Journal of Machine Learning Research 27 (2026) 1-24 Submitted 1/25; Revised 4/26; Published 5/26

Semi-supervised learning for linear extremile regression

Rong Jiang JIANGRONG@SUIBE.EDU.CN
School of Statistics and Data Science, Shanghai University of International Business and Economics, China

Jiangfeng Wang* WJF2929@163.COM
School of Statistics and Data Science, Zhejiang Gongshang University, China
Laboratory for Statistical Monitoring and Intelligent Governance of Common Prosperity, China

Keming Yu* KEMING.YU@BRUNEL.AC.UK
Department of Mathematical Sciences, Brunel University of London, UK

: Corresponding author

Editor: Arindam Banerjee

Abstract

Extremile regression, as a least squares analog of quantile regression, is potentially a useful tool for modeling and understanding the extreme tails of a distribution. However, existing extremile regression methods, as nonparametric approaches, may face challenges in high-dimensional settings due to data sparsity, computational inefficiency, and the risk of overfitting. While linear regression, particularly in high-dimensional settings, serves as the foundation for many other statistical and machine learning models due to its simplicity, interpretability, and relatively easy implementation, this paper introduces a novel definition of linear extremile regression along with an accompanying estimation methodology. The regression coefficient estimators of this method achieve root n consistency, which nonparametric extremile regression may not provide. In particular, while semi-supervised learning can leverage unlabeled data to make more accurate predictions and avoid overfitting to small labeled datasets in high-dimensional spaces, we propose a semi-supervised learning to enhance estimation efficiency, even when the specified linear extremile regression model may be misspecified. Both simulation studies and real data analyses demonstrate the finite sample performance of our proposed methods.

Keywords: Semi-supervised learning, extremile regression, quantile regression

1. Introduction

Assessing the extreme behavior of random phenomena is a significant challenge across various fields, including finance, extreme weather and climate events, as well as medicine (Leblanc et al., 2006; Chen et al., 2024). A common approach to analyzing extreme events involves estimating the extreme quantile (Koenker and Bassett, 1978) of a relevant random variable, such as the daily return of a stock market index or the intensity of an earthquake. Meanwhile, expectiles (Newey and Powell, 1987) can also serve this purpose. Although both quantiles and expectiles have proven to be valuable tools, they have faced criticism in the literature for various axiomatic or practical reasons. Quantiles rely solely on whether an

observation is below or above a specific threshold, while expectiles may suffer from a lack of transparent interpretation owing to the absence of an explicit expression.

From quantiles and expectiles, Daouia et al. (2019) proposed a new least squares analogue of quantiles: extremile. They found that the τ-th quantile can be derived from the alternative minimization problem: $q_\tau = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\mathrm{J}_\tau\{\mathrm{F}(\boldsymbol{Y})\}\cdot|\boldsymbol{Y}-\theta|\right]$, where

$$\mathrm{J}_\tau(t) = \begin{cases} s(\tau)(1-t)^{s(\tau)-1}, & \text{if } 0<\tau\le 1/2, \\ r(\tau)t^{r(\tau)-1}, & \text{if } 1/2\le\tau<1, \end{cases}$$

$r(\tau) = s(1-\tau) = \log(1/2)/\log(\tau)$ and $\mathrm{F}(\cdot)$ is the distribution of $\boldsymbol{Y}$. Therefore, Daouia et al. (2019) changed the absolute value term in the above quantile q_τ to a squared term, following the idea proposed by expectile, and proposed extremile as the following formula:

$$\xi_\tau = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\mathrm{J}_\tau\{\mathrm{F}(\boldsymbol{Y})\}\cdot(\boldsymbol{Y}-\theta)^2\right].$$

Like expectiles, extremiles take into account both the distance from the selected location and the frequency, but they employ a distinct weighting scheme. They occupy a middle ground between the robustness of ordinary quantile estimators and the outlier sensitivity of extreme quantile estimators. Notably, extremiles exhibit greater tail sensitivity than quantile estimators at moderate or central τ levels, while becoming more resistant at extremely high or low τ levels in the far tails, thereby avoiding the issue of crossing regression lines. The study in Furno (2023) provides a comprehensive comparative analysis of extremiles, quantiles, and expectiles specifically within tail regions.

Extremiles can be of considerable importance for modeling extremes of natural phenomena, which are weighted expectations rather than tail probabilities. Of special interest is their intuitive meaning in terms of expected minima and maxima as following:

$$\xi_\tau = \begin{cases} \mathbb{E}\left\{\max(\boldsymbol{Y}^1,\ldots,\boldsymbol{Y}^r)\right\}, & \text{when } \tau = 0.5^{1/r} \text{ with } r\in\mathbb{N}\backslash\{0\} \\ \mathbb{E}\left\{\min(\boldsymbol{Y}^1,\ldots,\boldsymbol{Y}^s)\right\}, & \text{when } \tau = 1-0.5^{1/s} \text{ with } s\in\mathbb{N}\backslash\{0\}, \end{cases}$$

where $\{\boldsymbol{Y}^i\}$ are independent observations and drawn from the distribution of $\boldsymbol{Y}$. Therefore, extremiles can be used in risk management, in contrast to quantiles, extremiles fulfill the coherency axiom and take the severity of tail losses into account. In addition, extremiles are comonotonically additive and belong to both the families of spectral risk measures and concave distortion risk measures. Drawing on this distinctive characteristic, Jiang et al. (2025) employed extremiles in the construction of a stock investment portfolio, subsequently ascertaining the weight assigned to each stock within the portfolio. Furthermore, Dieter et al. (2025) champions the use of extremiles and introduces a more broadly applicable tool for measuring risk.

The original work on extremile regression (Daouia et al., 2022) has recently introduced the concept of the conditional order-τ extremile of $\boldsymbol{Y}$ given $\boldsymbol{X}=\boldsymbol{x}$ as

$$\xi_\tau(\boldsymbol{x}) = \arg\min_{\theta\in\mathbb{R}} \mathrm{E}\left[\mathrm{J}_\tau\{\mathrm{F}(\boldsymbol{Y}|\boldsymbol{X})\}\cdot(\boldsymbol{Y}-\theta)^2|\boldsymbol{X}=\boldsymbol{x}\right], \tag{1.1}$$

where $\mathrm{F}(\cdot|\boldsymbol{X})$ is the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Subsequently, the study Chen et al. (2023) delves into the statistical inference of extremiles within the realm of heavy tailed heteroscedastic regression. The research Geng (2024) puts forward an additive extremile

regression approach suitable for high dimensional situations. Additionally, the paper Sun and Wang (2024) presents a robust linear extremile regression model that employs the Huber loss function.

Linear regression occupies a prominent position among regression techniques in both statistics and machine learning, largely owing to its remarkable simplicity, high interpretability, and straightforward implementation, even when dealing with high-dimensional datasets. Given these advantages, we hereby introduce the following linear extremile regression: $\xi_\tau(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\beta}_\tau$, then, from the equation (1.1) and Sun and Wang (2024), we can obtain the estimator of $\boldsymbol{\beta}_\tau$ as:

$$\bar{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}} \sum_{i=1}^{n} \mathrm{J}_\tau\{\hat{\mathrm{F}}(Y_i|\boldsymbol{X}_i)\} \cdot (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2. \tag{1.2}$$

The estimator $\bar{\boldsymbol{\beta}}_\tau$ encounters considerable hurdles in achieving $\sqrt{n}$-consistency. For example, when using the Nadaraya-Watson method to calculate $\hat{\mathrm{F}}(Y_i|\boldsymbol{X}_i)$, the convergence rate of $\bar{\boldsymbol{\beta}}_\tau$ incorporating $\hat{\mathrm{F}}(Y_i|\boldsymbol{X}_i)$ will be slower than $\sqrt{n}$ because $\hat{\mathrm{F}}(Y_i|\boldsymbol{X}_i)$ is $\sqrt{nh}$-consistent with $h \to 0$. This poses a challenge since parameter estimators are generally expected to be $\sqrt{n}$-consistent. While alternative parameter estimation methods for estimating $\mathrm{F}(\boldsymbol{Y}|\boldsymbol{X})$ could be explored, they often demand strict conditions. To overcome this, we propose a new estimation method that avoids including the unknown non-parametric component, as shown in equation (1.2), enabling us to construct a $\sqrt{n}$-consistent estimator for the unknown parameters in the linear extremile regression model.

In real world applications, linear models struggle to match the complex non-linearity of actual data due to their linear assumptions, resulting in low fitting accuracy, limited performance, and difficulty in meeting application requirements. Semi-supervised learning (SSL), as an emerging and promising machine learning method, can fully mine and utilize the information contained in a large amount of unlabeled data. When the working model has biases, it can construct effective estimation methods to mitigate the impact and ensure accurate results. Given its advantages in handling complex data and enhancing performance, this study proposes a method to construct semi-supervised learning for linear extremile regression by leveraging unlabeled data.

The SSL setting encompasses two distinct datasets: (i) a labeled data set comprising observations for an outcome $\boldsymbol{Y}$ and a set of covariates $\boldsymbol{X}$, and (ii) a significantly larger unlabeled data set where only covariates $\boldsymbol{X}$ are observed. This fundamental distinction sets SSL settings apart from standard missing data problems, where the proportion is always assumed to be bounded away from 0, a condition commonly referred to as the positivity (or overlap) assumption in the missing data literature, which is inherently violated in this context.

For instance, in biomedical applications, SSL settings are gaining increasing prominence in contemporary integrative genomics, particularly in the investigation of expression quantitative trait loci (eQTL) (Michaelson et al., 2009). This methodology integrates genetic association studies with gene expression profiles. Nevertheless, a significant challenge in such studies arises from the fact that the limited availability of costly gene expression data often restricts their analytical capabilities (Flutre et al., 2013). In contrast, the recording of genetic variations is more cost effective and can be readily applied to large scale datasets,

thereby naturally giving rise to SSL settings. Moreover, SSL settings are becoming increasingly pertinent across various fields, including image processing (Cheplygina et al., 2019), anomaly detection (Wang et al., 2019), and empirical risk analysis (Yuval and Rosset, 2022). A thorough review of SSL and its recent advancements can be found in Chapelle et al. (2010) and Cannings (2021). For related statistical theory research literature, refer to (Chakrabortty and Cai, 2018; Zhang et al., 2019; Cai and Guo, 2020; Song et al., 2024a,b; Wen et al., 2025; Hou et al., 2025).

To summarize, our key statistical contributions are outlined below:

(i) Unlike the nonparametric framework employed in extremile regression (Daouia et al., 2022), we propose a linear extremile regression model due to its simplicity, interpretability, and ease of implementation, especially in high-dimensional contexts. For the estimation approach, we focus on estimating the quantile function (the inverse of the conditional distribution function $\mathrm{F}(\cdot|\boldsymbol{X})$) rather than directly estimating $\mathrm{F}(\cdot|\boldsymbol{X})$. Furthermore, we adopt a parametric method to estimate the quantile regression coefficients, enabling us to achieve $\sqrt{n}$-consistency for the unknown parameters in the linear extremile regression model.

(ii) In scenarios involving model misspecification, we have devised an estimation method for the unknown parameters in the linear extremile regression model by leveraging unlabeled data. We show that the resulting estimator outperforms one based solely on labeled data. To our knowledge, this represents the first application of semi-supervised learning techniques to extremile regression.

The remainder of this paper is structured as follows: Section 2 introduces the new definition of linear extremile regression and its corresponding estimation method. Section 3 delves into the development of semi-supervised learning strategies. Section 4 presents simulation studies and applies real-world data to validate the proposed methods. Finally, Section 5 concludes the paper with a concise discussion, while all technical proofs are provided in the Appendix.

2. The definition of a new linear extremile regression and its estimation method

2.1 The definition of linear extremile regression

We introduce a novel linear extremile regression and construct a $\sqrt{n}$-consistent estimator for the parameter vector $\boldsymbol{\beta}_\tau$ within this linear extremile regression. Note that an alternative perspective on the τth extremile of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is the weighted quantile function (Proposition 1 in Daouia et al. (2022)):

$$\xi_\tau(\boldsymbol{X}) = \int_0^1 q_{\bar{\tau}}(\boldsymbol{X})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}, \tag{2.1}$$

where $q_{\bar{\tau}}(\boldsymbol{X})$ is the conditional $\bar{\tau}$-th quantile of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Due to the linear assumption $\xi_\tau(\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{\beta}_\tau$, only $q_{\bar{\tau}}(\boldsymbol{X})$ is related to $\boldsymbol{X}$ in the combination (2.1), thus it follows that

$$q_{\bar{\tau}}(\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{\gamma}(\bar{\tau}), \tag{2.2}$$

where $\boldsymbol{\gamma}(\cdot)$ is a p dimensional unknown function. Furthermore, we parameterize $\boldsymbol{\gamma}(\bar{\tau})$ as

$$\boldsymbol{\gamma}(\bar{\tau}) = \boldsymbol{\alpha}_0\boldsymbol{b}(\bar{\tau}), \tag{2.3}$$

where $\boldsymbol{b}(\cdot)$ is the q dimensional basis function and $\boldsymbol{\alpha}_0$ is a $p \times q$ unknown matrix.

As mentioned in Frumento et al. (2021), the choice of $\boldsymbol{b}(\cdot)$ is not as crucial as it appears. We recommend choosing $\boldsymbol{b}(\tau) = (\tau, 3/2\tau^2 - 1/2, 5/2\tau^3 - 3/2\tau)$, as it is a 3rd-degree shifted Legendre polynomial and is the default setting of the `iqr` function in the `qrcm R` package. Other valid choices of $\boldsymbol{b}(\tau)$ are, for example, $\boldsymbol{b}(\tau) = (1, \tau, \tau^2, \tau^3)$ simply consists of polynomials of increasing orders, $\boldsymbol{b}(\tau) = (1, \log(\tau), -\log(1-\bar{\tau}))$ relates to an asymmetric logistic distribution and $\boldsymbol{b}(\tau) = (1, \Phi^{-1}(\tau), \sqrt{-2\log(1-\tau)})$ is a combination of quantile functions of standard normal Φ and Rayleigh distributions.

Then, from (2.1)–(2.3), we can re-define the linear extremile regression as

$$\xi_\tau(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\alpha}_0 \int_0^1 \boldsymbol{b}(\bar{\tau}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau} \equiv \boldsymbol{X}^\top \boldsymbol{\beta}_\tau, \tag{2.4}$$

where

$$\boldsymbol{\beta}_\tau = \boldsymbol{\alpha}_0 \int_0^1 \boldsymbol{b}(\bar{\tau}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau}. \tag{2.5}$$

By comparing (1.2) and (2.5), we can see that only the estimation of unknown parameter $\boldsymbol{\alpha}_0$ is present in (2.5), which avoids a typically nonparametric estimation of $\mathrm{F}(\boldsymbol{Y}|\boldsymbol{X})$ in (1.2). The following Proposition 2.1 confirms that the τ-th extremile $\xi_\tau(\boldsymbol{X})$ defined in (2.4) is the same as that in Daouia et al. (2022).

Proposition 2.1. Let $\boldsymbol{Y}$ given $\boldsymbol{X}$ have a finite absolute first moment. Then, for any $\tau \in (0,1)$, we have the following equivalent form expression:

$$\xi_\tau(\boldsymbol{X}) = \begin{cases} \mathbb{E}\left\{\max(\boldsymbol{Y}^1_{\boldsymbol{X}}, \ldots, \boldsymbol{Y}^r_{\boldsymbol{X}})\right\}, & \text{when } \tau = 0.5^{1/r} \text{ with } r \in \mathbb{N}\backslash\{0\} \\ \mathbb{E}\left\{\min(\boldsymbol{Y}^1_{\boldsymbol{X}}, \ldots, \boldsymbol{Y}^s_{\boldsymbol{X}})\right\}, & \text{when } \tau = 1 - 0.5^{1/s} \text{ with } s \in \mathbb{N}\backslash\{0\}, \end{cases}$$

where $\{\boldsymbol{Y}^i_{\boldsymbol{X}}\}$ are independent observations and drawn from the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Specifically, when $\boldsymbol{X} = \boldsymbol{1}$, the linear extremile regression (2.4) is equal to extremile in Daouia et al. (2019) according to $\xi_\tau(\boldsymbol{X}) = \boldsymbol{\beta}_\tau = \int_0^1 \boldsymbol{\gamma}(\bar{\tau}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau} = \int_0^1 q_{\bar{\tau}} \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau} = \xi_\tau$. Therefore, the proposed definition of linear extremile regression (2.4) is reasonable.

2.2 Estimation method

In this section, we estimate the unknown parameter $\boldsymbol{\beta}_\tau$ in (2.5) under dataset $\mathcal{L} = \{Y_i, \mathbf{X}_i\}_{i=1}^n$, which are n independent and identically distributed (i.i.d) observations from $\{\boldsymbol{Y}, \boldsymbol{X}^\top\}^\top$ in model (2.4).

Based on (2.2) and (2.3), we have $q_{\bar{\tau}}(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\gamma}(\bar{\tau}) = \boldsymbol{X}^\top \boldsymbol{\alpha}_0 \boldsymbol{b}(\bar{\tau})$. Therefore, we can estimate $\boldsymbol{\alpha}_0$ as the minimizer of the integrated objective function:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^n L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^n \int_0^1 \rho_{\bar{\tau}}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\alpha} \boldsymbol{b}(\bar{\tau})) d\bar{\tau}, \tag{2.6}$$

where $L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) = \int_0^1 \rho_{\bar{\tau}}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\alpha} \boldsymbol{b}(\bar{\tau})) d\bar{\tau}$ and $\rho_{\bar{\tau}}(r) = \bar{\tau} r - r\mathrm{I}(r < 0)$ is the quantile check function. The objective function $L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})$ in (2.6) can be regarded as an average

loss function, achieved by marginalizing $\rho_{\bar{\tau}}(\boldsymbol{Y}_i - \boldsymbol{X}_i^\top \boldsymbol{\alpha}\boldsymbol{b}(\bar{\tau}))$ over the entire interval $(0,1)$. In addition, the solution of minimizing (2.6) is currently implemented by the `iqr` function in the `qrcm R` package. Then, the estimator $\hat{\boldsymbol{\beta}}_\tau$ of $\boldsymbol{\beta}_\tau$ (2.5) with $\hat{\boldsymbol{\alpha}}$ in (2.6) is

$$\hat{\boldsymbol{\beta}}_\tau = \hat{\boldsymbol{\alpha}} \int_0^1 \boldsymbol{b}(\bar{\tau}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau}. \tag{2.7}$$

Note that equation (2.7) enables the estimation of the entire extremile process, rather than merely yielding a discrete set of extremiles. This is because $\hat{\boldsymbol{\alpha}}$ is independent of $\bar{\tau}$. This particular property allows the method proposed in this paper to demonstrate its advantages when dealing with large-scale data, significantly reducing the computational time required for multi-quantile estimation problems. In addition, if $\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ in (2.7) is not integrable, we can use $n^{-1}\sum_{i=1}^n \boldsymbol{b}(i/n)\mathrm{J}_\tau(i/n)$ to calculate it approximately.

2.3 Large sample properties

To facilitate the presentation of the derivation of the asymptotic theories, let us introduce some notations. Let $\mathrm{Vec}(\cdot)$ be the vectoring operation, which creates a column vector by stacking the column vectors of below one another, that is, $\mathrm{Vec}(\boldsymbol{\alpha}) = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_q^\top)^\top$ with $\boldsymbol{\alpha}_j = (\alpha_{1,j}, \ldots, \alpha_{p,j})^\top$. Denote $S(\boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^n \nabla_{\mathrm{Vec}(\boldsymbol{\alpha})} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^n \int_0^1 \boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i[\mathrm{I}(Y_i < \{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i\}^\top \mathrm{Vec}(\boldsymbol{\alpha})) - \bar{\tau}]d\bar{\tau}$, where $\otimes$ is the Kronecker product.

C1: The true unknown parameter vector $\boldsymbol{\alpha}_0$ in (2.4) is an interior point of a compact set Θ and satisfies $\mathbb{E}\{S(\boldsymbol{\alpha}_0)|\boldsymbol{X}\} = \boldsymbol{0}$.

C2: The loss function $L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})$ satisfies $\mathbb{E}[\sup_{\boldsymbol{\alpha}\in\Theta}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}^2] < \infty$. $S(\boldsymbol{\alpha})$ is continuously differentiable, $\mathbb{E}\{\sup_{\boldsymbol{\alpha}\in\Theta} \|S(\boldsymbol{\alpha})\|_2^2\} < \infty$ and $\mathbb{E}\{\sup_{\boldsymbol{\alpha}\in\Theta} \|\nabla_{\mathrm{Vec}(\boldsymbol{\alpha})} S(\boldsymbol{\alpha})\|\} < \infty$ with $\|\cdot\|$ is the spectral norm.

C3: $\mathbf{H} = \mathbb{E}\{\nabla_{\mathrm{Vec}(\boldsymbol{\alpha})} S(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}\} = \mathbb{E}[\int_0^1 \{\boldsymbol{X}^\top \boldsymbol{\alpha}_0 \nabla_{\bar{\tau}} \boldsymbol{b}(\bar{\tau})\}^{-1} \{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}\}\{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}\}^\top d\bar{\tau}]$ is nonsingular.

Remark 2.1. The validity of the conditions **C1** and **C2** depends on the structure of $\boldsymbol{b}(\bar{\tau})$, which should induce a well-defined quantile function such that Θ is not empty (conditions **C1**); $\boldsymbol{b}(\bar{\tau})$ is continuous and ensure that a central limit theorem can be applied to $S(\boldsymbol{\alpha})$ (conditions **C2**). Conditions **C1** and **C2** are also used in Frumento and Bottai (2016). Condition **C3** is needed to establish the asymptotic normality.

Theorem 2.1. Suppose that the conditions **C1** and **C2** hold. Then as $n \to \infty$, we have

$$\mathrm{Vec}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{\mathbb{P}} \boldsymbol{0},$$

where $\xrightarrow{\mathbb{P}}$ represents the convergence in the probability. Moreover, if condition **C3** holds, we can obtain

$$\sqrt{n}\mathrm{Vec}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{\mathbb{L}} \mathbb{N}\left(\boldsymbol{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}\right),$$

where $\xrightarrow{\mathbb{L}}$ represents the convergence in the distribution and $\boldsymbol{\Sigma} = \mathbb{E}\{S(\boldsymbol{\alpha}_0)S(\boldsymbol{\alpha}_0)^\top\}$.

Theorem 2.1 shows that the parameter $\boldsymbol{\alpha}_0$ is identified and its estimator $\hat{\boldsymbol{\alpha}}$ has a normal distribution in large samples. The large sample distribution of the plug-in estimator of $\boldsymbol{\beta}_\tau = \boldsymbol{\alpha}_0 \int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ can also be obtained as

$$\hat{\boldsymbol{\beta}}_\tau = \hat{\boldsymbol{\alpha}} \int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau} = \tilde{\boldsymbol{b}}(\tau)^\top \mathrm{Vec}(\hat{\boldsymbol{\alpha}}), \tag{2.8}$$

where $\tilde{\boldsymbol{b}}(\tau) = \int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau} \otimes \boldsymbol{I}_p$ and $\boldsymbol{I}_p$ is the identity matrix of size p. Then, we consider the large sample distribution of $\hat{\boldsymbol{\beta}}_\tau$ by (2.8) in the following theorem.

Theorem 2.2. Suppose that the conditions **C1-C3** hold and $\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ is finite, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \xrightarrow{\mathbb{L}} \mathbb{N}\left(\mathbf{0}, \tilde{\boldsymbol{b}}(\tau)^\top \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}\tilde{\boldsymbol{b}}(\tau)\right).$$

3. Semi-supervised learning

3.1 Data representation and target parameter

Let $\mathbb{F}$ denote the joint distribution of $\{\boldsymbol{Y}, \boldsymbol{X}^\top\}^\top$, and let $\mathbb{F}_X$ represent the marginal distribution of $\boldsymbol{X}$. In semi-supervised setting, the data available are $\mathcal{D} = \mathcal{B} \cup \mathcal{M}$, where $\mathcal{B} = \{Y_i, \mathbf{X}_i\}_{i=1}^n$ is from $\mathbb{F}$ and $\mathcal{M} = \{\mathbf{X}_i\}_{i=n+1}^N$ with $N \geq 1$ are N i.i.d observations from $\mathbb{F}_X$. The $n/N \to \rho$ for some constant $\rho \in [0, +\infty)$ as $n \to \infty$ and $N \to \infty$. Note that the semi-supervised setting allows $n/N \to 0$, that means that the unlabeled dataset can be of much larger size than the labeled one in various practical problems, as labeling of the outcomes is often very costly. However, the missing completely at random assumption is that $\rho > 0$, which is the major difference between semi-supervised setting and missing data.

In most real-world data analyses, the model (2.4) may be misspecified due to its overly strong linear structure assumptions. But due to the simplicity and interpretability of the linear structure, it is often continued to be used (Box, 1976). Therefore, consider a τ-th linear extremile working regression model $\xi_\tau(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\alpha}^* \int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$, where the unknown parameter $\boldsymbol{\alpha}^*$ is defined as

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}. \tag{3.1}$$

It is noteworthy that in supervised framework (Section 2), $\boldsymbol{\alpha}^*$ is equal to $\boldsymbol{\alpha}_0$ in (2.4) when the outcome variable $\boldsymbol{Y}$ is fully observed and the working model is correctly specified.

3.2 Semi-supervised learning

To incorporate the unlabeled data $\{\mathbf{X}_i\}_{i=n+1}^N$ into the loss function $\mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}$ defined in equation (3.1), the key idea is to construct a consistent estimator of $\mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}$ utilizing the unlabeled data. To achieve this, we introduce a function of $\boldsymbol{X}$, denoted as $\boldsymbol{Z}$, which is a d-dimensional vector and provides flexibility to adapt to various data characteristics. Additionally, we define a parameter $\boldsymbol{\varphi}$ as a function of $\boldsymbol{\alpha}$, denoted $\boldsymbol{\varphi}(\boldsymbol{\alpha})$, which satisfies the following key equality:

$$\mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\} = \mathbb{E}\{\boldsymbol{Z}^\top \varphi(\boldsymbol{\alpha})\}. \tag{3.2}$$

Here, $\varphi(\boldsymbol{\alpha})$ is determined by minimizing the expected squared difference between $L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})$ and $\boldsymbol{Z}^\top\varphi$, that is,

$$\varphi(\boldsymbol{\alpha}) = \arg\min_{\varphi} \mathbb{E}\left\{\left(L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha}) - \boldsymbol{Z}^\top\varphi\right)^2\right\} = \left\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\right\}^{-1}\mathbb{E}\left\{\boldsymbol{Z}L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\right\}. \quad (3.3)$$

Different selections of $\boldsymbol{Z}$ yield distinct strategies for incorporating unlabeled data. A widely-adopted approach employs: $\boldsymbol{Z} = (1, \boldsymbol{X}^\top, \ldots, (\boldsymbol{X}^\theta)^\top)^\top$. To ascertain the polynomial order θ, we draw inspiration from the generalized Bayesian information criterion with prior probability (GBIC_{ppo}) proposed by Lv and Liu (2014). This criterion involves minimizing the following objective function with respect to θ:

$$\sum_{k=1}^{q}\sum_{j=1}^{p}\frac{1}{n}\left[\sum_{i=1}^{n}\frac{e_{i,j,k}^2}{\hat{\sigma}_{j,k}^2} + n\log(\hat{\sigma}_{j,k}^2) + p\log(n)\theta + \text{tr}(A_{i,j,k}) - \log|A_{i,j,k}|\right],$$

where $A_{i,j,k} = (\hat{\sigma}_{j,k}^2\sum_{i=1}^n Z_i Z_i^\top)^{-1}(\sum_{i=1}^n e_{i,j,k}^2 Z_i Z_i^\top)$, $e_{i,j,k} = \partial L(Y_i, X_i, \hat{\boldsymbol{\alpha}})/\partial\alpha_{j,k} - Z_i^\top\hat{\gamma}_{j,k}$, $\hat{\boldsymbol{\alpha}}$ is the supervised estimator defined in equation (2.6), $\hat{\gamma}_{j,k} = \left(\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\sum_{i=1}^n Z_i\partial L(Y_i, X_i, \hat{\boldsymbol{\alpha}})/\partial\alpha_{j,k}$ and $\hat{\sigma}_{j,k}^2 = \sum_{i=1}^n e_{i,j,k}^2/(n - p\theta - 1)$.

Therefore, we propose a new loss functions based on equations (3.1)-(3.3), which systematically integrate unlabeled data information into the supervised estimation process as follows:

$$\begin{aligned}\tilde{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}}\left\{\sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) + \sum_{i=n+1}^{n+N}\boldsymbol{Z}_i^\top\hat{\varphi}(\boldsymbol{\alpha})\right\} \\ &= \arg\min_{\boldsymbol{\alpha}}\sum_{i=1}^{n}\omega_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}),\end{aligned} \quad (3.4)$$

where $\hat{\varphi}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^\top/n\right)^{-1}\sum_{i=1}^n \boldsymbol{Z}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})/n$ only involves the labeled data, and

$$\omega_i = 1 + N/n\left(\sum_{i=n+1}^{n+N}\boldsymbol{Z}_i/N\right)^\top\left(\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top/n\right)^{-1}\boldsymbol{Z}_i.$$

In addition, the solution of minimizing (3.4) can also be achieved by the `iqr` function in the `qrcm R` package. From the equations (2.5) and (3.4), we can obtain

$$\tilde{\boldsymbol{\beta}}_\tau = \tilde{\boldsymbol{\alpha}}\int_0^1 \boldsymbol{b}(\bar{\tau})\text{J}_\tau(\bar{\tau})d\bar{\tau}. \quad (3.5)$$

3.3 Large sample properties

The following conditions are needed to establish the consistency and asymptotic normality of $\tilde{\boldsymbol{\alpha}}$.

C4: The unknown parameter vector $\boldsymbol{\alpha}^*$ in (3.1) is an interior point of a compact set Θ, and $\tilde{\mathbf{H}} = \mathbb{E}\{\nabla_{\text{Vec}(\boldsymbol{\alpha})}S(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}\}$ is nonsingular.

C5: The random vector $\boldsymbol{Z}$ is bounded almost surely and $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)$ is nonsingular.

Remark 3.1. The condition **C4** ensures the uniqueness of $\boldsymbol{\alpha}^*$ under the strict convexity of $\mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}$. Condition **C5** is regularity condition for the unlabeled data. Conditions **C4** and **C5** are also used in Song et al. (2024a).

Theorem 3.1. Suppose that the conditions **C2**, **C4** and **C5** hold and $n/N \to \rho$ for some constant $\rho \in [0, +\infty)$. Then as $n \to \infty$ and $N \to \infty$, we have

$$\mathrm{Vec}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \xrightarrow{\mathbb{P}} \mathbf{0},$$

and

$$\sqrt{n}\mathrm{Vec}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) \xrightarrow{\mathbb{L}} \mathbb{N}\left(\mathbf{0}, \tilde{\mathbf{H}}^{-1}\boldsymbol{\Sigma}_\rho\tilde{\mathbf{H}}^{-1}\right),$$

where $\boldsymbol{\Sigma}_\rho = \mathbb{E}(\boldsymbol{W}\boldsymbol{W}^\top) + \rho\mathbb{E}(\boldsymbol{V}\boldsymbol{V}^\top)$, $\boldsymbol{W} = S(\boldsymbol{\alpha}^*) - N(n+N)^{-1}\boldsymbol{A}^\top\boldsymbol{Z}$, $\boldsymbol{V} = N(n+N)^{-1}\boldsymbol{A}^\top\boldsymbol{Z}$ and $\boldsymbol{A} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbb{E}\{\boldsymbol{Z}S(\boldsymbol{\alpha}^*)^\top\}$.

Theorem 3.2. Suppose that conditions in Theorem 3.1 hold and $\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ is finite, we have

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^*) \xrightarrow{\mathbb{L}} \mathbb{N}\left(\mathbf{0}, \tilde{\boldsymbol{b}}(\tau)^\top\tilde{\mathbf{H}}^{-1}\boldsymbol{\Sigma}_\rho\tilde{\mathbf{H}}^{-1}\tilde{\boldsymbol{b}}(\tau)\right),$$

where $\tilde{\boldsymbol{\beta}}_\tau = \tilde{\boldsymbol{\alpha}}\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ and $\boldsymbol{\beta}_\tau^* = \boldsymbol{\alpha}^*\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$.

3.4 Comparison between supervised learning and semi-supervised learning

Theorems 2.2 and 3.2 have important implication on the asymptotic efficiency comparison between the supervised estimator $\hat{\boldsymbol{\beta}}_\tau$ in (2.7) and the semi-supervised estimator $\tilde{\boldsymbol{\beta}}_\tau$ in (3.5). From Theorem 2.2, $\boldsymbol{\Sigma}$ can be written as the following form as $\boldsymbol{U} = S(\boldsymbol{\alpha}^*) - \boldsymbol{A}^\top\boldsymbol{Z}$ and $\boldsymbol{A}^\top\boldsymbol{Z}$ are uncorrelated,

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{U}\boldsymbol{U}^\top) + \mathbb{E}\left\{(\boldsymbol{A}^\top\boldsymbol{Z})(\boldsymbol{A}^\top\boldsymbol{Z})^\top\right\}.$$

Moreover, $\boldsymbol{\Sigma}_\rho$ can be rewrite as:

$$\boldsymbol{\Sigma}_\rho = \mathbb{E}(\boldsymbol{U}\boldsymbol{U}^\top) + \frac{n}{n+N}\mathbb{E}\left\{(\boldsymbol{A}^\top\boldsymbol{Z})(\boldsymbol{A}^\top\boldsymbol{Z})^\top\right\}.$$

Note that $n/(n+N) \leq 1$. Therefore, the semi-supervised estimator $\tilde{\boldsymbol{\beta}}_\tau$ is equally or more efficient than the supervised estimator $\hat{\boldsymbol{\beta}}_\tau$ according to $\boldsymbol{\Sigma}_\rho \leq \boldsymbol{\Sigma}$.

3.5 Estimation of covariance

Finally, we provide consistent analytical estimators of the components of the variances below:

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n \int_0^1 \{\boldsymbol{X}_i^\top\boldsymbol{\alpha}\nabla_{\bar{\tau}}\boldsymbol{b}(\bar{\tau})\}^{-1}\{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i\}\{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i\}^\top d\bar{\tau},$$

$$\hat{\boldsymbol{\Sigma}}_\rho = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{W}}_i\hat{\boldsymbol{W}}_i^\top + \frac{1}{N}\sum_{i=n+1}^{n+N} \hat{\boldsymbol{V}}_i\hat{\boldsymbol{V}}_i^\top,$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \hat{S}_i(\hat{\boldsymbol{\alpha}})\hat{S}_i(\hat{\boldsymbol{\alpha}})^\top,$$

where $\hat{\boldsymbol{W}}_i = \hat{S}_i(\tilde{\boldsymbol{\alpha}}) - N(n+N)^{-1}\hat{\boldsymbol{A}}^\top \boldsymbol{Z}_i$, $\hat{\boldsymbol{V}}_i = N(n+N)^{-1}\hat{\boldsymbol{A}}^\top \boldsymbol{Z}_i$, $\hat{S}_i(\boldsymbol{\alpha}) = \boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i \int_0^1 [\mathrm{I}(Y_i < \{\boldsymbol{b}(\bar{\tau}) \otimes \boldsymbol{X}_i\}^\top \mathrm{Vec}(\boldsymbol{\alpha})) - \bar{\tau}] d\bar{\tau}$ and $\hat{\boldsymbol{A}} = (\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i)^{-1} \sum_{i=1}^n \{\boldsymbol{Z}_i \hat{S}_i(\tilde{\boldsymbol{\alpha}})^\top\}$. The consistency of these estimators follows from the law of large numbers and consistency of $\mathrm{Vec}(\hat{\boldsymbol{\alpha}})$ and $\mathrm{Vec}(\tilde{\boldsymbol{\alpha}})$, as all the components are continuous functions of the parameters. Therefore, the limiting covariance matrices $\tilde{\boldsymbol{b}}(\tau)^\top \mathbf{H}^{-1} \boldsymbol{\Sigma} \mathbf{H}^{-1} \tilde{\boldsymbol{b}}(\tau)$ and $\tilde{\boldsymbol{b}}(\tau)^\top \tilde{\mathbf{H}}^{-1} \boldsymbol{\Sigma}_\rho \tilde{\mathbf{H}}^{-1} \tilde{\boldsymbol{b}}(\tau)$ in Theorems 2.2 and 3.2 can be estimated by $\tilde{\boldsymbol{b}}(\tau)^\top \hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\Sigma}} \hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})^{-1} \tilde{\boldsymbol{b}}(\tau)$ and $\tilde{\boldsymbol{b}}(\tau)^\top \hat{\mathbf{H}}(\tilde{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\Sigma}}_\rho \hat{\mathbf{H}}(\tilde{\boldsymbol{\alpha}})^{-1} \tilde{\boldsymbol{b}}(\tau)$, respectively.

4. Numerical studies

In this section, we initially employ Monte Carlo simulation studies to evaluate the finite sample performance of the proposed procedures. Subsequently, we illustrate the practical application of these methods through the analysis of three real datasets. All the programs utilized are coded in R.

4.1 Simulation example 1: the performance of new linear extremile regression

In this section, we analyze the performance of the new linear extremile regression method introduced in Section 2. Specifically, we compare the supervised learning (SL) approach for estimating $\hat{\boldsymbol{\beta}}_\tau$ with the ordinary estimator (OE) given in (1.2), utilizing the Nadaraya-Watson method to estimate $\mathrm{F}(\cdot|\mathbf{X})$. Data are generated from the following linear model:

$$\mathbf{Y} = \mathbf{X}^\top \boldsymbol{\beta}_0 + \sigma(\mathbf{X})(\boldsymbol{\varepsilon} - \hat{e}_\tau), \tag{4.1}$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2)^\top$, and $\mathbf{X}_1$ and $\mathbf{X}_2$ are independently drawn from a uniform distribution $U(0,1)$. The true parameter value is set as $\boldsymbol{\beta}_0 = (1,2,3)^\top$. The estimator $\hat{e}_\tau$, defined as $\hat{e}_\tau = \sum_{i=1}^{\tilde{n}} \left\{\mathrm{H}_\tau\left(\frac{i}{\tilde{n}}\right) - \mathrm{H}_\tau\left(\frac{i-1}{\tilde{n}}\right)\right\} \varepsilon_{i,\tilde{n}}$, represents the τ-th extremile of $\boldsymbol{\varepsilon}$ (Daouia et al., 2019). This ensures that the true value of $\boldsymbol{\beta}_0$ remains $(1,2,3)^\top$ across different τ values. Here, $\varepsilon_{1,\tilde{n}} \le \varepsilon_{2,\tilde{n}} \le \cdots \le \varepsilon_{\tilde{n},\tilde{n}}$ denotes the ordered sample with $\tilde{n} = 10^6$, and $\mathrm{H}_\tau(t) = \{1 - (1-t)^{s(\tau)}\} \cdot I(0 < \tau \le 1/2) + t^{r(\tau)} \cdot I(1/2 \le \tau < 1)$. Consequently, under the model setup (4.1), we have $\xi_\tau(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}_0$. We consider two error distributions for $\boldsymbol{\varepsilon}$: a standard normal distribution $N(0,1)$ and a t-distribution with 5 degrees of freedom $t(5)$. Additionally, we examine two scenarios for $\sigma(\mathbf{X})$: (i) $\sigma(\mathbf{X}) = 0.25$ and (ii) $\sigma(\mathbf{X}) = 0.1\sqrt{1 + |\mathbf{X}_1| + |\mathbf{X}_2|}$.

To evaluate the performance of the estimation method, we compute the total absolute error : $\mathrm{TAE} = \sum_{j=1}^3 |\hat{\boldsymbol{\beta}}_{\tau,j} - \boldsymbol{\beta}_{0,j}|$ and the percentage of relative TAE between OE and SL as $\mathrm{PRTAE} = (\mathrm{TAE}_{OE} - \mathrm{TAE}_{SL})/\mathrm{TAE}_{SL} \times 100\%$. The simulation results, showing the means of TAEs and PRTAEs based on $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ and a sample size of $n = 500$, are presented in Table 4.1. These results are based on 500 simulation replications.

The simulation results in Table 4.1 indicate that the performance of SL is superior to that of OE under different error distributions, τ values, and Cases. This suggests that our proposed SL method has indeed enhanced estimation accuracy compared to OE, which is consistent with the theoretical findings. Our proposed estimator exhibits $\sqrt{n}$ consistency, whereas OE demonstrates a lower level of consistency than $\sqrt{n}$. Furthermore, it is evident that at extreme quantile levels (the focus of extremile regression), our results (SL) significantly outperform those of OE. Although the values of SL and OE are quite similar at

$\tau = 0.5$, this is because at this point ($\tau = 0.5$), the extremile is equivalent to the mean. Therefore, in extreme value regression, the case of $\tau = 0.5$ receives less attention.

Table 1: The means and standard deviations (in parentheses) of TAEs with different errors, τs, Cases and methods.

Case	Error	Method	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
(i)	N(0,1)	TAE_{OE}	1.155 (0.219)	0.359 (0.081)	0.086 (0.047)	0.254 (0.030)	0.775 (0.091)
		TAE_{SL}	0.131 (0.067)	0.093 (0.051)	0.086 (0.048)	0.093 (0.050)	0.130 (0.065)
		PRTAE	781.7%	286.0%	0.0%	173.1%	496.2%
(i)	t(5)	TAE_{OE}	1.969 (0.491)	0.510 (0.119)	0.110 (0.055)	0.339 (0.044)	1.197 (0.205)
		TAE_{SL}	0.193 (0.098)	0.118 (0.060)	0.105 (0.053)	0.116 (0.058)	0.188 (0.091)
		PRTAE	920.2%	332.2%	4.8%	192.2%	536.7%
(ii)	N(0,1)	TAE_{OE}	0.406 (0.106)	0.129 (0.023)	0.048 (0.026)	0.145 (0.018)	0.427 (0.046)
		TAE_{SL}	0.076 (0.065)	0.052 (0.030)	0.047 (0.026)	0.051 (0.027)	0.071 (0.035)
		PRTAE	434.2%	148.1%	2.1%	184.3%	501.4%
(ii)	t(5)	TAE_{OE}	0.679 (0.245)	0.173 (0.037)	0.061 (0.031)	0.195 (0.026)	0.649 (0.111)
		TAE_{SL}	0.103 (0.051)	0.064 (0.032)	0.057 (0.029)	0.063 (0.032)	0.102 (0.048)
		PRTAE	559.2%	170.3%	7.0%	209.5%	536.3%

4.2 Simulation example 2: the performance of semi-supervised learning

In this section, we evaluate the performance of the semi-supervised learning introduced in Section 3 by conducting experiments on datasets generated from two nonlinear regression models:

$$\mathbf{Y} = \alpha_0 + \boldsymbol{X}^\top \boldsymbol{\alpha}_1 + 5|\boldsymbol{X}_1 + \boldsymbol{X}_2| + 0.5\boldsymbol{\varepsilon}_1, \tag{4.2}$$

and

$$\mathbf{Y} = \alpha_0 + \boldsymbol{X}^\top \boldsymbol{\alpha}_1 + 5\log|\boldsymbol{X}_1 + \boldsymbol{X}_2| + 0.5\{1 + |\cos(\sum_{j=1}^{5} \boldsymbol{X}_j)|\}\boldsymbol{\varepsilon}_2, \tag{4.3}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_5)^\top$ with each $\{\boldsymbol{X}_j\}_{j=1}^5 \sim N(0,1)$ independently, the true parameter vector is set as $(\alpha_0, \boldsymbol{\alpha}_1^\top)^\top = (1, 0.5, 0.5, 0.5, 0.5, 0.5)^\top$, $\boldsymbol{\varepsilon}_1 \sim N(0,1)$ and $\boldsymbol{\varepsilon}_2 \sim t(5)$. We consider a labeled sample of size $n = 500$ and two unlabeled samples of sizes $N = 500$ and $N = 800$, respectively.

Given that the data are generated from the nonlinear models (4.2) and (4.3), the linear extremile regression model becomes misspecified. Consequently, we evaluate the performance of the semi-supervised learning (SSL) estimator $\tilde{\boldsymbol{\beta}}_\tau$ defined in (3.5) by introducing the percentage asymptotic relative efficiency (PARE), defined as $\text{PARE} = (\text{ESD}_{SL} - \text{ESD}_{SSL})/\text{ESD}_{SSL} \times 100\%$, where ESD_{SL} and ESD_{SSL} denote the empirical standard deviations of the supervised learning (SL) estimator in (2.7) and SSL estimator respectively. The simulation results of the means of estimators, their empirical standard deviations and PAREs based on $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ are shown in Tables 4.2 and 4.3, which are based on 500 simulation replications.

Empirical analyses based on models (4.2) and (4.3) indicate that model misspecification predominantly affects $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. As demonstrated by the PARE metric, the semi-supervised estimator $\tilde{\boldsymbol{\beta}}_\tau$ achieves standard deviation reductions of over 20% for the coefficients associated with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in the majority of cases; by contrast, its estimation precision for all other covariates remains statistically comparable to that of the supervised estimator (SL). Further examination of $N = 500$ and $N = 800$ configurations indicates that SSL's efficiency gains become more pronounced with larger unlabeled datasets, as reflected in the upward trend of PARE values. These results collectively demonstrate that SSL effectively mitigates estimator volatility under model misspecification, particularly for the most affected variables, while maintaining stable performance for specified components.

Table 2: Means, standard deviations (in parentheses), and PAREs of coefficient estimates under model (4.2) with different τs and Ns.

		SSL		PARE	
τ	SL	N=500	N=800	N=500	N=800
0.1	2.095 (0.125)	2.116 (0.115)	2.112 (0.120)	8.7%	12.5%
	0.618 (1.108)	0.524 (0.841)	0.442 (0.792)	31.7%	42.9%
	0.617 (1.110)	0.520 (0.842)	0.438 (0.787)	31.8%	43.6%
	0.494 (0.110)	0.493 (0.113)	0.499 (0.110)	-2.7%	-2.7%
	0.502 (0.107)	0.502 (0.110)	0.502 (0.109)	-2.7%	-0.9%
	0.511 (0.112)	0.511 (0.114)	0.504 (0.112)	-1.8%	-1.8%
0.3	4.371 (0.155)	4.392 (0.151)	4.381 (0.156)	2.6%	9.0%
	0.574 (0.688)	0.518 (0.519)	0.460 (0.492)	32.6%	44.7%
	0.580 (0.695)	0.521 (0.524)	0.456 (0.479)	32.6%	47.2%
	0.488 (0.146)	0.489 (0.150)	0.501 (0.148)	-2.7%	-2.0%
	0.504 (0.146)	0.504 (0.150)	0.503 (0.147)	-2.7%	0.0%
	0.511 (0.146)	0.510 (0.149)	0.496 (0.148)	-2.0%	-1.4%
0.5	6.553 (0.189)	6.574 (0.174)	6.561 (0.182)	8.6%	14.3%
	0.548 (0.505)	0.504 (0.383)	0.472 (0.366)	31.9%	44.8%
	0.562 (0.518)	0.522 (0.394)	0.470 (0.356)	31.5%	45.8%
	0.486 (0.187)	0.489 (0.192)	0.500 (0.192)	-2.6%	-2.1%
	0.507 (0.192)	0.507 (0.198)	0.503 (0.193)	-3.0%	-2.1%
	0.510 (0.181)	0.508 (0.183)	0.490 (0.190)	-1.1%	-1.6%
0.7	8.720 (0.245)	8.739 (0.222)	8.721 (0.233)	10.4%	14.6%
	0.523 (0.354)	0.495 (0.282)	0.481 (0.274)	25.5%	38.3%
	0.543 (0.367)	0.522 (0.294)	0.482 (0.265)	24.8%	36.6%
	0.484 (0.247)	0.488 (0.254)	0.501 (0.257)	-2.8%	-2.3%
	0.511 (0.256)	0.511 (0.264)	0.502 (0.255)	-3.0%	-2.0%
	0.510 (0.240)	0.506 (0.242)	0.484 (0.254)	-0.8%	-2.0%
0.9	12.341 (0.375)	12.356 (0.347)	12.331 (0.352)	8.1%	12.5%
	0.499 (0.359)	0.481 (0.315)	0.486 (0.312)	14.0%	22.1%
	0.528 (0.367)	0.518 (0.322)	0.497 (0.309)	14.0%	18.8%
	0.483 (0.386)	0.487 (0.397)	0.500 (0.403)	-2.8%	-2.7%
	0.516 (0.393)	0.512 (0.405)	0.501 (0.405)	-3.0%	-3.0%
	0.509 (0.391)	0.496 (0.392)	0.478 (0.403)	-0.3%	-2.5%

Table 3: Means, standard deviations (in parentheses), and PAREs of coefficient estimates under model (4.3) with different τs and Ns.

		SSL		PARE	
τ	SL	N=500	N=800	N=500	N=800
0.1	-1.935 (0.340)	-1.917 (0.325)	-1.934 (0.316)	4.6%	5.1%
	0.478 (0.909)	0.470 (0.753)	0.480 (0.701)	20.7%	27.2%
	0.514 (0.874)	0.498 (0.739)	0.523 (0.721)	18.3%	21.2%
	0.470 (0.317)	0.468 (0.325)	0.470 (0.341)	-2.5%	-4.7%
	0.515 (0.312)	0.516 (0.326)	0.496 (0.315)	-4.3%	-2.5%
	0.501 (0.309)	0.498 (0.320)	0.490 (0.315)	-3.4%	-3.2%
0.3	0.830 (0.213)	0.829 (0.196)	0.814 (0.178)	8.7%	12.9%
	0.489 (0.502)	0.484 (0.406)	0.489 (0.381)	23.6%	28.1%
	0.508 (0.476)	0.494 (0.396)	0.509 (0.383)	20.2%	24.8%
	0.488 (0.197)	0.486 (0.206)	0.484 (0.201)	-4.4%	-3.0%
	0.510 (0.200)	0.511 (0.209)	0.493 (0.194)	-4.3%	-3.1%
	0.505 (0.190)	0.504 (0.197)	0.493 (0.196)	-3.6%	-3.1%
0.5	2.521 (0.167)	2.523 (0.140)	2.507 (0.122)	19.3%	25.4%
	0.492 (0.338)	0.492 (0.267)	0.492 (0.253)	26.6%	28.1%
	0.507 (0.317)	0.497 (0.264)	0.504 (0.253)	20.1%	26.1%
	0.494 (0.157)	0.492 (0.164)	0.492 (0.153)	-4.3%	-2.6%
	0.509 (0.157)	0.509 (0.165)	0.494 (0.152)	-4.8%	-4.6%
	0.506 (0.154)	0.504 (0.161)	0.493 (0.156)	-4.3%	-3.8%
0.7	4.222 (0.141)	4.230 (0.108)	4.219 (0.096)	30.6%	35.4%
	0.495 (0.188)	0.496 (0.149)	0.496 (0.137)	26.2%	25.5%
	0.505 (0.172)	0.500 (0.142)	0.497 (0.138)	21.1%	26.8%
	0.502 (0.141)	0.499 (0.147)	0.499 (0.132)	-4.1%	-3.0%
	0.507 (0.134)	0.508 (0.139)	0.494 (0.130)	-3.6%	-4.6%
	0.507 (0.139)	0.506 (0.146)	0.492 (0.139)	-4.8%	-4.3%
0.9	6.185 (0.159)	6.196 (0.140)	6.189 (0.127)	13.6%	18.1%
	0.497 (0.128)	0.495 (0.115)	0.498 (0.103)	11.3%	13.6%
	0.504 (0.117)	0.502 (0.106)	0.496 (0.100)	10.4%	18.0%
	0.504 (0.166)	0.501 (0.168)	0.507 (0.154)	-1.2%	-3.2%
	0.511 (0.147)	0.513 (0.151)	0.497 (0.144)	-2.6%	-2.8%
	0.508 (0.163)	0.507 (0.169)	0.489 (0.156)	-3.6%	-3.8%

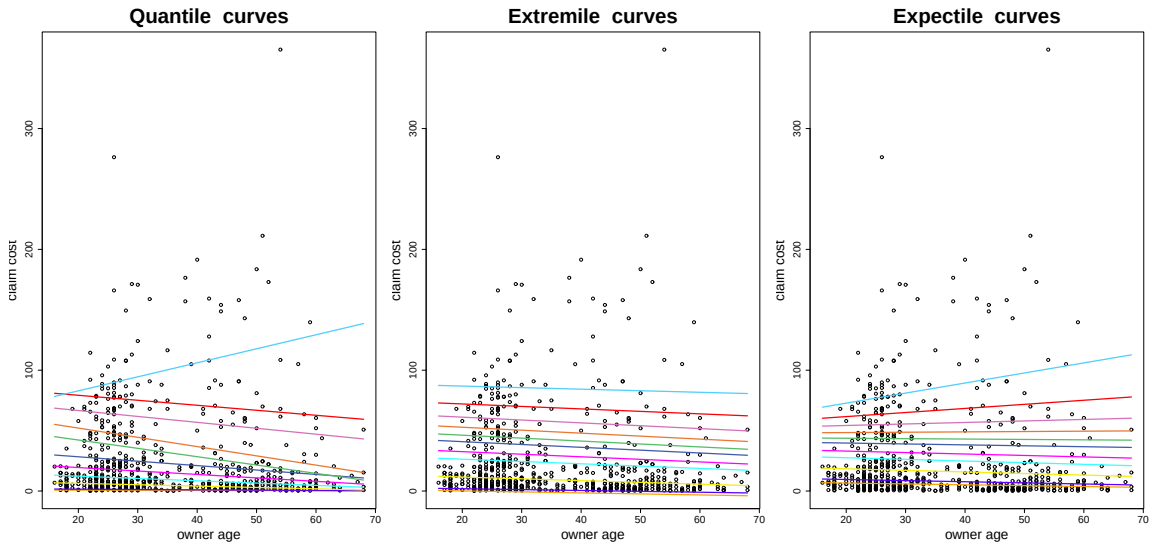

Figure 1: The quantile (left), extremile (middle) and expectile (right) curves of quantile levels 0.05, 0.1, 0.3, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95.

4.3 Real data application 1: the motorcycle insurance data

To demonstrate the application of the linear extremile regression model proposed in Section 2, we analyzed a motorcycle insurance dataset sourced from the `dataOhlsson` dataset within the R package `insuranceData`. This dataset originates from Wasa, a former Swedish insurance company, and encompasses partial casco insurance claims for motorcycles, including 670 motorcycle-related claims recorded between 1994 and 1998.

In this study, we explored the linear relationship between claim costs (expressed in thousands of US dollars) and owner age (ranging from 0 to 99 years). The scatter plot in Figure 4.1 reveals the presence of multiple outliers, suggesting that quantiles, expectiles, and extremiles are more appropriate for analyzing this dataset. Furthermore, Daouia et al. (2022) has previously examined this dataset using extremiles based on a non-parametric model. Figure 4.1 presents the linear fits of quantiles, expectiles, and extremiles at quantile levels of 0.05, 0.1, 0.3, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. Notably, the quantile curves exhibit two instances of crossover: between the 0.7 and 0.75 quantile curves, and between the 0.9 and 0.95 quantile curves. In contrast, extremile and expectile curves do not display this phenomenon. Additionally, extremiles appear to strike a balance between quantiles and expectiles. Through this example analysis, it becomes evident that extremiles avoid the unreasonable crossover phenomenon commonly observed in quantiles and produce an effect that lies between quantiles and expectiles.

4.4 Real data application 2: the mass body index (BMI) data

We compare the proposed supervised learning (SL) method in (2.7) with the ordinary estimator (OE) in (1.2) using the BMI dataset. The dataset comprises 2,111 records for estimat-

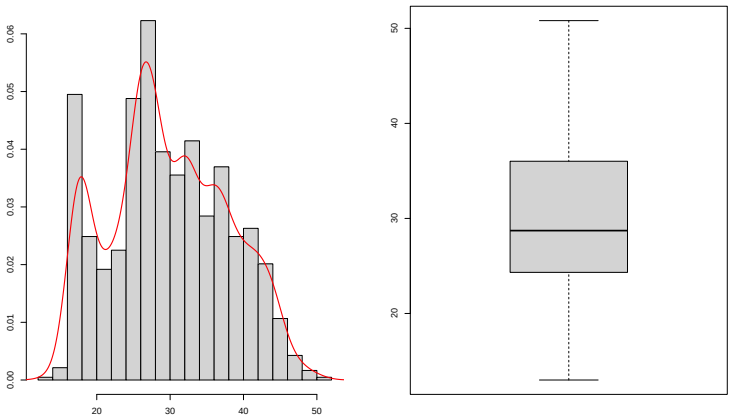

Figure 2: Histogram and Box plot of BMI.

ing obesity levels in individuals from Mexico, Peru, and Colombia. Obesity is measured via BMI = weight(kg)/height^2(m). Detailed dataset descriptions are available in Mendoza and De la Hoz Manotas (2019), and the data can be accessed at https://archive.ics.uci.edu/dataset/544/estimation+ +and+physical+condition.

This study investigates the linear relationships between BMI and key predictors: gender (0=female, 1=male), age (14–61 years), physical activity frequency (PAF), and technology usage time (TUE). Given the left heavy-tailed phenomenon observed in Figure 4.2, we employ quantile, expectile, and extremile regression (based on both SL and OE) for analysis. The estimated coefficients for quantile levels ranging from 0.05 to 0.95 (with 0.05 increments) are presented in Figure 4.3.

As shown in Figure 4.3, the performance of the proposed supervised learning method, extremile-SL, closely matches the relationship between quantiles and expectiles observed in Example 4.3. In contrast, extremile-OE exhibits notable discrepancies at extreme quantiles (Age variable) and minimum quantiles (TUE variable), particularly when compared with the other three methods. For the Gender variable, extremile-OE produces a trend entirely opposite to that of the other methods. This behavior may arise from the suboptimal performance of conditional distribution-based approaches when handling extreme quantiles and binary classification data. Analysis of coefficient estimates further reveals significant fluctuations in quantile estimates for extremile-OE, whereas extremile-SL and expectile methods display relatively stable and smooth patterns. Collectively, these observations indicate that the SL method (2.7) outperforms the OE method (1.2).

4.5 Real data application 3: the homeless data in Los Angeles County

To illustrate the proposed semi-supervised learning method in Section 3, we utilize a dataset containing the total number of individuals estimated by the Los Angeles Homeless Services Administration (LAHSA) to be living on the streets, in shelters, or in “almost homeless” conditions across Los Angeles County from 2004 to 2005. Given that the county comprises 2,054 census tracts, a complete survey would be prohibitively expensive. Therefore, a strat-

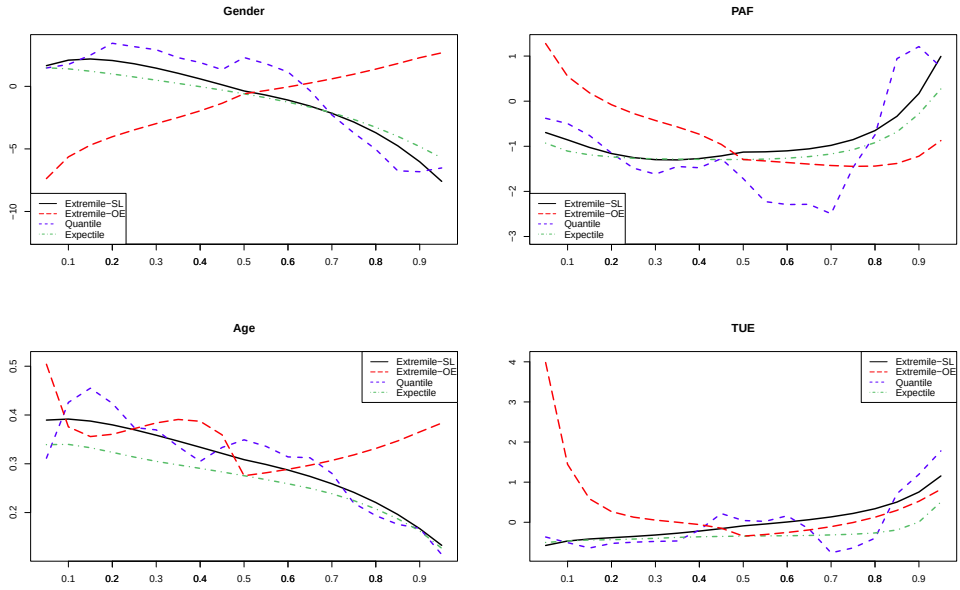

Figure 3: The estimated coefficients (vertical axis) by quantile, expectile, extremile-SL and extremile-OE under quantile levels (horizontal axis) from 0.05 to 0.95 with a step size of 0.05.

ified spatial sampling approach was employed: first targeting areas with high concentrations of homeless populations ("hot tracts") and then randomly selecting stratified samples from non-hot areas. This process resulted in 265 surveyed tracts, while the remaining 1,545 tracts remained unsurveyed. The final dataset consists of 1,810 observations (265 labeled and 1,545 unlabeled), which is available in the supplemental material of Song et al. (2024a) (https://www.tandfonline.com/doi/suppl/10.1080/01621459.2023.2169699?scroll=top).

The histogram and box plot in Figure 4.4 reveal multiple large outliers in the homeless count data. Consequently, we employ the linear extremile regression model (2.4) to analyze the relationship between street homeless counts and four key predictors: PctVacant (percentage of unoccupied housing units), PctOwnerOcc (percentage of owner-occupied housing units), PctMinority (percentage of non-Caucasian population), and MHI (median household income). These variables were identified as significant in Kriegler and Berk (2010). Table 4.4 presents the estimated regression coefficients, their standard deviations (estimated using the methods in Section 3.5), and the corresponding PARE values (defined in Section 4.2).

The results demonstrate that the proposed semi-supervised estimators generally outperform their supervised counterparts, as evidenced by predominantly positive PARE values (with the exception of the intercept term, which is not of practical concern). In most cases, the performance improvement exceeds 10%, with a maximum increase of 44.3%. These findings indicate that semi-supervised learning approaches are particularly effective for analyzing this homeless count dataset, likely due to its semi-labeled nature and the presence of significant outliers that pose challenges to traditional supervised methods.

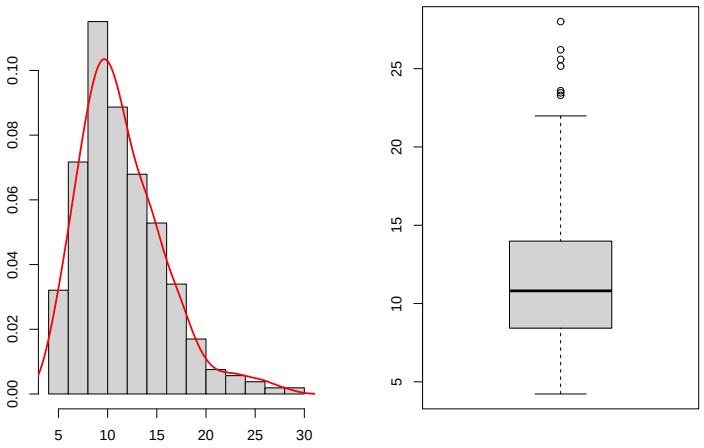

Figure 4: Histogram and Box plot of the homeless.

Table 4: The estimators and theirs estimated standard deviations (ESD) of SL and SSL and PAREs with different τs for the homeless data in Los Angeles County.

τ	Method	Intercept	PctVacant	PctOwnerOcc	PctMinority	MHI
0.1	Estimator of SL	6.909	1.027	0.894	0.822	0.713
	Estimator of SSL	7.212	0.961	0.886	0.744	1.015
	ESD of SL	0.430	0.395	0.430	0.413	0.376
	ESD of SSL	0.459	0.367	0.418	0.373	0.350
	PARE	-12.5%	15.9%	5.7%	22.1%	15.6%
0.3	Estimator of SL	9.286	0.757	0.810	0.514	0.833
	Estimator of SSL	9.687	0.794	0.913	0.471	1.207
	ESD of SL	0.558	0.508	0.533	0.539	0.533
	ESD of SSL	0.574	0.474	0.515	0.493	0.488
	PARE	-5.7%	14.8%	7.2%	19.4%	19.5%
0.5	Estimator of SL	11.352	0.710	0.812	0.326	0.887
	Estimator of SSL	11.815	0.780	1.054	0.238	1.151
	ESD of SL	0.608	0.558	0.567	0.559	0.609
	ESD of SSL	0.627	0.520	0.546	0.517	0.539
	PARE	-6.2%	15.5%	7.9%	16.8%	27.3%
0.7	Estimator of SL	13.407	0.659	0.812	0.136	0.943
	Estimator of SSL	13.933	0.763	1.193	0.008	1.101
	ESD of SL	0.674	0.620	0.609	0.587	0.701
	ESD of SSL	0.694	0.576	0.584	0.549	0.606
	PARE	-5.8%	15.9%	8.6%	14.2%	34.1%
0.9	Estimator of SL	16.703	0.790	0.867	-0.089	0.983
	Estimator of SSL	17.298	0.916	1.488	-0.443	0.825
	ESD of SL	0.752	0.681	0.664	0.582	0.810
	ESD of SSL	0.777	0.622	0.634	0.551	0.675
	PARE	-6.4%	19.6%	9.5%	11.4%	44.3%

5. Conclusion

The article introduces a novel approach to linear extremile regression that circumvents the need for estimating unknown distribution functions, as demonstrated in Daouia et al. (2022). This method achieves $\sqrt{n}$-consistent estimators for unknown regression parameters, offering significant theoretical advantages that align with standard expectations in parametric regression analysis. Moreover, the proposed estimation framework demonstrates remarkable flexibility by accommodating various τ-extremiles, making it particularly suitable for big data applications. The study further explores semi-supervised learning scenarios and, through Theorems 2.2 and 3.2 along with comprehensive simulation studies, validates the effectiveness of estimates derived from both labeled and unlabeled data. These findings collectively establish a foundation for extending extremile-based methods to more complex modeling frameworks, such as single-index models (Jiang and Yu, 2023), and to diverse data types including massive datasets (Ma and Xia, 2025).

Acknowledgments

This research was supported by the Humanities and Social Sciences Research Planning Fund of the Ministry of Education (Grant No. 25YJA910003); the National Social Science Fund of China (Grant No. 25BTJ041); the National Key R&D Program of China (Grant No. 2024YFA1013502); the National Natural Science Foundation of China (Grant Nos. U23A2064, 12531013); the Natural Science Foundation of Zhejiang Province (Grant No. LY24A010004); and the Chern Institute of Mathematics Visiting Scholar Program.

Appendix A. Proof of Theorems

Proof of Proposition 2.1. Based on (2.1)-(2.5), we can obtain that

$$\begin{aligned}\xi_\tau(\boldsymbol{X}) =& \boldsymbol{X}^\top \boldsymbol{\beta}_\tau = \int_0^1 \boldsymbol{X}^\top \boldsymbol{\gamma}(\bar{\tau}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau} = \int_0^1 q_{\bar{\tau}}(\boldsymbol{X}) \mathrm{J}_\tau(\bar{\tau}) d\bar{\tau} \\ =& \int_{y\in\mathbb{R}} y \mathrm{J}_\tau\{\mathrm{F}(y|\boldsymbol{X})\} d\mathrm{F}(y|\boldsymbol{X}) = \mathbb{E}[\boldsymbol{Y} \mathrm{J}_\tau\{\mathrm{F}(\boldsymbol{Y}|\boldsymbol{X})\}] = \mathbb{E}(\boldsymbol{Z}_{\boldsymbol{X},\tau}),\end{aligned}$$

where $\boldsymbol{Z}_{\boldsymbol{X},\tau}$ has cumulative distribution function $\mathrm{F}_{\boldsymbol{Z}_{\boldsymbol{X},\tau}} = \mathrm{H}_\tau\{\mathrm{F}(\cdot|\boldsymbol{X})\}$. When $\tau = 0.5^{1/r}$ and $r \in \mathbb{N}\backslash\{0\}$, for any $z \in \mathbb{R}$, we have

$$\mathrm{F}_{\boldsymbol{Z}_{\boldsymbol{X},\tau}}(z) = \mathrm{H}_\tau\{\mathrm{F}(z|\boldsymbol{X})\} = \{\mathrm{F}(z|\boldsymbol{X})\}^r = \mathbb{P}\left(\max(Y_{\boldsymbol{X}}^1, \ldots, Y_{\boldsymbol{X}}^r) \leqslant z\right).$$

So that $\xi_\tau(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Z}_{\boldsymbol{X},\tau}) = \mathbb{E}\{\max(Y_{\boldsymbol{X}}^1, \ldots, Y_{\boldsymbol{X}}^r)\}$ according to $\boldsymbol{Z}_{\boldsymbol{X},\tau} = \max(Y_{\boldsymbol{X}}^1, \ldots, Y_{\boldsymbol{X}}^r)$. Similarly, we can prove that $\xi_\tau(\boldsymbol{X}) = \mathbb{E}\{\min(Y_{\boldsymbol{X}}^1, \ldots, Y_{\boldsymbol{X}}^r)\}$ under $\tau = 1 - 0.5^{1/s}$ and $s \in \mathbb{N}\backslash\{0\}$. Because, for any $z \in \mathbb{R}$,

$$1 - \mathrm{F}_{\boldsymbol{Z}_{\boldsymbol{X},\tau}}(z) = 1 - \mathrm{H}_\tau\{\mathrm{F}(z|\boldsymbol{X})\} = \{1 - \mathrm{F}(z|\boldsymbol{X})\}^s = \mathbb{P}\left(\min(Y_{\boldsymbol{X}}^1, \ldots, Y_{\boldsymbol{X}}^r) > z\right).$$

Proof of Theorems 2.1 and 2.2. The results of Theorems 2.1 and 2.2 can be obtained directly from the following proofs of theorems 3.1 and 3.2 under $N = 0$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_0$.

Proof of Theorem 3.1. To establish consistency. Denote

$$\tilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) + \sum_{i=n+1}^{n+N} \boldsymbol{Z}_i^\top \hat{\varphi}(\boldsymbol{\alpha}).$$

Note that $\tilde{L}(\boldsymbol{\alpha})$ is the loss function in equation (3.4). We first show that $\tilde{L}(\boldsymbol{\alpha})$ is invariant under affine transformation on $\boldsymbol{Z} = (1, \tilde{\boldsymbol{Z}}^\top)^\top$, where $\tilde{\boldsymbol{Z}}$ is the remainder of $\boldsymbol{Z}$ after removing the first element. Assume that $\tilde{\boldsymbol{Z}}_i = \boldsymbol{M}\tilde{\boldsymbol{Q}}_i + \boldsymbol{b}$ for any fixed nonsingular $(d-1) \times (d-1)$ matrix $\boldsymbol{M}$ and $d-1$ vector $\boldsymbol{b}$, where $\tilde{\boldsymbol{Z}}_i$ is the i-th observation of $\tilde{\boldsymbol{Z}}$. Then, we can rewrite $\boldsymbol{Z}_i$ as

$$\boldsymbol{Z}_i = \begin{pmatrix} 1 & \mathbf{0}_{q-1}^\top \\ \boldsymbol{b} & \boldsymbol{M} \end{pmatrix} \boldsymbol{Q}_i,$$

where $\boldsymbol{Q}_i = (1, \tilde{\boldsymbol{Q}}_i^\top)^\top$. Then, for any $\boldsymbol{\alpha}$, we can obtain

$$\begin{aligned}\tilde{L}(\boldsymbol{\alpha}) =& \sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) + \sum_{i=n+1}^{n+N} \boldsymbol{Z}_i^\top \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) \\ =& \sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) + \sum_{i=n+1}^{n+N} \boldsymbol{Q}_i^\top \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_i \boldsymbol{Q}_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}).\end{aligned}$$

Therefore, we consider $\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top) = \boldsymbol{I}_d$ and $\mathbb{E}(\boldsymbol{Z}) = (1, \mathbf{0}_{d-1}^\top)^\top$ in the following proofs according to the above affine transformation invariant property. Let

$$\begin{aligned}\bar{L}(\boldsymbol{\alpha}) =& \sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) + N\mathbb{E}(\boldsymbol{Z}^\top)\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) \\ =& \frac{n+N}{n}\sum_{i=1}^{n} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha}).\end{aligned} \tag{A.1}$$

Next, we proof that $\sup_{\boldsymbol{\alpha}\in\Theta}|\tilde{L}(\boldsymbol{\alpha}) - \bar{L}(\boldsymbol{\alpha})| = o_p(1)$. Then, by Lemma 1 in Tauchen (1985) and conditions **C2** and **C5**, for large enough constants c_1 and c_2, we have

$$\begin{aligned}&|\tilde{L}(\boldsymbol{\alpha}) - \bar{L}(\boldsymbol{\alpha})| \\ =& \left|\left[\sum_{i=n+1}^{n+N}\boldsymbol{Z}_i^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)^{-1} - N\mathbb{E}(\boldsymbol{Z}^\top)\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\}^{-1}\right]\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})\right| \\ \le& \left\|\sum_{i=n+1}^{n+N}\boldsymbol{Z}_i^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)^{-1} - N\mathbb{E}(\boldsymbol{Z}^\top)\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\}^{-1}\right\|_2\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})\right\|_2 \\ \le& c_1\left\|\left\{\sum_{i=n+1}^{n+N}\boldsymbol{Z}_i^\top - N\mathbb{E}(\boldsymbol{Z}^\top)\right\}\left(\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)^{-1}\right\|_2 \\ &+ c_1 N\left\|\mathbb{E}(\boldsymbol{Z}^\top)\left[\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right)^{-1} - \mathbb{E}(\boldsymbol{Z}^\top)\{\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\}^{-1}\right]\right\|_2 \\ \le& c_2(N^{1/2} + Nn^{-1/2}),\end{aligned} \tag{A.2}$$

where $\|\cdot\|_2$ is L_2 norm. Then, by equations (A.1) and (A.2), and $n \to \infty$, we have

$$\begin{aligned}&\sup_{\boldsymbol{\alpha}\in\Theta}|\tilde{L}(\boldsymbol{\alpha})/(n+N) - \mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}| \\ \le& \sup_{\boldsymbol{\alpha}\in\Theta}|\tilde{L}(\boldsymbol{\alpha}) - \bar{L}(\boldsymbol{\alpha})|/(n+N) + \sup_{\boldsymbol{\alpha}\in\Theta}|\bar{L}(\boldsymbol{\alpha})/(n+N) - \mathbb{E}\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}| = o_p(1).\end{aligned} \tag{A.3}$$

(A.3) implies that $\|\mathrm{Vec}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)\|_2 = o_p(1)$ according to $\boldsymbol{\alpha}^*$ is the unique minimizer of $E\{L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\alpha})\}$. Thus, the consistency is proved.

To show asymptotic normality. By equation (3.4) and Taylor's expansion of $\nabla_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\tilde{\boldsymbol{\alpha}}}$ at $\boldsymbol{\alpha}^*$ as

$$\mathbf{0} = \nabla_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\tilde{\boldsymbol{\alpha}}} = \nabla_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} + \nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}\mathrm{Vec}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*), \tag{A.4}$$

where $\bar{\boldsymbol{\alpha}}$ is between $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$. We first consider $\nabla_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}$. Denote $\tilde{\boldsymbol{Z}}_N = \sum_{i=n+1}^{n+N}\boldsymbol{Z}_i/N$, $\tilde{\boldsymbol{Z}}_n = \sum_{i=1}^{n}\boldsymbol{Z}_i/n$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top/n$ and $\boldsymbol{U}_i = S_i(\boldsymbol{\alpha}^*) - \boldsymbol{A}^\top\boldsymbol{Z}_i$ with

$S_i(\boldsymbol{\alpha}^*) = \nabla_{\mathrm{Vec}(\boldsymbol{\alpha})} L(Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}$. Thus, we can obtain

$$\begin{aligned}
\nabla_{\mathrm{Vec}(\boldsymbol{\alpha})} \tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} =& nS(\boldsymbol{\alpha}^*) + \frac{N}{n}\sum_{i=1}^{n} S_i(\boldsymbol{\alpha}^*)\boldsymbol{Z}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\tilde{\boldsymbol{Z}}_N \\
=& nS(\boldsymbol{\alpha}^*) + \frac{N}{n}\sum_{i=1}^{n} (\boldsymbol{A}^\top \boldsymbol{Z}_i + \boldsymbol{U}_i)\boldsymbol{Z}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\tilde{\boldsymbol{Z}}_N \\
=& nS(\boldsymbol{\alpha}^*) + N\boldsymbol{A}^\top \left\{\tilde{\boldsymbol{Z}}_N - E(\boldsymbol{Z})\right\} + N\boldsymbol{A}^\top E(\boldsymbol{Z}) \\
&+ \frac{N}{n}\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{Z}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\tilde{\boldsymbol{Z}}_n + \frac{N}{n}\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{Z}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\left\{\tilde{\boldsymbol{Z}}_N - \tilde{\boldsymbol{Z}}_n\right\} \qquad \text{(A.5)} \\
=& \left\{nS(\boldsymbol{\alpha}^*) + \frac{N}{n}\sum_{i=1}^{n}\boldsymbol{U}_i\right\} + N\boldsymbol{A}^\top\left\{\tilde{\boldsymbol{Z}}_N - E(\boldsymbol{Z})\right\} \\
&+ o_p(n^{-1/2}(n+N)) \\
=& \left\{(n+N)S(\boldsymbol{\alpha}^*) - N\boldsymbol{A}^\top\tilde{\boldsymbol{Z}}_n\right\} + N\boldsymbol{A}^\top\left\{\tilde{\boldsymbol{Z}}_N - E(\boldsymbol{Z})\right\} \\
&+ o_p(n^{-1/2}(n+N)),
\end{aligned}$$

where the forth equality holds because of $\boldsymbol{A}^\top E(\boldsymbol{Z}) = \boldsymbol{0}$ by the definition of $\boldsymbol{\alpha}^*$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\tilde{\boldsymbol{Z}}_n = (1, \boldsymbol{0}_{d-1}^\top)^\top$ by Lemma 2 in Song et al. (2024a), and

$$\frac{N}{n}\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{Z}_i^\top \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\left\{\tilde{\boldsymbol{Z}}_N - \tilde{\boldsymbol{Z}}_n\right\} = o_p(n^{-1/2}(n+N)),$$

by proof similar to Theorem 1 in Song et al. (2024a). Finally, we consider $\nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}$ as

$$\begin{aligned}
\nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}} =& \nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\bar{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}} + \left\{\nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\tilde{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}} - \nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\bar{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}}\right\} \\
=& (n+N)\mathbf{H} + \left\{\nabla^2_{\mathrm{Vec}(\boldsymbol{\alpha})}\bar{L}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\bar{\boldsymbol{\alpha}}} - (n+N)\mathbf{H}\right\} + O_p(N^{1/2} + Nn^{-1/2}) \\
=& (n+N)\{\mathbf{H} + o_p(1)\},
\end{aligned} \qquad \text{(A.6)}$$

where the second equation is similar to (A.2) by conditions **C2** and **C4**, and the last equation is according to (A.1). Then, from (A.4)-(A.6), we have

$$\mathrm{Vec}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) = -\mathbf{H}^{-1}\left[\left\{S(\boldsymbol{\alpha}^*) - \frac{N}{n+N}\boldsymbol{A}^\top\tilde{\boldsymbol{Z}}_n\right\} + \frac{N}{n+N}\boldsymbol{A}^\top\left\{\tilde{\boldsymbol{Z}}_N - \mathbb{E}(\boldsymbol{Z})\right\}\right] + o_p(n^{-1/2}). \qquad \text{(A.7)}$$

Therefore, we can prove the theorem.

Proof of Theorem 3.2. From the $\tilde{\beta}_\tau = \tilde{\alpha}\int_0^1 \boldsymbol{b}(\bar{\tau})\mathrm{J}_\tau(\bar{\tau})d\bar{\tau}$ and (A.7), the theorem can be directly proven.

References

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976. doi: 10.1080/01621459.1976.10480949.

Tianxi Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:391–419, 2020. doi: 10.1111/rssb.12357.

Timothy Cannings. Random projections: Data perturbation for classification problems. *WIREs Computational Statistics*, 13:e1499, 2021. doi: 10.1002/wics.1499.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *Annals of Statistics*, 46:1541–1572, 2018. doi: 10.1214/17-AOS1594.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *the MIT Press*, 2010.

Haoyu Chen, Tiantian Mao, and Fan Yang. Estimation of the adjusted standard-deviatile for extreme risks. *Scandinavian Journal of Statistics*, 51:643–671, 2024. doi: https://doi.org/10.1111/sjos.12693.

Yu Chen, Mengyuan Ma, and Hongfang Sun. Statistical inference for extreme extremile in heavy-tailed heteroscedastic regression model. *Insurance: Mathematics and Economics*, 111:142–162, 2023. doi: 10.1016/j.insmatheco.2023.04.001.

Veronika Cheplygina, Marleen de Bruijne, and Josien Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. doi: 10.1016/j.media.2019.03.009.

Abdelaati Daouia, Irene Gijbels, and Gilles Stupfler. Extremiles: A new perspective on asymmetric least squares. *Journal of the American Statistical Association*, 114:1366–1381, 2019. doi: 10.1080/01621459.2018.1498348.

Abdelaati Daouia, Irene Gijbels, and Gilles Stupfler. Extremile regression. *Journal of the American Statistical Association*, 117:1579–1586, 2022. doi: 10.1080/01621459.2021.1875837.

Debrauwer Dieter, Gijbels Irene, and Herrmann Klaus. On a general class of functionals: Statistical inference and application to risk measures. *Electronic Journal of Statistics*, 19:2456–2510, 2025.

Timothee Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS genetics*, 9:e1003486, 2013. doi: 10.1371/journal.pgen.1003486.

Paolo Frumento and Matteo Bottai. Parametric modeling of quantile regression coefficient functions. *Biometrics*, 72:74–84, 2016. doi: 10.1111/biom.12410.

Paolo Frumento, Matteo Bottai, and Iván Fernández-Val. Parametric modeling of quantile regression coefficient functions with longitudinal data. *Journal of the American Statistical Association*, 116:783–797, 2021. doi: 10.1080/01621459.2021.1892702.

Marilena Furno. Extremiles, quantiles and expectiles in the tails. *Journal of Computational Finance*, 27:87–113, 2023. doi: 10.21314/JCF.2023.011.

Ziwen Geng. Modelling additive extremile regression by iteratively penalized least asymmetric weighted squares and gradient descent boosting. *Statistics*, 58:576–595, 2024. doi: 10.1080/02331888.2024.2348077.

Jue Hou, Rajarshi Mukherjee, and Tianxi Cai. Efficient and robust semi-supervised estimation of average treatment effect with partially annotated treatment and response. *Journal of Machine Learning Research*, 26:1–77, 2025.

Rong Jiang and Keming Yu. No-crossing single-index quantile regression curve estimation. *Journal of Business & Economic Statistics*, 41:309–320, 2023. doi: 10.1080/07350015.2021.2013245.

Rong Jiang, M. C. Jones, Keming Yu, and Jiangfeng Wang. Average quantile regression: a new non-mean regression model and coherent risk measure. *DOI: 10.48550/arXiv.2506.23059*, 2025.

Roger Koenker and Gilbert Bassett. Regression quantile. *Econometrica*, 46:33–50, 1978. doi: 10.2307/1913643.

Brian Kriegler and Richard Berk. Small area estimation of the homeless in los angeles: an application of cost-sensitive stochastic gradient boosting. *Annals of Applied Statistics*, 4:1234–1255, 2010. doi: 10.1214/10-AOAS328.

Michael Leblanc, James Moon, and Charles Kooperberg. Extreme regression. *Biostatistics*, 7:71–84, 2006. doi: 10.1093/biostatistics/kxi041.

Jinchi. Lv and Jun. S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society, Series B*, 76:141–167, 2014.

Xuejun Ma and Xiaochao Xia. An algorithm for distributed parameter estimation in modal regression models. *Statistical Theory and Related Fields*, 9:101–123, 2025. doi: 10.1080/24754269.2025.2483553.

Fabio Mendoza and Alexis De la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, 25:104344, 2019. doi: 10.1016/j.dib.2019.104344.

Jacob Michaelson, Salvatore Loguercio, and Andreas Beyer. Detection and interpretation of expression quantitative trait loci (eqtl). *Methods (San Diego, Calif.)*, 48:265–76, 2009. doi: 10.1016/j.ymeth.2009.03.004.

Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847, 1987.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119:1065–1075, 2024a. doi: 10.1080/01621459.2023.2169699.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. Semi-supervised inference for block-wise missing data without imputation. *Journal of Machine Learning Research*, 25:1–36, 2024b.

Weixi Sun and Shanshan Wang. Remire: Robust extremile regression in high dimensions. *Doi: 10.21203/rs.3.rs-5161987/v1*, 2024.

George Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30:415–443, 1985. doi: 10.1016/0304-4076(85)90149-6.

Daixin Wang, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, Shuang Yang, and Yuan Qi. A semi-supervised graph attentive network for nancial fraud detection. *2019 IEEE International Conference on Data Mining*, pages 598–607, 2019.

Mengtao Wen, Yinxu Jia, Haojie Ren, Zhaojun Wang, and Changliang Zou. Semi-supervised distribution learning. *Biometrika*, 112:asae056, 2025. doi: 10.1093/biomet/asae056.

Oren Yuval and Saharon Rosset. Semi-supervised empirical risk minimization: using unlabeled data to improve prediction. *Electronic Journal of Statistics*, 16:1434–1460, 2022.

Anru Zhang, Lawrence Brown, and Tianxi Cai. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47:2538–2566, 2019. doi: 10.1214/18-AOS1756.