Electronic Supplementary Material

Identifying Risk Factors for Severe Childhood Malnutrition
by Boosting Additive Quantile Regression

Nora Fenske, Thomas Kneib, Torsten Hothorn

A Additional Results for the Empirical Evaluation

This section contains additional results obtained from the empirical evaluations for the additive quantile regression model. For a detailed description of the simulation setups, see Section 3.1 of the manuscript.

For each of the considered distributions, i.e., standard normal, t and gamma distribution, we show results for the 'sin'-setup, the 'log'-setup, and the multivariable setup. The following boxplots (see Figure 1 – 11) display the empirical distributions of the considered performance measures from 100 simulation replications.

The multivariable setup was carried out several times based on different correlations ρ between the simulated covariate vectors, namely $\rho \in \{0, 0.2, 0.5, 0.8\}$. For all three response distributions, the results were almost identical for all correlation coefficients. Therefore, we only show results for all correlation coefficients in case of the gamma distribution and restrict the presentation to one of the possibly correlation coefficients for the remaining two distributions in the multivariable setup.

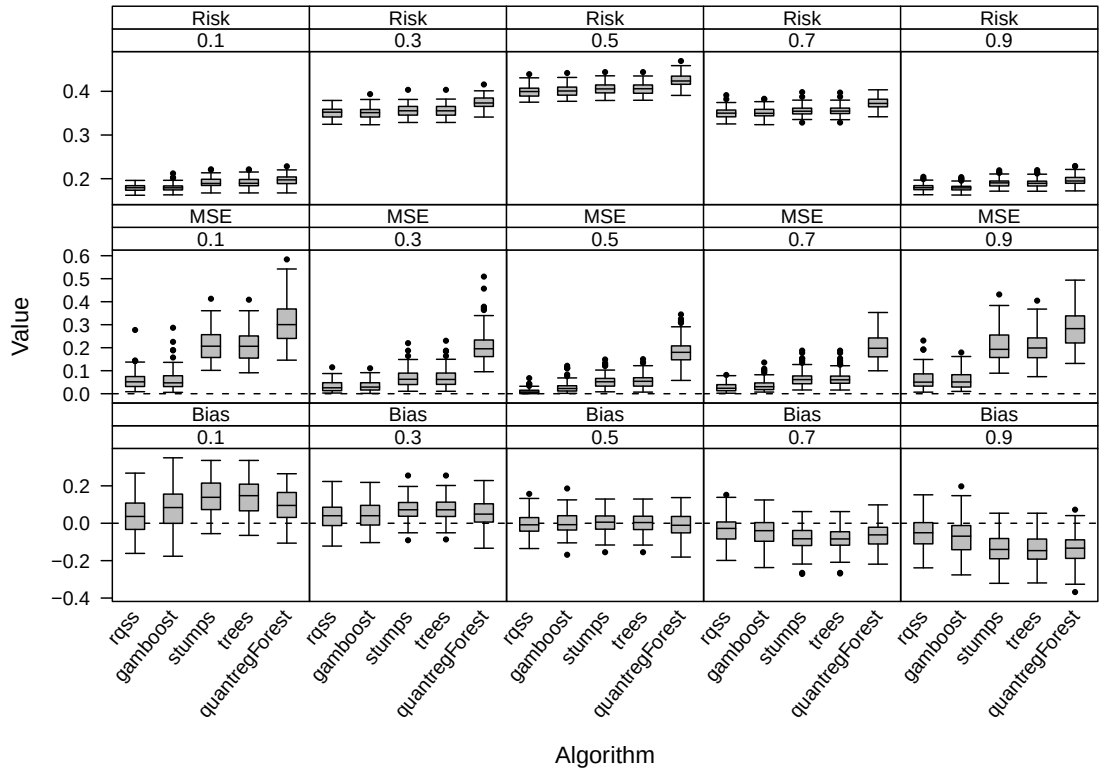


Figure 1: Simulation results for the 'sin'-setup with standard normal-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

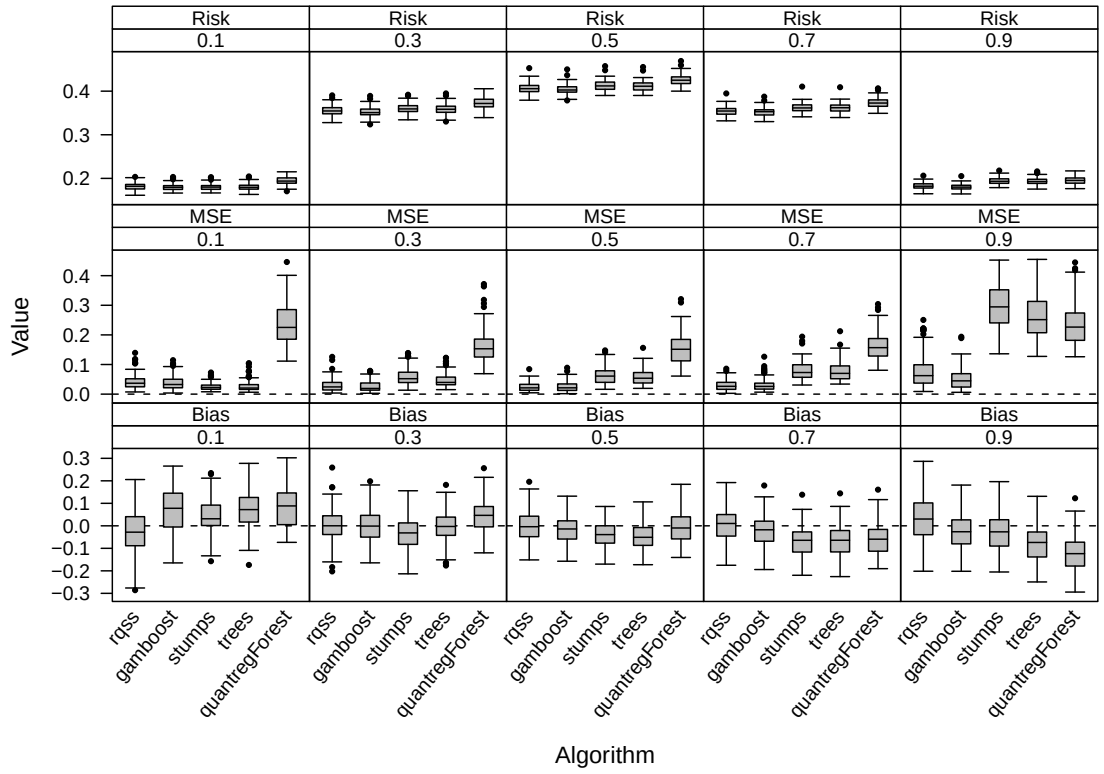


Figure 2: Simulation results for the 'log'-setup with standard normal-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

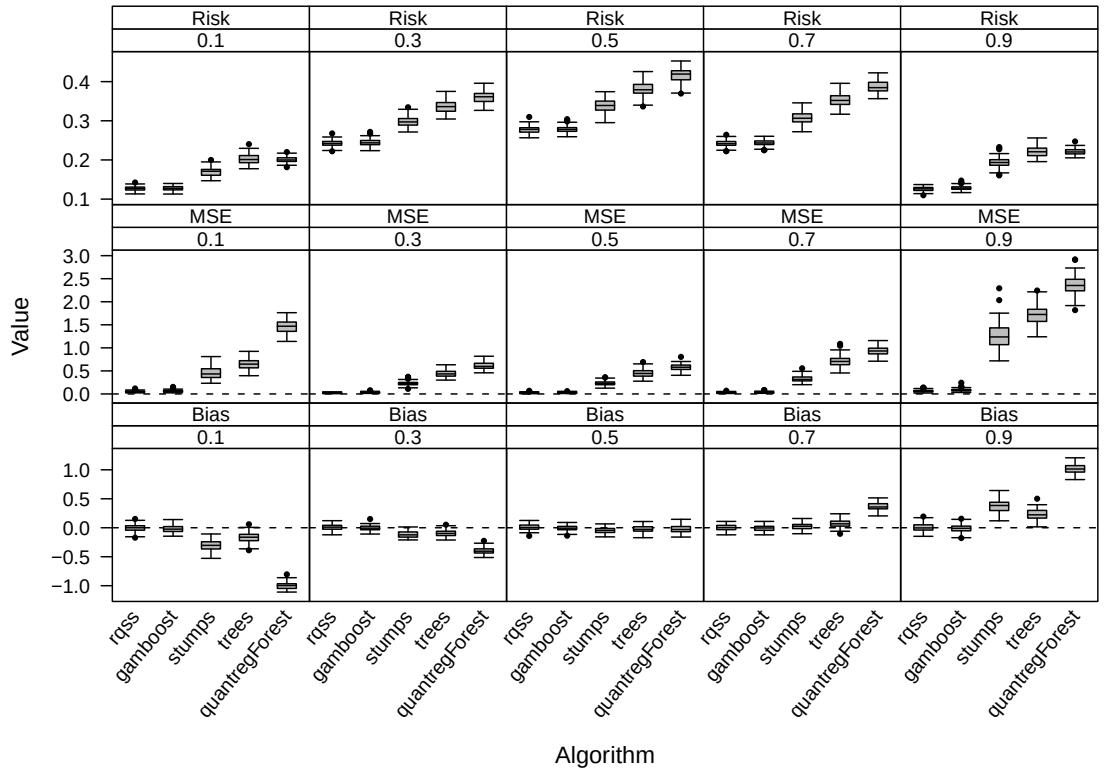


Figure 3: Simulation results for the multivariable setup with standard normal-distributed error terms and a correlation coefficient of 0.0. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

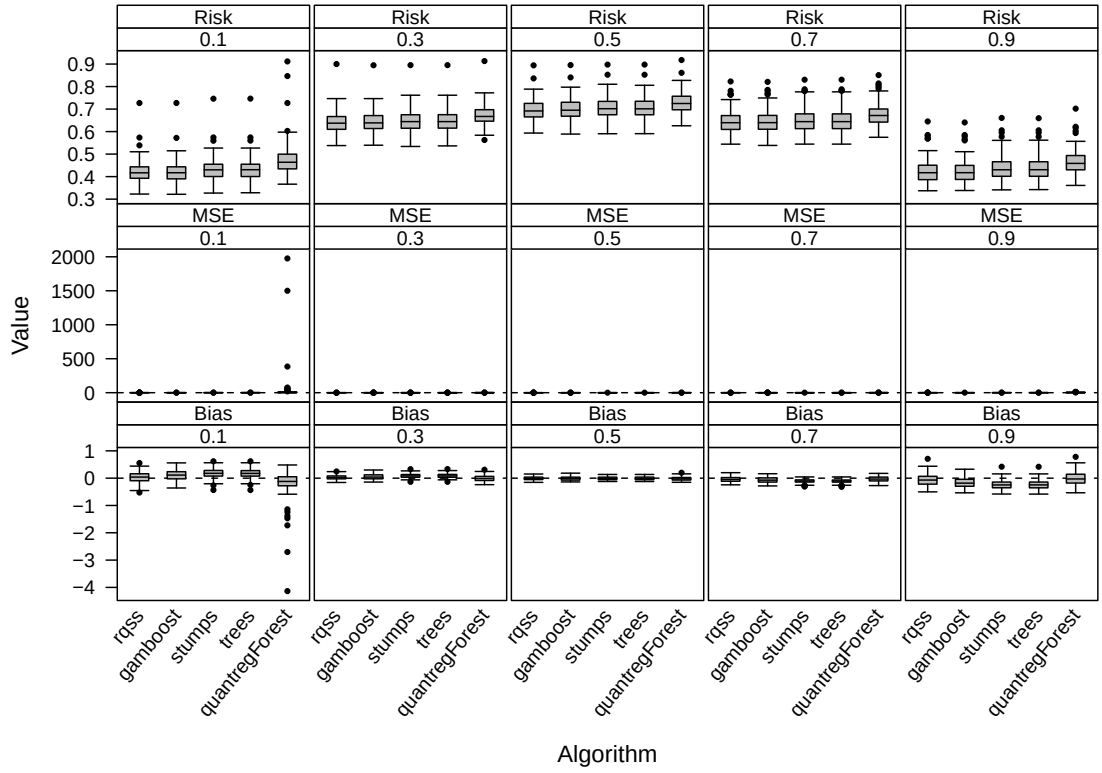


Figure 4: Simulation results for the 'sin' setup with t-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

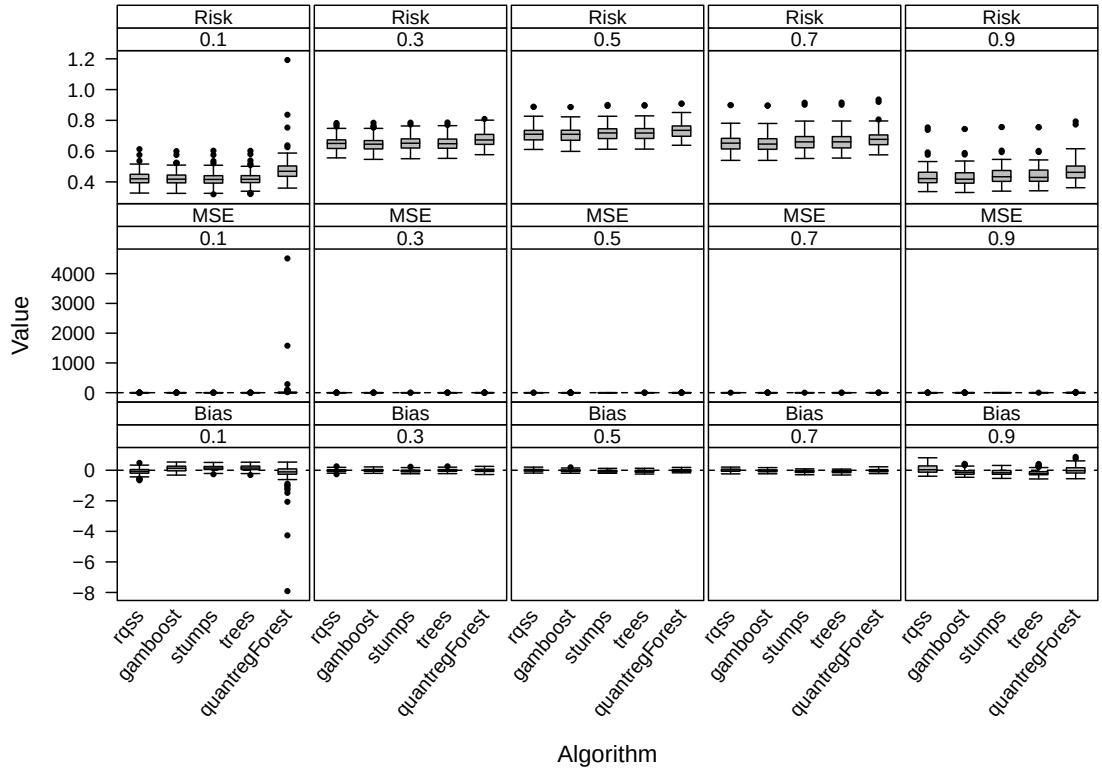


Figure 5: Simulation results for the 'log'-setup with t-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

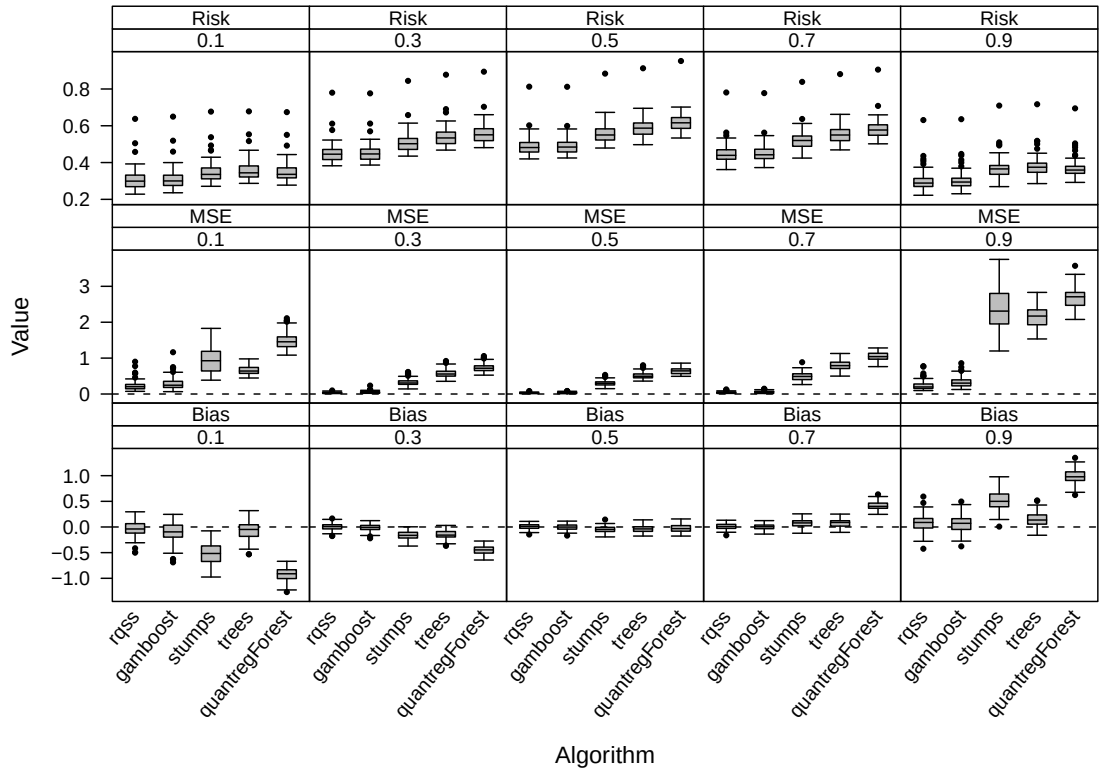


Figure 6: Simulation results for the multivariable setup with t-distributed error terms and a correlation coefficient of 0.2. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

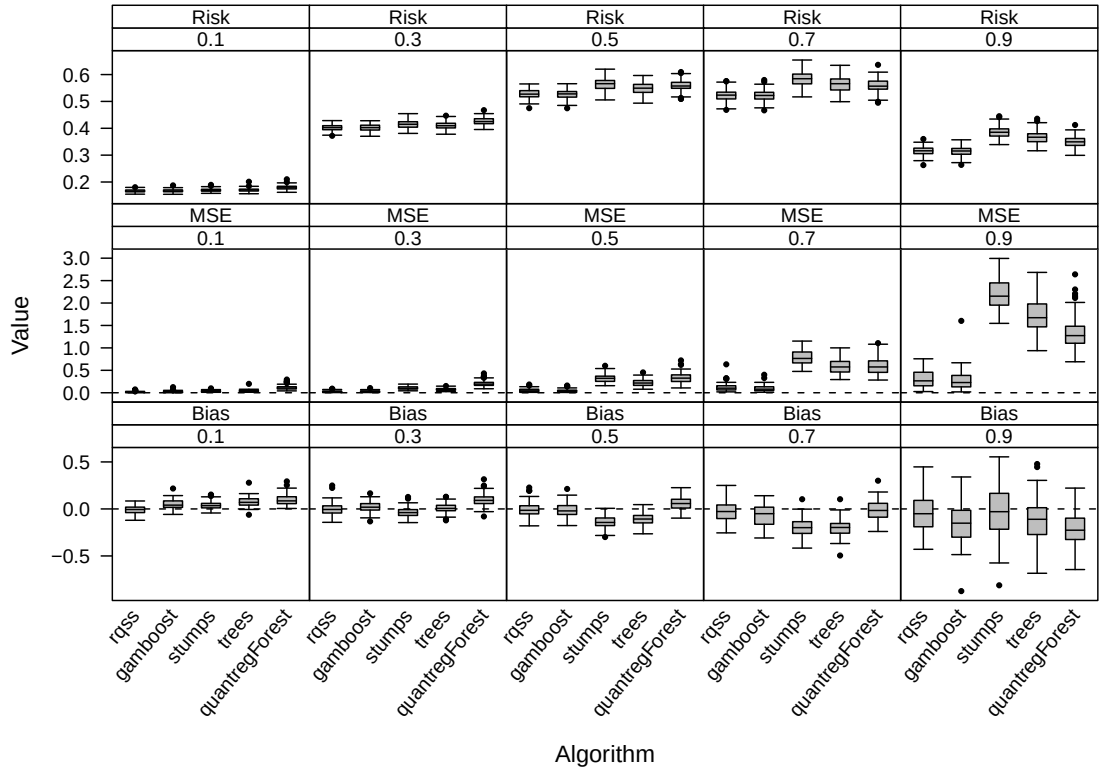


Figure 7: Simulation results for the 'sin' setup with gamma-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

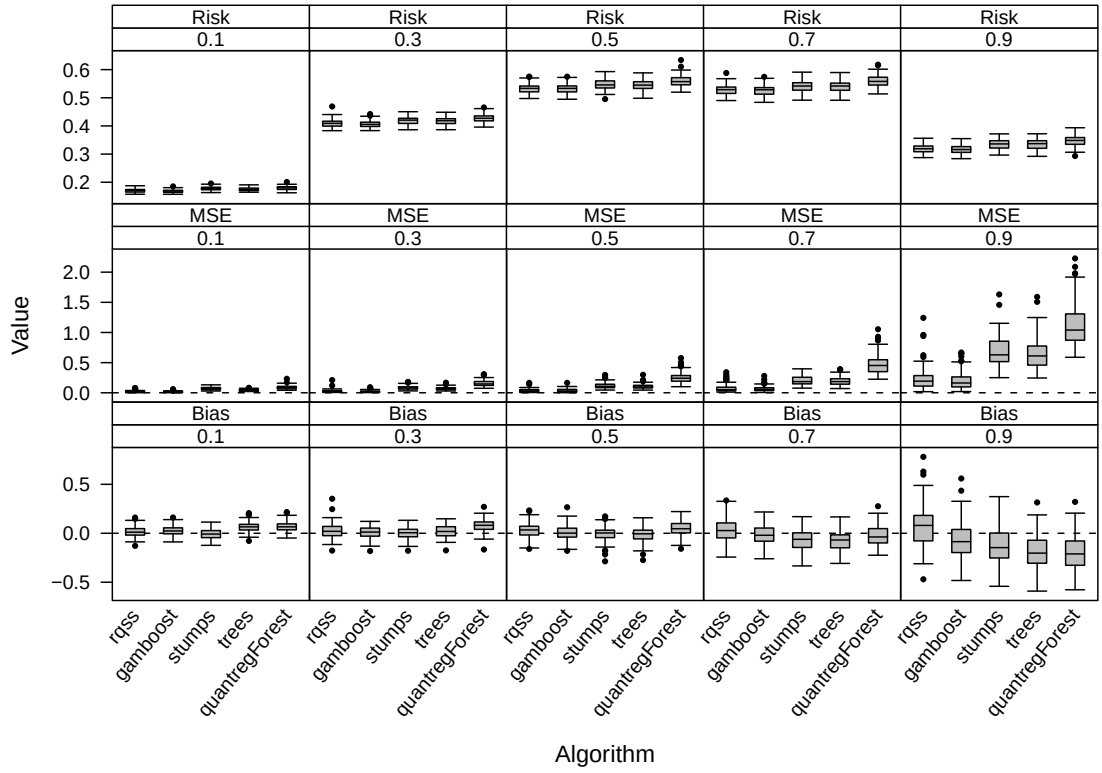


Figure 8: Simulation results for the 'log'-setup with gamma-distributed error terms. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

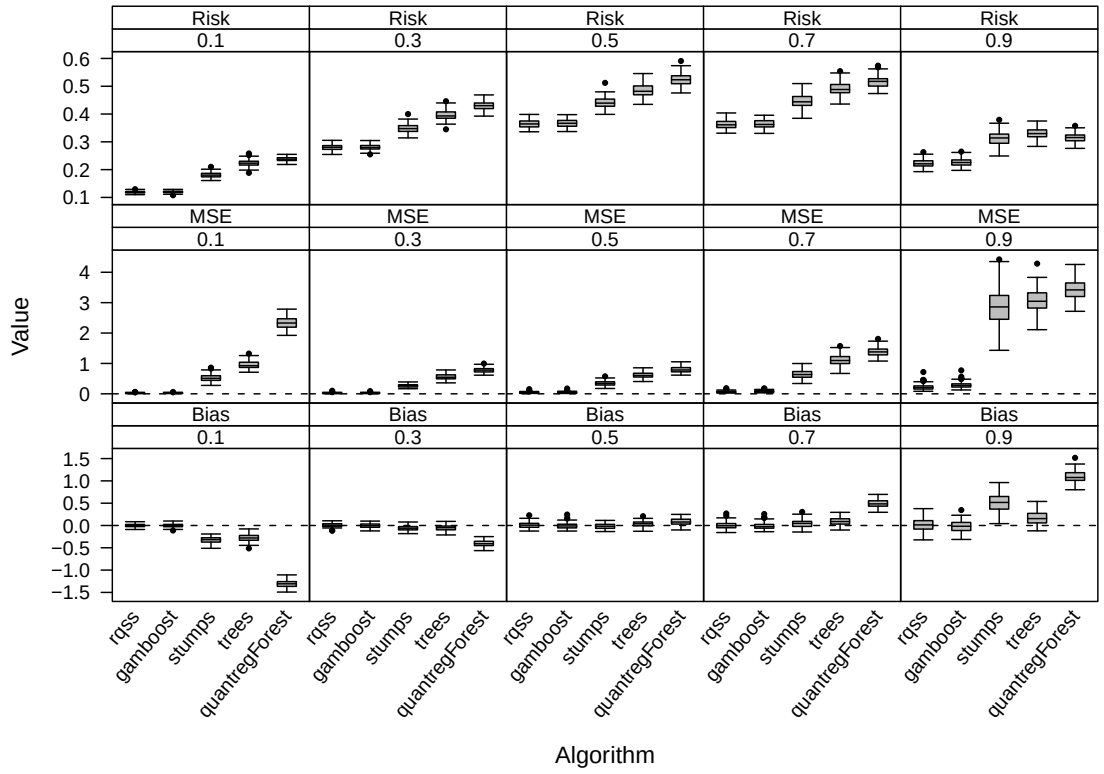


Figure 9: Simulation results for the multivariable setup with gamma-distributed error terms and a correlation coefficient of 0.0. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

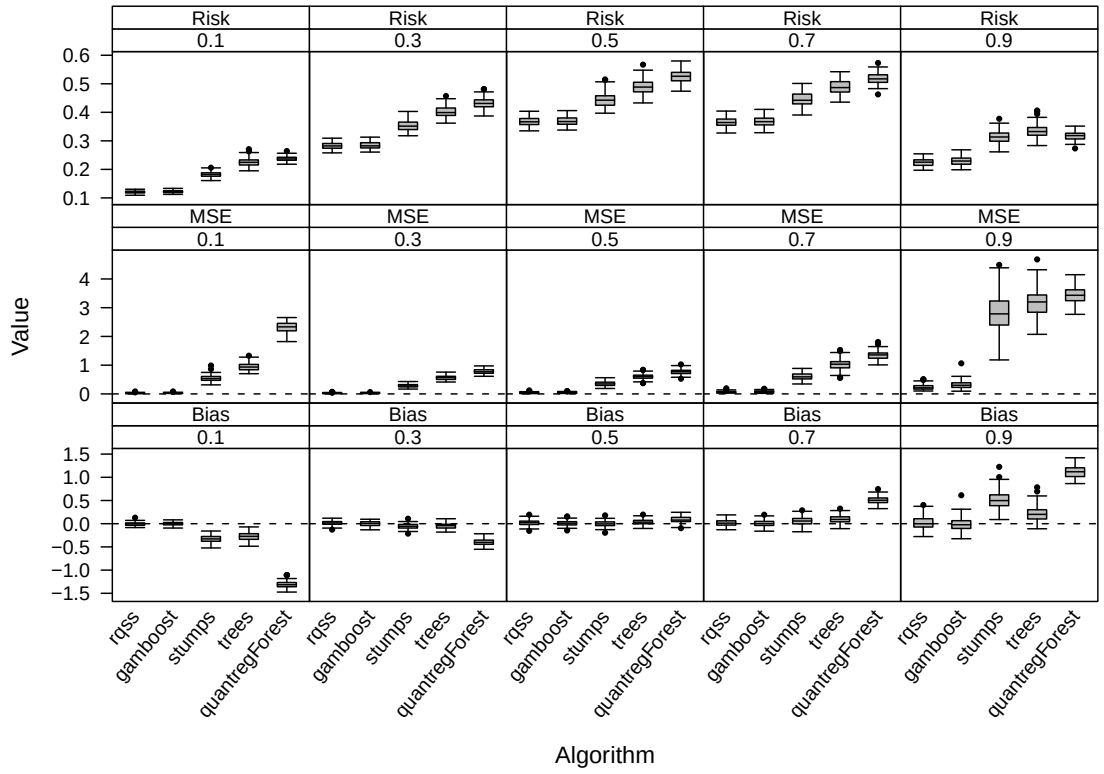


Figure 10: Simulation results for the multivariable setup with gamma-distributed error terms and a correlation coefficient of 0.2. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

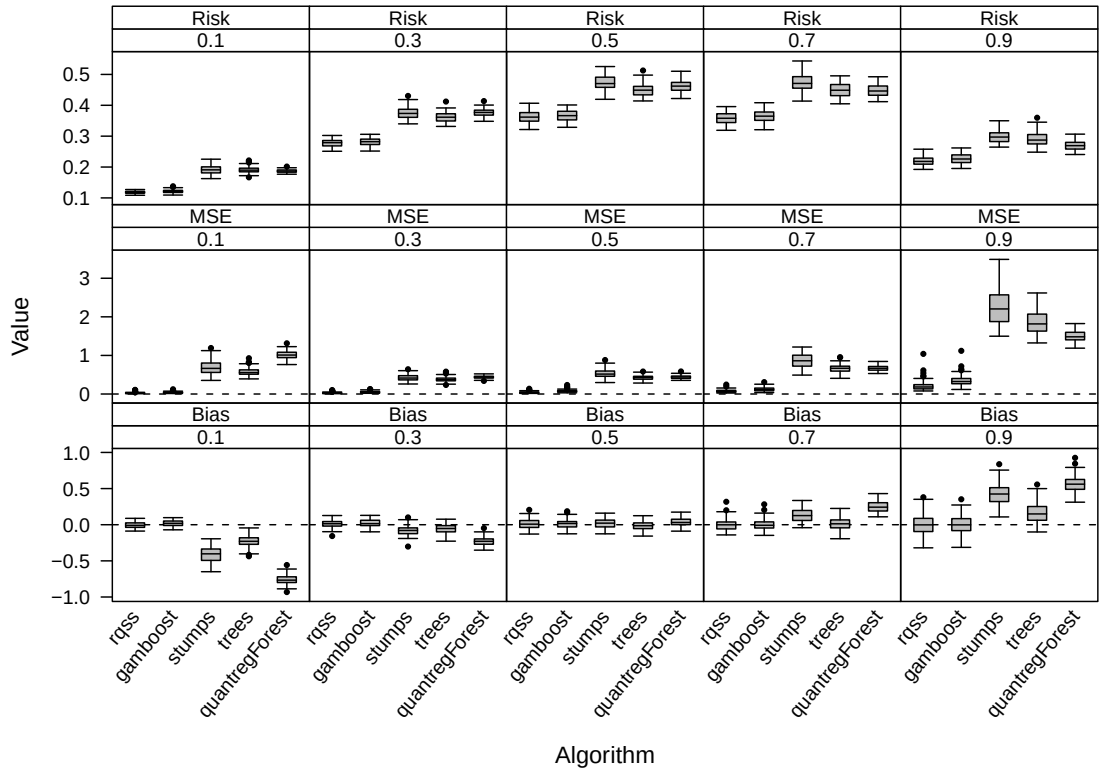


Figure 11: Simulation results for the multivariable setup with gamma-distributed error terms and a correlation coefficient of 0.8. Boxplots display the empirical distribution of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

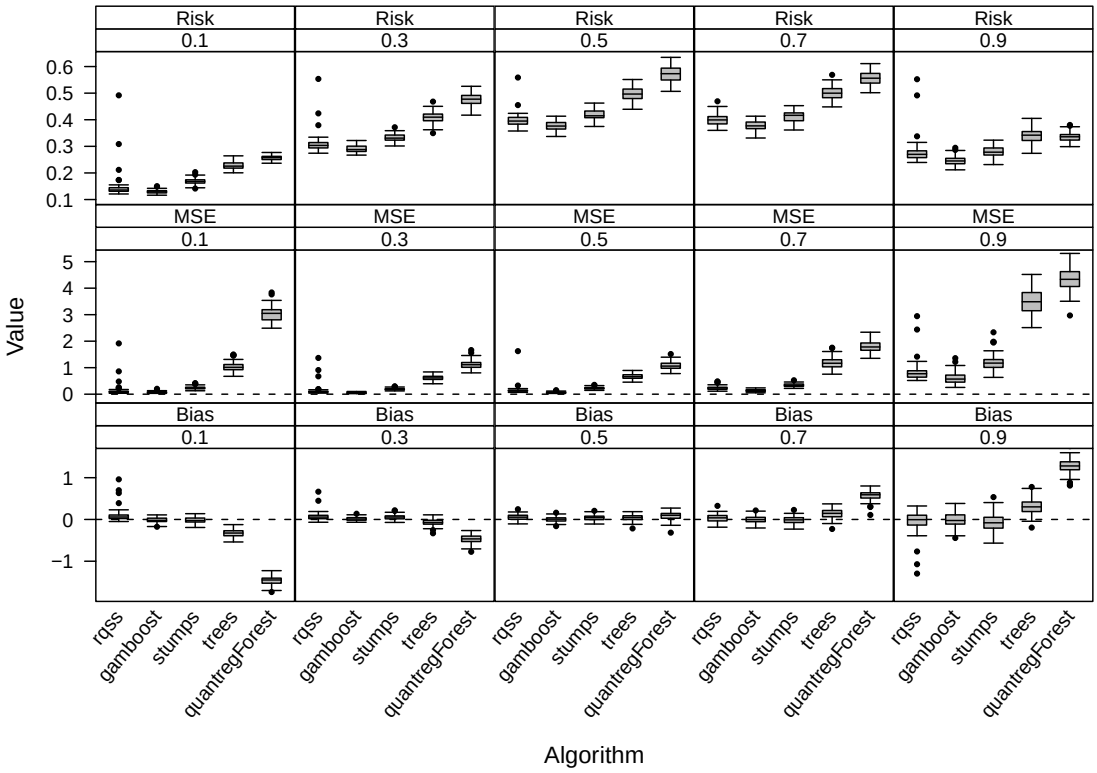


Figure 12: Simulation results for the higher-dimensional setup with $K = 20$ non-informative covariates, gamma-distributed error terms and a correlation coefficient of 0.5. Boxplots display empirical distributions of the performance criteria from 100 replications, depending on quantile τ and estimation algorithm.

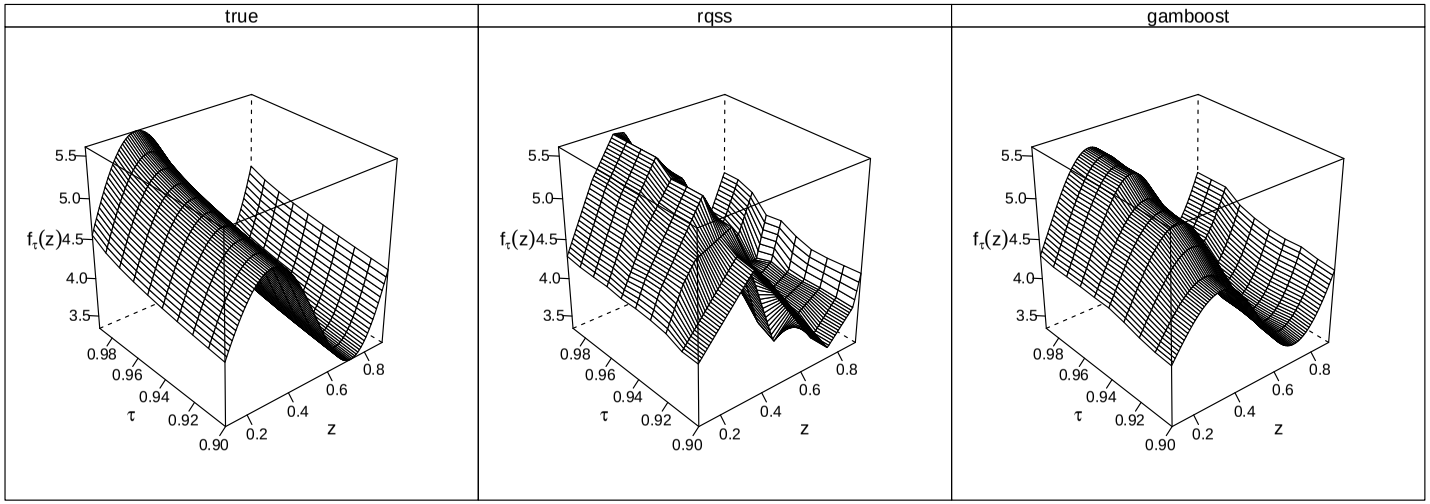


Figure 13: Example of true and estimated quantile functions for varying τ and z based on the simulation model in Section 3.2 of the manuscript. `gamboost` captures the smooth true quantile function better than `rqss`.

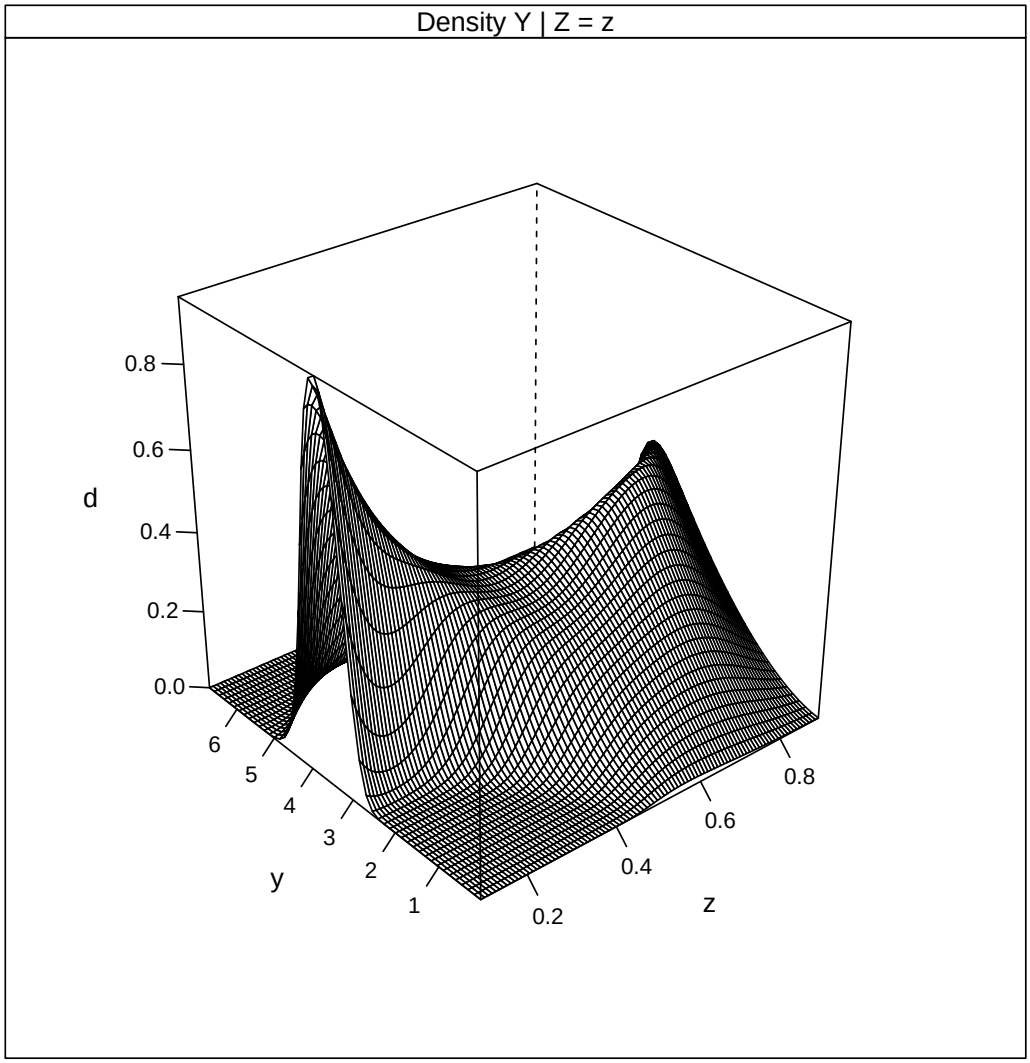


Figure 14: Conditional density $Y|Z = z$ for simulation in Section 3.2 of the manuscript. The first four moments depend on z in a non-linear smooth way.

B Empirical Evaluation: Linear Quantile Regression Model

In addition to the empirical evaluation for the additive quantile regression model, described in Section 3.1 of the manuscript, we addressed the special case of linear quantile regression with a separate linear simulation setup. With this simulation setup we wanted to check whether our boosting algorithm works in situations with linear effects on the response's quantile function. Our aim was in particular to compare the performance of our algorithm with well-established approaches for estimating linear quantile regression based on linear programming (as implemented in the function `rq()` from package **quantreg** Koenker, 2005). This linear simulation setup and its results will be described in the following.

Model. We considered the following location-scale-model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, \varepsilon_i \qquad \text{where} \quad \varepsilon_i \overset{iid}{\sim} H \quad \text{for } i = 1, \ldots, n \tag{1}$$

Here, the location as well as the scale of the response y_i depend in linear form on a covariate vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ and an error term ε_i with distribution function H_ε not depending on covariates. The coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ affects the response's location while $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ affects its scale. The resulting quantile function has a linear predictor structure and can be written as

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\boldsymbol{x}_i^\top \boldsymbol{\alpha})\, H^{-1}(\tau) = \boldsymbol{x}_i^\top (\boldsymbol{\beta} + \boldsymbol{\alpha} H^{-1}(\tau)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \ .$$

Hence, quantile specific coefficients can be determined as $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + \boldsymbol{\alpha} H^{-1}(\tau)$.

Based on the linear model in (1), we draw 100 data sets with the following parameter combinations:

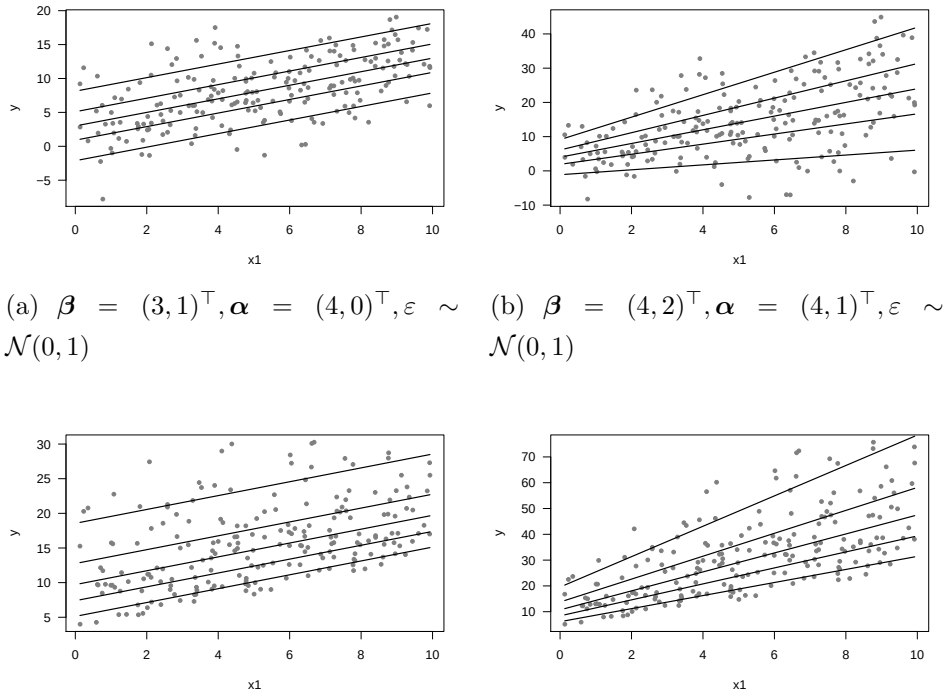
- Homoscedastic setup: $n = 200, \boldsymbol{\beta} = (3, 1)^\top, \boldsymbol{\alpha} = (4, 0)^\top$
- Heteroscedastic setup: $n = 200, \boldsymbol{\beta} = (4, 2)^\top, \boldsymbol{\alpha} = (4, 1)^\top$

- Multivariable setup: $n = 500, \boldsymbol{\beta} = (5, 8, -5, 2, -2, 0, 0)^\top, \boldsymbol{\alpha} = (1, 0, 2, 0, 1, 0, 0)^\top$

All required covariates were independently drawn from a continuously uniform distribution $\mathcal{U}[0, 10]$. We repeated all setups for three different distributions of the error terms: a standard normal distribution, a t-distribution with 2 degrees of freedom and a gamma distribution, where $\mathbb{E}(\varepsilon_i) = \mathbb{V}(\varepsilon_i) = 2$. Figure 15 visualizes data examples from the first two setups with one covariate for normal or gamma distributed error terms. Note that $\boldsymbol{\alpha} = (4, 1)$ leads to a heteroscedastic data structure where the quantile curves are no longer parallel shifted as for $\boldsymbol{\alpha} = (4, 0)$.

Estimation. For each of the generated data sets, we estimated the parameter vector $\boldsymbol{\beta}_\tau$ for a fixed quantile grid on $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ by our algorithm (function `glmboost()` from package **mboost**) and by linear programming (function `rq()` from package **quantreg**). In the boosting case, we fixed the step length at $\nu = 0.1$ and determined the optimal number of boosting iterations m_{stop} by evaluating the empirical risk on a test data set with 1000 observations drawn from the respective simulation setup and by choosing the point of minimal risk on the test data. We did not consider boosting trees, boosting stumps or quantile regression forests as competitors since these do not assume a linear model and would therefore naturally lead to a degraded fit when being compared to approaches that assume a linear model a priori.

As already mentioned in Section 2.2 of the manuscript, we decided to take the median as starting value for the intercept instead of the τ-th sample quantile of the response variable. This decision was based on the following empirical results. For quantiles smaller than $\tau = 0.5$, we explored hardly any differences between resulting m_{stop} criteria and estimators for $\boldsymbol{\beta}_\tau$ depending on the starting values. However, for quantiles larger than $\tau = 0.5$ the m_{stop} criterion was dramatically increased when taking the τ-th sample quantile as starting value. As an example, Figure 16 illustrates the stepwise approach of



(a) $\boldsymbol{\beta} = (3,1)^\top, \boldsymbol{\alpha} = (4,0)^\top, \varepsilon \sim \mathcal{N}(0,1)$

(b) $\boldsymbol{\beta} = (4,2)^\top, \boldsymbol{\alpha} = (4,1)^\top, \varepsilon \sim \mathcal{N}(0,1)$

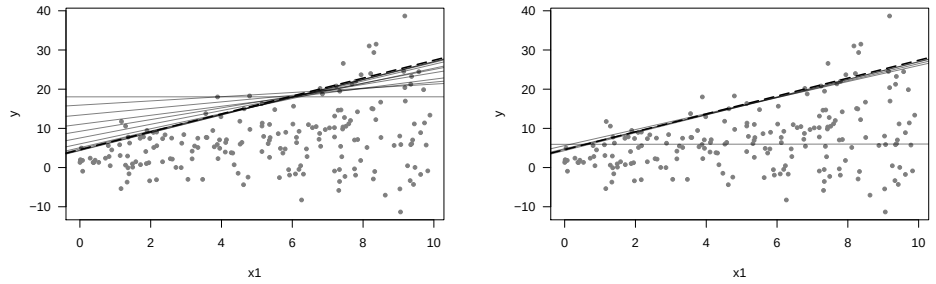
(c) $\boldsymbol{\beta} = (3,1)^\top, \boldsymbol{\alpha} = (4,0)^\top, \varepsilon \sim \mathcal{G}(1,2)$

(d) $\boldsymbol{\beta} = (4,2)^\top, \boldsymbol{\alpha} = (4,1)^\top, \varepsilon \sim \mathcal{G}(1,2)$

Figure 15: Data examples for linear simulation setups with $n = 200$ and one covariate in a homoscedastic (left) or heteroscedastic (right) data structure with normal (top) or gamma (bottom) distributed error terms. Lines designate true underlying quantile curves for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

the boosting estimation to the true underlying 90% quantile curves depending on the starting value. Note that it takes considerably more iterations until the estimation approaches the true quantile curve when beginning at the 0.9-th sample quantile, shown in Figure 16(a). On the contrary, Figure 16(b) displays that the estimation converges much faster when beginning at the median.

Performance results. In order to evaluate and to compare estimation



(a) Starting value = 90% quantile (horizontal line), $m_{\text{stop}} = 18474$

(b) Starting value = median (horizontal line), $m_{\text{stop}} = 7513$

Figure 16: Data example with parameters $n = 200$, $\boldsymbol{\beta} = (2, 1)^\top$ and $\boldsymbol{\alpha} = (2, 1)^\top$ and normal-distributed error terms. Dashed black lines show the true underlying quantile curve for $\tau = 0.9$, grey lines illustrate the stepwise boosting fit after each 2000 iterations beginning at the horizontal line.

results of the two considered algorithms, we estimated Bias and MSE for each quantile specific parameter $(\beta_{\tau 0}, \beta_{\tau 1}, \ldots, \beta_{\tau p})^\top$ by the following formulae:

$$\text{Bias}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau kj} - \beta_{\tau j}) \ , \qquad \text{MSE}(\hat{\beta}_{\tau j}) = \frac{1}{100} \sum_{k=1}^{100} (\hat{\beta}_{\tau kj} - \beta_{\tau j})^2 \ , \tag{2}$$

where $k = 1, \ldots, 100$ indexes the simulation replication and $j = 0, \ldots, p$ the number of covariates. Note that when the mean bias and MSE over all 100 iterations are calculated, those values can be interpreted as Monte Carlo estimators of the true bias and MSE of the non-linear functions. In case of boosting, we also considered the m_{stop} criteria.

In the following, we will focus on a short summary of the results by just showing some typical examples. Figure 17 displays boxplots for the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$ in the heteroscedastic setup with normal-distributed error terms. Note that estimators resulting from linear programming (`rq`)

are less biased but have a larger variance than those resulting from boosting (`boost`). This is consistent to previously reported results and to the fact that boosting estimators are usually shrunken towards zero, which can be traced back to the implicit regularization property of boosting estimation (Bühlmann and Hothorn, 2007).

Regarding the MSE, Table 1 shows estimators for setups with one covariate and gamma-distributed error terms, obtained according to (2). For the slope estimator $\hat{\beta}_{\tau 1}$, boosting achieves smaller MSE estimators on almost the whole quantile grid. Concerning the intercept estimator $\hat{\beta}_{\tau 0}$, boosting performs better in the homoscedastic setup while linear programming obtains better results in the heteroscedastic setup.

Table 1: Estimated MSE criteria from 100 replications of linear simulation setups with one covariate and gamma distributed error terms. Shown in bold are quantile and parameter specific smaller estimators.

	Homoscedastic setup				Heteroscedastic setup			
	MSE($\beta_{\tau 0}$)		MSE($\beta_{\tau 1}$)		MSE($\beta_{\tau 0}$)		MSE($\beta_{\tau 1}$)	
τ	rq	boost	rq	boost	rq	boost	rq	boost
0.1	**0.328**	0.350	0.010	**0.008**	**0.762**	1.007	0.050	**0.038**
0.3	0.676	**0.582**	0.016	**0.012**	**1.417**	1.475	0.063	**0.052**
0.5	0.732	**0.685**	0.020	**0.015**	**1.627**	1.962	0.099	**0.074**
0.7	1.751	**1.595**	0.048	**0.040**	4.168	**4.165**	0.229	**0.157**
0.9	4.983	**2.992**	0.129	**0.066**	**10.404**	17.971	**0.618**	0.657

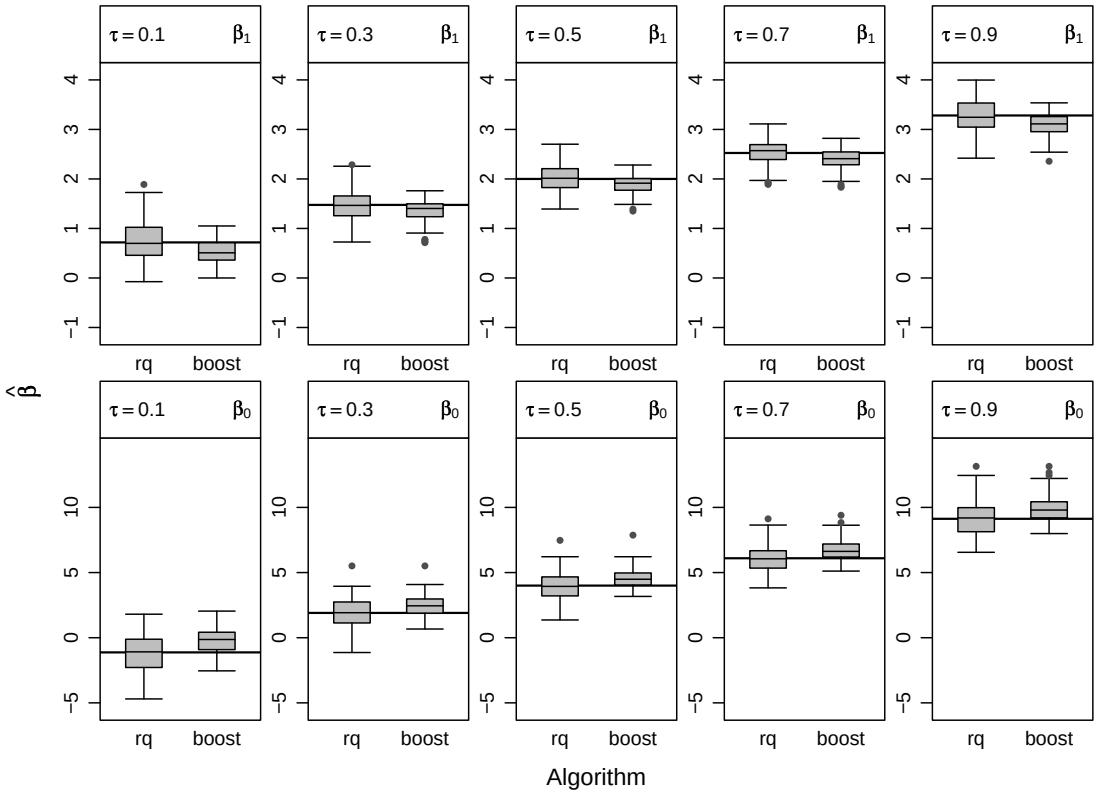


Figure 17: Simulation results for heteroscedastic linear setup with one covariate and normal-distributed error terms. Boxplots display the empirical distribution of the estimated parameters $(\hat{\beta}_{\tau 0}, \hat{\beta}_{\tau 1})^\top$ from 100 replications, depending on quantile τ and estimation algorithm (`rq` for linear programming and `boost` for boosting). Horizontal lines designate true underlying parameters $(\beta_{\tau 0}, \beta_{\tau 1})^\top$.

In addition, Table 2 shows mean m_{stop} criteria for all setups with t-distributed error terms. The optimal number of boosting iterations, determined by means of test data, ranges roughly between 3000 and 10000 in cases with one covariate and is considerably increased (30000 – 70000) for

the multivariable model with six covariates.

Table 2: Mean m_{stop} criteria from 100 replications of linear simulation setups with t-distributed error terms.

Setup	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Homoscedastic	10886.29	6415.06	7371.50	4762.08	3589.35
Heteroscedastic	4183.34	8935.24	9133.33	10038.65	6387.43
Multivariable	68054.97	43882.78	40540.50	42316.77	30254.83

We observed similar results for all other simulation setups, i.e., with more covariates or alternative error distributions. Therefore, we conclude that boosting estimation is competitive to linear programming estimation in situations with linear effects on the response's quantile function, i.e., when linear quantile regression is appropriate.

C Additional Results from Analyzing Childhood Malnutrition in India with `gamboost()`

This section contains additional results for the additive model in the analysis of the India malnutrition data based on the function `gamboost()`, see Section 4 of the manuscript for estimation details.

All effects plotted in Figure 18–28 are adjusted for the overall quantile levels by inserting the average values for the remaining continuous covariates and the reference category for the remaining categorical covariates.

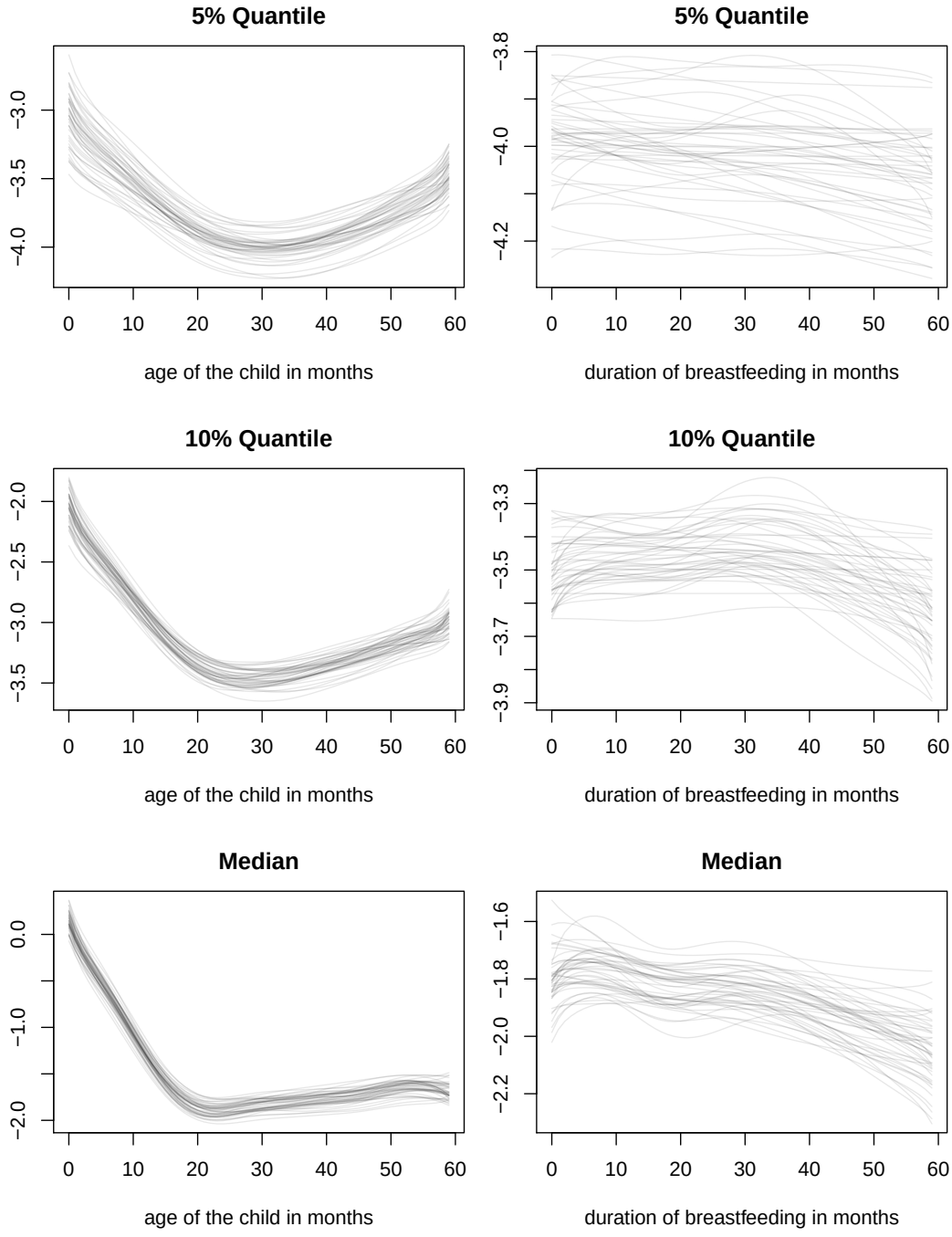


Figure 18: Quantile-specific estimated non-linear effects for age of the child (left) and duration of breastfeeding (right). Each gray line corresponds to one curve estimated by `gamboost` for one of the 50 samples.

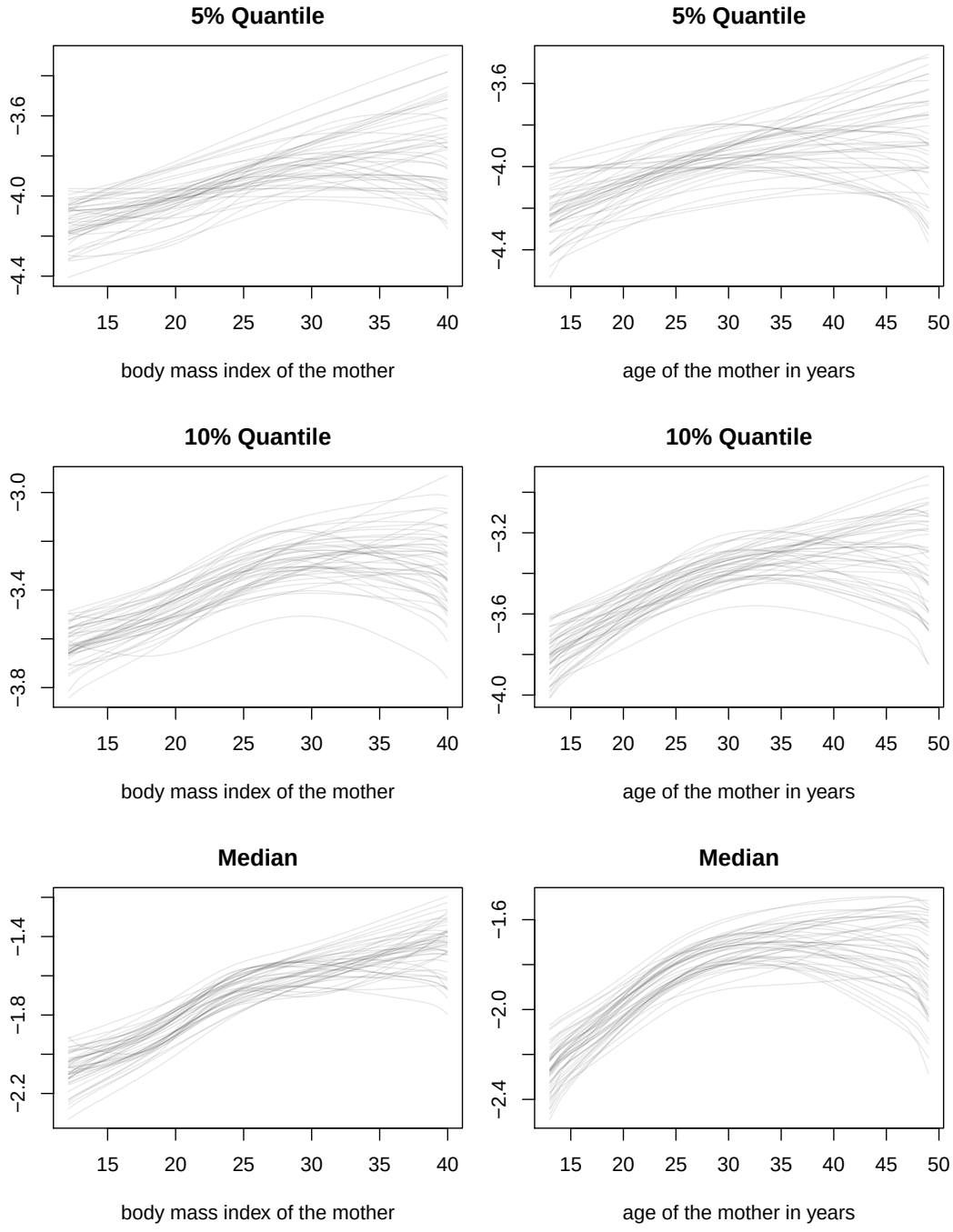


Figure 19: Quantile-specific estimated non-linear effects for body mass index of the mother (left) and age of the mother (right). Each gray line corresponds to one curve estimated by `gamboost` for one of the 50 samples.

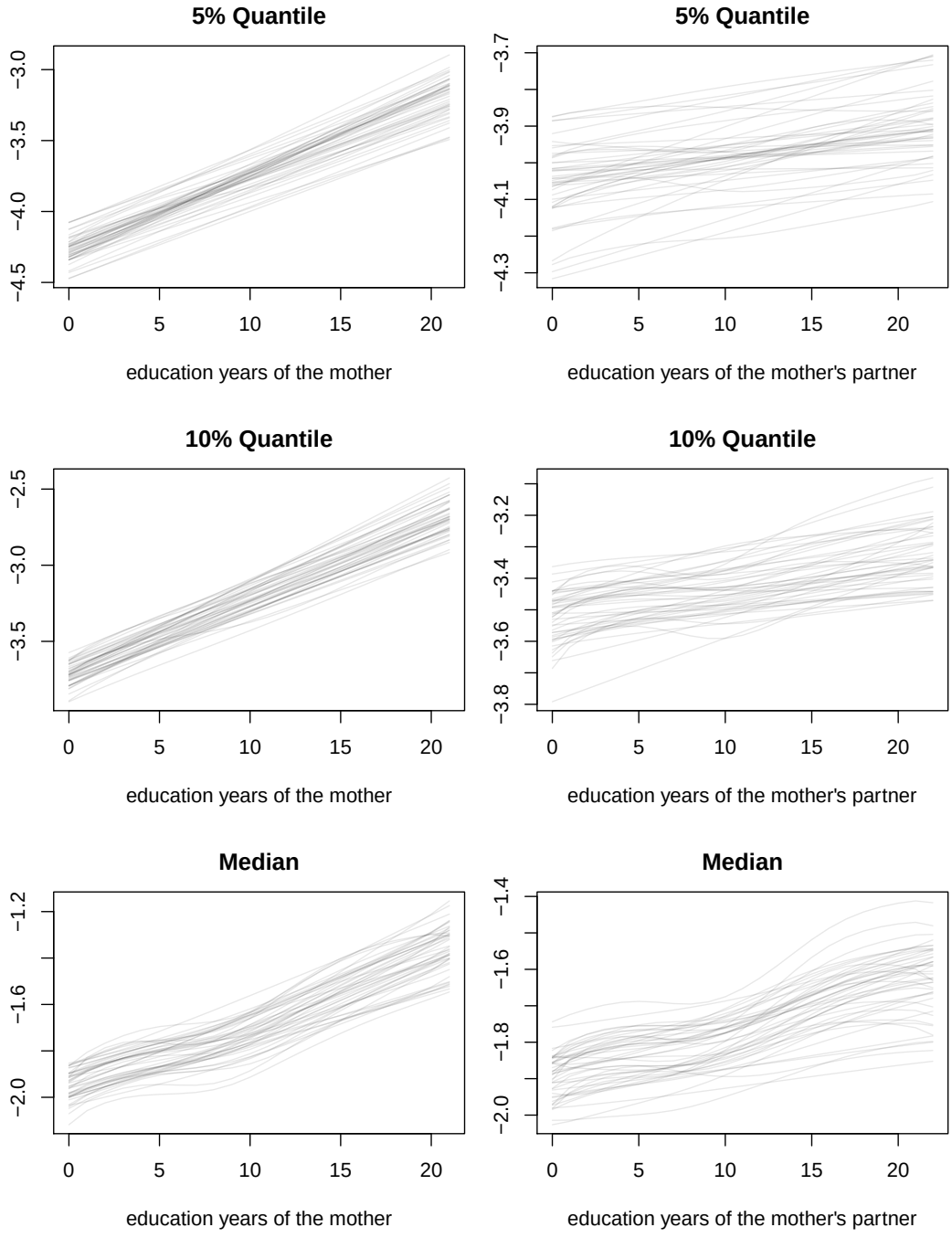


Figure 20: Quantile-specific estimated non-linear effects for education years of the mother (left) and education years of the mother's partner (right). Each gray line corresponds to one curve estimated by `gamboost` for one of the 50 samples.

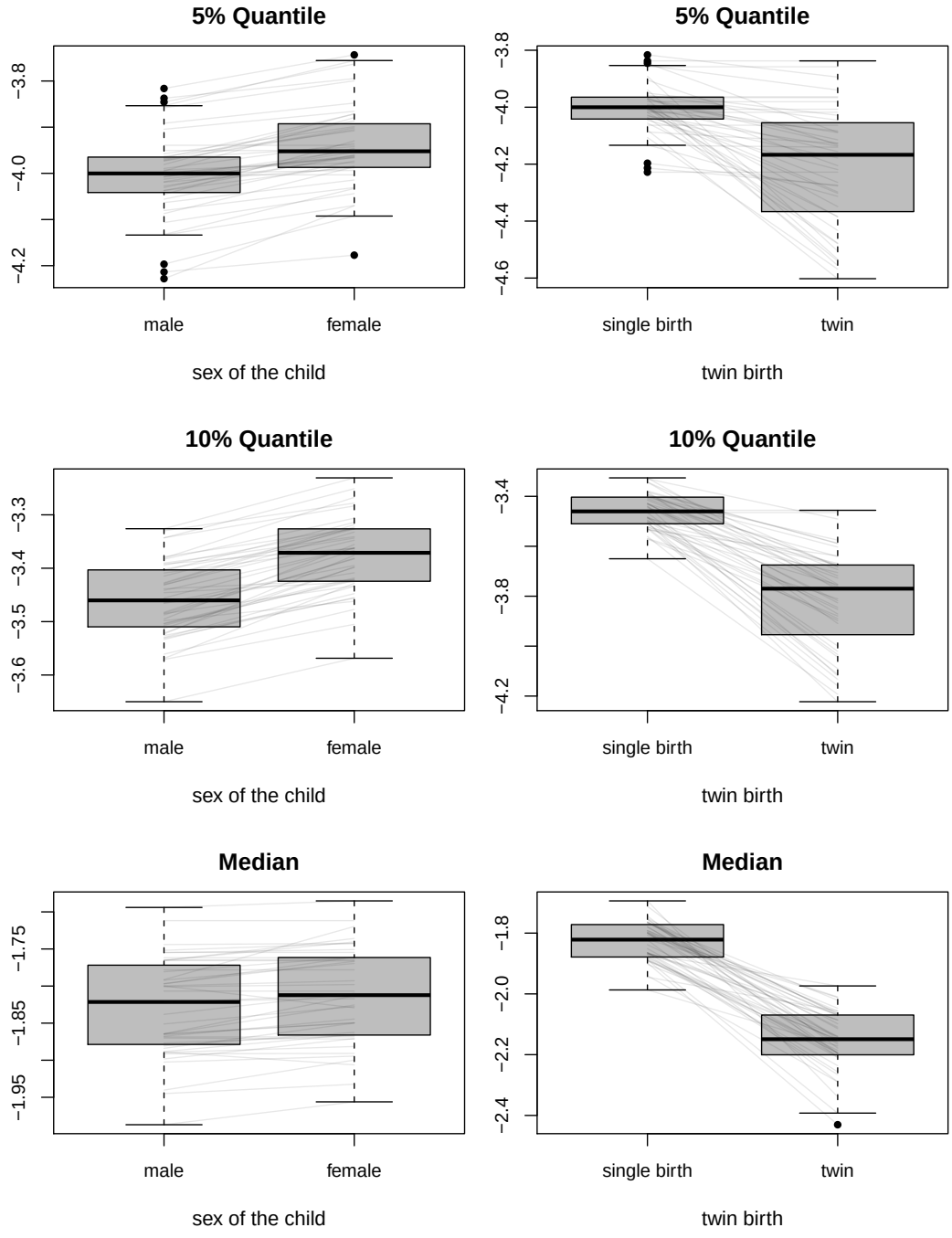


Figure 21: Quantile-specific estimated effects for sex of the child (left) and twin birth (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

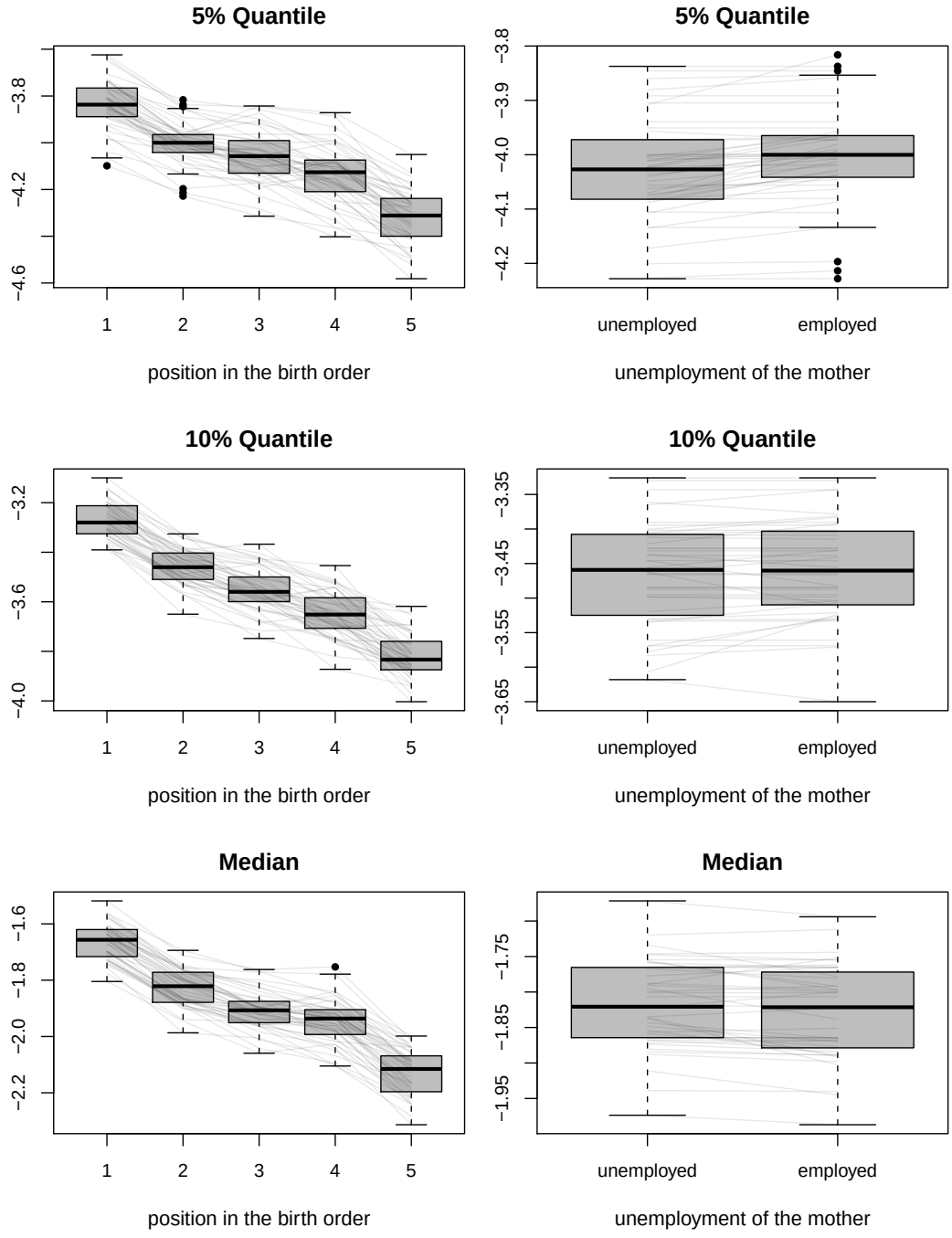


Figure 22: Quantile-specific estimated effects for position in birth order (left) and unemployment of the mother (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

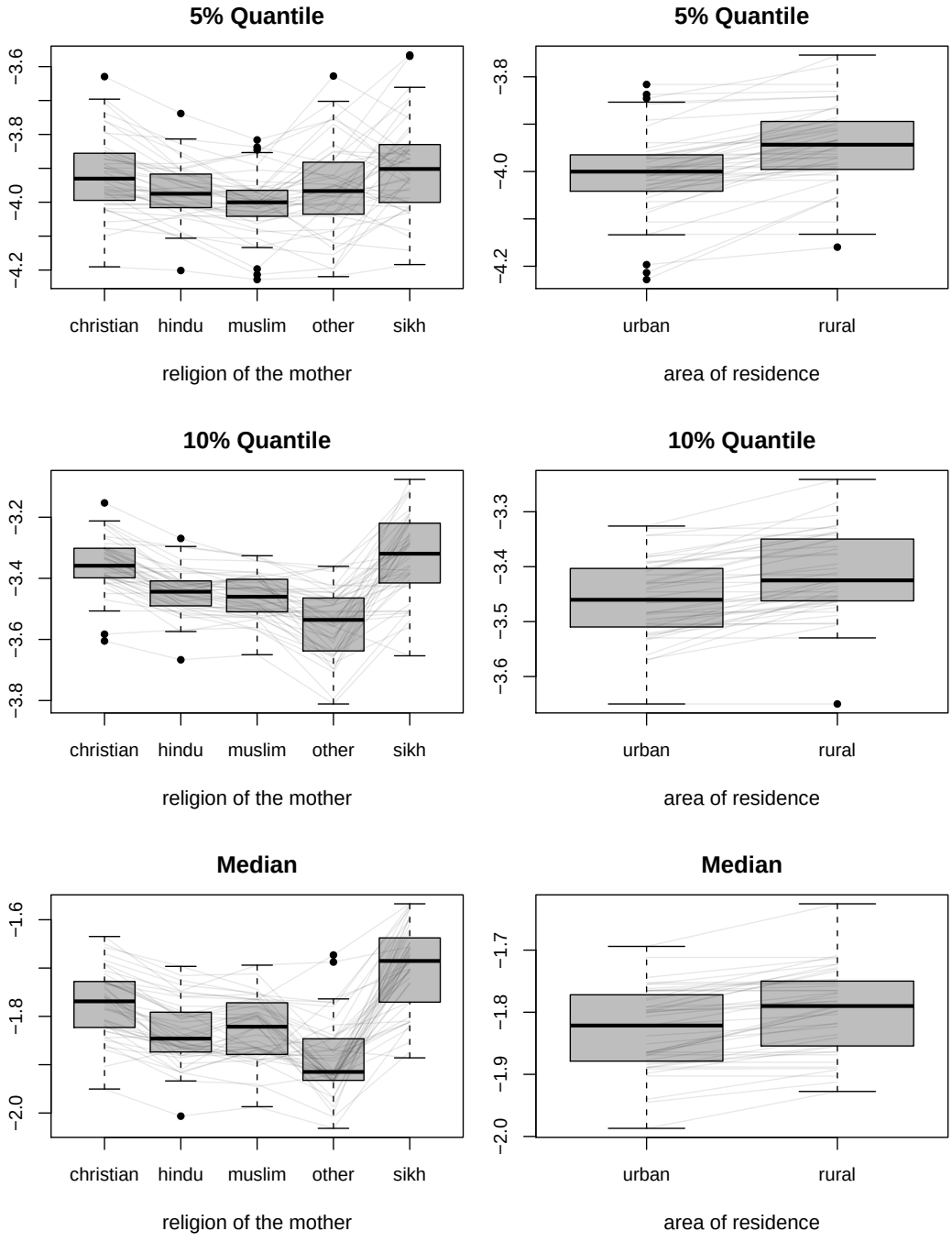


Figure 23: Quantile-specific estimated effects for religion of the mother (left) and area of residence (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

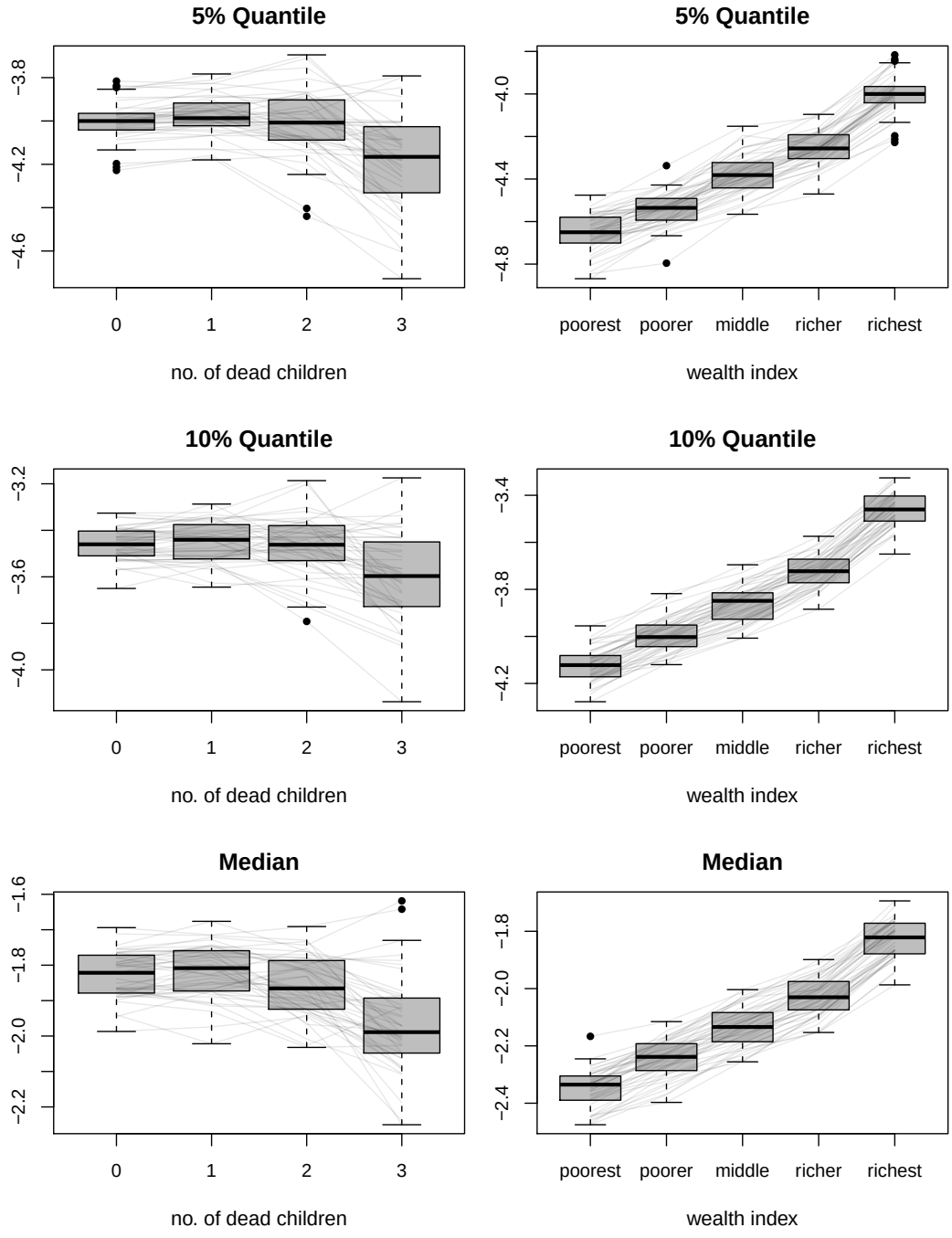


Figure 24: Quantile-specific estimated effects for number of dead children (left) and wealth index (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

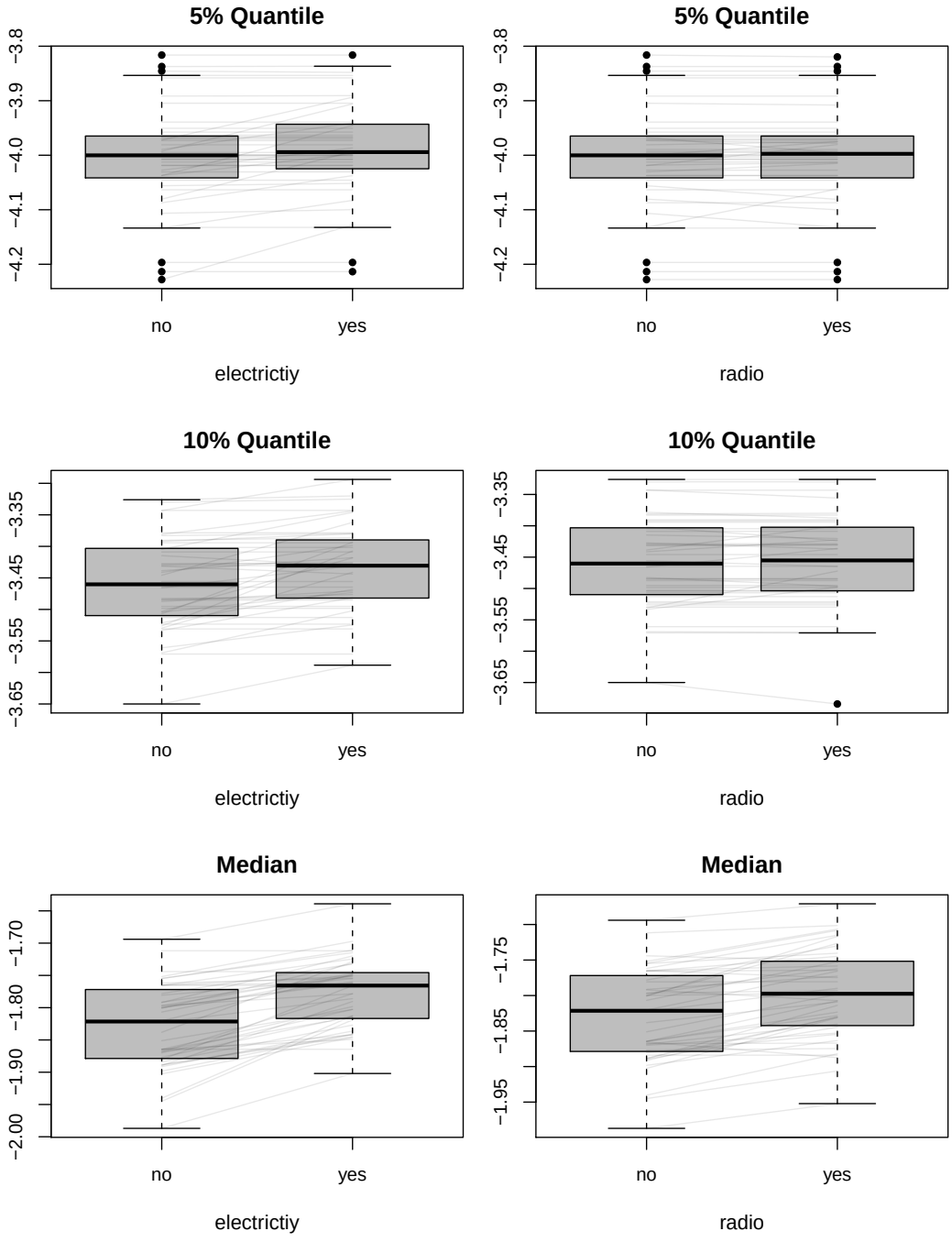


Figure 25: Quantile-specific estimated effects for electricity (left) and radio (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

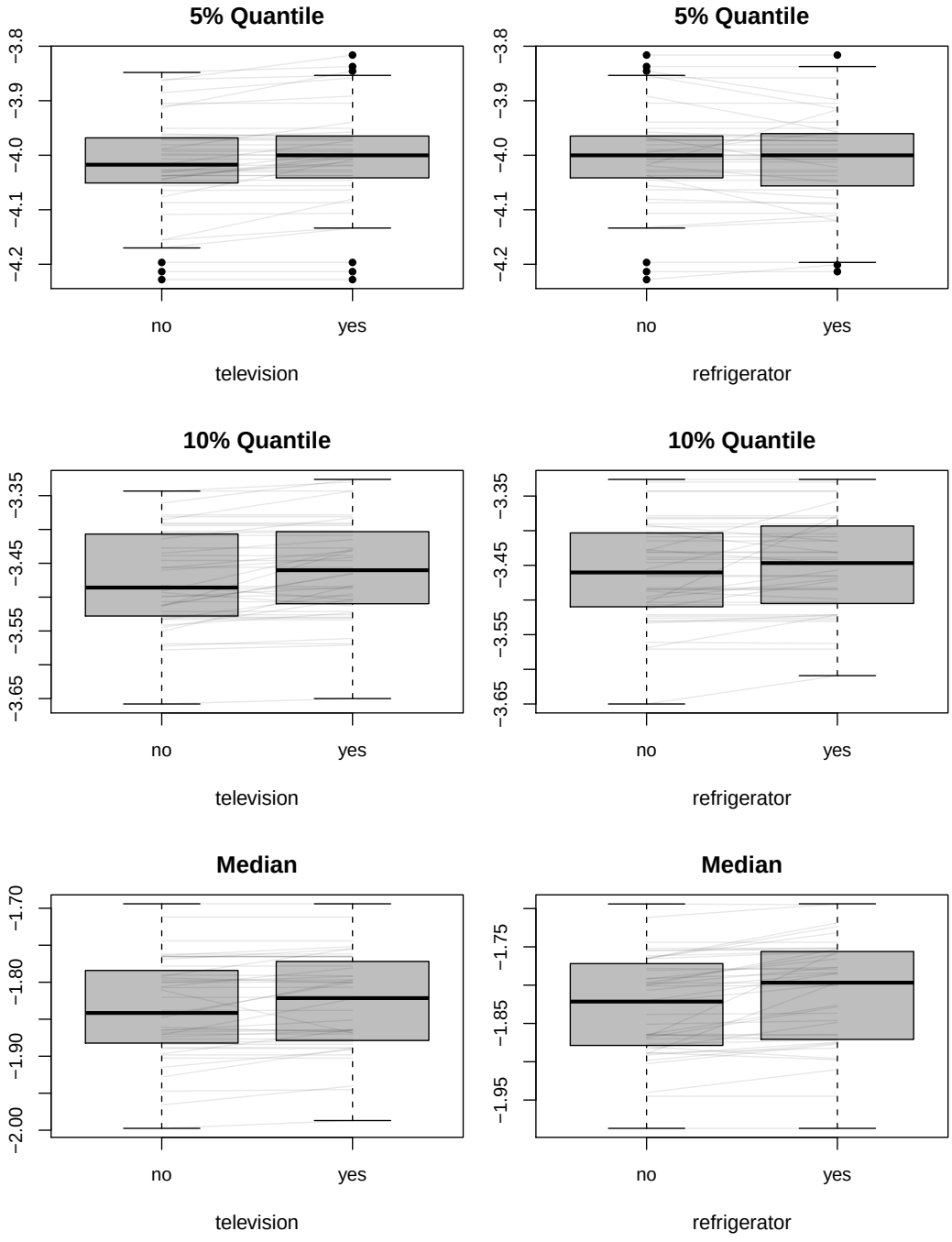


Figure 26: Quantile-specific estimated effects for television (left) and refrigerator(right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

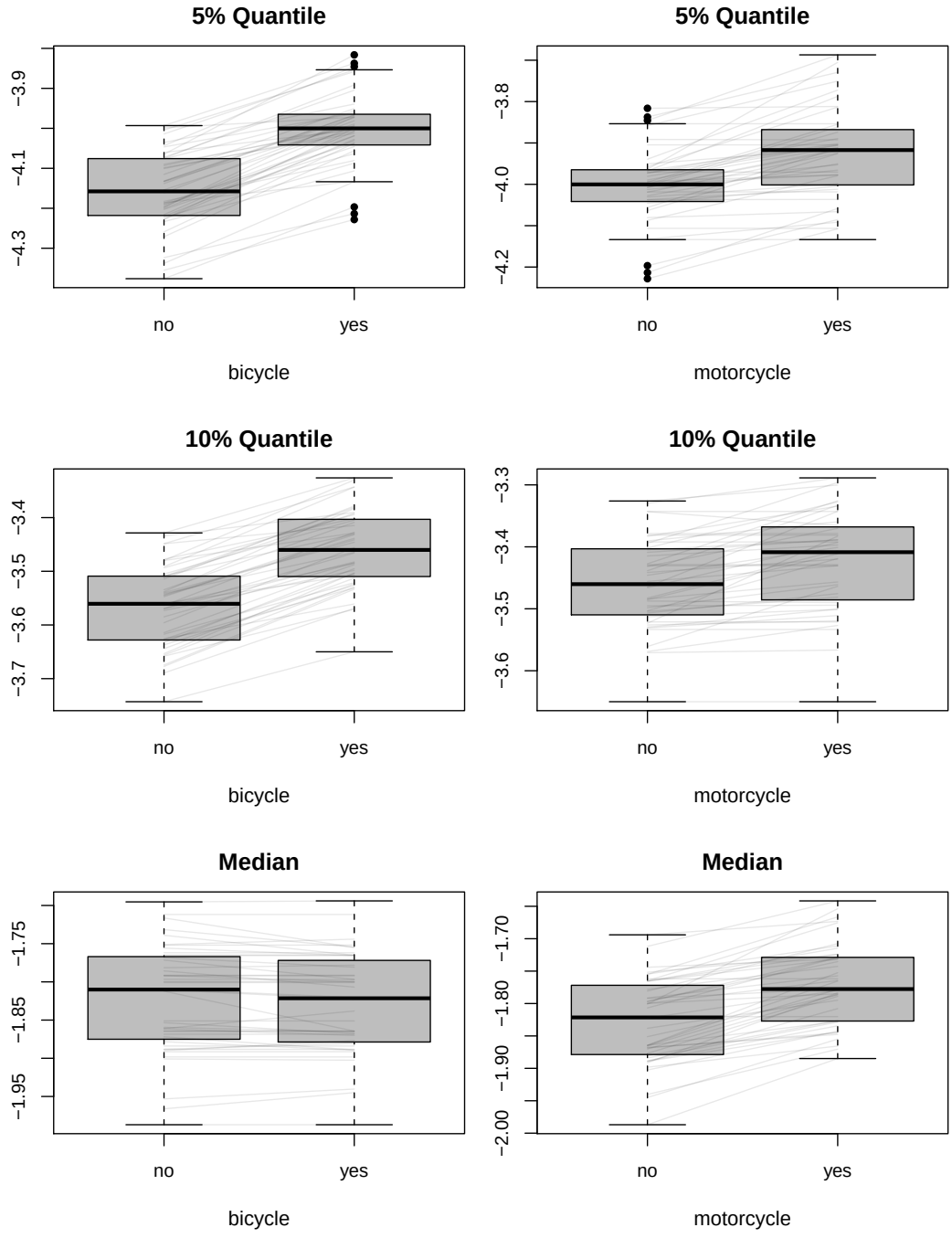


Figure 27: Quantile-specific estimated effects for bicycle (left) and motorcycle (right). Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

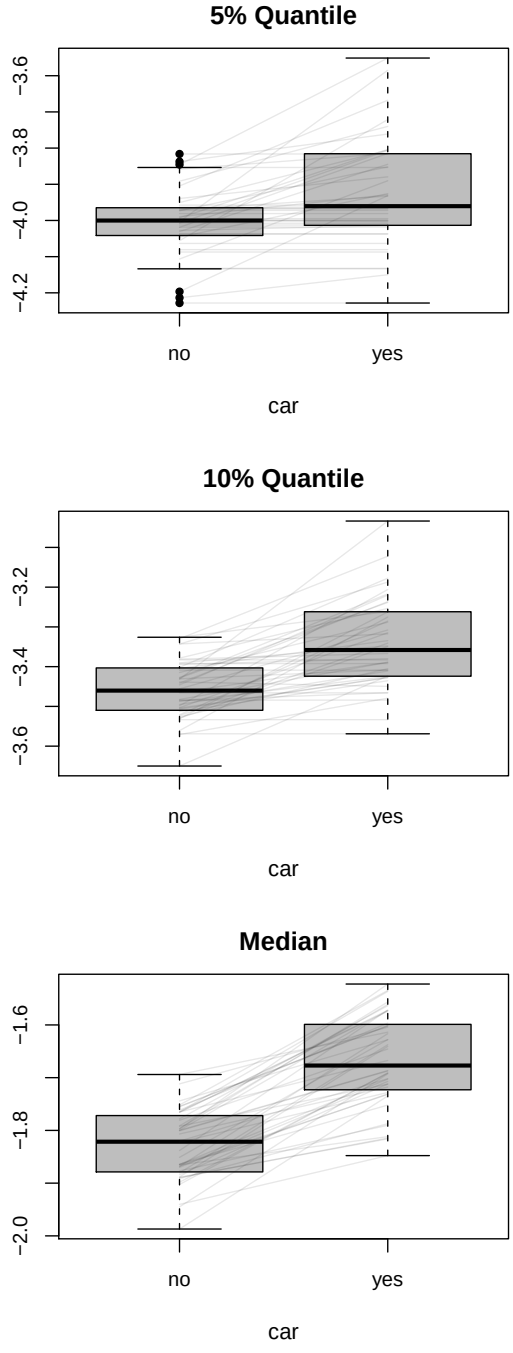


Figure 28: Quantile-specific estimated effects for car. Boxplots display the empirical distribution of effects estimated by `gamboost` obtained from the 50 samples.

D Additional Results from Analyzing Childhood Malnutrition in India with `rqss()`

This section contains computational details for the India malnutrition analysis with `rqss()`. In addition, results from an `rqss()` analysis on the complete India dataset with an informal choice of the tuning parameters are presented.

For the India malnutrition analysis in Section 4 of the manuscript, model fits were compared by regarding the empirical risks on the evaluation parts of the 50 different data splits. Note that with `rqss()`, predictions can only be obtained for observations of continuous covariates lying strictly within the convex hull of the fitting data. We therefore restricted each evaluation data to the convex hull of its corresponding fitting data and then calculated the empirical risk on the reduced evaluation data for all algorithms. This led to reduced evaluation sample sizes lying between 6440 and 6660 instead of the original sample size of $\frac{37623}{3} = 12541$.

Figure 29 shows empirical risks on the 50 complete evaluation datasets – without results for `rqss()`. One can see that the order of the algorithm performances remains the same as for the reduced evaluation datasets in Figure 8 of the manuscript, even though the absolute values of the risks seem to be higher for complete datasets.

Note that there are a number of replications where the differences between the additive model and the varying coefficient model are rather low. This is not too surprising since the additive model is contained in the varying coefficient model as a special case. If the boosting iteration stops early enough, it is therefore possible that results from the additive model and the varying coefficients model coincide.

Another difficult task when conducting the India analysis with `rqss()` is the choice of the smoothing parameters $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$. We determined them based on the test parts of the 50 different data splits. Since we included

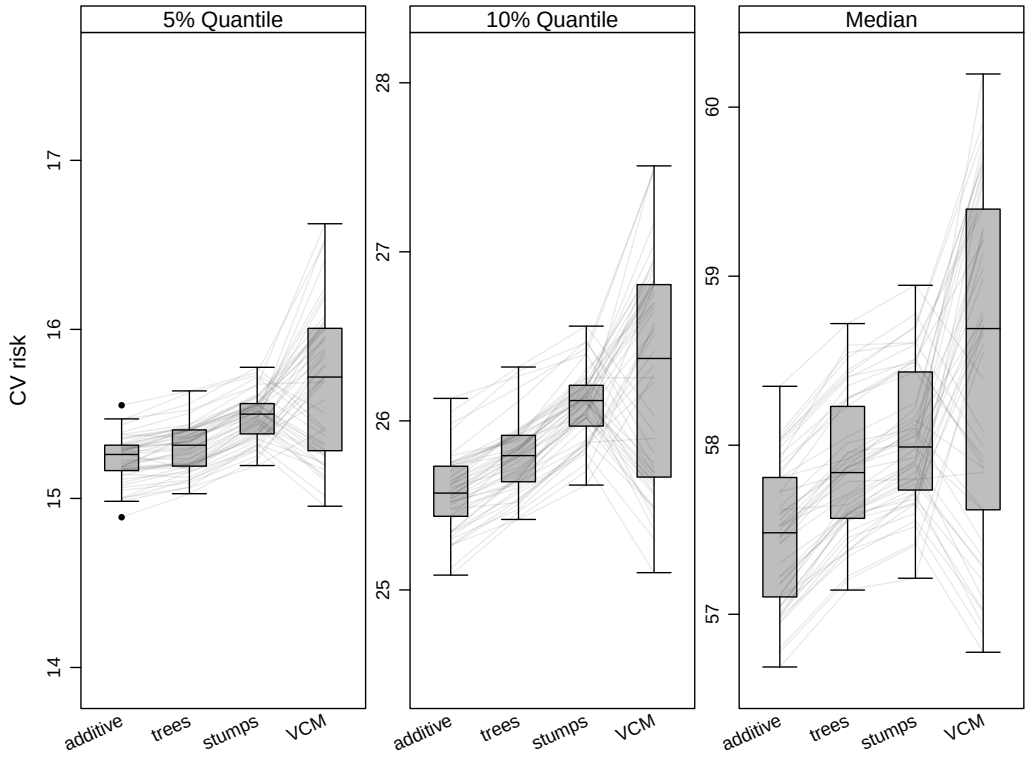


Figure 29: Boxplots display empirical distributions of the cross-validated empirical risks for the evaluation parts of the 50 data splits. Results for one split are connected by gray lines.

six continuous covariates with potentially non-linear effects, we had to optimize the six-dimensional vector of $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$. Figure 30 shows the empirical distribution of optimized parameters $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$ on the 50 different test parts with starting parameters of $\lambda_{\text{cage},0} = 10, \ldots, \lambda_{\text{medupart},0} = 10$. Figure 31 contains the empirical λ distributions for starting parameters of $\lambda_{\text{cage},0} = 100, \ldots, \lambda_{\text{medupart},0} = 100$. One can see that the optimized smooth-

ing parameters strongly depend on the starting parameters. Therefore, these "optimized" parameters have to be interpreted with care.

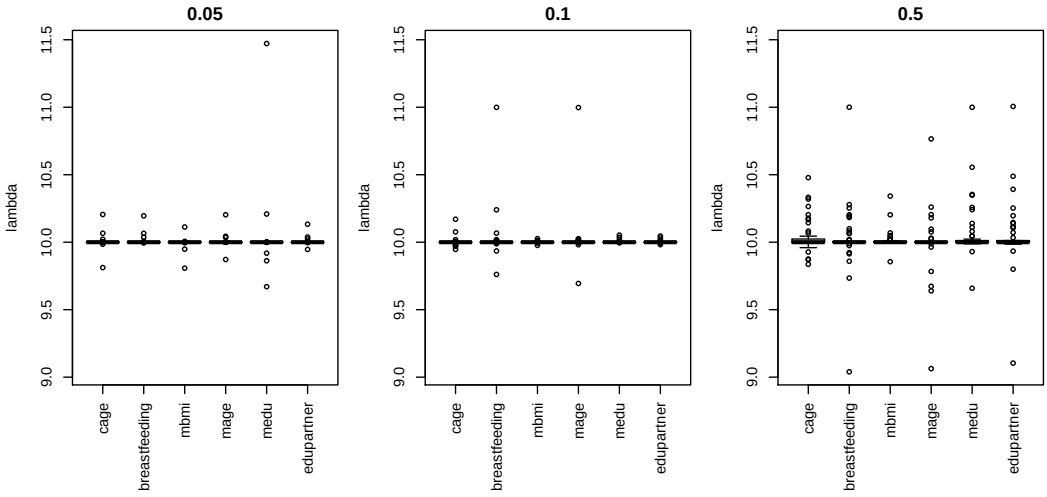


Figure 30: Boxplots display empirical distributions of the optimized parameters $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$ on the 50 different test parts with starting parameters of $\lambda_{\text{cage},0} = 10, \ldots, \lambda_{\text{medupart},0} = 10$.

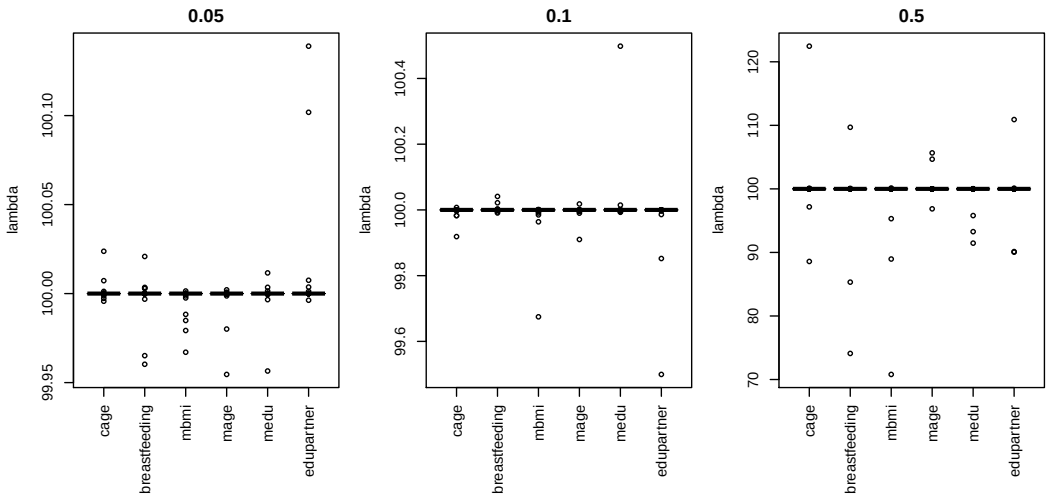


Figure 31: Boxplots display empirical distributions of the optimized parameters $\lambda_{\text{cage}}, \ldots, \lambda_{\text{medupart}}$ on the 50 different test parts with starting parameters of $\lambda_{\text{cage},0} = 100, \ldots, \lambda_{\text{medupart},0} = 100$.

In addition to the models on the 50 data splits, we fitted a saturated model on the complete India data. The choice of the tuning parameters was based on Koenker (2010). For the 5%-quantile, the R-specification of the model is given by the following command:

```
rqssIndia <- rqss(stunting ~ qss(cage,lambda = 20) +
                  qss(breastfeeding,lambda = 80) +
                  qss(mbmi, lambda = 80) + qss(mage, lambda = 80) +
                  qss(medu, lambda = 80) + qss(edupartner, lambda = 80) +
                  munemployed + csex + ctwin + cbirthorder + mreligion +
                  mresidence + deadchildren + wealth + electricity +
                  radio + television + refrigerator + bicycle +
                  motorcycle + car,
                  tau = .05, method = "fn", data = india)
```

The resulting estimated effects are shown in Figure 32 and in Figure 33. Note that all plotted effects are centered around the average effect of all further continuous covariates and that the reference category has been inserted for categorical covariates to make the levels comparable. For the continuous covariates, the estimated effects are either linear or piecewise linear functions. For the categorical covariates, the estimated coefficients from `rqss()` are qualitatively similar to the `gamboost()` results. However, removing variables from this saturated model would require a formal variable selection procedure, e.g., based on AIC- or SIC-type criteria, which is currently not available for `rqss()`. When using `rqss()` for model fitting, it is also not possible to separate the linear and non-linear contributions of one covariate effect, as can be done in `gamboost`.

In a recent case study for the India data based on additive quantile regression, Koenker (2010) uses the unstandardized height of the children as the response variable. For this reason, we cannot directly compare our results to those described in Koenker (2010), also because we provide results from a formal variable selection procedure offered by boosting in combination with stability selection.

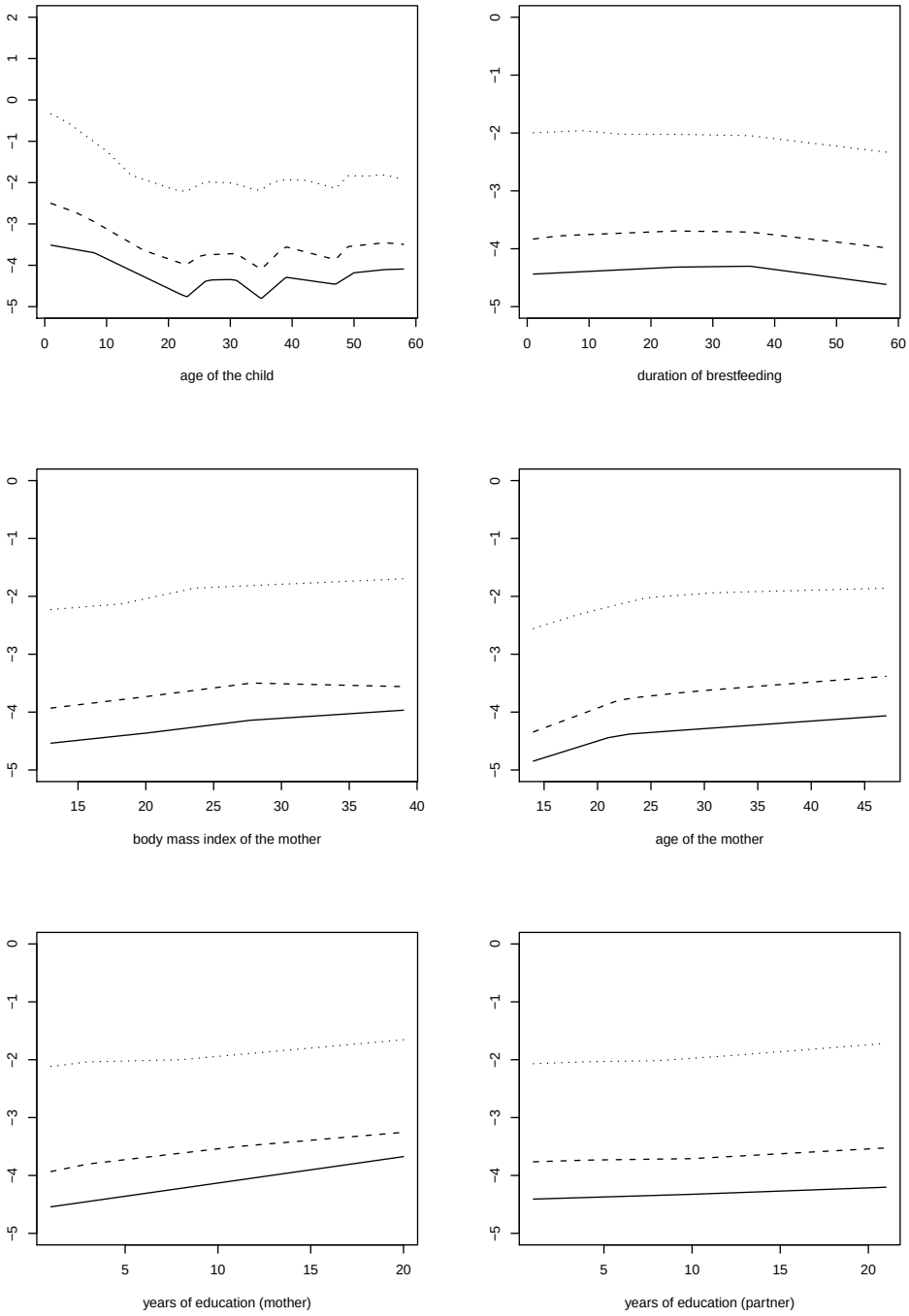


Figure 32: Estimated non-linear effects for the 50% quantile (dotted line), the 10% quantile (dashed line) and the 5% quantile (solid line) of the stunting score obtained by `rqss()`.

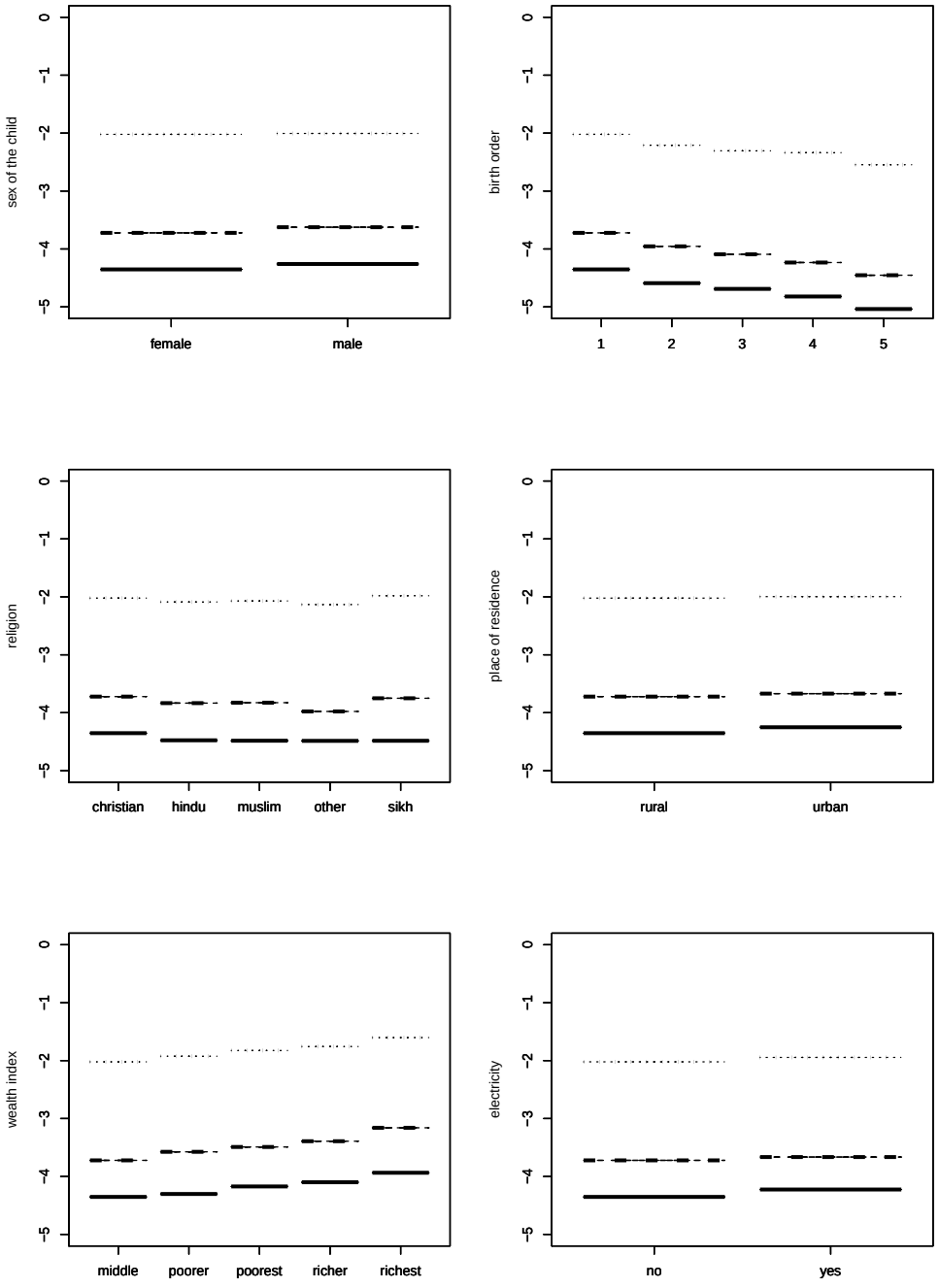


Figure 33: Selected estimated effects of categorical covariates for the 50% quantile (dotted line), the 10% quantile (dashed line) and the 5% quantile (solid line) of the stunting score obtained by `rqss()`.

References

Bühlmann, P. and Hothorn, T. (2007), "Boosting algorithms: Regularization, prediction and model fitting (with discussion)," *Statistical Science*, 22, 477–505.

Koenker, R. (2005), *Quantile Regression*, Economic Society Monographs, New York: Cambridge University Press.

Koenker, R. (2010), "Additive models for quantile regression: An analysis of risk factors for malnutrition in india," in *Advances in Social Science Research Using R*, ed. H. D. Vinod, New York: Springer-Verlag.