Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Unconditional Quantile Regression for Streaming Datasets

Rong Jiang & Keming Yu

To cite this article: Rong Jiang & Keming Yu (2024) Unconditional Quantile Regression for Streaming Datasets, Journal of Business & Economic Statistics, 42:4, 1143-1154, DOI: 10.1080/07350015.2023.2293162

To link to this article: https://doi.org/10.1080/07350015.2023.2293162

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 05 Jan 2024.

Submit your article to this journal

Article views: 1767

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ubes20

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2024, VOL. 42, NO. 4, 1143–1154
https://doi.org/10.1080/07350015.2023.2293162

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Unconditional Quantile Regression for Streaming Datasets

Rong Jiang[a] and Keming Yu[b]

[a]Shanghai Polytechnic University, Shanghai, China; [b]Brunel University London, UK

ABSTRACT

The Unconditional Quantile Regression (UQR) method, initially introduced by Firpo et al. has gained significant traction as a popular approach for modeling and analyzing data. However, much like Conditional Quantile Regression (CQR), UQR encounters computational challenges when it comes to obtaining parameter estimates for streaming datasets. This is attributed to the involvement of unknown parameters in the logistic regression loss function used in UQR, which presents obstacles in both computational execution and theoretical development. To address this, we present a novel approach involving smoothing logistic regression estimation. Subsequently, we propose a renewable estimator tailored for UQR with streaming data, relying exclusively on current data and summary statistics derived from historical data. Theoretically, our proposed estimators exhibit equivalent asymptotic properties to the standard version computed directly on the entire dataset, without any additional constraints. Both simulations and real data analysis are conducted to illustrate the finite sample performance of the proposed methods.

ARTICLE HISTORY
Received January 2023
Accepted December 2023

KEYWORDS
Renewable estimation;
Smoothing method;
Streaming datasets;
Unconditional quantile regression

1. Introduction

Quantile regression (QR) models proposed by Koenker and Bassett (1978) are more robust to outliers than the classical mean regression models, and any quantile can be used in any part of the outcome distribution. The most commonly used QR framework is the conditional quantile regression (CQR). It is used to assess the impact of a covariate on a quantile of the outcome conditional on specific values of other covariates (Jiang and Yu 2023). CQR is widely seen as an ideal tool to understand complex predictor-response relations, however, CQR models do not average up to their unconditional population counterparts. As a result, the estimates obtained cannot be used to estimate the impacts of an explanatory variable $\boldsymbol{X}$ on the corresponding unconditional quantile of the outcome variable $\boldsymbol{Y}$. To overcome this restriction, Firpo et al. (2009) proposed a regression of the (recentered) influence function of the unconditional quantile of the $\boldsymbol{Y}$ on the $\boldsymbol{X}$, or UQR estimates the impact of changing the distribution of $\boldsymbol{Y}$ on marginal distribution of $\boldsymbol{X}$. The advantage of the UQR model is that the quantiles are defined preregression; therefore, the model is not influenced by any right-hand-side variables. In UQR, one can, for instance, include fixed effects to adjust for selection bias without redefining the quantiles (Borgen 2016). The UQR method has attracted substantial attention in statistics and econometrics with many applications in different fields. By January 2023, Firpo et al. (2009) has attracted 2500+ Google Scholar citations, such as Ghosh (2021), Inoue, Li, and Xu (2021), Sasaki, Ura, and Zhang (2022), and Martinez-Iriarte, Montes-Rojas, and Sun (2022) and so on.

In spite of its rapidly growing popular method for modeling and analyzing data, however, UQR faces challenges to obtaining parameter estimates from "big data." The concept of "big data" may have different meanings to people from different fields and has since become a dominant topic in nearly all academic disciplines and in applied fields. In a broad sense, big data is data on a larger scale in terms of volume, variety, velocity, variability, and veracity. In this article, we consider one type of big data: streaming data, which grows rapidly in volume and velocity. Due to the explosive growth of data onto nontraditional sources such as mobile phones, social networks and e-commerce, streaming data is becoming a core component of big data analysis.

Streaming data grows rapidly in volume and velocity. Then storing and combing data becomes increasingly challenging. To reduce the demand on computing memory and achieve real-time processing, the nature of streaming data calls for the development of algorithms which require only "one pass" over the data. This means that in order to reduce storage requirements and computation time, data is only used once. Therefore, the primary goal of processing such streaming data is to sequentially update some statistics of interest upon the arrival of a new data batch, in the hope to not only free up space for the storage of massive historical individual-level data, but also provide real-time inference and decision-making. Online updating approaches are distinct from the massive data analysis because they target problem where data arrive in streams or large chunks and address statistical problems in an updating framework without storage requirements for previous data, as

CONTACT Keming Yu keming.yu@brunel.ac.uk Brunel University, London UB83PH, UK.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.
© 2024 The Authors. Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

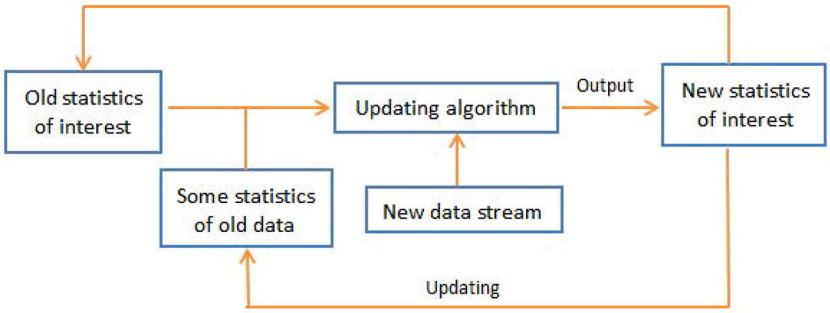

Figure 1. Online-updating algorithm for streaming datasets.

shown in Figure 1. There are three online updating methods for analyzing streaming data in the literature as follows: average updating methods (Schifano et al. 2016), subsampling methods (Xie, Bai, and Ma 2023) and renewable methods based on estimation equations (Luo and Song 2020). The average updating methods developed in Schifano et al. (2016) require the total number of batches smaller than the sample size of each batch to establish the same statistical properties as that of the oracle estimators with the full datasets, see Theorem B.1 in the Appendix, supplementary materials. This means that streaming data cannot be unlimited, which is not very suitable for the practical application of streaming data. The estimators obtained by subsampling-based approaches are $\sqrt{\tilde{n}}$-consistent instead of $\sqrt{n}$-consistent, where $\tilde{n}$ is the sample size of subsample and n is the sample size of all data, see Wang, Min, and Stufken (2019) and Xie, Bai, and Ma (2023). This means that there is information loss in this method. The estimators obtained by the renewable methods based on estimation equations in Luo and Song (2020) can achieve $\sqrt{n}$-consistent and overcome the above unnatural restriction for average updating methods. Other references on the online updating methods can see Deshpande, Javanmard, and Mehrabi (2023), Luo, Zhou, and Song (2023), Yang and Yao (2023) and so on.

Specifically, the difficulties faced in analyzing UQR under streaming data are as follows.

First, it is difficult to perform standard logistic regression based on the loss function (3.4) in Section 3 under streaming data according to the term $\mathrm{I}(Y_i > \tilde{q}_\tau)$. Although the method in Luo and Song (2020) can be used to construct a renewable estimator, due to the term $\mathrm{I}(Y_i > \tilde{q}_\tau)$, the error of the estimator of $\boldsymbol{\beta}_{q_\tau}$ increases as the number of streams increases. To solve the above problem, we adopt a smoothing technique to smooth the above indicative function, which helps to reduce the error from $O_p(|\tilde{q}_\tau - q_\tau|)$ to $O_p(|\tilde{q}_\tau - q_\tau|^2)$, so that the error can be ignored. The smoothing technique are often used in QR, see Horowitz (1998), Chen, Liu, and Zhang (2019), Fernandes, Guerre, and Horta (2021), He et al. (2023) and so on. But, to the best of our limited knowledge, there is no literature on the application of smoothing technique to UQR.

Second, as we all known that the bandwidth h is important for the kernel density estimator (3.6) in Section 3, and it depends on the sample size. For streaming data, we cannot know the total sample size at the beginning, so h needs to change with the arrival of new data. Kong and Xia (2019) developed an online density estimate with a single point of update. In this article, we extend the single point update estimation method to batch update estimation. Moreover, the kernel density estimator (3.6) contains $\tilde{q}_\tau$, thus, we take Taylor expansion to construct a renewable estimator.

Finally, the unconditional quantile partial effect defined in (3.3) involves the quantile of $\boldsymbol{Y}$. The above methods for streaming data based on the least squares or estimating equations are not suitable for the QR because the quantile regression estimator has no display expression like the least squares estimator and the loss function of the quantile regression is not differentiable, even though loss function needs to be second-order differentiable in the estimation equation (Luo and Song 2020). In order to overcome the non-differentiable of the QR loss function, Jiang and Yu (2022) used a convolution-type smoothing method to develop a renewable estimation. Chen, Liu, and Zhang (2019) and Wang, Wang, and Li (2022) also studied QR estimation for streaming data. However, their methods are all required additional strict conditions on the sample size of each bath. In this article, we adjust method of Jiang and Yu (2022) to estimate q_τ in the (3.3).

To summarize, we develop a renewable estimation for UQR. Our statistical contributions include: (i) Note that the loss function (3.4) in Section 3 is different to the standard logistic regression according to the term $\mathrm{I}(Y_i > \tilde{q}_\tau)$. Therefore, the method of Luo and Song (2020) does not work. To solve the above problem, we adopt a smoothing technique to smooth the above indicative function, which helps to produce a renewable estimator. (ii) We develop a renewable kernel density estimator and a renewable QR estimator. (iii) We propose a renewable UQR estimation that only requires the availability of the current data batch in the data stream and sufficient statistics on the historical data at each stage of the analysis. The asymptotic properties of the proposed renewable estimator under the conditions are similar to those in an offline setting and no restrictions on number of batches, which means that the new methods are adaptive to the situation where streaming datasets arrive fast and perpetually.

The remainder of this article is organized as follows. Section 2 presents a motivational example. The review of the standard UQR is given in Section 3. In Section 4, the streaming datasets analysis method is proposed. Both simulation studies and empirical applications are given in Sections 5 and 6 to illustrate the proposed procedures. We conclude the article with a brief discussion in Section 7. All technical proofs are deferred to the Appendix, supplementary materials.

2. Motivating Example

We exemplify the application of streaming data in economics using the following stock price and exchange rate data.

As the process of financial globalization continues to deepen, the financial markets of various countries become increasingly interconnected, underscoring the pivotal role of the exchange rate system in the capital market. The exchange rate represents the international price of a country's currency. Changes in the exchange rate signify alterations in the international purchasing power of the currency, making it a crucial policy tool for maintaining national economic security and ensuring financial stability. Stock prices act as a "barometer" of macroeconomics, offering timely reflections of microeconomic changes. The fluctuation in exchange rates not only impacts the macroeconomic operations of a country but also influences the behavior of microeconomic entities, subsequently affecting the stock prices of companies. A comprehensive understanding of the relationship between exchange rates and stock indexes is instrumental for countries and international organizations to manage their exposure to foreign exchange risks. Moreover, it proves invaluable for investors seeking to hedge or predict returns on their foreign investments.

The relationship between the foreign exchange market and the stock market has garnered significant attention from scholars. Bahmani-Oskooee and Sohrabian (1992) used Granger causality tests and co-integration methods to investigate the connection between the foreign exchange market and the stock market in the United States. The research indicates the presence of a short-term two-way causal relationship. Pan, Fok, and Liu (2007) delved into East Asia using data from January 1988 to October 1998, discovering that, before the Asian financial crisis in 1997, there was a causal relationship from the exchange rate to the stock market in Hong Kong, Japan, Malaysia, and Thailand. Additionally, there was a causal relationship from the stock market to the exchange rate in Hong Kong, South Korea, and Singapore. Amba and Nguyen (2019) examined the relationship between stock prices and exchange rates in the Mexican and Canadian markets, employing weekly data from January 2013 to December 2018. The Granger causality test affirmed the existence of a short-term one-way causal relationship between exchange rates and stock prices in the Mexican market.

In this section, we will provide a detailed introduction to the Chinese A-share stock market and explore the relationship between daily stock returns and exchange rates. China's A-share stock market, encompassing the Shenzhen Stock Exchange and Shanghai Stock Exchange, was officially established in 1990. The trading data within the stock industry is substantial, reaching the gigabyte level. As of March 31, 2023, the A-share market in China comprises 4495 stocks, with 1824 listed on the Shanghai Stock Exchange and 2671 on the Shenzhen Stock Exchange. A study by Zhang and Li (2010) investigated the correlation between exchange rate changes and the stock market in China post the reform of the exchange rate system and the split structure of the stock market in 2005. The study revealed a long-term co-integration relationship between the exchange rate and the stock market. The results indicate that, in the long run, the relationship between exchange rate changes and the stock market primarily follows the flow-oriented model, with the Shanghai A-share index being more noticeably affected by exchange rates. Both stock transaction and exchange rate datasets undergo real-time changes with each transaction, adhering to the typical characteristics of streaming data: (a) the data object is real-time and online; (b) the scale of the data is extensive and theoretically limitless, making it optimal to read each data object only once, thereby reducing data storage requirements. Simultaneously, both institutional and individual investors seek real-time insights into the stock market. Consequently, the analysis of stock trading data should offer real-time analysis functions. Fulfilling these requirements is unattainable with traditional data analysis techniques.

For instance, when examining the relationship between the daily returns in the Shanghai Stock Exchange in China and the exchange rate between the Chinese currency and the US dollar, we can collect approximately 1600 data points per day (excluding suspended trading and stocks under special treatment). Considering the 871 trading days between January 1, 2020, and August 25, 2023, the total data volume amounts to 1,339,591. If we break down the data per minute, the volume becomes $1339591 \times 4 \times 60 = 321,501,840$ (accounting for the 4 hr of daily trading). As established in Section 6.2, analyzing 1,339,591 data points takes 65.37 sec, indicating that processing minute-level data, totaling 321,501,840, would certainly exceed a minute, which is deemed unacceptable. However, employing stream data analysis techniques, specifically update estimation methods, allows for the processing of the last batch of incoming data and some statistics from past data, resulting in an analysis time of just 0.02 sec. Whether considering daily or minute data, the volume of the last batch typically hovers around 1600, making this approach highly efficient.

3. Standard Unconditional Quantile Regression

In this section, we first review the standard unconditional quantile regression with full data (assuming that streaming data can be pooled into a dataset and can be analyzed and stored by a computer). Consider a general structural model:

$$\boldsymbol{Y} = g(\boldsymbol{X}, \boldsymbol{\varepsilon}), \tag{3.1}$$

where the unknown mapping $g(\cdot,\cdot)$ is invertible on the second argument, $\boldsymbol{\varepsilon}$ is an unobservable determinant of the outcome variable $\boldsymbol{Y}$ and $\boldsymbol{X}$ is a p-dimensional covariates.

According to the definition in Firpo et al. (2009), we use another name unconditional quantile partial effect (**UQPE**) for UQR. The **UQPE** at quantile level τ proposed by Firpo et al. (2009) is defined as

$$\begin{aligned}\mathbf{UQPE}_\tau &= \int \frac{d\mathrm{E}\{\mathbf{RIF}(\boldsymbol{Y}, q_\tau)|\boldsymbol{X}=\boldsymbol{x}\}}{d\boldsymbol{x}} dF_{\boldsymbol{X}}(\boldsymbol{x}) \\ &= \frac{1}{f_{\boldsymbol{Y}}(q_\tau)} \int \frac{d\mathrm{E}\{\mathrm{I}(\boldsymbol{Y} > q_\tau)|\boldsymbol{X}=\boldsymbol{x}\}}{d\boldsymbol{x}} dF_{\boldsymbol{X}}(\boldsymbol{x}),\end{aligned}$$

where $F_{\boldsymbol{X}}(\cdot)$ is the distribution function of $\boldsymbol{X}$, $\mathbf{RIF}(y, q_\tau) = \mathrm{I}(y > q_\tau)/f_{\boldsymbol{Y}}(q_\tau) + q_\tau - (1-\tau)/f_{\boldsymbol{Y}}(q_\tau)$ is the recentered influence function, $\mathrm{I}(\cdot)$ is the indicator function, q_τ and $f_{\boldsymbol{Y}}(\cdot)$ are the τth quantile and the density function of $\boldsymbol{Y}$, respectively. Assume that

$$\mathrm{P}\left(\boldsymbol{Y} > q_\tau|\boldsymbol{X}=\boldsymbol{x}\right) = \Lambda(\boldsymbol{x}^\top \boldsymbol{\beta}_{q_\tau}), \tag{3.2}$$

where $\boldsymbol{\beta}_{q_\tau} = \arg\max_{\boldsymbol{\beta}} \mathrm{E}\left[\mathrm{I}(\boldsymbol{Y} > q_\tau) \cdot \boldsymbol{X}^\top \boldsymbol{\beta} + \log\{1 - \Lambda(\boldsymbol{X}^\top \boldsymbol{\beta})\}\right]$ and $\Lambda(r) = 1/(1 + e^{-r})$ is the logistic distribution function. Then, by the assumption (3.2), $\mathbf{UQPE}_\tau$ is equal to

$$\mathbf{UQPE}_\tau = \frac{\boldsymbol{\beta}_{q_\tau}}{f_{\boldsymbol{Y}}(q_\tau)} \int \Lambda'(\boldsymbol{x}^\top \boldsymbol{\beta}_{q_\tau}) dF_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.3}$$

where $\Lambda'(r) = \Lambda(r)\{1 - \Lambda(r)\}$ is the derivative of $\Lambda(r)$. Note that the assumption (3.2) is the assumption 11 in Firpo et al. (2009) and (3.3) is the RIF-Logit in Firpo et al. (2009).

We first review the standard estimation method of $\mathbf{UQPE}_\tau$ in Firpo et al. (2009). Let $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ be an iid sample from $(\boldsymbol{X}, \boldsymbol{Y})$ in model (3.1). Based on the assumption (3.2), the estimator of $\boldsymbol{\beta}_{q_\tau}$ based on $\tilde{q}_\tau$ is

$$\tilde{\boldsymbol{\beta}}_{\tilde{q}_\tau} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^n \left[\mathrm{I}(Y_i > \tilde{q}_\tau) \cdot \boldsymbol{X}_i^\top \boldsymbol{\beta} + \log\{1 - \Lambda(\boldsymbol{X}_i^\top \boldsymbol{\beta})\}\right], \tag{3.4}$$

where $\tilde{q}_\tau$ is the estimator of q_τ as

$$\tilde{q}_\tau = \arg\min_q \sum_{i=1}^n \rho_\tau\left(Y_i - q\right), \tag{3.5}$$

where $\rho_\tau(r) = \tau r - r\mathrm{I}(r < 0)$ is the check loss function. Moreover, the kernel density estimator for the density of $\boldsymbol{Y}$ at $\tilde{q}_\tau$ is

$$\tilde{f}_{\boldsymbol{Y}}(\tilde{q}_\tau) = \frac{1}{n} \sum_{i=1}^n K_h\left(Y_i - \tilde{q}_\tau\right), \tag{3.6}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a smooth kernel function and h is a bandwidth.

Then, the estimator of $\mathbf{UQPE}_\tau$ based on (3.3)–(3.6) is

$$\widetilde{\mathbf{UQPE}}_\tau = \frac{\tilde{\boldsymbol{\beta}}_{\tilde{q}_\tau}}{\tilde{f}_{\boldsymbol{Y}}(\tilde{q}_\tau)} \frac{1}{n} \sum_{i=1}^n \Lambda'(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}_{\tilde{q}_\tau}). \tag{3.7}$$

4. Streaming Datasets Analysis

Now let us discuss how to develop a renewable estimator for **UQPE** based on streaming datasets. Assume we have the streaming datasets $\{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_b\}$ up to the bth batch, where $\boldsymbol{D}_j = \{(\boldsymbol{X}_{i,j}, Y_{i,j}), i = 1, \ldots, N_j\}$ is the jth batch dataset with a sample size of N_j. We suppose that the $(\boldsymbol{X}_{i,j}, Y_{i,j})$ for all is and js are iid samples from $(\boldsymbol{X}, \boldsymbol{Y})$ in model (3.1). The sample size up to the bth batch is $\bar{N}_b = \sum_{j=1}^b N_j$.

The key idea of the following renewable estimation is to use the Taylor expansion of the score function so that the new estimation equation uses only the combined information of the previous data and the data of the current batch. In the non-differentiable case, smoothing techniques will be used to enable Taylor expansion of the scoring function.

4.1. Estimate q_τ for Streaming Datasets

Note that for a quantile regression, the loss function $\rho_\tau(r) = \tau r - r\mathrm{I}(r < 0)$ is non-differentiable. Therefore, the QR estimator has no display expression, so it is impossible to construct a renewable estimator for streaming data. To circumvent the non-differentiable of the QR loss function, we smooth quantile regression loss function $\rho_\tau(r)$ to a twice continuously differentiable function (Nadaraya 1964; Fernandes, Guerre, and Horta 2021):

$$\mathbf{Q}_h(r) = \int \rho_\tau(t) K_h(t - r) dt. \tag{4.1}$$

For example, we take Logistic kernel $K(u) = e^{-u}/(1 + e^{-u})^2$ in (4.1), the explicit expression of $\mathbf{Q}_h(r)$ is $\tau r + h \log(1 + e^{-r/h})$. By (4.1), the $\hat{q}_\tau^1$ based on $\boldsymbol{D}_1$ satisfies,

$$\sum_{i \in \boldsymbol{D}_1} \left\{\tilde{K}((\hat{q}_\tau^1 - Y_i)/h_{q,1}) - \tau\right\} = 0, \tag{4.2}$$

which is the derivative of $\mathbf{Q}_h(\cdot)$ on dataset $\boldsymbol{D}_1$, and where $\tilde{K}(r) = \int_{-\infty}^r K(u) du$ and $h_{q,j}$ is a bandwidth for jth batch. We propose a new estimator $\hat{q}_\tau^2$ for streaming data $\{\boldsymbol{D}_1, \boldsymbol{D}_2\}$ as a solution to the equation of the form

$$\begin{aligned} &\sum_{i \in \boldsymbol{D}_1} K_{h_{q,1}}(\hat{q}_\tau^1 - Y_i)(\hat{q}_\tau^2 - \hat{q}_\tau^1) \\ &+ \sum_{i \in \boldsymbol{D}_2} \left\{\tilde{K}((\hat{q}_\tau^2 - Y_i)/h_{q,2}) - \tau\right\} = \mathbf{0}, \end{aligned} \tag{4.3}$$

which is according to

$$\begin{aligned} &\sum_{i \in \boldsymbol{D}_1} \left\{\tilde{K}((\hat{q}_\tau^2 - Y_i)/h_{q,1}) - \tau\right\} \\ &= \sum_{i \in \boldsymbol{D}_1} \left\{\tilde{K}((\hat{q}_\tau^1 - Y_i)/h_{q,1}) - \tau\right\} \\ &+ \sum_{i \in \boldsymbol{D}_1} K_{h_{q,1}}(\hat{q}_\tau^1 - Y_i)(\hat{q}_\tau^2 - \hat{q}_\tau^1) + O_p(N_1 |\hat{q}_\tau^2 - \hat{q}_\tau^1|^2) \\ &= \sum_{i \in \boldsymbol{D}_1} K_{h_{q,1}}(\hat{q}_\tau^1 - Y_i)(\hat{q}_\tau^2 - \hat{q}_\tau^1) + O_p(N_1 |\hat{q}_\tau^2 - \hat{q}_\tau^1|^2), \end{aligned}$$

where the last equation is according to (4.2), $O_p(\cdot)$ means bounded with probability and the error term $O_p(N_1 |\hat{q}_\tau^2 - \hat{q}_\tau^1|^2)$ is asymptotically ignored.

Generalizing (4.3) to streaming datasets $\{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_b\}$, a renewable estimator $\hat{q}_\tau^b$ of q_τ is defined as a solution to the following incremental estimation equation:

$$\hat{\boldsymbol{K}}^{b-1}(\hat{q}_\tau^b - \hat{q}_\tau^{b-1}) + \sum_{i \in \boldsymbol{D}_b} \left\{\tilde{K}\left(\frac{\hat{q}_\tau^b - Y_i}{h_{q,b}}\right) - \tau\right\} = \mathbf{0}, \tag{4.4}$$

where $\hat{\boldsymbol{K}}^{b-1} = \sum_{j=1}^{b-1} \sum_{i \in \boldsymbol{D}_j} K_{h_{q,j}}(\hat{q}_\tau^j - Y_i)$. The asymptotic property of $\hat{q}_\tau^b$ can see Lemma 2 in the Appendix, supplementary materials. Numerically, it is quite straightforward to find $\hat{q}_\tau^b$ from (4.4) using the Newton-Raphson method as

$$\begin{aligned} \hat{q}_\tau^b = \hat{q}_\tau^{b-1} - &\left\{\hat{\boldsymbol{K}}^{b-1} + \sum_{i \in \boldsymbol{D}_b} K_{h_{q,j}}(\hat{q}_\tau^{b-1} - Y_i)\right\}^{-1} \\ &\sum_{i \in \boldsymbol{D}_b} \left\{\tilde{K}\left(\frac{\hat{q}_\tau^{b-1} - Y_i}{h_{q,b}}\right) - \tau\right\}. \end{aligned} \tag{4.5}$$

Only one iteration in (4.5) is due to that $\hat{q}_\tau^b$ is $\sqrt{\bar{N}_b}$-consistent by Lemma 2 in the Appendix, supplementary materials and condition $N_j = O(N_1)$ for $j = 1, \ldots, b$. Therefore, the estimators $\hat{q}_\tau^b$ can converge in one iteration.

4.2. Estimate $f_Y(q_\tau)$ for Streaming Datasets

It is easy to estimate $f_{\boldsymbol{Y}}(q_\tau)$ based on the first batch as $\hat{f}_{\boldsymbol{Y}}^1 = {N_1}^{-1}\sum_{i\in\boldsymbol{D}_1} K_{h_{f,1}}(Y_i - \hat{q}_\tau^1)$, where $h_{f,j}$ is a bandwidth for the jth batch. When the next data $\boldsymbol{D}_2$ arrive, the estimator of $f_{\boldsymbol{Y}}(q_\tau)$ should be

$$\frac{1}{\bar{N}_2}\left\{\sum_{i\in\boldsymbol{D}_1} K_{h_{f,2}}(Y_i-\hat{q}_\tau^2)+\sum_{i\in\boldsymbol{D}_2} K_{h_{f,2}}(Y_i-\hat{q}_\tau^2)\right\}.$$

As we all known that we can not know the sample size of next batch, thus, it is difficult to use $h_{f,2}$ in $\boldsymbol{D}_1$ because of $h_{f,2}$ always depending on $\bar{N}_2$. We will prove in Theorem 4.1 that the following estimator is also effective

$$\frac{1}{\bar{N}_2}\left\{\sum_{i\in\boldsymbol{D}_1} K_{h_{f,1}}(Y_i-\hat{q}_\tau^2)+\sum_{i\in\boldsymbol{D}_2} K_{h_{f,2}}(Y_i-\hat{q}_\tau^2)\right\}.$$

Moreover, we use Taylor expansion to $\hat{q}_\tau^2$ in $\boldsymbol{D}_1$ as

$$\begin{aligned}&\sum_{i\in\boldsymbol{D}_1} K_{h_{f,1}}(Y_i-\hat{q}_\tau^2)=\sum_{i\in\boldsymbol{D}_1} K_{h_{f,1}}(Y_i-\hat{q}_\tau^1)\\&+\sum_{i\in\boldsymbol{D}_1} h_{f,1}^{-1}K'_{h_{f,1}}(Y_i-\hat{q}_\tau^1)(\hat{q}_\tau^1-\hat{q}_\tau^2)+O_p(N_1h_{f,1}^{-2}|\hat{q}_\tau^1-\hat{q}_\tau^2|^2),\end{aligned}$$

where $K'(\cdot)$ is the derivative of $K(\cdot)$ and $K'_h(\cdot) = K'(\cdot/h)/h$. Then, we can obtain the estimator of $f_{\boldsymbol{Y}}(q_\tau)$ based on $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as

$$\begin{aligned}\hat{f}_{\boldsymbol{Y}}^2=&\frac{1}{\bar{N}_2}\sum_{j=1}^{2}\sum_{i\in\boldsymbol{D}_j} K_{h_{f,j}}(Y_i-\hat{q}_\tau^j)+\frac{1}{\bar{N}_2}\\&\sum_{i\in\boldsymbol{D}_1} h_{f,1}^{-1}K'_{h_{f,1}}(Y_i-\hat{q}_\tau^1)(\hat{q}_\tau^1-\hat{q}_\tau^2).\end{aligned}$$

Generalizing the above method to streaming datasets $\{\boldsymbol{D}_1,\ldots,\boldsymbol{D}_b\}$, a renewable estimator $\hat{f}_{\boldsymbol{Y}}^b$ of $f_{\boldsymbol{Y}}(q_\tau)$ is defined as

$$\begin{aligned}\hat{f}_{\boldsymbol{Y}}^b=&\frac{1}{\bar{N}_b}\sum_{j=1}^{b}\sum_{i\in\boldsymbol{D}_j} K_{h_{f,j}}(Y_i-\hat{q}_\tau^j)\\&+\frac{1}{\bar{N}_b}\sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j} h_{f,j}^{-1}K'_{h_{f,j}}(Y_i-\hat{q}_\tau^j)(\hat{q}_\tau^j-\hat{q}_\tau^b)\qquad(4.6)\\=&\frac{1}{\bar{N}_b}\left\{\boldsymbol{K}_1^{b-1}+\boldsymbol{K}_2^{b-1}-\hat{q}_\tau^b\boldsymbol{K}_3^{b-1}+\sum_{i\in\boldsymbol{D}_b} K_{h_{f,b}}(Y_i-\hat{q}_\tau^b)\right\},\end{aligned}$$

where $\boldsymbol{K}_1^{b-1} = \sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j} K_{h_{f,j}}(Y_i-\hat{q}_\tau^j)$, $\boldsymbol{K}_2^{b-1} = \sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j} h_{f,j}^{-1}K'_{h_{f,j}}(Y_i-\hat{q}_\tau^j)\hat{q}_\tau^j$ and $\boldsymbol{K}_3^{b-1} = \sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j} h_{f,j}^{-1}K'_{h_{f,j}}(Y_i-\hat{q}_\tau^j)$.

The item $\bar{N}_b^{-1}\sum_{j=1}^{b}\sum_{i\in\boldsymbol{D}_j} K_{h_{f,j}}(Y_i-\hat{q}_\tau^j)$ in (4.6) is to extend the single point update estimation method in Kong and Xia (2019) to batch update estimation, and the term $\bar{N}_b^{-1}\sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j} h_{f,j}^{-1}K'_{h_{f,j}}(Y_i-\hat{q}_\tau^j)(\hat{q}_\tau^j-\hat{q}_\tau^b)$ makes approximate to the density estimation of point $\hat{q}_\tau^b$. To reveal the merits of the proposed method, we now establish the asymptotic normality of $\hat{f}_{\boldsymbol{Y}}^b$.

To establish the asymptotic properties of the proposed estimator, the following technical conditions are imposed.

C1. Density function $f_{\boldsymbol{Y}}(\cdot)$ is positive and has second-order derivative, whose second-order derivative is bounded and continuous in a neighborhood of a grid of selected points $q_\tau \in \Re$. Moreover, $\int |f_{\boldsymbol{Y}}(y)|dy < \infty$.

C2. The kernel function $K(\cdot)$ is even, integrable, and twice differentiable with bounded first and second derivatives such that $\int K(u)du = 1$, $\int |u^2K(u)|du < \infty$, $\int uK(u)du = 0$ and $\int u^2K(u)du \neq 0$.

Remark 4.1. Conditions **C1** and **C2** are Assumptions 2, 3, 6, and 7 in Firpo et al. (2009). The Logistic kernel $K(u) = e^{-u}/(1+e^{-u})^2$ satisfies condition **C2**.

Theorem 4.1. Assume that conditions **C1** and **C2** hold. If $N_j = O(N_1)$ and $N_1 \to \infty$, $h_{f,j} = O(\bar{N}_j^{-c_1})$ with $0 < c_1 < 1/4$ and $\bar{N}_j = \sum_{i=1}^{j} N_i$ for $j = 1,\ldots,b$, where b can be a fixed number or a divergent number, we have

$$\sqrt{\bar{N}_b h_{\bar{N}_b}}\left(\hat{f}_{\boldsymbol{Y}}^b - f_{\boldsymbol{Y}}(q_\tau) - \frac{1}{2}f''_{\boldsymbol{Y}}(q_\tau)\mu_K\sum_{j=1}^{b}\frac{N_j}{\bar{N}_b}h_{f,j}^2\right)\xrightarrow{L}$$
$$\mathrm{N}\left(\boldsymbol{0}, f_{\boldsymbol{Y}}(q_\tau)\nu_K\right),$$

where $h_{\bar{N}_b} = \bar{N}_b/\sum_{j=1}^{b} N_j h_{f,j}^{-1}$, $\mu_K = \int u^2K(u)du$, $\nu_K = \int K^2(u)du$ and $a_n = O(b_n)$ means $\sup_n |a_n/b_n| < c < \infty$ with a positive and bounded constant c.

If $1/5 < c_1 < 1/4$, we have

$$\sqrt{\bar{N}_b h_{\bar{N}_b}}\left(\hat{f}_{\boldsymbol{Y}}^b - f_{\boldsymbol{Y}}(q_\tau)\right)\xrightarrow{L}\mathrm{N}\left(\boldsymbol{0}, f_{\boldsymbol{Y}}(q_\tau)\nu_K\right).$$

Note that $h_{\bar{N}_b} = O(\bar{N}_b^{-c_1}) = O(h_{f,b})$ by Lemma 1 in the Appendix, supplementary materials, we obtain the same convergence rate and asymptotic variance as the full data estimator (3.6).

4.3. Estimate β_{q_τ} for Streaming Datasets

Note that (3.2), the estimate $\boldsymbol{\beta}_{q_\tau}$ contains q_τ. Thus, it is difficult to construct the estimator of $\boldsymbol{\beta}_{q_\tau}$ according to indicative function $\mathrm{I}(Y > q_\tau)$ with unknown parameter q_τ. Therefore, we approximate the indicator factor $\mathrm{I}(Y > q_\tau)$ in the score equation with a smooth function $\mathrm{H}\left((Y-q_\tau)/h\right)$ and h is the bandwidth. For the first streaming data $\boldsymbol{D}_1$, the smoothing logistic regression estimator $\hat{\boldsymbol{\beta}}_1$ satisfies,

$$\sum_{i\in\boldsymbol{D}_1}\boldsymbol{X}_i\left\{\Lambda(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_1)-\mathrm{H}\left(\frac{Y_i-\hat{q}_\tau^1}{h_1}\right)\right\}=\boldsymbol{0},\qquad(4.7)$$

where h_j is a bandwidth for the jth batch and $\hat{q}_\tau^j$ is the up to jth batch estimator of q_τ in Section 4.1. In the Lemma 3 in the Appendix, supplementary materials, we prove that $\hat{\boldsymbol{\beta}}_1$ achieves optimal efficiency and its asymptotic covariance matrix is the same as that of estimator in (3.4) by ordinary logistic regression estimator.

Then, $\hat{\boldsymbol{\beta}}_2^*$ for streaming data $\{\boldsymbol{D}_1,\boldsymbol{D}_2\}$ satisfies the following aggregated score equation:

$$\boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_1)+\boldsymbol{S}(\boldsymbol{D}_2;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_2)=\boldsymbol{0},\qquad(4.8)$$

where $\boldsymbol{S}(\boldsymbol{D}_j;q;\boldsymbol{\beta};h) = \sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_i\{\Lambda(\boldsymbol{X}_i^\top\boldsymbol{\beta}) - \mathrm{H}((Y_i-q)/h)\}$. Note that $\boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_1)$ in (4.8) should be $\boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_2)$, we can prove that the difference between h_1 and h_2 in (4.8) can be ignored, see the proof of Theorem 4.2 in the Appendix, supplementary materials.

Solving (4.8) for $\hat{\boldsymbol{\beta}}_2^*$ actually involves the use of subject-level data in both $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, where $\boldsymbol{D}_1$ may no longer to accessible. Our renewable estimation is able to handle this issue. To proceed, we take the first-order Taylor series expansion of the $\boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_1)$ around $\hat{q}_\tau^1$ and $\hat{\boldsymbol{\beta}}_1$ as

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2^*;h_1) =& \boldsymbol{S}(\boldsymbol{D}_1;\hat{q}_\tau^1;\hat{\boldsymbol{\beta}}_1;h_1) + \boldsymbol{S}_q'(\boldsymbol{D}_1;\hat{q}_\tau^1;h_1)(\hat{q}_\tau^2-\hat{q}_\tau^1)\\ &+ \boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_1;\hat{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_2^*-\hat{\boldsymbol{\beta}}_1)\\ &+ N_1O_p(|\hat{q}_\tau^2-\hat{q}_\tau^1|^2+\|\hat{\boldsymbol{\beta}}_2^*-\hat{\boldsymbol{\beta}}_1\|_2^2)\\ =& \boldsymbol{S}_q'(\boldsymbol{D}_1;\hat{q}_\tau^1;h_1)(\hat{q}_\tau^2-\hat{q}_\tau^1)\\ &+ \boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_1;\hat{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_2^*-\hat{\boldsymbol{\beta}}_1)\\ &+ N_1O_p(|\hat{q}_\tau^2-\hat{q}_\tau^1|^2+\|\hat{\boldsymbol{\beta}}_2^*-\hat{\boldsymbol{\beta}}_1\|_2^2),\end{aligned} \tag{4.9}$$

where the last equation is according to (4.7), $\boldsymbol{S}_q'(\boldsymbol{D}_j;q;h) = \sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_i\mathrm{H}_h'((Y_i-q)/h)$, $\boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_j;\boldsymbol{\beta}) = \sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_i\boldsymbol{X}_i^\top\Lambda'(\boldsymbol{X}_i^\top\boldsymbol{\beta})$, $\mathrm{H}'(\cdot)$ is the derivative of $\mathrm{H}(\cdot)$ and $\mathrm{H}_h'(\cdot) = H'(\cdot/h)/h$.

By (4.8) and (4.9), and removing the asymptotically ignored term $N_1O_p(|\hat{q}_\tau^2-\hat{q}_\tau^1|^2+\|\hat{\boldsymbol{\beta}}_2^*-\hat{\boldsymbol{\beta}}_1\|_2^2)$, we propose a new estimator $\hat{\boldsymbol{\beta}}_2$ as a solution to the equation of the form

$$\begin{aligned}&\boldsymbol{S}_q'(\boldsymbol{D}_1;\hat{q}_\tau^1;h_1)(\hat{q}_\tau^2-\hat{q}_\tau^1) + \boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_1;\hat{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_2-\hat{\boldsymbol{\beta}}_1)\\ &\quad+\boldsymbol{S}(\boldsymbol{D}_2;\hat{q}_\tau^2;\hat{\boldsymbol{\beta}}_2;h_2)=\boldsymbol{0}.\end{aligned} \tag{4.10}$$

Generalizing the (4.10) to streaming datasets $\{\boldsymbol{D}_1,\ldots,\boldsymbol{D}_b\}$, a renewable estimator $\hat{\boldsymbol{\beta}}_b$ of $\boldsymbol{\beta}_{q_\tau}$ is defined as a solution to the following incremental estimation equation:

$$\hat{\boldsymbol{S}}_q^{b-1}(\hat{q}_\tau^b-\hat{q}_\tau^{b-1}) + \hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^{b-1}(\hat{\boldsymbol{\beta}}_b-\hat{\boldsymbol{\beta}}_{b-1}) + \boldsymbol{S}(\boldsymbol{D}_b;\hat{q}_\tau^b;\hat{\boldsymbol{\beta}}_b;h_b)=\boldsymbol{0}, \tag{4.11}$$

where $\hat{\boldsymbol{S}}_q^{b-1} = \sum_{j=1}^{b-1}\boldsymbol{S}_q'(\boldsymbol{D}_j;\hat{q}_\tau^j,h_j)$ and $\hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^{b-1} = \sum_{j=1}^{b-1}\boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_j;\hat{\boldsymbol{\beta}}_j)$. Through (4.11), the initial $\hat{\boldsymbol{\beta}}_{b-1}$ is renewed by $\hat{\boldsymbol{\beta}}_b$ only using the historical summary statistics, including sample variance matrices $\{\hat{\boldsymbol{S}}_q^{b-1},\hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^{b-1}\}$ and estimators $\{\hat{q}_\tau^b,\hat{q}_\tau^{b-1},\hat{\boldsymbol{\beta}}_{b-1}\}$ instead of the subject-level raw datasets $\{\boldsymbol{D}_1,\ldots,\boldsymbol{D}_{b-1}\}$. Numerically, it is quite straightforward to find $\hat{\boldsymbol{\beta}}_b$ from (4.11) using the Newton-Raphson method as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_b =& \hat{\boldsymbol{\beta}}_{b-1} - \left\{\hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^{b-1} + \boldsymbol{S}_{\boldsymbol{\beta}}'(\boldsymbol{D}_b;\hat{\boldsymbol{\beta}}_{b-1})\right\}^{-1}\\ &\left\{\hat{\boldsymbol{S}}_q^{b-1}(\hat{q}_\tau^b-\hat{q}_\tau^{b-1}) + \boldsymbol{S}(\boldsymbol{D}_b;\hat{q}_\tau^b;\hat{\boldsymbol{\beta}}_{b-1};h_b)\right\},\end{aligned} \tag{4.12}$$

where only one iteration in (4.12) is due to that $\hat{\boldsymbol{\beta}}_b$ is $\sqrt{\bar{N}_b}$-consistent and condition $N_j = O(N_1)$ for $j = 1,\ldots,b$ in the following Theorem 4.2. Therefore, the estimators $\hat{\boldsymbol{\beta}}_b$ can converge in one iteration.

To establish the asymptotic property of the proposed estimator $\hat{\boldsymbol{\beta}}_b$, the following technical conditions are imposed.

C3. Conditional density function $f_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot)$ is bounded away from zero and Lipschitz continuous in a neighborhood of a grid of selected points $q_\tau \in \Re$.

C4. $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{E}\left\{\boldsymbol{X}\boldsymbol{X}^\top\Lambda'(\boldsymbol{X}^\top\boldsymbol{\beta}_{q_\tau})\right\}$ is a positive definite matrix.

C5. The smoothing function $\mathrm{H}(\cdot)$ is twice differentiable and its second derivative is bounded. Moreover, (i) $\mathrm{H}(u) = 1$ if $u > 1$ and $\mathrm{H}(u) = 0$ if $u < -1$, $\int \mathrm{H}'(u)du = 1$, $\int u\mathrm{H}'(u)du = 0$ and $\int u^2\mathrm{H}'(u)du < \infty$. (ii) $\int \mathrm{H}''(u)du = 0$, $\int u\mathrm{H}''(u)du < \infty$ and $\int\{\mathrm{H}''(u)\}^2du < \infty$, where $\mathrm{H}''(\cdot)$ is the second derivative of $\mathrm{H}(\cdot)$.

Remark 4.2. Condition **C3** is a smoothing condition of the conditional density function $f_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot)$, which is a standard condition for smoothing method, see Jiang and Yu (2021). The condition **C4** ensures that $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ exists. Condition **C5** is a mild condition on $\mathrm{H}(\cdot)$ for smoothing approximation. For example, a biweight kernel $\mathrm{H}(u) = \{1/2 + 15/16(u - 2/3u^3 + 1/5u^5)\}\mathrm{I}(|u| \le 1) + \mathrm{I}(u > 1)$ satisfies condition **C5**.

Theorem 4.2. Assume that conditions **C1**–**C5** hold. If $N_j = O(N_1)$ and $N_1 \to \infty$, $h_j = O(\bar{N}_j^{-c_2})$ with $1/4 < c_2 < 1/3$, for $j = 1,\ldots,b$, where b can be a fixed number or a divergent number, we have

$$\sqrt{\bar{N}_b}\left(\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_{q_\tau}\right) \xrightarrow{L} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{\beta}}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})^\top\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{\beta}} = E(\boldsymbol{\psi}_\beta^2\boldsymbol{X}\boldsymbol{X}^\top) + \tau(1-\tau)E(\boldsymbol{X}|\boldsymbol{Y}=q_\tau)E(\boldsymbol{X}|\boldsymbol{Y}=q_\tau)^\top - 2E(\boldsymbol{\psi}_\beta^2\boldsymbol{X})E(\boldsymbol{X}|\boldsymbol{Y}=q_\tau)^\top$ and $\boldsymbol{\psi}_{\boldsymbol{\beta}} = \boldsymbol{X}\left\{\mathrm{I}(\boldsymbol{Y} > q_\tau) - \Lambda(\boldsymbol{X}^\top\boldsymbol{\beta}_{q_\tau})\right\}$.

Through the result of Theorem 4.2, it is interesting to notice that the renewable estimator $\hat{\boldsymbol{\beta}}_b$ achieves optimal efficiency and its asymptotic covariance matrix is the same as that of estimator $\tilde{\boldsymbol{\beta}}_{\tilde{q}_\tau}$ in (3.4) which is computed directly on all the samples, as shown in Firpo et al. (2009). This implies that the proposed renewable estimator achieves the same asymptotic distribution as $\tilde{\boldsymbol{\beta}}_{\tilde{q}_\tau}$.

4.4. Estimate $\mathbf{UQPE}_\tau$ for Streaming Datasets

Finally, we estimate $\mathbf{UQPE}_\tau$ based on the streaming datasets $\{\boldsymbol{D}_1,\ldots,\boldsymbol{D}_b\}$ by (4.5), (4.6), (4.12) and Taylor series expansion of the $\Lambda'(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_b)$ as

$$\begin{aligned}\widehat{\mathbf{UQPE}}_\tau^{*b} =& \frac{\hat{\boldsymbol{\beta}}_b}{\hat{f}_{\boldsymbol{Y}}}\frac{1}{\bar{N}_b}\sum_{j=1}^{b}\sum_{i\in\boldsymbol{D}_j}\Lambda'(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_b) = \frac{\hat{\boldsymbol{\delta}}_\tau^b}{\hat{f}_{\boldsymbol{Y}}}\\ &+ O_p\left(\frac{1}{\bar{N}_b}\sum_{j=1}^{b-1}N_j\|\hat{\boldsymbol{\beta}}_b-\hat{\boldsymbol{\beta}}_j\|_2^2\right),\end{aligned}$$

where $\hat{\boldsymbol{\delta}}_\tau^b = \hat{\boldsymbol{\beta}}_b\bar{N}_b^{-1}\left\{B_b + \hat{\boldsymbol{B}}_{b-1}^\top\hat{\boldsymbol{\beta}}_b - \tilde{B}_{b-1}\right\}$, $\tilde{B}_{b-1} = \sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j}\Lambda''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j$, $\hat{\boldsymbol{B}}_{b-1} = \sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j}\Lambda''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\boldsymbol{X}_i$, $B_b = \sum_{j=1}^{b}\sum_{i\in\boldsymbol{D}_j}\Lambda'(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)$ and $\Lambda''(r) = \Lambda'(r)\{1-2\Lambda(r)\}$ is the derivative of $\Lambda'(r)$. Thus, we can obtain a renewable estimator of $\mathbf{UQPE}_\tau$ as

$$\widehat{\mathbf{UQPE}}_\tau^{b} = \hat{\boldsymbol{\delta}}_\tau^b/\hat{f}_{\boldsymbol{Y}}. \tag{4.13}$$

Theorem 4.3. Assume that conditions in Theorems 4.1 and 4.2 hold, we have

$$\sqrt{\bar{N}_b h_{\bar{N}_b}}\left(\widehat{\mathbf{UQPE}}_\tau^b - \mathbf{UQPE}_\tau\right) \xrightarrow{L} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{h_{\bar{N}_b}}\right), \tag{4.14}$$

where $\boldsymbol{\Sigma}_{h_{\bar{N}_b}} = \nu_K f_{\boldsymbol{Y}}^{-3}(q_\tau)\boldsymbol{\delta}_\tau\boldsymbol{\delta}_\tau^\top + \lim_{h_{\bar{N}_b}\to 0}\Big[h_{\bar{N}_b}\{\mathrm{var}(\boldsymbol{\psi}) - f_{\boldsymbol{Y}}^{-3}(q_\tau)\boldsymbol{\delta}_\tau\boldsymbol{\delta}_\tau^\top\}\Big]$, $\boldsymbol{\psi} = f_{\boldsymbol{Y}}^{-1}(q_\tau)\boldsymbol{\Sigma}_{\boldsymbol{X}\Lambda}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\beta}} - \{\boldsymbol{\Sigma}_{\boldsymbol{X}\Lambda}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y} = q_\tau) + f_{\boldsymbol{Y}}^{-2}(q_\tau)\boldsymbol{\delta}_\tau f_{\boldsymbol{Y}}'(q_\tau)\}\boldsymbol{\psi}_q$, $\boldsymbol{\Sigma}_{\boldsymbol{X}\Lambda} = \mathrm{E}\left\{\Lambda'(\boldsymbol{X}^\top\boldsymbol{\beta}_{q_\tau}) + \boldsymbol{\beta}_{q_\tau}\boldsymbol{X}^\top\Lambda''(\boldsymbol{X}^\top\boldsymbol{\beta}_{q_\tau})\right\}$, $\boldsymbol{\delta}_\tau = \boldsymbol{\beta}_{q_\tau}\mathrm{E}\left\{\Lambda'(\boldsymbol{X}^\top\boldsymbol{\beta}_{q_\tau})\right\}$ and $\boldsymbol{\psi}_q = f_{\boldsymbol{Y}}^{-1}(q_\tau)\{I(\boldsymbol{Y} > q_\tau) + \tau - 1\}$.

From the analysis in Theorem 4.1, $h_{\bar{N}_b} = O(\bar{N}_b^{-c_1}) = O(h_{f,b})$ which is the same convergence rate as the full data estimator h. Thus, $\boldsymbol{\Sigma}_{h_{\bar{N}_b}}$ is equal to $\boldsymbol{\Sigma}_h$. Therefore, the renewable estimator $\widehat{\mathbf{UQPE}}_\tau^b$ achieves optimal convergence speed and its asymptotic covariance matrix is the same as that of the estimator $\widehat{\mathbf{UQPE}}_\tau$ in (3.7) which is computed directly on all the samples.

We can estimate the $\boldsymbol{\Sigma}_{h_{\bar{N}_b}}$ in (4.14) by a renewable estimator as

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{h_{\bar{N}_b}} =& \nu_K\{\hat{f}_{\boldsymbol{Y}}^b\}^{-3}\hat{\boldsymbol{\delta}}_\tau^b\{\hat{\boldsymbol{\delta}}_\tau^b\}^\top + h_{\bar{N}_b}\{\hat{f}_{\boldsymbol{Y}}^b\}^{-4}\hat{\boldsymbol{\delta}}_\tau^b\{\hat{\boldsymbol{\delta}}_\tau^b\}^\top\left\{\omega_K\hat{f}_{\boldsymbol{Y}}'^b - \hat{f}_{\boldsymbol{Y}}^b\right\}\\ &+ h_{\bar{N}_b}\{\hat{f}_{\boldsymbol{Y}}^b\}^{-2}\frac{1}{\bar{N}_b}\sum_{j=1}^b\sum_{i\in\boldsymbol{D}_j}\mathbf{A}_i\mathbf{A}_i^\top,\end{aligned}$$

where $\omega_K = \int uK^2(u)du$, $\hat{f}_{\boldsymbol{Y}}'^b = \bar{N}_b^{-1}\sum_{j=1}^b\sum_{i\in\boldsymbol{D}_j}h_{f,j}^{-1}K_{h_{f,j}}'(Y_i - \hat{q}_\tau^j) + \bar{N}_b^{-1}\sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j}h_{f,j}^{-2}K_{h_{f,j}}''(Y_i - \hat{q}_\tau^j)(\hat{q}_\tau^j - \hat{q}_\tau^b)$, $\mathbf{A}_i = \hat{\psi}_i^b - \{\hat{f}_{\boldsymbol{Y}}^b\}^{-1}(\hat{H}^b - 1 + \tau)(\hat{\boldsymbol{\delta}}_\tau^b\hat{f}_{\boldsymbol{Y}}'^b/\hat{f}_{\boldsymbol{Y}}^b + \hat{G}_X^b\{\hat{G}_{XX}^b\}^{-1}\hat{f}_{\boldsymbol{YX}}^b)$, $\hat{\psi}_i^b = \hat{G}_X^b\{\hat{G}_{XX}^b\}^{-1}\boldsymbol{X}_i\{\hat{H}^b - \Lambda'(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j) - \Lambda''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}}_b - \hat{\boldsymbol{\beta}}_j)\} + \hat{\boldsymbol{\delta}}_{\tau,i}^b - \hat{\boldsymbol{\delta}}_\tau^b$ $\hat{G}_X^b = 2\bar{N}_b^{-1}\sum_{j=1}^b\sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_b\Lambda(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\{1 - \Lambda(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\}^2[1 + \{1 - 3\Lambda(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\}\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}}_b - \hat{\boldsymbol{\beta}}_j)]$, $\hat{G}_{XX}^b = \bar{N}_b^{-1}\sum_{j=1}^b\sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_i\boldsymbol{X}_i^\top[\Lambda'(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j) + \Lambda''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_j)\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}}_b - \hat{\boldsymbol{\beta}}_j)]$, $\hat{H}^b = H((Y_i - \hat{q}_\tau^j)/h_j) + H_{h_j}'(Y_i - \hat{q}_\tau^j)(\hat{q}_\tau^j - \hat{q}_\tau^b)$, $\hat{\boldsymbol{\delta}}_{\tau,i}^b$ is the ith of $\hat{\boldsymbol{\delta}}_\tau^b$ and $\hat{f}_{\boldsymbol{YX}}^b = \bar{N}_b^{-1}\sum_{j=1}^b\sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_iK_{h_{f,j}}(Y_i - \hat{q}_\tau^j) + \bar{N}_b^{-1}\sum_{j=1}^{b-1}\sum_{i\in\boldsymbol{D}_j}\boldsymbol{X}_ih_{f,j}^{-1}K_{h_{f,j}}'(Y_i - \hat{q}_\tau^j)(\hat{q}_\tau^j - \hat{q}_\tau^b)$.

From Theorems 4.1–4.3, we can see that there are no restrictions on the number of batches b, so b can be a very large number, even greater than $\max_j N_j$.

4.5. Algorithm

We summarize the general algorithm for the proposed renewable method to estimate $\mathbf{UQPE}_\tau$ by (4.13)) as follows.

Note that in step 9 in Algorithm, we only need to save $\hat{q}_\tau^b$, $\hat{\boldsymbol{\beta}}_b$ and $\boldsymbol{A}^b$. The scale of the data to be stored is $4 + p + p^2$ instead of $\bar{N}_b \times p$, which is the sample size of the streaming datasets up to b batches. Because p is assumed to be a fixed number in this article, our method greatly reduces the amount of data storage.

Algorithm 1: Renewable estimation for streaming datasets.

1: **Input:** streaming datasets $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_b, \ldots$, the quantile level τ, kernel function $K(\cdot)$, smoothing function $\mathrm{H}(\cdot)$ and bandwidths $h_{q,b}$, $h_{f,b}$, h_b with $b = 1, 2 \ldots$;
2: **Initialize:** calculate $\hat{q}_\tau^1$ by Fernandes, Guerre, and Horta (2021) with $\boldsymbol{D}_1$, $\hat{\boldsymbol{\beta}}_1$ by (4.7) with $\hat{q}_\tau^1$ and compute $\hat{\boldsymbol{K}}^1, \hat{f}_{\boldsymbol{Y}}^1, \hat{\boldsymbol{S}}_q^1$, and $\hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^1$;
3: **for:** $b = 2, 3, \ldots$ **do**
4: read in dataset $\boldsymbol{D}_b$;
5: obtain $\hat{q}_\tau^b$ by (4.5);
6: obtain $\hat{f}_{\boldsymbol{Y}}^b$ and $\hat{\boldsymbol{\beta}}_b$ by (4.6) and (4.12) with $\hat{q}_\tau^b$, respectively;
7: compute $\widehat{\mathbf{UQPE}}_\tau^b$ by (4.13);
8: update $\boldsymbol{A}^{b-1} = \{\hat{\boldsymbol{K}}^{b-1}, \hat{\boldsymbol{S}}_q^{b-1}, \hat{\boldsymbol{S}}_{\boldsymbol{\beta}}^{b-1}, \boldsymbol{K}_1^{b-1}, \boldsymbol{K}_2^{b-1}, \boldsymbol{K}_3^{b-1}\}$ to $\boldsymbol{A}^b$;
9: save $\hat{q}_\tau^b$, $\hat{\boldsymbol{\beta}}_b$, and $\boldsymbol{A}^b$, and release dataset $\boldsymbol{D}_b$ and other statistics from the memory;
10: **end**
11: **Output:** $\widehat{\mathbf{UQPE}}_\tau^b$.

5. Simulation Studies

In this section, we use Monte Carlo simulation studies to assess the finite sample performance of the proposed procedures in Sections 4. All programs are written in R code. We generate data from the following linear model:

$$\boldsymbol{Y} = 1 + \boldsymbol{X}^\top\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{5.1}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)^\top$ is a covariate vector and $\boldsymbol{X}_j, j = 1, 2, 3$ are drawn from a normal distribution $N(0, 1)$. The true value of the parameter is $\boldsymbol{\beta}_0 = (1, -2, 1)^\top$. Three error distributions of $\boldsymbol{\varepsilon}$ are considered: a standard normal distribution $N(0, 1)$, a t distribution with 3 degrees of freedom $t(3)$ which is a symmetric thick-tailed distribution and a Chi-square distribution with 1 degree of freedom $\chi^2(1)$ which is a skewed distribution. Quantile levels $\tau \in \{0.1, 0.5, 0.9\}$ are considered in all of the simulation experiments. Simulation results are all the average of 200 simulation replications.

For streaming datasets, we fix the sample size of each batch $N_j = 500$ for $j = 1, \ldots, b$ and vary the number of batches $b = 100, 200, 500, 1000, 2000, 5000, 10{,}000$. Then the total sample size is $\bar{N}_b = 500b$. We take the Logistic kernel $K(u) = e^{-u}/(1 + e^{-u})^2$ and $\tilde{K}(u) = 1/(1 + e^{-u})$ as used in He et al. (2023), and choose $h_{q,j} = \bar{N}_j^{-1/4}/\log\bar{N}_j$ for $j = 1, \ldots, b$ as in Jiang and Yu (2022).

5.1. Choosing the Bandwidths for the Density Estimations

We first study the selection of bandwidths for density estimations by f-Streaming (4.6). From Theorems 4.1, we choose $h_{f,j}$ as

$$h_{f,j} = C \times (0.5 + |\tau - 0.5|) \times \bar{N}_j^{-1/5}/\log\bar{N}_j, \tag{5.2}$$

where $C > 0$ is the scaling constant. We vary the constant C from 0.1 to 100. We use the relative absolute errors (RAE=$|\hat{f}_{\boldsymbol{Y}}(\hat{q}_\tau) -$

$f_{\boldsymbol{Y}}(q_\tau)|/|f_{\boldsymbol{Y}}(q_\tau)|)$ to evaluate the performance of the different estimation methods. We only consider $\boldsymbol{\varepsilon} \sim N(0,1)$ for model (4.1) because of $\boldsymbol{Y} \sim N(1,7)$ under this case. Then $f_{\boldsymbol{Y}}(q_\tau)$ at $\tau = 0.1, 0.5, 0.9$ are 0.0663, 0.1508, and 0.0663, respectively.

The simulation results of RAEs are shown in Table 1. (i) As can be seen from Table 1 that $C = 10$ is a good choice for $h_{f,j}$ because of smallest RAEs in most cases. (ii) The method (4.6) of density estimation for streaming data is effective and very close to f-All (the full data estimation by (3.6)) in case $C = 10$.

5.2. Study the Sensitivity of $\widehat{\mathbf{UQPE}}_\tau^b$ to Bandwidths $\{h_j\}_{j=1}^b$

We study the sensitivity of $\widehat{\mathbf{UQPE}}_\tau^b$ in (4.13) to bandwidths $\{h_j\}_{j=1}^b$. From Theorem 4.3, we choose $h_j = C \times \bar{N}_j^{-1/4}/\log \bar{N}_j$ similar to $h_{q,j}$ for $j = 1, \ldots, b$, where $C > 0$ is the scaling constant. We vary the constant C from 0.01 to 100.

To evaluate the performance of the different estimation methods, we calculate the root-mean-square error (RMSE):

$$\text{RMSE} = \frac{1}{3}\sqrt{\sum_{j=1}^{3}(\widehat{\mathbf{UQPE}}_{\tau,j}^b - \mathbf{UQPE}_{\tau,j})^2}, \tag{5.3}$$

where the true value $\mathbf{UQPE}_\tau$ is $\boldsymbol{\beta}_0$ under settings of model (5.1). According to the analysis of Section 5.1, we choose $h_{f,j} = 10 \times (0.5 + |\tau - 0.5|) \times \bar{N}_j^{-1/5}/\log \bar{N}_j$ for $j = 1, \ldots, b$. The simulation results of the RMSE in Table 2 show that the performances of $\widehat{\mathbf{UQPE}}_\tau^b$ are better under $C = 0.1, 1, 10$ than those of $C = 0.01, 100$, and $\widehat{\mathbf{UQPE}}_\tau^b$ is insensitive to bandwidths $\{h_j\}_{j=1}^b$ under $C = 0.1, 1, 10$. Therefore, we can choose $h_j = 10 \times \bar{N}_j^{-1/4}/\log \bar{N}_j$ based on the smallest RMSEs in most cases for $j = 1, \ldots, b$.

Table 1. The means and standard deviations (in parentheses) of the RAEs ($\times 100$) for f-Streaming under different C, quantile levels $\tau = 0.1, 0.5, 0.9$ and $b = 100, 10{,}000$ for simulation study 5.1.

		$b = 100$		$b = 10{,}000$	
τ	C	f-All	f-Streaming	f-All	f-Streaming
0.1	0.1	18.38 (15.46)	118.36 (104.86)	4.58 (3.32)	9.00 (7.60)
	0.5	7.96 (5.97)	13.20 (11.26)	1.63 (1.12)	1.67 (1.07)
	1	5.93 (4.73)	7.70 (6.16)	1.17 (0.82)	1.26 (0.86)
	5	2.50 (1.85)	2.69 (2.16)	0.44 (0.45)	0.45 (0.40)
	10	1.73 (1.28)	2.18 (1.54)	0.32 (0.26)	0.30 (0.19)
	50	2.88 (0.97)	3.92 (1.76)	0.24 (0.14)	0.56 (0.22)
	100	6.12 (0.69)	4.70 (1.30)	1.00 (0.12)	1.67 (0.18)
0.5	0.1	17.12 (13.63)	134.72 (126.04)	2.57 (2.32)	8.30 (6.24)
	0.5	7.49 (5.90)	15.03 (13.59)	1.53 (1.31)	1.95 (1.28)
	1	5.33 (3.88)	7.05 (5.62)	0.77 (0.91)	1.21 (1.06)
	5	2.20 (1.70)	1.96 (1.58)	0.47 (0.31)	0.41 (0.25)
	10	1.59 (1.18)	1.30 (1.04)	0.26 (0.22)	0.26 (0.18)
	50	1.61 (0.73)	3.25 (0.66)	0.14 (0.11)	0.26 (0.13)
	100	5.85 (0.47)	10.57 (0.38)	0.53 (0.12)	1.08 (0.10)
0.9	0.1	19.88 (14.34)	111.97 (103.34)	3.09 (2.37)	11.24 (7.13)
	0.5	8.41 (6.35)	15.16 (12.08)	1.50 (1.32)	1.40 (0.90)
	1	5.88 (4.89)	7.28 (5.80)	1.26 (0.68)	0.96 (0.76)
	5	2.65 (1.84)	3.01 (2.14)	0.46 (0.31)	0.42 (0.26)
	10	1.63 (1.24)	2.20 (1.61)	0.27 (0.22)	0.28 (0.19)
	50	2.72 (0.93)	4.02 (1.69)	0.25 (0.15)	0.51 (0.31)
	100	6.01 (0.67)	4.65 (1.28)	1.05 (0.13)	1.81 (0.21)

5.3. Simulation Studies for Renewable Estimation Methods

Building upon the analysis in Sections 5.1 and 5.2, we proceed to investigate the performance of the proposed renewable estimation method. In order to assess the effectiveness of various estimation methods, we compute the Root Mean Square Error (RMSE) as outlined in (5.3), along with the corresponding computation time in seconds. It's worth noting that, for brevity, we only present the computation time for the normal error, as the computation time for different errors is closely comparable.

The simulation results presented in Tables 3–6 lead to the following conclusions:

(i) Regarding the RMSEs in Tables 3–5, both UQPE-A (using the all data estimator (3.7)) and UQPE-S (our proposed renewable estimator for streaming data (4.13)) closely approximate the true values. The RMSE results are consistently small across various numbers of batches b, quantile levels τ, and errors. Notably, UQPE-S demonstrates proximity to UQPE-A in terms of accuracy. (ii) Examining the computation time t in Table 6, it is evident that UQPE-S is significantly faster to compute than UQPE-A across all scenarios. (iii) As the number of batches b

Table 2. The means and standard deviations (in parentheses) of the RMSEs ($\times 100$) under different C, quantile levels $\tau = 0.1, 0.5, 0.9$, $b = 100, 10{,}000$ and errors for simulation study 5.2.

		$b = 100$			$b = 10000$		
Error	C	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
$N(0,1)$	0.01	3.69 (7.06)	1.88 (2.19)	5.32 (8.05)	0.33 (0.39)	0.20 (0.17)	0.30 (0.18)
	0.1	2.51 (1.85)	1.37 (0.91)	2.93 (4.76)	0.31 (0.20)	0.27 (0.16)	0.30 (0.22)
	1	2.15 (1.36)	1.23 (0.83)	2.09 (1.46)	0.25 (0.17)	0.21 (0.15)	0.33 (0.20)
	10	2.06 (1.38)	1.22 (0.79)	1.87 (1.19)	0.22 (0.15)	0.20 (0.18)	0.30 (0.19)
	100	2.38 (1.93)	1.59 (0.98)	2.69 (3.67)	0.39 (0.36)	0.26 (0.21)	0.24 (0.21)
$t(3)$	0.01	3.86 (6.12)	2.05 (1.85)	4.07 (7.96)	0.30 (0.21)	0.26 (0.19)	0.28 (0.15)
	0.1	2.72 (2.08)	1.38 (0.77)	2.77 (1.91)	0.29 (0.21)	0.22 (0.17)	0.33 (0.25)
	1	2.29 (1.56)	1.41 (0.84)	2.20 (1.39)	0.36 (0.17)	0.25 (0.22)	0.34 (0.19)
	10	2.24 (1.38)	1.29 (0.84)	2.24 (1.40)	0.29 (0.20)	0.21 (0.18)	0.33 (0.26)
	100	2.00 (1.13)	1.54 (0.85)	2.17 (1.40)	0.35 (0.23)	0.21 (0.16)	0.28 (0.20)
$\chi^2(1)$	0.01	4.26 (9.82)	2.01 (3.92)	4.05 (4.98)	0.30 (0.18)	0.22 (0.14)	0.40 (0.27)
	0.1	3.64 (6.95)	1.46 (0.86)	2.86 (1.88)	0.27 (0.12)	0.17 (0.13)	0.31 (0.28)
	1	2.01 (1.39)	1.27 (0.75)	2.49 (1.57)	0.25 (0.12)	0.24 (0.20)	0.36 (0.22)
	10	1.91 (1.56)	1.19 (0.75)	2.49 (1.56)	0.34 (0.22)	0.24 (0.15)	0.35 (0.25)
	100	4.95 (6.70)	1.47 (0.99)	2.19 (1.27)	1.15 (1.76)	0.27 (0.24)	0.33 (0.20)

Table 3. The means and standard deviations (in parentheses) of the RMSEs (×100) under different estimation methods, quantile levels $\tau = 0.1, 0.5, 0.9$ and b for simulation study 5.3 with $\varepsilon \sim N(0, 1)$.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
b	UQPE-A	UQPE-S	UQPE-A	UQPE-S	UQPE-A	UQPE-S
100	1.56 (1.04)	1.85 (1.30)	1.31 (0.89)	1.17 (0.74)	1.61 (1.02)	1.96 (1.22)
200	1.25 (0.81)	1.51 (1.00)	1.15 (0.78)	1.01 (0.70)	1.26 (0.82)	1.54 (1.06)
500	0.82 (0.54)	0.85 (0.65)	0.68 (0.43)	0.62 (0.38)	0.85 (0.59)	1.00 (0.63)
1000	0.74 (0.47)	0.75 (0.48)	0.63 (0.41)	0.54 (0.33)	0.66 (0.40)	0.74 (0.49)
2000	0.55 (0.29)	0.54 (0.39)	0.48 (0.33)	0.45 (0.30)	0.53 (0.34)	0.65 (0.39)
5000	0.31 (0.21)	0.33 (0.22)	0.35 (0.23)	0.29 (0.22)	0.38 (0.30)	0.39 (0.26)
10,000	0.30 (0.20)	0.29 (0.21)	0.14 (0.09)	0.14 (0.08)	0.28 (0.18)	0.27 (0.16)

Table 4. The means and standard deviations (in parentheses) of the RMSEs (×100) under different estimation methods, quantile levels $\tau = 0.1, 0.5, 0.9$ and b for simulation study 5.3 with $\varepsilon \sim t(3)$.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
b	UQPE-A	UQPE-S	UQPE-A	UQPE-S	UQPE-A	UQPE-S
100	1.84 (1.14)	2.15 (1.36)	1.48 (1.03)	1.32 (0.84)	1.79 (1.04)	1.98 (1.24)
200	1.33 (0.83)	1.70 (1.10)	1.28 (0.81)	1.10 (0.70)	1.43 (0.92)	1.61 (1.08)
500	0.98 (0.64)	1.22 (0.73)	0.78 (0.47)	0.71 (0.41)	0.92 (0.63)	1.03 (0.70)
1000	0.82 (0.50)	0.88 (0.55)	0.63 (0.43)	0.55 (0.39)	0.71 (0.51)	0.88 (0.58)
2000	0.57 (0.38)	0.63 (0.43)	0.48 (0.29)	0.45 (0.27)	0.61 (0.45)	0.74 (0.53)
5000	0.40 (0.24)	0.40 (0.26)	0.29 (0.21)	0.27 (0.19)	0.38 (0.25)	0.41 (0.24)
10,000	0.34 (0.23)	0.34 (0.25)	0.24 (0.14)	0.19 (0.14)	0.27 (0.15)	0.23 (0.14)

Table 5. The means and standard deviations (in parentheses) of the RMSEs (×100) under different estimation methods, quantile levels $\tau = 0.1, 0.5, 0.9$ and b for simulation study 5.3 with $\varepsilon \sim \chi^2(1)$.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
b	UQPE-A	UQPE-S	UQPE-A	UQPE-S	UQPE-A	UQPE-S
100	1.62 (0.99)	1.89 (1.28)	1.44 (0.97)	1.27 (0.84)	2.03 (1.24)	2.24 (1.30)
200	1.07 (0.67)	1.44 (1.07)	1.05 (0.72)	0.93 (0.62)	1.58 (0.85)	1.76 (1.11)
500	0.87 (0.58)	0.95 (0.67)	0.78 (0.55)	0.72 (0.50)	0.98 (0.60)	1.17 (0.69)
1000	0.68 (0.46)	0.69 (0.48)	0.63 (0.42)	0.58 (0.34)	0.79 (0.50)	0.94 (0.57)
2000	0.47 (0.29)	0.55 (0.40)	0.42 (0.34)	0.37 (0.30)	0.64 (0.37)	0.58 (0.37)
5000	0.38 (0.27)	0.41 (0.29)	0.34 (0.25)	0.32 (0.22)	0.40 (0.24)	0.38 (0.27)
10,000	0.29 (0.20)	0.30 (0.25)	0.21 (0.17)	0.22 (0.15)	0.36 (0.19)	0.35 (0.21)

Table 6. The means of computing time t (in seconds) under different estimation methods, quantile levels $\tau = 0.1, 0.5, 0.9$ and b for simulation study 5.3 with $\varepsilon \sim N(0, 1)$.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
b	UQPE-A	UQPE-S	UQPE-A	UQPE-S	UQPE-A	UQPE-S
100	0.31	0.09	0.28	0.10	0.24	0.07
200	0.72	0.18	0.57	0.18	0.46	0.12
500	2.07	0.41	1.44	0.41	1.17	0.27
1000	4.30	0.82	3.07	0.82	2.73	0.53
2000	7.91	1.51	6.52	1.72	5.44	1.03
5000	18.76	3.56	15.48	4.15	14.70	2.77
10,000	39.03	7.50	29.16	8.06	42.42	7.94

increases, RMSEs decrease, and computation time t expands, aligning with expectations.

6. Empirical Application

6.1. Labor Income and Minimum Wage Dataset

To illustrate the proposed methods in Sections 4, we employ a substantial sample consisting of 941,174 observations derived from the 2011 to 2020 Current Population Survey (CPS)-merged outgoing rotation group earnings data. This dataset is accessible online for replication at *https://www.nber.org/research/data/current-population-survey-cps-merged-outgoing-rotation-group-earnings-data*. Additionally, we use a dataset documenting the minimum wage (minimum hourly wage) set by U.S. states from 2011 to 2020, obtainable at *http://www.dol.gov/whd/state/stateMinWageHis.htm*. This dataset is also considered streaming data, as data continually enters the stream over time, resulting in a substantial volume of data. The objective is to evaluate the effects of *Minwage* (minimum wage) on the quantile of the unconditional distribution of log wages. In this application, $\boldsymbol{Y} = lwage$ (log hourly wage), $\boldsymbol{X}_1 = Minwage$ (our focal covariate), and other covariables include $\boldsymbol{X}_2 = Age$, $\boldsymbol{X}_3 = Sex$ (1 for female and 0 for male), $\boldsymbol{X}_4 = Grade92$ (the highest grade completed), $\boldsymbol{X}_5 = Race$, $\boldsymbol{X}_6 = Marital$ (marital status), and $\boldsymbol{X}_7 = Ftpt94$ (full-time or part-time status). Additional details on data processing can be found at *https://data.nber.org/morg/docs/cpsx.pdf*.

The minimum wage system serves as a government policy tool aimed at adjusting income distribution in the primary stage and is often a crucial means of poverty alleviation. Numerous scholars in the literature have delved into the CPS dataset and examined the minimum wage. For instance, Lee (1999) used CPS data spanning from 1979 to 1989 to explore the relationship between labor income (hourly wage) and the minimum wage. Their analysis suggests that the minimum wage can significantly contribute to the increase in dispersion in the lower tail

Table 7. The estimated coefficients by methods OLS, CQR, UQPE-A, UQPE-S-Year, and UQPE-S-Month for Labor income and minimum wage dataset.

τ	Method	Minwage	Age	Sex	Grade92	Race	Marital	Ftpt94
–	OLS	0.0181	0.0093	−0.1802	0.1750	0.0048	−0.0027	−0.0354
0.1	CQR	0.0228	0.0057	−0.1123	0.1625	0.0043	0.0045	−0.0414
	UQPE-A	0.0013	0.0019	−0.0347	0.0079	−0.0011	−0.0126	−0.0262
	UQPE-S-Year	0.0013	0.0027	−0.0468	0.0115	−0.0011	−0.0163	−0.0333
	UQPE-S-Month	0.0015	0.0028	−0.0467	0.0118	−0.0011	−0.0167	−0.0343
0.5	CQR	0.0159	0.0088	−0.1805	0.1764	0.0024	−0.0053	−0.0370
	UQPE-A	−0.0028	0.0037	−0.1166	0.0081	−0.0121	−0.0373	−0.0679
	UQPE-S-Year	−0.0026	0.0055	−0.1500	0.0144	−0.0157	−0.0523	−0.1014
	UQPE-S-Month	−0.0027	0.0057	−0.1534	0.0149	−0.0160	−0.0537	−0.1042
0.9	CQR	0.0184	0.0154	−0.2519	0.1832	0.0101	−0.0003	−0.0124
	UQPE-A	−0.0093	0.0023	−0.1004	−0.0070	−0.0129	−0.0474	−0.0346
	UQPE-S-Year	−0.0090	0.0025	−0.1185	−0.0053	−0.0145	−0.0496	−0.0443
	UQPE-S-Month	−0.0090	0.0025	−0.1168	−0.0056	−0.0144	−0.0496	−0.0429

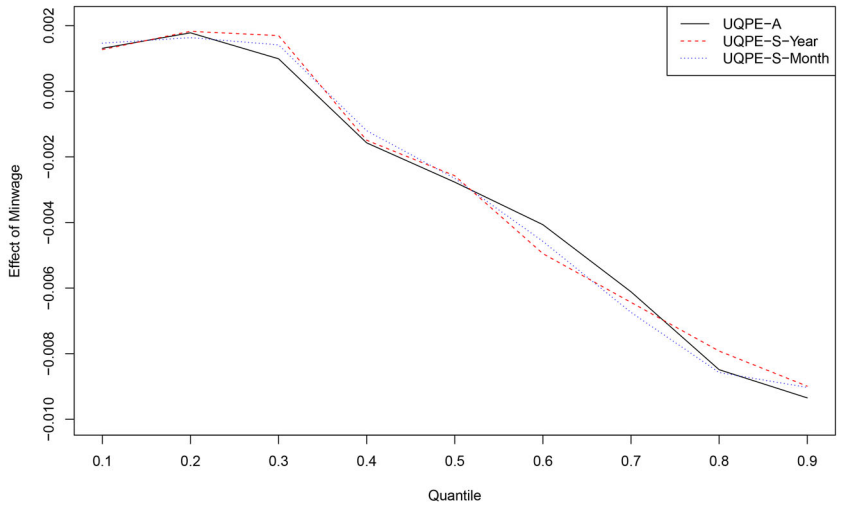

Figure 2. The estimates of the effect of *Minwage* on log wages by UQPE-A, UQPE-S-Year, and UQPE-S-Month for Labor income and minimum wage dataset.

of the wage distribution, especially for women. In a similar vein, Dube (2019) examined individual-level data from the CPS covering the period between 1984 and 2013. Their study provided an assessment of how U.S. minimum wage policies impact the distribution of family incomes for the non-elderly population. They comprehensively characterized how minimum wage increases shift the cumulative distribution of family incomes, subsequently using this information to estimate the unconditional quantile partial effects (UQPE) of the policy.

For comparison with standard OLS (conditional mean) estimates and with standard (conditional) quantile regressions (CQR), we use the following linear model for OLS and CQR:

$$\boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.1}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_7)^\top$. The mean relative absolute errors (MRAE=$n^{-1}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|/|Y_i|$) of OLS and CQR with quantile level 0.5 are 5.318% and 5.317%, respectively. Therefore, model (6.1) is assumed to be reasonable for OLS and CQR. Table 7 reports the estimated coefficients of model (6.1) by methods OLS, UQPE-A, and CQR for the 10th, 50th, and 90th quantiles, which shows the difference between conditional and unconditional quantiles regressions.

Next, we consider the proposed method UQPE-S in Sections 4. Since the data of minimum wage is recorded in years and the data of CPS is recorded in months, we consider $b = 10$ (year) and 120 (month), respectively. The data of 2011 and January 2011 are regarded as $\boldsymbol{D}_1$ for UQPE-S-Year ($b = 10$) and UQPE-S-Month ($b = 120$), respectively. The difference between the estimated effect of *Minwage* for UQPE-A, UQPE-S-Year and UQPE-S-Month is illustrated in Figure 2, which plots at nine different quantiles (from the 10th to the 90th). From Table 7 and Figure 2, we can see that the estimated coefficients and the effects of *Minwage* on log wages under different quantiles by UQPE-A, UQPE-S-Year, and UQPE-S-Month are all very close.

Finally, as with the streaming data setup, when the last batch of data arrives (2020 and December 2020 are regarded as $\boldsymbol{D}_b$ for UQPE-S-Year ($b = 10$) and UQPE-S-Month ($b = 120$), respectively), the total running time of UQPE-A with nine quantiles is also 52.98 sec, which needs to compute all the data. However, UQPE-S-Year is 0.87 sec and UQPE-S-Month is 0.08 sec, because only $\boldsymbol{D}_b$ and some past statistics are required. In addition, for the setting of massive data, that is, all data can be stored, the cumulative time from 1 to b of UQPE-S-Year ($b = 10$) and UQPE-S-Month ($b = 120$) with nine quantiles is 21.58 sec and 18.51 sec, respectively. Therefore, UQPE-S (UQPE-S-Year and UQPE-S-Month) is much faster to compute than UQPE-A.

6.2. Daily Return of Stocks and Exchange Rate Dataset

In order to illustrate the proposed methods in Sections 4 for large batches b, we used data from 871 trading days between

Table 8. The estimated coefficient of ERF by methods CQR, UQPE-A, and UQPE-S for Daily return of stocks and exchange rate dataset.

Method	$\tau = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CQR	−0.6335	−0.6257	−0.6094	−0.5740	−0.5190	−0.4916	−0.4628	−0.4384	−0.4071
UQPE-A	−0.5625	−0.5693	−0.5930	−0.6043	−0.6201	−0.6187	−0.6443	−0.7395	−1.1355
UQPE-S	−0.5466	−0.5674	−0.5812	−0.6121	−0.6252	−0.6322	−0.6749	−0.7856	−1.1572

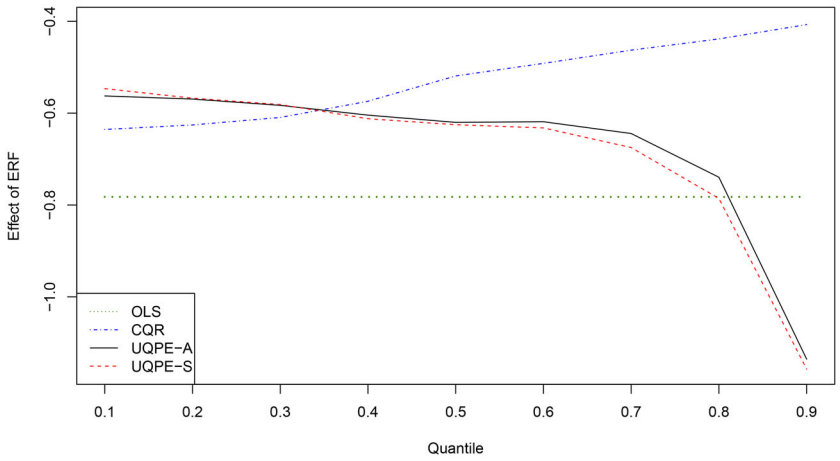

Figure 3. The estimates of the effect of *ERF* on *Return* by OLS, CQR, UQPE-A, and UQPE-S.

January 1, 2020 and August 25, 2023 for all companies listed on the Shanghai Stock Exchange in China and exchange rate between the U.S. dollar and the Chinese currency. The dataset contains 1,339,591 observations, where suspension and special treatment stock transactions have been deleted. There were 1546 stocks in 2020, 1629 stocks in 2021, 1665 stocks in 2022 and 1689 stocks in 2023. We focus on the effects of *Return* (daily return of stock) on the quantile of the unconditional distribution of *ERF* (exchange rate fluctuation).

In this application, $\boldsymbol{Y} = \textit{Return}$, which is (closing price of the day − closing price of the previous day)/closing price of the previous day, $\boldsymbol{X}_1 = \textit{ERF}$, which is our interested covariate, $\boldsymbol{X}_2 = \textit{TR}$ (daily turnover rate of stock) and $\boldsymbol{X}_3 = \textit{Vol}$ (daily trading volume of stock). Similar to the analysis of streaming time series data conducted in Deshpande, Javanmard, and Mehrabi (2023) and Jiang, Choy, and Yu (2023), the dataset in this study is not independent and identically distributed. However, we employ OLS, UQPE-A, CQR, and UQPE-S to analyze the dataset. It is important to note that this application poses a limitation on the iid assumption commonly used in practical time series data analysis. Table 8 and Figure 3 report the estimated coefficients $\boldsymbol{X}_1 = \textit{ERF}$ of model (6.1) with $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)^\top$ by methods OLS, CQR, UQPE-A, and UQPE-S under nine different quantiles (from the 10th to the 90th), which shows the difference between conditional and unconditional quantiles regressions.

Next, we consider the proposed method UQPE-S in Sections 4. For different stocks, the exchange rate is the same every day, so we choose $\boldsymbol{D}_1$ for the first 30 trading days. Since the data between January 1, 2021 and August 25, 2023 is recorded in days, we consider $b = 842$. The difference between the estimated effect of *ERF* for UQPE-A and UQPE-S is illustrated in Table 8 and Figure 3. From Table 8 and Figure 3, we can see that the effects of *ERF* on *Return* under different quantiles by UQPE-A and UQPE-S are all very close.

Finally, as with the streaming data setup, when the last batch of data arrives (the transaction data as of August 25, 2023 is regarded as $\boldsymbol{D}_b$), the total running time of UQPE-A with nine quantiles is also 65.37 sec and UQPE-S is 0.02 sec. In addition, for the setting of massive data, that is, all data can be stored, the cumulative time from 1 to b of UQPE-S ($b = 842$) with nine quantiles is 18.11 sec. Therefore, UQPE-S is much faster to compute than UQPE-A.

7. Discussion

In this article, we delve into renewable parameter estimation for unconditional quantile regression applied to streaming datasets. A pivotal insight derived from our work is the introduction of a smoothing logistic regression estimator, a crucial tool in generating renewable estimators for unconditional quantile regression. This innovative approach necessitates only the availability of the current data batch within the stream, along with sufficient statistics on the historical data at each stage of analysis. Notably, our proposed renewable methods do not impose constraints on the number of batches, allowing them to adapt seamlessly to situations where streaming data arrives rapidly and continuously.

Theoretical analysis reveals that the proposed estimators for streaming datasets attain optimal efficiency, with asymptotic covariance matrices mirroring those of estimators derived from full data. The algorithm's swiftness stems from its reliance on the Newton-Raphson method. Empirical results presented in Sections 5 and 6 demonstrate that our proposed method closely approximates the estimator derived from complete data, yet boasts a shorter running time.

Furthermore, the smoothing technique employed for the logistic regression estimator in this article can be extended to benefit other estimation methods, such as quantile regression and Huber estimation. As highlighted in Kong and Xia (2019)

and Jiang and Yu (2022), our renewable estimation method is not confined solely to streaming data, but is also apt for the analysis of massive data. "Massive data" denotes data that exceeds a computer's storage or computational capacity, often originating from a distributed system. Specifically, we can employ the divide-and-conquer method to partition the dataset into b blocks, or the dataset itself may stem from b sub-devices. This empowers us to analyze massive data with the aid of our renewable estimation method.

Supplementary Materials

The proofs of the proposed theorems are given in the supplementary material file.

Acknowledgments

We thank the editor, associate editor and two reviewers for their constructive comments, which helped us improve the article.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This research is supported by the National Social Science Foundation of China (Series number: 21BTU040) and the Ministry of Education of the People's Republic of China, Humanities and Social Science Foundation (Series number: 22YJC910005) and London Mathematics Society (Scheme 42206).

References

Amba, S. M., and Nguyen, B. H. (2019), "Exchange Rate and Equity Price Relationship: Empirical Evidence from Mexican and Canadian Markets," *The International Journal of Business and Finance Research*, 13, 33–43. [1145]

Bahmani-Oskooee, M., and Sohrabian, A. (1992), "Stock Prices and the Effective Exchange Rate of the Dollar," *Applied Economics*, 24, 459–64. [1145]

Borgen, N. (2016), "Fixed Effects in Unconditional Quantile Regression," *The Stata Journal*, 16, 403–415. [1143]

Chen, X., Liu, W., and Zhang, Y. (2019), "Quantile Regression Under Memory Constraint," *Annals of Statistics*, 47, 3244–3273. [1144]

Deshpande, Y., Javanmard, A., and Mehrabi, M. (2023), "Online Debiasing for Adaptively Collected High-Dimensional Data with Applications to Time Series Analysis," *Journal of the American Statistical Association*, 118, 1126–1139. [1144,1153]

Dube, A. (2019), "Minimum Wages and the Distribution of Family Incomes," *American Economic Journal: Applied Economics*, 11, 268–304. [1152]

Fernandes, M., Guerre, E., and Horta, E. (2021), "Smoothing Quantile Regressions," *Journal of Business & Economic Statistics*, 39, 338–357. [1144,1146,1149]

Firpo, S., Fortin, N. M., and Lemieux, T. (2009), "Unconditional Quantile Regressions," *Econometrica*, 77, 953–973. [1143,1145,1146,1147,1148]

Ghosh, P. (2021), "Box-Cox Power Transformation Unconditional Quantile Regressions with An Application on Wage Inequality," *Journal of Applied Statistics*, 48, 3086–3101. [1143]

He, X., Pan, X., Tan, K. M., and Zhou, W. (2023), "Smoothed Quantile Regression with Large Scale Inference," *Journal of Econometrics*, 232, 367–388. [1144,1149]

Horowitz, J. (1998), "Bootstrap Methods for Median Regression Models," *Econometrica*, 66, 1327–1352. [1144]

Inoue, A., Li, T., and Xu, Q. (2021), "Two Sample Unconditional Quantile Effect," arXiv:2105.09445v1. [1143]

Jiang, R., Choy, S. K., and Yu, K. (2023), "Non-Crossing Quantile Double-Autoregression for the Analysis of Streaming Time Series Data," *Journal of Time Series Analysis*. DOI:10.1111/jtsa.12725. [1153]

Jiang, R., and Yu, K. (2021), "Smoothing Quantile Regression for a Distributed System," *Neurocomputing*, 466, 311–326. [1148]

——— (2022), "Renewable Quantile Regression for Streaming Data Sets," *Neurocomputing*, 508, 208–224. [1144,1149,1154]

——— (2023), "No-Crossing Single-Index Quantile Regression Curve Estimation," *Journal of Business & Economic Statistics*, 41, 309–320. [1143]

Koenker, R., and Bassett, G. (1978), "Regression Quantile," *Econometrica*, 46, 33–50. [1143]

Kong, E., and Xia, Y. (2019), "On the Efficiency of Online Approach to Nonparametric Smoothing of Big Data," *Statistica Sinica*, 29, 185–201. [1144,1147,1153]

Lee, D. S. (1999), "Wage Inequality in the United States during the 1980s: Rising Dispersion or Falling Minimum Wage?" *The Quarterly Journal of Economics*, 114, 977–1023. [1151]

Luo, L., and Song, P. (2020), "Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Data Sets," *Journal of the Royal Statistical Society*, Series B, 82, 69–97. [1144]

Luo, L., Zhou, L., and Song, P. (2023), "Real-Time Regression Analysis of Streaming Clustered Data with Possible Abnormal Data Batches," *Journal of the American Statistical Association*, 118, 2029–2044. [1144]

Martinez-Iriarte, J., Montes-Rojas, G., and Sun, Y. (2022), "Location-Scale and Compensated Effects in Unconditional Quantile Regressions," arXiv:2201.02292v1. [1143]

Nadaraya, E. (1964), "Some New Estimates for Distribution Functions," *Theory of Probability and Its Applications*, 9, 497–500. [1146]

Pan, M.-S., Fok, R., and Liu, Y. (2007), "Dynamic Linkages between Exchange Rates and Stock Prices: Evidence from East Asian Markets," *International Review of Economics & Finance*, 16, 503–520. [1145]

Sasaki, Y., Ura, T., and Zhang, Y. (2022), "Unconditional Quantile Regression with High-Dimensional Data," *Quantitative Economics*, 13, 955–978. [1143]

Schifano, E., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016), "Online Updating of Statistical Inference in the Big Data Setting," *Technometrics*, 58, 393–403. [1144]

Wang, H., Min, Y., and Stufken, J. (2019), "Information-based Optimal Subdata Selection for Big Data Linear Regression," *Journal of the American Statistical Association*, 114, 393–405. [1144]

Wang, K., Wang, H., and Li, S. (2022), "Renewable Quantile Regression for Streaming Datasets," *Knowledge-Based Systems*, 235, 107675. [1144]

Xie, R., Bai, S. Y., and Ma, P. (2023), "Optimal Sampling Designs for Multi-Dimensional Streaming Time Series with Application to Power Grid Sensor Data," arXiv:2303.08242v1. [1144]

Yang, Y., and Yao, F. (2023), "Online Estimation for Functional Data," *Journal of the American Statistical Association*, 118, 1630–1644. [1144]

Zhang, B., and Li, X. (2010), "Currency Appreciation and Stock Market Performance: Evidence from China," *Frontiers of Economics in China*, 5, 393–411. [1145]