Distributed Subdata Selection for Big Data via Sampling-Based Approach

Haixiang Zhang

Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

HaiYing Wang*

Department of Statistics, University of Connecticut, Storrs, Mansfield, CT 06269, USA

Abstract

With the development of modern technologies, it is possible to gather an extraordinarily large number of observations. Due to the storage or transmission burden, big data are usually scattered at multiple locations. It is difficult to transfer all of data to the central server for analysis. A distributed subdata selection method for big data linear regression model is proposed. Particularly, a two-step subsampling strategy with optimal subsampling probabilities and optimal allocation sizes is developed. The subsample-based estimator effectively approximates the ordinary least squares estimator from the full data. The convergence rate and asymptotic normality of the proposed estimator are established. Simulation studies and an illustrative example about airline data are provided to assess the performance of the proposed method.

Keywords: Allocation sizes, Big data, Distributed subsampling, Optimal subsampling, Regression diagnostic

1. Introduction

Nowadays, big data analysis has become an interesting and important research field. The most common feature of big data is the abundant number of observations (large n), which lays heavy burden on storage and computation facilities. To deal with this difficulty, many efforts have been devoted in the literature. There are mainly three directions from the view of statistical applications: divide-and-conquer, online updating, and subsampling. The basic idea of divide-and-conquer approach is to split the whole data into many small data sets, and conduct statistical inference separately over each individual data set. The final estimator can be obtained by merging estimators from all small data sets. For example, Zhao et al. (2016) considered a partially linear framework for modeling massive heterogeneous data. Battey et al. (2018) studied the topic on hypothesis testing and parameter estimation with a divide and conquer algorithm. Jordan et al. (2019) presented a communication-efficient surrogate likelihood framework for distributed statistical inference problems. Shi et al. (2018) studied the divide and conquer method for cubic-rate estimators under massive

*Corresponding author
215 Glenbrook Rd. U-4120, Storrs, Mansfield, CT 06269, USA;
Tel: 1.860.486.6142; Fax: 1.860.486.4113.
Email address: haiying.wang@uconn.edu (HaiYing Wang)

data framework. Volgushev et al. (2019) proposed a two-step distributed method for quantile regression with massive data sets. The online updating method focuses on big data that observations are not available all at once. Instead, they arrive in chunks from a data stream. Schifano et al. (2016) developed some iterative estimating algorithms and statistical inference procedures for linear models and estimating equations with streaming data. Wang et al. (2018a) proposed an online updating method that could incorporate new variables for big data streams. Xue et al. (2019) proposed an online updating approach for testing the proportional hazards assumption with big survival data.

Another popular method is the subsampling approach, where the basic idea is to select a subsample for the purpose of statistical inference. Ma et al. (2015) proposed an algorithmic leveraging-based sampling procedure. Wang et al. (2018b) and Wang (2019) developed some optimal subsampling methods for logistic regression with massive data. Wang et al. (2019) provided a novel information-based optimal subdata selection approach. Ai et al. (2019) studied the optimal subsampling algorithms for big data generalized linear models. Wang & Ma (2020) considered the optimal subsampling for quantile regression in big data. For the above-mentioned subsampling methods, one common assumption is that the data is stored in one location. However, massive data are often partitioned across multiple servers due to the storage burden or privacy limit. For example, Wal-mart stores generate a large number of data from different locations around the world, and it is difficult to transmit these data to a central location; the clinical information of patients are usually stored at different hospitals, and it is impractical to touch these data simultaneously due to personal privacy. To deal with this problem, we propose a distributed subdata selection method in the framework of linear models. The main features of our approach are as follows: First, we focus on the subsampling method for several distributed big data, where each data source has a huge number of observations. Second, we present the optimal subsampling probabilities and allocation sizes that minimize the asymptotic mean squared error of the resultant estimator. Third, we establish consistency and asymptotic normality of the proposed estimator, which are useful for statistical inference.

The remainder of this article is organized as follows: In Section 2, we propose a distributed subdata selection algorithm in the framework of sampling technique. We establish consistency and asymptotic normality of the subsample estimator based on a general subsampling algorithm. In Section 3, we provide an optimal sampling criterion with the focus on developing subsampling probabilities and allocation sizes, which minimize the asymptotic mean squared error of the resultant estimator. Then we develop a two-step algorithm to approximate the optimal subsampling procedure. In Section 4, we conduct an extensive simulation study to verify the effectiveness of our method. An application to airline data is presented in Section 5. In Section 6, we provide concluding remarks and future research topics. All proof details are given in the Appendix.

2. Distributed Subdata Selection Algorithm

Given a covariate $\mathbf{X} \in \mathbb{R}^p$, we consider the linear regression model

$$Y = \boldsymbol{\beta}^T \mathbf{X} + \epsilon, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a vector of regression coefficients, ϵ is an error term with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. We assume that $\mathbf{X}$ is nonrandom, and the first component of $\mathbf{X}$ is 1 (β_1 *is the intercept*). Assume that there are K large data sources, denoted as $\mathcal{F}_n = \{(Y_{ik}, \mathbf{X}_{ik}), i = 1, \cdots, n_k; k = 1, \cdots, K\}$, which are stored separately at different locations. Here the sample size of the kth data source n_k is much larger than the dimension of the covariate $\mathbf{X}$ ($n_k \gg p$), and $n = \sum_{k=1}^{K} n_k$ is the total sample size.

If we are able to analyze the full data $\mathcal{F}_n$, the ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{ik} - \boldsymbol{\beta}^T \mathbf{X}_{ik})^2. \tag{2.2}$$

It is straightforward to deduce an explicit expression of the OLS estimator with full data,

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Psi}, \tag{2.3}$$

where

$$\boldsymbol{\Gamma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{X}_{ik} \mathbf{X}_{ik}^T, \quad \text{and} \quad \boldsymbol{\Psi} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ik} \mathbf{X}_{ik}.$$

Due to storage capacity or network limitation, it is impractical to send raw data to a central server for statistical analysis. If the estimate of $\boldsymbol{\beta}$ is the only thing we care about, then sending some summary statistics from each data locations would be sufficient. However, practical data analysis seldom ends with an estimate. For example, regression diagnostic is an integrated part of regression analysis, and it requires to access the raw data. In addition, calculating $\sum_{i=1}^{n_k} \mathbf{X}_{ik} \mathbf{X}_{ik}^T$ still takes $O(n_k p^2)$ CPU time on each data location. Our aim is to select a subdata with subsample size $r \ll n$ from the K data sets, where each data set $\{(Y_{ik}, \mathbf{X}_{ik}), i = 1, \cdots, n_k\}$ provides r_k observations. In Algorithm 1, we present our distributed subsampling procedure towards multiple large-scale data sets.

To establish asymptotic properties of the subsample-based estimators, we need the following regularity conditions. Here we point out that we allow π_{ik} and r_k to be dependent on the responses, hence they may be random.

Condition (C.1). *As $n \to \infty$, $\boldsymbol{\Gamma}$ goes to a positive-definite matrix, where $\boldsymbol{\Gamma}$ is defined in (2.3).*

Condition (C.2). $\frac{1}{n^2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^4}{r_k \pi_{ik}} = O_P\Big(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \Big)$ and $\frac{1}{n^2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{Y_{ik}^2 \|\mathbf{X}_{ik}\|^2}{r_k \pi_{ik}} = O_P\Big(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \Big)$, where $\|\cdot\|$ is the Euclidean norm.

Condition (C.3). $\frac{1}{n^3} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{Y_{ik}^3 \|\mathbf{X}_{ik}\|^3}{r_k^2 \pi_{ik}^2} = o_P(1)$ and $\frac{1}{n^3} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^6}{r_k^2 \pi_{ik}^2} = o_P(1)$.

Condition (C.1) is a common assumption in linear regression models; Conditions (C.2) and (C.3) are used to determine the convergence rate of $\tilde{\boldsymbol{\beta}}$, together with its asymptotic distribution. These regularity conditions are mild and can be satisfied in many practical situations. For example, if we consider the uniform sampling with $n_k = n/K$, $r_k = r/K$ and $\pi_{ik} = 1/r_k$, then the condition $\frac{1}{n^2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^4}{r_k \pi_{ik}} = O_P\Big(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \Big)$ is ensured by the moment restriction with $\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^4 = O(1)$.

Below we establish the convergence rate, together with the asymptotic normality of $\tilde{\boldsymbol{\beta}}$, which play an important role in determining the optimal subsampling probabilities π_{ik} and optimal allocation sizes r_k for $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$.

Algorithm 1 Distributed Subsampling Algorithm

• *Sampling*: Assign subsampling probabilities π_{ik}, $i = 1, \cdots, n_k$ for the kth data source $\{(Y_{ik}, \mathbf{X}_{ik}), i = 1, \cdots, n_k\}$ with $\sum_{i=1}^{n_k} \pi_{ik} = 1$, where $k = 1, \cdots, K$. Given total sampling size r, draw a random subsample of size r_k from the kth data source according to $\{\pi_{ik}; i = 1, \cdots, n_k\}$, where r_k is the allocation size with $\sum_{k=1}^{K} r_k = r$. We denote the corresponding responses, covariates, and subsampling probabilities as Y_{ik}^*, $\mathbf{X}_{ik}^*$ and π_{ik}^*, respectively, $i = 1, \cdots, r_k$, and $k = 1, \cdots, K$.

• *Estimation*: Minimize the following weighted least squares criterion function $S^*(\boldsymbol{\beta})$ to get an estimate $\tilde{\boldsymbol{\beta}}$ based on the subsample $\{(Y_{ik}^*, \mathbf{X}_{ik}^*), i = 1, \cdots, r_k; k = 1, \cdots, K\}$,

$$S^*(\boldsymbol{\beta}) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n r_k} \left\{ \sum_{i=1}^{r_k} \frac{1}{\pi_{ik}^*} (Y_{ik}^* - \boldsymbol{\beta}^T \mathbf{X}_{ik}^*)^2 \right\}. \tag{2.4}$$

By solving $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S^*(\boldsymbol{\beta})$, we get a weighted least squares estimator with closed-form $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^{*-1} \boldsymbol{\Psi}^*$, where

$$\boldsymbol{\Gamma}^* = \sum_{k=1}^{K} \sum_{i=1}^{r_k} \frac{1}{n r_k \pi_{ik}^*} \mathbf{X}_{ik}^* (\mathbf{X}_{ik}^*)^T, \quad \text{and} \quad \boldsymbol{\Psi}^* = \sum_{k=1}^{K} \sum_{i=1}^{r_k} \frac{1}{n r_k \pi_{ik}^*} Y_{ik}^* \mathbf{X}_{ik}^*. \tag{2.5}$$

Theorem 1. *If Conditions (C.1) - (C.3) hold, and suppose that* $n^{-2} \sum_{k=1}^{K} n_k^2 / r_k = o(1)$*, then with probability approaching one, for any* $\delta > 0$*, there exists a finite* $\Delta_\delta > 0$ *such that*

$$P\left(||\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}|| \geq \left\{ \sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \right\}^{1/2} \Delta_\delta \,\middle|\, \mathcal{F}_n \right) < \delta. \tag{2.6}$$

Moreover, conditional on $\mathcal{F}_n$ *in probability, as* $r \to \infty$ *and* $n \to \infty$ *we have*

$$\boldsymbol{\Sigma}^{-1/2} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{2.7}$$

where $\xrightarrow{d}$ *denotes convergence in distribution,*

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Phi} \boldsymbol{\Gamma}^{-1} = O_{P|\mathcal{F}_n} \left(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \right) \tag{2.8}$$

and

$$\boldsymbol{\Phi} = \sum_{k=1}^{K} \frac{1}{n^2 r_k} \sum_{i=1}^{n_k} \frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \mathbf{X}_{ik} \mathbf{X}_{ik}^T}{\pi_{ik}}. \tag{2.9}$$

Remark 1. *The condition* $n^{-2} \sum_{k=1}^{K} n_k^2 / r_k = o(1)$ *indicates that* K *can go to infinity, and the consistency of* $\tilde{\boldsymbol{\beta}}$ *still holds in this situation. This condition is very reasonable. For example, with the uniform sampling where* $n_k = n/K$ *and* $r_k = r/K$*, we know that* $n^{-2} \sum_{k=1}^{K} n_k^2 / r_k = 1/r$ *goes to zero as* $r \to \infty$*.*

3. Optimal Subsampling Criterion

Given the subsampling size r, we need to specify the subsampling probabilities π_{ik} together with the allocation sizes r_k in Algorithm 1. An easy choice is to adopt the uniform subsampling with $\pi_{ik} = \{n_k^{-1}\}_{i=1}^{n_k}$

and $r_k = [r/K]$, for $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$, where $[\cdot]$ denotes the rounding operation to the closest integer. However, this uniform subsampling method may not be optimal, and it is more suitable to propose a nonuniform subsampling procedure for better performance. As suggested by Wang et al. (2018b), we can determine the optimal subsampling probabilities and optimal allocation sizes by "minimizing" the asymptotic variance-covariance matrix $\mathbf{\Sigma}$ in (2.8). Here we adopt the idea of $\mathcal{A}$-optimality from Kiefer (1959) to define the "minimizing" of matrix $\mathbf{\Sigma}$. Specifically, this is to minimize the trace of covariance matrix, $\mathrm{tr}(\mathbf{\Sigma})$, to derive the optimal sampling strategy. However, the calculation burden of $\mathbf{\Gamma}^{-1}$ is heavy. For two positive definite matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we define the partial ordering as $\mathbf{M}_1 \geq \mathbf{M}_2$ if and only if $\mathbf{M}_1 - \mathbf{M}_2$ is a nonnegative definite matrix. This definition is called the Loewner-ordering. Note that the covariance matrix $\mathbf{\Sigma}$ in (2.8) depends on π_{ik} and r_k through $\mathbf{\Phi}$ in (2.9), while $\mathbf{\Gamma}$ does not involve π_{ik} and r_k, $i = 1, \cdots, n_k$, $k = 1, \cdots, K$. Moreover, for two given $\mathcal{S}^{(1)} \triangleq \{(\pi_{ik}^{(1)}, r_k^{(1)})|\ i = 1, \cdots, n_k; k = 1, \cdots, K\}$ and $\mathcal{S}^{(2)} \triangleq \{(\pi_{ik}^{(2)}, r_k^{(2)})|\ i = 1, \cdots, n_k; k = 1, \cdots, K\}$, it can be checked that $\mathbf{\Sigma}(\mathcal{S}^{(1)}) \geq \mathbf{\Sigma}(\mathcal{S}^{(2)})$ if and only if $\mathbf{\Phi}(\mathcal{S}^{(1)}) \geq \mathbf{\Phi}(\mathcal{S}^{(2)})$. From this point of view, we only focus on minimizing $\mathrm{tr}(\mathbf{\Phi})$ to specify the optimal subsampling rule (Wang et al., 2018b). The following theorem presents the explicit expressions for optimal subsampling probabilities and optimal allocation sizes.

Theorem 2. *In Algorithm 1, the optimal subsampling probabilities (OSP) minimizing $tr(\mathbf{\Phi})$ are*

$$\pi_{ik}^{OSP} = \frac{|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\|}{\sum_{j=1}^{n_k} |Y_{jk} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{jk}| \|\mathbf{X}_{jk}\|}, \text{ for } i = 1, \cdots, n_k \text{ and } k = 1, \cdots, K. \tag{3.1}$$

For a given subsample size r, the optimal allocation sizes (OAS) r_k are

$$r_k^{OAS} = r \cdot \frac{\sum_{i=1}^{n_k} |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\|}{\sum_{\ell=1}^{K} \sum_{i=1}^{n_\ell} |Y_{i\ell} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{i\ell}| \|\mathbf{X}_{i\ell}\|}, \text{ for } k = 1, \cdots, K. \tag{3.2}$$

For practical implementation, we can use $[r_k^{OAS}]$ as the allocation sizes for $k = 1, \cdots, K$, where $[\,\cdot\,]$ denotes the rounding operation to the closest integer. Furthermore, we need to provide $\hat{\boldsymbol{\beta}}_{OLS}$ for the subsampling probabilities $\pi_{ik}^{OSP}$and allocation sizes r_k^{OAS}, where $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$. However, $\hat{\boldsymbol{\beta}}_{OLS}$ is unavailable beforehand. To deal with this problem, we can use a pilot estimator $\tilde{\boldsymbol{\beta}}_0$ to replace $\hat{\boldsymbol{\beta}}_{OLS}$. In addition, for those data points with Y_{ik} being close to $\hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}$, the corresponding subsampling probabilities are very small. To protect the weighted criterion function $S^*(\boldsymbol{\beta})$ from being inflated, we can truncate the $|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}|$ by $\max(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}|, c)$, where $c > 0$ is a specified low-bound, e.g. $c = 10^{-6}$. To summarize the above procedure, we present the implementation details in Algorithm 2.

Algorithm 2 first takes a pilot subsample of size r_0 and then selects a second subsample of size r. We do not recommend combining the two step subsamples for data analysis. The reason is that if we could conduct statistical analysis with size $r_0 + r$, then we could have had a better subsample by setting the second step sample size to $r_0 + r$ directly. Below we establish the consistency and asymptotic normality of $\breve{\boldsymbol{\beta}}$ based on the subsampling procedure in Algorithm 2. First we need the following regularity moment conditions:

Condition (C.4). $\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} |Y_{ik}|^6 = O_P(1)$ and $\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^6 = O(1)$.

Algorithm 2 Two-Step Strategy

Step 1.

- Take a subsample size r_0 by using uniform subsampling probability $\pi_{ik} = \{1/n_k\}_{i=1}^{n_k}$ to take $r_{0k} = [r_0/K]$ data points from each of the K distributed units. Send the r_0 data points to the central unit and obtain a pilot estimator $\tilde{\boldsymbol{\beta}}_0$.

- Send $\tilde{\boldsymbol{\beta}}_0$ to the K distributed units to calculate $u_{ik} = \max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|$, $U_{0k} = \sum_{i=1}^{n_k} u_{ik}$, and

$$\pi_{ik}(\tilde{\boldsymbol{\beta}}_0) = \frac{u_{ik}}{U_{0k}}, \text{ for } i = 1, \cdots, n_k \text{ and } k = 1, \cdots, K. \tag{3.3}$$

- Send U_{0k}, $k = 1, ..., K$, to the central unit to calculate

$$r_k(\tilde{\boldsymbol{\beta}}_0) = \frac{rU_{0k}}{\sum_{j=1}^{K} U_{0j}}, \text{ for } k = 1, \cdots, K. \tag{3.4}$$

  (If the distributed units can communicate, then U_{0k}'s are shared by all units and each $r_k(\tilde{\boldsymbol{\beta}}_0)$ is calculated in each distributed unit.)

- Send $r_k(\tilde{\boldsymbol{\beta}}_0)$'s to the corresponding K distributed units.

Step 2.

- Draw a subsample of size $r_k(\tilde{\boldsymbol{\beta}}_0)$ with replacement using the subsampling probabilities $\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)$ for $i = 1, \cdots, n_k$ from the k-th unit for $k = 1, \cdots, K$.

- Denote the selected subsamples and associated probabilities as $\{(Y_{ik}^*, \mathbf{X}_{ik}^*, \pi_{ik}^*),\ i = 1, \cdots, r_{0k},\ k = 1, \cdots, K\}$. Send them to the central unit. Let

$$S_{\tilde{\boldsymbol{\beta}}_0}^*(\boldsymbol{\beta}) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{\pi_{ik}^*(\tilde{\boldsymbol{\beta}}_0)} (Y_{ik}^* - \boldsymbol{\beta}^T \mathbf{X}_{ik}^*)^2. \tag{3.5}$$

  An explicit expression of the subsample-based estimator is $\breve{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0)\boldsymbol{\Psi}^*(\tilde{\boldsymbol{\beta}}_0)$, where

$$\boldsymbol{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0) = \sum_{k=1}^{K} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{n r_k(\tilde{\boldsymbol{\beta}}_0) \pi_{ik}^*(\tilde{\boldsymbol{\beta}}_0)} \mathbf{X}_{ik}^* (\mathbf{X}_{ik}^*)^T, \text{ and } \boldsymbol{\Psi}^*(\tilde{\boldsymbol{\beta}}_0) = \sum_{k=1}^{K} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{n r_k(\tilde{\boldsymbol{\beta}}_0) \pi_{ik}^*(\tilde{\boldsymbol{\beta}}_0)} Y_{ik}^* \mathbf{X}_{ik}^*.$$

- Perform necessary regression diagnostics based on the selected subsample.

Theorem 3. *Under Conditions (C.1) and (C.4), for any $\delta > 0$ there exists a finite $\Delta_\delta > 0$ such that with probability approaching one,*

$$P\left(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}\| \geq r^{-1/2}\Delta_\delta \,\middle|\, \mathcal{F}_n\right) < \delta. \tag{3.6}$$

Moreover, if $r_0 \to \infty$ and $r \to \infty$, then conditional on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we have

$$\boldsymbol{\Sigma}_{opt}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{3.7}$$

where $\boldsymbol{\Sigma}_{opt} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{opt}\boldsymbol{\Gamma}^{-1}$ with

$$\boldsymbol{\Phi}_{opt} = \frac{1}{rn^2}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\max(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|}\right\}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\max(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|\right\},$$

and c is a pre-specified truncation value, e.g. $c = 10^{-6}$.

Remark 2. *From the expressions of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}_{opt}$, we know that they do not depend on the value of K, i.e., $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}_{opt}$ remain the same for different values of K. Thus, the limiting distribution of the subsample-based estimator $\breve{\boldsymbol{\beta}}$ remains the same if all data are stored in one location ($K = 1$). There is no statistical efficiency loss due to distributed computing.*

From the view of statistical application, it is desirable to provide an estimator for the covariance matrix $\boldsymbol{\Sigma}_{opt}$. A straightforward method is to replace $\hat{\boldsymbol{\beta}}_{OLS}$ with $\breve{\boldsymbol{\beta}}$ in the asymptotic variance-covariance matrix in Theorem 3. However, this approach requires calculations based on the full data. We propose to estimate the variance-covariance matrix $\boldsymbol{\Sigma}_{opt}$ using a subsample,

$$\breve{\boldsymbol{\Sigma}}_{opt} = \breve{\boldsymbol{\Gamma}}^{-1}\breve{\boldsymbol{\Phi}}_{opt}\breve{\boldsymbol{\Gamma}}^{-1}, \tag{3.8}$$

where

$$\breve{\boldsymbol{\Gamma}} = \sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{nr_k\pi_{ik}^*}\mathbf{X}_{ik}^*(\mathbf{X}_{ik}^*)^T,$$

and

$$\breve{\boldsymbol{\Phi}}_{opt} = \sum_{k=1}^{K}\frac{1}{n^2r_k^2}\sum_{i=1}^{r_k}\frac{(Y_{ik}^* - \breve{\boldsymbol{\beta}}^T\mathbf{X}_{ik}^*)^2\mathbf{X}_{ik}^*(\mathbf{X}_{ik}^*)^T}{(\pi_{ik}^*)^2}.$$

The above $\breve{\boldsymbol{\Gamma}}$ and $\breve{\boldsymbol{\Phi}}_{opt}$ are directly motivated from the method of moments. Note that $E(\breve{\boldsymbol{\Gamma}}|\mathcal{F}_n) = \boldsymbol{\Gamma}$, and $E(\breve{\boldsymbol{\Phi}}_{opt}|\mathcal{F}_n) = \boldsymbol{\Phi}_{opt}$ if we replace $\breve{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{OLS}$ in $\breve{\boldsymbol{\Phi}}_{opt}$. We will check the performance of the estimated variance-covariance matrix in (3.8) via numerical simulation in next section.

4. Numerical Simulation

In this section, we conduct simulations to check the rationality of our proposed method. For convenience, we choose the sample size of each data set as $n_k = n/K$ for $k = 1, \cdots, K$. The number of data sets is

$K = 2$ and 100, respectively. The vector of regression coefficients is $\boldsymbol{\beta} = (0.5, \cdots, 0.5)'$ with $p = 5$ and 50, respectively. Denote $\mathbf{X} = (1, \tilde{\mathbf{X}}')'$ with $\tilde{\mathbf{X}} = (X_2, \cdots, X_p)'$. Let $\boldsymbol{\Sigma}_X = (0.5^{|i-j|})$ be a covariance matrix with $1 \le i, j \le p-1$. We consider the following three cases for the generation of covariate $\tilde{\mathbf{X}}$:

Case I: The $\tilde{\mathbf{X}}$ has a multivariate normal distribution, that is, $\tilde{\mathbf{X}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$.

Case II: The $\tilde{\mathbf{X}}$ has a multivariate lognormal distribution, that is, $\tilde{\mathbf{X}} \sim \mathrm{LN}(\mathbf{0}, \boldsymbol{\Sigma}_X)$.

Case III: The $\tilde{\mathbf{X}}$ has a multivariate t distribution with degrees of freedom $v = 2$, that is, $\tilde{\mathbf{X}} \sim t_2(\mathbf{0}, \boldsymbol{\Sigma}_X)$. We generate the error term ϵ from $N(0, 1)$. The r_0 in Step 1 of Algorithm 2 is chosen as $r_0 = 500$, and the subsample size $r = 500$, 800 and 1000, respectively. All the simulation results in Tables 1 and 2, together with Figures 4.2 and 4.3 are based on 1000 replications with total sample size $n = 10^6$.

In Tables 1 and 2, we present the results of $\breve{\beta}_1$, which include the estimated biases (BIAS) given by the sample mean of the estimates minus the OLS estimator in (2.2), the sampling standard error (SE) of the estimates, the sample mean of the estimated standard errors (ESE), and the empirical 95% coverage probabilities (CP) based on normal approximation. From the results in Tables 1 and 2, we can see that the subsample-based estimators are unbiased, the estimated and empirical standard errors are similar, and the coverage probabilities of the 95% confidence intervals are satisfactory. Furthermore, the performances of subsample-based estimators become better as r increases. Results for other components of $\breve{\boldsymbol{\beta}}$ are similar and thus are omitted.

Table 1: Simulation results on the two-step subsample estimator $\breve{\beta}_1$ with $K = 2$†.

		$p = 5$				$p = 50$			
	r	BIAS	SE	ESE	CP	BIAS	SE	ESE	CP
Case I	500	−0.0021	0.0385	0.0389	0.958	−0.0067	0.0470	0.0435	0.927
	800	−0.0008	0.0298	0.0305	0.960	−0.0053	0.0349	0.0335	0.933
	1000	−0.0014	0.0281	0.0276	0.949	0.0002	0.0299	0.0297	0.945
Case II	500	0.0051	0.0740	0.0742	0.952	0.0115	0.1789	0.1728	0.942
	800	−0.0022	0.0576	0.0584	0.955	0.0007	0.1382	0.1304	0.931
	1000	0.0016	0.0522	0.0520	0.954	0.0016	0.1219	0.1172	0.940
Case III	500	−0.0024	0.0470	0.0469	0.948	0.0043	0.0696	0.0674	0.939
	800	−0.0011	0.0381	0.0377	0.949	−0.0027	0.0581	0.0556	0.941
	1000	−0.0002	0.0380	0.0373	0.945	−0.0006	0.0490	0.0477	0.947

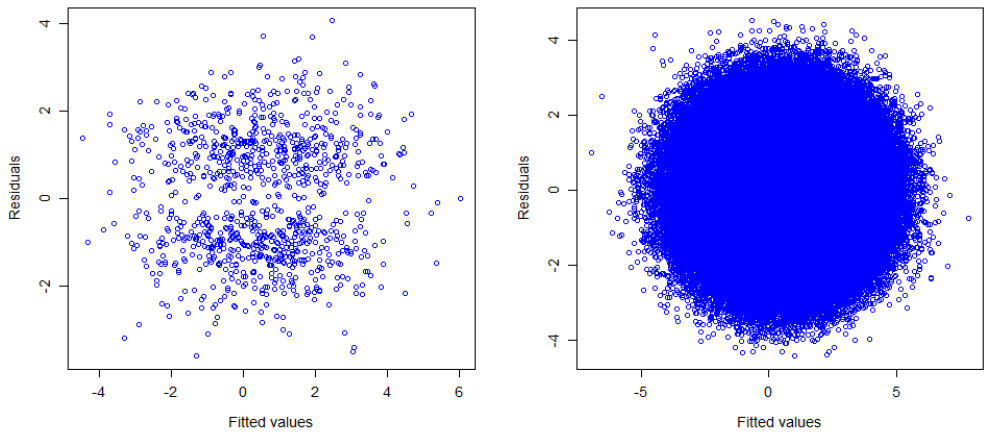
†"BIAS" denotes the biases of subsample estimates; " SE" denotes the sampling standard error of the estimates; "ESE" denotes the sample mean of the estimated standard errors; "CP" denotes the empirical 95% coverage probabilities.

Table 2: Simulation results on the two-step subsample estimator $\breve{\beta}_1$ with $K = 100^{\dagger}$.

		$p = 5$				$p = 50$			
	r	BIAS	SE	ESE	CP	BIAS	SE	ESE	CP
Case I	500	−0.0021	0.0397	0.0388	0.941	−0.0036	0.0470	0.0434	0.928
	800	0.0005	0.0305	0.0306	0.956	−0.0046	0.0357	0.0337	0.932
	1000	0.0006	0.0269	0.0272	0.954	0.0014	0.0317	0.0301	0.931
Case II	500	0.0002	0.0720	0.0745	0.966	0.0212	0.1826	0.1706	0.938
	800	0.0025	0.0564	0.0583	0.956	−0.0207	0.1395	0.1319	0.925
	1000	−0.0002	0.0537	0.0528	0.943	0.0068	0.1199	0.1169	0.942
Case III	500	−0.0030	0.0525	0.0501	0.942	0.0032	0.0870	0.0829	0.939
	800	0.0014	0.0390	0.0376	0.939	0.0006	0.0590	0.0566	0.940
	1000	−0.0026	0.0344	0.0337	0.942	0.0062	0.0574	0.0560	0.948

†"BIAS" denotes the biases of subsample estimates; " SE" denotes the sampling standard error of the estimates; "ESE" denotes the sample mean of the estimated standard errors; "CP" denotes the empirical 95% coverage probabilities.

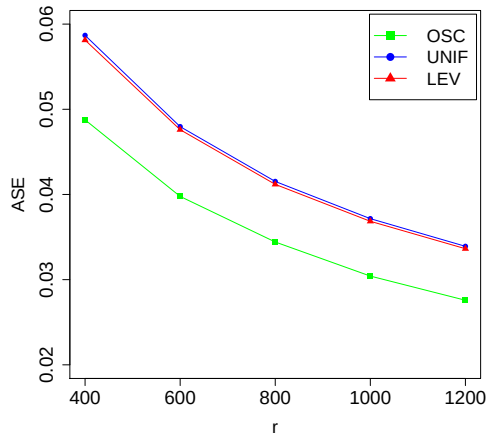
In addition to parameter estimates, we also consider the usefulness of a subsample in regression diagnostics. Figure 4.1 is a residual plot with residuals vs. fitted values for Case I, where $K = 2$, $p = 5$ and $r = 1000$. The residual plot from the subsample has a similar pattern with that from full data, which reveals potential usefulness of the OSC-based subsample in regression diagnostics.



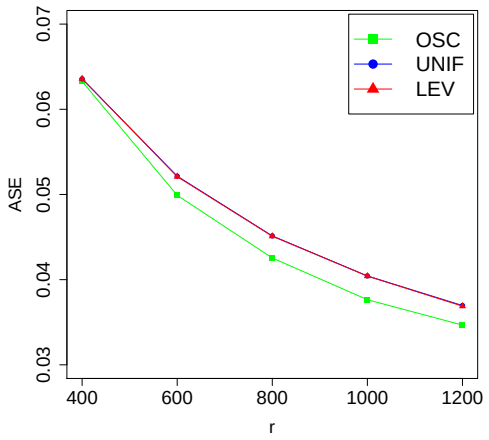
(a) Fitted values vs. residuals with subsample.

(b) Fitted values vs. residuals with full data.

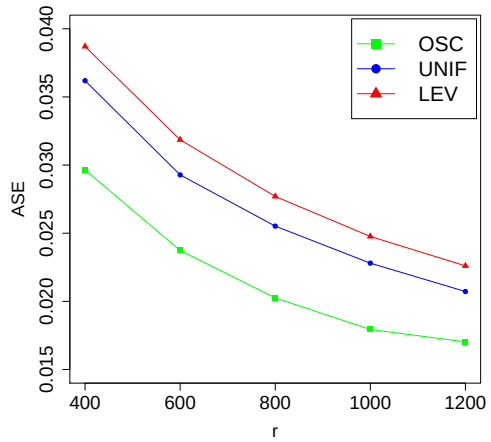
Figure 4.1: Residual plots with $K = 2$, $p = 5$ and $r = 1000$ (Case I).



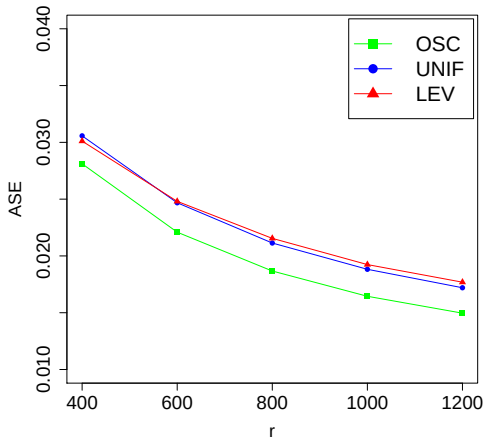
(a) Case I with $K = 2$, $p = 5$.



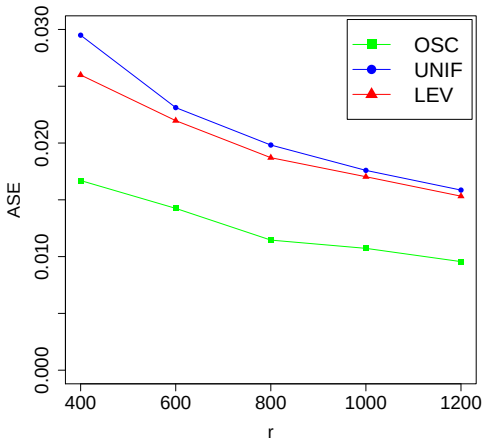
(b) Case I with $K = 100$, $p = 50$.



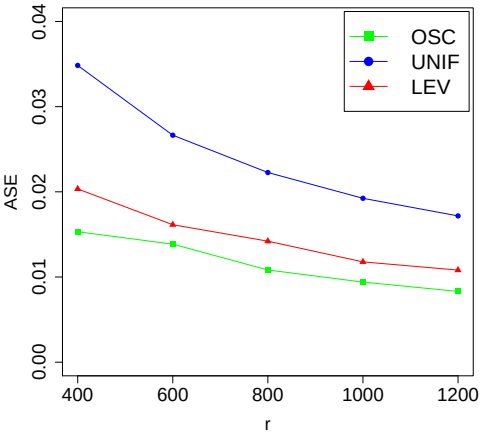
(c) Case II with $K = 2$, $p = 5$.



(d) Case II with $K = 100$, $p = 50$.

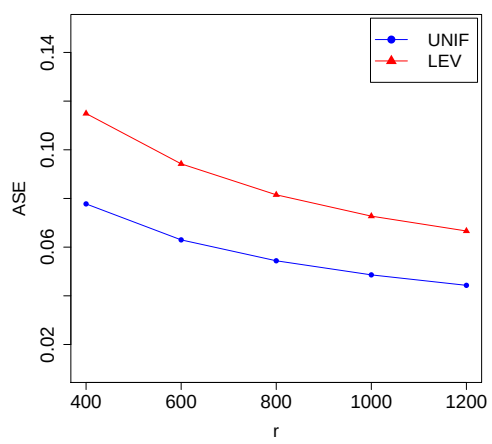


(e) Case III with $K = 2$, $p = 5$.

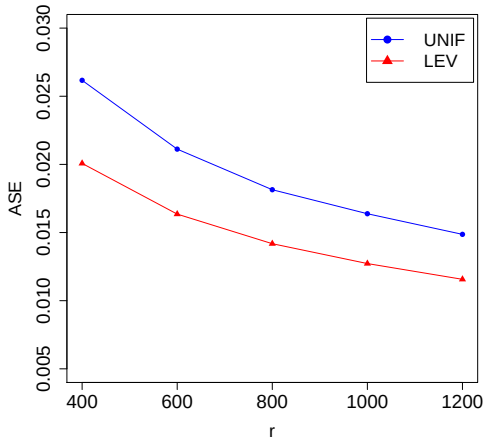


(f) Case III with $K = 100$, $p = 50$.

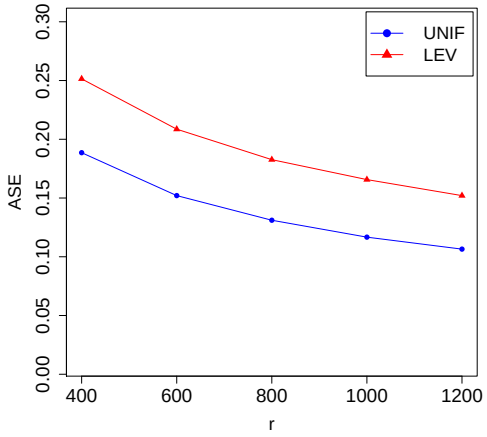
Figure 4.2: The ASEs for different subsampling methods.



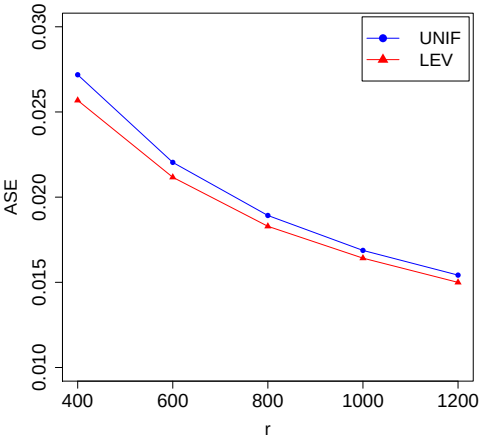
(a) Intercept with $K = 2$, $p = 5$.



(b) Slope with $K = 2$, $p = 5$.



(c) Intercept with $K = 100$, $p = 50$.



(d) Slope with $K = 100$, $p = 50$.

Figure 4.3: ASEs of UNIF and LEV for estimating the intercept and slope parameters with Case II.

To investigate the superiority of our optimal subsampling criterion (OSC), we compare the OSC with uniform subsampling (UNIF), whose subsampling probabilities $\pi_{ik} = 1/n_k$ and allocation sizes $r_k = [r/K]$, for $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$. Moreover, we also consider the leverage-based subsampling (LEV; Ma, et al., 2015). To ensoure the best performance of the LEV method in parameter estimation, exact full data statistical leverage scores are used to derive the subsampling probabilities, i.e., $\pi_{ik} = h_{ik}/\sum_{j=1}^{n_k} h_{jk}$ for $i = 1, ..., n_k$, where $h_{ik} = \mathbf{X}_{ik}^T(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_{ik}$ and $\mathbf{X}_{full}$ is the full data design matrix. The allocation sizes are taken as $r_k = [r/K]$, $k = 1, \cdots, K$. Let SE_{lj} denote the estimated standard error for $\breve{\beta}_j$ in the lth repetition of the simulation, and define $SE_l = \frac{1}{p}\sum_{j=1}^{p} SE_{lj}$. Here we calculate the average of SE_l based on 1000 repetition of the simulation i.e., ASE $= \sum_{l=1}^{1000} SE_l/1000$. For $p = 5$ and 50, the results are

presented in Figure 4.2. It is seen that the ASEs of OSC are much smaller than those of the UNIF and LEV methods. The performances of LEV are better than UNIF in Cases I and III, while the LEV has a much larger ASE than UNIF in Case II. In Case II the covariate distribution is skewed, and we see that the UNIF out performs the LEV. To better understand this unexpected performance of the LEV method with asymmetrically distributed covariates, we calculated ASE's of the intercept and slope estimators separately. Figure 4.3 shows the results for the LEV and UNIF methods in Case II. We see that the LEV has a much larger ASE than the UNIF towards the estimation of the intercept, which leads to the corresponding results in Case II of Figure 4.2.

We also use simulation to evaluate the computation efficiency of our method. First we generate data using the same mechanism as the above situation with Case I. All computations are carried out on a laptop running R software with 16GB random-access memory (RAM). Given a pilot subsample size r_0, we select $r_{0k} = [r_0/K]$ data points from each of the K data units and send them back to the central place to obtain a pilot estimator $\tilde{\boldsymbol{\beta}}_0$. The pilot $\tilde{\boldsymbol{\beta}}_0$ is then send to all data units to calculate U_{0k} and $\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)$ as in (3.3). Only the K scalars U_{0k}'s need to be send to the central unit to calculate scalars $r_k(\tilde{\boldsymbol{\beta}}_0)$'s as in (3.4), which are then send to distributed units as subsample sizes. From each data unit, $r_k(\tilde{\boldsymbol{\beta}}_0)$ data points are selected and they are sent to the central unit along with their associate $\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)$'s for final data analysis. In Table 3, we report the required CPU times (in seconds) to obtain $\breve{\boldsymbol{\beta}}$ with $K = 2$, $r_0 = 1000$, $p = 5$, 50, 300 and 500, where Algorithm 2 is implemented on a single core. Of note, these times are CPU times for implementing each method in the RAM, while the times to generate data are not counted. Moreover, these times are the mean CPU times of 10 repetitions. For the LEV method, the leverage scores are approximated using the fast algorithm in Drineas et al. (2012). The computing times for the UNIF and the full data methods are also reported for comparison. For the full data method, $\boldsymbol{\Gamma}_k$'s and $\boldsymbol{\Psi}_k$'s for $k = 1, \cdots, K$, are calculated from each data block, and then the estimator is calculated as $\hat{\boldsymbol{\beta}}_{OLS} = (\sum_{k=1}^{K} \boldsymbol{\Gamma}_k)^{-1} \sum_{k=1}^{K} \boldsymbol{\Psi}_k$. That is to say the CPU times for full data method mainly consist of the calculation of the summary statistics, and the communication cost is negligible. It is seen from the results that the UNIF is much faster than the other methods. The reason is that there is no need for UNIF to calculate subsampling probabilities and allocation sizes, so it requires less RAM and CPU times as well. Our proposed OSC method has significant computational advantages over the LEV and full data methods. In Table 4, we show the comparisons of ASEs for OSC and UNIF methods when the computation times are similar in Case III, where $K = 2$ and $r_0 = 1000$. The results indicate that the OSC and UNIF may have similar estimation efficiency with similar CPU times. However, since UNIF uses larger sample sizes, it requires larger memory. Hence, our proposed method achieves the same estimation efficiency with less computing resources.

Table 3: The CPU times of subsampling methods with $r = 1000$ (seconds)†.

	Method	$n = 3 \times 10^5$	$n = 5 \times 10^5$	10^6
p=5	OSC	0.043	0.068	0.129
	UNIF	0.008	0.016	0.033
	LEV	0.739	1.288	3.462
	Full data	0.061	0.111	0.255
p=50	OSC	0.180	0.325	0.610
	UNIF	0.016	0.018	0.022
	LEV	1.452	2.469	4.963
	Full data	1.200	2.134	3.809
p=300	OSC	0.965	1.403	5.293
	UNIF	0.061	0.062	0.161
	LEV	3.906	6.593	21.865
	Full data	21.037	32.589	85.873
p=500	OSC	1.827	2.359	5.934
	UNIF	0.180	0.181	0.203
	LEV	6.562	10.636	32.853
	Full data	53.996	85.730	178.572

† "OSC" denotes our optimal subsampling criterion;"UNIF" denotes uniform sampling; "LEV" denotes leverage-based subsampling.

Table 4: Comparisons of CPU times between OSC and UNIF with Case III (in seconds)†.

		$n = 3 \times 10^5$			$n = 5 \times 10^5$		
		CPU	r	ASE	CPU	r	ASE
$p = 5$	OSC	0.034	17500	0.00246	0.069	27500	0.00192
	UNIF	0.033	45000	0.00234	0.059	75000	0.00172
$p = 50$	OSC	0.204	17500	0.00209	0.278	15000	0.00214
	UNIF	0.211	35000	0.00241	0.236	35000	0.00236
$p = 300$	OSC	1.348	5500	0.00695	1.762	7500	0.00552
	UNIF	1.210	7500	0.00680	1.580	9700	0.00579
$p = 500$	OSC	3.922	6000	0.00784	5.186	7000	0.00663
	UNIF	4.086	7500	0.00736	4.747	9000	0.00651

† "OSC" denotes the optimal subsampling criterion;"UNIF" denotes uniform sampling; "CPU" denotes computation time.

5. Application

We apply our proposed method to a real data example bout the USA airline (DVN, 2008). The data consist of flight details for all commercial flights within the USA from October 1987 to April 2008. Specifically, the flights information are separately deposited in 22 files by years, where the raw dataset is as large as 12 GB on a hard drive. The response variable Y_{ik} denotes the arrival delay time of an airline (continuous, in minutes). The covariates $\mathbf{X}_{ik} = (X_{ik1}, X_{ik2})^T$ are departure delay times (continuous, in minutes) and distances (continuous, in thousands of miles), respectively. In Table 5, we report the number of flights with full informations about Y_{ik} and $\mathbf{X}_{ik}$ by years, where the total sample size is $n = \sum_{k=1}^{K} n_k = 120748239$ with $K = 22$.

Table 5: The yearly airline data and allocation sizes ($r = 1000$)†.

Years	n_k	r_k	Years	n_k	r_k
1987	1287333	10	1998	5227051	70
1988	5126498	30	1999	5360018	74
1989	4925482	30	2000	5481303	55
1990	5110527	31	2001	5723673	34
1991	4995005	31	2002	5197860	28
1992	5020651	36	2003	6375689	37
1993	4993587	38	2004	6987729	47
1994	5078411	41	2005	6992838	55
1995	5219140	51	2006	7003802	59
1996	5209326	52	2007	7275288	65
1997	5301999	67	2008	6855029	59

† The n_k and r_k denote the number of airline and optimal allocation size towards the kth year, respectively.

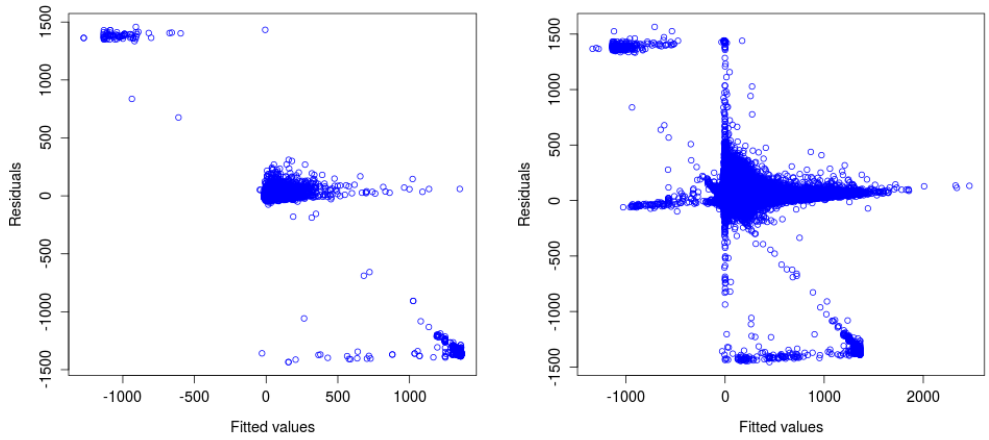
For this data, the full data OLS estimator is $\hat{\boldsymbol{\beta}}_{OLS} = (0.0945, 0.9484, -1.0531)^T$, where the first term $\hat{\beta}_1$ is the intercept estimate. We implement the OSC and UNIF estimators and calculate the SE and ESE based on 1000 subsamples with $r = 1000$, 3000 and 5000, respectively. As an example, we report the subsample sizes of OAS for different blocks with $r = 1000$ in Table 5. The results for the SE and ESE are given in 6. From Table 6, we see that the OSC-based estimator is unbiased, and its SE and ESE are close to each other. The overall performance of OSC-based method is much better than that of the UNIF method.

Table 6: Subsample estimator and (SE, ESE) for the airline data†.

	$\boldsymbol{\beta}$	OSC	UNIF
$r = 1000$	β_1	0.1166 (1.2541, 1.2489)	−0.1714 (1.1272, 0.7129)
	β_2	0.9479 (0.0108, 0.0108)	0.9940 (0.1377, 0.0294)
	β_3	−0.9984 (1.3725, 1.3897)	−1.1700 (1.0399, 0.9530)
$r = 3000$	β_1	0.1333 (0.7568, 0.7355)	−0.1310 (0.9320, 0.4988)
	β_2	0.9478 (0.0067, 0.0069)	0.9829 (0.1309, 0.0338)
	β_3	−1.0759 (0.8108, 0.7927)	−1.1082 (0.6425, 0.5673)
$r = 5000$	β_1	0.0876 (0.5606, 0.5451)	−0.0413 (0.8769, 0.4732)
	β_2	0.9482 (0.0050, 0.0051)	0.9709 (0.1259, 0.0409)
	β_3	−1.0527 (0.6043, 0.5973)	−1.0978 (0.4867, 0.4478)

† β_1 is an intercept; "OSC" denotes optimal subsampling criterion;"UNIF" denotes uniform subsampling.

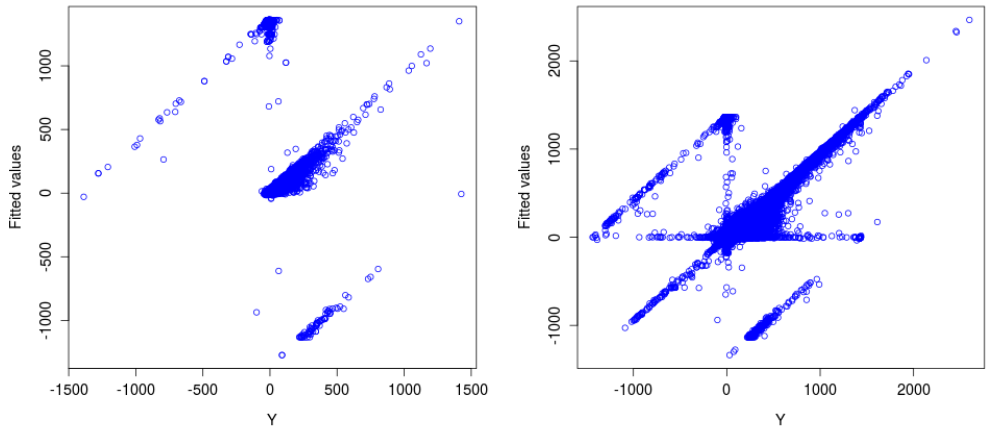
Finally, we use a subsample to create a residual plot to check the appropriateness of using a linear regression model on the data. For comparison, we also use a server with 128G RAM to create the same plot from the full data. We present the results in Figure 5.1, where OSC-based subsample size r= 5000. It is seen that there is some potential cluster pattern in the data, and the subsample residual plot also identifies this. This indicates that a more complicated model, such as a mixture model, may improve the goodness-of-fit. However, this is beyond the scope of the current paper and we will investigate it in a future project. In fact, the R square statistic for the whole data is $R^2 = 0.7607$ (calculated on the server), which is not very small, indicating that a linear regression is still a useful model for the data. The clustering patter for the data is more evident if we plot the fitted values against the observed responses as shown in Figure 5.2.



(a) Residual plot from a subsample of $r = 5000$.

(b) Residual plot from the full data.

Figure 5.1: Residual plots from an OSC-based subsample (r= 5000) and from the full data.



(a) Fitted values vs. observed responses in a subsample of $r = 5000$.

(b) Fitted values vs. observed responses in the full data.

Figure 5.2: Fitted values vs. observed responses in an OSC-based subsample (r= 5000) and in the full data.

6. Concluding Remarks

In this paper, we have proposed an optimal subsampling method for distributed and massive data in the context of linear models. The convergence rate and asymptotic normality of the subsample-based estimators were established. The optimal subsampling probabilities and optimal allocation sizes were provided. We have

also used numerical results to show that the OSC-based subsample selection procedure was more efficient than the UNIF-based method in parameter estimation. Simulations and a real data example have revealed the effectiveness of our method.

There are several topics to investigate in the future. First, we have illustrated the usefulness of OSC-based subsample in regression diagnostics by numerical studies. However, it is difficult to theoretically validate that the OSC-based subsample is optimal towards regression diagnostics. The optimal subsample selection for regression diagnostics with big data needs further research. Second, it is interesting to extend our distributed subsampling method to other models, such as the logistic regression of Wang et al. (2018b) and the generalized linear models in Ai et al. (2019). Third, a larger π_i indicates that the data point $(\mathbf{X}_i, Y_i)$ contains more information about $\boldsymbol{\beta}$, but it has a smaller weight in the object function (2.4). To improve contributions of those more informative points for parameter estimation, it is desirable to propose an un-weighted estimator as Wang (2019) in the framework of distributed subsampling.

Acknowledgements

The authors would like to thank the Editor, the Associate Editor and the reviewer for their constructive and insightful comments that greatly improved the manuscript.

7. Appendix

Below, we give the proofs for Theorems $1-3$. Note that from Cauchy-Schwarz inequality Condition (C.2) implies that $\frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{Y_{ik}\|\mathbf{X}_{ik}\|^3}{r_k\pi_{ik}} = O_P\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)$, and from Hölder's inequality, Condition (C.3) implies that $\frac{1}{n^3}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{Y_{ik}^2\|\mathbf{X}_{ik}\|^4}{r_k^2\pi_{ik}^2} = o_P(1)$ and $\frac{1}{n^3}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{Y_{ik}\|\mathbf{X}_{ik}\|^5}{r_k^2\pi_{ik}^2} = o_P(1)$. We first establish the following lemma.

Lemma 1. *If Conditions (C.1) and (C.2) hold, then conditionally on $\mathcal{F}_n$ we have*

$$\mathbf{\Gamma}^* - \mathbf{\Gamma} = O_{P|\mathcal{F}_n}\left(\left\{\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right\}^{1/2}\right), \tag{A.1}$$

$$\mathbf{\Psi}^* - \mathbf{\Psi} = O_{P|\mathcal{F}_n}\left(\left\{\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right\}^{1/2}\right), \tag{A.2}$$

and

$$\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}\left(\left\{\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right\}^{1/2}\right). \tag{A.3}$$

Proof. It is straightforward to deduce that $E(\mathbf{\Gamma}^*|\mathcal{F}_n) = \mathbf{\Gamma}$ holds. Moreover, for any component $\mathbf{\Gamma}^*_{j_1 j_2}$ of $\mathbf{\Gamma}^*$ with $1 \le j_1 \le j_2 \le p$, we have

$$\begin{aligned} Var(\mathbf{\Gamma}^*_{j_1 j_2}|\mathcal{F}_n) &= \sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\left\{\frac{1}{n_k^2}\sum_{i=1}^{n_k}\frac{1}{\pi_{ik}}\left(\mathbf{X}_{ikj_1}\mathbf{X}_{ikj_2}\right)^2 - \Big(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{ikj_1}\mathbf{X}_{ikj_2}\Big)^2\right\} \\ &\le \sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\left\{\frac{1}{n_k^2}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^4}{\pi_{ik}}\right\} \\ &= O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right), \end{aligned} \tag{A.4}$$

where the equality (A.4) holds by Condition (C.2). Then Markov's inequality leads to (A.1). Similarly, we know $E(\mathbf{\Psi}^*|\mathcal{F}_n) = \mathbf{\Psi}$. Besides, for any component $\mathbf{\Psi}^*_j$ of $\mathbf{\Psi}^*$ with $1 \le j \le p$,

$$\begin{aligned} Var(\mathbf{\Psi}^*_j|\mathcal{F}_n) &= E(\mathbf{\Psi}^*_j - \mathbf{\Psi}_j|\mathcal{F}_n)^2 \\ &= \sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\left\{\frac{1}{n_k^2}\sum_{i=1}^{n_k}\frac{1}{\pi_{ik}}Y_{ik}^2\mathbf{X}_{ikj}^2 - \left(\frac{1}{n_k}\sum_{i=1}^{n_k}Y_{ik}\mathbf{X}_{ikj}\right)^2\right\} \\ &\le \sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\left\{\frac{1}{n_k^2}\sum_{i=1}^{n_k}\frac{1}{\pi_{ik}}Y_{ik}^2\|\mathbf{X}_{ik}\|^2\right\} \\ &= O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right), \end{aligned}$$

where the last equality is due to Condition (C.2). Thus, (A.2) holds using the Markov's inequality.

Finally, direct calculation yields that

$$E\left(\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \,\middle|\, \mathcal{F}_n\right) = \frac{\partial S(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} = 0. \tag{A.5}$$

By Condition (C.2), for $j = 1, \cdots, p$, we can derive that

$$\begin{aligned} Var\left(\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}_j} \,\middle|\, \mathcal{F}_n\right) &= \sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\Big\{\sum_{i=1}^{n_k} \frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \mathbf{X}_{ikj}^2}{n_k^2 \pi_{ik}} - \Big[\frac{1}{n_k}\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ikj}\Big]^2\Big\} \\ &\le \sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \sum_{i=1}^{n_k} \frac{Y_{ik}^2 \|\mathbf{X}_{ik}\|^2 - 2Y_{ik}\|\mathbf{X}_{ik}\|^3 \|\hat{\boldsymbol{\beta}}_{OLS}\| + \|\hat{\boldsymbol{\beta}}_{OLS}\|^2 \|\mathbf{X}_{ik}\|^4}{n_k^2 \pi_{ik}} \\ &= O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\right). \end{aligned} \tag{A.6}$$

Using (A.5), (A.6) and Markov's inequality, (A.3) holds. This ends the proof. □

Proof of Theorem 1. Note that

$$\begin{aligned} \frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} &= -\sum_{k=1}^{K} \frac{1}{r_k}\left\{\sum_{i=1}^{r_k} \frac{(Y_{ik}^* - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}^*)\mathbf{X}_{ik}^*}{n\pi_{ik}^*}\right\} \\ &= -(\boldsymbol{\Psi}^* - \boldsymbol{\Gamma}^* \boldsymbol{\Gamma}^{-1} \boldsymbol{\Psi}). \end{aligned} \tag{A.7}$$

From (A.1) in Lemma 1,

$$\boldsymbol{\Gamma}^{*-1} - \boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{*-1}(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*)\boldsymbol{\Gamma}^{-1} = O_{P|\mathcal{F}_n}\left(\left\{\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\right\}^{1/2}\right). \tag{A.8}$$

After some direct calculation, by (A.3), (A.7) and (A.8) we can deduce the expression

$$\begin{aligned} \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS} &= \boldsymbol{\Gamma}^{*-1}\boldsymbol{\Psi}^* - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi} \\ &= -\boldsymbol{\Gamma}^{*-1}\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \\ &= -\boldsymbol{\Gamma}^{-1}\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} + (\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{*-1})\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \\ &= -\boldsymbol{\Gamma}^{-1}\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} + O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\right) \\ &= O_{P|\mathcal{F}_n}\left(\left\{\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\right\}^{1/2}\right). \end{aligned} \tag{A.9}$$

Thus, the convergence rate of $\tilde{\boldsymbol{\beta}}$ in (2.6) is established. In what follows, we need to establish the asymptotic distribution of $\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}}$. Since

$$\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} = -\sum_{k=1}^{K} \frac{1}{r_k}\left\{\sum_{i=1}^{r_k} \frac{(Y_{ik}^* - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}^*)\mathbf{X}_{ik}^*}{n\pi_{ik}^*}\right\}$$

$$= -\sum_{k=1}^{K}\sum_{i=1}^{r_k}\zeta_{ik}^*,$$

where $\zeta_{ik}^* = \frac{(Y_{ik}^* - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}^*)\mathbf{X}_{ik}^*}{nr_k\pi_{ik}^*}$, $i = 1, \cdots, r_k$, and $k = 1, \cdots, K$. Given $\mathcal{F}_n$, the ζ_{ik}^* are independent, and for every $\eta > 0$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{r_k} E\left\{\|\zeta_{ik}^*\|^2 I(\|\zeta_{ik}^*\| > \eta) \,\middle|\, \mathcal{F}_n\right\} \\
&\leq \sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{\eta}E\left\{\|\zeta_{ik}^*\|^3 I(\|\zeta_{ik}^*\| > \eta) \,\middle|\, \mathcal{F}_n\right\} \\
&\leq \sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{\eta}E\left\{\|\zeta_{ik}^*\|^3 \,\middle|\, \mathcal{F}_n\right\} \\
&= \sum_{k=1}^{K}\frac{1}{r_k^2\eta n^3}\sum_{i=1}^{n_k}\frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^3\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2} \\
&\leq \sum_{k=1}^{K}\frac{1}{n^3r_k^2}\left\{\frac{1}{\eta}\sum_{i=1}^{n_k}\frac{Y_{ik}^3\|\mathbf{X}_{ik}\|^3 - 3Y_{ik}^2\|\hat{\boldsymbol{\beta}}_{OLS}\|\|\mathbf{X}_{ik}\|^4 + 3Y_{ik}\|\hat{\boldsymbol{\beta}}_{OLS}\|^2\|\mathbf{X}_{ik}\|^5 - \|\hat{\boldsymbol{\beta}}_{OLS}\|^3\|\mathbf{X}_{ik}\|^6}{\pi_{ik}^2}\right\} \\
&= o_{P|\mathcal{F}_n}(1),
\end{aligned}$$

where the last equation is from Condition (C.3). Moreover, we have that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{i=1}^{r_k}\mathrm{Cov}(\zeta_{ik}^*|\mathcal{F}_n) &= \sum_{k=1}^{K}\frac{n_k^2}{n^2r_k}\Big\{\sum_{i=1}^{n_k}\frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{n_k^2\pi_{ik}} \\
&\qquad - \Big[\frac{1}{n_k}\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}\Big]\Big[\frac{1}{n_k}\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}^T\Big]\Big\} \\
&= \boldsymbol{\Phi} - \Delta_1, \qquad \text{(A.10)}
\end{aligned}$$

where

$$\Delta_1 = \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big[\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}\Big]\Big[\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}^T\Big]. \qquad \text{(A.11)}$$

For convenience, we denote $\mathbf{X}_k$ and $\mathbf{Y}_k$ as the design matrix and responses of kth data sets, $k = 1, \cdots, K$. Similarly, let $\mathbf{X}_{full}$ and $\mathbf{Y}_{full}$ be the design matrix and responses of full data. It is straightforward to deduce that

$$E\left[\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}\right] = E(\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{OLS}) = 0.$$

Hence,

$$E\left\{\left[\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}\right]\left[\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})\mathbf{X}_{ik}^T\right]\right\} = \mathrm{Cov}(\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{OLS}). \qquad \text{(A.12)}$$

Notice that

$$\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\sum_{\ell=1}^{K}\mathbf{X}_\ell^T\mathbf{Y}_\ell$$

$$= \underbrace{\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T\mathbf{Y}_k}_{R_1} - \underbrace{\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\sum_{\ell\neq k}^{K}\mathbf{X}_\ell^T\mathbf{Y}_\ell}_{R_2}.$$

In view of the following expressions,

$$\mathrm{Cov}(R_1) = \sigma^2\mathbf{X}_k^T\{\mathbf{I} - 2\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T + \mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T\}\mathbf{X}_k,$$

and

$$\mathrm{Cov}(R_2) = \sigma^2\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\sum_{\ell\neq k}^{K}\mathbf{X}_\ell^T\mathbf{X}_\ell(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T\mathbf{X}_k,$$

we have

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{OLS}) &= \sigma^2\{\mathbf{X}_k^T\mathbf{X}_k - \mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_k^T\mathbf{X}_k\} \\ &= \sigma^2\left\{\sum_{i=1}^{n_k}\mathbf{X}_{ik}\mathbf{X}_{ik}^T - \sum_{i=1}^{n_k}\mathbf{X}_{ik}\mathbf{X}_{ik}^T(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_{ik}\mathbf{X}_{ik}^T\right\}.\end{aligned} \tag{A.13}$$

Noting that $\mathbf{X}_{full}^T\mathbf{X}_{full}$ is positive-definite for large n and thus

$$\left\|\mathbf{X}_{ik}\mathbf{X}_{ik}^T(\mathbf{X}_{full}^T\mathbf{X}_{full})^{-1}\mathbf{X}_{ik}\mathbf{X}_{ik}^T\right\| \leq \left\|\mathbf{X}_{ik}\mathbf{X}_{ik}^T\right\|, \tag{A.14}$$

we know that

$$\left\|\mathrm{Cov}(\mathbf{X}_k^T\mathbf{Y}_k - \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{OLS})\right\| \leq 2\sigma^2\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^2. \tag{A.15}$$

Thus, from (A.11), (A.12) and (A.15),

$$E(\Delta_1) \leq \frac{2\sigma^2}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^2}{r_k} \leq \frac{2\sigma^2}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^4}{r_k} = o(1), \tag{A.16}$$

where the last step is from Condition (C.2). Since Δ_1 is positive semidefinite, we know that $\Delta_1 = o_p(1)$, and therefore (A.10) implies that

$$\sum_{k=1}^{K}\sum_{i=1}^{r_k}\mathrm{Cov}(\zeta_{ik}^*) = \boldsymbol{\Phi} + o_P(1).$$

From the Lindeberg-Feller central limit theorem in Proposition 2.27 of van der Vaart (1998) and Slutsky's theorem, conditionally on $\mathcal{F}_n$,

$$\boldsymbol{\Phi}^{-1/2}\frac{\partial S^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial\boldsymbol{\beta}} \xrightarrow{d} N(\mathbf{0},\mathbf{I}). \tag{A.17}$$

By Conditions (C.2) and (C.3), it can be proved that

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1} = O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right).$$

Therefore, (A.9), (A.17) and the Slutsky's theorem lead to the asymptotic distribution in (2.7). This ends the proof. □

Proof of Theorem 2. From (2.8) and (2.9), it can be calculated that

$$\begin{aligned} tr(\mathbf{\Phi}) &= \sum_{k=1}^{K} \frac{1}{r_k n^2} \sum_{i=1}^{n_k} \left[\frac{1}{\pi_{ik}} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \|\mathbf{X}_{ik}\|^2 \right] \\ &= \sum_{k=1}^{K} \frac{1}{r_k n^2} \cdot \left(\sum_{i=1}^{n_k} \pi_{ik} \right) \cdot \sum_{i=1}^{n_k} \left[\frac{1}{\pi_{ik}} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \|\mathbf{X}_{ik}\|^2 \right] \\ &\geq \sum_{k=1}^{K} \frac{1}{r_k n^2} \left[\sum_{i=1}^{n_k} |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\| \right]^2 \qquad \text{(A.18)} \\ &= \left(\sum_{k=1}^{K} \frac{r_k}{r} \right) \cdot \sum_{k=1}^{K} \frac{1}{r_k n^2} \left[\sum_{i=1}^{n_k} |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\| \right]^2 \\ &\geq \left\{ \sum_{k=1}^{K} \frac{1}{n r^{1/2}} \sum_{i=1}^{n_k} |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\| \right\}^2, \end{aligned}$$

where the inequality (A.18) is from the Cauchy-Schwarz inequality and the equality of it holds if and only if when π_{ik} are proportional to $|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\|$, namely $\pi_{ik} = C_1 |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\|$ for some constant $C_1 > 0$. Similarly, the last equality holds if and only if when $r_k = C_2 \sum_{i=1}^{n_k} |Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}| \|\mathbf{X}_{ik}\|$ for some $C_2 > 0$. This completes the proof. □

Proof of Theorem 3. First, we need to prove the following two conclusions

$$\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0) - \mathbf{\Gamma} = O_{P|\mathcal{F}_n}(r^{-1/2}), \qquad \text{(A.19)}$$

$$\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_n}(r^{-1/2}), \qquad \text{(A.20)}$$

where $\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0) = \sum_{k=1}^{K} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{n r_k(\tilde{\boldsymbol{\beta}}_0) \pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \mathbf{X}^*_{ik} (\mathbf{X}^*_{ik})^T$. By direct calculation,

$$E(\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n) = E_{\tilde{\boldsymbol{\beta}}_0}\{E(\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0)\} = E_{\tilde{\boldsymbol{\beta}}_0}(\mathbf{\Gamma}|\mathcal{F}_n) = \mathbf{\Gamma}. \qquad \text{(A.21)}$$

Here $E_{\tilde{\boldsymbol{\beta}}_0}$ denotes the expectation with respect to the distribution of $\tilde{\boldsymbol{\beta}}_0$ given $\mathcal{F}_n$. For any component $\mathbf{\Gamma}^*_{j_1 j_2}(\tilde{\boldsymbol{\beta}}_0)$ of $\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0)$ with $1 \leq j_1 \leq j_2 \leq p$,

$$\begin{aligned} Var(\mathbf{\Gamma}^*_{j_1 j_2}(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) &= \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \left\{ \frac{1}{n^2} \sum_{i=1}^{n_k} \frac{1}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} (\mathbf{X}_{ikj_1} \mathbf{X}_{ikj_2})^2 - \left(\sum_{i=1}^{n_k} \frac{1}{n} \mathbf{X}_{ikj_1} \mathbf{X}_{ikj_2} \right)^2 \right\} \\ &\leq \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \left\{ \frac{1}{n^2} \sum_{i=1}^{n_k} \frac{1}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} \|\mathbf{X}_{ik}\|^4 \right\} \\ &\leq \frac{1}{r} \left(\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^3}{c} \right) \left\{ \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}| + c) \|\mathbf{X}_{ik}\| \right\} \\ &= O_P(r^{-1}), \qquad \text{(A.22)} \end{aligned}$$

where the equality (A.22) is from Condition (C.4). Thus,

$$Var(\mathbf{\Gamma}^*_{j_1j_2}(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n) = E_{\tilde{\boldsymbol{\beta}}_0}\{Var(\mathbf{\Gamma}^*_{j_1j_2}(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0)\} = O_{P|\mathcal{F}_n}(r^{-1}). \tag{A.23}$$

From (A.21) and (A.23) together with Markov's inequality, (A.19) follows. Note that

$$\begin{aligned}\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} &= -\sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \cdot \left\{ \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{(Y^*_{ik} - \hat{\boldsymbol{\beta}}^T_{OLS}\mathbf{X}^*_{ik})\mathbf{X}^*_{ik}}{n\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \right\} \\ &= -[\mathbf{\Psi}^*(\tilde{\boldsymbol{\beta}}_0) - \mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0)\mathbf{\Gamma}^{-1}\mathbf{\Psi}].\end{aligned}$$

Similar to (A.21) and (A.23), by Condition (C.4) we can get that

$$E\left\{ \frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \,\middle|\, \mathcal{F}_n \right\} = 0, \quad \text{and} \quad Var\left\{ \frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \,\middle|\, \mathcal{F}_n \right\} = O_{P|\mathcal{F}_n}(r^{-1}). \tag{A.24}$$

Then, (A.20) holds from (A.24) and Markov's inequality.

Next, we begin to prove the convergence rate of $\breve{\boldsymbol{\beta}}$. From (A.19),

$$\begin{aligned}\mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0) &= \mathbf{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0)\{\mathbf{\Gamma}^*(\tilde{\boldsymbol{\beta}}_0) - \mathbf{\Gamma}\}\mathbf{\Gamma}^{-1} \\ &= O_{P|\mathcal{F}_n}(r^{-1/2}).\end{aligned} \tag{A.25}$$

By careful calculation, we can derive that

$$\begin{aligned}\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS} &= \mathbf{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0)\mathbf{\Psi}^*(\tilde{\boldsymbol{\beta}}_0) - \mathbf{\Gamma}^{-1}\mathbf{\Psi} \\ &= -\mathbf{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0)\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \\ &= -\mathbf{\Gamma}^{-1}\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} + \{\mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{*-1}(\tilde{\boldsymbol{\beta}}_0)\}\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} \\ &= -\mathbf{\Gamma}^{-1}\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} + O_{P|\mathcal{F}_n}\left(r^{-1}\right) \\ &= O_{P|\mathcal{F}_n}\left(r^{-1/2}\right).\end{aligned} \tag{A.26}$$

Lastly, we start to establish the asymptotic distribution of $\breve{\boldsymbol{\beta}}$. Note that

$$\begin{aligned}\frac{\partial S^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{OLS})}{\partial \boldsymbol{\beta}} &= -\sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \cdot \left\{ \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{(Y^*_{ik} - \hat{\boldsymbol{\beta}}^T_{OLS}\mathbf{X}^*_{ik})\mathbf{X}^*_{ik}}{n\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \right\} \\ &= -\sum_{k=1}^{K} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \zeta^*_{ik}(\tilde{\boldsymbol{\beta}}_0),\end{aligned}$$

where $\zeta^*_{ik}(\tilde{\boldsymbol{\beta}}_0) = \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \cdot \left\{ \frac{(Y^*_{ik} - \hat{\boldsymbol{\beta}}^T_{OLS}\mathbf{X}^*_{ik})\mathbf{X}^*_{ik}}{n\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \right\}$, for $i = 1, \cdots, n_k$, and $k = 1, \cdots, K$. Given $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, the $\zeta_{ik}(\tilde{\boldsymbol{\beta}}_0)$ are independent. For every $\eta > 0$,

$$\sum_{k=1}^{K} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} E\left\{ \|\zeta^*_{ik}(\tilde{\boldsymbol{\beta}}_0)\|^2 I(\|\zeta^*_{ik}(\tilde{\boldsymbol{\beta}}_0)\| > \eta) \,\middle|\, \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0 \right\}$$

$$
\begin{aligned}
&\le \sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{\eta} E\left\{\|\zeta_{ik}^*(\tilde{\boldsymbol{\beta}}_0)\|^3 I(\|\zeta_{ik}^*(\tilde{\boldsymbol{\beta}}_0)\| > \eta) \,\middle|\, \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\} \\
&\le \sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{\eta} E\left\{\|\zeta_{ik}^*(\tilde{\boldsymbol{\beta}}_0)\|^3 \,\middle|\, \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\} \\
&\le \sum_{k=1}^{K} \frac{1}{r_k^2(\tilde{\boldsymbol{\beta}}_0)\eta} \sum_{i=1}^{n_k} \frac{|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}|^3 \|\mathbf{X}_{ik}\|^3}{n^3 \pi_{ik}^2(\tilde{\boldsymbol{\beta}}_0)} \\
&= \frac{1}{\eta r^2} \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{1}{n} \frac{|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}|^3 \|\mathbf{X}_{ik}\|}{\max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}|^2, c^2)} \cdot \left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|\right)^2 \\
&\le \frac{1}{\eta r^2} \left\{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik}|^3 \|\mathbf{X}_{ik}\|}{c^2}\right\} \cdot \left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} (|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}| + c)\|\mathbf{X}_{ik}\|\right)^2 \\
&= O_P(r^{-2}),
\end{aligned}
$$

where the last equality is from Condition (C.4). Moreover, we can prove that

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{i=1}^{r_k} \mathrm{Cov}\{\zeta_{ik}^*(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} &= \sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k(\tilde{\boldsymbol{\beta}}_0)} \Big\{ \sum_{i=1}^{n_k} \frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \mathbf{X}_{ik}\mathbf{X}_{ik}^T}{n_k^2 \pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} \\
&\qquad - \Big[\frac{1}{n_k}\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}\Big]\Big[\frac{1}{n_k}\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}^T\Big]\Big\} \\
&= \boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0} + \Delta_2,
\end{aligned}
$$

where

$$
\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0} = \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{n_k} \frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})^2 \mathbf{X}_{ik}\mathbf{X}_{ik}^T}{n^2 \pi_{ik}(\tilde{\boldsymbol{\beta}}_0)},
$$

and

$$
\Delta_2 = \underbrace{\frac{1}{r}\left[\frac{1}{n}\sum_{k=1}^{K} \frac{1}{\sum_{i=1}^{n_k} \max(c, |Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}|)\|\mathbf{X}_{ik}\|} \left\{\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}\right\}\left\{\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}^T\right\}\right]}_{R_3}
$$
$$
\times \underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|}_{R_4}.
$$

Note that

$$
\|R_3\| \le \left\| \frac{1}{cn}\sum_{k=1}^{K} \frac{1}{\sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|} \left\{\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}\right\}\left\{\sum_{i=1}^{n_k} (Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T \mathbf{X}_{ik})\mathbf{X}_{ik}^T\right\}\right\|.
$$

Similar to (A.16), we can derive that

$$
E(R_3) \le \frac{2\sigma^2}{nc}\sum_{k=1}^{K} \frac{\sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^2}{\sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|} \le \frac{2\sigma^2}{nc}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^2 = O(1),
$$

where the equality is from Condition (C.4). Hence, $R_3 = O_P(1)$. In addition, we have

$$\|R_4\| \leq \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}| + c)\|\mathbf{X}_{ik}\| = O_P(1). \tag{A.27}$$

Then, we know that $\Delta_2 = O_P(r^{-1})$, and

$$\sum_{k=1}^{K}\sum_{i=1}^{r_k}\mathrm{Cov}\{\zeta_{ik}^*(\tilde{\boldsymbol{\beta}}_0)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} = \boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0} + o_P(1).$$

By the Lindeberg-Feller central limit theorem in Proposition 2.27 of van der Vaart (1998) and Slutsky's theorem, conditionally on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we have

$$\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}^{-1/2}\frac{\partial S_{\tilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial\boldsymbol{\beta}} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}). \tag{A.28}$$

Note that

$$\boldsymbol{\Phi}_{opt} = \frac{1}{r}\underbrace{\left[\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\max(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|}\right]}_{E_1}\underbrace{\left[\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\max(|Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|\right]}_{E_2},$$

and

$$\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{r}\underbrace{\left[\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|}\right]}_{E_3}\underbrace{\left[\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\max(|Y_{ik} - \tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_{ik}|, c)\|\mathbf{X}_{ik}\|\right]}_{E_4}.$$

The distance between $\boldsymbol{\Phi}_{opt}$ and $\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}$ can be described as

$$\|\boldsymbol{\Phi}_{opt} - \boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}\| \leq r^{-1}\|E_1 - E_3\| \cdot \|E_2\| + r^{-1}\|E_2 - E_4\| \cdot \|E_3\|. \tag{A.29}$$

By Condition (C.4) and $\|\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{OLS}\| = O_P(r_0^{-1/2})$, we can deduce that $\|E_2\| = O_P(1)$ and

$$r^{-1}\|E_1 - E_3\| \leq \frac{1}{r}\left[\frac{1}{c^2 n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - \hat{\boldsymbol{\beta}}_{OLS}^T\mathbf{X}_{ik})^2\|\mathbf{X}_{ik}\|^2\right] \cdot \|\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{OLS}\| = O_P(r^{-1}r_0^{-1/2}).$$

Similarly, we have $r^{-1}\|E_2 - E_4\| = O_P(r^{-1}r_0^{-1/2})$ and $\|E_3\| = O_P(1)$. Therefore,

$$\|\boldsymbol{\Phi}_{opt} - \boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}\| = O_P\left(r^{-1}r_0^{-1/2}\right). \tag{A.30}$$

Now the asymptotic property of $\breve{\boldsymbol{\beta}}$ is from (A.26), (A.28) and (A.30)

$$\begin{aligned}\boldsymbol{\Sigma}_{opt}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) &= -\boldsymbol{\Sigma}_{opt}^{-1/2}\boldsymbol{\Gamma}^{-1}\frac{\partial S_{\tilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial\boldsymbol{\beta}} + o_{P|\mathcal{F}_n}(1)\\ &= -\boldsymbol{\Sigma}_{opt}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}^{1/2}\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}^{-1/2}\frac{\partial S_{\tilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_{OLS})}{\partial\boldsymbol{\beta}} + o_{P|\mathcal{F}_n}(1).\end{aligned}$$

Furthermore, we notice that

$$\boldsymbol{\Sigma}_{opt}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}^{1/2}(\boldsymbol{\Sigma}_{opt}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}^{1/2})^T = \boldsymbol{\Sigma}_{opt}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\tilde{\boldsymbol{\beta}}_0}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}_{opt}^{-1/2}$$

$$= \mathbf{\Sigma}_{opt}^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{\Phi}_{opt}\mathbf{\Gamma}^{-1}\mathbf{\Sigma}_{opt}^{-1/2} + \mathbf{\Sigma}_{opt}^{-1/2}\mathbf{\Gamma}^{-1}(\mathbf{\Phi}_{\tilde{\boldsymbol{\beta}}_0} - \mathbf{\Phi}_{opt})\mathbf{\Gamma}^{-1}\mathbf{\Sigma}_{opt}^{-1/2}$$
$$= \mathbf{I} + O_{P|\mathcal{F}_n}\left(r_0^{-1/2}\right),$$

where the last equality is from (A.30) and the fact that $\mathbf{\Sigma}_{opt} = O_{P|\mathcal{F}_n}(r^{-1})$. This ends the proof. □

References

(2008). Data Expo 2009: Airline on time data. URL: https://doi.org/10.7910/DVN/HG7NV7. doi:10.7910/DVN/HG7NV7.

Ai, M., Yu, J., Zhang, H., & Wang, H. (2019). Optimal subsampling algorithms for big data regressions. *Statistica Sinica,*, . doi:10.5705/ss.202018.0439.

Battey, H., Fan, J., Liu, H., Lu, J., & Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, *46*, 1352–1382. doi:10.1214/17-aos1587.

Drineas, P., Magdon-Ismail, M., Mahoney, M., & Woodruff, D. (2012). Faster approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, *13*, 3475–3506.

Jordan, M. I., Lee, J. D., & Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, *114*, 668–681. doi:10.1080/01621459.2018.1429274.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, *21*, 272–319.

Ma, P., Mahoney, M., & Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, *16*, 861–911.

Schifano, E. D., Wu, J., Wang, C., Yan, J., & Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, *58*, 393–403.

Shi, C., Lu, W., & Song, R. (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, *113*, 1698–1709. doi:10.1080/01621459.2017.1360779.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, London.

Volgushev, S., Chao, S.-K., & Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics*, *47*, 1634–1662. doi:10.1214/18-aos1730.

Wang, C., Chen, M.-H., Wu, J., Yan, J., Zhang, Y., & Schifano, E. (2018a). Online updating method with new variables for big data streams. *Canadian Journal of Statistics*, *46*, 123–146. doi:10.1002/cjs.11330.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsample. *Journal of Machine Learning Research,*, *20*, 1–59.

Wang, H., & Ma, Y. (2020). Optimal subsampling for quantile regression in big data. *Biometrika*, .. doi:10.1093/biomet/asaa043.

Wang, H., Yang, M., & Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, *114*, 393–405.

Wang, H., Zhu, R., & Ma, P. (2018b). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, *113*, 829–844. doi:10.1080/01621459.2017.1292914.

Xue, Y., Wang, H., Yan, J., & Schifano, E. D. (2019). An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics*, *76*, 171–182. doi:10.1111/biom.13137.

Zhao, T., Cheng, G., & Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, *44*, 1400–1437. doi:10.1214/15-aos1410.