052	0.009
radio	–	–	–	–	0.236	0.003
fridge	–	–	0.948	0.000	0.036	0.011
bicycle	0.061	0.044	0.074	0.034	0.302	0.008
mcycle	–	–	0.610	0.002	0.045	0.016
car	0.255	0.019	0.100	0.018	0.047	0.013
	$m_{\text{stop}} = 107754$		$m_{\text{stop}} = 84966$		$m_{\text{stop}} = 41702$	

$$\text{FI} = \frac{\text{First iteration where variable is selected}}{m_{\text{stop}}} \qquad \text{PI} = \frac{\text{Number of iterations where variable is selected}}{m_{\text{stop}}}$$

Summary

- Quantile regression models are a valuable tool for exploring risk factors for extreme forms of stunting.
- Estimation of quantile regression by boosting offers the following advantages:
  - Flexible predictor specification including nonlinear effects of continuous covariates.
  - Implicit variable selection via early stopping.
  - Implemented in the R add-on package `mboost`, freely available from CRAN (`http://www.r-project.org`).
- Future work: Extensions of the algorithm to include random and spatial effects.

- The proposed approach has also been benchmarked against a previous suggestion based on total variation regularisation by Koenker et al. (1994) in a simulation study.



- The boosting approach is a strong competitor in terms of estimation accuracy and allows for considerably more complex model specifications.

- Reference: Fenske, N., Kneib, T. & Hothorn, T. (2009). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Department of Statistics, Technical Report 52, Ludwig-Maximilians-University Munich.
- A place called home:

`http://www.staff.uni-oldenburg.de/thomas.kneib`