Received: 29 June 2020 | Revised: 6 December 2020 | Accepted: 20 January 2021

DOI: 10.1002/bimj.202000203

Biometrical Journal

RESEARCH PAPER

Generalized expectile regression with flexible response function

Elmar Spiegel[1,2] | **Thomas Kneib**[2] | **Petra von Gablenz**[3] | **Fabian Otto-Sobotka**[4]

[1] Helmholtz Zentrum München GmbH, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

[2] University of Goettingen, Chair of Statistics, Göttingen, Germany

[3] Jade University of Applied Sciences, Institute for Hearing Technology and Audiology, Oldenburg, Germany

[4] Carl von Ossietzky University Oldenburg, Division of Epidemiology and Biometry, Oldenburg, Germany

Correspondence
Elmar Spiegel, Helmholtz Zentrum München GmbH, German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany.
Email: elmar.spiegel@helmholtz-muenchen.de

Funding information
Research Association of German Hearing Aid Acousticians; Federal Resources of Niedersächsisches Vorab, Grant/Award Number: Hearing in everyday life (HALLO); European Regional Funding; Lower Saxony Department of Science and Culture; Deutsche Forschungsgemeinschaft, Grant/Award Number: KN 922/4-2

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Expectile regression, in contrast to classical linear regression, allows for heteroscedasticity and omits a parametric specification of the underlying distribution. This model class can be seen as a quantile-like generalization of least squares regression. Similarly as in quantile regression, the whole distribution can be modeled with expectiles, while still offering the same flexibility in the use of semiparametric predictors as modern mean regression. However, even with no parametric assumption for the distribution of the response in expectile regression, the model is still constructed with a linear relationship between the fitted value and the predictor. If the true underlying relationship is nonlinear then severe biases can be observed in the parameter estimates as well as in quantities derived from them such as model predictions. We observed this problem during the analysis of the distribution of a self-reported hearing score with limited range. Classical expectile regression should in theory adhere to these constraints, however, we observed predictions that exceeded the maximum score. We propose to include a response function between the fitted value and the predictor similarly as in generalized linear models. However, including a fixed response function would imply an assumption on the shape of the underlying distribution function. Such assumptions would be counterintuitive in expectile regression. Therefore, we propose to estimate the response function jointly with the covariate effects. We design the response function as a monotonically increasing P-spline, which may also contain constraints on the target set. This results in valid estimates for a self-reported listening effort score through nonlinear estimates of the response function. We observed strong associations with the speech reception threshold.

KEYWORDS
distributional regression, generalized additive models, monotonicity constraints, P-spline, single index models

1 | INTRODUCTION

In classical linear regression, we model the expected value μ_i for a response variable y_i based on covariates $\boldsymbol{x}_i$ via the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ as

$$\mathrm{E}(y_i) = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

with $\boldsymbol{\beta}$ the vector of coefficients. However, in a linear model we often assume a specific distribution of the response and homoscedasticity of the error terms. Expectile regression (Newey & Powell, 1987) is one possible model to overcome these restrictions. With expectiles we do not only model the expected value of the response, but the whole distribution. Therefore, several models are defined, where emphasis is placed on different parts of the distribution. Overall, an expectile is a weighted mean, where the weights depend on the responses, the fitted values, and the current asymmetry level $\tau \in (0, 1)$, where a value of $\tau = 0.5$ results in the expectation/arithmetic mean. The idea of asymmetrically weighting the contributions to the least squares criterion is conceptually close to the notion of asymmetric maximum likelihood (Efron, 1992) but has the advantage to avoid an explicit distributional assumption.

More formally, an expectile e_τ is defined as the solution of

$$\underset{e_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i)(y_i - e_\tau)^2$$

with weight

$$w_\tau(y_i) = \begin{cases} \tau & \text{if} \quad y_i \geq e_\tau \\ 1 - \tau & \text{if} \quad y_i < e_\tau \end{cases}.$$

Similarly, as with the expected value in linear regression, expectiles can be used to set up expectile regression. We model the expectile as a linear combination

$$e_{i,\tau} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau,$$

with $\boldsymbol{\beta}_\tau$ the coefficient depending on the chosen asymmetry τ. Since expectiles are weighted means, the restriction on linear covariate effects can be relaxed as in the linear model by using semiparametric predictors like splines or Gaussian Markov random fields. This model class is called semiparametric expectile regression and was introduced by Schnabel and Eilers (2009) and Sobotka and Kneib (2012). As expectiles use the least squares framework, it is easier to include smooth effects and complex covariate structures with quadratic penalties in expectile regression than in quantile regression (Koenker & Bassett, 1978). Expectiles also relate to the expected shortfall, a very intuitive tail expectation measure. Hence, we chose expectile regression for this paper over quantile regression. Formally expectiles generalize the mean in a similar way as quantiles do for the median. More details on semiparametric expectile regression are given in Section 3.

Even if semiparametric expectile regression is able to capture nonlinear effects of the covariates, it still assumes that the relationship between the predictor and the expectile is linear. In particular we assume that the underlying link function is the identity link. If the true underlying relationship is nonlinear, then severe biases can be observed in the parameter estimates as well as in quantities derived from them such as model predictions. While modeling the distribution of scores derived from a questionnaire on hearing abilities with expectiles, we observed the problem that expectiles in the upper tail were fitted to larger values than the maximum of the hearing ability score considered in the questionnaire (von Gablenz et al., 2018). In theory, expectiles should only be fitted to values inside the support of the response. However, due to numerical errors and the automatic selection of smoothing parameters for nonlinear effects, some of the estimates went beyond the maximum. The aim of our analysis was to model associations of sociodemographic covariates as

well as results from a diagnostic hearing test and scores from the self-administered questionnaire on hearing abilities. We expected a strong heteroscedasticity in the distribution and aimed to show the different associations in the upper and lower tail of the hearing score distribution. However, without the possibility to cover restrictions to the target set, expectile regression may give useless outputs (Rigby et al., 2013). Another application where constraints on the target set are useful, is when analyzing values with natural lower limit in zero and a distribution that is close to this boundary. As an illustration for such behavior, we analyze the concentration of mercury in blood samples. Without restricting the fitted values to the target set, frequently negative (i.e., impossible) predictions occur (see Section 6.2 and in particular Figure 6).

So a similar trick as in generalized linear models (McCullagh & Nelder, 1989) could be applied in expectile regression to fulfill both constraints, the limits of the response's support and the nonlinearity between the fitted value and the predictor. In a generalized linear model a prespecified response function defines the relationship between the predictor and the fitted value and also ensures the constraints on the target set. In a generalized linear or additive model the expected value is defined as

$$\mathrm{E}(y_i) = \mu_i = h(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where h is the predefined response function. Though h will result in biased estimates, if the choice of h does not correspond to the underlying true response function as it has been discussed by Czado and Santner (1992). Single index (SI) models as introduced by Ichimura (1993) are often applied to remove the assumptions based on the specification of the response function. In SI models, kernel density estimates are used to determine the response function. Others like Klein and Spady (1993), Weisberg and Welsh (1994), Carroll et al. (1997), and Wang et al. (2010) developed further ideas and applied them to the partially linear SI framework. Several articles including Wu et al. (2010) and Jiang and Qian (2016) provide ideas on how to include SI models in quantile regression. Moreover, all discussed SI models rely on kernel methods, which regularly result in wiggly estimates. One approach to obtain smooth response functions is to construct a monotonic P-spline basis to estimate the response function, as Yu and Ruppert (2002) and Muggeo and Ferrara (2008) suggest. Though these models still rely on linear predictor structures. Consequently the idea of combining generalized additive models with SI models might be useful to reduce the bias of assuming a fixed response function or the bias from the linear predictor structure. This idea has been discussed by Marx (2015), Tutz and Petry (2016), and Spiegel et al. (2019). They all use P-splines to estimate the response function, but apply different tools to generate the estimates. Overall, we modeled the expected value as

$$\mu_i = \tilde{h}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

with a semiparametric predictor and P-splines for the response function $\tilde{h}$. Note that appropriate restrictions are required in this setup to render the model identifiable.

The aim of this paper is to introduce a new type of expectile regression in which we are able to consider constraints on the target set. Therefore we include a response function in the model definition. Thus, we introduce the new generalized expectile as

$$e_\tau = \tilde{h}_\tau(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau),$$

where $\tilde{h}_\tau$ is the flexible response function. The idea of using a fixed response function in expectile regression would imply the same assumptions as in generalized linear models on the shape of the distribution. We generally want to avoid such an assumption in expectile regression. Therefore, we propose to estimate the response function jointly with the (semiparametric) predictor similarly as in our previous work on generalized additive models with flexible response function (Spiegel et al., 2019). So $\tilde{h}_\tau$ will be determined as a monotonic P-spline while the predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau$ may also be semiparametric, comprising splines as well as spatial information, etc. We still estimate the coefficients based on the least asymmetric weighted squares as

$$\underset{\boldsymbol{\beta}_\tau, \tilde{h}_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i)\left(y_i - \tilde{h}_\tau(\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau)\right)^2 + \text{Penalty}$$

including an appropriate quadratic penalty for the smooth components. These smooth components are on the one hand the estimated response function, but on the other hand the semiparametric covariate effects. The exact definition of the penalty is described in Section 4.2. As denoted in the index τ, both the response function and the covariate effects depend

on the asymmetry level τ. We define this dependence to avoid distributional assumptions and to capture heteroscedasticity in both parts. Moreover, we will include shifts and scalings of the predictor later on to ensure identifiability, see Section 4.3 for details. This includes the omission of an intercept. However, these constraints prevent that we can transfer one (overall) response function to specific asymmetry levels.

In summary, we propose an approach in which all three assumptions on generalized linear models—linearity of the predictor, fixed response functions, and underlying distribution function—are relaxed jointly. Therefore, we combine generalized additive models with single index models and semiparametric expectile regression.

To explain generalized expectile regression properly, we need to first recap some established methods. Thus, in Section 2 we will sum up the ideas of splines and additive models. The basic theory on classical expectile regression is shown in Section 3. Afterward, we introduce our new approach on generalized expectile regression in Section 4. Next, we compare the behavior of classical expectile regression and our new approach in a simulation study in Section 5. Finally, in Section 6.1, we use our approach in the analysis of self-reported hearing scores in a population sample from northern Germany. There, we also illustrate in detail how to interpret this complex setting. The analysis of mercury concentrations in blood samples of the general adult population of the United States is further described in Section 6.2. Section 7 concludes the paper with a discussion of our findings.

2 | SPLINES AND ADDITIVE MODELS

In many cases the assumptions of linear covariate effects are not fulfilled. Therefore, several ideas on how to deal with nonlinear effects have been introduced. One of them are spline bases (Duchon, 1977). We will use the concept of splines not only for estimating smooth covariate effects, but also for estimating the smooth response function. The methods for determining a smooth response function and smooth covariate effects are similar, thus we explain the concept of splines only based on covariate effects. Later we will describe the concept of estimating the response function in more detail.

Let us assume that the effect of the covariate x is not linear, but any smooth function. In order to approximate an unknown smooth functional effect $s(x)$ with a spline basis, first several basis functions $B_j^{(l)}(x)$, for example, B-spline basis functions, are evaluated at the observed covariate values x_i. The degree $l > 0$ of the basis functions is chosen in advance. Then these evaluated basis functions are scaled by regression coefficients $\tilde{\gamma}_j$ and summed up to be the representation of the smooth effect $s(x)$ such that the estimated approximation of the function can be written as

$$s(x) = \sum_{j=1}^{J} B_j^{(l)}(x)\tilde{\gamma}_j.$$

For a classical spline basis, the number of basis functions J and the position of these basis functions need to be optimized to obtain a suitable estimate. Alternatively, a much larger set of basis functions can be constructed to achieve a flexible curve and a penalty can be added that ensures a smooth estimate by penalizing differences between neighboring coefficients. This results in P-splines (Eilers & Marx, 1996). The penalty is usually based on the second-order differences between the neighboring coefficients

$$\lambda \sum_{j=3}^{J} (\tilde{\gamma}_j - 2\tilde{\gamma}_{j-1} + \tilde{\gamma}_{j-2})^2.$$

Moreover, the penalty can be expressed as a quadratic form of the coefficients and a penalty matrix $\tilde{\boldsymbol{K}}$ as $\tilde{\boldsymbol{\gamma}}^\top \tilde{\boldsymbol{K}} \tilde{\boldsymbol{\gamma}}$. In the following, we will include the smoothing parameter $\lambda \geq 0$ into the penalty matrix $\tilde{\boldsymbol{K}}$, to simplify the notation. For estimation, the evaluated basis functions are treated as covariates and the coefficients are the solution of minimizing a penalized least squares criterion (for more details, see Wood, 2017).

In the estimation of models with a flexible response function, we need the derivative of P-splines. Based on the local polynomial structure of B-splines, the calculation of the derivative is rather simple (Wood, 2017). Thus, we can obtain the derivative as

$$s'(x) = \frac{\mathrm{d}s(x)}{\mathrm{d}x} = (l-1) \sum_{j=1}^{J} \left(\frac{B_j^{(l-1)}(x)}{\kappa_{j+l-1} - \kappa_j} - \frac{B_{j+1}^{(l-1)}(x)}{\kappa_{j+l} - \kappa_{j+1}} \right) \tilde{\gamma}_j,$$

where $B_j^{(l-1)}$ are the basis functions at the same knots κ_j as for the estimation of $s(x)$, but of one degree less. To simplify the notation, we skip the degree of the basis functions in the following.

Another characteristic of the local polynomial structure of B-splines is that monotonicity constraints and constraints on the target set can be transferred to constraints of the coefficients. Thus, we get a monotonically increasing B-spline by ensuring that the coefficients are monotonically increasing. Moreover, B-spline basis functions are nonnegative. Positive coefficients $\tilde{\boldsymbol{\gamma}}$ induce a spline with only positive values. Finally, constraints mentioned in the previous section can be imposed on B-splines by recognizing that the sum of the unscaled basis functions is 1, such that coefficients $\tilde{\boldsymbol{\gamma}} < 1$ imply that $s(x)$ is smaller than 1. All these constraints are linear inequality constraints on the coefficients and can be included in the estimation as shown by Wood (1994).

To ease the notation we combine the covariate effects in a design matrix $\boldsymbol{Z}$. In a semiparametric or additive model, several smooth effects are combined and also classical linear or categorical effects can be attached to the predictor. The design matrix $\boldsymbol{Z}$ consists of covariates and dummies as in an ordinary linear model and evaluated basis functions for smooth effects. So the row of $\boldsymbol{Z}$ corresponding to the i-th observation is defined as

$$\boldsymbol{z}_i^\top = (1, \dots, x_{iq}, \dots, B_1(x_{ip}), \dots, B_{J_p}(x_{ip}), \dots),$$

where covariate x_q has a linear effect on the response while covariate x_p is modeled to have a smooth effect. Here, J_p is the number of basis functions for the smooth effect $s_p(x_p)$. The coefficients are similarly concatenated in one vector $\boldsymbol{\gamma}$. Furthermore, the penalty matrices are combined to one block diagonal matrix $\boldsymbol{K}$, which has dimensions corresponding to $\boldsymbol{\gamma}$. Therefore, unpenalized coefficients correspond to 0 values in the penalty matrix. In our notation the smoothing parameter λ_p of the smooth effect $s_p(x_p)$ is part of the corresponding penalty matrix $\boldsymbol{K}_p$. However, we also discuss the set of smoothing parameters in a jointed way as vector $\boldsymbol{\lambda}$.

In summary, we model the expected value in a linear model with semiparametric predictor as

$$\mathrm{E}(y_i) = \mu_i = \gamma_0 + \cdots + x_{iq}\gamma_q + \cdots + s_p(x_{ip}) + \cdots = \boldsymbol{z}_i^\top \boldsymbol{\gamma}.$$

For the optimal predictions we minimize the penalized least squares criterion

$$\mathrm{PLS}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma})^2 + \boldsymbol{\gamma}^\top \boldsymbol{K} \boldsymbol{\gamma}$$

via

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{K})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}.$$

The smoothing parameters $\boldsymbol{\lambda}$, which are included in the penalty matrix $\boldsymbol{K}$ need to be optimized additionally to the model estimation. Therefore, cross-validation or model selection criteria as generalized cross-validation criterion (GCV) or Akaike information criterion (AIC) can be applied.

3 | EXPECTILE REGRESSION

In a linear model, we assume that the response follows a specific distribution and the error terms are homoscedastic. One approach to overcome these restrictions is expectile regression (Newey & Powell, 1987). Expectiles allow us to model the whole distribution without assuming a parametric distribution family. Based on the similarity between expectile regression and linear models in their estimation, all smooth covariate structures of the latter can also be used in expectile regression. Then we are talking about semiparametric expectile regression as introduced by Schnabel and Eilers (2009) and Sobotka and Kneib (2012). A weight is included in the standard penalized least squares approach when estimating expectile regression such that now the least asymmetrically weighted squares (LAWS) criterion is optimized

$$\mathrm{LAWS}(\boldsymbol{\gamma}_\tau) = \sum_{i=1}^{n} w_\tau(y_i)\left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau\right)^2 + \boldsymbol{\gamma}_\tau^\top \boldsymbol{K} \boldsymbol{\gamma}_\tau. \tag{1}$$

Here the weight $w_\tau(y_i)$ depends on the chosen asymmetry $\tau \in (0,1)$, the response y_i, and the fitted value $\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau$ via

$$w_\tau(y_i) = \begin{cases} \tau & \text{if} \quad y_i \geq \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau \\ 1-\tau & \text{if} \quad y_i < \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau \end{cases}.$$

The minimizer of the LAWS criterion defines the expectile

$$e_{i,\tau} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau.$$

For the calculation of the coefficients $\boldsymbol{\gamma}_\tau$ the interdependence between the weight and the coefficients needs to be solved. Therefore, an iteratively reweighted least squares (IRLS) procedure is applied.

The smoothing parameters $\boldsymbol{\lambda}$ are optimized from outside the LAWS algorithm based on model selection criteria. This means we fix the first set of smoothing parameters and run the expectile regression based on these parameters. Then we evaluate the goodness-of-fit and iterate to the next set of smoothing parameters. In this article we use 5-fold cross-validation with sum of predictive asymmetric weighted squared errors as criterion for all models to have a fair comparison. The optimization procedure of the smoothing parameters starts with a small grid search whose output is used as initial values for a Nelder–Mead optimization. Details on further approaches on optimizing smoothing parameters are described in Sobotka and Kneib (2012).

For the interpretation of expectiles, the comparison with the ordinary mean is useful. The mean specifies the center of gravity to the point where the sum of the distances between the fitted value and the observations located above that point is equal to the sum of distances located below that point. For expectiles we now put more emphasis on a specific part of the distribution by introducing the asymmetry τ. For the derivative of Equation (1) we get

$$\sum_{i:\ y_i < \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau} (1-\tau)(y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau) = \sum_{i:\ y_i \geq \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau} \tau(y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau). \tag{2}$$

So $\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau$ determines the point where the weighted sum of distances located above is equal to the weighted sum of distances located below. Based on Equation (2) we see that the expectile is the weighted center of gravity. Rearranging Equation (2) results in

$$\tau = \frac{\sum_{i:\ y_i < \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau} |y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau|}{\sum_{i=1}^{n} |y_i - \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau|}.$$

Thus, the expectile defines the point where the fraction of distances below the predictor is τ. This is in contrast to quantile regression where τ is the fraction of the number of observations located below the fitted value. So quantiles are a generalization of the median while expectiles are a generalization of the mean. Thus, expectile regression is, in practice, a weighted least squares idea, with asymmetric weights. Due to this, expectile regression can easily comprise the complex covariate structures that are available for the linear models.

4 | EXPECTILE REGRESSION WITH FLEXIBLE RESPONSE FUNCTION

4.1 | Introduction to generalized expectiles

As discussed in the introduction, we generalize the classical semiparametric expectile regression to allow for nonlinear relationships between the predictor and the expectile and to consider constraints on the support of the response. Therefore, we introduce generalized expectiles or expectiles with flexible response function as

$$e_{i,\tau} = \tilde{h}_\tau(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$$

with $\tilde{h}_\tau$ a strictly monotonically increasing function, that may be constrained by the support of the response variable. Basically, a fixed response function h could be used to consider the constraints on the support of the response. This would be in line with classical generalized linear (additive) models, see, for example, Efron (1992) where asymmetric maximum likelihood estimation is combined with a fixed response function to study the distribution of count responses. However, these models would be dependent on specifying the underlying distribution correctly. We provide more details on this in Section 4.2. In the following, we always mean generalized expectile regression with flexible response function, when speaking of generalized expectiles if not specified otherwise. The version with fixed response function is always indicated as such.

So here we emphasize to estimate, similarly as in generalized additive models with flexible response function (Spiegel et al., 2019), the function $\tilde{h}$ using a P-spline basis

$$\tilde{h}_\tau(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau) = \sum_{j_h=1}^{J_h} B_{j_h}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau)\nu_{j_h,\tau}$$

with B_{j_h} $(j_h = 1, \dots, J_h)$ B-spline basis functions. We impose monotonicity on $\tilde{h}$ by forcing monotonically increasing coefficients $\boldsymbol{\nu}_\tau = (\nu_{1,\tau}, \dots, \nu_{J_h,\tau})$, which can be written as a linear constraint on the model. The constraints on the support can also be described in terms of linear constraints on the coefficients $\boldsymbol{\nu}_\tau$. We restrict the coefficients to be positive only in order to achieve only nonnegative predictions. Based on the fact that B-spline basis functions are always nonnegative, the combination will result in nonnegative predictions. Additionally, constraints from above could also be included as linear constraints. Due to the fact that unweighted B-splines basis functions sum up to 1, we only allow for coefficients smaller than 1. This results in a response function, which is always smaller or equal to 1. Once we apply the linear constraints to the coefficients, the response function can be estimated based on methods of Wood (1994) for fixed predictors $\eta_{i\tau} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau$.

The response function $\tilde{h}_\tau$ may depend on the asymmetry in a similar way as the coefficients of the covariate effects $\boldsymbol{\gamma}_\tau$. We construct one response function per expectile to be able to detect heteroscedasticity in both parts, the predictor and the response function. We will show the flexibility in a scenario in the simulation study.

Generally, using one estimated response function for all asymmetry levels would ease the interpretation of the results. However, the estimation is nontrivial. First, the scaling of the predictor prevents the use of one response function estimated jointly for all asymmetry levels (see next sections for details on the scaling due to identifiability). Since we need for example to cancel the intercept. Second, to get one response function that is valid for all parts of the distribution, we would first need to run flexible expectile regression models for all different asymmetry levels before we could aggregate the different response functions. Figure 5 provides some impression that the estimated response functions might vary strongly between the asymmetry levels. So just using, for example, the estimated response function of the 50% expectile will not be sufficient for the lower or upper tail of the distribution. Furthermore, Figure 5 also shows that it is nontrivial to combine the different response functions. Therefore, we leave this approach to further research.

4.2 | Estimation of generalized expectiles

For the estimation we set up the new generalized least asymmetrically weighted squares (GLAWS) as

$$\text{GLAWS}(\boldsymbol{\gamma}_\tau, \boldsymbol{\nu}_\tau) = \sum_{i=1}^{n} w_\tau(y_i)\big(y_i - \tilde{h}_\tau(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau)\big)^2 + \boldsymbol{\gamma}_\tau^\top \boldsymbol{K} \boldsymbol{\gamma}_\tau + \boldsymbol{\nu}_\tau^\top \boldsymbol{K}_\nu \boldsymbol{\nu}_\tau,$$

where $\boldsymbol{K}$ comprises the penalty matrices for the covariate effects and $\boldsymbol{K}_\nu$ is the penalty matrix for the smooth response function. Both penalty matrices include their smoothing parameters. Similarly as in generalized additive models with flexible response functions, the interdependence of the response function and the predictor prevents the existence of an analytical solution. Thus, we need to adopt the iterative procedure from Spiegel et al. (2019) for the determination of estimates for generalized additive models with flexible response function to expectile regression. In the new procedure we additionally need to adjust the asymmetry weights in each step. Informally we define a pseudo log-likelihood $l(\boldsymbol{\gamma}_\tau, \boldsymbol{\nu}_\tau)$ based on the GLAWS as

$$l(\boldsymbol{\gamma}_\tau, \boldsymbol{\nu}_\tau) = -\frac{1}{2}\text{GLAWS}(\boldsymbol{\gamma}_\tau, \boldsymbol{\nu}_\tau).$$

Derivatives of this pseudo log-likelihood define pseudo score functions and pseudo Fisher information matrices. Details on the pseudo functions are displayed in the supplementary material in Section A. We then calculate working weights and working responses for the modified Fisher scoring algorithm from these pseudo functions.

We start with the estimation of a classical expectile regression model. This yields the first (scaled) predictor. Using this predictor we can estimate the first response function $\tilde{h}^{(m)}$ as a smooth effect in a classical expectile regression (the predictor of the initial expectile regression). This smooth effect includes the constraints on the support of the response and the monotonicity. The estimation of the response function also needs several iterations to account for the asymmetry. Based on this estimated response function we can build a Fisher-Scoring–type algorithm to estimate the new covariate effects and cycle back to estimate a new response function. This iteration continues until the estimate of the response function converges or the goodness-of-fit does not increase anymore. The Fisher-Scoring–type algorithm uses the working response $y_i^{(k)}$ and working weights $\omega_i^{(k)}$

$$y_i^{(k)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)} + \frac{y_i - \tilde{h}_\tau^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})}{\tilde{h}_\tau'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})},$$

$$\omega_i^{(k)} = w_\tau(y_i) \cdot \left(\tilde{h}_\tau'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})\right)^2,$$

where $w_\tau(y_i)$ is the weight depending on the asymmetry and the fitted value via

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \tilde{h}_\tau^{(m)}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)}\right) \\ (1-\tau) & \text{if } y_i \leq \tilde{h}_\tau^{(m)}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)}\right) \end{cases}.$$

Furthermore $\boldsymbol{\gamma}_\tau^{(k-1)}$ are the coefficients of the previous iteration and $\tilde{h}'$ is the derivative of the response function with respect to $\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau$. The new coefficients are then estimated as

$$\boldsymbol{\Omega}^{(k)} = \text{diag}(\omega_i^{(k)}),$$

$$\boldsymbol{\gamma}_\tau^{(k)} = \left(\boldsymbol{Z}^\top \boldsymbol{\Omega}^{(k)} \boldsymbol{Z} + \boldsymbol{K}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{\Omega}^{(k)} \boldsymbol{y}^{(k)}.$$

In each step the asymmetry weights, the working response, and the working weights need to be updated and the coefficients must be scaled as discussed in the next section. A detailed scheme for the estimation is presented in the Appendix. For the estimation we use the software `R` (R Core Team, 2019). In detail we built an `R`-package called `FlexER` (see description of the supplement material), which internally makes use of the `mgcv`-package (Wood, 2017). In our package the covariate effects are estimated via the `gam` function and the response function considering the constraints with the `pcls` function.

In principle, the estimation of the response function $\tilde{h}_\tau$ can be ignored, and generalized expectile regression with *fixed* response function could be used. In consequence, $\tilde{h}_\tau$ is replaced with a prespecified response function h in all previous equations. This could be the exponential or the inverse of the logit function, for example. So the GLAWS would result in

$$\text{GLAWS}^*(\boldsymbol{\gamma}_\tau) = \sum_{i=1}^{n} w_\tau(y_i)\left(y_i - h(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau)\right)^2 + \boldsymbol{\gamma}_\tau^\top \boldsymbol{K} \boldsymbol{\gamma}_\tau$$

and a similar Fisher-scoring–type algorithm can be implemented. The only difference would be that we could omit the estimation of the response function. Furthermore, the constraints to ensure identifiability as introduced in the next section can be ignored. We also implemented this approach in the package `FlexER`. We thereby detected that generalized expectile regression with *fixed* response function is superior to classical expectile regression when the response function is nonlinear. However, we figured out that the fixed version is similar to the flexible version only in specific settings and it is easy to construct cases where the flexible version is superior to the fixed version. We provide corresponding examples in Section 5 and the supplementary material Section B.

4.3 | Identifiability of generalized expectiles

In our proposed model with flexible response function, the identification of different effects is a major issue. We use similar constraints as in generalized additive models with flexible response functions (Spiegel et al., 2019) to achieve identifiability. There we denote with x_{ir} the i-th observation of the r-th covariate and with $s_{r,\tau}(x_{ir})$ the corresponding effect. We use the index $r \in \{q, p\}$ for the covariate to highlight that the constraints 4 and 5 are used independent whether the covariate is modeled as linear or smooth effect.

1. We require that at least two additive covariate effects are contained in the model specification. For illustration, consider a regression predictor with only one single continuous covariate with nonlinear effect modeled as $s(x)$. If we plug this single effect into the response function $\tilde{h}(\cdot)$, we obtain $\tilde{h}(s(x))$, where changes in $s(x)$ can be offset by changes in $\tilde{h}$. This is not possible for $\tilde{h}(s_1(x_1) + s_2(x_2))$ due to the additive combination of the two effects $s_1(x_1)$ and $s_2(x_2)$.
2. The intercept has to be removed, that is, we restrict $\gamma_0 = 0$. Otherwise, a shift on the x-axis of the response function could be compensated by a corresponding change in the intercept. In expectile regression, different parts of the distribution are usually analyzed. In classical expectile regression this also results in very different intercepts for different expectiles. Those are not available in generalized expectile regression, but their effect is part of the estimated response function.
3. All smooth effects have to be centered around zero, that is, $\sum_{i=1}^{n} s_{p,\tau}(x_{ip}) = 0$. In this way, all the intercepts inherent to the P-spline bases are removed as well such that no constant shift invalidates the identifiability. This is incorporated in the setup of the smooth effects by including a QR decomposition in their basis functions (for details, see Wood, 2017).
4. The predictor has to be scaled. Otherwise, the predictor can be shifted and stretched arbitrarily and the effect can be compensated by a shift or stretching in the response function. We scale each effect $s_{r,\tau}$ by the model's standard deviation. In detail we apply (similarly as in Tutz & Petry, 2016)

$$s_{r,\tau}(x_{ir}) = \frac{s_{r,\tau}(x_{ir})}{\left(\sum_{r} \sum_{i=1}^{n} \left(s_{r,\tau}(x_{ir}) - \frac{1}{n} \sum_{i=1}^{n} s_{r,\tau}(x_{ir}) \right)^2 \right)^{1/2}}.$$

5. We scale the coefficients a second time via

$$s_{r,\tau}(x_{ir}) = \frac{s_{r,\tau}(x_{ir}) - \overline{s_{r,\tau}(x_{ir})}}{\max(\eta_\tau) - \min(\eta_\tau)}$$

   to be able to fix the knots of the spline of the response function. Fixing the knots also speeds up the convergence of the algorithm.
6. The response function has to be monotonically increasing, that is, $\tilde{h}'_\tau(\eta_\tau) > 0$. Otherwise interpretations are not possible.

During the estimation we apply step-halving to prevent estimation steps that do not improve the model fit. Otherwise, we end up with wrong solutions or the estimates change back and forth without converging. This problem is well known from the ordinary generalized linear model without the canonical response function (see, e.g., Jørgensen, 1984, Marschner, 2011, or Yu et al., 2017). The smoothing parameters are optimized from outside the algorithm based on true cross-validation with test and training data sets. The usage of the GCV criterion is not possible, since we lack the definition of the effective degrees of freedom in this complex interdependent two-stage model definition. The problem is similar to what we specify in Spiegel et al. (2019).

In classical expectile regression, similar to quantile regression, the crossing of expectiles is a well-known problem (Schulze-Waltrup et al., 2015). Generally, crossing expectiles are hard to interpret, since in theory they should not appear. In practice, they occur due to numerical deficits. In classical expectile regression they occur less often than in classical quantile regression. This is based on the fact that expectile regression with its L2-norm uses more information than quantile regression with its L1-norm. In the new approach we cannot exclude the possibility of crossing expectiles. For classical expectile regression there are approaches like expectile sheets (Schulze-Waltrup et al., 2015), which ensure monotonicity between the asymmetry levels. We have to leave the implementation of those ideas to generalized expectile regression for

future research. However, we conclude based on our examples (see Section 6 and supplementary material Sections C and D), that fixing the smoothing parameters for all asymmetry levels to one set of values reduces the risk of crossing expectiles.

For the interpretation of the effects, we need to combine both the response function and the estimated effects. Therefore, we check the effect of one covariate ceteris paribus, that is, keep all other covariates and the response function fixed. Due to the centering of the covariate effects all curves would be overlying. To ease the view on the different asymmetries, we suggest two methods to adjust the covariates effects. In the first approach, we combine the estimate of the covariate effect with the estimate of the response function. In the second approach an asymmetry depending constant, similarly to the classical intercept, is added to the pure covariate effects. Both are applied in Figure 4.

In detail, we display with the first method the effect of the metric covariates of the generalized expectile regression combined with the response function. Thus, we display $\tilde{h}_\tau(s_{r,\tau}(x) + c_{r,\tau})$ with r any covariate. Here $c_{r,\tau}$ is the typical effect of the other covariates. In our examples $c_{r,\tau} \equiv 0$, since all covariates are either determined as splines or as categories. In the general case, so with including linear effects of metric covariates (which we do not recommend), $c_{r,\tau}$ is specific for each covariate r, which can be any kind of covariate (smooth, categorical, linear). In detail we determine $c_{r,\tau}$ via

$$c_{r,\tau} = \sum_{p \in P} \frac{1}{n} \sum_{i=1}^{n} s_{p,\tau}(x_{ip})$$

with P the set of smooth or linear effects, but $r \notin P$. Categorical effects are not considered in P, since we suggest to use their effect of the reference category instead, which is 0. Indeed, if P consists only of smooth effects $c_{r,\tau}$ is 0.

In the second approach the covariate effects $s_{r,\tau}(x)$ are displayed together with another constant: $s_{r,\tau}(x) + \bar{\hat{e}}_\tau$. This constant $\bar{\hat{e}}_\tau$ is the average of all the fitted values $\hat{e}_{i\tau}$:

$$\bar{\hat{e}}_\tau = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{i\tau}.$$

Both approaches help to identify the effects depending on the asymmetry levels. In the second method, the pure covariate effects can be interpreted, while the first approach allows, to directly see the effect on the response. More details on the interpretation of the estimates are given in the analysis of the hearing scores in Section 6.1.

4.4 | Uncertainty quantification and pointwise confidence bands

Due to the complex model set up and the distribution-free approach we are not giving theoretic approximations or asymptotics for the variance of the estimates. The key problem in the estimation of the variance is the missing definition of the effective degrees of freedom. Instead, we apply a nonparametric bootstrap (Efron & Tibshirani, 1994) to consider how sure we can be about our estimates. Therefore, we draw a nonparametric bootstrap sample of the same size randomly with replacement from our original data set. Using the bootstrap sample we estimate the model again. We use the same configurations as for the original model. In particular we use the final smoothing parameters of the original model as fixed smoothing parameters to save computing time. We find that the variation of the coefficients can be estimated with 1000 to 2000 bootstrap replications. Estimating the smoothing parameters in each bootstrap sample would allow to consider the uncertainty of those. However, with the current algorithm optimizing 1000 times the smoothing parameters takes too long. One would need to reduce the number of bootstrap samples to 10–20. Moreover, with flexible smoothing parameters, it cannot be ensured that the estimated curve of the main model without bootstrap is always inside the bootstrap interval.

For the smooth covariate effects and the smooth response function, we applied some additional steps to obtain pointwise confidence bands that may be included in the plots. First, for each covariate a regular grid based on the range of the original data set is defined. We predict all smooth covariate effects on their grid from the bootstrap models in order to calculate their pointwise confidence bands. Thus, for each point on the grid the quantiles of the predicted effects can be used to determine the pointwise confidence bands. For the calculation of the pointwise confidence band for the response function we use a grid based on the linear predictor of the original model. This grid is used to predict the effect of the response function in each bootstrap replication. Again the pointwise quantiles of the fitted values determine the pointwise confidence bands.

5 | SIMULATION STUDY

5.1 | Design of the simulation study

The ideas that we are presenting in this article are rather complex. Thus, the simulation design is complex, too, since it needs to assess several aspects. Similar to classical expectile regression, generalized expectiles are free from an assumption of the underlying distribution. Therefore, we model five types of responses: Normal, Gamma, Beta, Lognormal, and Pseudo-Beta distributed responses. Overall, we want to show that the classical expectile regression is only valid if we assume a linear relationship between the predictor and the expectile. Thus, we present several scenarios with different response functions in these simulations: the identity link (id), the logarithmic link (log), the logit link (logit), and a link function that depends on the asymmetry level, thus a heteroscedasticity inducing link function (ht). The identity link function is the benchmark process to show that the classical expectile regression is a special case of our new approach for Normal and Gamma distributions. The logarithmic link simulation represents the generalization of the classical generalized additive model. To check whether our new approach is identifiable, when both the response function and the coefficients are varying, we included the third scenario where the response function depends on the simulated error terms. The latter are also simulated with heteroscedasticity. The Lognormal and Pseudo-Beta distributions indicate cases where the generalized expectile regression with flexible response function returns similar values as with a prespecified fixed response function.

In total we provide three approaches for Normal distributed data (`id`, `log`, `ht` link), two approaches for Gamma distributed data (`id`, `log` link), one approach for Beta distributed data (`logit` link), one approach for Lognormal distributed data (`id` link), and one approach for Pseudo-Beta distributed data (`logit` link).

The covariates x_1 and x_2 are drawn uniformly from a grid between 0 and 1 with step size 0.05. The predictor for the mean effect $\eta_i^{(\mu)}$ is then constructed as

$$\begin{aligned} s_1(x_{i1}) &= \sin(2\pi x_{i1}) \\ s_2(x_{i2}) &= \exp(3x_{i2})/20 \\ \eta_i^{(\mu)} &= s_1(x_{i1}) + s_2(x_{i2}) \end{aligned}$$

and for the standard deviation as

$$\eta_i^{(\sigma)} = 0.25 + 0.7x_{i1} + 0.5x_{i2}.$$

To simulate *Gamma* distributed responses we calculate the mean μ_i either with the identity or the exponential function similar to the standard deviation σ_i

$$\mu_i = \begin{cases} \eta_i^{(\mu)} + 2.5 \\ \exp(\eta_i^{(\mu)} + 1) \end{cases} \quad \& \quad \sigma_i = \begin{cases} \eta_i^{(\sigma)} & \texttt{id link} \\ \exp(\eta_i^{(\sigma)}) & \texttt{log link} \end{cases}.$$

Those parameters are used to calculate the shape $\alpha_i = \frac{\mu_i^2}{\sigma_i^2}$ and rate $\beta_i = \frac{\mu_i}{\sigma_i^2}$ of the Gamma distribution. Thus, the response y_i is simulated via

$$y_i \sim Ga(\alpha_i, \beta_i).$$

To generate *Normal* distributed responses we first generate the standard deviation

$$\sigma_i = \begin{cases} \eta_i^{(\sigma)} & \text{for } \texttt{id} \text{ or } \texttt{ht} \text{ link} \\ \exp(\eta_i^{(\sigma)}) & \text{for } \texttt{log} \text{ link} \end{cases}.$$

Then we draw the error term $\varepsilon_i \sim N(0, \sigma_i)$ before we generate the mean value via $\mu_i = h(\eta_i^{(\mu)})$ with h either the identity (id), the exponential (log), or the variance dependent response function (ht).

$$\mu_i = \begin{cases} \eta_i^{(\mu)} + 2.5 & \texttt{id link} \\ \exp(\eta_i^{(\mu)} + 1) & \texttt{log link} \\ \begin{cases} \eta_i^{(\mu)} + 1 & \text{for } \varepsilon_i \leq 0 \\ \exp(\eta_i^{(\mu)} + 1) & \text{for } \varepsilon_i > 0 \end{cases} & \texttt{ht link} \end{cases}.$$

Finally, we calculate the response $y_i = \mu_i + \varepsilon_i$.

To simulate *Beta* distributed responses we calculate the mean μ_i with the `logit` function similar to the standard deviation σ_i

$$\mu_i = \frac{\exp(\eta_i^{(\mu)} + 0.6)}{1 + \exp(\eta_i^{(\mu)} + 0.6)} \;\&\; \sigma_i = \frac{\exp(\eta_i^{(\sigma)} - 4)}{1 + \exp(\eta_i^{(\sigma)} - 4)}.$$

Those parameters are used to calculate the shapes $\alpha_i = -\frac{\mu_i(\mu_i^2 - \mu_i + \sigma_i^2)}{\sigma_i^2}$ and $\beta_i = \frac{(\mu_i - 1)(\mu_i^2 - \mu_i + \sigma_i^2)}{\sigma_i^2}$ of the Beta distribution. Thus, the response y_i is simulated via

$$y_i \sim Beta(\alpha_i, \beta_i).$$

To generate *Lognormal* distributed responses we first generate the standard deviation

$$\sigma_i = \eta_i^{(\sigma)}/1.5.$$

Then we draw the error term $\varepsilon_i \sim N(0, \sigma_i)$ before we generate the mean value μ_i via

$$\mu_i = \eta_i^{(\mu)} + 1.$$

Finally, we calculate the response $y_i = \exp((\mu_i + \varepsilon_i)/1.5)$.

To simulate *Pseudo-Beta* distributed responses with the `logit` function we start with simulating the mean μ_i and standard deviation σ_i

$$\mu_i = \eta_i^{(\mu)} + 0.6 \;\&\; \sigma_i = \eta_i^{(\sigma)}.$$

Before we draw normal distributed error terms ε_i

$$\varepsilon_i \sim N(0, \sigma_i).$$

Finally, we calculate the Pseudo-Beta distributed response y_i using the logit link

$$y_i^* = \mu_i + \varepsilon_i,$$
$$y_i = \frac{\exp(y_i^*)}{\exp(y_i^*) + 1}.$$

For more details on the design of the data, please read the attached R-code in the supplementary material.

5.2 | Comparison of classical expectile regression and generalized expectile regression

In the first part of the simulation study we compare classical expectile regression (ER), generalized expectile regression with flexible response function (FlexER), and generalized expectile regression with fixed response function (Fixed_ER) based on their goodness-of-fit. Therefore, we use a sample size of 2000 for training the models. We estimate all types of expectile regression for the asymmetries $\tau = 10\%, 50\%, 90\%$. For each scenario, the simulations are repeated 100 times. We estimate the covariate effects $s_{\cdot,\tau}$ with penalized B-spline bases of order 3 with maximal 15 degrees of freedom. The smoothing parameters are optimized via 5-fold cross-validation. For the evaluation of the simulations we calculate the predictive mean weighted squared error (PMWSE) via a validation data set of size 10,000.

For normally distributed responses, we are not implying a constraint on the response function of FlexER, while for Gamma distributed responses we include the positivity constraint in the estimation of the response function of FlexER. In the case of Beta distribution, we specify in FlexER the constraint to have values in the interval [0,1]. In the estimation of Normal and Gamma distributed data via Fixed_ER we consider a log-link independently of the true link function since the identity link is represented by the classical expectile regression. For Beta distributed data we specify a logit-link while applying Fixed_ER. The results for Normal, Gamma, and Beta distributed data are shown in Figure 1.

These results indicate that for the benchmark data with identity link, the PMWSE of the new FlexER approach is similar to the results by classical expectile regression. If the response function is the log link, our new approach outperforms the classical expectile regression for all asymmetry levels. In case of heteroscedasticity-inducing link function, the differences get clearer with increasing asymmetry level. This is based on the simulation design, where the link function is set as linear for small error terms (lower asymmetry level) and as log link for larger error terms (higher asymmetry levels). Thus, the pattern of difference between classical and flexible expectile regression is valid for both Normal and Gamma distributions. In the Beta distributed design we see a similar pattern. Generalized expectile regression with flexible response function outperforms classical expectile regression. In the supplementary material (Section B.2.2 and B.2.3) we also show the estimated response functions and covariate effects, as well as additional asymmetry levels τ. These plots indicate that our new FlexER approach estimates both the response function and the covariate effects correctly, while the classical expectile regression results in biased estimates or predicts values, which are impossible by theory.

In all the above settings, except the Gamma distribution with log-link, the generalized expectile regression with flexible response function has a lower PMWSE than the version with fixed response function. Thus, we now focus on cases when both approaches are similar. The key point is that the model with fixed response function is dependent on the correct specification of the response function. Therefore, we provide two additional cases where both models coincide, the design with Lognormal and Pseudo-Beta distributed response. The remaining design of the simulation study is similar to above. In detail for the Lognormal case we do not place a constraint for the flexible model, while we specify $h(\eta) = \exp(\eta)$ as response function for the fixed version. For the Pseudo-Beta case we place the constraint to be between 0 and 1 for the flexible approach and specify $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$ for the fixed version. The resulting PMWSE is displayed in Figure 2.

The similarity of both generalized approaches is visible. Additionally, we see that both approaches are superior to the classical expectile regression when considering all asymmetries in total. For the Pseudo-Beta approach and low asymmetry levels, all three approaches perform similarly, which is based on the linear shape of the inverse logit function for values of the predictor around 0. The estimated response functions and covariate effects are shown in the supplementary material Section B. Based on these results we emphasize to only use the flexible approach, as we do for the remainder of this article.

5.3 | Identifiability of generalized expectile regression

The last part of the simulation study is designed to check whether our new approach is identified. Therefore, we build one data set of size 100,000 and estimate expectiles for the asymmetry levels $\tau = 10\%, 50\%, 90\%$ with ER and FlexER. Based on these models the prediction/expectiles for each combination of values of x_1 and x_2 are calculated as $(\hat{e}_\tau^{(ER)}(x_{u1}, x_{v2}), \hat{e}_\tau^{(FlexER)}(x_{u1}, x_{v2}))$. Due to the grid structure of the covariates also the empirical expectiles for each combination can be estimated without building a model $(\hat{y}(x_{u1}, x_{v2}))$. Then we compare the fitted values for each combination of x_1 and x_2 $(\hat{e}_\tau^{(\cdot)}(x_{u1}, x_{v2}))$ with the empirical/marginal expectile for this combination $(\hat{y}(x_{u1}, x_{v2}))$ by calculating their

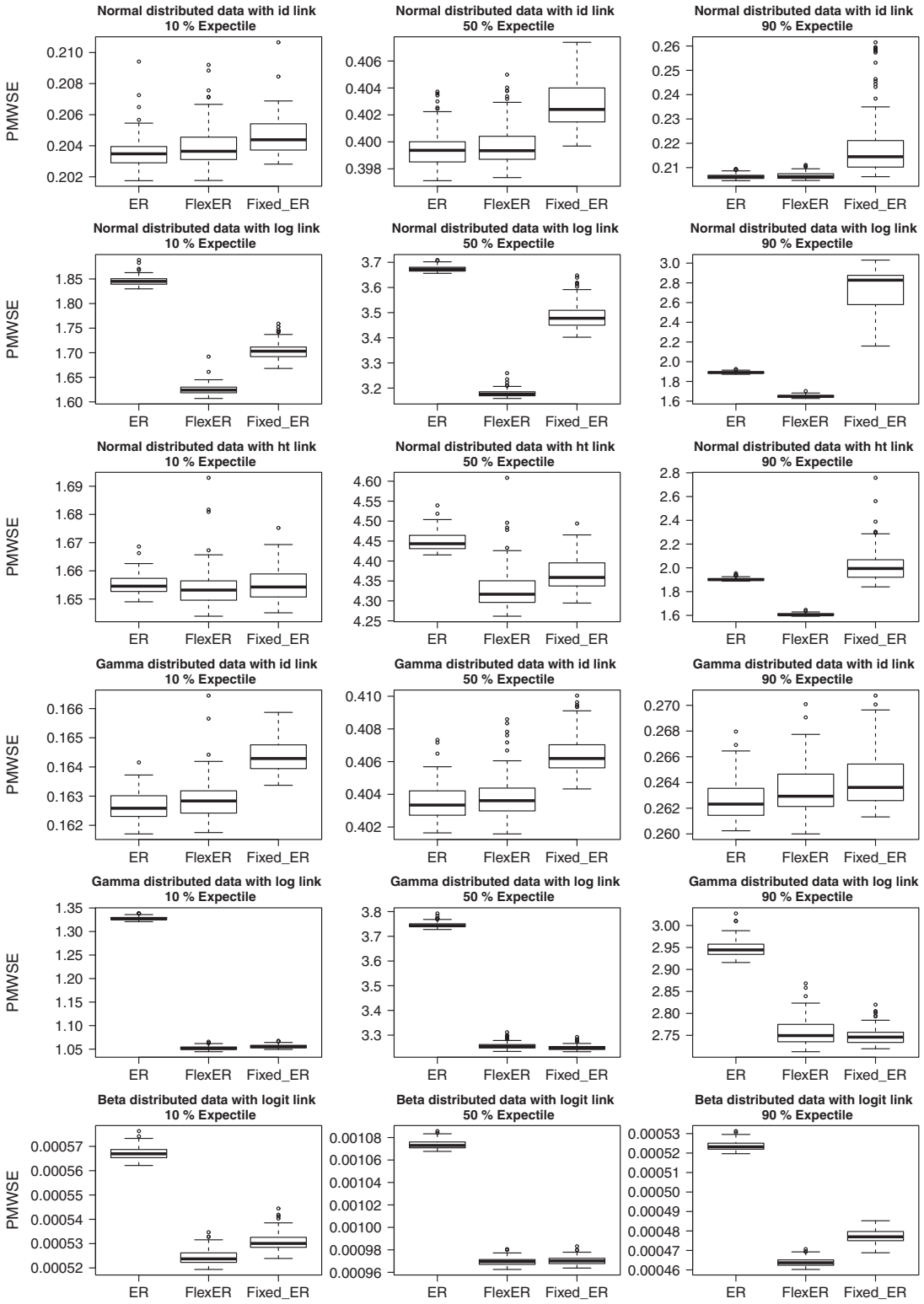

FIGURE 1 Estimated predictive mean weighted squared errors (PWMSE) for the 100 replications of the simulation study with Normal (top panels), Gamma (middle panels), and Beta distribution (bottom panel)

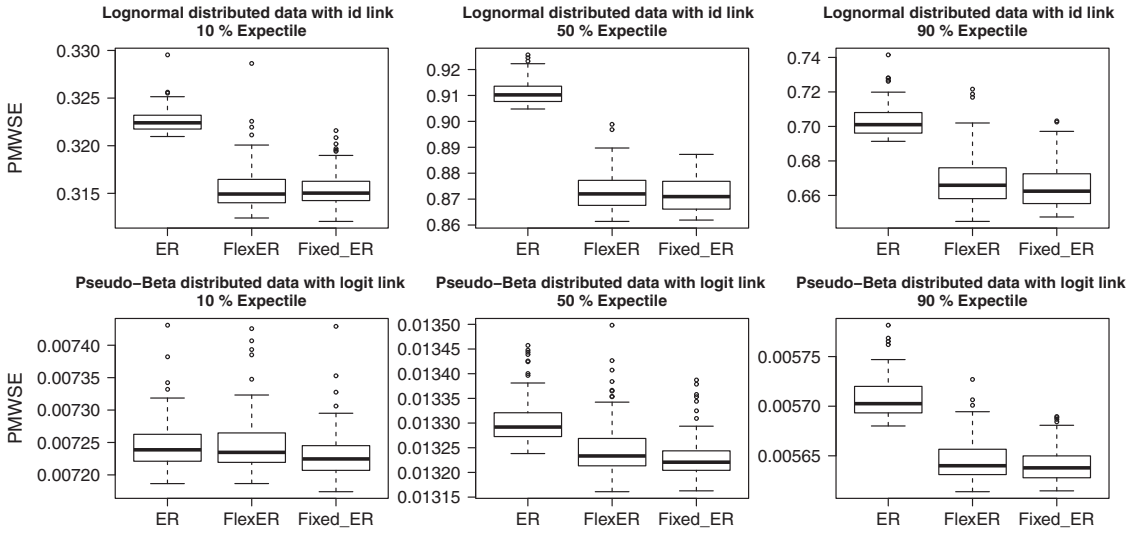

FIGURE 2 Estimated predictive mean weighted squared errors (PWMSE) for the 100 replications of the simulation study with Lognormal and Pseudo-Beta distribution

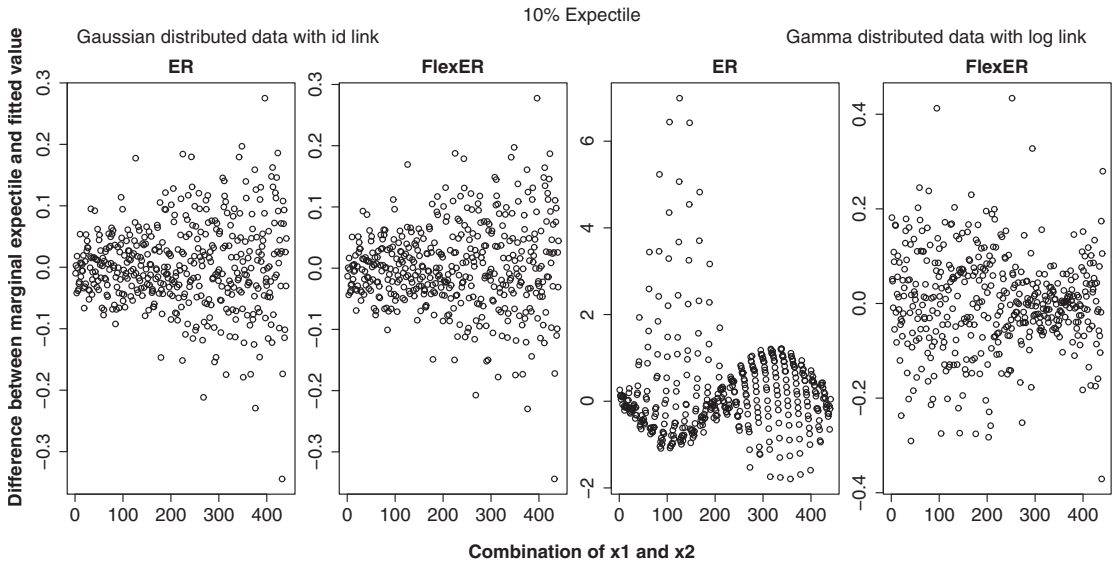

FIGURE 3 Difference between the predicted expectiles and the marginal expectiles for each combination of x_1 and x_2 with Normal distributed response with identity link and Gamma distributed response with log link

differences

$$\hat{y}(x_{u1}, x_{v2}) - \hat{e}_{\tau}^{(\cdot)}(x_{u1}, x_{v2}), \tag{3}$$

where u and v specify elements of the grid of covariates. In Figure 3 the remaining differences of the models are displayed. In the graphs each unique combination of x_1 and x_2 is one index on the horizontal axis, while the corresponding difference according to Equation (3) is shown on the y-axis. In a correct model the difference between the empirical and the predicted expectiles should be zero. In the paper we show the results of 10% expectile regression for Normal distributed response with identity link and Gamma distributed response with log link. The remaining figures are collected in the supplementary material (Section B.3.2), since the overall trends are the same.

In the base scenario in which we simulated the data with the identity link both models give a similar estimate. Their differences show no specific pattern. Moreover, the values of the differences between the fitted values based on classical or generalized expectile regression and the empirical/marginal expectiles shrink toward zero. However, for the cases with

nonidentity links the classical expectile regression predicts expectiles that are systematically different from the marginal ones. The residuals do not shrink toward zero and there are some trends remaining. Contrarily, the generalized expectile regression predicts expectiles that coincide with the marginal values and has no remaining trends. Furthermore, the residuals shrink to zero. Overall, we conclude that the new approach forecasts the correct values while the classical approach does not. So our new approach is identified. In the supplementary material (Section B.3.3) the estimated covariate effects and response functions are displayed. Those also indicate that the new model recaptures the underlying effect, while the classical expectile regression results in biased estimates. Generalized expectile regression with fixed response function (Fixed_ER) behaves similarly as classical expectile regression (as displayed in the supplementary material Section B.3), as long as the response function is correctly specified there is no pattern visible and unbiased estimates occur. If the response function is not specified correctly the differences do not shrink toward zero and patterns remain.

6 | EMPIRICAL APPLICATIONS

The new model class generalized expectile regression allows us to consider constraints on the target set. To emphasize the use of applying these constraints two examples are discussed in the following two subsections. In the first example, we apply our model in a study conducted in northwest Germany on the hearing scores (values limited to the range 0 to 10) of the general population. The second example uses publicly available data from the general population of the United States on mercury concentrations in blood samples. In this data set, the values of the response are strictly positive. Hence, in both cases our model is more appropriate to use than classical expectile regression.

6.1 | Analysis of self-reported hearing

The HÖRSTAT study was a cross-sectional study with 1903 participants recruited in the northwest of Germany between 2010 and 2012. The participants passed a comprehensive protocol including hearing tests of pure tones and speech, an interview, and a questionnaire survey on hearing abilities. Compared to the general population, middle-aged and older cohorts as well as high levels of educational attainment were overrepresented in the sample. Details on the conduct of the study and the sample composition can be found in detail elsewhere (von Gablenz & Holube, 2017).

After the application of exclusion criteria such as nonnative speaker status or missing covariate values, 1 737 complete cases were obtained. The data for 74 further hearing aid–wearing participants were excluded to avoid the mismatch of self-reported abilities in the aided condition and audiometric measurements in the unaided condition. In total, data of 1 663 adults (55% women) were included for the present analysis. In examining hearing performance, the participants conducted the Goettingen Sentence Test in noise (GÖSA, Kollmeier & Wesselkamp, 1997). Everyday sentences were presented in background noise to the participants whose task was to repeat as many sentences or single words as possible. The speech level was continuously adapted to the individual participant's performance to estimate the speech reception threshold (SRT), which refers to a speech intelligibility of 50%. The better-ear SRT for each individual as well as the absolute difference between the left and the right ear SRT as a quantification of hearing asymmetry (hereafter called "SRT asymmetry") were used for the following model calculations. The main focus of this analysis is on the self-reported hearing abilities as collected in the Speech, Spatial and Qualities of Hearing Scale (SSQ; Gatehouse & Noble, 2004) which was used in its German short form (SSQ17; Kießling et al., 2011). The SSQ17 contains five items each for the three subscales as well as an item addressing hearing abilities in quiet and overall listening effort, respectively. The items describe everyday situations involving certain listening tasks. The respondents rate how well they could accomplish these tasks on a discrete scale from 0 to 10 points. In this paper, we concentrate on one single item, which raises the question of listening effort: "Do you have to put in a lot of effort to hear what is being said in conversation with others?" In previous analysis we found that some expectile regression models had fitted values above the possible maximum score of 10 points (no effort). The violation can be seen in the left plot of Figure 5 for the exemplary item listening effort. The fitted values exceed 1 for $\tau = 0.9$ and $\tau = 0.95$. Hence, we rescaled the listening effort score to the standard unit interval in order to apply a generalized expectile regression. We constructed the model as follows:

$$e_\tau = \tilde{h}_\tau\big(\beta_{1,\tau}\text{Gender} + \boldsymbol{\beta}_{2,\tau}\text{Schooling} + s_{3,\tau}(\text{Age}) + s_{4,\tau}(\text{SRT}) + s_{5,\tau}(\text{SRT asymmetry})\big).$$

We approximated the unknown functions $s_{\cdot,\tau}$ with penalized B-spline bases of order 3 with maximal 20 degrees of freedom. The smoothing parameters were optimized via 5-fold cross-validation for the $\tau = 0.5$ expectile and the results were used in the estimation of all asymmetry levels. This simplification is based on the idea that interpretation of the results, especially between the different asymmetry levels is easier, if at least the smoothing parameters are fixed. We need the response functions to vary between the different asymmetry level to capture the different shape of the distributions in the tails, however fixing the smoothing parameters does not impact the flexibility too much. Using models with individually optimized smoothing parameters would result in a better individual fit, but generalizations are relaxed using the same set. Another reason, as described in Section 4.3, for using the same set of smoothing parameters for all asymmetry levels is that this reduced the risk of crossing expectiles. The results of the estimation with individually optimized smoothing parameters can be seen in the supplementary material Section C.2. There we see that using the individually optimized smoothing parameters results in similar trends, but more expectile crossings occur.

During the estimation, the smooth covariate effects are all centered around 0. To provide a better impression of the resulting effect we add constants to the metric covariate effects, when displaying the estimates. For the metric covariates of the classical expectile regression, as displayed in Figure 4 in the top row, the classical intercept is added to the estimates.

As described in Section 4.3, we display the covariate effects of the generalized expectile regression in two ways. First, we display them together with the estimated response function $\tilde{h}_\tau(s_{r,\tau}(x) + c_{r,\tau})$ (Figure 4, middle row). Second, we show the pure covariate effects, where we add some intercept based on the fitted values to ease the view on the different asymmetries $s_{r,\tau}(x) + \bar{\hat{e}}_\tau$ (Figure 4, bottom row). For the original effects see the supplementary material (Section C.2).

As can be seen in the classical expectile regression (Figure 4, top row), a worse performance in the Goettingen Sentence Test (higher SRT) was associated with higher listening effort (lower scores). Due to a very small number of observations for very low and very high SRT there is an increased uncertainty to the estimated effects in the extremes. We saw an increase in variance for higher SRT and a relatively symmetric distribution as shown by the expectiles. Further, we observed a small negative effect of SRT asymmetry. The effect of age was small and showed a counterintuitive increase at older age, that is, elderly adults reported less listening effort than middle-aged adults. When interpreting this effect, it is important to bear in mind that data from participants who were fitted with hearing aids were excluded from the analysis. As a consequence, the data for this analysis refer to participants who retained comparatively good hearing for their age.

In the generalized expectile regression model, the interpretation of the covariate effects is similar as in generalized linear models, that is, the direction of the effects can be interpreted straightforwardly while the size of the effect depends on the response function. Overall, we found less change in the variance of the listening effort across the covariates than in the classical expectile regression model. As can be seen in the results (Figure 4, bottom row), the distance between the expectiles is much larger in the lower tail than in the upper tail. Hence, the expectiles gave additional information about the skewness of the listening effort scores. The size of the observed effect was similar to classical expectile regression with negative associations of SRT and SRT asymmetry. Again, there was basically no relevant effect of age. Combining the covariate effect and the response function (Figure 4, middle row) results in similar estimates as the classical expectile regression, but ensuring the limits of the support of the scores. The estimated response functions are shown in the right plot of Figure 5. They are mostly linear in the lower tail. At the upper tail toward the maximum of the listening effort scale, a small curvature can be found. This curvature helps to ensure that the predicted values stay within the limits of the support of the response.

The heteroscedasticity of the response is mostly captured by the estimated response functions. Those are not parallel. In the right plot of Figure 5, we see linear effects for the response functions at the lower tail, while for the upper tail they comprise the constraint on the target set. Overall, the curves on the lower tail have a steeper slope than at the upper tail. This highlights the differences in the variance. For example, the variance is increased for combinations of covariates, which results in a low value of the predictor. For low SRT (read: better hearing) and low SRT asymmetry combined with an average age we observed a higher variance of the listening effort than for high SRT (read: worse hearing) and more pronounced SRT asymmetry. This indicates that participants who performed comparatively good in the speech in noise test have a more volatile impression of their listing effort than participants with poorer test results.

The estimates of this model show that the heteroscedasticity is mainly captured with the response function. We conclude that a purely additive effect of SRT and SRT asymmetry is not sufficient. Moreover, the estimation of the response functions allows for implying some kind of interaction effect between both variables.

We further applied bootstrap as discussed in Section 4.4 to estimate pointwise confidence bands for the estimated effects and response functions. The results are collected in the supplementary material Section C.1. Here we provide four different confidence intervals, either using smoothing parameters that are estimated for the 50% asymmetry level or separately

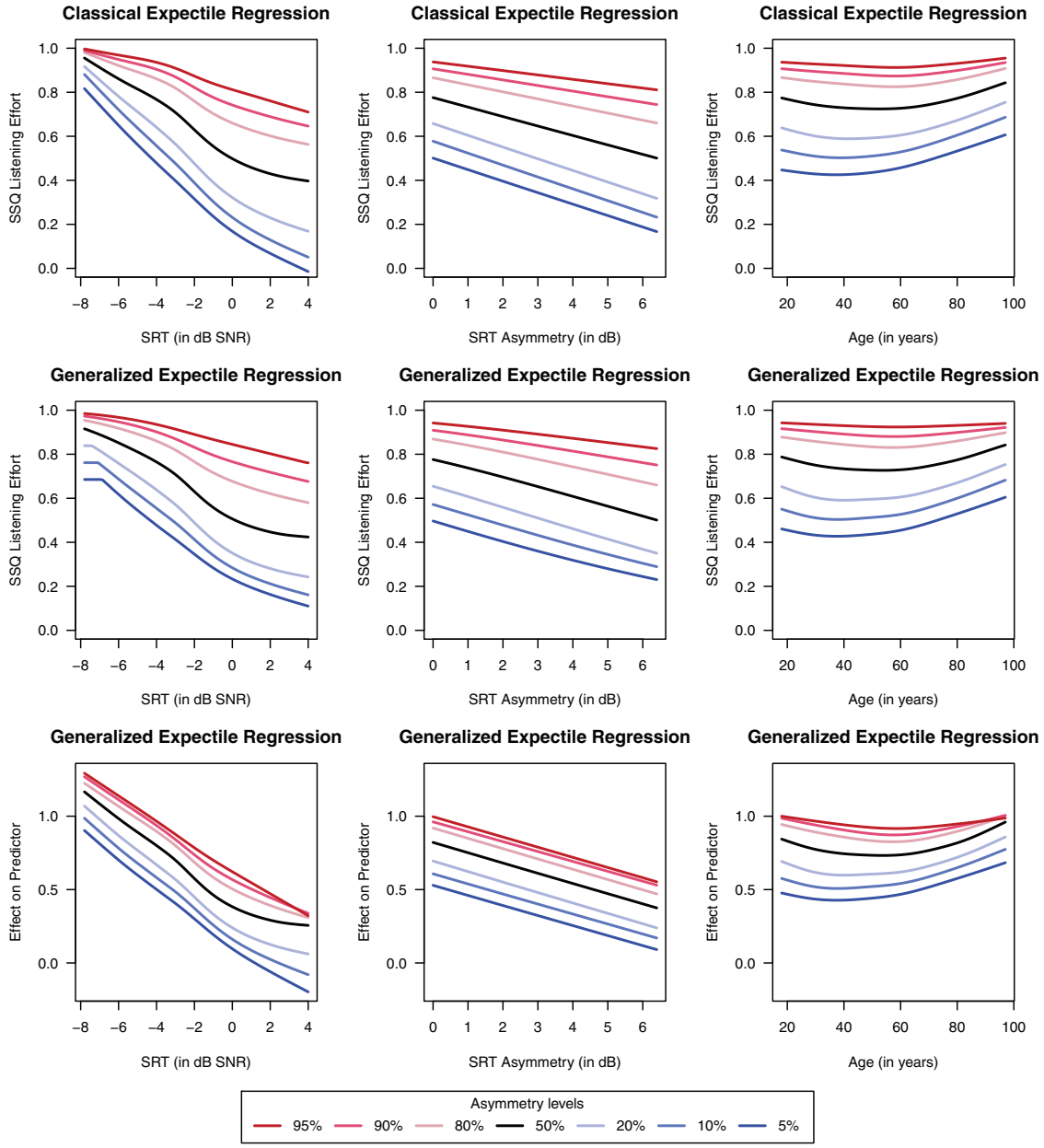

FIGURE 4 Influence of SRT, SRT asymmetry, and age on the self-reported listening effort. In the first row the results from the classical expectile regression are displayed, while the second and third rows show the results from generalized expectile regression. In detail, the second row shows the combined effect of the covariate and the response function $\tilde{h}_\tau(s_{r,\tau}(x) + 0)$, while the third row displays the pure covariate effects on the predictor $s_{r,\tau}(x) + \tilde{\bar{e}}_\tau$

for each asymmetry level. Moreover, we discriminate, whether the smoothing parameters are optimized just outside of the bootstrapping or also inside. Generally, the results provide similar figures, while the more flexible the smoothing parameters are, the wider are the intervals. In our example this is especially visible for smaller effects of the covariates.

6.2 | Modeling of mercury concentration in blood

The NHANES study (Centers for Disease Control and Prevention (CDC), 2020) consists of a general health survey that is conducted in the United States on a regular basis. Participants have to fill in a detailed sociodemographic questionnaire and multiple measurements are performed. This includes classical examinations, like weight and height measurements, as

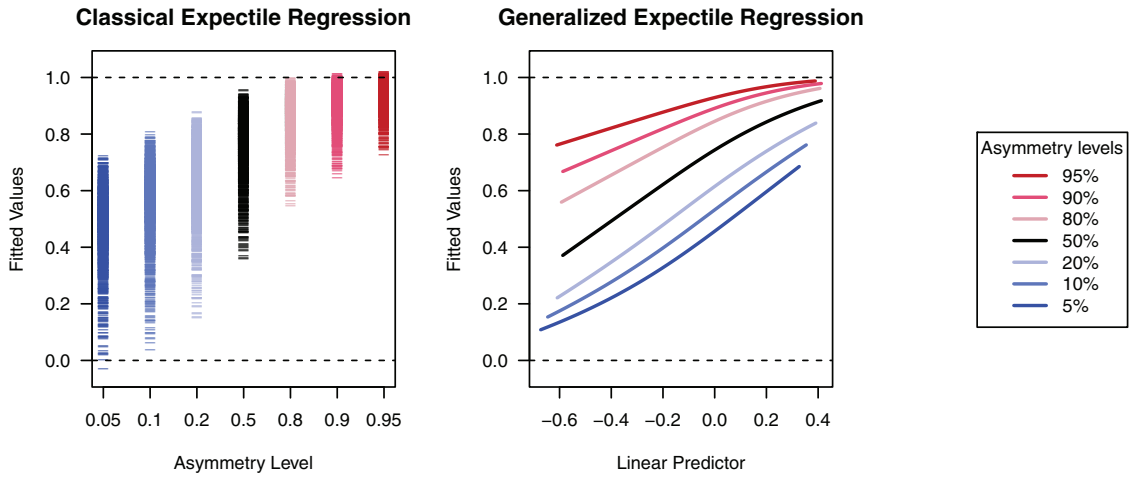

FIGURE 5 Results of the self-reported hearing model. On the left: Fitted values for classical expectile regression with several asymmetry levels. On the right: Estimated response function of generalized expectile regression with several asymmetry levels

well as laboratory analyses based on blood and urine samples. Additionally, interviews are conducted to evaluate, among others, the nutritional intake over the previous weeks. In our analysis, we focused on the concentration of mercury in blood samples. Mercury is toxic for humans and can cause harmful effects on multiple parts of the body. This includes effects on the nervous, digestive, and immune systems, as well as lungs or kidneys (WHO, 2017). Here we estimated associations between concentrations of mercury and a number of covariates characterizing the study participants. Therefore, we did not only focus on mean effects, but evaluated whether there are differences between effects on the lower or upper end on the concentration scale to detect, for example, heteroscedasticity. For a better comparison, we applied both, classical expectile regression as well as generalized expectile regression with flexible response function.

In this application, we modeled differences of mercury concentration in the human body measured by the total amount of mercury in a liter of blood in μg/L (THg). The values range from 0.11 to 50.81. A histogram of the distribution of mercury is displayed in the supplementary material Section D. It shows a clearly right skewed distribution with higher frequency of low concentration, but without actually reaching the boundary of the target set. In the analysis, we used extracted data from the NHANES study from 2011 and 2012. In the inclusion criteria, we limited the sample to participants who were between 21 and 79 years old and had a body mass index (BMI) between 16 and 50. We excluded children as their BMI has to be handled differently. We further excluded extreme values of the BMI for a more consistent estimate of the nonlinear effect of BMI. Further, we required the income and the highest education to be available as well as self-reported values for the consumption of fish and shellfish from the questionnaire. The latter are known to be associated with mercury concentration as described in Garí et al. (2013) and Buchanan et al. (2015). Finally, we required available information in the variables age and sex. After the removal of missing values and further plausibility checks, we had a sample of size 3 965.

The variable household income of the participants is coded in three categories, while the educational level is coded in five categories. Sex is a binary variable. Fish and shellfish consumption are coded in four categories, depending on the frequency of intake. THg, age, and BMI are coded as numerical variables. The exact details of sampling and data preparation are provided in the supplementary material.

Overall, we built generalized expectile models for $\tau = 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95$ with the following structure:

$$e_\tau = \tilde{h}_\tau\big(\beta_{1,\tau}\text{Sex} + \boldsymbol{\beta}_{2,\tau}\text{Education} + \boldsymbol{\beta}_{3,\tau}\text{Income} + \boldsymbol{\beta}_{4,\tau}\text{Shellfish} + \boldsymbol{\beta}_{5,\tau}\text{Fish} + s_{6,\tau}(\text{BMI}) + s_{7,\tau}(\text{Age})\big).$$

Again, the unknown functions $s_{\cdot,\tau}$ were approximated with cubic penalized B-spline bases and a maximum of 20 degrees of freedom. The smoothing parameters were optimized via 5-fold cross-validation for the $\tau = 0.5$ expectile and the results were used in the estimation of all asymmetry levels. For comparison, a similarly specified classical expectile regression model was evaluated.

The resulting predictions for the classical model as well as the estimated response functions are displayed in Figure 6 in the left column. The other columns of this figure show the estimated smooth covariate effects of both models without

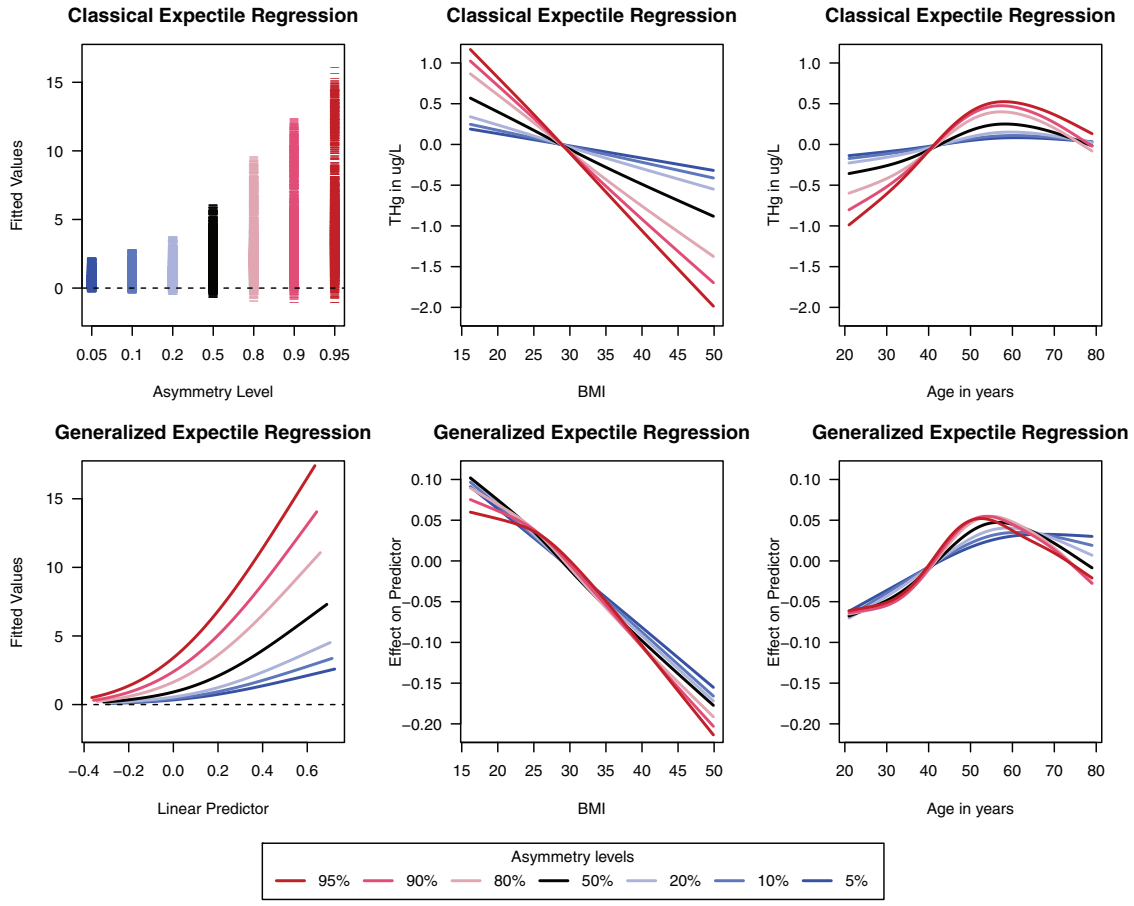

FIGURE 6 In the top row: Estimates of classical expectile regression. In the bottom row: Estimates of generalized expectile regression. In the left column: Fitted values and estimated response functions for several asymmetry levels. In the two right columns: Estimated smooth covariate effects

applying the response function, or any kind of shifting. The covariate effects including the response function or intercepts are displayed in the supplementary material Section D.2. Based on the predictions of the expectiles in the classical regression, we conclude that this model is not sufficient. Frequently, values below the natural limit of zero are predicted for all asymmetry levels. Contrarily, generalized expectile regression considers the positivity constraint. The estimated response functions show a nonlinear effect, which is also increasing more steeply in the upper tail. The covariate effects based on classical expectile regression would therefore be biased.

In the first row of Figure 6 the effect of age on the mercury concentration (THg) for low asymmetry levels would be considered as being close to constant compared with the upper levels. However, with generalized expectile regression, we see different estimates. While the concentration is constantly increasing in the lower tail, for higher ages there is a maximum of the concentration around 50 years of age and then a clear decrease in the upper asymmetry levels. In the generalized model, the flattening of the effects for low asymmetry levels is captured with the estimate of the response function and allows us to check the pure covariate effects for differences. Thus, the misspecification of the response function hinders the detection of potentially interesting patterns due to the shrinkage of the effects on the lower asymmetry levels. For BMI, both models show a negative impact of increasing BMI on the concentration of THg. Comparing the sizes of the effects of age and BMI in both models for the upper asymmetry level shows that in classical expectile regression the impact of BMI is slightly underestimated in comparison to age. The values range from -2 to 1 for BMI and -1 to 1 for age. In the generalized expectile regression model the values for BMI range from -0.2 to 0.1, while the age effect varies from -0.07 to 0.07. Both models show that the variation of the expectiles between the different asymmetry levels is lower for the lower tail of the distribution, while the variation increases with increasing asymmetry levels.

Bootstrap confidence intervals, as described in Section 4.4, provide an impression of the variability of the effects. The estimated effects and their confidence intervals are displayed in the supplementary material Section D.1. We also included

the estimated regression coefficients of the categorical covariates and their bootstrap confidence intervals in the supplementary material.

7 | CONCLUSION

In this paper, we introduced a new approach to estimate expectile regression models for response variables with compact or otherwise restricted support. Our new generalized expectile regression solves the open question on how to deal with constraints on the target set in expectile regression and allows for a nonlinear relationship between the predictor and the expectile. Therefore, we built a bridge between expectile regression and single index models. This results in interpretable estimates from a very flexible model. The analysis of data on hearing abilities shows the necessity to consider the constraints of the response. A comparison of generalized and classical expectile regression in the hearing questionnaire data shows that both approaches estimate similar trends while the generalized expectiles are restricted to the support of the response. The latter ensures valid audiological results. Similarly, the analysis of mercury concentration in blood samples is less biased, when considering the positivity of the estimates. Due to its flexibility and freedom from any distributional assumption generalized expectile regression is widely applicable, but the drawbacks are a computationally burdensome algorithm and rather slow convergence rates. Moreover, the interpretation of the results is not straightforward.

The model is a little more complex than usual such that model selection is of major importance. Furthermore, the model is sensitive to the choice of the smoothing parameters. We started by using cross-validation to select the optimal values. However, using a GCV criterion for expectiles (Schnabel & Eilers, 2009), or a modified Schall algorithm (Wood & Fasiolo, 2017) could reduce the computation time significantly. Since the smoothing parameter selection offers a wide variety of options, the best criterion and algorithm for the smoothing parameter choice should be the next step in the research of this model class.

Overall, we reduced the assumptions in classical expectile regression by showing a path to generalized additive models. Hence, generalized expectiles could also be a valuable alternative to Generalized Additive Models for Location Scale and Shape (GAMLSS) (Stasinopoulos et al., 2017). Both models allow for the estimation of all details of a response's distribution. However, with generalized expectiles we can avoid a possibly false parametric distributional assumption or the selection of an inappropriate link function. Instead we now estimate the link semiparametrically. In practice, we also often find that the convergence of GAMLSS is not guaranteed despite its likelihood-based approach. Nevertheless, GAMLSS have found a wide field of applications in recent years. Thus we conclude that generalized expectiles can be the model of choice in various scenarios.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the comments from three anonymous referees and an associate editor who provided valuable suggestions for improving the original submission of this paper. In addition, we especially thank Inga Holube for her support and the design of the HÖRSTAT study. This work was supported by the Research Association of German Hearing Aid Acousticians, the Lower Saxony Department of Science and Culture, and the European Regional Funding with HÖRSTAT. Further analysis was financed from the federal resources of Niedersächsisches Vorab within the research focus "Hearing in everyday life (HALLO)." We also thank Mercè Garí for her support with analyzing mercury concentrations. Moreover, we acknowledge financial support by the German Research Foundation (DFG), grant KN 922/4-2.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Most data that support the findings of this study are openly available in Open Science Framework.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the https://osf.io/3zsqe/.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Elmar Spiegel https://orcid.org/0000-0002-5834-2383
Thomas Kneib https://orcid.org/0000-0003-3390-0972

REFERENCES

Buchanan, S., Anglen, J., & Turyk, M., (2015). Methyl mercury exposure in populations at risk: analysis of NHANES 2011–2012. *Environmental Research*, *140*, 56–64.

Carroll, R. J.J. Fan, GijbelsI. & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Centers for Disease Control and Prevention (CDC) (2020). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2011.

Czado, C., & Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, *33*(2), 213–231.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*, 85–100.

Efron, B. (1992). Poisson overdispersion estimates based on the method of asymmetric maximum likelihood. *Journal of the American Statistical Association*, *87*(417), 98–107.

Efron, B., & Tibshirani, R. J., (1994). An *introduction to the bootstrap*. CRC Press.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121.

Garí, M., Grimalt, J. O., Torrent, M., & Sunyer, J., (2013). Influence of socio-demographic and diet determinants on the levels of mercury in preschool children from a Mediterranean island. *Environmental Pollution*, *182*, 291–298.

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, *43*(2), 85–99.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*(1), 71–120.

Jiang, R., & Qian, W.-M., (2016). Quantile regression for single-index-coefficient regression models. *Statistics & Probability Letters*, *110*, 305–317.

Jørgensen, B. (1984). The delta algorithm and GLIM. *International Statistical Review/Revue Internationale de Statistique*, *52*(3), 283–300.

Kießling, J., Grugel, L., Meister, H., & Meis, M. (2011). Übertragung der Fragebögen SADL, ECHO und SSQ ins Deutsche und deren Evaluation. *Z Audiol*, *50*(1), 6–16.

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, *61*(2), 387–421.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, *46*(1), 33–50.

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, *102*(4), 2412–2421.

Marschner, I. C. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal*, *3*(2), 12–15.

Marx, B. D., (2015). Varying-coefficient single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, *143*, 111–121.

McCullagh, P., & Nelder, J. A., (1989). Generalized *linear models*. CRC Press.

Muggeo, V. M., & Ferrara, G. (2008). Fitting generalized linear models with unspecified link function: a P-spline approach. *Computational Statistics & Data Analysis*, *52*(5), 2529–2537.

Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, *55*(4), 819–847.

Rigby, R., Stasinopoulos, D., & Voudouris, V. (2013). Discussion: A comparison of GAMLSS with quantile regression. *Statistical Modelling*, *13*(4), 335–348.

R Core Team, (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Schnabel, S. K., & Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, *53*(12), 4168–4177.

Schulze-Waltrup, L., Sobotka, F., Kneib, T., & Kauermann, G. (2015). Expectile and quantile regression—David and Goliath? *Statistical Modelling*, *15*(5), 433–456.

Sobotka, F., & Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, *56*(4), 755–767.

Spiegel, E., Kneib, T., & Otto-Sobotka, F. (2019). Generalized additive models with flexible response functions. *Statistics and Computing*, *29*(1), 123–138.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F., (2017). Flexible *regression and smoothing: Using GAMLSS in R*. CRC Press.

Tutz, G., & Petry, S. (2016). Generalized additive models with unknown link function including variable selection. *Journal of Applied Statistics*, *43*(15), 2866–2885.

von Gablenz, P., & Holube, I. (2017). Social inequalities in pure-tone hearing assessed using occupational stratification schemes. *International Journal of Audiology*, *56*(7), 443–452.

von Gablenz, P., Otto-Sobotka, F., & Holube, I., (2018). Adjusting expectations: Hearing abilities in a population-based sample using an SSQ short form. *Trends in Hearing*, *22*, 2331216518784837.

Wang, J.-L., Xue, L., Zhu, L., & Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *The Annals of Statistics*, *38*(1), 246–274.
Weisberg, S., & Welsh, A. (1994). Adapting for the missing link. *The Annals of Statistics*, *22*(4), 1674–1700.
WHO (2017). Mercury and health. *Fact Sheets*. https://www.who.int/en/news-room/fact-sheets/detail/mercury-and-health [2020-11-06].
Wood, S. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, *15*(5), 1126–1133.
Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.
Wood, S. N., & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, *73*(4), 1071–1081.
Wu, T. Z., Yu, K., and Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, *101*(7), 1607–1621.
Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042–1054.
Yu, Y., Wu, C., & Zhang, Y. (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing*, *27*(2), 571–582.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Spiegel E, Kneib T, von Gablenz P, Otto-Sobotka F. Generalized expectile regression with flexible response function. *Biometrical Journal*. 2021;1–24. https://doi.org/10.1002/bimj.202000203

APPENDIX

Algorithm

Please consider the centering in each step.

1. Initialize Algorithm: Estimate Classical Expectile Regression:
   - Start:

$$\boldsymbol{\gamma}_\tau^{\text{temp}} = (\boldsymbol{Z}^\top \boldsymbol{Z}^\top + \boldsymbol{K})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$$

   - Iterate until convergence of $\boldsymbol{W}$ and $\boldsymbol{\gamma}_\tau^{\text{temp}}$:

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{\text{temp}} \\ (1-\tau) & \text{if } y_i \leq \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{\text{temp}} \end{cases}$$

$$\boldsymbol{W} = \text{diag}(w_\tau(y_i))$$

$$\boldsymbol{\gamma}_\tau^{\text{temp}} = (\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z}^\top + \boldsymbol{K})^{-1} \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{y}$$

   - After convergence the first estimates are:

$$\boldsymbol{W}, \boldsymbol{\gamma}_\tau^{(0)}$$

2. Iterate (a) and (b) until the convergence of $\tilde{h}_\tau^{(m)}$ and $\boldsymbol{\gamma}_\tau^{(m)}$
   (a) Estimation of response function $\tilde{h}_\tau^{(m)}$ for fixed $\boldsymbol{\gamma}_\tau^{(m-1)}$:
      - Iterate until convergence of $\tilde{h}_\tau^{\text{temp}}$ and $\boldsymbol{W}$:

$$\tilde{h}_\tau^{\text{temp}}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(m-1)}) = \texttt{pcls}(\boldsymbol{y} \sim s(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(m-1)}, \texttt{ bs =}'' \texttt{ps}'') \mid \boldsymbol{A}\boldsymbol{\nu} \geq \boldsymbol{b}, \texttt{ weights=}\boldsymbol{W})$$

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \tilde{h}_\tau^{\text{temp}}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(m-1)}\right) \\ (1-\tau) & \text{if } y_i \leq \tilde{h}_\tau^{\text{temp}}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(m-1)}\right) \end{cases}$$

$$\boldsymbol{W} = \text{diag}(w_\tau(y_i))$$

- After convergence the estimates are: $\tilde{h}_\tau^{(m)}, \boldsymbol{W}$

(b) Estimation of covariate effects $\boldsymbol{\gamma}_\tau^{(m)}$ for fixed $\tilde{h}_\tau^{(m)}$:

- Initial $\boldsymbol{\gamma}_\tau^{(k-1)} = \boldsymbol{\gamma}_\tau^{(m-1)}$:
- Iterate until convergence of $\boldsymbol{\gamma}_\tau^{(k)}$:

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \tilde{h}_\tau^{(m)}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)}\right) \\ (1-\tau) & \text{if } y_i \leq \tilde{h}_\tau^{(m)}\left(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)}\right) \end{cases}$$

$$\omega_i^{(k)} = w_\tau(y_i) \cdot \left(\tilde{h}_\tau'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})\right)^2$$

$$y_i^{(k)} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)} + \frac{y_i - \tilde{h}_\tau^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})}{\tilde{h}_\tau'^{(m)}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_\tau^{(k-1)})}$$

$$\boldsymbol{W} = \text{diag}(w_\tau(y_i))$$

$$\boldsymbol{\Omega}^{(k)} = \text{diag}(\omega_i^{(k)})$$

$$\boldsymbol{\gamma}_\tau^{(k)} = \left(\boldsymbol{Z}^\top \boldsymbol{\Omega}^{(k)} \boldsymbol{Z} + \boldsymbol{K}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{\Omega}^{(k)} \boldsymbol{y}^{(k)}$$

- After convergence the estimates are: $\boldsymbol{\gamma}_\tau^{(m)}, \boldsymbol{W}$

3. The final estimates after iterating (2) are $\boldsymbol{\gamma}_\tau$ and $\tilde{h}_\tau$

Centering

Here $s_r^{(k)}(x_{ir})$ is any kind of predictor using the current estimates $\boldsymbol{\gamma}^{(k)}$. Here r is the index of an arbitrary covariate. To avoid cluttering the index τ is skipped.

$$\zeta_r^{(k)} = \frac{1}{n} \sum_{i=1}^{n} s_r^{(k)}(x_{ir}),$$

$$\theta^{(k)} = \left(\sum_r \sum_{i=1}^{n} \left(s_r^{(k)}(x_{ir}) - \zeta_r^{(k)} \right)^2 \right)^{1/2},$$

$$\boldsymbol{\gamma}_\tau^{(k)} = \frac{\boldsymbol{\gamma}_\tau^{(k)}}{\theta^{(k)}} \text{ and } s_r^{(k)}(x_{ir}) = \frac{s_r^{(k)}(x_{ir}) - \zeta_r^{(k)}}{\theta^{(k)}},$$

$$\eta_i^{(k)} = \sum_r s_r(x_{ir}),$$

$$\boldsymbol{\gamma}_\tau^{(k)} = \frac{\boldsymbol{\gamma}_\tau^{(k)}}{\max(\boldsymbol{\eta}^{(k)}) - \min(\boldsymbol{\eta}^{(k)})} \text{ and } \eta_i^{(k)} = \frac{\eta_i^{(k)}}{\max(\boldsymbol{\eta}^{(k)}) - \min(\boldsymbol{\eta}^{(k)})}.$$