Optimal Subsampling for Multiplicative Regression with Massive Data

Tianzhen Wang and Haixiang Zhang*

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

Abstract

Faced with massive data, subsampling is a popular way to downsize the data volume for reducing computational burden. The key idea of subsampling is to perform statistical analysis on a representative subsample drawn from the full data. It provides a practical solution to extracting useful information from big data. In this article, we develop an efficient subsampling method for large-scale multiplicative regression model, which can largely reduce the computational burden due to massive data. Under some regularity conditions, we establish consistency and asymptotic normality of the subsample-based estimator, and derive the optimal subsampling probabilities according to the L-optimality criterion. A two-step algorithm is developed to approximate the optimal subsampling procedure. Meanwhile, the convergence rate and asymptotic normality of the two-step subsample estimator are established. Numerical studies and two real data applications are carried out to evaluate the performance of our subsampling method.

Keywords: Asymptotic normality; Big data; Multiplicative regression; Optimal subsampling; Positive responses.

1 Introduction

With the rapid development of data capturing and storage techniques, the sizes of available datasets have grown exponentially, which motivate urgent demands for building statistical methods to analyze huge datasets. However, it is often computationally infeasible to conduct statistical analysis on such big data with relatively limited computing resources. Basically,

*Corresponding author: haixiang.zhang@tju.edu.cn (Haixiang Zhang)

there are two bottlenecks when performing big data analysis: (i) the dataset is too large to be held in a computer's memory; (ii) the computation takes too long time to output the desired results. To deal with the two challenges, various techniques have been developed to conduct statistical inference for big data, such as divide-and-conquer method (Chen and Xie, 2014; Battey *et al.*, 2018; Shi *et al.*, 2018) and online updating method for streaming data (Schifano *et al.*, 2016; Wang *et al.*, 2018a; Xue *et al.*, 2019; Luo and Song, 2020).

Another popular technique is the subsampling method, and its basic idea is to select a tractable and representative subsample for conducting statistical inference. Many researchers have been devoted to the development of subsampling methods for big data. For example, Ma *et al.* (2015) proposed an algorithmic leveraging-based sampling procedure for linear model. Wang *et al.* (2018b) developed an optimal subsampling method for logistic model based on A-optimality criterion. Wang *et al.* (2019) provided a novel information-based optimal subdata selection (IBOSS) approach. Wang (2019) further proposed a more efficient estimator for logistic model based on the optimal subsample. Han *et al.* (2020) provided a subsampling scheme for large-scale multi-class logistic regression. Ma *et al.* (2020) studied the asymptotic properties of randomized numerical linear algebra sampling estimators for linear models. Meng *et al.* (2021) proposed a low condition number pursuit subsampling algorithm in the misspecified linear model. Wang and Ma (2021) studied the optimal subsampling for quantile regression with big data. Ai *et al.* (2021a) investigated the Poisson subsampling for large-scale quantile regression. In addition, Ai *et al.* (2021b) and Lee *et al.* (2021) studied the optimal subsampling for big data generalized linear models. Zuo *et al.* (2021a) proposed a sampling-based estimation method for massive survival data with additive hazards model. Yu *et al.* (2020) developed a distributed Poisson subsampling method for maximum quasi-likelihood estimator. Zhang and Wang (2021) and Zuo *et al.* (2021b) proposed some distributed subsampling procedures for big data linear and logistic regression models, respectively. For more papers on subsampling methods for massive datasets, we refer to a review paper by Yao and Wang (2021).

The multiplicative regression model plays an important role in economic/financial or biomedical studies. Compared with generalized linear model, the multiplicative regression has the following two advantages: First, the multiplicative regression model is suitable to

directly analyse a dataset with positive responses, such as stock prices or life times. However, the generalized linear model is not able to describe this special characteristic of positive outcomes. Second, in many practical applications, we are interested in the size of relative error (e.g. stock price data), rather than that of error itself. The multiplicative regression has the ability to capture the size of relative error, while the generalized linear model fails to complete this task. There have been several papers on the statistical analysis with multiplicative regression. e.g., Chen *et al.* (2010) proposed a least absolute relative errors (LARE) estimation criterion for multiplicative regression model. Li *et al.* (2014) considered an empirical likelihood approach towards constructing confidence intervals of the regression parameters in multiplicative regression model. Chen *et al.* (2016) proposed a least product relative error (LPRE) estimation criterion for multiplicative regression model. Xia *et al.* (2016) studied the variable selection for multiplicative regression model. Faced with large-scale data with positive responses, we adopt the optimal subsampling to resolve the computational challenges for multiplicative regression model. The main features of our approach are as follows: First, the computational speed of our method is much faster than the full data approach. Second, we provide an explicit expression for the optimal subsampling distribution in the context of L-optimality criterion. Third, we establish consistency and asymptotic normality of the subsample estimator, which is helpful for performing statistical inference.

The remainder of this article is organized as follows. In Section 2, we briefly review some natations for the multiplicative regression model. In Section 3, we present a general subsampling algorithm, and establish consistency together with asymptotic normality of the subsample estimator. Based on the L-optimality, we derive the optimal subsampling probabilities. Meanwhile, a two-step subsample-based estimator and its asymptotic properties are given. In Section 4, we conduct extensive simulations to demonstrate the effectiveness of our method. Section 5 provides two real data examples. In Section 6, we provide some conclusions and future research topics. All proofs are given in the Appendix.

2 Model and Notations

The multiplicative regression model is widely used when analyzing data with positive responses, such as incomes, stock prices and survival times, etc. Suppose that there are n independent and identically distributed samples $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$, where $Y_i > 0$ for $i = 1, \cdots, n$. We consider the following multiplicative regression model (Chen *et al.*, 2010),

$$Y_i = \exp(\boldsymbol{\beta}^T \mathbf{X}_i)\epsilon_i, \tag{2.1}$$

where Y_i is a positive response variable, $\mathbf{X}_i \in \mathbb{R}^p$ is a vector of covariates with the first component being 1 (intercept), $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ is a vector of regression parameters, and $\epsilon_i > 0$ is an error term. The true parameter value $\boldsymbol{\beta}_t$ is in the interior of a compact set $\Theta \subset \mathbb{R}^p$.

For convenience, we denote the full data as $\mathcal{F}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. To estimate the parameters in model (2.1), Chen *et al.* (2016) proposed a least product relative error (LPRE) criterion

$$\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \exp(-\boldsymbol{\beta}^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T \mathbf{X}_i) - 2\right\}, \tag{2.2}$$

which is infinitely differentiable and strictly convex. The full data LPRE estimator is $\hat{\boldsymbol{\beta}}_{LPRE} = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$. There is no closed-form of $\hat{\boldsymbol{\beta}}_{LPRE}$, and a Newton-Raphson method is usually adopted with the following iterative formula:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - \left\{\frac{\partial^2 \ell(\hat{\boldsymbol{\beta}}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right\}^{-1}\left\{\frac{\partial \ell(\hat{\boldsymbol{\beta}}^{(m)})}{\partial \boldsymbol{\beta}}\right\}. \tag{2.3}$$

Of note, the computational complexity when calculating $\hat{\boldsymbol{\beta}}_{LPRE}$ is about $O(Knp^2)$, where K is the number of iterations until convergence. As we can see, the computational burden is heavy when the full data size is very large. To deal with this issue, we propose a subsampling-based method for the purpose of reducing computational burden in next section.

3 Subsample-Based Estimation Method

3.1 A General Subsampling Algorithm

In Algorithm 1, we present a general subsampling procedure for the multiplicative regression model. To establish asymptotic properties of the subsample-based estimator $\tilde{\boldsymbol{\beta}}$, we need the following regularity assumptions.

(H.1) *As* $n \to \infty$, $\boldsymbol{\Lambda} = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}\mathbf{X}_i\mathbf{X}_i^T$ *goes to a positive-definite matrix in probability.*

(H.2) $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_i\| = O_P(1)$ and $\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\mathbf{X}_i\|^2}{\pi_i} = O_P(1)$, where $\|\cdot\|$ is the Euclidean norm.

(H.3) $sup_{\beta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^2 = O_P(1)$.

(H.4) $sup_{\beta\in\Theta}\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{Y_i \exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^k = O_P(1)$, where $k = 2$ and 4.

(H.5) $sup_{\beta\in\Theta}\frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{\pi_i^2}\left\{Y_i \exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^3\|\mathbf{X}_i\|^3 = O_P(1)$.

Assumption (H.1) is commonly used for the multiplicative regression model (Chen *et al.*, 2010); Assumption (H.2) is a condition on both subsampling probabilities and the covariates; Assumptions (H.3)-(H.5) are used to determine the convergence rate of $\tilde{\boldsymbol{\beta}}$, together with its asymptotic distribution.

The following theorem presents the consistency and asymptotic normality of the subsample estimator $\tilde{\boldsymbol{\beta}}$ towards the true value $\boldsymbol{\beta}_t$, which is useful for conducting statistical inference.

Theorem 1 *If the assumptions (H.1)-(H.5) hold, for the subsample estimator* $\tilde{\boldsymbol{\beta}}$ *in Algorithm 1 and any* $\delta > 0$*, there exists a finite* $\Delta_\delta > 0$ *such that with probability approaching one,*

$$P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t\| \geq r^{-1/2}\Delta_\delta \,\middle|\, \mathcal{F}_n\right) < \delta. \tag{3.2}$$

Moreover, as $r \to \infty$ *and* $n \to \infty$ *we have*

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{3.3}$$

where $\xrightarrow{d}$ *denotes convergence in distribution, and* $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c\boldsymbol{\Lambda}^{-1}$ *with*

$$\boldsymbol{\Sigma}_c = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\mathbf{X}_i\mathbf{X}_i^T. \tag{3.4}$$

Algorithm 1 General Subsampling Algorithm

• *Sampling*: Assign subsampling probabilities $\{\pi_i\}_{i=1}^n$ to the full data $\mathcal{F}_n$ with $\sum_{i=1}^n \pi_i = 1$, and $\pi_i > 0$. Draw a random subsample of size $r(\ll n)$ with replacement based on $\{\pi_i\}_{i=1}^n$ from $\mathcal{F}_n$. For $i = 1, \ldots, r$, we denote the covariates, responses and corresponding subsampling probabilities in this subsample as $\mathbf{X}_i^*, Y_i^*$ and π_i^*, respectively.

• *Estimation*: Based on the subsample $\{(Y_i^*, \mathbf{X}_i^*), i = 1, \ldots, r\}$, we minimize the following weighted least product relative error criterion function $\ell^*(\boldsymbol{\beta})$ to get an estimate $\tilde{\boldsymbol{\beta}}$, where

$$\ell^*(\boldsymbol{\beta}) = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n\pi_i^*} \left\{ Y_i^* \exp(-\boldsymbol{\beta}^T \mathbf{X}_i^*) + Y_i^{*-1} \exp(\boldsymbol{\beta}^T \mathbf{X}_i^*) - 2 \right\}. \tag{3.1}$$

Due to the convexity of $\ell^*(\boldsymbol{\beta})$, a Newton's method is adopted until $\tilde{\boldsymbol{\beta}}^{(m+1)}$ and $\tilde{\boldsymbol{\beta}}^{(m)}$ are closed enough,

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^{(m+1)} &= \tilde{\boldsymbol{\beta}}^{(m)} - \left[\sum_{i=1}^{r} \frac{1}{\pi_i^*} \left\{ \omega_i^*(\tilde{\boldsymbol{\beta}}^{(m)}) + \omega_i^*(\tilde{\boldsymbol{\beta}}^{(m)})^{-1} \right\} \mathbf{X}_i^* (\mathbf{X}_i^*)^T \right]^{-1} \\ &\quad \times \left[\sum_{i=1}^{r} \frac{1}{\pi_i^*} \left\{ -\omega_i^*(\tilde{\boldsymbol{\beta}}^{(m)}) + \omega_i^*(\tilde{\boldsymbol{\beta}}^{(m)})^{-1} \right\} \mathbf{X}_i^* \right] \end{aligned}$$

with $\omega_i^*(\boldsymbol{\beta}) = Y_i^* \exp(-\boldsymbol{\beta}^T \mathbf{X}_i^*)$, $i = 1, \cdots, r$.

3.2 Optimal Subsampling Strategy

To practically implement the Algorithm 1, we need to specify the subsampling probabilities π_i's. A simple choice is to use the uniform sampling with $\{\pi_i = n^{-1}\}_{i=1}^n$. Because this uniform sampling does not distinguish different data points, it is less effective compared with nonuniform sampling approach. Theorem 1 shows that the distribution of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t$ is approximated by a normal random vector u with mean zero and covariace $\boldsymbol{\Sigma}$. The asymptotic mean squared error (AMSE) of $\tilde{\boldsymbol{\beta}}$ is equal to the trace of $\boldsymbol{\Sigma}$. i.e.,

$$\text{AMSE}(\tilde{\boldsymbol{\beta}}) = E(\|u\|^2) = \text{tr}(\boldsymbol{\Sigma}), \tag{3.5}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix. Therefore, we derive optimal subsampling probabilities by minimizing the trace of the variance-covariance matrix $\boldsymbol{\Sigma}$. However, the calculation burden of $\boldsymbol{\Lambda}^{-1}$ is heavy due to big n. To further reduce the computational burden, we adopt the idea of Loewner-ordering to define the partial ordering of positive definite matrices. For two positive definite matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, we define the partial ordering as $\boldsymbol{\Gamma}_1 \geq \boldsymbol{\Gamma}_2$ if and only if $\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2$ is a nonnegative definite matrix. Following Wang *et al.* (2018b), we suggest to specify the optimal subsampling probabilities by minimize $\text{tr}(\boldsymbol{\Sigma}_c)$ rather than $\text{tr}(\boldsymbol{\Sigma})$, which is also referred to as the L-optimality criterion (Atkinson *et al.*, 2007).

Theorem 2 *Under the assumptions (H.1)-(H.5), if the subsampling probabilities are chosen as*

$$\pi_i^{OSP} = \frac{|-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)| \|\mathbf{X}_i\|}{\sum_{j=1}^n |-Y_j \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_j) + Y_j^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_j)| \|\mathbf{X}_j\|}, \quad i = 1, \ldots, n, \tag{3.6}$$

then $\text{tr}(\boldsymbol{\Sigma}_c)$ *attains its minimum.*

3.3 Two-Step Subsampling Algorithm

The optimal subsampling probabilities $\{\pi_i^{OSP}\}_{i=1}^n$ cannot be used directly because they depend on the unavailable $\boldsymbol{\beta}_t$. As suggested by Wang *et al.* (2018b), we use a pilot estimator $\tilde{\boldsymbol{\beta}}_0$ to replace the $\boldsymbol{\beta}_t$ in (3.6). In addition, for those data points with $Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)$ being close to $Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i)$, the corresponding subsampling probabilities are very small. If these data point are selected into a subsample, the weighted criterion function $\ell^*(\boldsymbol{\beta})$ given in

(3.1) would be dominated by them. To protect the $\ell^*(\boldsymbol{\beta})$ from being inflated by these data points, we truncate $|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|$ by $max(|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon)$, where $\upsilon>0$ is a specified bound, e.g. $\upsilon=10^{-6}$. We summarize the above procedure in the following Algorithm 2.

Algorithm 2 Two-Step Strategy

Step 1.

• Take a pilot subsample with size r_0 from $\mathcal{F}_n$ by uniform subsampling probabilities $\{\pi_i=n^{-1}\}_{i=1}^n$, we obtain a pilot estimate $\tilde{\boldsymbol{\beta}}_0$ through minimizing (3.1).

• For $i=1,\cdots,n$, we calculate the subsampling probabilities

$$\pi_i(\tilde{\boldsymbol{\beta}}_0)=\frac{\max(|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon)}{\sum_{j=1}^n\max(|-Y_j\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_j)+Y_j^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_j)|\|\mathbf{X}_j\|,\upsilon)}. \tag{3.7}$$

Step 2.

• Draw a subsample of size r with replacement from $\mathcal{F}_n$ based on the subsampling probabilities $\pi_i(\tilde{\boldsymbol{\beta}}_0)$'s. Using the subsample $\mathcal{F}_r=\{(\mathbf{X}_i^*,Y_i^*,\pi_i^*)\}_{i=1}^r$, we can obtain a two-step subsample estimator $\breve{\boldsymbol{\beta}}$ through minimizing

$$\ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})=\frac{1}{r}\sum_{i=1}^r\frac{1}{n\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i^*\exp(-\boldsymbol{\beta}^T\mathbf{X}_i^*)+Y_i^{*-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i^*)-2\right\}. \tag{3.8}$$

Note that we first takes a pilot subsample of size r_0 and then selects an optimal subsample of size r in the Algorithm 2. We do not recommend to combine the two subsamples together for outputting estimator. The reason is that if we can perform statistical analysis using a combined subsample with size r_0+r, then we would own a better subsample by setting the subsample size in the second step as r_0+r directly. Below we establish the consistency and asymptotic normality of $\breve{\boldsymbol{\beta}}$ towards $\boldsymbol{\beta}_t$. We need the following regularity moment conditions:

(H.6) $sup_{\beta\in\Theta}\frac{1}{n}\sum_{i=1}^n\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^6=O_P(1)$.

(H.7) $\frac{1}{n}\sum_{i=1}^n\|\mathbf{X}_i\|^8=O_P(1)$.

Assumptions (H.6) and (H.7) impose some moment conditions on the covariates and the responses, which are needed for the development of theoretical properties for the two-step subsample-based estimator. The assumptions (H.6) and (H.7) are reasonable in most practical situations.

Theorem 3 *Under assumptions (H.1), (H.3), (H.6) and (H.7), for any $\delta > 0$, there exists a finite $\Delta_\delta > 0$ such that with probability approaching one,*

$$P\left(\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t\| \geq r^{-1/2}\Delta_\delta \,\middle|\, \mathcal{F}_n\right) < \delta. \tag{3.9}$$

Moreover, as $r \to \infty$ and $n \to \infty$, we have

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{3.10}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{opt}\boldsymbol{\Lambda}^{-1}$ with

$$\begin{aligned}\boldsymbol{\Sigma}_{opt} = \frac{1}{r}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\mathbf{X}_i\mathbf{X}_i^T}{\max\left\{|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|, \upsilon\right\}}\right] \\ \times\left[\frac{1}{n}\sum_{i=1}^{n}\max\left\{|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|, \upsilon\right\}\right],\end{aligned} \tag{3.11}$$

and $\upsilon > 0$ is a pre-specified truncation value, e.g. $\upsilon = 10^{-6}$.

In order to perform statistical inference, we need to provide an estimator for the variance-covariance matrix $\boldsymbol{\Sigma}$. A simple way is to replace $\boldsymbol{\beta}_t$ with $\breve{\boldsymbol{\beta}}$ for the asymptotic variance-covariance matrix in Theorem 3. However, this approach based on the full data $\mathcal{F}_n$ with heavy calculations. To alleviate computational burden, we adopt the method of moment to estimate the variance-covariance matrix $\boldsymbol{\Sigma}$ with a subsample $\mathcal{F}_r = \{(\mathbf{X}_i^*, Y_i^*, \pi_i^*(\tilde{\boldsymbol{\beta}}_0))\}_{i=1}^r$ in the second step of Algorithm 2,

$$\breve{\boldsymbol{\Sigma}} = \breve{\boldsymbol{\Lambda}}^{-1}\breve{\boldsymbol{\Sigma}}_{opt}\breve{\boldsymbol{\Lambda}}^{-1}, \tag{3.12}$$

where

$$\breve{\boldsymbol{\Lambda}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i^*\exp(-\breve{\boldsymbol{\beta}}^T\mathbf{X}_i^*) + Y_i^{*^{-1}}\exp(\breve{\boldsymbol{\beta}}^T\mathbf{X}_i^*)\right\}\mathbf{X}_i^*(\mathbf{X}_i^*)^T,$$

and

$$\breve{\boldsymbol{\Sigma}}_{opt} = \frac{1}{n^2r^2}\sum_{i=1}^{r}\frac{1}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)^2}\left\{-Y_i^*\exp(-\breve{\boldsymbol{\beta}}^T\mathbf{X}_i^*) + Y_i^{*^{-1}}\exp(\breve{\boldsymbol{\beta}}^T\mathbf{X}_i^*)\right\}\mathbf{X}_i^*(\mathbf{X}_i^*)^T.$$

Note that $E(\breve{\boldsymbol{\Lambda}}|\mathcal{F}_n) = \boldsymbol{\Lambda}$ and $E(\breve{\boldsymbol{\Sigma}}_{opt}|\mathcal{F}_n) = \boldsymbol{\Sigma}_{opt}$, if we replace $\breve{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}_t$ in $\breve{\boldsymbol{\Sigma}}_{opt}$. That is to say, both $\breve{\boldsymbol{\Lambda}}$ and $\breve{\boldsymbol{\Sigma}}_{opt}$ are unbiased estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_{opt}$, respectively. We will check the performance of this estimated variance-covariance matrix in (3.12) via numerical simulation.

4 Simulation

In this section, we conduct extensive simulations to demonstrate the effectiveness of our proposed subsampling method. The true parameter value is chosen as $\boldsymbol{\beta}_t = (0.5, 1, 0.5, -0.5, 0.3)^T$. Denote $\mathbf{X} = (1, \tilde{\mathbf{X}}^T)^T$ with $\tilde{\mathbf{X}} = (X_1, \ldots, X_4)^T$, i.e. $p = 5$. We consider the following four cases for the generation of covariate $\tilde{\mathbf{X}}$,

Case 1. $\tilde{\mathbf{X}} \sim N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}_{ij} = 0.5^{|i-j|}$.

Case 2. $\tilde{\mathbf{X}} \sim 0.5N(\mathbf{1}, \boldsymbol{\Omega}) + 0.5N(-\mathbf{1}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}_{ij} = 0.5^{|i-j|}$.

Case 3. $\tilde{\mathbf{X}} \sim t_5(\mathbf{0}, \boldsymbol{\Omega})$. i.e., $\tilde{\mathbf{X}}$ follows a multivariate t distribution with degree of freedom 5 and covariance matrix $\boldsymbol{\Omega}_{ij} = 0.5^{|i-j|}$.

Case 4. $\tilde{\mathbf{X}} = (X_1, \ldots, X_4)^T$, and X_i's are independent and identically distributed exponential random variables with probability density function $f(x) = e^{-x}$.

We consider two cases for the error term: $\log(\epsilon)$ follows N(0, 1), and $\log(\epsilon)$ follows Uniform(-2, 2). The r_0 in step 1 of Algorithm 2 is chosen as $r_0 = 500$, and the subsample size is $r = 600$, 800 and 1000, respectively. All the simulation results in Tables 1$-$5, together with Figure 1 and 2 are based on 500 replications with $n = 10^6$.

We evaluate the performance of our optimal subsampling criterion (OSC) given in the Algorithm 2. Note that Ma *et al.* (2020) presented eight different nonuniform subsampling probabilities in the context of linear models. However, it is not clear how to directly extend these sampling methods to the multiplicative regression due to its nonlinearity. Anyway, we have tried the eight subsampling probabilities of Ma *et al.* (2020), and found two of them are better than the uniform subsampling for multiplicative regression. Therefore, we consider the uniform subsampling (UNIF) probabilities, root leverage subsampling (RL) probabilities and inverse-covariance subsampling (IC) probabilities for comparison, i.e., $\pi_i^{UNIF} = \frac{1}{n}$, $\pi_i^{RL} = \frac{\|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$ and $\pi_i^{IC} = \frac{\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}{\sum_{i=1}^n \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\|}$, respectively. It is worth to pointing out that π_i^{RL} and π_i^{IC} are two ad hoc subsampling probabilities for our method, because their expressions are derived from the linear model rather than the multiplicative regression model. In Tables 1-4, we present the estimation results for $\boldsymbol{\beta}_1$ (intercept) and $\boldsymbol{\beta}_2$ ($\boldsymbol{\beta}_i$'s are

similar to $\boldsymbol{\beta}_2$ and omitted, $i = 3, 4, 5$), which include the estimated bias (BIAS) given by the sample mean of the estimates minus the true value $\boldsymbol{\beta}_t$, the sampling standard error (SSE) of the estimates, the sample mean of the estimated standard errors (ESE), and the empirical 95% coverage probabilities (CP) towards the true value $\boldsymbol{\beta}_t$ based on normal approximation. From the results in Tables 1-4 we can see that four subsample-based estimators seem to be unbiased, the SSE and ESE are similar, and the coverage probabilities of confidence intervals are satisfactory. Furthermore, these estimators become better as the subsample size r increases. Moreover, the SSE and ESE of OSC-based estimator are much smaller than those of other subsampling-based estimators. The performances of RL and IC are better than UNIF for $\boldsymbol{\beta}_2$, while the RL and IC are not uniformly better than the UNIF towards the intercept term $\boldsymbol{\beta}_1$. Similar conclusions are also found in the linear model (Wang *et al.*, 2019; Zhang and Wang, 2021).

To further investigate the superiority of our proposed subsampling method, we calculate the MSEs of $\breve{\boldsymbol{\beta}}$ from 500 subsamples using MSE $= \frac{1}{500}\sum_{d=1}^{500} \|\breve{\boldsymbol{\beta}}^{(d)} - \boldsymbol{\beta}_t\|^2$, where $\breve{\boldsymbol{\beta}}^{(d)}$ is from the dth subsample. We present the MSEs of each method in Figures 1 and 2. It is clear to see that the RL, IC and OSC always result in smaller MSEs than the UNIF. Furthermore, the OSC leads to the smallest MSEs compared with other sampling probabilities. The MSEs decrease as r increases, which confirms the consistency of our subsampling method. From Theorem 3, the convergence rate of the proposed subsample estimator is $r^{-1/2}$. To improve the estimation efficiency, we suggest to choose a subsample as large as possible according to the available computing resources. As suggested by one reviewer, it is interesting to explore the influence of the pilot subsample size r_0. In Table 5, we present the estimation results for case 1 with different sample size $r_0 = 400,\ 500$ and 600, respectively. As mentioned before, the pilot subsample with size r_0 does not come into the estimation step in Algorithm 2. The results for other cases are similar and omitted. The results in Table 5 indicate that the performances of subsampling methods are similar if r_0 is relatively large (e.g. $r_0 = 400$).

We conduct the second simulation to assess the computational efficiency of our proposed subsampling method. We generate data using the same mechanism as the first simulation with Case 1, except that $\boldsymbol{\beta}_t = (0.5, ..., 0.5)^T$ with $p = 5,\ 50$ and 100, respectively. Table 6 reports the required CPU times (in seconds) to obtain $\breve{\boldsymbol{\beta}}$ with $r_0 = 500$, $r = 1000$,

$n = 10^6,\ 3 \times 10^6,\ 5 \times 10^6$ and 10^7, where the Algorithm 2 is implemented on a single core. All computations are carried out on a laptop running R software with 16GB random-access memory (RAM). The computing time for the full data method is also reported for comparison. Of note, the results are the CPU time when implementing each method in the RAM, while the time to generate data is not counted. Moreover, the results are the mean CPU time of ten replications. The subsampling probabilities of RL and IC are approximated by the fast algorithm in Drineas *et al.* (2012). It can be seen from the results that the UNIF is much faster than the other methods. The main reason is that the UNIF does not require an additional step to calculate the subsampling probability. It is clear that our proposed OSC takes less computing time compared with the RL, IC and full data methods, and its advantage is more significant as the full data size n increases.

5 Application

5.1 The Bike Sharing Data

In this section, we apply our proposed method to the bike sharing dataset, which contains 17,379 observations. We consider four covariates : a binary variable "workingday"(X_1) to indicate whether a certain day is a working day or not(1 = working day; 0 = non-working day), three continuous variables: temperature(X_2), humidity(X_3) and windspeed(X_4). The square of the number of bikes rented hourly is used as the response. For comparison, we also provide the full data LPRE estimator $\hat{\boldsymbol{\beta}}_{LPRE} = (2.2142, -0.0342, 1.4525, -1.1379, 0.1816)^T$, where the first term is an intercept. As we can see, the rented bikes in non-working days are more than that of working days. The temperature and windspeed have a positive influence on the number of rented bikes, and the humidity has a negative effect. The r_0 in step 1 of Algorithm 2 is chosen as $r_0 = 200$, and we give the OSC, UNIF, RL and IC subsampling-based estimators and calculate the SSE and ESE based on 1000 subsamples with $r = 200, 400$ and 600, respectively. The BIAS is given by the sample mean of the estimates minus the full data LPRE estimator. The results in Table 7 indicate that four subsample estimators are unbiased, and the SSE and ESE are similar. Moreover, we report the subsampling-based

estimators and the 95% confidence intervals with one subsample in Table 8. As expected, the confidence intervals constructed by our proposed method are shorter compared with other subsampling method. For all subsampling methods, as r increases, the length of confidence interval decreases. To further check the rationality of our method, We calculate MSEs of $\breve{\beta}$ with MSE $= \frac{1}{1000}\sum_{d=1}^{1000} \|\breve{\boldsymbol{\beta}}^{(d)} - \hat{\boldsymbol{\beta}}_{LPRE}\|^2$, where $\breve{\boldsymbol{\beta}}^{(d)}$ is a two-step subsampling-based estimator from the dth subsample. In Figure 3, we present the MSEs of four subsampling-based estimators. It is observed that the OSC results in the smallest MSE.

5.2 The Electric Power Consumption Data

We apply our proposed method to an electric power consumption dataset, which contains 2,049,280 completed measurements for a house located at Sceaux between December 2006 and November 2010. For analysis, the minute-averaged current intensity(in ampere) is used as the response. We consider three covariates: active electrical energy in the kitchen(X_1, in watt-hour), active electrical energy in the laundry room(X_2, in watt-hour), and active electrical energy for an electric water-heater and an air-conditioner(X_3, in watt-hour). All covariates are centered and scaled with mean 0 and variance 1. The full data LPRE estimator is $\hat{\boldsymbol{\beta}}_{LPRE} = (1.1162, 0.2205, 0.2045, 0.6326)^T$, where the first term is an intercept. The r_0 in step 1 of Algorithm 2 is chosen as $r_0 = 500$, and we give the OSC, UNIF, RL and IC subsampling-based estimators and calculate the corresponding SSE and ESE based on 1000 subsamples with $r = 600, 800$ and 1000, respectively. Similar to section 5.1, we report the BIAS, SSE and ESE in Table 9. Moreover the subsampling-based estimators and 95% confidence intervals with one subsample are presented in Table 10. The corresponding MSEs are reported in Figure 3. It is clear to see that the overall performance of OSC is much better than those of the other three subsampling probabilities.

6 Conclusion

In this paper, we have studied the statistical properties of a general subsampling algorithm for multiplicative regression model with big data. Based on the asymptotic property of subsample estimator, we derived optimal subsampling probabilities under the L-optimality

criterion. For practical implementation, a two-step algorithm was developed, for which we also derived some theoretical properties. Simulation studies and two real data examples were used to verify the effectiveness of our method. In recent years, many sampling methods have been developed, such as Meng *et al.* (2021) and Ma *et al.* (2020). The two papers mainly focused on subsampling methods in the context of linear models. As suggested by one reviewer, it is desirable to consider the Randomized Numerical Linear Algebra algorithms for the multiplicative regression model. Moreover, how to design optimal subsampling for misspecified multiplicative regression is an interesting research topic.

Availability of data

The data that support the findings of this study are openly available at
http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset,
and
http://archive.ics.uci.edu/ml/datasets/ Individual+household+electric+power+consumption.

Acknowledgement

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.

Appendix

In this Appendix, we give the proof details of Theorems 1-3.

Lemma 1 *If the assumptions (H.1)-(H.5) hold, then conditionally on $\mathcal{F}_n$ we have*

$$\dot{\ell}^* (\boldsymbol{\beta}_t) = O_{P|\mathcal{F}_n} \left(r^{-1/2} \right), \tag{A.1}$$

and

$$\tilde{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda} = O_{P|\mathcal{F}_n} \left(r^{-1/2} \right), \tag{A.2}$$

where

$$\dot{\ell}^*(\boldsymbol{\beta}_t) = \frac{\partial \ell^*(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}} = \frac{1}{r}\sum_{i=1}^{r} \frac{1}{n\pi_i^*}\left\{-Y_i^* \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i^*) + Y_i^{*^{-1}} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i^*)\right\} \mathbf{X}_i^*,$$

and

$$\tilde{\boldsymbol{\Lambda}} = \ddot{\ell}^*(\boldsymbol{\beta}_t) = \frac{\partial^2 \ell^*(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \frac{1}{r}\sum_{i=1}^{r} \frac{1}{n\pi_i^*}\left\{Y_i^* \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i^*) + Y_i^{*^{-1}} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i^*)\right\} \mathbf{X}_i^* (\mathbf{X}_i^*)^T.$$

Proof. Direct calculation yields that

$$\begin{aligned} E\left\{\dot{\ell}^*(\boldsymbol{\beta}_t) \,|\mathcal{F}_n\right\} &= \frac{1}{n}\sum_{i=1}^{n} \left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\} \mathbf{X}_i \\ &= O_P\left(n^{-1/2}\right). \end{aligned} \tag{A.3}$$

This is because for each element of $\dot{\ell}^*(\boldsymbol{\beta}_t)$, say $\dot{\ell}_j^*(\boldsymbol{\beta}_t)$, for $1 \le j \le p$, we have

$$E\left[E\left\{\dot{\ell}_j^*(\boldsymbol{\beta}_t) \,|\mathcal{F}_n\right\}\right] = 0, \tag{A.4}$$

and

$$\begin{aligned} Var\left[E\left\{\dot{\ell}_j^*(\boldsymbol{\beta}_t) \,|\mathcal{F}_n\right\}\right] &= E\left[\frac{1}{n^2}\sum_{i=1}^{n} \left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij}\right] \\ &\le \frac{1}{n} E\left[\frac{1}{n}\sum_{i=1}^{n} \left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \|\mathbf{X}_i\|^2\right] \\ &= O_P\left(n^{-1}\right), \end{aligned} \tag{A.5}$$

where the last equality is from the assumption (H.3). Combining (A.4), (A.5) and Chebyshev's inequality, we have $E\left\{\dot{\ell}^*(\boldsymbol{\beta}_t) \,|\mathcal{F}_n\right\} = O_P\left(n^{-1/2}\right)$.

By the assumption (H.4), for $j = 1, \ldots, p$, we can derive that

$$\begin{aligned} Var\left\{\dot{\ell}_j^*(\boldsymbol{\beta}_t) \,|\mathcal{F}_n\right\} &= \frac{1}{r}\Bigg[\sum_{i=1}^{n} \frac{1}{n^2\pi_i} \left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij} \\ &\qquad - \left\{\frac{1}{n}\sum_{i=1}^{n} \left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\} \mathbf{X}_{ij}\right\}^2\Bigg] \\ &\le \frac{1}{n^2 r}\sum_{i=1}^{n} \frac{1}{\pi_i} \left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \|\mathbf{X}_i\|^2 \\ &= O_P\left(r^{-1}\right). \end{aligned} \tag{A.6}$$

From (A.3), (A.6) and the Markov's inequality, we have

$$P\left\{\left|\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)-E\{\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)\}\right|\geq B_\epsilon r^{-1/2}|\mathcal{F}_n\right\}\leq\frac{Var\left\{\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)|\mathcal{F}_n\right\}}{\left(B_\epsilon r^{-1/2}\right)^2}$$
$$=\epsilon,$$

where

$$B_\epsilon=\left[\frac{Var\left\{\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)|\mathcal{F}_n\right\}}{r^{-1}\epsilon}\right]^{1/2}.$$

This implies $\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)-E\{\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)\}=O_P\left(r^{-1/2}\right)$. Combining this and (A.3), we have $\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)=E\{\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)\}+O_P\left(r^{-1/2}\right)=O_P\left(n^{-1/2}\right)+O_P\left(r^{-1/2}\right)=O_P\left(r^{-1/2}\right)$.

To prove (A.2), direct calculation yields that

$$E\left(\tilde{\boldsymbol{\Lambda}}|\mathcal{F}_n\right)=\boldsymbol{\Lambda}. \tag{A.7}$$

For any component $\tilde{\boldsymbol{\Lambda}}^{j_1j_2}$ of $\tilde{\boldsymbol{\Lambda}}$ with $1\leq j_1\leq j_2\leq p$,

$$\begin{aligned}
&Var\left(\tilde{\boldsymbol{\Lambda}}^{j_1j_2}|\mathcal{F}_n\right)=E\left(\tilde{\boldsymbol{\Lambda}}^{j_1j_2}-\boldsymbol{\Lambda}^{j_1j_2}|\mathcal{F}_n\right)^2\\
&=\frac{1}{r}\sum_{i=1}^n\pi_i\left[\frac{1}{n\pi_i}\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}x_{ij_1}x_{ij_2}-\boldsymbol{\Lambda}^{j_1j_2}\right]^2\\
&=\frac{1}{n^2r}\sum_{i=1}^n\frac{1}{\pi_i}\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\left(x_{ij_1}x_{ij_2}\right)^2-\frac{1}{r}\left(\boldsymbol{\Lambda}^{j_1j_2}\right)^2\\
&\leq\frac{1}{n^2r}\sum_{i=1}^n\frac{1}{\pi_i}\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^4\\
&=O_P\left(r^{-1}\right),
\end{aligned} \tag{A.8}$$

where the last equality is from the assumption (H.4). The Markov's inequality, (A.7) and (A.8) imply (A.2). This ends the proof. □

Proof of Theorem 1. Note that

$$\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)=\frac{1}{r}\sum_{i=1}^r\frac{1}{n\pi_i^*}\left\{-Y_i^*\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i^*)+Y_i^{*^{-1}}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i^*)\right\}\mathbf{X}_i^*=\frac{1}{r}\sum_{i=1}^r\boldsymbol{\xi}_i, \tag{A.9}$$

where

$$\boldsymbol{\xi}_i = \frac{1}{n\pi_i^*}\left\{-Y_i^* \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i^*) + Y_i^{*^{-1}} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i^*)\right\} \mathbf{X}_i^*,\ i = 1, \ldots, r.$$

For each element of $\boldsymbol{\xi}_i$, say $\boldsymbol{\xi}_{ij}$, for $1 \leq j \leq p$. Given $\mathcal{F}_n$, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$ are i.i.d with

$$E\left(\boldsymbol{\xi}_{ij} | \mathcal{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\} \mathbf{X}_{ij},$$

and

$$\begin{aligned}
Var\left(\boldsymbol{\xi}_{ij} | \mathcal{F}_n\right) = r\boldsymbol{\Sigma}_c &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij} \\
&\quad - \left[\frac{1}{n}\sum_{i=1}^{n}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\} \mathbf{X}_{ij}\right]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij} + o_P(1) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^2 \|\mathbf{X}_i\|^2 + o_P(1) \\
&= O_P\left(1\right), \qquad \text{(A.10)}
\end{aligned}$$

Meanwhile, for every $\tau > 0$,

$$\begin{aligned}
&\sum_{i=1}^{r} E\left\{\|r^{-1/2}\boldsymbol{\xi}_i\|^2 I\left(\|r^{-1/2}\boldsymbol{\xi}_i\| > \tau\right) | \mathcal{F}_n\right\} \\
&= \sum_{i=1}^{r} E\left\{\|r^{-1/2}\boldsymbol{\xi}_i\|^2 I\left(\|\boldsymbol{\xi}_i\| > r^{1/2}\tau\right) | \mathcal{F}_n\right\} \\
&\leq \frac{1}{r}\sum_{i=1}^{r} E\left(\|\boldsymbol{\xi}_i\|^2 \cdot \frac{\|\boldsymbol{\xi}_i\|}{r^{1/2}\tau} | \mathcal{F}_n\right) \\
&\leq \frac{1}{r^{1/2}\tau} E\left(\|\boldsymbol{\xi}_i\|^3 | \mathcal{F}_n\right) \\
&= \frac{1}{r^{1/2}\tau}\frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{{\pi_i}^2}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^3 \|\mathbf{X}_i\|^3.
\end{aligned}$$

By the assumption (H.5), as $r \to \infty$ we have

$$\sum_{i=1}^{r} E\left\{\|r^{-1/2}\boldsymbol{\xi}_i\|^2 I\left(\|r^{-1/2}\boldsymbol{\xi}_i\| > \tau\right) | \mathcal{F}_n\right\} \leq \frac{1}{r^{1/2}\tau} \cdot O_P(1) = o_P(1).$$

This and (A.10) show that the Lindeberg-Feller conditions are satisfied in probability. From (A.9) and (A.10), by the Lindeberg-Feller central limit theorem in Proposition 2.27 of Van der Vaart (1998), conditionally on $\mathcal{F}_n$,

$$\boldsymbol{\Sigma}_c{}^{-1/2}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) = \frac{1}{r^{1/2}}\left\{Var\left(\boldsymbol{\xi}_i|\mathcal{F}_n\right)\right\}^{-1/2}\sum_{i=1}^{r}\boldsymbol{\xi}_i \to N\left(\mathbf{0},\mathbf{I}\right) \tag{A.11}$$

in distribution.

Because the estimator $\tilde{\boldsymbol{\beta}}$ is the minimizer of $\ell^*(\boldsymbol{\beta})$, $\sqrt{r}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ is the minimizer of $D^*\left(\boldsymbol{\lambda}\right) = \ell^*\left(\boldsymbol{\beta}_t + \boldsymbol{\lambda}/\sqrt{r}\right) - \ell^*\left(\boldsymbol{\beta}_t\right)$, where $\boldsymbol{\lambda} \in \mathbb{R}^p$. By Taylor's expansion,

$$\begin{aligned} D^*\left(\boldsymbol{\lambda}\right) &= \frac{1}{\sqrt{r}}\boldsymbol{\lambda}^T\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + \frac{1}{2r}\boldsymbol{\lambda}^T\ddot{\ell}^*\left(\boldsymbol{\beta}_t\right)\boldsymbol{\lambda} + o_P(1) \\ &= \boldsymbol{\lambda}^T\boldsymbol{Z}^* + \frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{H}^*\boldsymbol{\lambda} + o_P(1), \end{aligned} \tag{A.12}$$

where $\boldsymbol{Z}^* = \frac{1}{\sqrt{r}}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)$ and $\boldsymbol{H}^* = \frac{1}{r}\ddot{\ell}^*\left(\boldsymbol{\beta}_t\right)$.

Due to $D^*\left(\boldsymbol{\lambda}\right)$ is convex, from the corollary in page 2 of Hjort and Pollard (2011), its minimizer $\sqrt{r}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ satisfies

$$\sqrt{r}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) = -\left(\boldsymbol{H}^*\right)^{-1}\boldsymbol{Z}^* + o_P(1). \tag{A.13}$$

Thus,

$$\begin{aligned} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t &= -r^{-1/2}\left\{\frac{1}{r}\ddot{\ell}^*\left(\boldsymbol{\beta}_t\right)\right\}^{-1}\left\{\frac{1}{\sqrt{r}}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right)\right\} + o_P(1) \\ &= -\tilde{\boldsymbol{\Lambda}}^{-1}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + o_P(1). \end{aligned} \tag{A.14}$$

From Lemma 1, $\tilde{\boldsymbol{\Lambda}}^{-1} = O_{P|\mathcal{F}_n}\left(1\right)$. Combining this with (A.1) and (A.14),

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1),$$

which implies that

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right). \tag{A.15}$$

From (A.2) of Lemma 1,

$$\tilde{\boldsymbol{\Lambda}}^{-1} - \boldsymbol{\Lambda}^{-1} = -\boldsymbol{\Lambda}^{-1}(\tilde{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\tilde{\boldsymbol{\Lambda}}^{-1} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right). \tag{A.16}$$

Based on (H.1) and (A.10), it is verified that

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c\boldsymbol{\Lambda}^{-1} = O_{P|\mathcal{F}_n}\left(r^{-1}\right). \tag{A.17}$$

In view of (A.14), (A.16) and (A.17), we can derive that

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) = \boldsymbol{\Sigma}^{-1/2}\left\{-\tilde{\boldsymbol{\Lambda}}^{-1}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + o_P(1)\right\} \\
&= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}^{-1}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) - \boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\Lambda}}^{-1} - \boldsymbol{\Lambda}^{-1})\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + o_P(1) \\
&= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c^{1/2}\boldsymbol{\Sigma}_c^{-1/2}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + O_{P|\mathcal{F}_n}\left(r^{1/2}\right)O_{P|\mathcal{F}_n}\left(r^{-1/2}\right)O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1) \\
&= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c^{1/2}\boldsymbol{\Sigma}_c^{-1/2}\dot{\ell}^*\left(\boldsymbol{\beta}_t\right) + O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1).
\end{aligned} \tag{A.18}$$

Due to the fact that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c^{1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_c^{1/2})^T = \boldsymbol{I}$, (A.11) and Slutsky's Theorem, we have $\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ converges to $N(\mathbf{0}, \mathbf{I})$ in distribution given $\mathcal{F}_n$ in probability. This means that for any $\boldsymbol{x} \in \mathbb{R}^p$,

$$P\{\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \leq \boldsymbol{x}|\mathcal{F}_n\} \to \Phi\left(\boldsymbol{x}\right), \tag{A.19}$$

in probability, where $\Phi(\boldsymbol{x})$ is the cumulative distribution function of standard multivariate normal vector. Note that the (A.19) is a bounded random variable, convergence in probability to a constant implies convergence to the mean. Therefore, the unconditional probability

$$P\left\{\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \leq \boldsymbol{x}\right\} = E\left[P\left\{\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \leq \boldsymbol{x}|\mathcal{F}_n\right\}\right] \to \Phi(\boldsymbol{x}).$$

This ends the proof. □

Proof of Theorem 2. Note that

$$\begin{aligned}
tr(\boldsymbol{\Sigma}_c) &= \frac{1}{r}\sum_{i=1}^{n} tr\left[\frac{1}{\pi_i}\left\{-Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\mathbf{X}_i{\mathbf{X}_i}^T\right] \\
&= \frac{1}{r}\sum_{i=1}^{n}\pi_i\sum_{i=1}^{n} tr\left[\frac{1}{\pi_i}\left\{-Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^2\right] \\
&\geq \frac{1}{r}\left\{\sum_{i=1}^{n}|-Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)|\|\mathbf{X}_i\|\right\}^2,
\end{aligned}$$

where the last inequality is from Cauchy-Schwarz inequality and the equality holds if and only if $\pi_i = C|-Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)|\|\mathbf{X}_i\|$ for some constant $C > 0$. This completes the proof. □

We first establish two lemmas that will be used in the proof of Theorem 3.

Lemma 2 *If Assumptions (H.3), (H.6)-(H.7) hold, then for* $k_1 = 2, 4$,

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^{k_1}=O_P(1). \tag{A.20}$$

Proof. From the expression of $\pi_i(\tilde{\boldsymbol{\beta}}_0)$,

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^{k_1}\\
&=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^{k_1}}{max\left\{|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right\}}\\
&\quad\times\sum_{i=1}^{n}max\left\{|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right\}\\
&\le\frac{1}{n\upsilon}\sum_{i=1}^{n}\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^{k_1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^{n}\left\{|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|+\upsilon\right\}.
\end{aligned}\tag{A.21}$$

Note that

$$\begin{aligned}
&E\left[\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^{k_1}\right]\\
&\le\left[E\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^4E\left(\|\mathbf{X}_i\|^{2k_1}\right)\right]^{1/2}\le\infty,
\end{aligned}\tag{A.22}$$

where the last inequality is from (H.6) and (H.7).

By the assumption (H.3), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|+\upsilon\right\}\\
&\le\frac{1}{n}\sum_{i=1}^{n}\left\{|Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|+\upsilon\right\}\\
&=O_P(1).
\end{aligned}$$

Combining this with (A.21), (A.22) and using the law of large number, (A.20) follows.

Lemma 3 *If Assumptions (H.1), (H.3), (H.6) and (H.7) hold, then conditionally on $\mathcal{F}_n$ in probability,*

$$\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{A.23}$$

and

$$\tilde{\boldsymbol{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0} - \boldsymbol{\Lambda} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right), \tag{A.24}$$

where

$$\tilde{\boldsymbol{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0} = \ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i^*\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i^*) + Y_i^{*^{-1}}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i^*)\right\}\mathbf{X}_i^*\mathbf{X}_i^{*^T}.$$

Proof. Direct calculation yields that

$$\begin{aligned} E\left\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\,|\mathcal{F}_n\right\} &= \frac{1}{n}\sum_{i=1}^{n}\left\{-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}\mathbf{X}_i \\ &= O_P\left(n^{-1/2}\right). \end{aligned} \tag{A.25}$$

This is because for the jth element of $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)$, say $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta}_t)$, $1 \le j \le p$, we get

$$E\left[E\left\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta}_t)\,|\mathcal{F}_n\right\}\right] = 0, \tag{A.26}$$

and

$$\begin{aligned} Var\left[E\left\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta}_t)\,|\mathcal{F}_n\right\}\right] &= E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left\{-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\mathbf{X}_{ij}\mathbf{X}_{ij}\right] \\ &\le \frac{1}{n}E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^2\right] \\ &= O_P\left(n^{-1}\right). \end{aligned} \tag{A.27}$$

Combining (A.26), (A.27) and Chebyshev's inequality, we have $E\left\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\,|\mathcal{F}_n\right\} = O_P\left(n^{-1/2}\right)$.

By direct calculation,

$$Var\left\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta}_t)\,|\mathcal{F}_n\right\} = \frac{1}{r}\left[\sum_{i=1}^{n}\frac{1}{n^2\pi_i^*(\tilde{\boldsymbol{\beta}}_0)}\left\{-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\mathbf{X}_{ij}\mathbf{X}_{ij}\right.$$

$$
\begin{aligned}
&- \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ -Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i) \right\} \mathbf{X}_{ij} \right\}^2 \Bigg] \\
&\leq \frac{1}{n^2 r} \sum_{i=1}^{n} \frac{1}{\pi_i^*(\tilde{\boldsymbol{\beta}}_0)} \left\{ Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i) \right\}^2 \|\mathbf{X}_i\|^2 \\
&= O_P\left(r^{-1}\right). \qquad \text{(A.28)}
\end{aligned}
$$

From (A.25), (A.28) and the Markov's inequality, we have $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) - E\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\} = O_P\left(r^{-1/2}\right)$. Combining this and (A.25), we have $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) = E\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\} + O_P\left(r^{-1/2}\right) = O_P\left(n^{-1/2}\right) + O_P\left(r^{-1/2}\right) = O_P\left(r^{-1/2}\right)$.

By direct calculation,

$$
E\left(\tilde{\boldsymbol{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n\right) = E_{\tilde{\boldsymbol{\beta}}_0}\left\{ E\left(\tilde{\boldsymbol{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right)\right\} = E_{\tilde{\boldsymbol{\beta}}_0}\left(\boldsymbol{\Lambda} | \mathcal{F}_n\right) = \boldsymbol{\Lambda}, \qquad \text{(A.29)}
$$

where $E_{\tilde{\boldsymbol{\beta}}_0}$ means the expectation is taken with respect to the distribution of $\tilde{\boldsymbol{\beta}}_0$ given $\mathcal{F}_n$. For any component $\tilde{\boldsymbol{\Lambda}}^{j_1 j_2}_{\tilde{\boldsymbol{\beta}}_0}$ of $\tilde{\boldsymbol{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0}$ with $1 \leq j_1 \leq j_2 \leq p$,

$$
\begin{aligned}
&Var\left(\boldsymbol{\Lambda}^{j_1 j_2}_{\tilde{\boldsymbol{\beta}}_0} | \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) \\
&= \frac{1}{n^2 r} \sum_{i=1}^{n} \frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} \left\{ Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i) \right\}^2 (x_{ij_1} x_{ij_2}^T)^2 - \frac{1}{r}\left(\boldsymbol{\Lambda}^{j_1 j_2}\right)^2 \\
&\leq \frac{1}{n^2 r} \sum_{i=1}^{n} \frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)} \left\{ Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i) \right\}^2 \|\mathbf{X}_i\|^4 \\
&= O_P\left(r^{-1}\right), \qquad \text{(A.30)}
\end{aligned}
$$

where the last equality is from Lemma 2. From (A.29) and (A.30) together with Markov's inequality, (A.24) follows.

Proof of Theorem 3. Note that

$$
\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n \pi_i^*(\tilde{\boldsymbol{\beta}}_0)} \left\{ -Y_i^* \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i^*) + Y_i^{*^{-1}} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i^*) \right\} \mathbf{X}_i^* = \frac{1}{r} \sum_{i=1}^{r} \boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}, \quad \text{(A.31)}
$$

where

$$
\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{n \pi_i^*(\tilde{\boldsymbol{\beta}}_0)} \left\{ -Y_i^* \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i^*) + Y_i^{*^{-1}} \exp(\boldsymbol{\beta}_t^T \mathbf{X}_i^*) \right\} \mathbf{X}_i^*, \ i = 1, \ldots, r.
$$

For each element of $\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}$, say $\boldsymbol{\xi}_{ij}^{\tilde{\boldsymbol{\beta}}_0}$. Given $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, $\boldsymbol{\xi}_1^{\tilde{\boldsymbol{\beta}}_0}, \ldots, \boldsymbol{\xi}_r^{\tilde{\boldsymbol{\beta}}_0}$ are i.i.d with

$$E\left(\boldsymbol{\xi}_{ij}^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}\mathbf{X}_{ij},$$

and

$$\begin{aligned}
Var\left(\boldsymbol{\xi}_{ij}^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) = r\boldsymbol{\Sigma}_c &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\left\{-Y_i \exp(-\boldsymbol{\beta}^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij} \\
&\quad - \left[\frac{1}{n}\sum_{i=1}^{n}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}\mathbf{X}_{ij}\right]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\left\{-Y_i \exp(-\boldsymbol{\beta}^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}^T \mathbf{X}_i)\right\}^2 \mathbf{X}_{ij}\mathbf{X}_{ij} + o_P(1) \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i \exp(-\boldsymbol{\beta}^T \mathbf{X}_i + Y_i^{-1}\exp(\boldsymbol{\beta}^T \mathbf{X}_i)\right\}^2 \|\mathbf{X}_i\|^2 + o_P(1) \\
&= O_P(1). \qquad \text{(A.32)}
\end{aligned}$$

Meanwhile, for every $\tau > 0$,

$$\begin{aligned}
&\sum_{i=1}^{r} E\left\{\|r^{-1/2}\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 I\left(\|r^{-1/2}\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\| > \tau\right)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\} \\
&= \sum_{i=1}^{r} E\left\{\|r^{-1/2}\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 I\left(\|\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\| > r^{1/2}\tau\right)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\} \\
&\leq \frac{1}{r}\sum_{i=1}^{r} E\left(\|\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\|^2 \cdot \frac{\|\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\|}{r^{1/2}\tau}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) \\
&\leq \frac{1}{r^{1/2}\tau} E\left(\|\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}\|^3|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right) \\
&= \frac{1}{r^{1/2}\tau}\frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)}\left\{-Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^3 \|\mathbf{X}_i\|^3 \\
&\leq \frac{1}{r^{1/2}\tau}\frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{\pi_i^2(\tilde{\boldsymbol{\beta}}_0)}\left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^3 \|\mathbf{X}_i\|^3 \\
&= \frac{1}{r^{1/2}\tau}\frac{1}{n^3}\sum_{i=1}^{n}\frac{\left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^3 \|\mathbf{X}_i\|^3}{\left[max\left\{|-Y_i \exp(-\tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_i) + Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_i)|\|\mathbf{X}_i\|, \upsilon\right\}\right]^2} \\
&\quad \times \left[\sum_{i=1}^{n} max\left\{|-Y_i \exp(-\tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_i) + Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T \mathbf{X}_i)|\|\mathbf{X}_i\|, \upsilon\right\}\right]^2 \\
&\leq \frac{1}{r^{1/2}\tau}\frac{1}{n\upsilon^2}\sum_{i=1}^{n}\left\{Y_i \exp(-\boldsymbol{\beta}_t^T \mathbf{X}_i) + Y_i^{-1}\exp(\boldsymbol{\beta}_t^T \mathbf{X}_i)\right\}^3 \|\mathbf{X}_i\|^3
\end{aligned}$$

$$\times \left[\frac{1}{n}\sum_{i=1}^{n}\left\{|Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|+\upsilon\right\}\right]^2$$
$$= o_P(1),$$

where the last equality is from the assumption (H.3), (H.6) and (H.7). This and (A.32) show that the Lindeberg-Feller conditions are satisfied in probability. From (A.31) and (A.32), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of Van der Vaart (1998)), conditionally on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$(\boldsymbol{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) = \frac{1}{r^{1/2}}\left\{Var(\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n)\right\}^{-1/2}\sum_{i=1}^{r}\boldsymbol{\xi}_i^{\tilde{\boldsymbol{\beta}}_0} \to N(\mathbf{0},\mathbf{I}) \tag{A.33}$$

in distribution.

Note that

$$\boldsymbol{\Sigma}_{opt} = \underbrace{\frac{1}{r}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\mathbf{X}_i\mathbf{X}_i^T}{max\left(|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right)}\right]}_{E_1}$$
$$\times \underbrace{\left\{\frac{1}{n}\sum_{i=1}^{n}max\left(|-Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right)\right\}}_{E_2},$$

and

$$\boldsymbol{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0} = \underbrace{\frac{1}{r}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\left\{Y_i\exp(-\boldsymbol{\beta}^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\right\}^2\mathbf{X}_i\mathbf{X}_i^T}{max\left(|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right)}\right]}_{E_3}$$
$$\times \underbrace{\left\{\frac{1}{n}\sum_{i=1}^{n}max\left(|-Y_i\exp(-\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)+Y_i^{-1}\exp(\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_i)|\|\mathbf{X}_i\|,\upsilon\right)\right\}}_{E_4}.$$

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

$$\|\boldsymbol{\Sigma}_{opt}-\boldsymbol{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0}\| \le r^{-1}\|E_1-E_3\|\cdot\|E_2\|+r^{-1}\|E_2-E_4\|\cdot\|E_3\|. \tag{A.34}$$

By the assumption (H.3) and $\|\tilde{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t\| = O_P(r_0^{-1/2})$, we can deduce that

$$r^{-1}\|E_1-E_3\| \le \frac{1}{r}\left[\frac{1}{n\upsilon^2}\sum_{i=1}^{n}\left\{Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)\right\}^2\|\mathbf{X}_i\|^2\right]\cdot\|\tilde{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t\|$$

$$= O_P(r^{-1}r_0^{-1/2}),$$

and

$$\begin{aligned}\|E_2\| &\leq \frac{1}{n}\sum_{i=1}^{n}\left(|Y_i\exp(-\boldsymbol{\beta}_t^T\mathbf{X}_i)+Y_i^{-1}\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)|\|\mathbf{X}_i\|+\upsilon\right)\\ &= O_P(1).\end{aligned}$$

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

$$\|\boldsymbol{\Sigma}_{opt} - \boldsymbol{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0}\| = O_P\left(r^{-1}r_0^{-1/2}\right). \tag{A.35}$$

The estimator $\breve{\boldsymbol{\beta}}$ is the minimizer of $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$, so $\sqrt{r}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ is the minimizer of $\breve{D}^*(\boldsymbol{\lambda}) = \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t + \boldsymbol{\lambda}/\sqrt{r}) - \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)$, where $\boldsymbol{\lambda} \in \mathbb{R}^p$. By Taylor's expansion,

$$\begin{aligned}\breve{D}^*(\boldsymbol{\lambda}) &= \frac{1}{\sqrt{r}}\boldsymbol{\lambda}^T\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t) + \frac{1}{2r}\boldsymbol{\lambda}^T\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\boldsymbol{\lambda} + o_P(1)\\ &= \boldsymbol{\lambda}^T\breve{\boldsymbol{Z}}^* + \frac{1}{2}\boldsymbol{\lambda}^T\breve{\boldsymbol{H}}^*\boldsymbol{\lambda} + o_P(1),\end{aligned} \tag{A.36}$$

where $\breve{\boldsymbol{Z}}^* = \frac{1}{\sqrt{r}}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)$, and $\breve{\boldsymbol{H}}^* = \frac{1}{r}\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)$. Since $D^*(\boldsymbol{\lambda})$ is convex, from the corollary in page 2 of Hjort and Pollard (2011), its minimizer $\sqrt{r}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ satisfies

$$\sqrt{r}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) = -(\breve{\boldsymbol{H}}^*)^{-1}\breve{\boldsymbol{Z}}^* + o_P(1). \tag{A.37}$$

Thus,

$$\begin{aligned}\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t &= -r^{-1/2}\left\{\frac{1}{r}\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\right\}^{-1}\left\{\frac{1}{\sqrt{r}}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}_t)\right\} + o_P(1)\\ &= -\tilde{\boldsymbol{\Lambda}}^{-1}_{\tilde{\boldsymbol{\beta}}_0}\dot{\ell}^*(\boldsymbol{\beta}_t) + o_P(1).\end{aligned} \tag{A.38}$$

From Lemma 1, $\tilde{\boldsymbol{\Lambda}}^{-1}_{\tilde{\boldsymbol{\beta}}_0} = O_{P|\mathcal{F}_n}(1)$. Combining this with (A.23) and (A.38)

$$\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1),$$

which implies that

$$\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right). \tag{A.39}$$

From (A.24) of Lemma 3,

$$\tilde{\mathbf{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0}^{-1} - \mathbf{\Lambda}^{-1} = -\mathbf{\Lambda}^{-1}(\tilde{\mathbf{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0} - \mathbf{\Lambda})\tilde{\mathbf{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0}^{-1} = O_{P|\mathcal{F}_n}\left(r^{-1/2}\right). \tag{A.40}$$

From (A.38),(A.39) and (A.40),

$$\begin{aligned}
&\mathbf{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) = \mathbf{\Sigma}^{-1/2}\left\{-\tilde{\mathbf{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0}^{-1}\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta}_t) + o_P(1)\right\} \\
&= -\mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta}_t) - \mathbf{\Sigma}^{-1/2}(\tilde{\mathbf{\Lambda}}_{\tilde{\boldsymbol{\beta}}_0}^{-1} - \mathbf{\Lambda}^{-1})\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta}_t) + o_P(1) \\
&= -\mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta}_t) + O_{P|\mathcal{F}_n}\left(r^{1/2}\right)O_{P|\mathcal{F}_n}\left(r^{-1/2}\right)O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1) \\
&= -\mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^{*}(\boldsymbol{\beta}_t) + O_{P|\mathcal{F}_n}\left(r^{-1/2}\right) + o_P(1).
\end{aligned}$$

The result in Theorem 3 follows from Slutsky's Theorem and the fact that

$$\begin{aligned}
&\mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\left\{\mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}(\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\right\}^T = \mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}\mathbf{\Sigma}_{opt}^{\tilde{\boldsymbol{\beta}}_0}\mathbf{\Lambda}^{-1}\mathbf{\Sigma}^{-1/2} \\
&= \mathbf{\Sigma}^{-1/2}\mathbf{\Lambda}^{-1}\mathbf{\Sigma}_{opt}\mathbf{\Lambda}^{-1}\mathbf{\Sigma}^{-1/2} + O_P(r^{-1}r_0^{-1/2}) \\
&= \boldsymbol{I} + O_P(r^{-1}r_0^{-1/2}),
\end{aligned}$$

which is obtained using (A.35). This means that for any $\boldsymbol{x} \in \mathbb{R}^p$

$$P\left\{\mathbf{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \le \boldsymbol{x}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\} \to \Phi(\boldsymbol{x}), \tag{A.41}$$

in probability. Since the conditional probability is a bounded random variable, convergence in probability to a constant implies convergence to the mean. Therefore, the unconditional probability

$$P\left\{\mathbf{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \le \boldsymbol{x}\right\} = E\left[P\left\{\mathbf{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t) \le \boldsymbol{x}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\right\}\right] \to \Phi(\boldsymbol{x}).$$

This ends the proof. □

References

Ai, M., Wang, F., Yu, J., and Zhang, H. (2021a). Optimal subsampling for large-scale quantile regression. *Journal of Complexity* **62**, 101512.

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021b). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31**, 2, 749–772.

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum Experimental Designs, With SAS*, vol. 34. Oxford University Press.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics* **46**, 3, 1352–1382.

Chen, K., Guo, S., Lin, Y., and Ying, Z. (2010). Least absolute relative error estimation. *Journal of the American Statistical Association* **105**, 491, 1104–1112.

Chen, K., Lin, Y., Wang, Z., and Ying, Z. (2016). Least product relative error estimation. *Journal of Multivariate Analysis* **144**, 91–98.

Chen, X. and Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**, 4, 1655–1684.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research* **13**, 1, 3475–3506.

Han, L., Tan, K. M., Yang, T., and Zhang, T. (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *The Annals of Statistics* **48**, 3, 1770–1788.

Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806* .

Lee, J., Schifano, E. D., and Wang, H. (2021). Fast optimal subsampling probability approximation for generalized linear models. *Econometrics and Statistics* DOI: 10.1016/j.ecosta.2021.02.007.

Li, Z., Lin, Y., Zhou, G., and Zhou, W. (2014). Empirical likelihood for least absolute relative error regression. *Test* **23**, 1, 86–99.

Luo, L. and Song, P. X.-K. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1, 69–97.

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

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. (2020). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In *International Conference on Artificial Intelligence and Statistics*, 1026–1035. PMLR.

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021). Lowcon: A design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics* **30**, 3, 694–708.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics* **58**, 3, 393–403.

Shi, C., Lu, W., and Song, R. (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association* **113**, 524, 1698–1709.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, C., Chen, M.-H., Wu, J., Yan, J., Zhang, Y., and Schifano, E. (2018a). Online updating method with new variables for big data streams. *Canadian Journal of Statistics* **46**, 1, 123–146.

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

Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika* **108**, 1, 99–112.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114**, 525, 393–405.

Wang, H., Zhu, R., and Ma, P. (2018b). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* **113**, 522, 829–844.

Xia, X., Liu, Z., and Yang, H. (2016). Regularized estimation for the least absolute relative error models with a diverging number of covariates. *Computational Statistics and Data Analysis* **96**, 104–119.

Xue, Y., Wang, H., Yan, J., and Schifano, E. D. (2019). An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics* **76**, 1, 171–182.

Yao, Y. and Wang, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science* **19**, 1, 151–172.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2020). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* 1–12. DOI: 10.1080/01621459.2020.1773832.

Zhang, H. and Wang, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics and Data Analysis* **153**, 107072.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021a). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine* **40**, 2, 441–450.

Zuo, L., Zhang, H., Wang, H., and Sun, L. (2021b). Optimal subsample selection for massive logistic regression with distributed data. *Computational Statistics* **36**, 2535–2562.

Table 1. Simulation results on the two-step subsample estimator $\breve{\beta}_1$ with $log(\varepsilon) \sim N(0,1)$.

		OSC				UNIF			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	−0.0012	0.0309	0.0302	0.942	−0.0027	0.0464	0.0428	0.932
	800	−0.0012	0.0261	0.0260	0.944	−0.0004	0.0386	0.0374	0.944
	1000	−0.0001	0.0210	0.0233	0.970	−0.0004	0.0354	0.0337	0.930
Case 2	600	0.0003	0.0308	0.0297	0.936	0.0009	0.0450	0.0429	0.948
	800	−0.0013	0.0260	0.0257	0.940	−0.0003	0.0374	0.0375	0.938
	1000	−0.0012	0.0235	0.0230	0.948	−0.0003	0.0337	0.0338	0.954
Case 3	600	−0.0025	0.0312	0.0313	0.952	−0.0011	0.0430	0.0431	0.956
	800	0.0014	0.0285	0.0272	0.948	−0.0005	0.0374	0.0376	0.942
	1000	0.0006	0.0256	0.0242	0.942	−0.0003	0.0337	0.0339	0.958
Case 4	600	−0.0022	0.0778	0.0764	0.948	−0.0013	0.1006	0.0955	0.928
	800	0.0034	0.0666	0.0661	0.948	−0.0016	0.0798	0.0834	0.964
	1000	0.0056	0.0593	0.0591	0.950	−0.0013	0.0754	0.0754	0.952
		RL				IC			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	−0.0002	0.0465	0.0447	0.928	0.0027	0.0477	0.0454	0.936
	800	−0.0018	0.0378	0.0387	0.962	−0.0030	0.0403	0.0395	0.940
	1000	−0.0001	0.0349	0.0350	0.944	−0.0002	0.0363	0.0356	0.944
Case 2	600	−0.0002	0.0452	0.0443	0.944	−0.0024	0.0476	0.0465	0.948
	800	−0.0001	0.0390	0.0385	0.944	0.0021	0.0425	0.0408	0.942
	1000	0.0006	0.0341	0.0346	0.960	−0.0001	0.0386	0.0368	0.938
Case 3	600	0.0002	0.0473	0.0458	0.940	−0.0031	0.0456	0.0455	0.948
	800	0.0026	0.0408	0.0400	0.942	0.0009	0.0428	0.0400	0.928
	1000	0.0012	0.0359	0.0359	0.956	−0.0024	0.0371	0.0359	0.950
Case 4	600	0.0039	0.1040	0.0990	0.928	−0.0032	0.0898	0.0871	0.938
	800	0.0053	0.0869	0.0865	0.946	−0.0007	0.0765	0.0759	0.954
	1000	−0.0013	0.0769	0.0774	0.950	−0.0027	0.0704	0.0683	0.936

Table 2. Simulation results on the two-step subsample estimator $\breve{\beta}_2$ with $log(\epsilon) \sim N(0,1)$.

		OSC				UNIF			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	0.0010	0.0313	0.0326	0.954	−0.0001	0.0499	0.0491	0.936
	800	0.0003	0.0283	0.0281	0.944	0.0007	0.0424	0.0430	0.944
	1000	−0.0004	0.0252	0.0250	0.946	−0.0016	0.0399	0.0387	0.942
Case 2	600	−0.0017	0.0667	0.0629	0.946	0.0032	0.0931	0.0902	0.952
	800	−0.0012	0.0545	0.0545	0.948	−0.0021	0.0827	0.0794	0.942
	1000	−0.0001	0.0492	0.0486	0.938	−0.0005	0.0767	0.0718	0.922
Case 3	600	0.0002	0.0248	0.0236	0.938	−0.0021	0.0376	0.0374	0.944
	800	−0.0018	0.0201	0.0203	0.964	−0.0003	0.0347	0.0333	0.946
	1000	−0.0009	0.0169	0.0180	0.972	−0.0011	0.0311	0.0299	0.942
Case 4	600	−0.0003	0.0257	0.0265	0.962	0.0012	0.0428	0.0418	0.940
	800	−0.0004	0.0233	0.0229	0.954	−0.0006	0.0385	0.0368	0.936
	1000	−0.0014	0.0203	0.0205	0.948	−0.0007	0.0345	0.0331	0.938
		RL				IC			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	−0.0010	0.0480	0.0466	0.948	−0.0009	0.0491	0.0466	0.934
	800	−0.0022	0.0408	0.0408	0.950	−0.0012	0.0457	0.0408	0.926
	1000	−0.0006	0.0361	0.0367	0.962	0.0003	0.0360	0.0366	0.946
Case 2	600	−0.0058	0.0879	0.0860	0.936	0.0006	0.0867	0.0850	0.950
	800	−0.0029	0.0778	0.0750	0.934	0.0068	0.0771	0.0749	0.940
	1000	0.0008	0.0674	0.0677	0.948	−0.0001	0.0695	0.0673	0.946
Case 3	600	−0.0027	0.0339	0.0336	0.942	−0.0020	0.0339	0.0335	0.944
	800	−0.0012	0.0293	0.0294	0.954	0.0021	0.0297	0.0295	0.944
	1000	−0.0003	0.0262	0.0264	0.948	0.0015	0.0282	0.0263	0.938
Case 4	600	−0.0011	0.0366	0.0379	0.952	−0.0001	0.0406	0.0397	0.944
	800	0.0001	0.0322	0.0331	0.950	0.0002	0.0368	0.0352	0.930
	1000	0.0002	0.0285	0.0298	0.946	0.0004	0.0336	0.0313	0.936

Table 3. Simulation results on the two-step subsample estimator $\breve{\beta}_1$ with $log(\epsilon) \sim Uniform(-2, 2)$.

		OSC				UNIF			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	−0.0015	0.0328	0.0332	0.964	0.0018	0.0386	0.0384	0.950
	800	0.0022	0.0295	0.0287	0.950	0.0007	0.0318	0.0333	0.962
	1000	−0.0014	0.0260	0.0256	0.958	−0.0017	0.0305	0.0298	0.938
Case 2	600	−0.0021	0.0337	0.0328	0.930	−0.0005	0.0393	0.0385	0.948
	800	−0.0020	0.0287	0.0284	0.930	−0.0015	0.0337	0.0333	0.944
	1000	−0.0019	0.0252	0.0254	0.944	−0.0002	0.0297	0.0298	0.956
Case 3	600	0.0004	0.0343	0.0346	0.956	−0.0004	0.0389	0.0385	0.948
	800	−0.0026	0.0294	0.0299	0.956	−0.0013	0.0338	0.0333	0.944
	1000	−0.0003	0.0262	0.0268	0.954	−0.0018	0.0306	0.0298	0.938
Case 4	600	−0.0055	0.0823	0.0842	0.960	−0.0050	0.0855	0.0859	0.954
	800	−0.0031	0.0680	0.0728	0.966	0.0006	0.0759	0.0744	0.948
	1000	−0.0019	0.0630	0.0652	0.958	0.0017	0.0706	0.0664	0.934
		RL				IC			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	0.0006	0.0396	0.0399	0.952	−0.0051	0.0395	0.0405	0.948
	800	−0.0002	0.0359	0.0346	0.930	−0.0035	0.0353	0.0351	0.944
	1000	−0.0012	0.0324	0.0309	0.936	0.0004	0.0328	0.0314	0.938
Case 2	600	−0.0015	0.0389	0.0396	0.954	0.0001	0.0426	0.0419	0.944
	800	−0.0001	0.0336	0.0342	0.958	−0.0007	0.0368	0.0364	0.956
	1000	0.0006	0.0313	0.0306	0.942	0.0008	0.0319	0.0325	0.952
Case 3	600	−0.0004	0.0411	0.0410	0.954	0.0003	0.0408	0.0410	0.952
	800	0.0013	0.0353	0.0354	0.950	−0.0004	0.0364	0.0355	0.942
	1000	−0.0019	0.0341	0.0317	0.938	−0.0001	0.0323	0.0317	0.942
Case 4	600	0.0001	0.0906	0.0885	0.968	0.0024	0.0796	0.0776	0.942
	800	0.0001	0.0748	0.0766	0.960	0.0050	0.0678	0.0673	0.944
	1000	0.0073	0.0674	0.0685	0.948	0.0001	0.0608	0.0601	0.942

Table 4. Simulation results on the two-step subsample estimator $\breve{\beta}_2$ with $log(\epsilon) \sim Uniform(-2, 2)$.

		OSC				UNIF			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	0.0019	0.0378	0.0358	0.940	−0.0032	0.0440	0.0445	0.954
	800	0.0027	0.0321	0.0309	0.938	0.0030	0.0374	0.0385	0.950
	1000	−0.0009	0.0275	0.0275	0.952	−0.0011	0.0357	0.0344	0.954
Case 2	600	−0.0016	0.0667	0.0694	0.950	−0.0027	0.0811	0.0819	0.950
	800	−0.0023	0.0593	0.0600	0.948	−0.0027	0.0707	0.0710	0.958
	1000	0.0008	0.0522	0.0536	0.964	−0.0032	0.0647	0.0636	0.944
Case 3	600	−0.0003	0.0254	0.0257	0.942	0.0001	0.0355	0.0345	0.920
	800	0.0001	0.0222	0.0223	0.964	−0.0005	0.0314	0.0299	0.942
	1000	−0.0005	0.0198	0.0199	0.952	−0.0009	0.0275	0.0267	0.954
Case 4	600	0.0025	0.0281	0.0292	0.964	−0.0014	0.0396	0.0387	0.936
	800	0.0011	0.0246	0.0252	0.970	−0.0014	0.0353	0.0333	0.932
	1000	0.0010	0.0225	0.0225	0.944	−0.0015	0.0301	0.0297	0.958
		RL				IC			
	r	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
Case 1	600	−0.0002	0.0423	0.0421	0.944	0.0009	0.0428	0.0419	0.952
	800	0.0007	0.0363	0.0364	0.962	−0.0008	0.0349	0.0363	0.952
	1000	0.0003	0.0326	0.0325	0.960	0.0012	0.0325	0.0324	0.964
Case 2	600	0.0022	0.0791	0.0774	0.948	0.0076	0.0758	0.0766	0.950
	800	0.0005	0.0679	0.0671	0.942	0.0062	0.0671	0.0664	0.930
	1000	0.0020	0.0601	0.0598	0.946	−0.0055	0.0578	0.0595	0.960
Case 3	600	−0.0008	0.0320	0.0301	0.936	0.0011	0.0313	0.0303	0.940
	800	0.0014	0.0279	0.0261	0.928	−0.0001	0.0251	0.0262	0.954
	1000	−0.0005	0.0241	0.0233	0.948	0.0008	0.0231	0.0234	0.938
Case 4	600	0.0004	0.0359	0.0341	0.954	0.0011	0.0380	0.0360	0.944
	800	0.0020	0.0300	0.0294	0.950	−0.0002	0.0334	0.0313	0.954
	1000	0.0002	0.0264	0.0264	0.950	0.0009	0.0265	0.0280	0.960

Table 5. Simulation results on the two-step subsample estimator $\breve{\beta}_1$ for case 1 with $log(\epsilon) \sim N(0,1)$.

		OSC				UNIF			
	r_0	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
$r = 600$	400	−0.0025	0.0310	0.0302	0.950	0.0005	0.0412	0.0429	0.952
	500	0.0007	0.0303	0.0302	0.946	−0.0014	0.0446	0.0431	0.930
	600	0.0023	0.0291	0.0302	0.948	−0.0001	0.0437	0.0430	0.934
$r = 800$	400	−0.0003	0.0250	0.0261	0.950	−0.0008	0.0387	0.0377	0.954
	500	0.0007	0.0263	0.0261	0.948	−0.0008	0.0402	0.0375	0.936
	600	−0.0005	0.0248	0.0260	0.950	−0.0004	0.0381	0.0375	0.948
$r = 1000$	400	−0.0015	0.0226	0.0233	0.956	−0.0006	0.0348	0.0337	0.948
	500	−0.0001	0.0243	0.0232	0.942	−0.0013	0.0339	0.0337	0.952
	600	−0.0005	0.0214	0.0232	0.972	0.0001	0.0334	0.0336	0.946
		RL				IC			
	r_0	BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
$r = 600$	400	−0.0032	0.0450	0.0447	0.952	−0.0003	0.0467	0.0455	0.948
	500	−0.0048	0.0465	0.0446	0.936	0.0003	0.0453	0.0453	0.950
	600	−0.0001	0.0448	0.0447	0.932	−0.0014	0.0458	0.0453	0.940
$r = 800$	400	−0.0021	0.0377	0.0389	0.956	0.0021	0.0420	0.0395	0.926
	500	−0.0025	0.0386	0.0390	0.952	0.0009	0.0403	0.0394	0.956
	600	−0.0021	0.0377	0.0387	0.960	0.0009	0.0410	0.0395	0.940
$r = 1000$	400	−0.0020	0.0350	0.0350	0.952	0.0015	0.0369	0.0355	0.940
	500	0.0018	0.0369	0.0350	0.940	−0.0015	0.0357	0.0356	0.952
	600	−0.0011	0.0347	0.0349	0.946	−0.0001	0.0333	0.0356	0.960

Table 6. The CPU time for Case 1 with $log(\epsilon) \sim N(0, 1)$ and $r = 1000$ (seconds).

	Methods	$n = 10^6$	$n = 3 \times 10^6$	$n = 5 \times 10^6$	$n = 10^7$
$p = 5$	UNIF	0.009	0.032	0.054	0.107
	OSC	0.173	0.521	0.920	1.792
	RL	1.647	5.281	9.385	19.288
	IC	1.684	5.301	9.411	20.513
	Full data	1.322	4.294	6.577	13.544
$p = 50$	UNIF	0.055	0.072	0.103	0.180
	OSC	0.752	2.116	3.728	7.565
	RL	2.841	8.962	16.371	34.041
	IC	2.893	9.115	16.814	35.356
	Full data	53.594	171.255	275.725	597.980
$p = 100$	UNIF	0.197	0.212	0.254	0.363
	OSC	1.547	4.415	6.918	15.257
	RL	3.742	13.700	22.343	46.529
	IC	3.778	13.819	22.726	47.943
	Full data	244.745	766.650	1353.470	2764.720

Table 7 . The BIAS and (SSE, ESE) for the bike sharing data†.

	$\boldsymbol{\beta}$	OSC	UNIF	RL	IC
$r = 200$	β_1	-0.0120(0.2117, 0.2016)	0.0155 (0.2546, 0.2489)	0.0036 (0.2473, 0.2402)	0.0167 (0.2472, 0.2303)
	β_2	0.0065 (0.0856, 0.0823)	0.0005 (0.0956, 0.0933)	-0.0029 (0.0955, 0.0934)	-0.0045 (0.1013, 0.0998)
	β_3	0.0011 (0.1911, 0.1894)	-0.0166 (0.2487, 0.2381)	-0.0027 (0.2437, 0.2353)	-0.0117 (0.2522, 0.2402)
	β_4	0.0062 (0.2016, 0.1979)	-0.0050 (0.2555, 0.2468)	-0.0036 (0.2470, 0.2384)	-0.0098 (0.2557, 0.2375)
	β_5	-0.0105(0.3275, 0.3146)	-0.0045 (0.4090, 0.3834)	0.0187 (0.3682, 0.3653)	0.0021 (0.3546, 0.3429)
$r = 400$	β_1	0.0045 (0.1411, 0.1381)	-0.0009 (0.1879, 0.1787)	0.0067 (0.1750, 0.1724)	0.0040 (0.1654, 0.1651)
	β_2	-0.0023(0.0556, 0.0561)	0.0023 (0.0673, 0.0667)	0.0037 (0.0675, 0.0666)	0.0005 (0.0721, 0.0717)
	β_3	-0.0163(0.1287, 0.1302)	-0.0035 (0.1806, 0.1701)	$-$ 0.0092 (0.1663, 0.1679)	-0.0034 (0.1705,0.1718)
	β_4	0.0050 (0.1396, 0.1349)	-0.0025 (0.1857, 0.1769)	$-$ 0.0059 (0.1795, 0.1710)	-0.0012 (0.1706, 0.1703)
	β_5	0.0012 (0.2145, 0.2139)	0.0189 (0.2744, 0.2743)	0.0079 (0.2629, 0.2609)	-0.0027 (0.2513, 0.2444)
$r = 600$	β_1	0.0002 (0.1101, 0.1120)	0.0015 (0.1463, 0.1468)	-0.0056 (0.1384, 0.1414)	0.0093 (0.1378, 0.1354)
	β_2	0.0010 (0.0465, 0.0458)	0.0011 (0.0549, 0.0546)	0.0038 (0.0559, 0.0547)	-0.0015 (0.0609, 0.0588)
	β_3	0.0035 (0.1057, 0.1056)	-0.0044 (0.1416, 0.1392)	0.0007 (0.1384, 0.1376)	-0.0068 (0.1442, 0.1404)
	β_4	-0.0045(0.1097, 0.1108)	0.0013 (0.1492, 0.1456)	0.0075 (0.1374, 0.1406)	-0.0027 (0.1409, 0.1397)
	β_5	-0.0087(0.1759, 0.1753)	0.0033 (0.2252, 0.2265)	-0.0006 (0.2054, 0.2137)	-0.0058 (0.2046, 0.2007)

Table 8. The estimator $\breve{\boldsymbol{\beta}}$ and 95% confidence interval with one subsample for the bike sharing data[†].

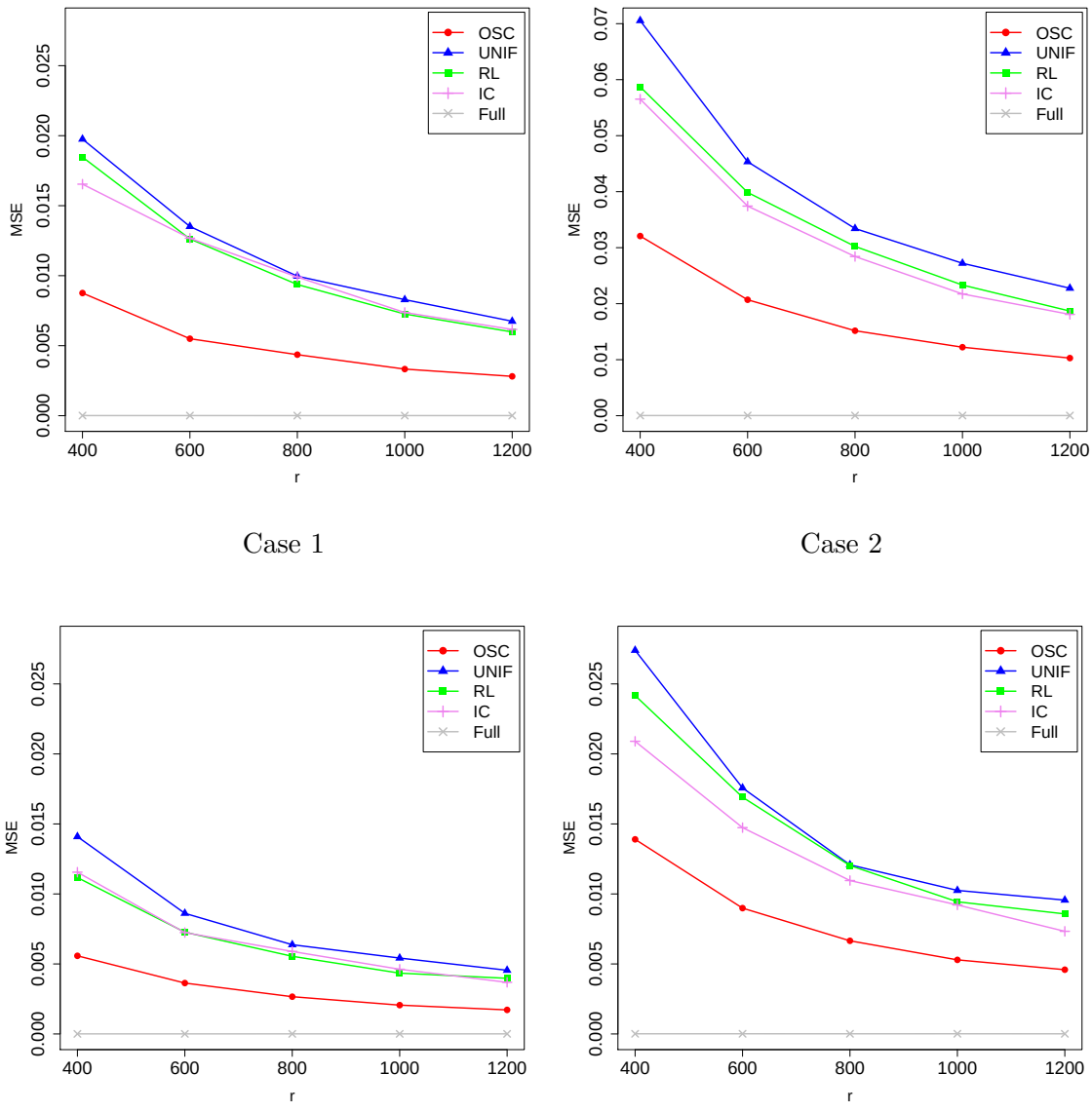
		OSC	UNIF	RL	IC
$r = 200$	β_1	2.3302 (1.9194, 2.7411)	1.8064 (1.2341, 2.3788)	1.6905 (1.2423, 2.1388)	1.8652 (1.4470, 2.2834)
	β_2	−0.0040 (−0.1513, 0.1433)	−0.0777 (−0.2606, 0.1052)	0.0098(−0.1763, 0.1959)	0.0353 (−0.1740, 0.2446)
	β_3	1.3101 (0.9465, 1.6737)	1.5161 (1.0628 , 1.9693)	1.6322 (1.1812, 2.0833)	1.9360 (1.4323, 2.4396)
	β_4	−1.2532 (−1.6564, −0.8499)	−0.6544 (−1.2051, −0.1037)	−0.8136(−1.2845,−0.3428)	−0.9910(−1.4739,−0.5082)
	β_5	0.0706 (−0.5870, 0.7282)	0.6645 (−0.1350, 1.4640)	1.2600 (0.5430, 1.9770)	0.1268 (−0.5919, 0.8455)
$r = 400$	β_1	2.1634 (1.9211, 2.4056)	2.1057 (1.7538, 2.4575)	2.2573 (1.9215, 2.5931)	2.0417 (1.7383, 2.3452)
	β_2	−0.0866 (−0.1948, 0.0215)	−0.0163 (−0.1513, 0.1188)	0.0147 (−0.1256, 0.1550)	−0.0682 (−0.2072, 0.0708)
	β_3	1.6468 (1.3828, 1.9107)	1.5091 (1.2018 , 1.8164)	1.4956 (1.1689, 1.8223)	1.5391 (1.1956 , 1.8827)
	β_4	−1.1452 (−1.4006, −0.8898)	−0.9813 (−1.3255, −0.6371)	−1.0674(−1.3995,−0.7353)	−0.9503(−1.2744,−0.6262)
	β_5	0.4386 (0.0409, 0.8364)	−0.0045 (−0.5489, 0.5399)	−0.3826 (−0.9057, 0.1405)	0.4234 (−0.0118, 0.8586)
$r = 600$	β_1	2.1932 (1.9773, 2.4090)	2.2479 (1.9409, 2.5549)	2.1530 (1.8962, 2.4098)	2.3682 (2.0976, 2.6389)
	β_2	−0.1071 (−0.2046, −0.0097)	−0.0416 (−0.1508, 0.0677)	−0.0443(−0.1518, 0.0633)	−0.0019 (−0.1155, 0.1118)
	β_3	1.4858 (1.2619, 1.7098)	1.5013 (1.2390 , 1.7635)	1.4236 (1.1502, 1.6970)	1.3914 (1.1247, 1.6580)
	β_4	−1.0288 (−1.2651, −0.7925)	−1.1969 (−1.5060, −0.8878)	−1.2204(−1.4703,−0.9706)	−1.2941(−1.5834,−1.0048)
	β_5	0.1637 (−0.1590, 0.4865)	0.1016 (−0.3361, 0.5394)	0.7835 (0.3905, 1.1765)	−0.0710 (−0.4378, 0.2959)

Table 9. The BIAS and (SSE, ESE) for the electric power consumption data†.

	$\boldsymbol{\beta}$	OSC	UNIF	RL	IC
$r = 600$	β_1	0.0006 (0.0191, 0.0186)	0.0016 (0.0262, 0.0258)	0.0009 (0.0283, 0.0288)	0.0015 (0.0286, 0.0290)
	β_2	−0.0001(0.0102, 0.0100)	0.0016 (0.0252, 0.0226)	0.0007 (0.0124, 0.0124)	0.0007 (0.0126, 0.0124)
	β_3	0.0019 (0.0134, 0.0125)	0.0053 (0.0304, 0.0254)	0.0005 (0.0137, 0.0134)	0.0017 (0.0140, 0.0134)
	β_4	0.0017 (0.0171, 0.0165)	0.0010 (0.0214, 0.0211)	0.0004 (0.0212, 0.0223)	−0.0008 (0.0221,0.0223)
$r = 800$	β_1	0.0010 (0.0158, 0.0158)	−0.0001(0.0234, 0.0224)	0.0001 (0.0249, 0.0249)	−0.0008 (0.0252, 0.0251)
	β_2	0.0003 (0.0087, 0.0088)	0.0011 (0.0223, 0.0199)	0.0006 (0.0109, 0.0108)	0.0001 (0.0109, 0.0108)
	β_3	0.0007 (0.0094, 0.0090)	0.0044 (0.0264, 0.0225)	0.0009 (0.0121, 0.0117)	0.0007 (0.0120, 0.0117)
	β_4	0.0020 (0.0145, 0.0140)	0.0008 (0.0192, 0.0183)	0.0009 (0.0197, 0.0193)	0.0010 (0.0198, 0.0194)
$r = 1000$	β_1	0.0002 (0.0145, 0.0145)	−0.0005(0.0204, 0.0201)	−0.0007(0.0226, 0.0225)	−0.0015 (0.0225, 0.0225)
	β_2	0.0003 (0.0076, 0.0075)	0.0024 (0.0193, 0.0180)	0.0008 (0.0098, 0.0097)	0.0008 (0.0101, 0.0097)
	β_3	0.0003 (0.0071, 0.0073)	0.0042 (0.0235, 0.0205)	0.0007 (0.0111, 0.0104)	0.0006 (0.0108, 0.0104)
	β_4	0.0001 (0.0127, 0.0127)	0.0012 (0.0163, 0.0164)	0.0001 (0.0178, 0.0173)	0.0005 (0.0165, 0.0173)

Table 10. Estimator $\breve{\boldsymbol{\beta}}$ and 95% confidence interval with one subsample for the electric power consumption data[†].

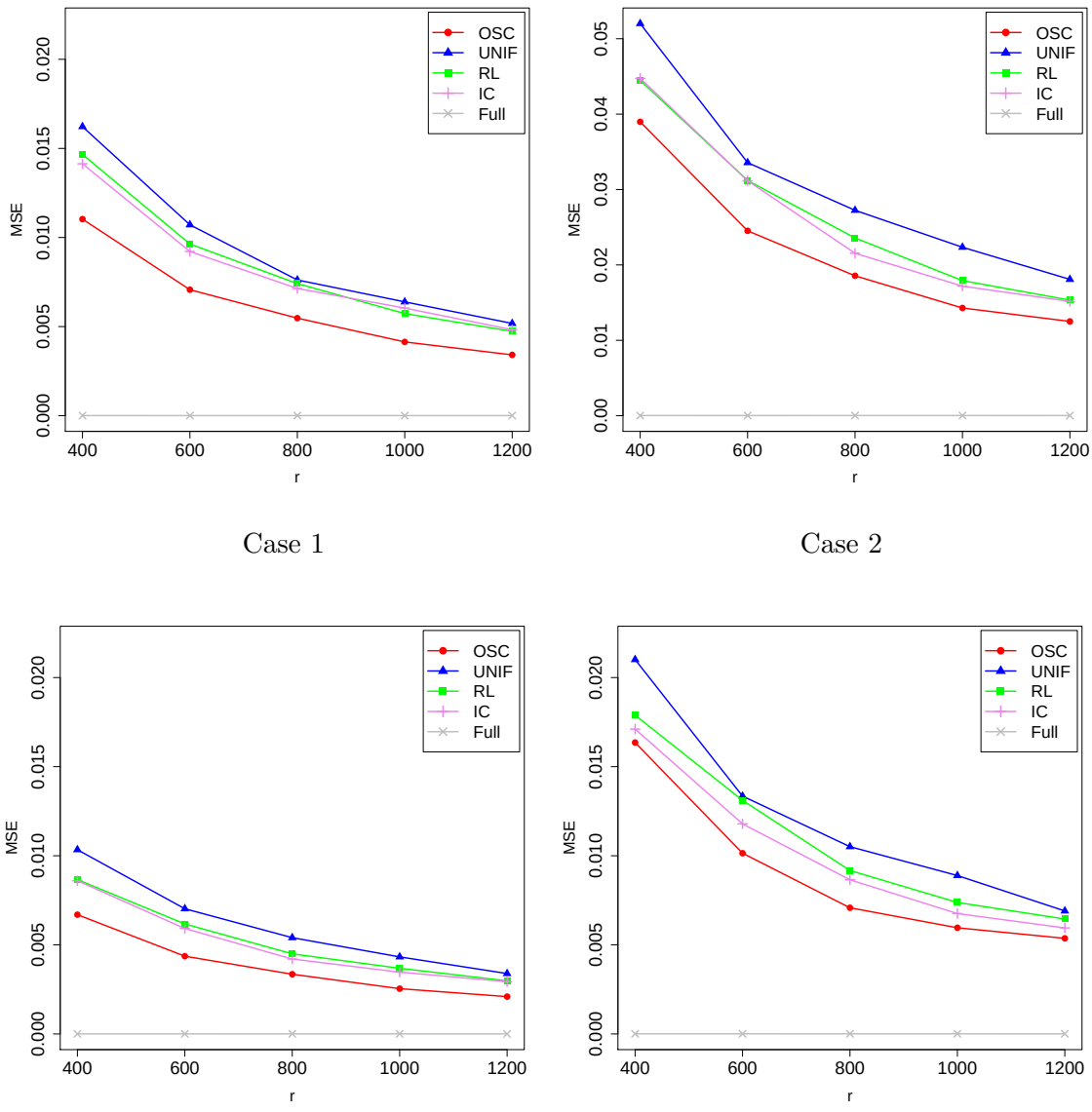
		OSC	UNIF	RL	IC
$r = 600$	β_1	1.1485 (1.1129, 1.1842)	1.1229 (1.0750, 1.1708)	1.0844 (1.0261, 1.1427)	1.1277 (1.0665, 1.1889)
	β_2	0.2287 (0.2075, 0.2499)	0.2119 (0.1554, 0.2685)	0.2187 (0.1940, 0.2434)	0.2034 (0.1791, 0.2277)
	β_3	0.1921 (0.1751, 0.2090)	0.1712 (0.1380, 0.2044)	0.1943 (0.1676, 0.2210)	0.2099 (0.1843, 0.2354)
	β_4	0.6261 (0.5949, 0.6574)	0.6532 (0.6138, 0.6925)	0.6575 (0.6119, 0.7031)	0.6273 (0.5811, 0.6734)
$r = 800$	β_1	1.1203 (1.0874, 1.1533)	1.0882 (1.0407, 1.1358)	1.1500 (1.1046, 1.1954)	1.1102 (1.0642, 1.1561)
	β_2	0.2088 (0.1890, 0.2285)	0.2208 (0.1820, 0.2596)	0.2118 (0.1943, 0.2294)	0.2047 (0.1840, 0.2254)
	β_3	0.1953 (0.1783, 0.2122)	0.1909 (0.1467, 0.2352)	0.1960 (0.1744, 0.2176)	0.1972 (0.1759, 0.2185)
	β_4	0.6347 (0.6064, 0.6630)	0.6575 (0.6192, 0.6959)	0.6123 (0.5764, 0.6482)	0.6287 (0.5934, 0.6640)
$r = 1000$	β_1	1.1264 (1.0980, 1.1549)	1.1191 (1.0797, 1.1584)	1.0799 (1.0363, 1.1236)	1.0890 (1.0490, 1.1290)
	β_2	0.2244 (0.2073, 0.2414)	0.2178 (0.1709, 0.2648)	0.2183 (0.1997, 0.2369)	0.2338 (0.2146, 0.2530)
	β_3	0.2188 (0.2025, 0.2350)	0.2145 (0.1779, 0.2511)	0.2087 (0.1880, 0.2294)	0.2117 (0.1906, 0.2329
	β_4	0.6257 (0.6007, 0.6507)	0.6447 (0.6128, 0.6765)	0.6666 (0.6328, 0.7004)	0.6420 (0.6105, 0.6735)



Case 1 Case 2

Case 3 Case 4

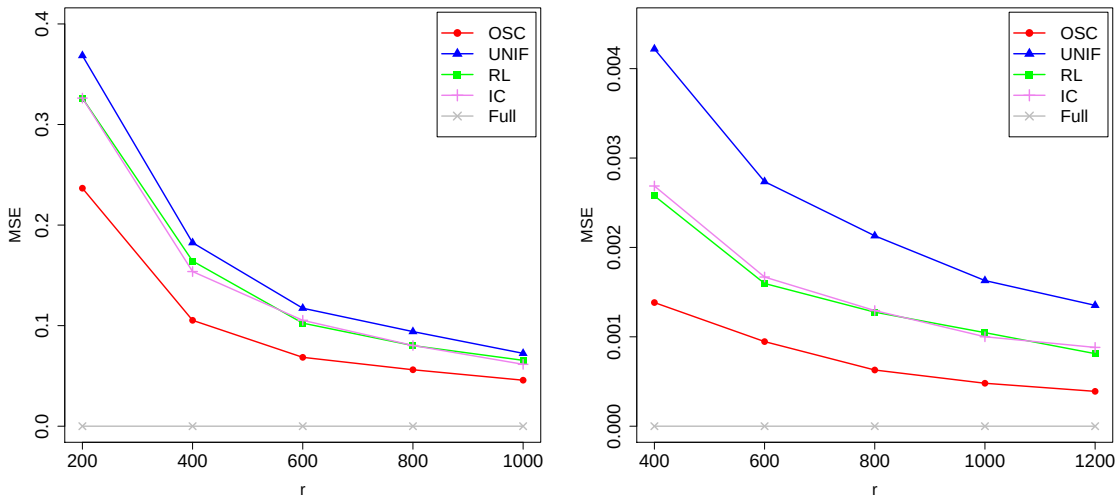
Figure 1. The MSEs for different subsampling probabilities with $log(\varepsilon) \sim N(0, 1)$.



Case 1 Case 2

Case 3 Case 4

Figure 2. The MSEs for different subsampling probabilities with $log(\varepsilon) \sim Uniform(-2, 2)$.



(a) The bike sharing data (b) The electric power consumption data

Figure 3. The results of MSEs in the real data analysis.