SIS | 2023
SIS Società Italiana di Statistica
Statistical LEArning, Sustainability and Impact EvaluatioN
SEAS IN
UNIVERSITÀ POLITECNICA DELLE MARCHE
June 21-23, 2023
Ancona (Italy)

Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB

Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM

ISBN 9788891935618AAVV

New advances in Regression Forests

Mila Andreani[a], Lea Petrella[b], and Nicola Salvati[c]

[a]Scuola Normale Superiore, Pisa, Italy; mila.andreani@sns.it
[b]MEMOTEF Depart., Sapienza University of Rome, Rome, Italy; lea.petrella@uniroma1.it
[c]Department of Economics and Management, University of Pisa, Pisa, Italy; nicola.salvati@unipi.it

Abstract

In this paper we propose a new Mixed-Effects Quantile Regression Forest by generalizing the Quantile Regression Forest approach to longitudinal data. The inferential procedure is based on the Nonparametric Maximum Likelihood exploiting the Asymmetric Laplace distribution tool. The performance of the ME-QRF is tested in a simulation study and compared with the results of standard quantile regression models. Finally, the ME-QRF is applied to a data set for analysing the effect of the tratment on lead-exposed children.

Keywords: Quantile Regression, Random Forests, mixed-effects, longitudinal data

1. Introduction

Mixed-effects quantile regression models are used in longitudinal studies to obtain a more complete picture of the response variable distribution with respect to standard linear regression while accounting for serial correlation among observations of the same statistical unit [5; 4; 2; 7]. This paper proposes a novel machine learning algorithm, denoted Mixed- Effects Quantile Regression Forest (ME-QRF) to estimate quantiles of longitudinal data generalizing the Quantile Regression Forest (QRF) algorithm of [9]. The inferential approach is based on the Asymmetric Laplace distribution tool by applying the Non Parametric Maximum Likelihood approach (NPML) of [6] already introduced in a quantile regression framework by [1; 10] to the Quantile Regression Forest contest. In particular, we develop an EM algorithm to estimate quantiles by decoupling the fixed-effects estimation part from the random-effects one without making any parametric assumption. The ME-QRF performance is tested by means of a simulation study and by comparing its performance with standard quantile regression models. The ME-QRF is also applied empirically using a dataset from the study of [13] conducted to assess whether the succimer treatment of children with Blood Lead Levels $(BLL) < 45\mu$g/dL is beneficial and safe.

2. Methodology

Let $y_{it}, i = 1, \ldots, N, t = 1, \ldots, T_i$ be the response variable for the i-th statistical unit observed at time t, and $\mathbf{x}_{it} \in \mathbb{R}^p$ be the vector of explanatory variables where $x_{it,1} \equiv 1$.

By indicating with $\tau \in (0, 1)$ the quantile probability level, the standard quantile regression linear mixed-model (LQMM) is:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_\tau + b_{i,\tau} + \varepsilon_{it} \quad \text{where} \quad Q_\tau(\varepsilon_{ij}|\mathbf{x}_{it}, \boldsymbol{\beta}_\tau, b_{i,\tau}) = 0, \quad \forall\tau \in (0, 1) \tag{1}$$

where $\boldsymbol{\beta}_\tau$ is a vector of τ- dependent regression coefficients common to all statistical units, the term $\mathbf{x}'_{it}\boldsymbol{\beta}_\tau$ is the "fixed-effects" part of the model, whereas the term $b_{i,\tau}$ is the "random-effects" part, represented by a time-constant parameter that varies across statistical units according to a distribution $f_b(\cdot)$ with support $\mathcal{B}$.

Here we avoid making a parametric assumption concerning the fixed-effects part as in (1) and formulate the quantile mixed model as follows:

$$y_{it} = g_\tau(\mathbf{x}_{it}) + b_{i,\tau} + \varepsilon_{it} \quad \text{where} \quad Q_\tau(\varepsilon_{it}|\mathbf{x}_{it}, b_{i,\tau}) = 0, \quad \forall \tau \in (0,1) \tag{2}$$

where $g_\tau : \mathbb{R}^p \to \mathbb{R}$ is a non-parametric unknown function. The terms $g_\tau(\mathbf{x}_{it})$ and $b_{i,\tau}$ are estimated via Maximum Likelihood by means of an EM algorithm based on QRF. We exploit the Asymmetric Laplace (AL) distribution as suitable tool [14], where $y_{it} \sim AL(\mu_{it}, \sigma_\tau, \tau)$:

$$f(y_{it}|\mu_{it,\tau}, \sigma_\tau, \tau) = \frac{\tau(1-\tau)}{\sigma_\tau} \exp\left\{ - \rho_\tau\left(\frac{y_{it} - \mu_{it,\tau}}{\sigma_\tau}\right)\right\}, \tag{3}$$

where $\sigma_\tau > 0$ is the scale parameter, the function $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u<0\}})$ is the quantile loss function of [5] and the location parameter $\mu_{it,\tau} = g_\tau(\mathbf{x}_{it}) + b_{i,\tau}$ represents the quantile at level τ.

The observed data likelihood is:

$$L(\boldsymbol{\Phi}_\tau) = \prod_{i=1}^{N} \left\{ \int_{\mathcal{B}} \prod_{t=1}^{T_i} f(y_{it}|\mu_{it,\tau}, \sigma_\tau, \tau) f_b(b_{i,\tau})\, \mathrm{d}b_{i,\tau} \right\} \tag{4}$$

where $\boldsymbol{\Phi}_\tau = \{\sigma, b_1, \ldots, b_N\}$. The main issue concerning the likelihood in (4) is that it involves a multidimensional integral that does not have a closed form solution and that it requires to specify the functional form of $f_b(\cdot)$. Thus, an EM algorithm is developed to estimate $g_\tau(\mathbf{x}_{it})$ with a QRF and to estimate $b_{i,\tau}$ by maximising (4) without making any parametric assumptions about the form of $f_b(\cdot)$.

In line with previous contributions [8; 11; 1], we approximate $f_b(\cdot)$ with a discrete distribution by exploiting the NPML approach of [6]. In particular, we consider a discrete distribution on $K < N$ locations $b_{k,\tau}$ such that $b_{i,\tau} \sim \sum_{k=1}^{K} \pi_{k,\tau}\delta_{b_{k,\tau}}$, where the probability $\pi_{k,\tau}$ is defined as $\pi_{k,\tau} = \mathbb{P}(b_{i,\tau} = b_{k,\tau})$ with $i = 1, \ldots, N$ and $k = 1, \ldots, K$ where $\delta_{b_{k,\tau}}$ is a one-point distribution putting a unit mass at $b_{k,\tau}$. The likelihood (4)is reformulated as:

$$L(\Phi_\tau) = \prod_{i=1}^{N} \left\{ \sum_{k=1}^{K} \prod_{t=1}^{T_i} f(y_{itk}|\mu_{itk,\tau}, \sigma_\tau, \tau)\pi_{k,\tau} \right\}, \tag{5}$$

where $\boldsymbol{\Phi}_\tau = \{\sigma, b_1, \ldots, b_K, \pi_1, \ldots, \pi_K\}$ is the parameter vector.

The next section described the EM algorithm based on (5) and QRF used to obtained $\hat{\Phi}_\tau$.

2.1 The EM algorithm

Given that each observation i in (5) can be considered as drawn from one of the K locations of the discrete distribution used to approximate $f_b(\cdot)$, we denote with w_{ik} the indicator variable equal to 1 if the i-th unit belongs to the k-th component of the finite mixture, and 0 otherwise. The component membership w_{ik} is considered as missing data and, from (5), the complete data log-likelihood is:

$$\ell_c(\boldsymbol{\Phi}_\tau) = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik,\tau}\left\{\sum_{t=1}^{T_i} \log(f(y_{it}|\mu_{itk,\tau}, \sigma_\tau)) + \log(\pi_{k,\tau})\right\} \tag{6}$$

Estimates $\widehat{g}_\tau(\mathbf{x}_{it})$ and $\widehat{b}_{i,\tau}$ in (2) are obtained from (6) in a EM algorithm by decoupling the fixed-effects estimation, obtained with a QRF, from the random-effects one as follows.

Initialization By indicating with r the generic iteration of the algorithm, in the first step $r = 0$, $\widehat{b}_{i,\tau}^{(0)}, \widehat{\sigma}_\tau^{(0)}, \widehat{\pi}_{k,\tau}^{(0)}, \widehat{g}_\tau(\mathbf{x}_{it})^{(0)}$ are initialised. In particular, the initial value $\widehat{g}_\tau(\mathbf{x}_{it})^{(0)}$ is computed as the τ-th quantile estimated with a QRF fitted with the training set $\mathcal{T}^{(0)} = \{(y_{it}, \mathbf{x}_{it})\}_{\substack{i=1,\ldots,N \\ t=1,\ldots,T_i}}$.

E-step The E-step consists in updating $\hat{w}_{ik,\tau}^{(r+1)}$ and $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$. In particular, $\hat{w}_{ik,\tau}^{(r+1)}$ is updated as:

$$\hat{w}_{ik,\tau}^{(r+1)} = \mathbb{E}[w_{ik,\tau}|y_{it}, \mathbf{x}_{it}, \hat{\Phi}_\tau^{(r)}] = \frac{\prod_{t=1}^{T_i} f_{itk,\tau}^{(r)} \hat{\pi}_{k,\tau}^{(r)}}{\sum_{l=1}^{K} \prod_{i=1}^{T_i} f_{itl,\tau}^{(r)} \hat{\pi}_{l,\tau}^{(r)}}, \tag{7}$$

where $f_{itk,\tau}^{(r)}$ is the response variable distribution when considering the k -th component of the finite mixture.

The estimate $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$ is updated by decoupling the random-effects from the fixed-effects. To this end, $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$ is estimated with the QRF fitted using the training set $\mathcal{T}^{(r+1)} = \left\{ \left(y_{it}^{*(r+1)}, \mathbf{x}_{it} \right) \right\}_{\substack{i=1,\dots,N \\ t=1,\dots,T_i}}$, in which $y_{it}^{*(r+1)} = y_{it} - \widehat{b}_{i,\tau}^{(r)}$.

M-step In the M-step, numerical optimisation techniques are applied to maximise $\mathbb{E}[\ell_c(\mathbf{\Phi}_\tau)|y_{it}, \mathbf{x}_{it}, \hat{\mathbf{\Phi}}_\tau^{(r)}]$ with respect to $\hat{\sigma}_\tau$ and $\widehat{b}_{k,\tau}$.

The E- and M-steps are alternated iteratively until convergence.

3. Simulation study

This section reports the results of a simulation study carried out to assess the performance of the ME-QRF in a non-linear setting. To this end, the ME-QRF is used to predict quantiles at levels $\tau \in \{0.1, 0.5, 0.9\}$ of an outcome variable simulated under the following non-linear data generating process (DGP) [3]:

$$y_{it} = g(\mathbf{x}_{it}) + b_i + \varepsilon_{it} \quad \text{where} \quad g(\mathbf{x}_{it}) = 2x_{it,1} + x_{it,1}^2 + 4 \cdot \mathbf{1}_{\{x_{it,3}>0\}} + 2x_{it,3} \log|x_{it,1}|$$

The covariates are generated as $x_{it,1},\ x_{it,2},\ x_{it,3} \sim \mathcal{N}(0,1)$. The random-effects parameters and the error terms are generated independently according to two DGPs:

$$\textbf{(NN)}\ b_i \sim N(0,1),\ \ \varepsilon_{it} \sim N(0,1) \qquad\qquad \textbf{(TT)}\ b_i \sim t(3),\ \ \varepsilon_{it} \sim t(3)$$

As in [3], for each scenario we consider a training set of 500 observation for $N = 100$ statistical units and $T_i = 5$ measurements each, and an unbalanced test set with $T_i \in \{9, 27, 45, 63, 81\}$ for a total of 4500 observations. Each scenario has been replicated $S = 100$ times.

The average performance of the ME-QRF across the 100 replications is assessed in terms of Average Mean Absolute Error (MAE) and average Mean Squared Error (MSE) with respect to the theoretical quantile of the DGP, computed as in [12]:

$$MAE_\tau = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}|Q_{it,\tau}^s - \hat{Q}_{it,\tau}^s| \qquad MSE_\tau = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}(Q_{it,\tau}^s - \hat{Q}_{it,\tau}^s)^2$$

where $Q_{it,\tau}^s = Q_\tau^s(y_{it}|\mathbf{x}_{it})$ and $\hat{Q}_{it,\tau}^s = \hat{Q}_\tau^s(y_{it}|\mathbf{x}_{it})$ are respectively the theoretical and estimated conditional quantiles of variable y_{it} at level τ of the s-th simulated dataset.

The ME-QRF is compared with three benchmark models: LQMM, Quantile Random Forest (QRF) and the Quantile Mixed Model (QMM) of [10]. The latter model exploits the same methodological approach of the ME-QRF in a linear setting. Results are reported in Table 1.

		$\tau = 0.1$				$\tau = 0.5$				$\tau = 0.9$			
		ME-QRF	**LQMM**	**RF**	**QMM**	**ME-QRF**	**LQMM**	**RF**	**QMM**	**ME-QRF**	**LQMM**	**RF**	**QMM**
NN	**MAE**	**1.83**	1.86	2.64	4.67	**1.65**	1.72	2.11	2.80	1.67	**1.57**	2.20	5.16
	MSE	**5.87**	5.89	10.98	40.62	**4.62**	5.27	7.15	14.61	4.86	**4.34**	7.93	61.12
TT	**MAE**	**1.80**	1.94	2.02	4.11	**1.43**	1.57	1.44	2.01	**1.87**	2.06	2.06	5.13
	MSE	6.66	**6.42**	7.88	33.16	**4.32**	4.48	4.51	7.82	7.10	**6.81**	8.02	50.79

Table 1: Loss values for each scenario computed on the test set of the four fitted models. Values in bold indicate the smallest loss.

The results highlight that the ME-QRF outperforms the benchmark models at almost all quantile levels in each scenario, especially when the data violate the Gaussianity assumptions. The only exception is represented by the quantile at $\tau = 0.9$ for the NN scenario. In this case, the LQMM outperforms the ME-QRF since the Gaussianity assumptions of the LQMM model hold. The ME-QRF and the LQMM perform similarly in terms of MSE and represent the two models with the lowest MAE and MSE values.

4. Empirical Application

In this section, the ME-QRF is applied to a dataset from a placebo-controlled, double-blind, randomised trial to study whether the succimer treatment of children with Blood Lead Levels (BLL) $<$ 45μg/dL is beneficial and safe. Following [1], three quantile levels are considered $\tau \in \{0.25, 0.5, 0.75\}$.

The dataset includes $T_i = 4$ weekly measurements of BLL for $N = 100$ children with BLL of 20–44 μg/dL. The covariates are the dummy variable Treatment (R_i) taking value 1 for children that have been treated and 0 otherwise, and Time $W_{it} \in \{0, 1, 4, 6\}$ representing week 0 -baseline-, week 1, week 4 and week 6. Given the results of the simulation study, the ME-QRF has been compared in terms of quantile loss with the LQMM with the following formulation:

$$BLL_{it} = \beta_{i1} R_i + \beta_{i2} W_{it}^2 + \beta_{i3}(R_i * W_{it}^2) + \beta_{i4}(R_i * W_{it}) + b_i \tag{8}$$

The quantile loss of [5] is computed as follows and the related results are presented in Table 2:

$$QLOSS_\tau = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{t=1}^{T_i} u_{it}(\tau - \mathbf{1}_{\{u_{it}<0\}}) \qquad \text{where} \qquad u_{it} = y_{it} - \hat{Q}_{it,\tau}. \tag{9}$$

τ	**0.25**	**0.5**	**0.75**
QLOSS$^{\text{ME-QRF}}$	**1.22**	**1.55**	**1.31**
QLOSS$^{\text{LQMM}}$	1.61	1.82	1.59

Table 2: Results in terms of QLOSS for the treatment of lead-exposed children dataset.

Results show that our model outperforms the LQMM at each quantile level. Figure 1 depicts the treatment and control group quantile trajectories estimated with the ME-QRF for each mixture component at level $\tau \in (0.25, 0.5, 0.75)$. Each trajectory is colour-coded according to the mixture component and the treatment and control group of each component are identified with the solid and dashed lines, respectively. The legend reports the number of statistical units in the control and treatment group of each mixture component. The trajectories estimated with our model are coherent with the findings of [1].

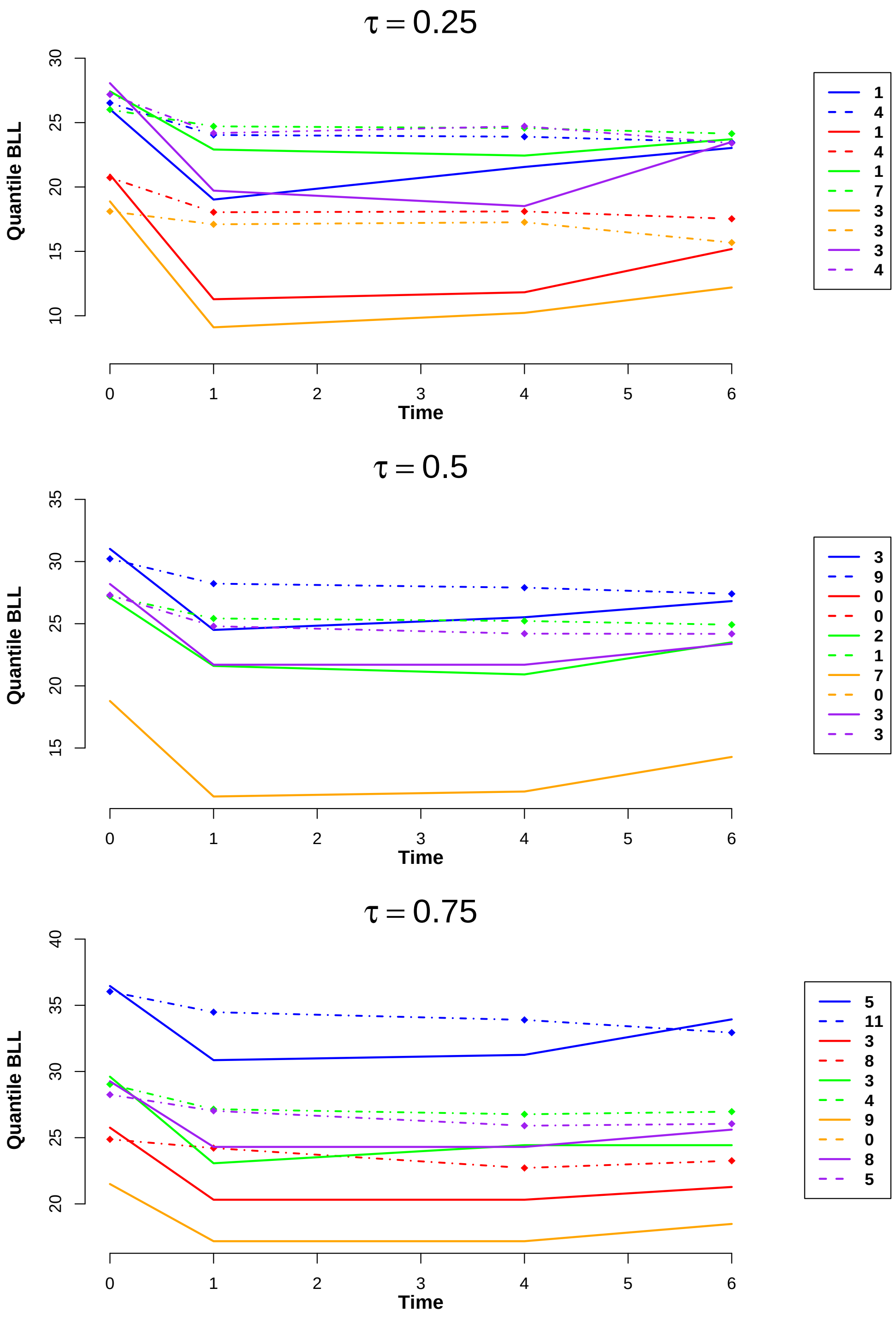

Figure 1: Estimated trajectories for ME-QRF for the first five mixture components. Each component is indicated with one colour, and the treatment (solid curves) and control (dashed lines) groups are represented separately. The legend reports the number of statistical units belonging to each mixture component.

5. Conclusions

This paper introduces the Mixed-Effects Quantile Regression Forest (ME-QRF) model, which combines QRF and mixed-models to estimate quantiles of longitudinal data wotjoud any parametric assumption on the fixed-effects and the random-effects distribution. Simulation results highlight that the ME-QRF outperforms benchmark models in non-linear settings, especially when Gaussianity assumptions are violated. The ME-QRF is applied empirically using data from the study of [13] in order to assess the effectiveness of the succimer treatment on lead-esposed children. Results show that the ME-QRF outperforms the LQMM model in terms of quantile loss and that findings are coherent with the ones presented in the previous literature.

References

[1] Alfò, M., Salvati, N., and Ranallli, M. G. (2017). Finite mixtures of quantile and m-quantile regression models. *Statistics and Computing*, 27(2):547–570.

[2] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

[3] Hajjem, A., Bellavance, F., and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328.

[4] Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

[5] Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

[6] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.

[7] Liu, Y. and Bottai, M. (2009). Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, 5(1).

[8] Marino, M. F., Tzavidis, N., and Alfò, M. (2018). Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical methods in medical research*, 27(7):2231–2246.

[9] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

[10] Merlo, L., Maruotti, A., and Petrella, L. (2021). Two-part quantile regression models for semi-continuous longitudinal data: A finite mixture approach. *Statistical Modelling*, page 1471082X21993603.

[11] Merlo, L., Petrella, L., and Tzavidis, N. (2022). Quantile mixed hidden markov models for multivariate longitudinal data: An application to children's strengths and difficulties questionnaire scores. *Journal of the Royal Statistical Society, Series C (Applied Statistics).*

[12] Min, I. and Kim, I. (2004). A monte carlo comparison of parametric and nonparametric quantile regressions. *Applied Economics Letters*, 11(2):71–74.

[13] Rogan, W., Bornschein, R., Chisolm, J., Damokosh, A., Dockery, D., Fay, M., Jones, R., Rhoads, G., Ragan, N., Salganik, M., et al. (2000). Safety and efficacy of succimer in toddlers with blood lead levels of 20-44 μg/dl. *Pediatric Research*, 48(5):593–599.

[14] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.