Stat Comput (2013) 23:135–148
DOI 10.1007/s11222-011-9297-1

On confidence intervals for semiparametric expectile regression

Fabian Sobotka · Göran Kauermann · Linda Schulze Waltrup · Thomas Kneib

Received: 10 June 2011 / Accepted: 20 October 2011 / Published online: 22 November 2011

Abstract In regression scenarios there is a growing demand for information on the conditional distribution of the response beyond the mean. In this scenario quantile regression is an established method of tail analysis. It is well understood in terms of asymptotic properties and estimation quality. Another way to look at the tail of a distribution is via expectiles. They provide a valuable alternative since they come with a combination of preferable attributes. The easy weighted least squares estimation of expectiles and the quadratic penalties often used in flexible regression models are natural partners. Also, in a similar way as quantiles can be seen as a generalisation of median regression, expectiles offer a generalisation of mean regression. In addition to regression estimates, confidence intervals are essential for interpretational purposes and to assess the variability of the estimate, but there is a lack of knowledge regarding the asymptotic properties of a semiparametric expectile regression estimate. Therefore confidence intervals for expectiles based on an asymptotic normal distribution are introduced. Their properties are investigated by a simulation study and compared to a boostrap-based gold standard method. Finally the introduced confidence intervals help to evaluate a geoadditive expectile regression model on childhood malnutrition data from India.

F. Sobotka (✉)
Department of Mathematics, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany
e-mail: fabian.sobotka@uni-oldenburg.de

G. Kauermann · L. Schulze Waltrup
Department of Statistics, Ludwig Maximilians University Munich, 80539 Munich, Germany

T. Kneib
Department of Economics, Georg August University Göttingen, 37073 Göttingen, Germany

Keywords Expectiles · Least asymmetrically weighted squares · P-splines · Confidence intervals · Semiparametric regression

1 Introduction

1.1 Expectiles

Recent interest in modern regression modelling has focused on extending available model specifications beyond mean regression by describing more general properties of the response distribution. For example, Rigby and Stasinopoulos (2005) proposed regression models for location, scale and skewness where separate predictors can be specified for various parameters of a response distribution. A completely distribution free approach is quantile regression (Koenker and Bassett 1978) where regression effects on the conditional quantile function of the response are assumed. Combining models for a large set of quantiles then allows to characterise the complete conditional distribution of the response.

Quantile regression for the τ-quantile with $\tau \in (0, 1)$ relies on the regression specification

$$y_i = \eta_{i,\tau} + \varepsilon_{i,\tau}, \quad i = 1, \ldots, n, \tag{1}$$

where $\eta_{i,\tau}$ is a (quantile-specific) predictor and $\varepsilon_{i,\tau}$ are independent error terms. Instead of imposing the usual mean regression model assumption that $E(\varepsilon_{i,\tau}) = 0$, quantile regression relies on the assumption that for the quantile function Q holds that $Q_{\varepsilon_{i,\tau}}(\tau) = 0$, i.e. the τ-quantile of the error distribution is zero. This implies that the conditional quantile of the response y_i is given by the predictor $\eta_{i,\tau}$. Note that no specific distribution is assumed for the error terms or responses and that in particular the error distribution

may differ between individuals. Estimation of quantile specific predictors now relies on minimising the asymmetrically weighted absolute residuals criterion $\sum_{i=1}^{n} w_{i,\tau}|y_i - \eta_{i,\tau}|$ with weights

$$w_{i,\tau} = w_{i,\tau}(\eta_{i,\tau}, y_i) = \begin{cases} \tau, & \text{for } y_i \geq \eta_{i,\tau} \\ 1-\tau, & \text{for } y_i < \eta_{i,\tau}. \end{cases} \tag{2}$$

A computationally attractive alternative to quantile regression is expectile regression, where absolute residuals are replaced with squared residuals yielding the fit criterion

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

with weights as defined in (2). The underlying assumption in regression model (1) is that the τ-expectiles μ_τ of the error terms are zero. They are implicitly defined by $\mu_\tau = \arg\min_m E[w_{i,\tau}(m, \varepsilon_i)(\varepsilon_{i,\tau} - m)^2]$. Least asymmetrically weighted squares (LAWS) estimation of expectiles dates already back to Newey and Powell (1987) but recently re-gained interest in the context of semiparametric or geoadditive regression (see for example Schnabel and Eilers 2009; Sobotka and Kneib 2010). Expectile estimation is thereby a special form of M-quantile estimation, see Breckling and Chambers (1988), Jones (1994). One of the advantages of expectile regression is that estimation basically reduces to (iteratively) weighted least squares fits since the optimality criterion is differentiable with respect to the regression effects while linear programming routines have to be used in case of quantile regression. This is of particular relevance when considering more flexible regression specifications as for example in geoadditive regression. The effects included here depend on a quadratic penalty for smooth estimates which can easily be included in a least squares estimation procedure. Further, when using expectiles (or quantiles) we try to get a complete picture of the conditional distribution of the response while at the same time avoiding a parametric specification for the distribution. To achieve this, we need to consider a set of expectiles or quantiles. In this scenario, a single estimate would not hold more information than the mean. Therefore we regard the reduced interpretability of expectiles as non-critical. Nevertheless, the estimation efficiency of expectiles and the interpretability of quantiles could be combined, if wished for, since Efron (1991) already proposed a method to obtain quantiles from a set of expectiles.

In summary, point estimates for expectile regression are easily derived for simple as well as complex models but their statistical properties are not yet well understood. In contrast, confidence intervals and significance tests for quantile regression have been studied extensively, relying for example on asymptotic considerations, the connection of quantiles to ranks or on bootstrap procedures (Koenker 2005; Kocherginsky and He 2005; Buchinsky 1998). In this paper, we derive asymptotic properties of expectile regression estimates and use them to construct corresponding confidence intervals. We continue the work of Newey and Powell (1987) by introducing a correction for the asymptotic results and extending them to semiparametric regression models. Further we determine the empirical properties of the asymptotic results. Therefore we state bootstrap-based confidence intervals as a computationally demanding gold standard for comparison with confidence intervals relying on asymptotic normality. Pointwise bootstrap percentile intervals have already been considered in Sobotka and Kneib (2010). However, the empirical properties were not determined and the method proved to be impractical for larger data sets due to the highly increased computational costs.

1.2 Geoadditive expectile regression

The need for our methodological innovations has arisen during a large-scale application on childhood malnutrition in developing countries where the impact of a large set of covariates should be assessed with respect to their impact on the nutritional status of children. Exploring not only the conditional mean but also extreme parts of the conditional distribution is of particular interest in this application since it allows to determine specific determinants of severe malnutrition by modelling lower expectiles. A comparable application is considered in Fenske et al. (2011) who use boosting to estimate regression quantiles in a high-dimensional additive quantile regression model, but spatial effects were not included and confidence intervals are not provided. For the assessment of estimation uncertainty they apply cross-validation in combination with the stability selection procedure recently proposed by Meinshausen and Bühlmann (2010). In this paper we use an extended, geoadditive model specification as introduced by Kammann and Wand (2003). The model definition combines parametric and nonlinear effects as well as spatial effects from geostatistics like kriging and can therefore be seen as a highly general semiparametric mixed model. For our application the geoadditive specification yields

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

where $\boldsymbol{\beta}_\tau$ corresponds to parametric effects of categorical covariates such as gender of the child (csex) or household-specific asset indicators (e.g. presence of a car), $f_{1,\tau}, \ldots, f_{6,\tau}$ are nonlinear effects of the continuous covariates age of the child in months (cage), duration of breastfeeding in months (cfeed), body mass index of the mother at birth (mbmi), age

of the mother at birth (mage) and education years of the mother and the mother's partner (medu, medupart) modeled via penalised splines and $f_{\text{spat},\tau}$ is a spatial effect corresponding to a Gaussian Markov random field.

The rest of this paper is structured as follows: Sect. 2 presents results on the asymptotic normality of expectile regression estimates in simple parametric models and for semiparametric extensions relying on penalised estimation. Required nonlinear and spatial effects are introduced alongside. Section 3 uses these asymptotic results to derive confidence intervals and also proposes bootstrap-based alternatives. Simulations and results for the childhood malnutrition data are presented in Sect. 4. The final Sect. 5 summarises the findings.

2 Asymptotics for least asymmetrically weighted squared error estimates

In the following, we assume that n metric observations $y_1, \ldots, y_n$ are given. For the underlying unknown distribution we require the existence of second moments. Further, all inverted matrices are assumed to have full rank.

2.1 Parametric models

We start our considerations with a simple, parametric model $\eta_{i,\tau} = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$ and study the asymptotic behaviour of

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} w_{i,\tau}(\boldsymbol{\beta}_\tau)\big(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau\big)^2,$$

where $w_{i,\tau}(\boldsymbol{\beta}_\tau) := w_{i,\tau}(\eta_{i,\tau}, y_i)$. Let $\boldsymbol{\beta}_\tau^0$ be the true parameter vector implicitly defined through

$$0 = \sum_{i=1}^{n}\Bigg\{(1-\tau)\int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0}\big(y - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big) f(y|\boldsymbol{x}_i)\mathrm{d}y + \tau\int_{\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0}^{\infty}\big(y - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big) f(y|\boldsymbol{x}_i)\mathrm{d}y\Bigg\}. \tag{4}$$

To avoid complexities arising from the dependence of the weights on the parameter vector, let for the moment $w_{i,\tau}^0 = w_{i,\tau}(\boldsymbol{\beta}_\tau^0)$ be the "true" weights and define $\hat{\boldsymbol{\beta}}_\tau^0$ as the minimiser of

$$\hat{\boldsymbol{\beta}}_\tau^0 = \arg\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} w_{i,\tau}^0\big(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau\big)^2 \tag{5}$$

which can easily be derived explicitly as

$$\hat{\boldsymbol{\beta}}_\tau^0 = \left(\sum_{i=1}^{n} w_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} w_{i,\tau}^0\boldsymbol{x}_i y_i\right). \tag{6}$$

Since the weights are considered as fixed we end up with standard weighted regression and obtain the following result:

Lemma 1 *The least asymmetrically weighted squares estimate with fixed weights is asymptotically normal, i.e.*

$$\hat{\boldsymbol{\beta}}_\tau^0 \overset{a}{\sim} N\big(\boldsymbol{\beta}_\tau^0, \operatorname{Var}\big(\hat{\boldsymbol{\beta}}_\tau^0\big)\big) \tag{7}$$

with covariance matrix

$$\begin{aligned}\operatorname{Var}\big(\hat{\boldsymbol{\beta}}_\tau^0\big) &= \left(\sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left\{\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\operatorname{Var}\big(\varpi_{i,\tau}^0\big(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau\big)\big)\right\}\\ &\quad\times\left(\sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\end{aligned} \tag{8}$$

with $\varpi_{i,\tau}^0 = E(w_{i,\tau}^0) = (1-\tau)P(y_i < \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0) + \tau\, P(y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0)$.

Proof With fixed weights, it is easy to show that

$$E\big(w_{i,\tau}^0 y_i\big) = (1-\tau)\int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0} y f(y|\boldsymbol{x}_i)\mathrm{d}y + \tau\int_{\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0}^{\infty} y f(y|\boldsymbol{x}_i)\mathrm{d}y,$$

which, combined with the implicit definition of the expectile (4), yields

$$E\left(\sum_{i=1}^{n} w_{i,\tau}^0\boldsymbol{x}_i y_i\right) = \sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0.$$

Applying standard expansion techniques to the weights in the first component in (6) yields

$$\left(\sum_{i=1}^{n} w_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} + O_p\big(n^{-1}\big),$$

so that we can extract the asymptotically leading components in (6) through

$$\hat{\boldsymbol{\beta}}_\tau^0 - \boldsymbol{\beta}_\tau^0 = \left(\sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\varpi_{i,\tau}^0\boldsymbol{x}_i\big(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big)\right) + O_p\big(n^{-1}\big). \tag{9}$$

This shows that $E(\hat{\boldsymbol{\beta}}_\tau^0) = \boldsymbol{\beta}_\tau^0 + O(n^{-1})$ and the variance of the weighted least squares estimate with fixed weights

equals (8). With the variance being of order $O(n^{-1})$, we obtain $\hat{\boldsymbol{\beta}}_\tau^0 - \boldsymbol{\beta}_\tau^0 = O_p(n^{-1/2})$ which, together with (9), yields the asymptotic normality (7). □

The next step in our consideration is to replace weights $w_{i,\tau}^0 = w_{i,\tau}(\boldsymbol{\beta}_\tau^0)$ in (5) by its estimate $\hat{w}_{i,\tau}^0 = w_{i,\tau}(\hat{\boldsymbol{\beta}}_\tau^0)$, that is we allow the weights to depend on the parameter estimate.

Theorem 1 *The least asymmetrically weighted squares estimate with estimated weights is asymptotically normal, i.e.*

$$\hat{\boldsymbol{\beta}}_\tau \overset{a}{\sim} N\big(\boldsymbol{\beta}_\tau^0, \mathrm{Var}(\hat{\boldsymbol{\beta}}_\tau^0)\big). \tag{10}$$

A proof is available under the assumptions stated in the beginning. It is provided in the appendix and follows a similar line of thought as in Newey and Powell (1987). The inner component (8) of the variance in (7) and (10), respectively, can easily be derived analytically, but the analytic form is hard to estimate. We therefore suggest to replace $\mathrm{Var}(\varpi_{i,\tau}^0(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0))$ by its empirical version

$$\big(w_{i,\tau}^0\big)^2\big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\big)^2. \tag{11}$$

Apparently, replacing (11) with its fitted version by substituting $\boldsymbol{\beta}_\tau^0$ with its estimate $\hat{\boldsymbol{\beta}}_\tau^0$ will lead to down-biased estimates since fitted squared expectile residuals underestimate the variance, like in classical regression. We therefore need to adjust (11) when applying its fitted version. From mean regression we already know that without further assumptions for the distribution of the residuals we have

$$\mathrm{Var}\big\{\big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\big)\big\} = \mathrm{Var}\big\{\big(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big)\big\}(1 - h_{ii})$$

with h_{ii} being the ith diagonal element of the hat matrix $\boldsymbol{H}$, say. For expectile regression we obtain the generalised hat matrix $H^\tau = (h_{ij}^\tau)_{ij}$ with

$$h_{ij}^\tau = w_{i,\tau}^0 \boldsymbol{x}_i'\left(\sum_{k=1}^n w_{k,\tau}^0 \boldsymbol{x}_k\boldsymbol{x}_k'\right)^{-1}\boldsymbol{x}_j,$$

that coincides with the OLS hat matrix for $\tau = 0.5$, i.e. $\boldsymbol{H} = H^{0.5}$. Therefore we use (11) but estimate the variance with the adjusted fitted residuals

$$(\hat{w}_{i,\tau})^2\frac{(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0)^2}{1 - h_{ii}^\tau}, \tag{12}$$

where $\hat{w}_{i,\tau} = w_{i,\tau}(\hat{\boldsymbol{\beta}}_\tau)$.

2.2 Semiparametric models

Now we extend the results from the previous section to semiparametric regression models with generic predictor

$$\eta_{i,\tau} = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \sum_{j=1}^r f_{j,\tau}(z_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \sum_{j=1}^r \boldsymbol{b}_{ij}'\boldsymbol{\gamma}_{j,\tau},$$

where $\boldsymbol{x}_i'\boldsymbol{\beta}_\tau$ summarises usual parametric, linear effects while $f_{1,\tau}(z_i), \ldots, f_{r,\tau}(z_i)$ represent generic semiparametric effects of covariates z_i. These may for example stand for nonlinear effects of continuous covariates or spatial effects as in our application (also compare (3)) but may also correspond to more complex terms such as varying coefficients or interaction surfaces (see Fahrmeir et al. 2004, for more details on available model terms). Each of the generic regression terms can then be expanded in terms of basis functions, yielding a representation as $f_{j,\tau}(z_i) = \boldsymbol{b}_{ij}'\boldsymbol{\gamma}_{j,\tau}$ where $\boldsymbol{b}_{ij}$ comprises the basis function evaluations while $\boldsymbol{\gamma}_{j,\tau}$ is a vector of basis coefficients.

To enforce specific properties of the resulting estimates such as smoothness, estimation then typically relies on penalised fit criteria. In case of expectile regression, this yields

$$\sum_{i=1}^n w_{i,\tau}(\eta_{i,\tau})(y_i - \eta_{i,\tau})^2 + \sum_{j=1}^r \lambda_{j,\tau}\boldsymbol{\gamma}_{j,\tau}'\boldsymbol{K}_j\boldsymbol{\gamma}_{j,\tau},$$

where $\lambda_{j,\tau} \geq 0$, $j = 1, \ldots, r$ are smoothing parameters and $\boldsymbol{K}_j$ are appropriate penalty matrices.

The two relevant examples of semiparametric model terms in the context of our application are penalised splines and Gaussian Markov random fields. The former enables estimation of nonlinear effects $f_{j,\tau}(z_i)$ of a single continuous covariate z_i and relies on a basis expansion in terms of B-splines in combination with a difference penalty for the basis coefficients. Therefore, in this case $\boldsymbol{b}_{ij}' = (B_1(z_i), \ldots, B_K(z_i))$ where $B_1, \ldots, B_K$ is a K-dimensional B-spline basis and $\boldsymbol{K}_j = \boldsymbol{D}'\boldsymbol{D}$ with a difference matrix $\boldsymbol{D}$. The penalty $\boldsymbol{\gamma}_{j,\tau}'\boldsymbol{K}_j\boldsymbol{\gamma}_{j,\tau}$ then consists of the sum of all squared differences of adjacent coefficient sequences and penalises large variation in the function estimate (compare Eilers and Marx 1996). Gaussian Markov random fields allow to estimate spatial effects based on geographical data. Suppose that each individual observation pertains to one region s_i from a fixed set of regions $\mathcal{S} = \{1, \ldots, S\}$. Then the design vector $\boldsymbol{b}_{ij}$ is an S-dimensional indicator vector with a one at the position of the region of observation i and zeros otherwise while the vector of coefficients $\boldsymbol{\gamma}_{j,\tau}$ simply collects all potential spatial effects. The penalty matrix should enforce spatial smoothness and therefore has the structure of an adjacency matrix such that the penalty $\boldsymbol{\gamma}_{j,\tau}'\boldsymbol{K}_j\boldsymbol{\gamma}_{j,\tau}$ consists of all squared differences between spatial effects of neighboring regions (see Rue and Held 2005, for details).

In any case, the estimates in semiparametric expectile regression models for fixed smoothing parameters can always be written as

$$\hat{\boldsymbol{\theta}}_\tau = \left(\sum_{i=1}^n \boldsymbol{u}_i' w_{i,\tau}\boldsymbol{u}_i + \boldsymbol{P}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{u}_i' w_{i,\tau} y_i\right)$$

where $\boldsymbol{\theta}_\tau = (\boldsymbol{\beta}_\tau', \boldsymbol{\gamma}_{1,\tau}', \ldots, \boldsymbol{\gamma}_{r,\tau}')'$ and $\boldsymbol{u}_i = (\boldsymbol{x}_i', \boldsymbol{b}_{i1}', \ldots, \boldsymbol{b}_{ir}')'$ collect all regression coefficients and design vectors, re-

spectively, and $\boldsymbol{P} = \text{blockdiag}(\mathbf{0}, \lambda_{1,\tau}\boldsymbol{K}_1, \ldots, \lambda_{r,\tau}\boldsymbol{K}_r)$ is the complete penalty matrix.

Theorem 2 *For fixed smoothing parameters, the penalised least asymmetrically weighted squares estimate is asymptotically normal, i.e.*

$$\hat{\boldsymbol{\theta}}_\tau \overset{a}{\sim} N\big(\boldsymbol{\theta}_\tau^0, \text{Var}\big(\hat{\boldsymbol{\theta}}_\tau^0\big)\big),$$

where $\boldsymbol{\theta}_\tau^0$ is defined in analogy to $\boldsymbol{\beta}_\tau^0$ and

$$\begin{aligned}\text{Var}\big(\hat{\boldsymbol{\theta}}_\tau^0\big) = &\left(\sum_{i=1}^n \varpi_{i,\tau}^0 \boldsymbol{u}_i\boldsymbol{u}_i' + \boldsymbol{P}\right)^{-1} \\ &\times \left\{\sum_{i=1}^n \boldsymbol{u}_i\boldsymbol{u}_i' \,\text{Var}\big(\big(w_{i,\tau}^0\big)\big(y_i - \boldsymbol{u}_i'\boldsymbol{\theta}_\tau^0\big)\big)\right\} \\ &\times \left(\sum_{i=1}^n \varpi_{i,\tau}^0 \boldsymbol{u}_i\boldsymbol{u}_i' + \boldsymbol{P}\right)^{-1}. \end{aligned} \tag{13}$$

The covariance matrix of the penalised estimate has the typical sandwich form arising from the inclusion of the penalty in the estimation objective.

As before, the residual terms $y_i - \boldsymbol{u}_i'\boldsymbol{\theta}_\tau^0$ in (13) can be replaced by empirical terms in order to estimate the variance, where in close analogy to (12) we divide the fitted version $(y_i - \boldsymbol{u}_i'\hat{\boldsymbol{\theta}}_\tau^0)^2$ by its generalised hat matrix entry

$$1 - w_{i,\tau}^0 \boldsymbol{u}_i'\left(\sum_{j=1}^n w_{j,\tau}^0 \boldsymbol{u}_j\boldsymbol{u}_j' + \boldsymbol{P}\right)^{-1}\boldsymbol{u}_i. \tag{14}$$

Of course, in practice the smoothing parameters will have to be determined jointly with the regression coefficients to obtain a data-driven amount of smoothness. A REML estimate based on the Schall algorithm (Schall 1991) has been adapted to expectiles by Schnabel and Eilers (2009). In our simulations and the example, we will use asymmetric cross-validation adapted for geoadditive expectile regression in Sobotka and Kneib (2010). The grid search over the smoothing parameter for the minimal cross-validation score is widened to an r-dimensional grid. The score itself is defined as

$$V_g^w = \frac{n\sum_{i=1}^n w_{i,\tau}(y_i - \eta_{\tau,i})^2}{[\text{tr}(\mathbb{1} - \boldsymbol{H}^\tau)]^2}$$

and the score is therefore independent from the number of functions r. The method is computationally demanding but accurate. We use the more accurate possibility to gain reliable informations on the confidence interval performance.

3 Confidence intervals

3.1 Asymptotic confidence intervals

Equation (13) together with the correction (14) provides us with the asymptotic covariance matrix of the complete estimate $\hat{\boldsymbol{\theta}}_\tau$ and therefore the covariance matrix of specific coefficient vectors of interest can immediately be obtained by extracting the appropriate sub-blocks. For example, for the variance of the estimated function evaluation $\hat{f}_{j,\tau}(z_i) = \boldsymbol{b}_{ij}'\hat{\boldsymbol{\gamma}}_{j,\tau}$, we obtain

$$\text{Var}\big(\hat{f}_{j,\tau}(z_i)\big) = \boldsymbol{b}_{ij}'\,\text{Var}(\hat{\boldsymbol{\gamma}}_{j,\tau})\boldsymbol{b}_{ij}$$

where $\text{Var}(\hat{\boldsymbol{\gamma}}_{j,\tau})$ is the block of $\text{Var}(\hat{\boldsymbol{\theta}}_\tau)$ corresponding to $\hat{\boldsymbol{\gamma}}_{j,\tau}$. Together with the asymptotic normality of the least asymmetrically weighted squares estimate, this yields the following confidence interval for the true function evaluation $f_{j,\tau}(z_i)$:

$$\text{CI}\big(\hat{f}_{j,\tau}(z_i)\big) = \big[\hat{f}_{j,\tau}(z_i) \pm z_{1-\frac{\alpha}{2}}\sqrt{\text{Var}\big(\hat{f}_{j,\tau}(z_i)\big)}\big]$$

where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. Note that a particular amount of undercoverage is inevitable since we work with normal but not t-distribution quantiles.

3.2 Bootstrap confidence intervals

A further possibility to fit pointwise $(1-\alpha)$-confidence intervals to expectile regression curves can be created with large computational expense. By conducting a nonparametric bootstrap, the distribution of the estimated expectiles can be approximated. At first, B bootstrap samples $(\boldsymbol{y}, \boldsymbol{X})^*_{b=1,\ldots,B}$ are drawn from the original data set. The expectiles are fitted independently for all B samples resulting in a bootstrapped sample $\mu_\tau(\boldsymbol{x}_{i,1}^*), \ldots, \mu_\tau(\boldsymbol{x}_{i,B}^*)$ from the unknown distribution of the true expectile $\mu_\tau(x_i)$. According to Efron and Tibshirani (1993) for a number of bootstrap replications $B \geq 1000$ we can construct bootstrap percentile intervals from $\mu_\tau(x_{i,1}^*), \ldots, \mu_\tau(x_{i,B}^*)$ with sufficient quality. This holds under the assumption that the empirical distribution formed by the observations $(y_i, x_i)_{i=1,\ldots,n}$, is a good estimate for the unknown true distribution. The resulting pointwise intervals are therefore constructed from the $(\frac{\alpha}{2}B)$-th and the $((1-\frac{\alpha}{2})B)$-th element of the sorted set of the expectile estimates for each of the effects f_j from the Bootstrap samples and $i = 1, \ldots, n$:

$$\text{CI}\big(\hat{f}_{j,\tau}^*(x_i)\big) = \big[\hat{f}_{j,\tau}\big(\boldsymbol{x}_{b_1,i}^*\big)_{(\frac{\alpha}{2}B)};\ \hat{f}_{j,\tau}\big(\boldsymbol{x}_{b_2,i}^*\big)_{((1-\frac{\alpha}{2})B)}\big].$$

An alternative would be to construct bootstrap-t-intervals. This would require an additional nonparametric bootstrap

inside every previously drawn bootstrap sample to estimate the variance of the expectiles. In consequence this method would take a lot of time or processor cores when used on large data sets. Therefore we restrict our analyses to the bootstrap percentile intervals.

4 Empirical evaluation

4.1 Simulation study

After introducing two estimation approaches for expectile regression confidence intervals, an asymptotic and a numerical method, their merits and disadvantages will now be investigated in terms of a simulation study. The data structures considered in the simulation study are linear on the one hand, mixed and additive nonlinear on the other in order to simulate different data scenarios. We will also investigate numerical properties of the estimation approaches in terms of computing time.

4.1.1 Design

The models used for the simulations are defined as

$$y = 0.75 + 0.9x_1 + \varepsilon \tag{15}$$

$$y = 3x_3 + 3\underbrace{\exp\left(-x_1^2\right)}_{f_{\text{p-spline}}(x_1)} + \varepsilon \tag{16}$$

$$y = \underbrace{x_1^2}_{f_{\text{p-spline}}(x_1)} + \underbrace{\sin(8x_2 - 4) + 2\exp\left(-(16x_2 - 8)^2\right)}_{f_{\text{p-spline}}(x_2)} + \varepsilon \tag{17}$$

where ε follows either a normal distribution $N(0, 3^2)$, a beta distribution or the so called "expectiles-meet-quantiles" (emq) distribution with distribution function

$$F_{\mu,s}(\varepsilon) = 0.5\left(1 + \text{sign}(\varepsilon - \mu)\sqrt{1 - \frac{2}{2 + (\frac{\varepsilon-\mu}{s})^2}}\right)$$

with expectation $\mu = 0$ and scaling parameter $s = \sqrt{2}$ (The variance itself is not finite regardless the value of s). The latter distribution has the desirable property that quantiles and expectiles coincide (see Koenker 2005, p. 67) for all parameters $\mu \in \mathbb{R}$ and $s > 0$. Also, due to the non-existing second moments a key assumption for the asymptotic results is violated. Here, we can examine the importance of the assumption. Note that both the normal and the "emq" distribution are homoscedastic while the beta distribution is variance heteroscedastic with Beta$(0.5x_1, 3x_1)$ for models (15) and (16) and Beta$(0.5x_1, 3x_2)$ for model (17). The true expectiles of the above distributions are obtained by numerically solving

$$\tau = \frac{G(\mu_\tau) - \mu_\tau F(\mu_\tau)}{2(G(\mu_\tau) - \mu_\tau F(\mu_\tau)) + (\mu_\tau - \mu_{0.5})},$$

where F is the cumulative distribution function, $G(\mu_\tau) = \int_{-\infty}^{\mu_\tau} x \, \mathrm{d}F(x)$ is the partial moment function and $G(\infty) = \mu_{0.5}$ is the expectation of ε.

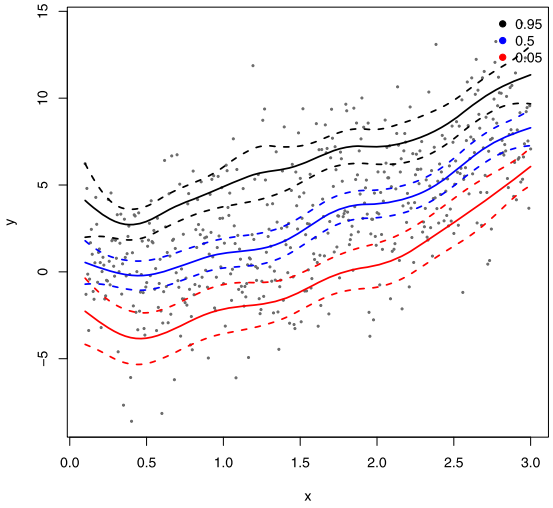
The binary covariate x_3 is drawn from a $B(1, 0.5)$ distribution. The values of the continuous covariates x_1 and x_2 are equally spaced over their domains [0; 3] and [0; 1], respectively. Therefore we have the same positions in every simulated data set where the confidence intervals are evaluated. The corresponding functions are modelled as cubic penalised splines with second order difference penalty and 20 inner knots. Figure 1 visualises simulated data for one replication to give an impression of the functional form of the effects considered. Based on sample sizes of $n = 100$, 250, 500 and 1000, we generated 1000 simulation replications for each of the 36 different data structures arising from the combination of (i) the model (linear, mixed and additive), (ii) the error distribution (normal, beta, emq), and (iii) the sample size. For each data set, we applied the two different approaches for the estimation of confidence intervals introduced in the previous section, i.e. asymptotic normality and bootstrap percentiles to determine confidence intervals for expectiles with asymmetries $\tau \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 0.98, 0.99\}$. The asymptotic normality was used to estimate confidence intervals from the regression coefficients obtained from least asymmetrically weighted squares (LAWS). The same is true for the bootstrap percentile intervals.

All simulations have been implemented using "expectreg" (Sobotka et al. 2011), a package for R (R Development Core Team 2010). The package also contains expectile functions for several distributions including those used in the simulations.

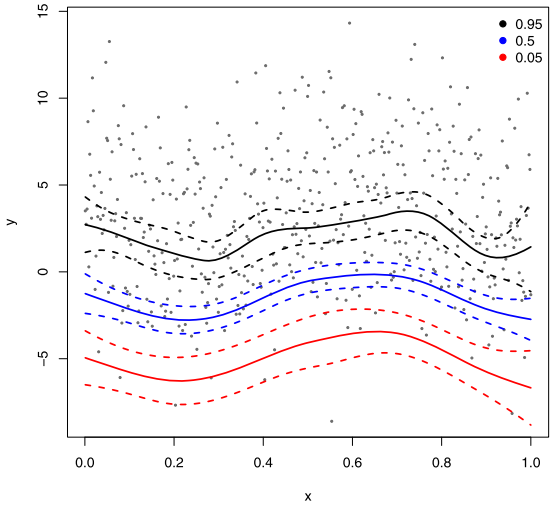
4.1.2 Performance measures

For the measurement of the quality of the results we evaluate the true expectile curve at the covariate values $x_{1,i}$ and $x_{2,i}$. Then the number of times the true expectile is covered by the interval are counted. Also the intervals will be compared according to their width. We therefore measure the coverage of the confidence intervals for $m = 1, 2$ at a given covariate value $x_{m,i}$ as

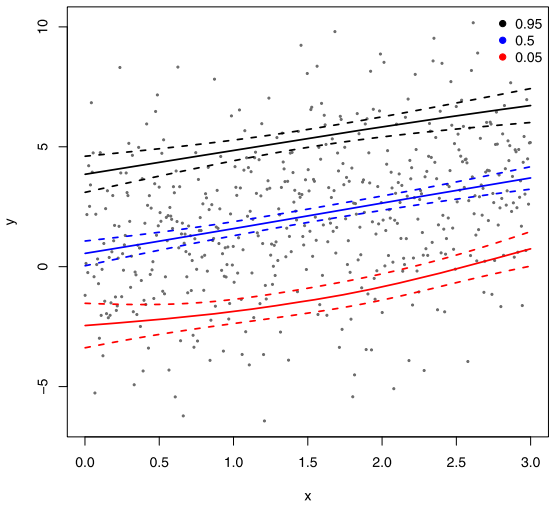
$$\widehat{\text{Cover}}(CI(\hat{f}_{j,\tau}(x_{m,i})) = \frac{1}{1000}\sum_{k=1}^{1000} \mathbb{1}_{\{\hat{f}_{j,\tau}(x_{m,i}) \in CI(\hat{f}_{j,\tau}^{[k]}(x_{m,i}))\}},$$



(a) Nonlinear effect $f_{\text{p-spline}}(x_1)$



(b) Nonlinear effect $f_{\text{p-spline}}(x_2)$



(c) Linear effect of x_1

Fig. 1 Exemplary data and fitted asymptotic confidence intervals for one simulated data set with $n = 500$ observations and $N(0, 3^2)$ distributed errors

the maximum width of all confidence intervals at all fixed $x_{m,i}$

$$\max \widehat{\text{Width}}\big(CI\big(\hat{f}_{j,\tau}(x_{m,i})\big)\big) = \max_k \big(\hat{f}^{[k]}_{j,\tau,U}(x_{m,i}) - \hat{f}^{[k]}_{j,\tau,L}(x_{m,i})\big)$$

and for a compact measure the mean coverage along the covariate x_m, $m = 1, 2$

$$\widehat{\text{Cover}}\big(CI\big(\hat{f}_{j,\tau}(x_m)\big)\big) = \frac{1}{1000n}\sum_{i=1}^{n}\sum_{k=1}^{1000} \mathbb{1}_{\{\hat{f}_{j,\tau}(x_{m,i}) \in CI(\hat{f}^{[k]}_{j,\tau}(x_{m,i}))\}} \tag{18}$$

as well as the mean interval width

$$\overline{\widehat{\text{Width}}}\big(CI\big(\hat{f}_{j,\tau}(x_m)\big)\big) = \frac{1}{1000n}\sum_{i=1}^{n}\sum_{k=1}^{1000} \hat{f}^{[k]}_{j,\tau,U}(x_{m,i}) - \hat{f}^{[k]}_{j,\tau,L}(x_{m,i}). \tag{19}$$

Here, $\hat{f}^{[k]}_{j,\tau}$ denotes the expectile estimate for the j-th effect in the k-th simulation run. Further, the upper or lower end of the interval is indicated by U and L, respectively. In order get a better hold of the actual quality of the confidence intervals guarantee identifiability of the expectiles in the additive model by centering $\tilde{f}_{j,\tau}(x_i) = f_{j,\tau}(x_i) - \bar{f}_{j,\tau}$.

4.1.3 Results

The first observation we can make is that the desired confidence level of 95% cannot be guaranteed for all situations. None of the introduced methods shows that quality. The best results are achieved for the special case of a mean regression ($\tau = 0.5$) and for covariate values near $\bar{x}$. The larger the asymmetry ($\tau \to 0$ or $\tau \to 1$) and the nearer to the edge of the covariates' support, the higher the probability that the confidence level will not be met. The former is displayed in Table 1, the latter is exemplarily shown for the beta distribution in Fig. 2. In addition, calculating the mean coverage for all covariates, as defined in Sect. 4.1.2, results in the simulated coverage probabilities shown in Table 1. Results for $n = 250$ and the width of the confidence intervals are available on request.

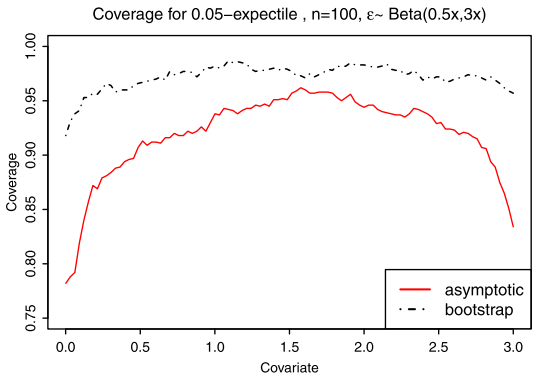
We also investigate the average width of the confidence intervals. Apparently, for symmetrical distributions the width increases towards the boundary of the covariate support. For the heteroscedastic scenario this needs not to be the case as the beta distribution shows. Table 1 shows an increasing coverage probability with growing sample size. The latter, however, is only partly true for the emq distribution due to the infinite variance. In comparison, the gain in coverage probability and the decrease in interval width is stronger for the confidence intervals constructed from the asymptotic properties. The latter is especially important since we want the narrowest interval width possible given a proper coverage. Analysing both measures together, the coverage (18) and the width (19), ensures that we select intervals for which the appropriate coverage is not gained by additional interval width.

Regarding the performance of the bootstrap percentile intervals one needs to bear in mind the increased computational demand. In fact, one needs to fit the complete set of considered expectiles in each nonparametric bootstrap samples which is a rather time-consuming method. After this computational burden, that can take more than an hour for a single data set, depending on the complexity of the data, the results are however satisfactory. The bootstrap intervals provide a coverage of nearly $1-\alpha$ with the limitations stated in the beginning. Especially for small samples, the provided coverage of the bootstrap method is better than from the asymptotic method without resulting in unreasonably wide intervals. Also for small samples the time required to conduct the bootstrap is within a few minutes depending on the possibilities for parallelisation.

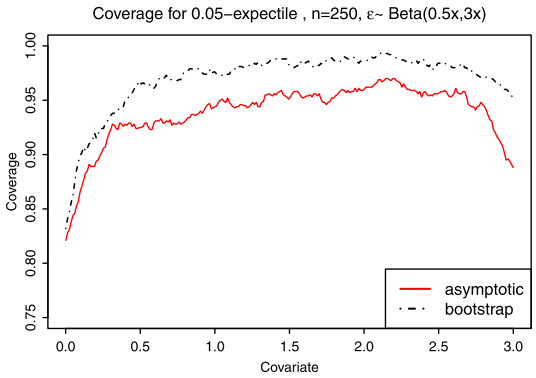
In conclusion, we can see that both methods have their merits and weaknesses. Small samples are best tackled with bootstrap intervals and for heteroscedastic errors or large samples we can recommend to use the asymptotic normality to construct confidence intervals for the expectile curves. If the variance does not exist, we can see that the violated assumption in the asymptotics leads to poor coverage. In simple cases, 500 observations will suffice. Otherwise and if extreme expectiles like $\tau = 0.01, 0.99$ shall be estimated, more are required.

Table 1 Mean relative coverage frequency as defined in (18) for the eleven asymmetry parameters, both estimation methods and all error distributions for a sample size of $n = 100, 500, 1000$

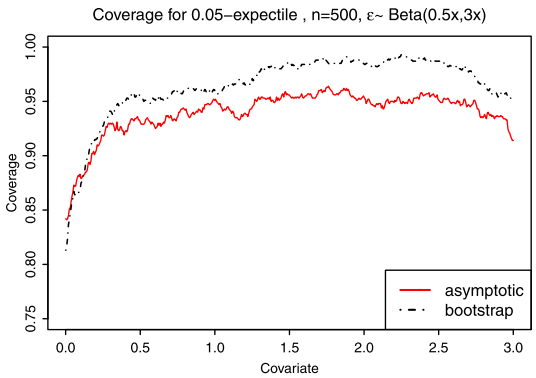
Error	$F_{emq}(0, \sqrt{2})$		$N(0, 3^2)$		Beta$(0.5x_1, 3x_1)$	
τ	boot	asympt	boot	asympt	boot	asympt
$n = 100$						
0.01	0.358	0.345	0.706	0.715	0.947	0.858
0.02	0.496	0.462	0.783	0.788	0.950	0.880
0.05	0.654	0.615	0.848	0.847	0.952	0.899
0.1	0.758	0.728	0.882	0.875	0.949	0.902
0.2	0.841	0.829	0.901	0.893	0.941	0.911
0.5	0.914	0.939	0.920	0.915	0.918	0.903
0.8	0.824	0.819	0.910	0.899	0.881	0.858
0.9	0.733	0.714	0.886	0.874	0.848	0.829
0.95	0.623	0.600	0.850	0.841	0.813	0.797
0.98	0.467	0.439	0.776	0.780	0.760	0.725
0.99	0.332	0.319	0.700	0.723	0.711	0.657
$n = 500$						
0.01	0.634	0.594	0.876	0.851	0.946	0.923
0.02	0.713	0.681	0.902	0.879	0.946	0.927
0.05	0.801	0.771	0.922	0.905	0.930	0.928
0.1	0.848	0.830	0.929	0.915	0.921	0.925
0.2	0.893	0.890	0.930	0.924	0.924	0.922
0.5	0.922	0.947	0.929	0.931	0.934	0.931
0.8	0.882	0.873	0.931	0.930	0.895	0.900
0.9	0.833	0.814	0.931	0.923	0.843	0.874
0.95	0.780	0.752	0.923	0.913	0.783	0.862
0.98	0.696	0.661	0.904	0.888	0.721	0.846
0.99	0.604	0.573	0.882	0.855	0.687	0.829
$n = 1000$						
0.01	0.714	0.665	0.937	0.879	0.929	0.934
0.02	0.790	0.734	0.939	0.905	0.933	0.937
0.05	0.858	0.802	0.932	0.919	0.921	0.939
0.1	0.897	0.843	0.932	0.926	0.911	0.934
0.2	0.919	0.890	0.936	0.932	0.923	0.927
0.5	0.925	0.942	0.946	0.935	0.947	0.934
0.8	0.889	0.888	0.942	0.931	0.912	0.910
0.9	0.876	0.846	0.933	0.926	0.824	0.894
0.95	0.856	0.808	0.930	0.915	0.697	0.888
0.98	0.783	0.741	0.913	0.898	0.603	0.873
0.99	0.708	0.681	0.884	0.877	0.563	0.861



(a) relative coverage for $n = 100$



(b) relative coverage for $n = 250$



(c) relative coverage for $n = 500$

Fig. 2 (Color online) Simulation results for $\tau = 0.05$ with $n = 100, 250, 500$ observations and Beta$(0.5x, 3x)$ distributed errors. The relative coverage frequency for both methods along the covariate is shown. The method of asymptotic normality is plotted in *red*, the LAWS bootstrap percentile intervals in *black* and *dotted*

4.2 Childhood malnutrition in India

Malnutrition is a severe problem in developing countries. Regular surveys are herefore conducted on national bases in order to determine risk factors for malnutrition. General and representative studies on health and population development are done by MEASURE Demographic and Health Surveys (DHS). Those include topics like HIV distribution, fertility or nutrition aspects. The data can be obtained from www.measuredhs.com free of charge for research purposes. In our case we use data on childhood malnutrition in India from the year 2001. After preprocessing and deleting observations with missing values, the data contains 24316 observations in 40 variables. In general, malnutrition of each individual i is measured as a score Z defined as

$$Z_i = \frac{AC_i - m}{s}$$

where AC is an anthropometric characteristic. Most of the time the weight in relation to the age is measured for this variable. This characteristic is standardised by subtracting the median m and dividing by the standard deviation s of the same attribute in a reference population. While a score based on weight is also an indicator for acute malnutrition, an insufficient height for a child's age, also called stunting, is a distinct indicator for chronic malnutrition. Therefore stunting is the variable that is modelled here. The score for stunting Z is neither normally distributed nor restricted to a certain support. In our data the value ranges from -600 to 600. The model is inspired by Fenske et al. (2011) and the predicted stunting η_τ for the τ-expectile is modelled as

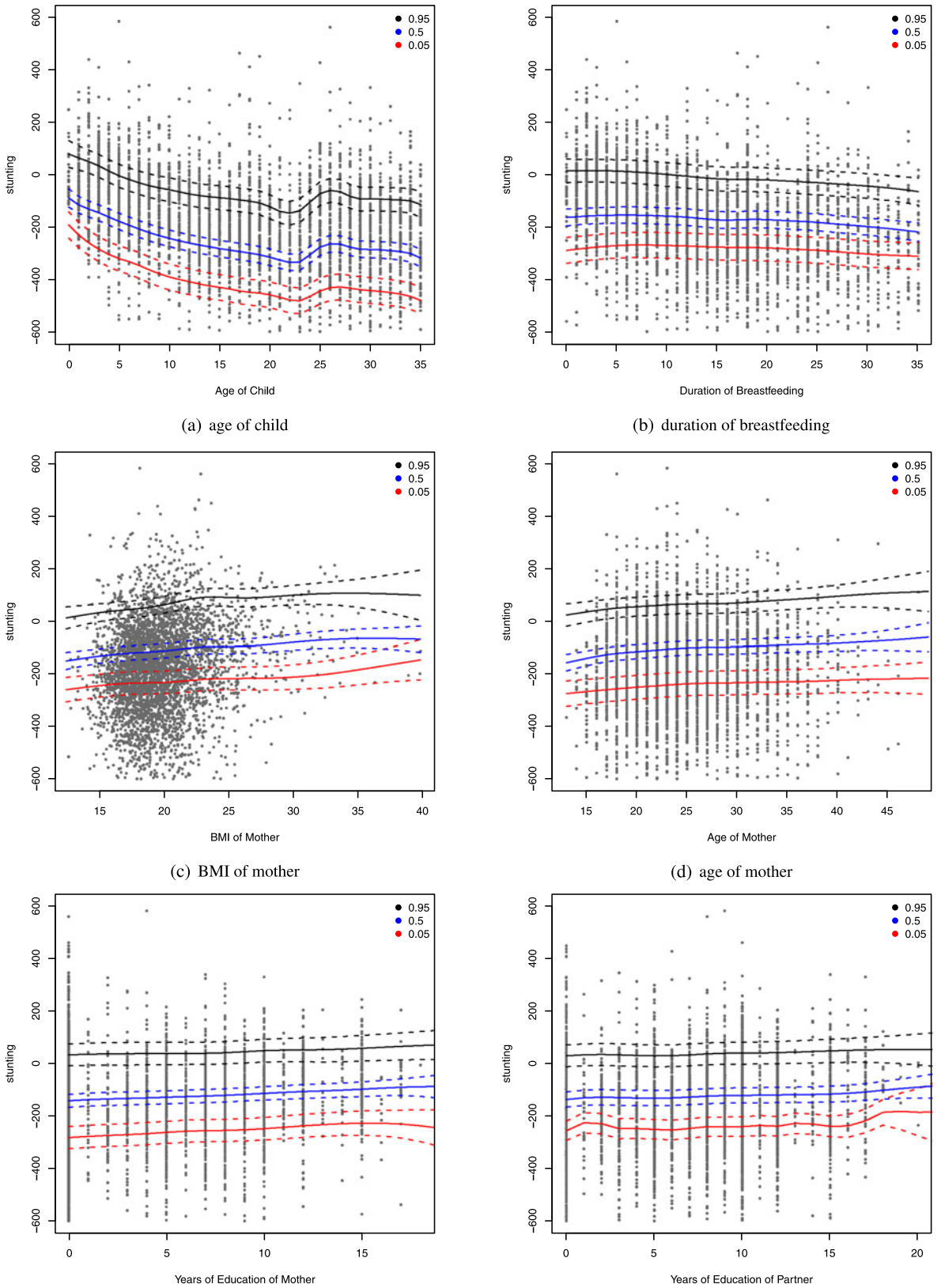
$$\begin{aligned}\eta_\tau = {} & \boldsymbol{x}'\boldsymbol{\beta}_\tau + f_{\tau,1}(\text{age of child}) \\ & + f_{\tau,2}(\text{duration of breastfeeding}) \\ & + f_{\tau,3}(\text{BMI of mother}) + f_{\tau,4}(\text{age of mother}) \\ & + f_{\tau,5}(\text{education years of mother}) \\ & + f_{\tau,6}(\text{education years of partner}) + f_{\tau,\text{spat}}(\text{district}).\end{aligned}$$

The parametric effects included in $\boldsymbol{x}$ are listed in Table 2. Further there are six nonlinear effects in the model that are fitted with a cubic P-spline basis constructed from 20 inner knots and penalised with second order differences. Also one spatial effect is included as a Markov random field. A special interest of this analysis lies in the lower tails of the conditional distribution of Z. The expectiles for small values of τ will show the relation of the covariates to the response for cases of severe malnutrition. Confidence intervals from a nonparametric bootstrap are not considered here as we expect a computing time of several weeks.

For the lower expectiles we can see that stunting gets worse if the child is later in the birth order. This as well as the insignificance of the residence region of the mother (rural/urban) is a result comparable to the lower quantiles computed by Fenske et al. (2011). The 0.8 and 0.95-expectiles show a different behaviour for these covariates. The family size is insignificant for children that do not suffer from stunting. For those children living in urban areas also has a positive effect. We can support this by the 0.95-expectile of the regions of India depicted in Fig. 4. The map shows a positive

Table 2 Estimated parametric effects for Childhood Malnutrition data. Reference categories and confidence intervals ($1-\alpha=0.95$) obtained by asymptotic normality are included in *italics*. Significant effects are set in **bold**

Variable/τ	0.05	0.2	0.8	0.95
sex of child	−2.91	−2.45	−1.35	3.52
reference: "male"	(*−8.63, 2.80*)	(*−7.31, 2.41*)	(*−6.46, 3.74*)	(*−2.88, 9.94*)
twin birth	**−67.53**	**−68.71**	**−72.21**	**−79.22**
reference: "single birth"	(*−91.01, −44.10*)	(*−88.25, −49.17*)	(*−93.84, −50.59*)	(*−112.10, −46.34*)
birth order:		*reference: "first"*		
"second"	−5.75	**−8.81**	**−7.66**	0.05
	(*−13.37, 1.87*)	(*−15.57, −2.06*)	(*−14.71, −0.61*)	(*−9.03, 9.13*)
" third"	**−15.70**	**−15.82**	**−14.55**	−7.45
	(*−25.28, −6.11*)	(*−23.82, −7.82*)	(*−23.18, −5.92*)	(*−19.30, 4.38*)
"fourth"	**−17.97**	**−17.25**	−4.07	**18.35**
	(*−29.12, −6.81*)	(*−26.64, −7.86*)	(*−13.90, 5.74*)	(*3.65, 33.05*)
"fifth"	**−35.54**	**−33.41**	**−24.00**	−9.91
	(*−47.59, −23.49*)	(*−43.11, −23.71*)	(*−34.18, −13.81*)	(*−25.56, 5.72*)
mother's work	−1.41	−3.48	−1.25	2.20
reference: "unemployed"	(*−7.81, 4.97*)	(*−9.33, 2.36*)	(*−7.46, 4.95*)	(*−6.37, 10.79*)
mother's religion		*reference: "christian"*		
"hindu"	**−7.96**	−4.91	−2.39	1.05
	(*−15.47, −0.46*)	(*−10.55, 0.72*)	(*−8.62, 3.83*)	(*−9.44, 11.55*)
"muslim"	**31.23**	**24.27**	**26.51**	**37.91**
	(19.14, 43.32)	*(13.01, 35.53)*	*(14.44, 38.59)*	*(22.16, 53.66)*
"sikh"	6.72	5.27	8.51	8.23
	(*−16.30, 29.75*)	(*−11.46, 22.00*)	(*−7.79, 24.82*)	(*−15.94, 32.41*)
"other"	**22.82**	**14.49**	7.77	5.41
	(6.24, 39.41)	*(0.37, 28.61)*	(*−7.15, 22.69*)	(*−15.40, 26.23*)
mother's residence	−1.61	−0.79	2.32	**8.98**
reference: "rural"	(*−8.34, 5.11*)	(*−5.64, 4.04*)	(*−2.44, 7.09*)	(*1.99, 15.97*)
# dead children:		*reference: "0"*		
"1"	−5.62	−2.89	−6.18	−10.43
	(*−12.94, 1.69*)	(*−8.23, 2.45*)	(*−13.05, 0.68*)	(*−21.28, 0.42*)
"2"	−3.80	−1.46	−6.05	−11.91
	(*−16.85, 9.24*)	(*−12.21, 9.28*)	(*−18.92, 6.80*)	(*−32.28, 8.45*)
"3"	−15.94	**−16.05**	−14.93	−16.26
	(*−33.82, 1.92*)	(*−31.88, −0.23*)	(*−35.07, 5.20*)	(*−44.51, 11.99*)
electricity supply	**16.71**	**12.65**	7.73	4.77
reference: "no"	*(9.47, 23.95)*	*(5.80, 19.50)*	(*−0.35, 15.82*)	(*−6.18, 15.72*)
radio	3.47	**4.51**	**5.58**	3.52
reference: "no"	(*−1.80, 8.75*)	*(0.72, 8.31)*	*(1.44, 9.73)*	(*−3.73, 10.78*)
television	**11.49**	**13.35**	**14.80**	**18.77**
reference: "no"	*(4.73, 18.25)*	*(8.57, 18.13)*	*(9.61, 19.99)*	*(10.50, 27.04)*
refrigerator	9.73	**10.63**	**9.29**	**6.93**
reference: "no"	(*−1.18, 20.65*)	*(2.17, 19.09)*	*(0.20, 18.37)*	*(10.50, 27.04)*
bicycle	−3.87	−3.53	**−7.36**	**−8.83**
reference: "no"	(*−8.80, 1.05*)	(*−7.35, 0.28*)	(*−12.60, −2.12*)	(*−16.66, −1.01*)
motorcycle	**11.13**	**9.80**	**9.56**	**11.61**
reference: "no"	*(2.26, 20.00)*	*(2.88, 16.72)*	*(1.91, 17.22)*	*(0.31, 22.90)*
car	−10.55	1.10	4.40	11.09
reference: "no"	(*−38.11, 17.00*)	(*−12.48, 14.69*)	(*−9.45, 18.26*)	(*−14.19, 36.38*)



(a) age of child

(b) duration of breastfeeding

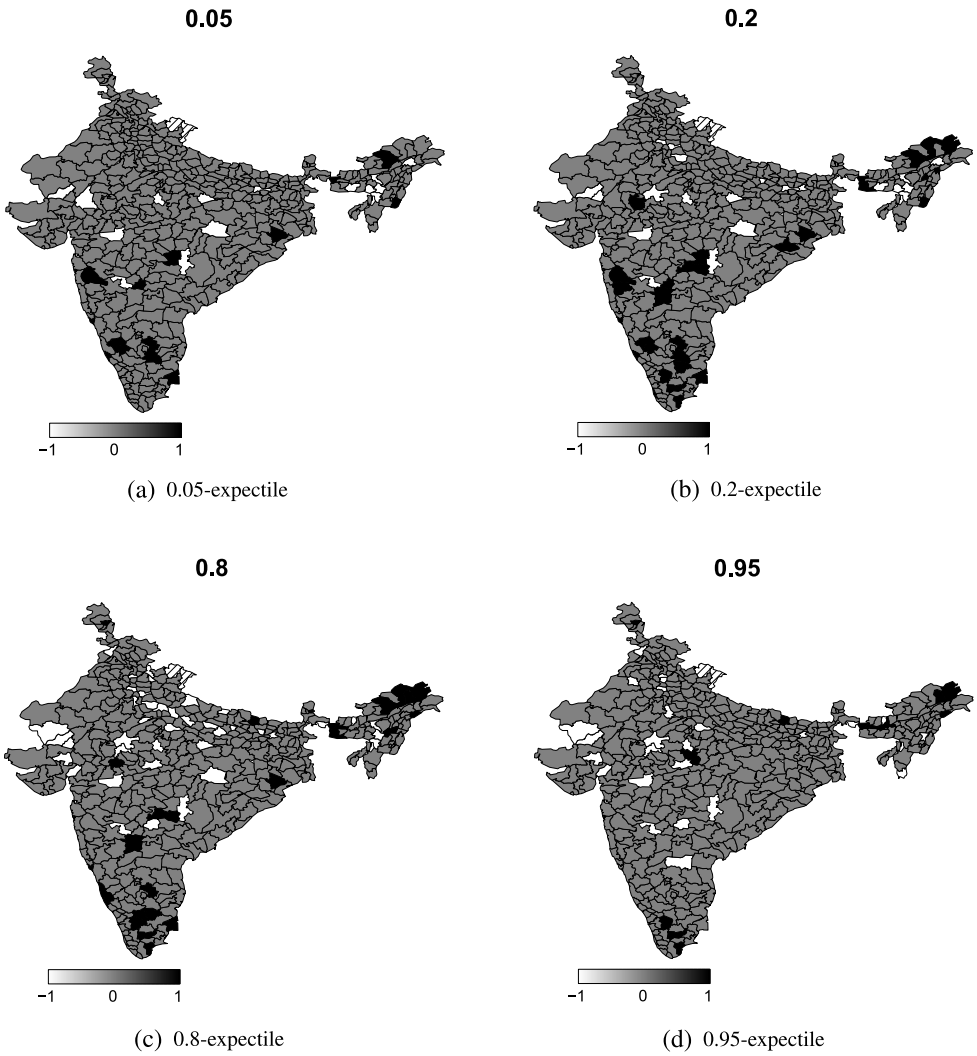
(c) BMI of mother

(d) age of mother

(e) years of mother's education

(f) partner's education

Fig. 3 Estimated nonlinear effects and confidence intervals for the six continuous covariates included in the model for the 0.05, 0.5 and 0.95-expectile



(a) 0.05-expectile (b) 0.2-expectile

(c) 0.8-expectile (d) 0.95-expectile

Fig. 4 Estimated significance indicators for the effects of the Markov random field in the regions of India for four expectiles. *White regions* indicate a significant negative effect on the response while *black regions* indicate a positive effect

effect on the nutritional status of the children for densely populated areas. Those regions are mainly in the northeast along the rivers Ganges and Brahmaputra. In consequence, we can assume a sufficient supply with fresh water for these children. We can also see a relation to the effects of the religion here since most of India's muslims live in the densely populated areas. The inclusion of an interaction term could be part of further research. In the additive model considered here, an increased correlation between two covariates will just result in wider confidence intervals. The effects nevertheless show us that muslim children suffer from stunting less than children from the other religions. This observation can be made throughout all expectiles and stands in contrast to the results from Fenske et al. (2011) whose results indicated no difference between the five religions. This might be due to the fact that no spatial effect could be included in the quantile regression model. They also performed variable selection in the quantile regression which led to the elimination of the television indicator variable from their model. The expectiles, however, show that the presence of a TV in a household is an indicator for less stunting. The reason for this result is probably that the whole family is provided with food before the money is spent on a TV. So we can take this variable as an indicator for wealth. Not yet mentioned was the positive influence of the presence of a motorcycle or a refrigerator to the stunting score.

From the six continuous covariates included in the model and shown in Fig. 3 we see that up to an age of two years the stunting gets worse and after that there's a consolidation. The remaining five continuous effects present less drastic changes along the covariates than the quantiles portrayed. For increasing age, BMI and years of education of the mother we observe a slight increase in the stunting score. Comparing both the education of the mother and of her partner we make the same observation as Fenske et al. (2011). The education of the partner is less important for the nutritional status of the child. For all continuous variables we can see a homoscedastic behaviour as the expectiles are almost parallel throughout the support of the covariates. Also we can conclude from the expectiles that the conditional distribution of the stunting score is right skewed. Further, the variation in the response is substantial. This leads to wide

confidence intervals to all expectiles even with the large amount of observations. The latter is nevertheless important for the high smoothness of the expectile curves. The analyses demonstrate several indicators that are associated with malnutrition in India. But we can see from the lower expectiles in Fig. 4, severe malnutrition can be found anywhere in India.

5 Conclusion

In this paper, we derived the asymptotic results supplementing the point estimators for geoadditive expectiles. The asymptotic normality of the LAWS method as well as the subsequent confidence intervals are an essential extension to the estimation methods introduced e.g.in Sobotka and Kneib (2010). Our simulations and the application to the malnutrition data have shown us that we can safely replace the computationally expensive method of the bootstrap with the usage of the asymptotic properties. As Fig. 2 has shown, both methods provide similar coverage for growing sample sizes.

Generally, we need to recollect that the advantages of expectile regression over mean regression can be exploited solely when regarding a set of expectiles. As seen in Sect. 4.2, by comparing different expectiles we gain information about the distribution of the response. The introduced confidence intervals help us by signifying the strength of the results. The data analysis has also shown that the obtained information is comparable to a quantile regression despite reduced interpretability. Hence, we use expectiles and gain computational advantages and flexible geoadditive models.

Acknowledgements We thank two anonymous referees and a coordinating editor for valuable comments that lead to considerable improvement upon the original version of this paper. Financial support from the German Research Foundation (DFG) grant KN 922/4-1 is gratefully acknowledged.

Appendix: Proof of Asymptotic Normality

Proof Note first that

$$\begin{aligned}&\sum_{i=1}^{n} w_{i,\tau}\big(\hat{\boldsymbol{\beta}}_\tau^0\big)\boldsymbol{x}_i\big(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\big)\\&\quad=\sum_{i=1}^{n} w_{i,\tau}\big(\boldsymbol{\beta}_\tau^0\big)\boldsymbol{x}_i\big(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\big)\\&\qquad+\sum_{i=1}^{n}\big(w_{i,\tau}\big(\hat{\boldsymbol{\beta}}_\tau^0\big)-w_{i,\tau}\big(\boldsymbol{\beta}_\tau^0\big)\big)\boldsymbol{x}_i\big(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\big)\end{aligned}\tag{20}$$

with

$$w_{i,\tau}\big(\hat{\boldsymbol{\beta}}_\tau^0\big)-w_{i,\tau}\big(\boldsymbol{\beta}_\tau^0\big)=\begin{cases}0, & \text{for } y_i\geq\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\\ & \text{and } y_i\geq\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\\ \tau-(1-\tau), & \text{for } y_i\geq\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\\ & \text{and } y_i<\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\\ (1-\tau)-\tau, & \text{for } y_i<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\\ & \text{and } y_i\geq\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\\ 0, & \text{for } y_i<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_\tau^0\\ & \text{and } y_i<\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0.\end{cases}$$

Since $\hat{\boldsymbol{\beta}}_\tau^0-\boldsymbol{\beta}_\tau^0=O_p(n^{-1/2})$, the last component in (20) is of ignorable asymptotic order $O_p(1)$ (while the other component is of order $O_p(n^{1/2})$). Following the same line of arguments, we can now derive the asymptotic properties for the final estimate

$$\begin{aligned}&\hat{\boldsymbol{\beta}}_\tau^0-\boldsymbol{\beta}_\tau^0\\&\quad=\left(\sum_{i=1}^{n} w_{i,\tau}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} w_{i,\tau}(\hat{\boldsymbol{\beta}}_\tau)\boldsymbol{x}_i\big(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big)\right)\\&\quad=\left(\sum_{i=1}^{n} \varpi_{i,\tau}^0\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \varpi_{i,\tau}^0\boldsymbol{x}_i\big(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_\tau^0\big)\right)\\&\qquad+O_p\big(n^{-1}\big)\end{aligned}$$

and therefore $\hat{\boldsymbol{\beta}}_\tau\overset{a}{\sim}N(\boldsymbol{\beta}_\tau^0,\mathrm{Var}(\hat{\boldsymbol{\beta}}_\tau^0))$. □

References

Breckling, J., Chambers, R.: M-quantiles. Biometrika **75**, 761–771 (1988)

Buchinsky, M.: Recent advances in quantile regression models: a practical guideline for empirical research. J. Hum. Resour. **33**, 88–126 (1998)

Efron, B.: Regression percentiles using asymmetric squared error loss. Stat. Sin. **1**, 93–125 (1991)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap, 1st edn. Chapman and Hall, New York (1993)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**, 89–121 (1996)

Fahrmeir, L., Kneib, T., Lang, S.: Penalized structured additive regression: a Bayesian perspective. Stat. Sin. **14**, 731–761 (2004)

Fenske, N., Kneib, T., Hothorn, T.: Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. J. Am. Stat. Assoc. **106**(494), 494–510 (2011)

Jones, M.: Expectiles and m-quantiles are quantiles. Stat. Probab. Lett. **20**(2), 149–153 (1994)

Kammann, E.E., Wand, M.P.: Geoadditive models. Appl. Stat. **52**, 1–18 (2003)

Kocherginsky, M., He, X., Mu, Y.: Practical confidence intervals for regression quantiles. J. Comput. Graph. Stat. **14**, 41–55 (2005)

Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)

Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–50 (1978)

Meinshausen, N., Bühlmann, P.: Stability selection. J. R. Stat. Soc., Ser. B **72**(4) (2010, in press), with discussion. doi:10.1111/j.1467-9868.2010.00740.x

Newey, W.K., Powell, J.L.: Asymmetric least squares estimation and testing. Econometrica **55**, 819–847 (1987)

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

Rigby, R., Stasinopoulos, D.: Generalized additive models for location, scale and shape. Appl. Stat. **54**, 507–554 (2005)

Rue, H., Held, L.: Gaussian Markov Random Fields. Chapman & Hall/CRC, Boca Raton (2005)

Schall, R.: Estimation in generalized linear models with random effects. Biometrika **78**, 719–727 (1991)

Schnabel, S., Eilers, P.: Optimal expectile smoothing. Comput. Stat. Data Anal. **53**, 4168–4177 (2009)

Sobotka, F., Kneib, T.: Geoadditive expectile regression. Comput. Stat. Data Anal. (2010). doi:10.1016/j.csda.2010.11.015

Sobotka, F., Schnabel, S., Schulze Waltrup, L.: expectreg: Expectile and Quantile Regression. http://CRAN.R-project.org/package=expectreg, *r* package version 0.26 (2011)