IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 23 May 2023, accepted 12 June 2023, date of publication 19 June 2023, date of current version 28 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287571

RESEARCH ARTICLE

Expectile Regression With Errors-in-Variables

XIAOXIA HE, XIAODAN ZHOU, AND CHUNLI LI
College of Science, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China

Corresponding author: Xiaoxia He (hexiaoxia@wust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11201356; and in part by the Hubei Province Key Laboratory of Systems Science in Metallurgical Process, Wuhan University of Science and Technology, under Grant Y202201.

ABSTRACT This paper studies the expectile regression with error-in-variables to reduce the data error and describe the overall data distribution. Specifically, the asymptotic normality of the proposed estimator is thoroughly investigated, and an IRWLS algorithm based on orthogonal distance expectile regression (ODER) is proposed to estimate the parameters. Extensive simulation studies and real data applications evaluate our method's capabilities in reducing the measurement error bias, demonstrating our model's parameter estimation effectiveness, and its capability in reducing the simulation error compared with linear and quantile regression schemes.

INDEX TERMS Errors-in-variables, expectile regression, IRWLS algorithm, orthogonal distance regression.

I. INTRODUCTION

The errors-in-variables model was proposed by Deaton [6], aiming to solve the problem when the covariates cannot be measured accurately. The measurement error originates from the instruments used to measure the variables of interest or the inadequacy of the measurements taken over the short term used as proxies for long-term variables. The linear errors-in-variables model is typically defined as

$$Y = X^T\beta + \epsilon, \quad W = X + U, \tag{1}$$

where $\{X_i\}$ is a sequence of p-dimensional regression vectors with first component $X_{i1} = 1, i = 1, 2, \cdots n$, Y is the response, $\beta = (\beta_1, \beta_2, \cdots, \beta_{p+1})^T$ is a vector of $p+1$-dimensional unknown parameters, ϵ is the error. Since the covariate X is measured with errors, we cannot observe X, but we can observe W. Besides, U is the measurement error, which is independent of X, ϵ. We also assume that $Cov(U) = \sum_U$, where $\sum_U$ is known and $(\epsilon, U^T)^T \in \mathbb{R}^{p+n}$ are independent with a common error distribution that is spherically symmetric. From LUDWIG (1991), it is known that the A $(p+1) \times 1$ random vector is said to have a spherically symmetric distribution if X and HX have the same distribution for all $(p+1) \times (p+1)$ orthogonal matrices.

It is known that regressing Y_i on the W_i using traditional methods leads to an inconsistent and biased estimate of β Carroll et al. [3]. Thus, current literature mostly focused on describing the relationship between the true covariates and the variables of interest accurately, by adjusting the measurement error Fuller [7], Cui and Li [5], Jiang et al. [12], Shen [1], Nghiem et al. [16].

Quantile regression, proposed by Koenker and Bassett [14], has become a popular paradigm to describe the complete conditional distribution information hidden in variables. Besides, He and Liang [8] considered the regression quantile estimates when the error variables for both the response and the manifest variables have a joint distribution that is spherically symmetric but is otherwise unknown. Wei and Carrol [27] proposed a new method to correct measurement errors by constructing joint estimating equations. Moreover, the regression quantile has also attracted attention Shim [13], Jiang [11], and Yang and Yang [28].

Expectile regression, proposed by Newey and Powell [20], is an alternative way to observe the tail of a distribution, which states that expectile regression is prior over quantile regression in some folds. Due to the asymmetric least squares loss function, the regression expectiles are easy to obtain via standard gradient optimization algorithms, with the related estimators providing a convenient and relatively efficient way to summarize the condition distribution of a dependent

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

variable under various repressor values. Additionally, Waltrup et al. [24] proposed that expectile regression tends toward less crossing and more robustness for heavy-tailed distributions than quantile regression. The θ-expectile of the random variable Y is defined as follows

$$e_\theta(Y) = argmin_\nu E[\phi_\theta(Y - \nu)], \tag{2}$$

where ϕ_θ is the asymmetric least squares loss function of the form

$$\phi_\theta(u) = \theta u^2 I_{[u\geq 0]} + (1-\theta)u^2 I_{[u<0]}, \tag{3}$$

and $I_{[\cdot]}$ denotes the indicator function. Since there is no direct and explicit explanation of the expectile, the regression expectile estimator has some good properties, such as translation invariance and homogeneity monotonicity with respect to the level θ(See Pan and Liu.(2021)).

The expectile regression properties have been thoroughly investigated Jones [12], Schnabel and Eilers [21], Sobotka and Kneib [22]; Kook-Lyeol et al. [4], Waltrup and Kauermann [25], with lly, Sobotka et al. [23] establishing the asymptotic normal properties of a geo-additive expectile regression estimator and introduced confidence intervals based on the asymptotic normal distribution results. For specific data, Zhao et al. [9] developed the expectile regression for ultrahigh-dimensional data, while Mohammedi et al. [18] obtained the consistency and asymptotic normality of the kernel-type expectile regression estimator for functional data. Pan et al. [29] developed a weighted expectile regression approach for estimating the conditional expectile when covariates are missing at random (MAR). Moreover, they obtained the asymptotic normality of the proposed weighted estimators.

This paper focuses on an expectile regression model for the errors-in-variables. Specifically, we propose a new class of estimators using the idea of ODR (Orthogonal Distance Regression, see Boggs and Rogers [2], Shim [13] and establish the asymptotic normality of the estimator.

The remainder of this paper is organized as follows. Section II proposes expectile regression with ODR when the covariates suffer from measurement errors. Besides, this section establishes the asymptotic normality of the estimator. Section III proposes an IRWLS algorithm to calculate the estimator in the error-in-variables model, and evaluates the proposed method's performance on simulation scenarios. Section IV applies our model to the ACTG315 dataset to study the relationship between HIV viral load and the number of CD4+T cells in AIDS patients. Then, the obtained results are analyzed. Finally, Section V concludes this paper.

II. METHODOLOGY AND MAIN RESULTS

Let $\{X_i, Y_i\}_{i=1}^n$ be an independent and identically distributed sample from the model. The true parameter vector β_0 satisfies that

$$\beta_0 = \text{argmin}_\beta E\left[\phi_\theta\left(\frac{Y-\beta^T W}{\sqrt{1+\|\beta\|^2}}\right)\right]. \tag{4}$$

where $\phi_\theta(u)$ is defined in equation (3).

The measurement error correction factor $\sqrt{1+\|\beta\|^2}$ is widely used in linear models See Lindley [15], He and Liang [8], Ma and Yin [19]. Similar to He and Liang [8], we propose the empirical loss function as follows

$$\beta_0 = \text{argmin}_\beta \sum_{i=1}^n \left[\phi_\theta\left(\frac{Y_i-\beta^T W_i}{\sqrt{1+\|\beta\|^2}}\right)\right]. \tag{5}$$

Let the following assumptions hold:

(A1: The distribution function F of ε is absolutely continuous with uniform continuous densities f and bounded away from 0 and ∞ at the points C_θ, and $E\varepsilon^2 < \infty$.

(A2: Let $EX = 0$, $\Sigma_0 = EXX^T$ is positive definite.

Theorem 1: Under conditions (A1) and (A2), we have

$$\sqrt{n}(\hat{\beta}-\beta_0) \rightarrow^L N(0, (g(\theta))^{-2}(1+\|\beta_0\|^2)\Sigma_1^{-1}S\Sigma_1^{-1}) \tag{6}$$

where $g(\theta) = 2(1-\theta)F(0) + 2\theta(1-F(0))$,

$$S = \theta(1-\theta)\Sigma_0 + Cov\left(\psi_\theta\left(\frac{\epsilon-\beta_0^T U}{\sqrt{1+\|\beta_0\|^2}}\right)\left(U + \frac{(\epsilon-\beta_0^T U)\beta_0}{1+\|\beta_0\|^2}\right)\right), \tag{7}$$

and $\Sigma_1 = \frac{\Sigma_0}{1+\|\beta_0\|^2}$.

Proof: Denote

$$Q_n(\beta) = \sum_{i=1}^n \left[\phi_\theta\left(\frac{Y_i - W_i^T\beta}{\sqrt{1+\|\beta\|^2}}\right) - \phi_\theta\left(\frac{Y_i - W_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\right],$$

$$\begin{aligned} Q_{1i}(\beta) = {} & \phi_\theta\left(\frac{\epsilon_i - U_i^T\beta}{\sqrt{1+\|\beta\|^2}} - \frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right) \\ & - \phi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right) \\ & - \psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\left(\frac{\epsilon_i - U_i^T\beta}{\sqrt{1+\|\beta\|^2}} - \frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}} - \frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right), \end{aligned}$$

$$Q_{2i}(\beta) = \psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\left(\frac{\epsilon_i - U_i^T\beta}{\sqrt{1+\|\beta\|^2}} - \frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}} - \frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right), \tag{8}$$

where $\psi_\theta(u) = 2|\theta - I_{[u<0]}|u$ is the gradient function of $\phi_\theta(u)$. Then we have $Q_n(\beta) = \sum_{i=1}^n Q_{1i}(\beta) + \sum_{i=1}^n Q_{2i}(\beta)$, and $EQ_n(\beta) = \sum_{i=1}^n [Q_{1i}(\beta) - EQ_{1i}(\beta)] + \sum_{i=1}^n [Q_{2i}(\beta) - EQ_{2i}(\beta)]$.

On the other hand,

$$\begin{aligned}\sum_{i=1}^{n} E\left[\phi_\theta\left(\frac{Y_i - W_i^T\beta}{\sqrt{1+\|\beta\|^2}}\right) - \phi_\theta\left(\frac{Y_i - W_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\right] \\ = \sum_{i=1}^{n} E\left[\phi_\theta\left(\epsilon_i - \frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right) - \phi_\theta(\epsilon_i)\right] \\ \equiv \sum_{i=1}^{n} E\left[M\left(\frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right)\right],\end{aligned}$$

where $M(t) \equiv E[\phi_\theta(\epsilon_i - t) - \phi_\theta(\epsilon_i)]$. From condition (A2), $M(t)$ has a unique minimizer at zero, and from the Taylor expansion at zero Pan et al. [29], we have $M'(0) = -E[\psi_\theta(\epsilon_i)] = 0, M''(0) = 2g(\theta)$, then $M(t) = g(\theta)t^2 + o(t^2)$.

For a large enough n,

$$\begin{aligned}\sum_{i=1}^{n} E\left[M\left(\frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right)\right] \\ = \sum_{i=1}^{n}\left[g(\theta)\left(\frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right)^2 \right.\\ \left. + o\left(\frac{X_i^T(\beta-\beta_0)}{\sqrt{1+\|\beta\|^2}}\right)\right] \\ \to g(\theta)(\sqrt{n}(\beta-\beta_0)))^T\Sigma_1(\beta-\beta_0))).\end{aligned}$$

According to Jiang [11] and Cui and Li [5]

$$\begin{aligned}&\sum_{i=1}^{n}[Q_{1i}(\beta) - EQ_{1i}(\beta)] \\ &\quad = o_P(\sqrt{n}(\beta-\beta_0)), \\ &\sum_{i=1}^{n} Q_{2i}(\beta) \\ &\quad = -\psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\frac{1}{\sqrt{1+\|\beta_0\|^2}}(X_i \\ &\qquad + U_i + \frac{(\epsilon_i - U_i^T\beta_0)\beta_0}{1+\|\beta_0\|^2})^T(\beta-\beta_0) + R, \\ &\sum_{i=1}^{n}[Q_{2i}(\beta) - EQ_{2i}(\beta)] \\ &\quad = -\psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\frac{1}{\sqrt{1+\|\beta_0\|^2}}(X_i \\ &\qquad + U_i + \frac{(\epsilon_i - U_i^T\beta_0)\beta_0}{1+\|\beta_0\|^2})^T(\beta-\beta_0) \\ &\qquad + o_P(\sqrt{n}(\beta-\beta_0)).\end{aligned}$$

It is easy to verify that

$$\sum_{i=1}^{n}[R - ER] = o_P(\sqrt{n}(\beta-\beta_0)).$$

Thus, it follows that

$$\begin{aligned}Q_n(\beta) \to Q_0(\beta) = g(\theta)(\sqrt{n}(\beta-\beta_0)))^T\Sigma_1(\sqrt{n}(\beta-\beta_0))) \\ - \frac{1}{\sqrt{n}}\psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\frac{1}{\sqrt{1+\|\beta_0\|^2}}(X_i \\ + U_i + \frac{(\epsilon_i - U_i^T\beta_0)\beta_0}{1+\|\beta_0\|^2})^T\sqrt{n}(\beta-\beta_0).\end{aligned}$$

From the convexity of the limiting objective function, we have that

$$\begin{aligned}&\sqrt{n}(\beta-\beta_0) \\ &= \frac{\Sigma_1^{-1}\sqrt{1+\|\beta_0\|^2}}{\sqrt{n}g(\theta)}\psi_\theta\left(\frac{\epsilon_i - U_i^T\beta_0}{\sqrt{1+\|\beta_0\|^2}}\right)\frac{1}{\sqrt{1+\|\beta_0\|^2}}(X_i \\ &\quad + U_i + \frac{(\epsilon_i - U_i^T\beta_0)\beta_0}{1+\|\beta_0\|^2}) + o_P(1).\end{aligned}$$

The proof is completed.

III. IRWLS ALGORITHM BASED ON ODER

Since β_0 lacks a convex, the standard optimal algorithm can not be adopted. Therefore, we extend the IRWLS algorithm proposed by Shim [13] to ODER (Orthogonal distance expectile regression). Specifically, by applying the orthogonal distance regression (ODR) principle Boggs and Rogers [2], we propose the error-in-variables objective function for expectile regression

$$L = \sum_{i=1}^{n} u_i(\theta)\left[(Y_i - \beta^T X_i)^2 + \sum_{j=1}^{p}(W_{ij} - X_{ij})^2\right], \quad (9)$$

where $u_i(\theta) = \theta I_{[Y_i \le W_i^T\beta]} + (1-\theta)I_{[Y_i > W_i^T\beta]}$, W_{ij}, $and X_{ij}, j = 1, 2, \cdots, p$ are the components of W_i, X_i respectively.

Our aim is to minimize L. Hence, we obtain the expectile regression in EIV models using (7), as it has the advantage over (5) because it not only affords estimating β but also $X_i, i = 1, 2, \cdots, n$. This is important, as such a strategy leads to the $\beta^T X$ estimate, where X was not observed in the training data. Given $u_i(\theta)$, taking the partial derivatives of (7) with regard to (β, X_{ij}) leads to the optimal values of (β, X_{ij}) to be the solution of

$$\begin{aligned}\frac{\partial L}{\partial\beta} &= X^T U(\theta)(Y - \beta^T X) = 0_p, \\ \frac{\partial L}{\partial X_{ij}} &= u_i(\theta)[(Y_i - \beta^T X_i)\beta_j + (W_{ij} - X_{ij})] = 0, \\ & i = 0, 1, 2, \cdots, n, \quad j = 1, 2, \cdots, p,\end{aligned}$$

where $U(\theta)$ is a $n \times n$ diagonal matrix composed of $u_i(\theta)$, $X^T = (X_1, X_2, \cdots, X_n)$, and $Y^T = (y_1, y_2, \cdots, y_n)^T$.

Note that the solutions of equations (8) and (9) cannot be obtained in a single step since $u_i(\theta)$ contains (β, X_{ij}), which is unknown. Thus we apply the iteratively reweighted least squares (IRWLS) procedure as follows: For small enough γ_1, γ_2:

Step 1. Random initialization of $\beta^{(0)}, X^{(0)}$.

step 1.1. Find $U^{(0)}$ with $\beta^{(0)}, X^{(0)}$.

step 1.2. Find $\beta^{(k)}$ from $\beta^{(k)} = ((X^{(0)})^T U^{(k-1)}(X^{(0)})^T)^{-1}$ $((X^{(0)})^T U^{(k-1)} Y)$.

step 1.3. Reiterate the above steps until convergence, such that $\|\beta^{(k+1)} - \beta^{(k)}\| < \gamma_1$.

Step 2. Find new estimates of X_{ij} from $X_i^{(k)} = (Y(\beta^{(k)}) + W)((\beta^{(k)})^T(\beta^{(k)}) + I)^{-1}$.
with β.

Step 3. Reiterate Steps 1 and 2 until convergence such that $\|X^{(k+1)} - X^{(k)}\| < \gamma_2$.

From the (β, X_{ij}) estimates obtained by the IRWLS procedure, we obtain as follows the estimator of the expectile regression function given the input vector X_i, which is not observed, while W is observed in the training data,

$$e(Y \mid X_i) = \beta^T X_i.$$

A. SIMULATION STUDIES

This section conducts some experiments to verify the efficiency of the proposed expectile regression using the errors-in-variables model. First, we describe our experimental setup, which relies on MATLAB R2019a. Next, we compare the proposed expectile regression with the errors-in-variables model, the quantile regression with errors in variables (ODQR), the linear regression(LR), and the θth quantile regression model (QR). The simulation results of the proposed expectile regression using the errors-in-variables model have been obtained using the iterative reweighted least squares (IRWLS) algorithm.

B. PERFORMANCE CRITERIA

The following evaluation criteria are used to evaluate the efficiency of the proposed expectile regression algorithm employing the errors-in-variables model. Let $\hat{e}_\theta(Y_i)$, $\hat{\beta}$, and $\hat{Y}_i$ be generated by the proposed expectile regression with the errors-in-variables model, where $e_\theta(Y_i)$, β, and Y_i are all from true data. Given that X_i is the unobserved value of the input vector, $e_\theta(Y_i)$ is the θth quantile of Y_i conditional on X_i, and Y_i is the response variable under the condition of X_i, we list the evaluation criteria as follows

$$MSE_e = \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_\theta(Y_i) - e_\theta(Y_i))^2,$$

$$RMSE_\beta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta} - \beta)^2},$$

$$SSE = \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

C. SIMULATED DATASETS

First, we prove our model's efficiency by observing the role of θ in expectile regression with an errors-in-variables model. For this trial, We consider simulated datasets under various noise types. In the following examples, $x_i, i = 1, 2, \cdots, n$ are unobservable input variables, $w_i, i = 1, 2, \cdots, n$ are the observable input variables, and $y_i, , i = 1, 2, \cdots, n$ are the observable response variables. In the above data, we consider the homogeneity of the random error term in the classical hypothesis and the existence of heteroscedasticity. The model is as follows

$$\begin{cases} Y = X\beta + \epsilon \\ W = X + U \\ E(\epsilon) = 0 \\ Var(\epsilon) = \Sigma = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2) \end{cases}$$

Example 1: x_i is from the uniform distribution $U(0, 1)$, w_i is from the normal distribution $N(x_i, 0.1)$, and y_i is from the normal distribution $N(1 + 2x_i, 0.1)$. Additionally, we consider the existence of heteroscedasticity and that the error originates from the normal distribution $N(0, {x_i}^2)$. The true $\beta_0(\theta)$ values are reported in Table 1, and the true θth expectile regression function is

$$e_\theta(x_i) = \beta_0(\theta) + 2x_i.$$

Example 2: (x_1, x_2), w_1, w_2, and Y_i come from the uniformly distribution $U(0, 1)$, normal distribution $N(X_1, 0.1)$, normal distribution $N(X_2, 0.1)$, and normal distribution $N(1 + x_1 - 2x_2, 0.1)$, respectively. Additionally, we consider the existence of heteroscedasticity and make the error come from the normal distribution $N(0, {x_i}^2)$. Table 2 reports the true $\beta_0(\theta)$, and the true θth expectile regression function is

$$e_\theta(x_1, x_2) = \beta_0(\theta) + x_1 - 2x_2.$$

Example 3: x_i, w_i, and y_i are from uniform distribution $U(0, 1)$, the normal distribution $N(x_i, 0.1)$, and the normal

TABLE 1. The true $\beta_0(\theta)$ in Example 1.

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\beta_0(\theta)$	0.5749	0.7343	0.8323	0.9750	1	1.0790	1.1676	1.2656	1.4053

TABLE 2. True $\beta_0(\theta)$ in Example 2.

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\beta_0(\theta)$	0.5749	0.7343	0.8323	0.9750	1	1.0790	1.1676	1.2656	1.4053

TABLE 3. True $\beta_0(\theta)$ in Example 3.

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\beta_0(\theta)$	0.3592	0.5891	0.7357	0.8759	1	1.1253	1.2659	1.4282	1.6408

TABLE 4. MSE_e of 100 experiments of $e_\theta(Y_i)$ for $\theta = i/10, i = 1, 2, \cdots, 9$.

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
EX1	0.0045	0.0034	0.0030	0.0028	**0.0027**	0.0028	0.0029	0.0032	0.0041
EX2	0.1384	0.0987	0.0737	0.0430	**0.0366**	0.0490	0.0591	0.0963	0.1369
EX3	0.1825	0.1573	0.0891	0.0843	**0.0945**	0.1046	0.1635	0.1881	0.2209

TABLE 5. Average of 100 estimates of β when θ=0.5 in Example 1~3. (RMSE is in parenthesis.)

Data	$\beta0$				$\beta1$				$\beta2$			
	TRUE	LR	QR	ODER	TRUE	LR	QR	ODER	TRUE	LR	QR	ODER
EX1	1	1.0719 (2.4084)	1.1078 (0.5179)	**1.0402** (0.4368)	2	1.7570 (5.8909)	1.7896 (0.7270)	**1.9796** (0.3841)	—	—	—	—
EX2	1	0.9971 (0.1110)	0.9047 (0.1929)	**1.0025** (0.0534)	-2	-2.1029 (0.4627)	-1.9955 (0.5837)	**-1.9993** (0.0101)	1	1.0176 (0.4228)	0.9849 (0.6145)	**0.9903** (0.0509)
EX3	1	0.9893 (0.1110)	1.1253 (0.6859)	**1.0010** (0.5036)	0.5	0.5503 (0.4627)	0.4949 (0.5837)	**0.5020** (0.0925)	1	1.2214 (0.4228)	1.0303 (0.6145)	**0.9964** (0.2804)

TABLE 6. The average of 100 estimates of β when θ=0.5 and heteroscedasticity is present in Example 1~3. (RMSE is in parenthesis.)

Data	$\beta0$				$\beta1$				$\beta2$			
	TRUE	LR	QR	ODER	TRUE	LR	QR	ODER	TRUE	LR	QR	ODER
EX1	1	0.9974 (0.0961)	0.9748 (0.3969)	**0.9961** (0.1737)	2	1.9799 (0.5230)	2.0081 (0.2748)	**1.9962** (0.0991)	—	—	—	—
EX2	1	0.9994 (0.1870)	1.0105 (0.3882)	**0.9907** (0.2372)	-2	-1.9660 (0.2812)	-1.9726 (0.2749)	**-2.0174** (0.1168)	1	1.0176 (0.1328)	0.9890 (0.1732)	**0.9959** (0.1346)
EX3	1	0.7749 (0.4427)	0.8194 (0.6424)	**0.9805** (0.2486)	0.5	0.4382 (0.5197)	0.5980 (0.3164)	**0.4909** (0.1319)	1	1.3688 (0.4376)	1.2220 (0.5713)	**0.9907** (0.1130)

distribution $N(1 + 0.5x_i + x_i^2, 0.25)$, respectively. Moreover, we consider the existence of heteroscedasticity, and the error originates from the normal distribution $N(0, {x_i}^2)$. Table 3 reports the true $\beta_0(\theta)$, and the true θth expectile regression function is given as

$$e_\theta(x_i) = \beta_0(\theta) + 0.5x_i + x_i^2.$$

Since the expectile regression is practically useful even when heterogeneity is present, we investigate the performance of the proposed method in the following Examples.

D. SIMULATED RESULTS

We conduct a small simulation study with n=100, where data are generated from Examples 1~3. We refer to Tables 1~3 to observe the influence of different θ values on the ODER model. It can be found that as θ increases from 0.1 to 0.9, the MSE_e of the ODER model first decreases and then increases by observing Table 4. So MSE_e is the smallest for $\theta = 0.5$. This simulation adopts the settings from Shim [13].

Table 4 highlights that the MSE_e of the ODER model is the smallest when $\theta = 0.5$. Therefore, we first study the parameter estimation results of different methods and the corresponding *RMSE* when $\theta = 0.5$. Table III-C compares LR, QR, and ODER, and Table III-C compares LR and ODER when heteroscedasticity is present. Then we investigate the various calculation results caused by the measurement error. Besides, Table III-D reports the relevant results using QR, ODQR (orthogonal distance quantile regression), and ODER when $\theta = 0.1, 0.5, 0.9$. When the data has heteroscedasticity, the relevant results are presented in Table III-D. Note that in all tables, bold-faced values indicate the result closest to the real value per expectile level.

Table III-C reveals that the estimated parameter using the ODER model is closer to the coefficients obtained through TRUE when compared with LR and ODQR. Moreover, the error of parameter estimates obtained with the ODER model is more diminutive. This infers that ODER has a better calculation effect than traditional calculation methods. Table III-C presents the experimental results when heteroscedasticity is

TABLE 7. The average of 100 estimates of β when $\theta = 0.1, 0.5, 0.9$ in Example 1~3. (RMSE is in parenthesis).

Data	θ	$\beta0$ TRUE	$\beta0$ QR	$\beta0$ ODQR	$\beta0$ ODER	$\beta1$ TRUE	$\beta1$ QR	$\beta1$ ODQR	$\beta1$ ODER	$\beta2$ TRUE	$\beta2$ QR	$\beta2$ ODQR	$\beta2$ ODER
EX1	0.1	0.5749	0.8178 (0.4826)	05489 (0.1072)	**0.5500** (0.1162)	2	1.8114 (1.1370)	1.9153 (0.1324)	**2.0252** (0.1106)	—	—	—	—
	0.5	1	1.1078 (0.5179)	1.0637 (0.7912)	**1.0402** (0.4368)	2	1.7896 (0.7270)	1.9577 (0.6208)	**1.9796** (0.3841)	—	—	—	—
	0.9	1.4053	1.3733 (0.4892)	1.6783 (0.6722)	**1.3893** (0.4248)	2	1.7889 (0.8690)	1.8934 (0.2873)	**1.9796** (0.4529)	—	—	—	—
EX2	0.1	0.5749	0.8693 (0.2636)	0.3340 (0.7692)	**0.5757** (0.9349)	-2	-1.9916 (0.8633)	-2.1285 (0.6763)	**-2.0071** (0.0966)	1	0.8913 (0.7961)	0.9768 (0.4436)	**1.0004** (0.1017)
	0.5	1	0.9047 (0.1929)	**1.0015** (0.4875)	1.0025 (0.0534)	-2	-1.9955 (0.5837)	-1.9772 (0.4371)	**-1.9993** (0.0101)	1	0.9849 (0.6145)	0.9794 (0.4382)	**0.9903** (0.0509)
	0.9	1.4053	1.1313 (0.2380)	1.6903 (0.5673)	**1.6401** (0.4229)	-2	-2.0112 (0.8456)	-2.0281 (0.4956)	**-1.9951** (0.0105)	1	1.0374 (0.8390)	1.0058 (0.3517)	**1.0024** (0.0510)
EX3	0.1	0.3592	0.4956 (0.6859)	0.4679 (0.3837)	**0.3697** (0.5136)	0.5	0.5334 (2.5018)	0.4642 (0.2980)	**0.4976** (0.0984)	1	0.9833 (2.5558)	0.9591 (0.3942)	**1.0001** (0.5030)
	0.5	1	1.1253 (0.6859)	0.9923 (0.1344)	**1.0010** (0.5036)	0.5	0.4949 (0.5837)	**0.5007** (0.3247)	0.5020 (0.0925)	1	1.0303 (0.6145)	0.9896 (0.2331)	**0.9964** (0.2804)
	0.9	1.6408	1.1921 (0.7712)	1.3247 (0.6281)	**1.6524** (0.5663)	0.5	0.4464 (4.1838)	0.4466 (0.4365)	**0.4990** (0.1019)	1	1.0288 (3.9214)	0.9919 (0.3552)	**1.0041** (0.1983)

TABLE 8. Average of 100 estimates of β when heteroscedasticity is present $\theta = 0.1, 0.5, 0.9$ and heteroscedasticity is present in Example 1~3. (RMSE is in parenthesis).

Data	θ	$\beta0$ TRUE	$\beta0$ QR	$\beta0$ ODQR	$\beta0$ ODER	$\beta1$ TRUE	$\beta1$ QR	$\beta1$ ODQR	$\beta1$ ODER	$\beta2$ TRUE	$\beta2$ QR	$\beta2$ ODQR	$\beta2$ ODER
EX1	0.1	0.5749	0.7450 (0.5908)	0.6911 (0.3067)	**0.4499** (0.2734)	2	1.9804 (0.4287)	2.0032 (0.3164)	**2.0026** (0.1241)	—	—	—	—
	0.5	1	0.9748 (0.3969)	0.9862 (0.1190)	**0.9961** (0.1737)	2	2.0081 (0.2748)	1.9542 (0.1615)	**1.9962** (0.0991)	—	—	—	—
	0.9	1.4053	1.1457 (0.4860)	**1.3911** (0.2511)	1.4617 (0.2475)	2	1.9978 (0.5218)	2.0032 (0.2489)	**0.9993** (0.2393)	—	—	—	—
EX2	0.1	0.5749	0.8906 (0.4570)	0.8818 (0.2566)	**0.6719** (0.1978)	-2	-1.9617 (0.3861)	-1.9785 (0.3169)	**-2.0157** (0.1281)	1	0.9608 (0.1874)	0.9961 (0.1587)	**0.9987** (0.1085)
	0.5	1	1.0105 (0.3882)	1.0091 (0.2163)	**0.9907** (0.2372)	-2	-1.9726 (0.2749)	-2.0332 (0.1199)	**-2.0174** (0.1168)	1	0.9522 (0.1732)	1.0222 (0.1289)	**0.9959** (0.1346)
	0.9	1.4053	1.6016 (0.4298)	1.5193 (0.3633)	**1.4973** (0.3247)	-2	-1.9775 (0.2614)	-2.0864 (0.2184)	**-1.9814** (0.1844)	1	1.0338 (0.1932)	0.9948 (0.1325)	**1.0019** (0.1172)
EX3	0.1	0.3592	0.7927 (0.6022)	0.4728 (0.4266)	**0.4712** (0.3693)	0.5	0.3395 (0.4153)	0.5732 (0.3621)	**0.5698** (0.2142)	1	0.7915 (0.5412)	0.9863 (0.2934)	**1.0148** (0.1635)
	0.5	1	0.8194 (0.6424)	0.8267 (0.3112)	**0.9805** (0.2486)	0.5	0.5980 (0.3164)	0.4892 (0.4197)	**0.4909** (0.1319)	1	1.2220 (0.5713)	0.9858 (0.2197)	**0.9907** (0.1130)
	0.9	1.6408	1.4450 (0.6654)	1.5164 (0.4437)	**1.7331** (0.2869)	0.5	0.6048 (0.3472)	0.5927 (0.3752)	**0.5213** (0.2419)	1	1.4324 (0.5619)	0.8148 (0.2844)	**0.9830** (0.1469)

present (boldfaced values represent the data closest to the real value), revealing that the results of different experiments are consistent with the above conclusions as a whole. Moreover, we find that even if heteroscedasticity is present, the calculated ODER result is still the closest to the true value providing the best effect. Under the same experiment, the ODER estimation error is significantly smaller than LR and QR.

Nevertheless, solely challenging the proposed scheme against LR and QR is inadequate. Thus, to verify this conclusion further, we select $\theta = 0.1, 0.5$, and 0.9, and present in Table III-D the experimental results of ODQR and QR. Given that in Table 4, the MSE_e of the ODER model is the smallest when $\theta = 0.5$, as θ increases, the calculation error first decreases and then increases. Therefore, representative θ values are 0.1, 0.5, and 0.9. In Table III-D, we compare QR, ODQR, ODER, and TRUE and conclude that the parameter estimation result of ODER is closer to TRUE most times under different θ values (boldfaced values indicate the best

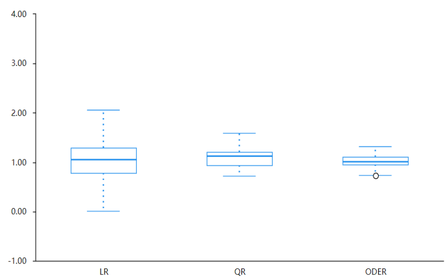


FIGURE 1. Box-plot of $\beta 0$ when $\theta = 0.5$ in Example 1.

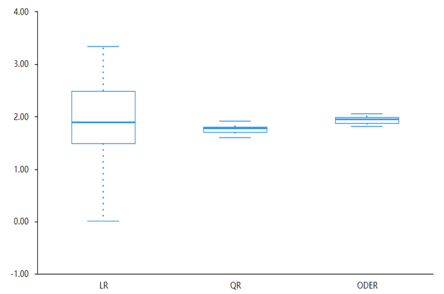


FIGURE 2. Box-plot of $\beta 1$ when $\theta = 0.5$ in Example 1.

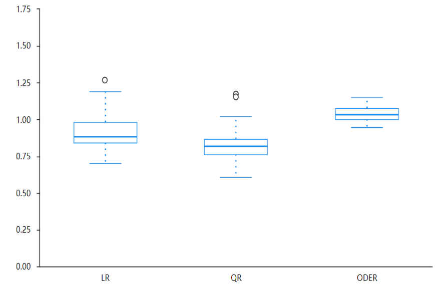


FIGURE 3. Box-plot of $\beta 0$ when $\theta = 0.5$ in Example 2.

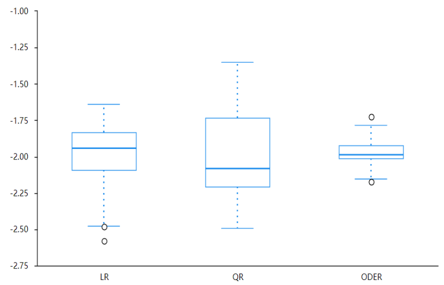


FIGURE 4. Box-plot of $\beta 1$ when $\theta = 0.5$ in Example 2.

results obtained between the different methods). The parameter estimation result of ODQR is twice as good as ODER, with the ODQR and ODER results under these two experiments being very close. Therefore, we conclude that ODER performs best in this experiment. In Table III-D, we present the experimental results with heteroscedasticity and reveal a similarity against Table III-D. Thus, we arrive at the same conclusion.

Overall, the ODER model has the best parameter estimation effect and effectively reduces the experimental error presented in Tables III-C~III-D.

Figures 1~8 are the Box-plot of the experimental parameter estimation results of the three methods when $\theta = 0.5$. Box-plot generally comprises a box with upper, middle, and lower lines. The middle line represents the median of the data, and the lower and upper lines correspond to the first and third-quartile values, respectively. Thus, the box plot represents the data point distribution in a data set, and the small black circles are the outliers.

Figures 1~2 illustrate the results of Example 1, revealing that the LR and QR results in the box plot have a wide range, inferring that the LR and QR have poor stability and cannot guarantee the model's reliability. Moreover, the box range of ODER is the smallest among the three methods, indicating that the data is relatively centralized, and the median is the closest to TRUE. Figures 3~5 depict the results of Example 2, suggesting that ODER is more accurate and demonstrating its superiority. Accordingly, Figures 6~8 illustrate the results of Example 3. Observing the vertical coordinates of the box plot of the three experiments reveals that the unit distance of the third experiment is the shortest, and thus the range of the boxes in the third experiment is the smallest. This is because of the special data structure of Example 3. For Example 3, the error between $\hat{Y}_i$ and Y_i is smaller and we calculated that $SSE_{Example1} = 5.6004$, $SSE_{Example2} = 7.9725$, $SSE_{Example3} = 3.8895$.

In order to more intuitively explore our model, we counted the 100 experimental results of Example 1~3 without the existence of heteroscedasticity. The corresponding mean value of MSE_e per ten experiments (as a group), provided ten values, with the corresponding results depicted in Figures 9~17. Since the experimental error is too small, to facilitate data visualization we enlarge the result of Example 1 by a factor of 100, and the experimental result of Example 2~3 is expanded to 10 times. The results reveal that as the number of iterations increases, the MSE_e of the experiment tends to stabilize, indicating that the gap between the experimental results and the real value reduces.

IV. REAL DATA ANALYSIS

Acquired immunodeficiency syndrome is caused by the human immunodeficiency virus (HIV), which has always been a significant research problem in medicine. HIV enters cells through the receptors on the surface of susceptible cells, directly or indirectly destroying the human immune system. The damage to the CD4+T cells is the most serious, and CD4+T lymphocytes are crucial immune cells in the human immune system. When the number of CD4+T cells in patients is small, cellular immunity is almost entirely

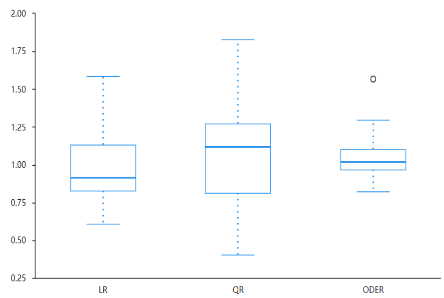


FIGURE 5. Box-plot of $\beta2$ when $\theta = 0.5$ in Example 2.

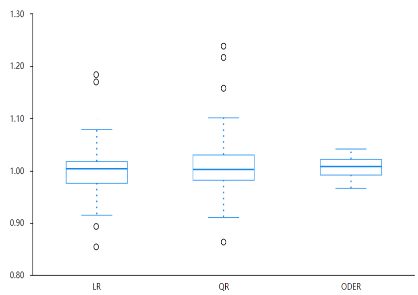


FIGURE 6. Box-plot of $\beta0$ when $\theta = 0.5$ in Example 3.

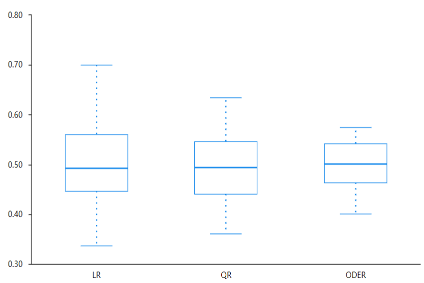


FIGURE 7. Box-plot of $\beta1$ when $\theta = 0.5$ in Example 3.

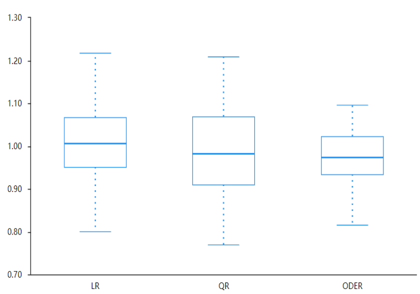


FIGURE 8. Box-plot of $\beta2$ when $\theta = 0.5$ in Example 3.

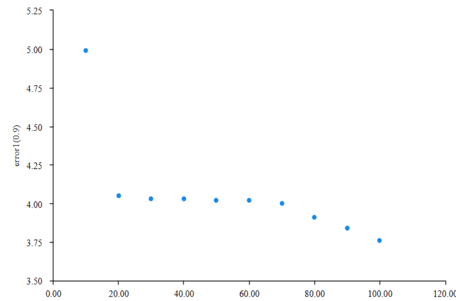


FIGURE 9. The MSE_e of 100 estimates when $\theta = 0.1$ in Example 1.

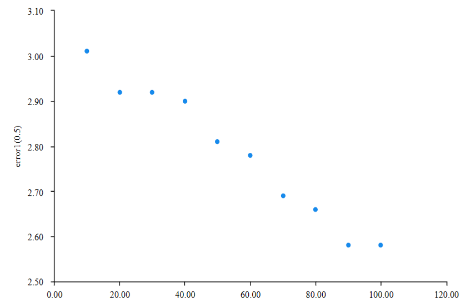


FIGURE 10. The MSE_e of 100 estimates when $\theta = 0.5$ in Example 1.

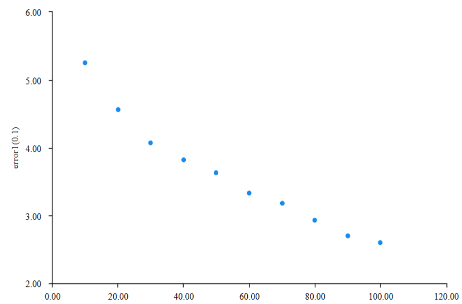


FIGURE 11. The MSE_e of 100 estimates when $\theta = 0.9$ in Example 1.

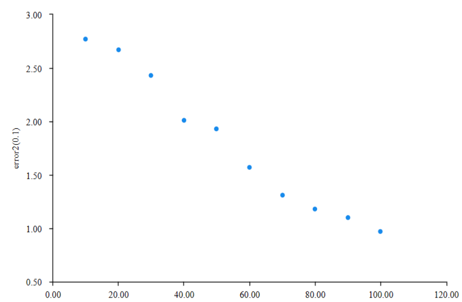


FIGURE 12. The MSE_e of 100 estimates when $\theta = 0.1$ in Example 2.

lost, significantly increasing the risk of malignancies due to virus invasion, i.e., AIDS. However, antiviral treatment can effectively reduce the incidence and mortality of AIDS. Since the viral load and CD4+ cell number are two key indicators for the AIDS treatment effect, it is necessary to study their relationship during AIDS treatment. The data in this paper are from ACTG315, which is clinical trial data of HIV-infected adults receiving antiretroviral therapy. The experiment lasted for 28 weeks and involved 46 adults with AIDS. The HIV viral load and CD4+T cell count of the subjects were spot

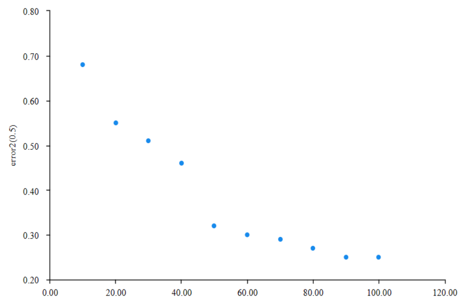


FIGURE 13. The MSE_e of 100 estimates when $\theta = 0.5$ in Example 2.

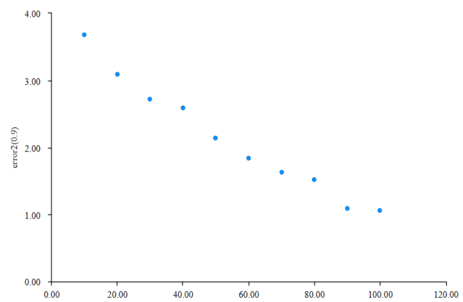


FIGURE 14. The MSE_e of 100 estimates when $\theta = 0.9$ in Example 2.

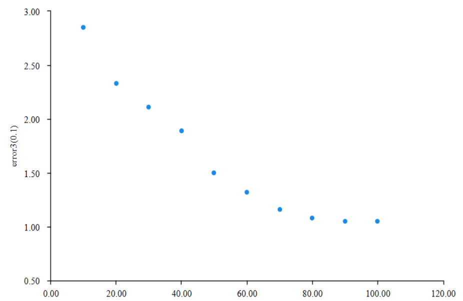


FIGURE 15. The MSE_e of 100 estimates when $\theta = 0.1$ in Example 3.

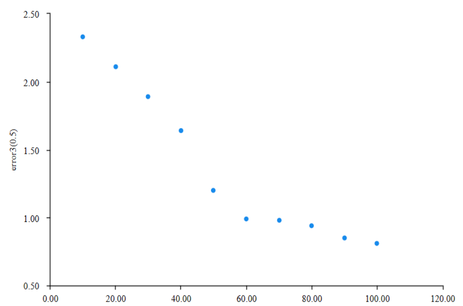


FIGURE 16. The MSE_e of 100 estimates when $\theta = 0.5$ in Example 3.

checked on a particular date during the experiment. Hence, the data label included the specific date, patient identifier, and CD4+T cell count.

This paper selects the measurement data of the second day, involving 35 groups of data with HIV viral load as the

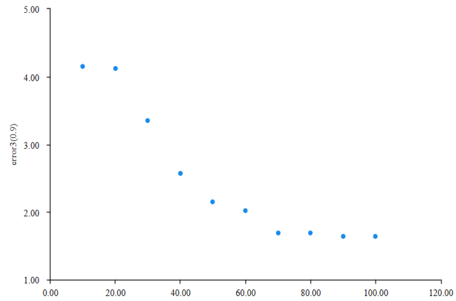


FIGURE 17. The MSE_e of 100 estimates when $\theta = 0.9$ in Example 3.

input variable and CD4+T cell count as the input response variable. The parameter estimation results obtained by our method are $\beta_0 = -0.0022$ and $\beta_1 = 184.6520$, for $\theta = 0.5$. The standard deviation of β_0 is 0.0020, the 95% confidence interval is $[-0.0031, -0.0024]$, and the standard deviation of β_1 is 28.5198, and the 95% confidence interval is [179.0872, 190.2663].

V. CONCLUSION

This paper proposes an expectile regression model with error-in-variables. A further extension to quantile regression is possible by adopting Shim [13]. However, there are two main difficulties in correcting the bias in QR caused by errors-in-variables Wang et al. [26]. One is that QR cannot entirely specify the parameter regression error, and the other is that the quantile of the sum of two random variables is not necessarily the sum of the two marginal quantiles. The same properties are also true for the expectile. Therefore, we transform the expectile regression model with error-in-variables to ODER to overcome the difficulties and establish our estimator's asymptotic normality. The simulation studies reveal that the proposed method is efficient for finite sample sizes and eliminates the estimation bias caused by measurement errors. Real data examples demonstrate the ease of implementing the proposed method.

FURTHER OUTLOOK

The proposed expectile regression with errors-in-variables model can be further studied to evaluate whether variable selection can be realized. It can also be combined with the latest machine learning methods to improve the algorithm's speed and accuracy.

CONFLICT OF INTEREST STATEMENT

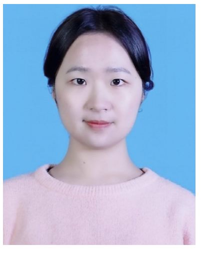
We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

REFERENCES

[1] A. Shen, "Asymptotic properties of LS estimators in the errors-in-variables model with MD errors," *Stat. Papers*, vol. 60, no. 4, pp. 1193–1206, Aug. 2019.
[2] P. T. Boggs and J. E. Rogers, "Orthogonal distance regression," Dept. Appl. Comput. Math. Division, NIST, Gaithersburg, MD, USA, Tech. Rep. NISTIR 89-4197, Nov. 1989.

[3] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL, USA: CRC Press, 2006.
[4] K.-L. Choi, J. Shim, and K. Seok, "Support vector expectile regression using IRWLS procedure," *J. Korean Data Inf. Sci. Soc.*, vol. 25, no. 4, pp. 931–939, Jul. 2014.
[5] C. Hengjian and L. Rongcai, "On parameter estimation for semi-linear errors-in-variables models," *J. Multivariate Anal.*, vol. 64, no. 1, pp. 1–24, Jan. 1998.
[6] A. Deaton, "Panel data from time series of cross-sections," *J. Econ.*, vol. 30, nos. 1–2, pp. 109–126, Oct. 1985.
[7] W. A. Fuller, "Measurement error models," *J. Appl. Econ.*, vol. 3, no. 4, pp. 315–317, 1987.
[8] X. He and H. Liang, "Quantile regression estimates for a class of linear and partially linear errors-in-variables models," *Stat. Sin.*, vol. 10, no. 1, pp. 129–140, Jan. 2000.
[9] J. Zhao, Y. Chen, and Y. Zhang, "Expectile regression for analyzing heteroscedasticity in high dimension," *Statist., Probab. Lett.*, vol. 137, pp. 304–311, Jun. 2018.
[10] R. Jiang, X. Yang, and W. Qian, "Random weighting M-estimation for linear errors-in-variables models," *J. Korean Stat. Soc.*, vol. 41, no. 4, pp. 505–514, Dec. 2012.
[11] R. Jiang, "Composite quantile regression for linear errors-in-variables models," *Hacettepe J. Math. Statist.*, vol. 44, no. 3, pp. 707–713, 2015.
[12] M. C. Jones, "Expectiles and M-quantiles are quantiles," *Statist., Probab. Lett.*, vol. 20, no. 2, pp. 149–153, May 1994.
[13] J. Shim, "Quantile regression with errors in variables," *J. Korean Data Inf. Sci. Soc.*, vol. 25, no. 2, pp. 439–446, Mar. 2014.
[14] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, Jan. 1978.
[15] D. B. Lindley, "Regression lines and the linear functional relationship," *Suppl. J. Roy. Stat. Soc. B*, vol. 9, no. 2, pp. 218–244, 1947.
[16] L. H. Nghiem, M. C. Byrd, and C. J. Potgieter, "Estimation in linear errors-in-variables models with unknown error distribution," *Biometrika*, vol. 107, no. 4, pp. 841–856, Dec. 2020.
[17] L. Baringhaus, "Testing for spherical symmetry of a multivariate distribution," *Ann. Statist.*, vol. 19, no. 2, pp. 899–917, Jun. 1991.
[18] M. Mohammedi, S. Bouzebda, and A. Laksaci, "The consistency and asymptotic normality of the kernel type expectile regression estimator for functional data," *J. Multivariate Anal.*, vol. 181, Jan. 2021, Art. no. 104673.
[19] Y. Ma and G. Yin, "Censored quantile regression with covariate measurement errors," *Stat. Sin.*, vol. 21, no. 2, pp. 949–971, Apr. 2011.
[20] W. K. Newey and J. L. Powell, "Asymmetric least squares estimation and testing," *Econometrica*, vol. 55, no. 4, pp. 819–847, Jul. 1987.
[21] S. K. Schnabel and P. H. C. Eilers, "Optimal expectile smoothing," *Comput. Statist., Data Anal.*, vol. 53, no. 12, pp. 4168–4177, Oct. 2009.
[22] F. Sobotka and T. Kneib, "Geoadditive expectile regression," *Comput. Statist., Data Anal.*, vol. 56, no. 4, pp. 755–767, Apr. 2012.
[23] F. Sobotka, G. Kauermann, L. S. Waltrup, and T. Kneib, "On confidence intervals for semiparametric expectile regression," *Statist. Comput.*, vol. 23, no. 2, pp. 135–148, Mar. 2013.
[24] L. S. Waltrup, F. Sobotka, T. Kneib, and G. Kauermann, "Expectile and quantile regression—David and Goliath?" *Stat. Model.*, vol. 15, no. 5, pp. 433–456, 2015.
[25] L. S. Waltrup and G. Kauermann, "Smooth expectiles for panel data using penalized splines," *Statist. Comput.*, vol. 27, no. 1, pp. 271–282, Jan. 2017.
[26] H. J. Wang, L. A. Stefanski, and Z. Zhu, "Corrected-loss estimation for quantile regression with covariate measurement errors," *Biometrika*, vol. 99, no. 2, pp. 405–421, Jun. 2012.
[27] Y. Wei and R. J. Carroll, "Quantile regression with measurement error," *J. Amer. Stat. Assoc.*, vol. 104, no. 487, pp. 1129–1143, Sep. 2009.
[28] W. Yang and Y. Yang, "Composite quantile regression estimation of linear error-in-variable models using instrumental variables," *Metrika*, vol. 83, no. 1, pp. 1–16, Jan. 2020.
[29] Y. Pan, Z. Liu, and G. Song, "Weighted expectile regression with covariates missing at random," *Commun. Statist.-Simul. Comput.*, vol. 52, no. 3, pp. 1057–1076, Jan. 2021.

XIAOXIA HE received the Ph.D. degree in probability theory and mathematical statistics from Wuhan University, China, in 2007. She is currently an Associate Professor with the Wuhan University of Science and Technology. Her research interest includes statistical modeling in industrial big data.



XIAODAN ZHOU is currently pursing the master's degree with the Wuhan University of Science and Technology. Her research interest includes artificial intelligence computing.

CHUNLI LI received the Ph.D. degree in probability theory and mathematical statistics from Wuhan University, China, in 2015. She is currently a Lecturer with the Wuhan University of Science and Technology. Her research interests include statistical learning and industrial big data analysis.