

# Semiparametric Quantile and Expectile Regression

Thomas Kneib

Department of Mathematics  
Carl von Ossietzky University Oldenburg

# Childhood Malnutrition in Developing and Transition Countries

- Malnutrition and childhood malnutrition in particular are among the main public health problems in developing and transition countries.
- **Halving the proportion of malnourished people** in developing countries until 2015 is one of the United Nations Millennium goals.
- Statistical analyses can help in the development and evaluation of interventions.
- We use data from the 1998/99 **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 24.316 children is available (after excluding observations with missing information).

- **Childhood malnutrition** is assessed by a Z-score formed from an appropriate anthropometric measure  $AI$  relative to a reference population:

$$Z_i = \frac{AI_i - \mu}{\sigma}$$

where  $\mu$  and  $\sigma$  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.

- 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 covariates in total).

- In addition, we have information on the district a child lives in

⇒ **Spatial alignment of the data.**

- 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

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

- **Difficulties:**
    - All effects are assumed to be linear while effects of continuous covariates may be suspected to be nonlinear.
    - The model does not allow for spatial effects.
    - The direct regression model explains the expectation of  $Z$ , i.e. it focusses on the average nutritional status.
    - Restrictive assumptions on the error terms  $\varepsilon$ .
- ⇒ **Semiparametric quantile and expectile regression models.**

## Quantile and Expectile Regression

- Quantile regression aims at describing **conditional quantiles** in terms of covariates instead of the mean.
- **Parametric quantile regression** for quantile  $\tau \in [0, 1]$ :

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

with independent errors

$$\varepsilon_{\tau i} \sim F_{\varepsilon i} \quad F_{\varepsilon i}(0) = \tau.$$

- The condition  $F_{\varepsilon i}$  ensures that the covariates act on the conditional quantiles of the response:

$$F_{y_i}(\mathbf{x}'_i \boldsymbol{\beta}) = F_{\varepsilon i}(0) = \tau \quad \Rightarrow \quad Q_{y_i}(\tau) = F_{y_i}^{-1}(\tau) = \mathbf{x}'_i \boldsymbol{\beta}_\tau.$$

- Properties of parametric quantile regression:
  - **No explicit distributional assumption** for the error terms.
  - In particular, the errors are **not identically distributed**.
  - Semiparametric approach including the possibility of **variance heteroscedasticity**.
- Estimation of quantile-specific parameters is based on minimising the loss function

$$\hat{\beta}_{\tau} = \operatorname{argmin}_{\beta_{\tau}} \sum_{i=1}^n w_i(\tau) |y_i - \mathbf{x}'_i \beta_{\tau}|$$

with weights

$$w_i(\tau) = \begin{cases} \tau & y_i - \mathbf{x}'_i \beta_{\tau} \geq 0 \\ (1 - \tau) & y_i - \mathbf{x}'_i \beta_{\tau} < 0 \end{cases}$$

(**asymmetrically weighted absolute residuals**).

- Empirical quantiles of an i.i.d. sample  $y_1, \dots, y_n$  can be characterised as

$$q_\tau = \operatorname{argmin}_q \sum_{i=1}^n w_i(\tau) |y_i - q|.$$

- In particular, the median is defined by

$$q_{0.5} = \operatorname{argmin}_q \sum_{i=1}^n |y_i - q|.$$

- Correspondingly, the arithmetic mean is given by

$$\bar{y} = \operatorname{argmin}_e \sum_{i=1}^n (y_i - e)^2.$$

- Asymmetrically weighted squared residuals yield **empirical expectiles**:

$$e_\tau = \operatorname{argmin}_e \sum_{i=1}^n w_i(\tau)(y_i - e)^2.$$

- Expectile-specific regression coefficients can be obtained via **asymmetrically weighted least squares estimation**:

$$\hat{\beta}_\tau = \operatorname{argmin}_{\beta_\tau} \sum_{i=1}^n w_i(\tau)(y_i - \mathbf{x}'_i \beta_\tau)^2.$$

- Theoretical expectiles (of a continuous distribution) are the solutions of

$$(1 - \tau) \int_{-\infty}^e |y - e| f(y) dy = \tau \int_e^{\infty} |y - e| f(y) dy.$$

# Semiparametric Regression

- **Semiparametric regression models** replace the parametric predictor

$$\eta_{\tau i} = \beta_0 + \beta_1 x_{i1} + \dots + x_{ip} \beta_p = \mathbf{x}'_i \boldsymbol{\beta}$$

with

$$\eta_{\tau i} = \beta_0 + f_1(\mathbf{z}_i) + \dots + f_p(\mathbf{z}_i)$$

where  $f_1, \dots, f_p$  are functions **of different type** depending on **generic covariates**  $\mathbf{z}$ .

- Examples:
  - Linear effects:  $f_j(\mathbf{z}) = \mathbf{x}'\boldsymbol{\beta}$ .
  - Nonlinear, smooth effects of continuous covariates:  $f_j(\mathbf{z}) = f(x)$ .
  - Varying coefficients:  $f_j(\mathbf{z}) = u f(x)$ .
  - Interaction surfaces:  $f_j(\mathbf{z}) = f(x_1, x_2)$ .
  - Spatial effects:  $f_j(\mathbf{z}) = f_{\text{spat}}(s)$ .
  - Random effects:  $f_j(\mathbf{z}) = b_c$  with cluster index  $c$ .

- Generic model description based on
  - a **design matrix**  $\mathbf{Z}_j$ , such that the vector of function evaluations  $\mathbf{f}_j = (f_j(\mathbf{z}_1), \dots, f_j(\mathbf{z}_n))'$  can be written as

$$\mathbf{f}_j = \mathbf{Z}_j \boldsymbol{\gamma}_j.$$

- a quadratic **penalty term**

$$\text{pen}(f_j) = \text{pen}(\boldsymbol{\gamma}_j) = \boldsymbol{\gamma}_j' \mathbf{K}_j \boldsymbol{\gamma}_j$$

which operationalises smoothness properties of  $f_j$ .

- From a Bayesian perspective, the penalty term corresponds to a **multivariate Gaussian prior**

$$p(\boldsymbol{\gamma}_j) \propto \exp \left( -\frac{1}{2\delta_j^2} \boldsymbol{\gamma}_j' \mathbf{K}_j \boldsymbol{\gamma}_j \right).$$

- Estimation then relies on a penalised fit criterion, e.g.

$$\sum_{i=1}^n w_i(\tau) |y_i - \eta_{\tau i}| + \sum_{j=1}^p \lambda_j \boldsymbol{\gamma}_j' \mathbf{K}_j \boldsymbol{\gamma}_j$$

with smoothing parameters  $\lambda_j \geq 0$ .

- Example 1. Penalised splines for nonlinear effects  $f(x)$ :
  - Approximate  $f(x)$  in terms of a linear combination of **B-spline basis functions**

$$f(x) = \sum_k \gamma_k B_k(x).$$

- Large variability in the estimates corresponds to large **differences in adjacent coefficients** yielding the penalty term

$$\text{pen}(\gamma) = \sum_k (\Delta_d \gamma_k)^2 = \gamma' \mathbf{D}'_d \mathbf{D}_d \gamma$$

with difference operator  $\Delta_d$  and difference matrix  $\mathbf{D}_d$  of order  $d$ .

- The corresponding Bayesian prior is a **random walk** of order  $d$ , e.g.

$$\gamma_k = \gamma_{k-1} + u_k, \quad \gamma_k = 2\gamma_{k-1} + \gamma_{k-2} + u_k$$

with  $u_k$  i. i. d.  $\text{N}(0, \delta^2)$ .

- Example 2. Markov random fields for the estimation of spatial effects based on regional data:
  - Estimate a **separate regression coefficient**  $\gamma_s$  for each region, i.e.  $\mathbf{f} = \mathbf{Z}\boldsymbol{\gamma}$  with

$$\mathbf{Z}[i, s] = \begin{cases} 1 & \text{observation } i \text{ belongs to region } s \\ 0 & \text{otherwise} \end{cases}$$

- Penalty term based on **differences of neighboring regions**:

$$\text{pen}(\boldsymbol{\gamma}) = \sum_s \sum_{r \in N(s)} (\gamma_s - \gamma_r)^2 = \boldsymbol{\gamma}' \mathbf{K} \boldsymbol{\gamma}$$

where  $N(s)$  is the set of neighbors of region  $s$  and  $\mathbf{K}$  is an **adjacency matrix**.

- An equivalent Bayesian prior structure is obtained based on **Gaussian Markov random fields**.

# Markov Chain Monte Carlo Simulations

- Quantile regression models

$$y_i = \eta_{\tau i} + \varepsilon_{\tau i}$$

can be embedded in a Bayesian framework based on a suitable **distributional assumption for the error terms**.

- Assume that  $\varepsilon_{\tau i} \sim \text{ALD}(0, \sigma^2, \tau)$  (**asymmetric Laplace distribution**), with density

$$p_{\varepsilon i}(\varepsilon) = \frac{\tau(1-\tau)}{\sigma^2} \exp\left(-w(\tau) \frac{|\varepsilon|}{\sigma^2}\right).$$

- For the responses, this yields  $y_i \sim \text{ALD}(\eta_{\tau i}, \sigma^2, \tau)$  with

$$p_{y i}(y) = \frac{\tau(1-\tau)}{\sigma^2} \exp\left(-w(\tau) \frac{|y - \eta_{\tau i}|}{\sigma^2}\right).$$

- The resulting likelihood is

$$p(\mathbf{y}|\boldsymbol{\eta}_\tau) \propto \exp\left(-\sum_{i=1}^n w_i(\tau) \frac{|y_i - \eta_{\tau i}|}{\sigma^2}\right).$$

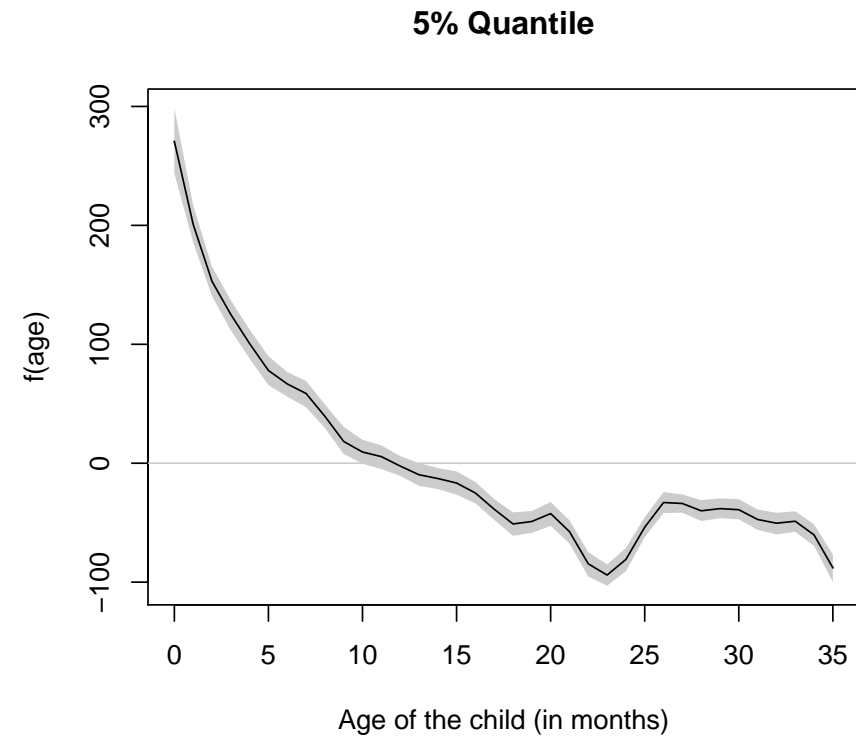
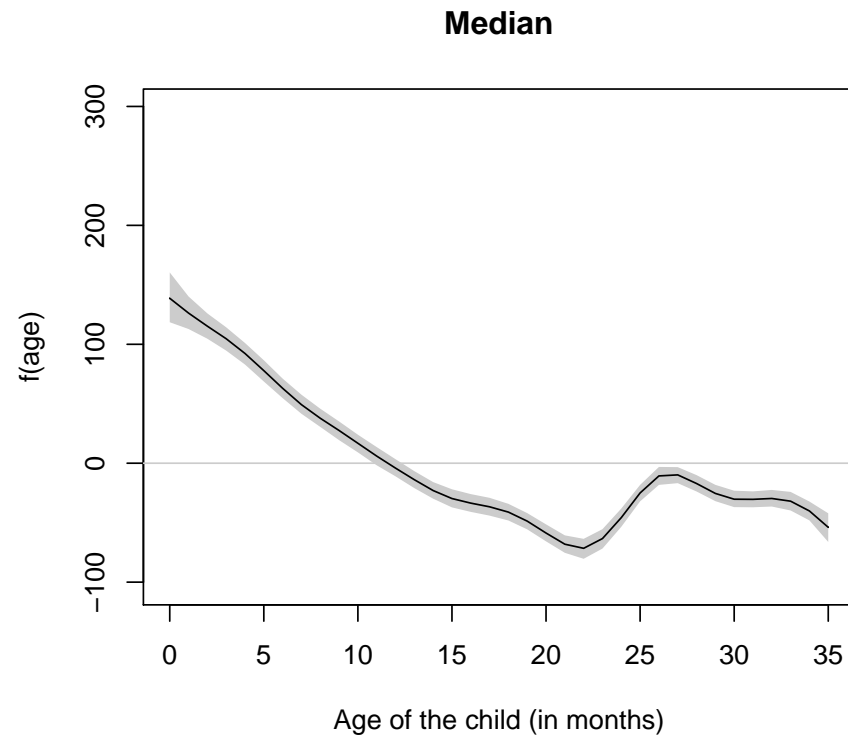
- Therefore the **posterior mode is equivalent to the penalised asymmetrically weighted absolute error estimate.**

- More precisely: The resulting point estimates coincide but the statistical estimates have different properties.
- In particular, Bayesian quantile regression additionally assumes that
  - the errors are **identically** distributed.
  - the errors follow an **asymmetric Laplace distribution**.
- Consequences:
  - The model is **no longer semiparametric** (with respect to the error distribution).
  - The posterior is usually **misspecified**, such that measures of uncertainty should be interpreted with care.

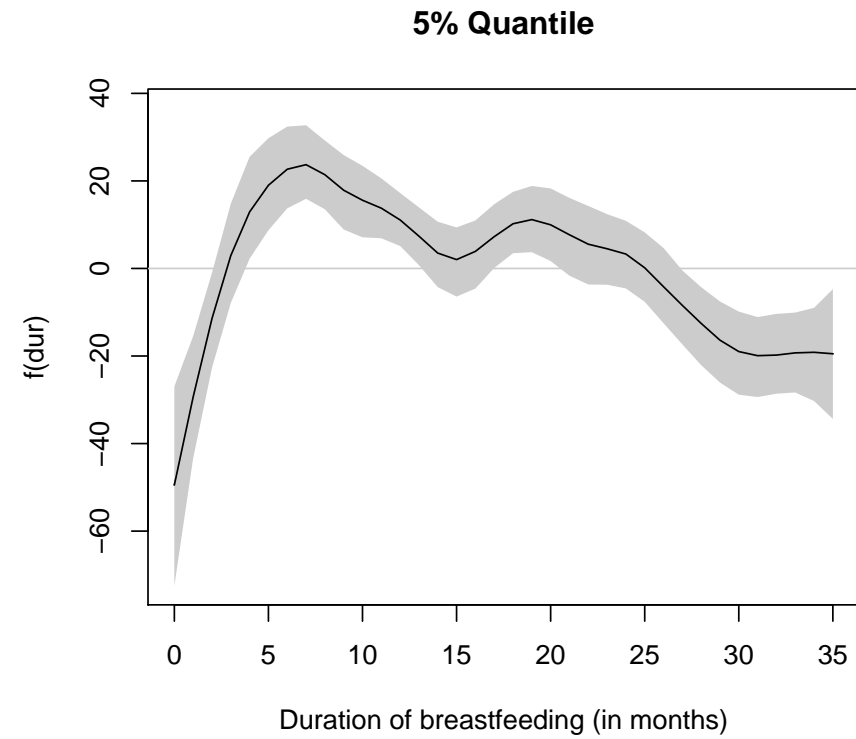
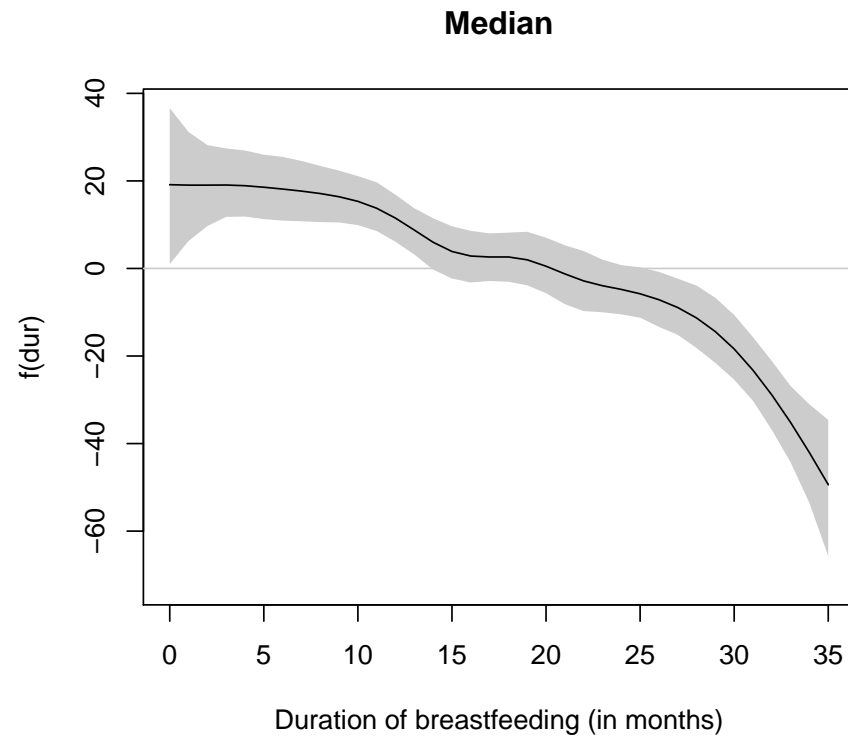
- However, the **posterior mean is easily obtained** based on Markov chain Monte Carlo simulation techniques (even for very complex predictor structures)
- A Gibbs sampler can be constructed based on a location scale mixture of normals representation of the asymmetric Laplace distribution.
- Results in the imputation of additional unknowns but yields simple Gibbs updates.

- Selected results for the malnutrition example:
  - Estimates for the 5% quantile and the median.
  - Note: Effects are centered and therefore the natural ordering of the 5% quantile and the median is not visible.

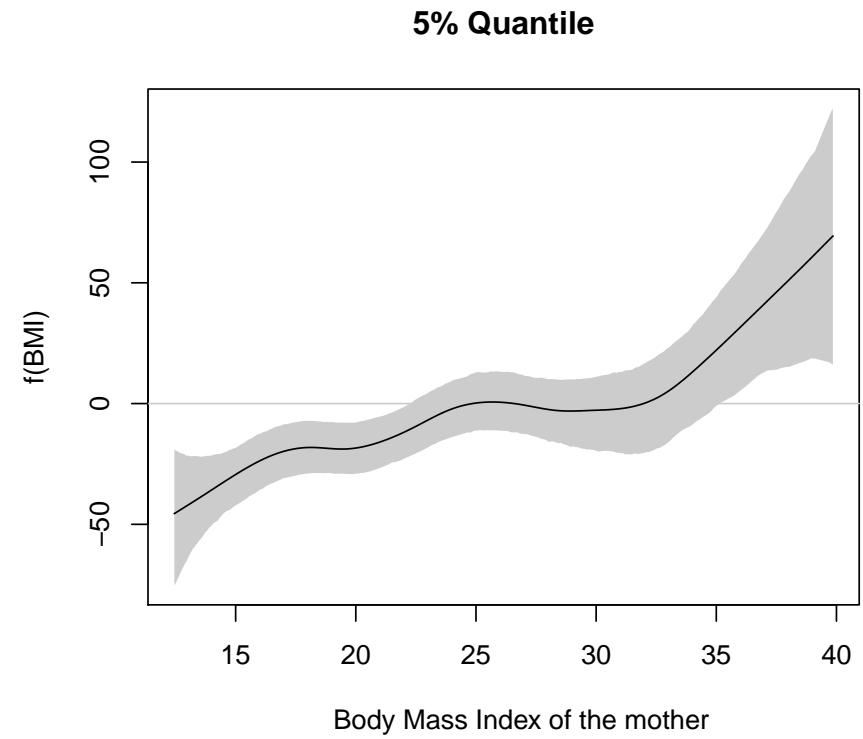
- Age of the child:



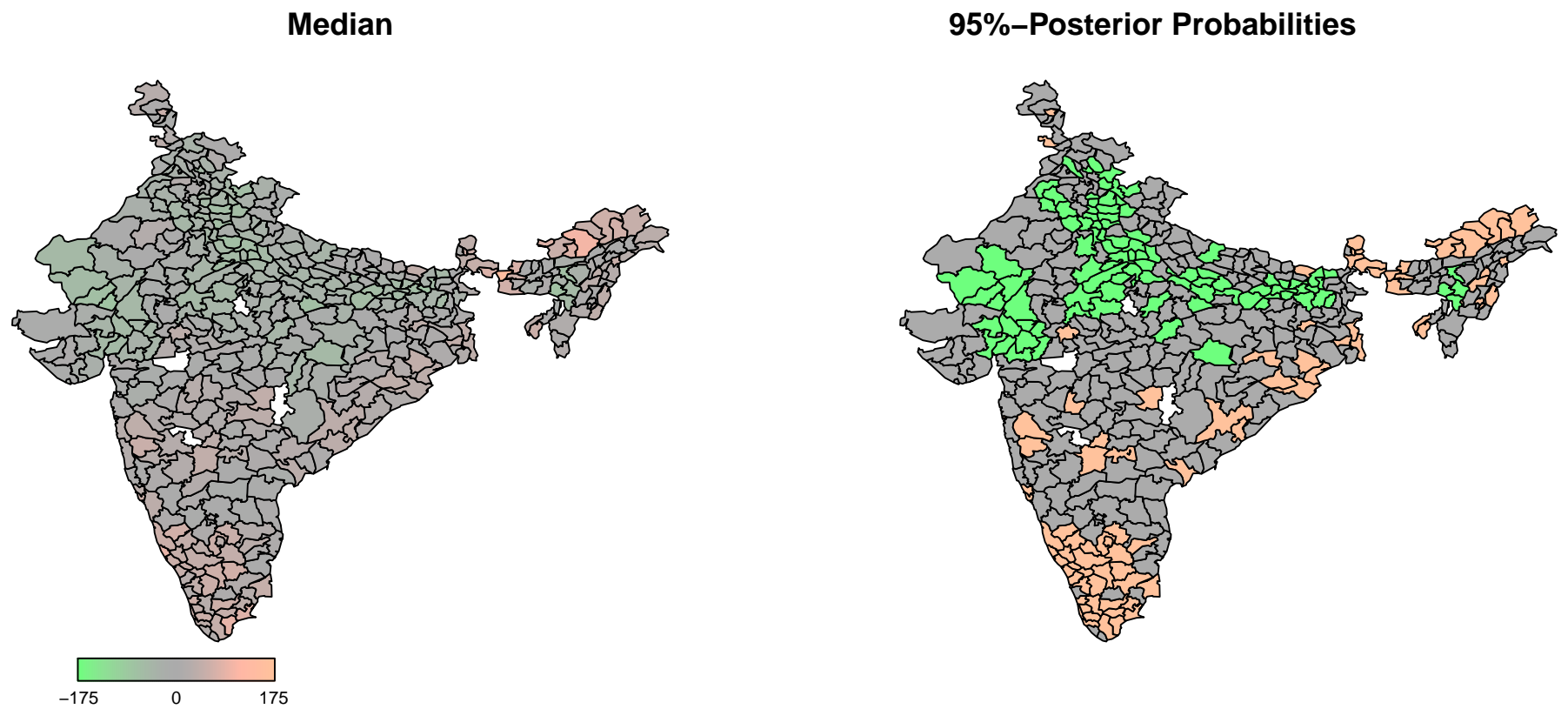
- Duration of breastfeeding:



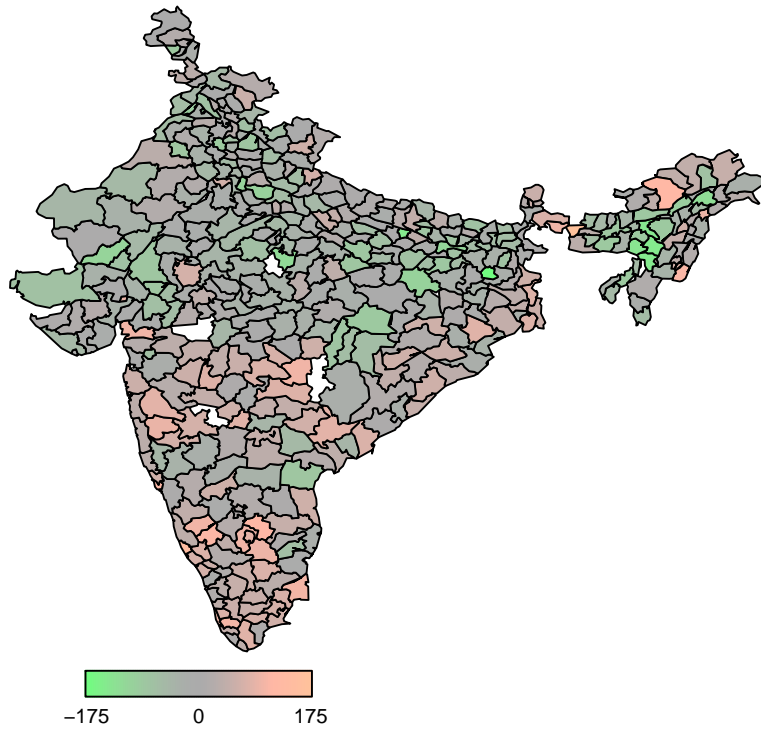
- Body mass index of the mother:



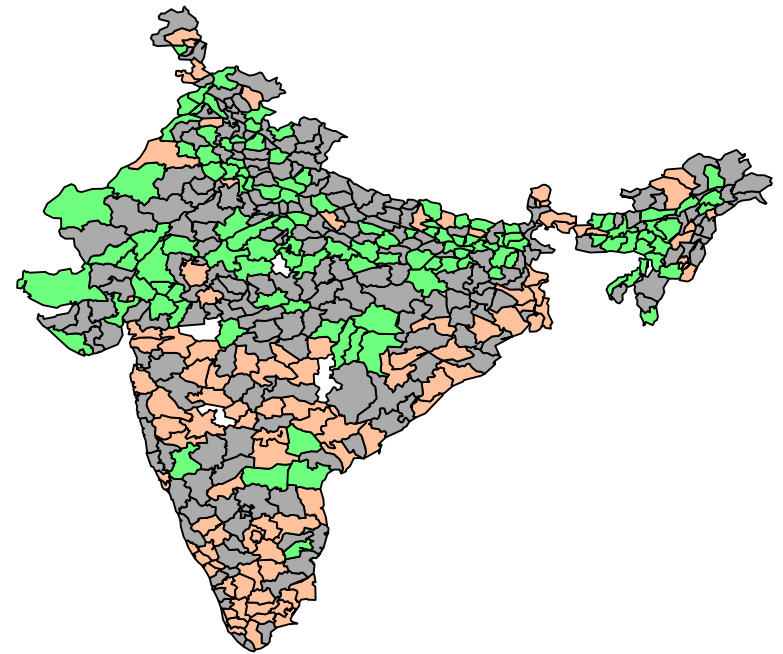
- Spatial effects:



**5% Quantile**



**95%-Posterior Probabilities**



## Asymmetrically Weighted Least Squares

- Expectile-specific parameters are easier to obtain since the criterion

$$\sum_{i=1}^n w_i(\tau)(y_i - \eta_{\tau i})^2 + \sum_{j=1}^p \lambda_j \boldsymbol{\gamma}'_{\tau j} \mathbf{K}_j \boldsymbol{\gamma}_{\tau j}$$

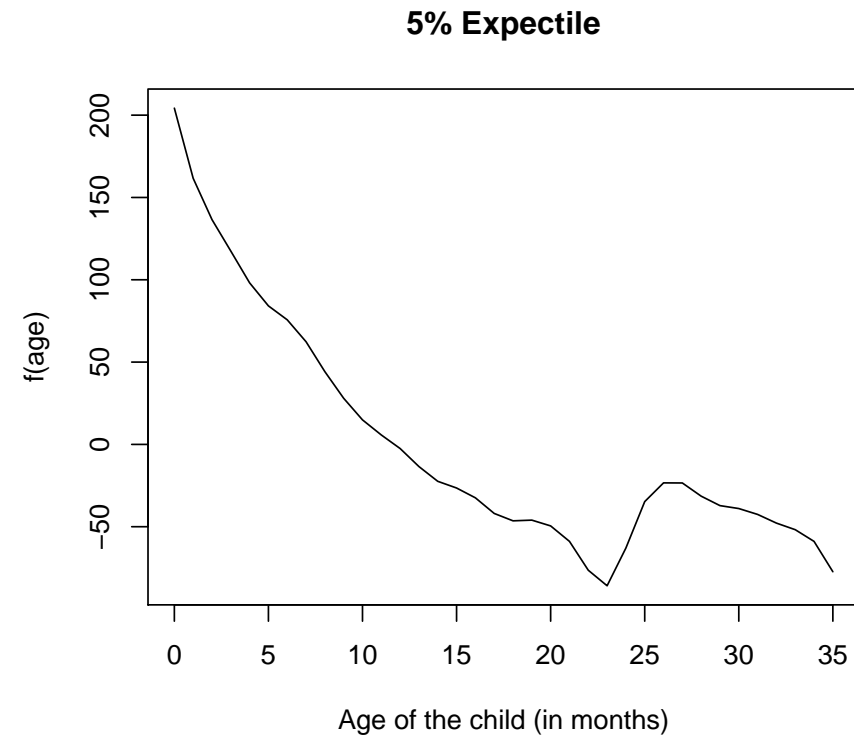
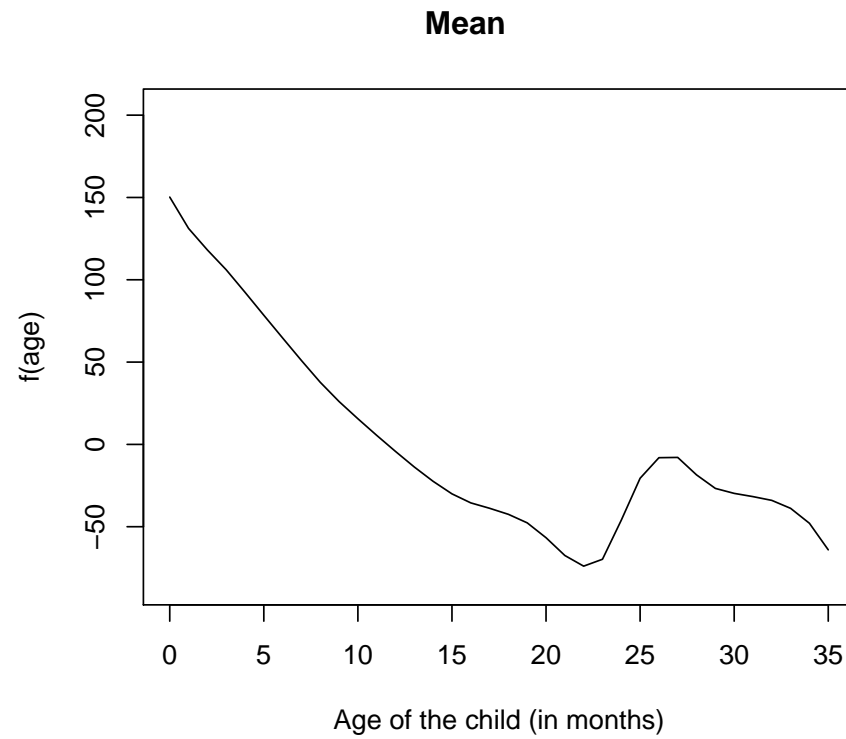
is differentiable with respect to the regression coefficients.

- Iteratively weighted penalised least squares estimation:

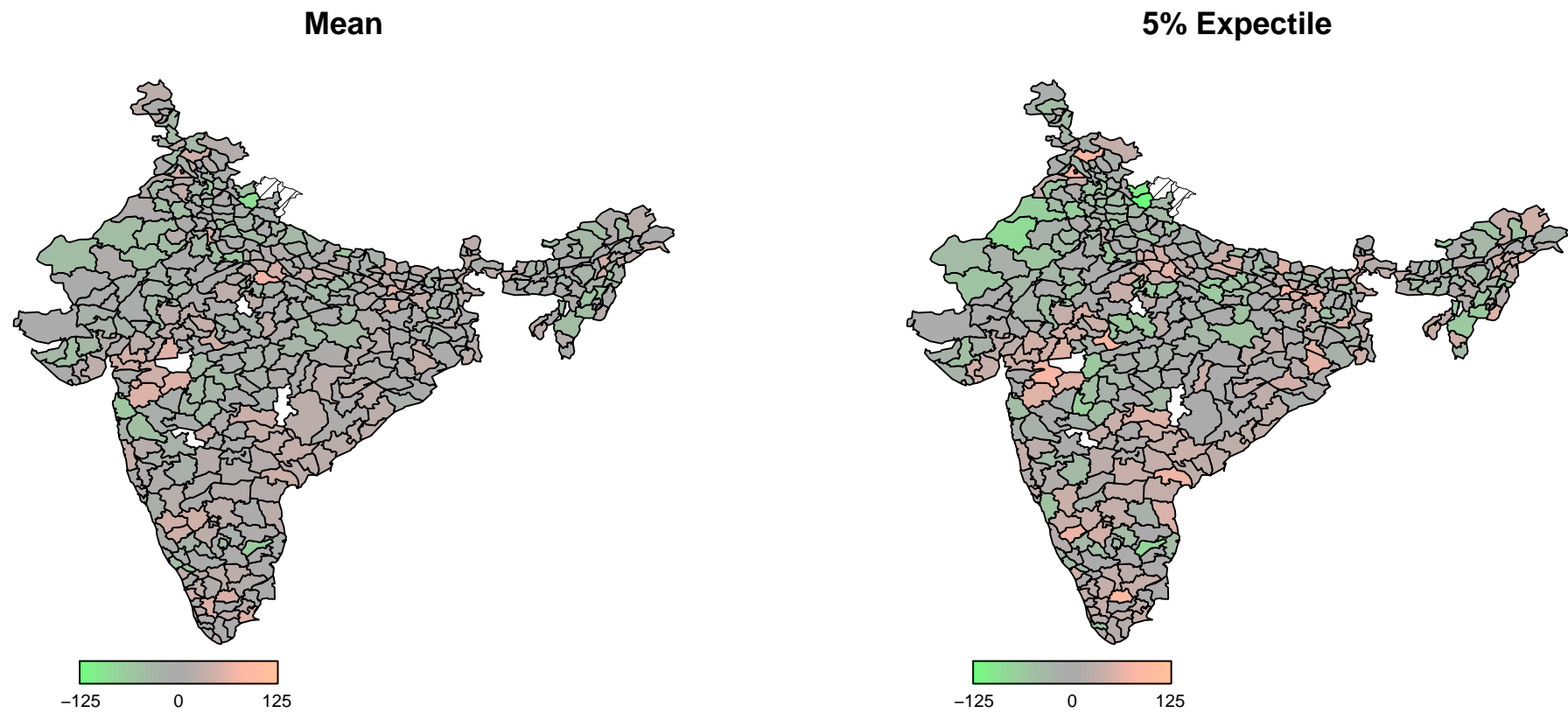
$$\hat{\boldsymbol{\gamma}}_{\tau j} = (\mathbf{Z}'_j \mathbf{W}(\tau) \mathbf{Z}_j + \lambda_j \mathbf{K}_j)^{-1} \mathbf{Z}'_j \mathbf{W}(\tau) (\mathbf{y} - \boldsymbol{\eta}_{\tau} + \mathbf{Z}_j \boldsymbol{\gamma}_j).$$

- Smoothing parameters can be estimated based on a **mixed model representation** similar as in mean regression.

- Age of the child:



- Spatial effect:



# Boosting

- Boosting yields a **generic approach** for both quantiles and expectiles.
- The estimation problem is formulated as an **empirical risk minimisation problem**:

$$\sum_{i=1}^n w_i(\tau) |y_i - \eta_{\tau i}| \rightarrow \min_{\eta_{\tau}} \quad \text{bzw.} \quad \sum_{i=1}^n w_i(\tau) (y_i - \eta_{\tau i})^2 \rightarrow \min_{\eta_{\tau}}$$

- Main components of a boosting approach:
  - A **loss function** defining the estimation problem.
  - Suitable **base-learning procedures** for the model components.
- Estimation relies on the repeated application of the base-learning procedures to negative gradients of the loss function (“residuals”).

- Componentwise boosting yields **structured, interpretable model**.
- **Penalised least squares estimates** yield suitable base-learners for semiparametric regression.

## Summary & Extensions

- Flexible, semiparametric regression beyond mean regression.
- More complex models than in our example are possible, including for example
  - interaction surfaces.
  - random effects.
  - different types of spatial effects.
- Different inferential procedures are available
  - MCMC simulation techniques.
  - Mixed Models.
  - Boosting.

- Future work:
  - Investigate properties of the statistical estimates resulting from quantile and expectile regression
  - In particular: How to perform inference for the estimated regression coefficients?
  - Investigate properties of theoretical expectiles.
  - Bayesian quantile regression based on flexible error distributions to avoid restrictive assumptions on the error terms.

## Acknowledgements

- Location scale mixtures for Bayesian quantile regression: Yu Yue (Zicklin School of Business, City University of New York), Elisabeth Waldmann (Department of Statistics, LMU Munich), Stefan Lang (Department of Statistics, University of Innsbruck).
- Expectile regression: Fabian Sobotka (Department of Mathematics, Carl von Ossietzky University Oldenburg), Paul Eilers (University of Rotterdam), Sabine Schnabel (Max-Planck-Institute for Demography, Rostock).
- Boosting approaches: Nora Fenske, Torsten Hothorn (Department of Statistics, LMU Munich)
- A place called home:

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